“I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust

Sunnie S. Y. Kim*
sunniesuhyoung@princeton.edu
Princeton University
USA

Q. Vera Liao
veraliao@microsoft.com
Microsoft
Canada

Mihaela Vorvoreanu
mihaela.vorvoreanu@microsoft.com
Microsoft
USA

Stephanie Ballard
stephballard@microsoft.com
Microsoft
USA

Jennifer Wortman Vaughan
jenn@microsoft.com
Microsoft
USA

This is the appendix for the paper:

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. “I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24), June 3–6, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA. https://doi.org/10.1145/3630106.3658941

APPENDIX

The appendix is structured in the following way.

- **Appendix A:** Participant Demographics and Background
- **Appendix B:** Data Collection and Exclusion
- **Appendix C:** Additional Analyses
  - **Appendix C.1:** Exploration of `LinkClick` and `UseLink`
  - **Appendix C.2:** Internal Consistency of Scales Used to Measure Trust and Perception of AI
- **Appendix D:** Full Wording Used in the Experiment
- **Appendix E:** Questions and AI Responses Used

A PARTICIPANT DEMOGRAPHICS AND BACKGROUND

In the exit questionnaire, participants optionally self-reported their gender, age, race, ethnicity, and level of education. Of 404 participants in the final sample, 51.7% identified as woman, 46.8% as man, and 0.5% as non-binary. The age distribution was as follows: 18–24 (5.2%), 25–34 (32.4%), 35–44 (35.1%), 45–54 (14.4%), 55-64 (9.7%), 65–74 (2.7%), and 75–84 (0.2%). Regarding race, 82.4% identified as white alone, followed by Asian alone (6.7%), Black or African American alone (5.9%), two or more races (2.0%), and American Indian and Alaska Native alone (1.0%). Regarding ethnicity, 12.4% identified as Hispanic or Latino, and 86.9% as not. For the highest level of education completed, 58.7% said 4 year degree, followed by professional degree (13.6%), high school graduate (9.4%), some college (7.9%), 2 year degree (7.7%), doctorate (1.7%), and less than high school (0.2%).

In Table A1, we compare the demographics of our study sample (N=404) with U.S. Census data [11, 12]. Compared to the U.S. population, our sample is younger and more educated, a common characteristic of samples recruited through online crowdsourcing platforms [6, 7, 9, 10]. Further, white respondents are over-represented, while Black and Hispanic/Latino respondents are under-represented.

In the exit questionnaire, we also measured the following variables based on participants’ self-reported ratings on a 5-point scale:

- `LLMFamiliarity`: Rating on the question “How familiar are you with LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”
- `LLMUse`: Rating on the question “How often do you use LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”
- `LLMAttitude`: Rating on the question “Overall, how do you feel about LLMs and LLM-infused applications such as OpenAI’s ChatGPT, Microsoft’s Bing Chat, and Google’s Bard?”

The mean and standard deviation of `LLMFamiliarity` are 3.0 ± 1.0 around “3: Moderately familiar, I know what they are and can explain.” For `LLMUse`, they are 2.8 ± 1.2 between “2: Rarely, about 1–2 times a month” and “3: Sometimes, about 3–4 times a month.” For `LLMAttidue`, they are 3.8 ± 1.0 between “3: Neutral” and “4: Somewhat positive.” This result consistent with that of Koonchanok et al. [8] who investigated public attitudes towards ChatGPT by analyzing Twitter data from December 2022 to March 2023 and found that the overall sentiment is largely neutral to positive.

B DATA COLLECTION AND EXCLUSION

In this section, we provide more details about our pre-registered data collection procedures and exclusion criteria.[3]

Sample size. We determined our target sample size (N=432) with an a priori power analysis using the statistical software G*Power [3, 4]. According to G*Power’s calculation, a conservative Wilcoxon-Mann-Whitney two-tailed test requires 90 participants per condition to detect medium-sized effects (Cohen’s d = 0.5) at the significance level of $\alpha = 0.05$ with power (1-β) = 0.90. Since we have four conditions, the required sample size is 360. We increased the number by 20% to 432 to account for possible exclusions.

Data collection plan. We planned to recruit U.S.-based participants from Amazon Mechanical Turk (MTurk) with a Masters qualification, 99% or higher approval rating, and at least 2000 completed human intelligence tasks (HITs). If we are unable to meet

---

[*] Most work done during an internship at Microsoft.

[3] Our pre-registration is viewable at https://osf.io/mnrp9.

Table A1: Demographics of our study sample (N=404) compared to the U.S. population. N/A indicates there is no data for the given category. * indicates multiple selection and self-description were possible.

Demographic Attribute		Sample	Census
Gender*	Woman	51.7%	N/A
	Man	46.8%	N/A
	Non-binary	0.5%	N/A
	Gender-diverse	0.0%	N/A
	Transgender	0.0%	N/A
	Prefer not to answer	1.0%	N/A
Sex	Female	N/A	50.9%
	Male	N/A	49.1%
Age	18–24	5.2%	9.4%
	25–34	32.4%	13.5%
	35–44	35.1%	12.7%
	45–54	14.4%	12.4%
	55–64	9.7%	13.1%
	65–74	2.7%	10.0%
	75–84	0.2%	4.9%
	85 or older	0.0%	1.9%
	Prefer not to answer	0.2%	N/A
Race*	White alone	82.4%	61.6%
	Black or African American alone	5.9%	12.4%
	American Indian or Alaska Native alone	1.0%	1.1%
	Asian alone	6.7%	6.0%
	Native Hawaiian or Other Pacific Islander alone	0.0%	0.2%
	Some Other Race	N/A	8.4%
	Two or More Races	2.0%	10.2%
	Prefer not to answer	1.7%	N/A
Ethnicity	Hispanic or Latino	12.4%	18.7%
	Not Hispanic or Latino	86.9%	81.3%
	Prefer not to answer	0.7%	N/A
Educational attainment	Less than high school	0.2%	9.6%
	High school graduate	9.4%	29.2%
	Some college	7.9%	16.5%
	2 year degree (Associate's degree in the Census)	7.7%	9.9%
	4 year degree (Bachelor's degree in the Census)	58.7%	22.1%
	Master's degree	N/A	9.5%
	Professional degree	13.6%	1.4%
	Doctorate	1.7 %	1.9%
	Prefer not to answer	0.7%	N/A

our target sample size in 7 days, we planned to remove the Masters qualification and release additional HITs in batches of size 50, checking the number of non-excluded data points after each batch, until we have hit a total of at least 360 non-excluded responses.

Data collection. We collected data over the course of two weeks in September 2023. From September 2 to 8, we collected 101 complete responses with the Masters qualification requirement. Since we were not able to recruit the target sample size of 432 in 7 days, following our pre-registered plan, we then removed the Masters qualification requirement and released additional HITs in batches of size 50, checking the number of non-excluded data points after each batch, until we hit a total of at least 360 non-excluded responses. Concretely, we collected 555 complete responses from September 8 to 15 without the Masters qualification requirement.

Data exclusion. In total, we collected 656 complete responses and excluded 252 (38.4%) — 5 from the 101 (5.0%) collected with the Masters qualification requirement and 247 from the 555 (44.5%) collected without the requirement — that satisfied any of the following criteria: (1) with answers to honeypot questions that are invisible to people but visible to bots; (2) with the same answers to all tasks; (3) with response time under 5 minutes; (4) with less than 75% accuracy on the post-task attention check where participants are shown a list of questions and asked to indicate which ones they answered and didn't during the task; and (5) with free-form answers that are clearly off-topic (e.g., answers are irrelevant, gibberish, spam) or identical across multiple responses. Note that (1)–(4) can be assessed automatically, while (5) requires a manual review. When determining whether a free-form answer was off-topic for (5), we did not consider the condition or answers to any other questions in the experiment.

In total, 1 response was flagged for (1), 86 responses for (2), 21 responses for (3), 180 responses for (4), and 101 (56 off-topic and 45 identical) responses for (5). Oftentimes, the same response was flagged for multiple criteria. In total, 205 responses were excluded based on (1)–(4), and 47 responses were additionally excluded based on (5). While time-consuming, we found manually reviewing free-form answers for (5) to be an effective data quality control measure. With it, we were able to exclude responses that seem to be from bots, and responses that seem to be from the same person using multiple MTurk accounts. (We set up our HITs such that each MTurk account can participate only once. Nonetheless, we saw responses from different MTurk accounts sometimes contain identical free-form answers.)

Examples of free-form answers that were off-topic include:

- "I cannot directly access or analyze the responses of other AI systems, including 'Al system A,' so I do not have insight into when or why it might express uncertainty."
- "The term AI, coined in the 1950s, refers to the simulation of human intelligence by machines."
- "Humans use their brains' memory, processing power, and cognitive abilities."

Examples of free-form answers that were identical across multiple responses from different MTurk accounts include:

- "I have given the AI's given answers by burning them into consideration." (appeared 3 times)
- "During some questions, the AI does not know the answer surely and the reason for its uncertainly is that it isn't updated to latest information." (appeared 4 times)
- "It is very useful and make my decision. I am only concentrate this task and use the AI's answer and make my decision." (appeared 3 times)

Manipulation check. According to our manipulation check in the exit questionnaire, 90.2% of UNCERTAIN1ST and 88.3% of UNCERTAINGENERAL participants answered that they noticed the AI system's uncertainty expression.

C ADDITIONAL ANALYSES

C.1 Exploration of `LinkClick` and `UseLink`

For the within-condition analysis, we pre-registered that we would fit the model `DV ~ AIUncertain + (1|participant) + (1|question)`. However, when we fit this model for `LinkClick` and `UseLink` looking at data from UNCERTAINGENERAL, the model does not fit properly (has a singular fit) due to large individual variance. For example, according to our tracking, 50 out of 94 participants in UNCERTAINGENERAL never clicked on any link, 17 participants clicked on one or more links for all eight tasks, and 27 participants clicked on links for some tasks and not for others. To explore how these variables differ between instances with AI responses that are not uncertain vs. uncertain, we fit a simpler model with the random effects removed: `DV ~ AIUncertain`. For `LinkClick`, the estimated means (and standard errors) from the model are 32.4% (2.4) vs. 33.0% (2.4). For `UseLink`, they are 65.2% (2.5) vs. 69.9% (2.4). The differences are not significant for both variables. This result is consistent with the result for UNCERTAIN1ST in Table 2 of the main paper.

C.2 Internal Consistency of Scales Used to Measure Trust and Perception of AI

The four DVs related to trust and perceptions of AI — `TrustBelief`, `TrustIntention`, `Anthropomorphism`, and `Transparency` — are indexes calculated based on participants' ratings on multi-item scales. In this section, we assess the internal consistency of the used scales by calculating Cronbach's alpha [2] for each variable. Cronbach's alpha is a popular measure of internal consistency, or reliability, of a scale. It ranges between 0 and 1 with higher values indicating higher internal consistency. Values over 0.8 are deemed good. The Cronbach's alpha values were 0.83 for `TrustBelief`, 0.85 for `TrustIntention`, 0.85 for `Anthropomorphism`, and 0.21 for `Transparency`. Recall that `TrustBelief`, `TrustIntention`, and `Anthropomorphism` were measured with 6, 4, and 4 items adapted from existing scales [1, 5]. The high Cronbach's alpha values suggest these items have good internal consistency. On the other hand, `Transparency` had a low Cronbach's alpha value. This may be because it was measured with 2 items that are not from an existing, tested scale. (To the best of our knowledge, there are no established scales for measuring perceived transparency of algorithmic systems.) For future researchers interested in using our 2 items, we suggest reporting their results separately. The item-level results are presented in Table C2.

D FULL WORDING USED IN THE EXPERIMENT

We show the full wording used in the experiment. If questions and descriptions varied across conditions, we indicated in the heading. We also added annotations in [brackets] to aid understanding.

About the study

CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL

Imagine you have a question that you don't know the answer to. You have access to an AI system that is connected to the internet and can answer a wide range of questions, so you asked the question and received an answer. What will you do next?

In this study, you will complete a set of information-seeking tasks with the help of an AI system prototype named A. Each task involves determining the correct answer to the presented question. You will see AI system A's answer to the question. If you want, you can also use other resources (e.g., internet, books, friends and family).

You will complete 8 tasks in total. After all tasks, you will complete an exit questionnaire that asks about your experience and perception of the AI system, demographics, and other background. In total, this study will take around 20 minutes to complete, with the exit questionnaire taking 5–7 minutes. Please complete everything in one sitting.

NO-AI

In this study, you will complete a set of information-seeking tasks. Each task involves determining the correct answer to the presented question. You may use any resource (e.g., internet, books, friends and family) to complete the tasks.

Table C2: Item-level results for `TrustBelief`, `TrustIntention`, `Anthropomorphism`, and `Transparency`. We report estimated means (and standard errors) from an ANOVA. See Appendix C.2 for the internal consistency of the used scales.

[†] These items are reverse-coded.

DV (Scale: 1-5)	Control	Uncertain1st	UncertainGeneral
`TrustBelief` (6 items)	3.90 (0.06)	3.86 (0.07)	4.00 (0.07)
AI system A is competent and effective in answering questions	4.06 (0.09)	3.83 (0.09)	3.98 (0.09)
Overall, AI system A is a capable and proficient information provider	3.91 (0.09)	3.71 (0.10)	4.00 (0.10)
I would characterize AI system A as honest	3.80 (0.09)	3.85 (0.09)	3.90 (0.09)
AI system A is NOT truthful in providing information to me[†]	3.45 (0.11)	3.60 (0.11)	3.85 (0.11)
I believe that AI system A was developed to act in my (or the users') best interest	3.99 (0.09)	3.92 (0.09)	4.00 (0.09)
AI system A was developed with good intentions to do its best to help me (or the users)	4.21 (0.07)	4.23 (0.08)	4.26 (0.08)
`TrustIntention` (4 items)	3.25 (0.10)	2.91 (0.10)	3.36 (0.10)
When an important question arises, I would NOT feel comfortable depending on the information provided by AI system A[†]	2.75 (0.12)	2.36 (0.12)	3.06 (0.12)
I can always rely on AI system A to find information	3.46 (0.12)	2.97 (0.12)	3.32 (0.12)
I would feel comfortable acting on the information given to me by AI system A	3.47 (0.12)	3.30 (0.13)	3.74 (0.13)
I would not hesitate to use the information AI system A supplied me	3.32 (0.12)	3.00 (0.13)	3.33 (0.12)
`Anthropomorphism` (4 items)	3.07 (0.10)	3.00 (0.11)	3.13 (0.11)
Fake – 1 – 2 – 3 – 4 – 5 – Natural	3.60 (0.10)	3.36 (0.10)	3.57 (0.10)
Machinelike – 1 – 2 – 3 – 4 – 5 – Humanlike	2.89 (0.12)	3.02 (0.13)	2.99 (0.13)
Unconscious – 1 – 2 – 3 – 4 – 5 – Conscious	2.99 (0.13)	2.86 (0.14)	3.13 (0.14)
Artificial – 1 – 2 – 3 – 4 – 5 – Lifelike	2.79 (0.13)	2.76 (0.14)	2.84 (0.14)
`Transparency` (2 items)	4.04 (0.06)	3.93 (0.07)	4.01 (0.07)
I feel I had a good understanding of what AI system A's answers were based on	4.29 (0.07)	4.17 (0.07)	4.34 (0.07)
I feel I had a good understanding of when AI system A's answers might be wrong	3.80 (0.10)	3.68 (0.10)	3.68 (0.10)

You will complete 8 tasks in total. After all tasks, you will complete an exit questionnaire that asks about your demographics and other background. In total, this study will take around 20 minutes to complete, with the exit questionnaire taking 5–7 minutes. Please complete everything in one sitting.

About the AI system

Control, Uncertain1st, UncertainGeneral

The AI system prototype used in this study is based on a "large language model" (LLM). When asked questions or given instructions, LLMs can generate human-like responses. These generated responses can sound convincing and fluent, but may not always be correct. LLMs can be used for many tasks, including question answering, text summarization, creative writing, and programming. You may have used or heard about some popular applications using LLMs, such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard. Some are connected to the Internet, while others are not.

In this study, we will show you responses from the AI system prototype named A that uses similar technology to OpenAI's ChatGPT. AI system A is connected to the internet and can answer a wide range of questions. It also sometimes provides sources for its answers. However, as with all AI systems, its responses, including the provided sources, may be inaccurate, incomplete, and inconsistent, even when they sound convincing.

Task example

[In Figure D1, we show the task example page shown for participants in the `Control`, `Uncertain1st`, `UncertainGeneral` conditions. Participants in the `No-AI` condition were only shown the task question and slightly different set of survey questions. See the Task section for details.]

Task comprehension questions

CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL

Q. Recall what you read earlier about this study and AI system A, and select TRUE or FALSE for each statement. (Options: TRUE, FALSE)

- AI system A is connected to the internet and sometimes provides sources in clickable links.
- AI system A uses technology similar to OpenAI's ChatGPT.
- You may not leave this Qualtrics survey page during the study.
- You will complete 8 information-seeking tasks in total.

These are the correct answers. Please review them before proceeding.

- (TRUE) AI system A is connected to the internet and sometimes provide sources in clickable links.
- (TRUE) AI system A uses technology similar to OpenAI's ChatGPT.
- (FALSE) You may not leave this Qualtrics survey page during the study.
  → You may leave the page to read sources in AI system A's answers or use another resource to complete the information-seeking tasks.
- (TRUE) You will complete 8 information-seeking tasks in total.

You will now proceed to the 8 information-seeking tasks. Please complete them in one sitting, and please do your best! You will be asked questions about them at the end.

NO-AI

Q. Recall what you read earlier about this study, and select TRUE or FALSE for each statement.

- You may not leave this Qualtrics survey page during the study.
- You will complete 8 information-seeking tasks in total.

These are the correct answers. Please review them before proceeding.

- (FALSE) You may not leave this Qualtrics survey page during the study.
  → You may leave the page to use another resource to complete the information-seeking tasks.
- (TRUE) You will complete 8 information-seeking tasks in total.

You will now proceed to the 8 information-seeking tasks. Please complete them in one sitting, and please do your best! You will be asked questions about them at the end.

Task (repeated 8 times)

[All participants answered the same 8 questions. However, participants in different conditions were provided with different AI responses (e.g., with vs. without uncertainty) or not provided with an AI response. In Figure D2, we show the study stimuli for one of the tasks. The full task questions and AI responses used are in Appendix E. Below we show the questions that participants answered after seeing the study stimuli.]

CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL

Q. How confident are you in AI system A's answer?

- Not confident at all
- Not very confident
- Neither
- Fairly confident
- Very confident

Q. What is your final answer to the question: [task question]?

- Yes
- No

Q. How confident are you in your final answer?

- Not confident at all
- Not very confident
- Neither
- Fairly confident
- Very confident

Q. What is your final answer based on? (Select all that apply)

- AI system A's answer
- Your own reading of the linked sources in AI system A's answer
- Your own knowledge
- Your own internet search
- Other (please specify):

NO-AI

Q. What is your final answer to the question: [task question]?

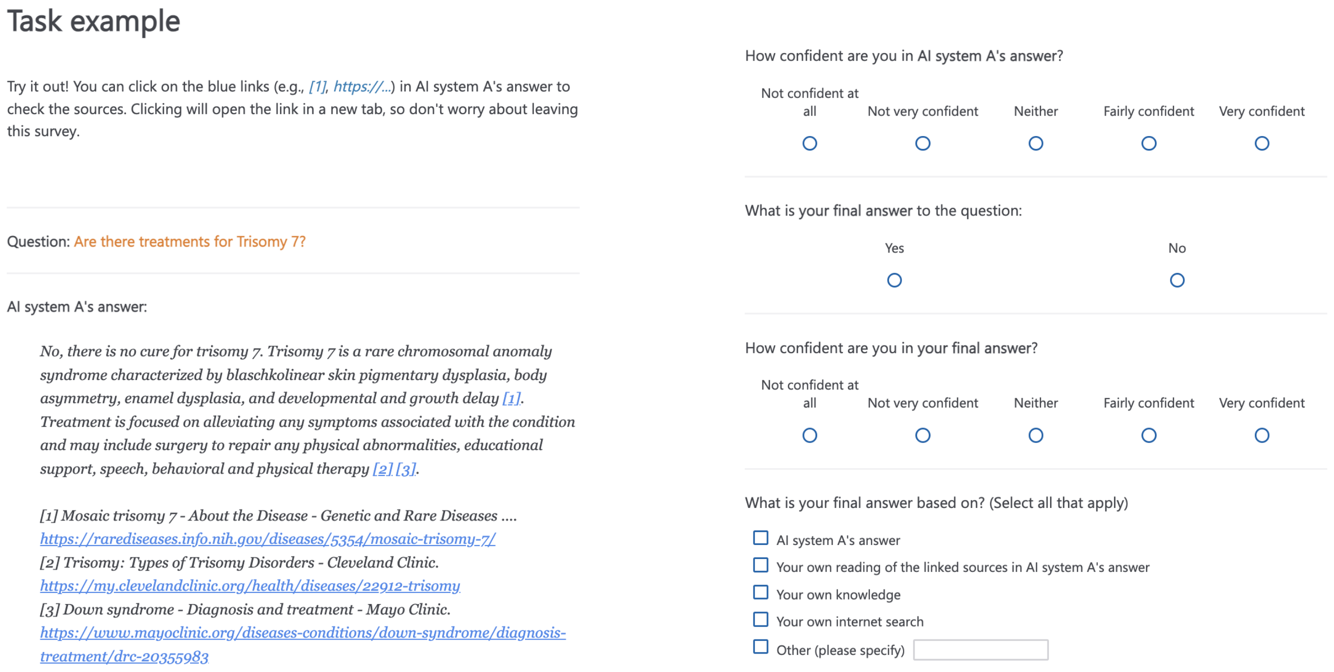
- Yes
- No

Q. How confident are you in your final answer?

- Not confident at all
- Not very confident
- Neither
- Fairly confident
- Very confident

Q. What is your final answer based on? (Select all that apply)

- AI system A's answer
- Your own reading of the linked sources in AI system A's answer
- Your own knowledge



Task example

Try it out! You can click on the blue links (e.g., *[1]*, *https://...*) in AI system A's answer to check the sources. Clicking will open the link in a new tab, so don't worry about leaving this survey.

Question: Are there treatments for Trisomy 7?

AI system A's answer:

No, there is no cure for trisomy 7. Trisomy 7 is a rare chromosomal anomaly syndrome characterized by blaschkolinear skin pigmentary dysplasia, body asymmetry, enamel dysplasia, and developmental and growth delay [1]. Treatment is focused on alleviating any symptoms associated with the condition and may include surgery to repair any physical abnormalities, educational support, speech, behavioral and physical therapy [2] [3].

[1] Mosaic trisomy 7 - About the Disease - Genetic and Rare Diseases ....
https://rarediseases.info.nih.gov/diseases/5354/mosaic-trisomy-7/
[2] Trisomy: Types of Trisomy Disorders - Cleveland Clinic.
https://my.clevelandclinic.org/health/diseases/22912-trisomy
[3] Down syndrome - Diagnosis and treatment - Mayo Clinic.
https://www.mayoclinic.org/diseases-conditions/down-syndrome/diagnosis-treatment/drc-20355983

How confident are you in AI system A's answer?

Not confident at all	Not very confident	Neither	Fairly confident	Very confident
○	○	○	○	○

What is your final answer to the question:

Yes	No
○	○

How confident are you in your final answer?

Not confident at all	Not very confident	Neither	Fairly confident	Very confident
○	○	○	○	○

What is your final answer based on? (Select all that apply)

☐ AI system A's answer
☐ Your own reading of the linked sources in AI system A's answer
☐ Your own knowledge
☐ Your own internet search
☐ Other (please specify)

Figure D1: Task example page shown for participants in the `Control`, `Uncertain1st`, `UncertainGeneral` conditions.

- Your own internet search
- Other (please specify):

Task attention check

Thank you for completing all the tasks! We're curious how well you remember the questions you answered. For each question, indicate whether you answered it or didn't.
(Options: Answered, Didn't answer)

- Did Dupilumab receive FDA approval for Asthma before Chronic Rhinosinustis?
- Is there more antihistamine in Benadryl than Rhinocort?
- Is Deep Vein Thrombosis a common side effect of Ocella?
- Is Spironolactone an FDA-approved drug for treating acne?
- Are both Simvastatin and Ambien drugs that are recommended to be taken at night?
- Is Uveitis a common symptom of Ankylosing Spondylitis?
- Is fever a common symptom of Jock Itch?
- Can an adult who has not had chickenpox get shingles?
- Comparing single tablets, does Tylenol Regular Strength Liquid Gels contain more Acetaminophen and Dextromethorphan than DayQuil Cold and Flu Liquicaps?
- Can an adult who has had measles get SSPE?
- Are there known drug-drug interactions between Gabapentin and Hydrocodone?
- Comparing single tablets, does Pepto-Bismol contain more bismuth subsalicylate than Imodium?
- Are Bisphosphonates the most common drug treatment for Osteoporosis?
- Is there a FDA-approved drug for treating peanut allergy in adults 18 years or older?
- Is tonsillectomy the most common surgical treatment for sleep apnea?
- Are there known drug-drug interactions between Phentermine and Dicyclomine?

[The correct answer is "Answered" for the first 8 questions, and "Didn't answer" for the latter 8 questions. For each participant, we randomly selected 8 questions from the list and showed them in random order.]

Open-ended questions

Welcome to the final part of the study! This questionnaire will take 5–7 minutes, and you will receive your payment code at the end.

Control

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

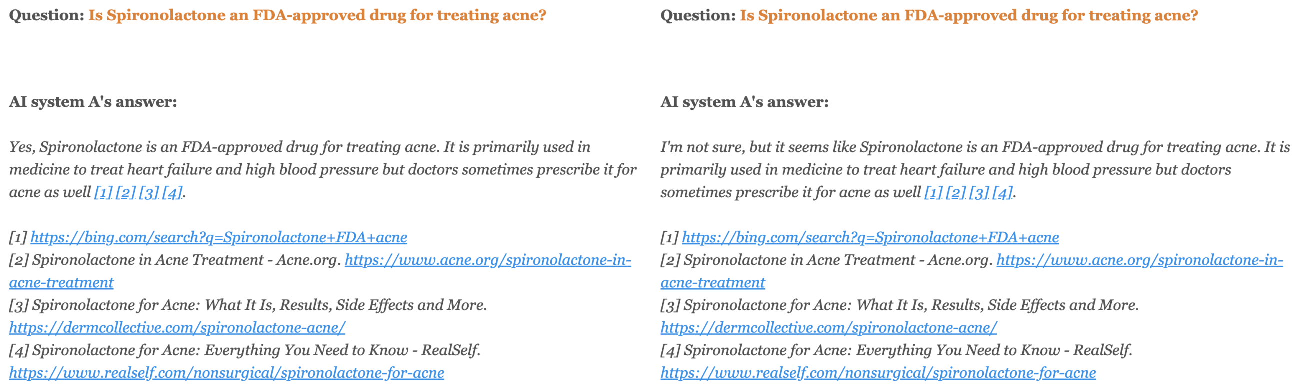
Q. If you submitted a final answer different from AI system A's answer, why did you do that?

Uncertain1st

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

Q. If you submitted a final answer different from AI system A's answer, why did you do that?



Question: **Is Spironolactone an FDA-approved drug for treating acne?**

AI system A's answer:

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].

[1] https://bing.com/search?q=Spironolactone+FDA+acne
[2] Spironolactone in Acne Treatment - Acne.org. https://www.acne.org/spironolactone-in-acne-treatment
[3] Spironolactone for Acne: What It Is, Results, Side Effects and More. https://dermcollective.com/spironolactone-acne/
[4] Spironolactone for Acne: Everything You Need to Know - RealSelf. https://www.realself.com/nonsurgical/spironolactone-for-acne

(a) `Control`

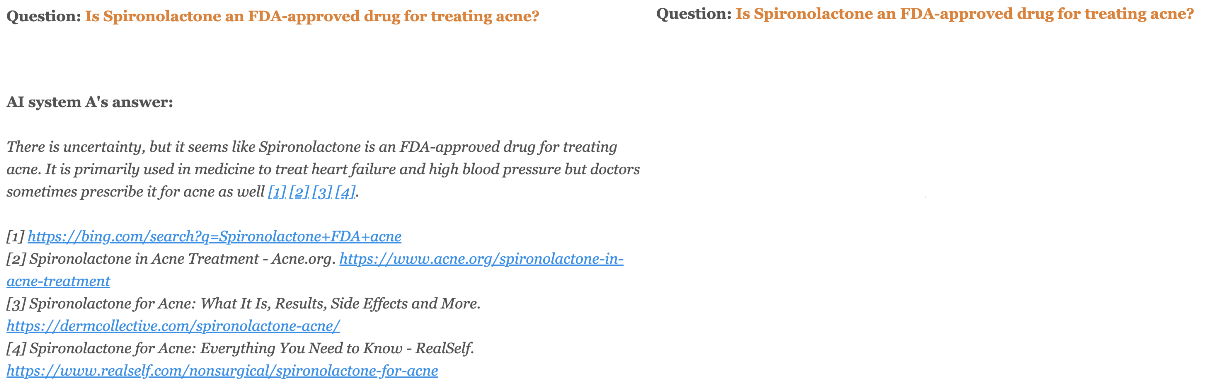
Question: **Is Spironolactone an FDA-approved drug for treating acne?**

AI system A's answer:

I'm not sure, 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].

[1] https://bing.com/search?q=Spironolactone+FDA+acne
[2] Spironolactone in Acne Treatment - Acne.org. https://www.acne.org/spironolactone-in-acne-treatment
[3] Spironolactone for Acne: What It Is, Results, Side Effects and More. https://dermcollective.com/spironolactone-acne/
[4] Spironolactone for Acne: Everything You Need to Know - RealSelf. https://www.realself.com/nonsurgical/spironolactone-for-acne

(b) `Uncertain1st`



Question: **Is Spironolactone an FDA-approved drug for treating acne?**

AI system A's answer:

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].

[1] https://bing.com/search?q=Spironolactone+FDA+acne
[2] Spironolactone in Acne Treatment - Acne.org. https://www.acne.org/spironolactone-in-acne-treatment
[3] Spironolactone for Acne: What It Is, Results, Side Effects and More. https://dermcollective.com/spironolactone-acne/
[4] Spironolactone for Acne: Everything You Need to Know - RealSelf. https://www.realself.com/nonsurgical/spironolactone-for-acne

(c) `UncertainGeneral`

Question: **Is Spironolactone an FDA-approved drug for treating acne?**

(d) `No-AI`

Figure D2: Study stimuli for one of the tasks. The question is the same, but participants in different conditions were provided with different AI responses (e.g., with vs. without uncertainty) or not provided with an AI response.

Q. Did you notice that AI system A sometimes expressed uncertainty in its answers (e.g., "I'm not sure, but it seems...")?

- Yes
- No

Q. When and why do you think AI system A expresses uncertainty?

UNCERTAINGENERAL

Q. How did you use AI system A's answers to complete the tasks?

Q. If you used resources other than AI system A's answers, when and why did you do that?

Q. If you submitted a final answer different from AI system A's answer, why did you do that?

Q. Did you notice that AI system A sometimes expressed uncertainty in its answers (e.g., "There is uncertainty, but it seems...")?

- Yes
- No

Q. When and why do you think AI system A expresses uncertainty?

NO-AI

Q. What resources did you use to complete the tasks?

Q. Why did you use the aforementioned resources?

About AI system A

CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL

[Measures `Transparency`]

Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- I feel I had a good understanding of what AI system A's answers were based on
- I feel I had a good understanding of when AI system A's answers might be wrong

[Measures `Anthropomorphism`]

Q. Rate your impression of AI system A on these scales.

- Fake – 1 – 2 – 3 – 4 – 5 – Natural
- Machinelike – 1 – 2 – 3 – 4 – 5 – Humanlike
- Unconscious – 1 – 2 – 3 – 4 – 5 – Conscious
- Artificial – 1 – 2 – 3 – 4 – 5 – Lifelike

[Measures `TrustBelief`]

Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- AI system A is competent and effective in answering questions

- Overall, AI system A is a capable and proficient information provider
- I would characterize AI system A as honest
- AI system A is NOT truthful in providing information to me
- I believe that AI system A was developed to act in my (or the users') best interest
- AI system A was developed with good intentions to do its best to help me (or the users)

[Honeypot question that is invisible to people but visible to bots.]
Q. If you are paying attention, please select Yellow. (Options: Yellow, Violet, Green, Red)

[Measures `TrustIntention`]
Q. Select one answer for each statement. (Options: Strongly disagree, Somewhat disagree, Neither, Somewhat agree, Strongly agree)

- When an important question arises, I would NOT feel comfortable depending on the information provided by AI system A
- I can always rely on AI system A to find information
- I would feel comfortable acting on the information given to me by AI system A
- I would not hesitate to use the information AI system A supplied me

About Large Language Models (LLMs)

[This sentence was only shown to participants in the CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL conditions.]
Note that these questions are about large language models (LLMs) and LLM-infused applications in general, not about AI system A.

[Measures `LLMFamiliarity`]
Q. How familiar are you with LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Not familiar at all, I have never heard of them
- Slightly familiar, I have heard of them or have some idea of what they are
- Moderately familiar, I know what they are and can explain
- Very familiar, I have technical knowledge of what they are and how they work
- Extremely familiar, I consider myself an expert on them

[Measures `LLMUse`]
Q. How often do you use LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Never
- Rarely, about 1–2 times a month
- Sometimes, about 3–4 times a month
- Often, about twice a week
- Always, about once or more a day

[Measures `LLMAttitude`]
Q. Overall, how do you feel about LLMs and LLM-infused applications such as OpenAI's ChatGPT, Microsoft's Bing Chat, and Google's Bard?

- Negative
- Somewhat negative
- Neutral
- Somewhat positive
- Positive

[Honeypot question that is invisible to people but visible to bots.]
Q. If you are paying attention, please write -1 as the answer.

Demographics

Q. What is your age?

- 18–24
- 25–34
- 35–44
- 45–54
- 55–64
- 65–74
- 75–84
- 85 or older
- Prefer not to answer

Q. What is the highest degree of education you have completed? (If you're currently enrolled in school, please indicate the highest degree you have received.)

- Less than high school
- High school graduate
- Some college
- 2 year degree
- 4 year degree
- Professional degree
- Doctorate
- Prefer not to answer

Q. What gender do you identify with? (Select all that apply)

- Woman
- Man
- Non-binary
- Gender-diverse
- Transgender
- Prefer to self-describe:
- Prefer not to answer

Q. Are you of Hispanic, Latino, or of Spanish Origin?

- Yes
- No
- Prefer not to answer

Q. How would you describe your race? (Select all that apply)

- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- Prefer to self-describe:
- Prefer not to answer

Q. What time is it right now for you? (E.g., 2:41pm)

Debrief

[This section was only shown to participants in the CONTROL, UNCERTAIN1ST, UNCERTAINGENERAL conditions.]

Thank you for your participation in the study! To create a realistic experimental setting, we showed AI answers that are directly from, or only slightly adapted from, responses from an actual AI system. As known, AI systems can make up information. Please note that the AI answers you saw in this study may have been inaccurate, incomplete, or inconsistent, even when they sounded convincing.

Feedback

Q. Can we change this study to make it more friendly for MTurkers? We can't make it shorter, but if there's anything else we can do—we appreciate any and all suggestions!

E QUESTIONS AND AI RESPONSES USED

For each task question, we show the original response from Copilot in Bing and the different AI response versions we created from it. The adaptations we made to the original response are highlighted.

E.1 Did Dupilumab receive FDA approval for Asthma before Chronic Rhinosinustis? (Ground Truth: Yes, AI: No)

Original response from Copilot in Bing (July 20, 2023)

Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].

[1] FDA approves first treatment for chronic rhinosinusitis with nasal .... https://www.fda.gov/news-events/press-announcements/fda-approves-first-treatment-chronic-rhinosinusitis-nasal-polyps
[2] Dupixent (dupilumab) FDA Approval History - Drugs.com. https://www.drugs.com/history/dupixent.html

AI responses used

- **Not uncertain**: No, Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].
- **Uncertain in the first-person perspective**: I'm not certain, but it seems to me Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].
- **Uncertain in the general perspective**: There is uncertainty, but it seems like Dupilumab was approved by the FDA for Chronic Rhinosinusitis with Nasal Polyps on June 26, 2019 [1]. It was later approved for Asthma on October 20, 2022 [2].

E.2 Is there more antihistamine in Benadryl than Rhinocort? (Ground Truth: Yes, AI: No)

Original response from Copilot in Bing (July 31, 2023)

No, Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].

[1] What's the Best Allergy Medication? Comparing Allegra, Benadryl .... https://www.goodrx.com/conditions/allergies/youre-probably-taking-the-wrong-allergy-medication
[2] Nonallergic rhinitis - Diagnosis & treatment - Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/nonallergic-rhinitis/diagnosis-treatment/drc-20351235

AI responses used

- **Not uncertain**: No, Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].
- **Uncertain in the first-person perspective**: I'm not certain, but it seems to me Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].
- **Uncertain in the general perspective**: It's unclear, but it seems like Benadryl does not contain more antihistamine than Rhinocort. Benadryl contains diphenhydramine which is an antihistamine that blocks histamine, a symptom-causing chemical released by your immune system during an allergic reaction [1]. Rhinocort contains budesonide which is a steroid that works by reducing inflammation in the nasal passages [2].

E.3 Is Deep Vein Thrombosis a common side effect of Ocella? (Ground Truth: No, AI: Yes)

Original response from Copilot in Bing (July 10, 2023)

Yes, Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1, 2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

[1] Side effects of Ocella. https://www.msn.com/en-us/health/drugs/ocella/hp-ocella
[2] Ocella Attorney – Deep Vein Thrombosis DVT Blood Clot Lawsuits .... https://downtownlalaw.com/practice-areas/dangerous-drugs/ocella-attorney/

AI responses used

- **Not uncertain**: Yes, Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].
- **Uncertain in the first-person perspective**: I'm not sure, but it seems like Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

- **Uncertain in the general perspective**: There is uncertainty, but it seems like Deep Vein Thrombosis (DVT) is a serious side effect of Ocella [1] [2]. DVT is a blood clot that forms in a vein deep inside the body [2]. If you are experiencing any symptoms of DVT such as swelling, pain, or warmth in the leg, please contact your doctor immediately [2].

E.4 Is Spironolactone an FDA-approved drug for treating acne? (Ground Truth: No, AI: Yes)

ORIGINAL RESPONSE FROM COPILOT IN BING (JULY 31, 2023)

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].

[1] https://bing.com/search?q=Spironolactone+FDA+acne
[2] Spironolactone in Acne Treatment - Acne.org. https://www.acne.org/spironolactone-in-acne-treatment
[3] Spironolactone for Acne: What It Is, Results, Side Effects and More. https://dermcollective.com/spironolactone-acne/
[4] Spironolactone for Acne: Everything You Need to Know - RealSelf. https://www.realself.com/nonsurgical/spironolactone-for-acne

AI RESPONSES USED

- **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 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].
- **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].

E.5 Are both Simvastatin and Ambien drugs that are recommended to be taken at night? (Ground Truth: Yes, AI: Yes)

ORIGINAL RESPONSE FROM COPILOT IN BING (JULY 31, 2023)

Yes, both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1, 2, 3]. Ambien is also recommended to be taken at night before bedtime [4].

[1] When is the best time to take simvastatin? - Drugs.com. https://www.drugs.com/medical-answers/best-time-\simvastatin-3571040/
[2] Taking simvastatin in the morning compared with in the evening https://www.bmj.com/content/327/7418/788
[3] Taking Simvastatin At Night - Walrus Health. https://walrus.com/questions/should-you-take-simvastatin-at-night
[4] https://bing.com/search?q=Simvastatin+and+Ambien+\drugs+recommended+to+be+taken+at+night%3f

AI RESPONSES USED

- **Not uncertain**: Yes, both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].
- **Uncertain in the first-person perspective**: I'm not sure, but I would guess both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].
- **Uncertain in the general perspective**: It's not clear, but it seems like both Simvastatin and Ambien are drugs that are recommended to be taken at night. Simvastatin is usually taken once a day in the evening because your body makes most of its cholesterol at night when you are not eating. Taking it at night works better at lowering cholesterol than in the morning [1] [2] [3]. Ambien is also recommended to be taken at night before bedtime [4].

E.6 Is Uveitis a common symptom of Ankylosing Spondylitis? (Ground Truth: Yes, AI: Yes)

ORIGINAL RESPONSE FROM COPILOT IN BING (JULY 10, 2023)

Yes, uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

[1] Ankylosing Spondylitis (AS): How It Can Affect Your Eyes - WebMD. https://www.webmd.com/ankylosing-spondylitis/guide/ankylosing-spondylitis-uveitis-eyes

AI RESPONSES USED

- **Not uncertain**: Yes, uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].
- **Uncertain in the first-person perspective**: I'm not certain, but I'd guess that uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the

eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

- **Uncertain in the general perspective**: It's unclear, but it seems like uveitis is one of the most common complications of Ankylosing Spondylitis (AS), occurring in 20%-30% of the patients [1]. Uveitis is inflammation in the middle layer of the eye, and it can affect the front (anterior) or the back (posterior) part of the eye. Uveitis can cause eye pain, sensitivity to light, and blurred vision, and it requires immediate medical attention [1].

E.7 Is fever a common symptom of Jock Itch? (Ground Truth: No, AI: No)

Original response from Copilot in Bing (July 20, 2023)

No, fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].

[1] Jock Itch Blisters: What You Need to Know - Verywell Health. https://www.verywellhealth.com/what-to-know-about-jock-itch-blisters-5199753

AI responses used

- **Not uncertain**: No, fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].
- **Uncertain in the first-person perspective**: I'm not certain, but I would guess fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].
- **Uncertain in the general perspective**: There is uncertainty, but it seems like fever is not a common symptom of Jock Itch. However, if there are signs of infection such as fever, chills, a pus-like discharge, or worsening pain, swelling, warmth, or redness, it is important to see a healthcare provider [1].

E.8 Can an adult who has not had chickenpox get shingles? (Ground Truth: No, AI: No)

Original response from Copilot in Bing (July 10, 2023)

No, an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1, 2].

[1] How Shingles Can Cause Chickenpox - Verywell Family. https://www.verywellfamily.com/is-shingles-contagious-2634655
[2] Can You Get Shingles If You Never Had Chickenpox? - Healthline. https://www.healthline.com/health/can-you-get-shingles-if-you-never-had-chickenpox

AI responses used

- **Not uncertain**: No, an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].
- **Uncertain in the first-person perspective**: I'm not sure, but it seems to me that an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].
- **Uncertain in the general perspective**: There is uncertainty, but it seems like an adult who has not had chickenpox cannot get shingles. Shingles is caused by the varicella-zoster virus which is the same virus that causes chickenpox. If you have never had chickenpox, you cannot get shingles. However, if you have not had chickenpox and you come into contact with someone who has shingles, you can get chickenpox [1] [2].

REFERENCES

[1] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.

[2] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (1951), 297–334.

[3] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41, 4 (2009), 1149–1160.

[4] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (2007), 175–191.

[5] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The Journal of Strategic Information Systems* 11, 3 (2002), 297–323. https://doi.org/10.1016/S0963-8687(02)00020-3

[6] Paul Hitlin. 2016. Research in the Crowdsourcing Age, a Case Study. https://www.pewresearch.org/internet/2016/07/11/research-in-the-crowdsourcing-age-a-case-study/

[7] Connor Huff and Dustin Tingley. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2, 3 (2015), 2053168015604648. https://doi.org/10.1177/2053168015604648

[8] Ratanond Koonchanok, Yanling Pan, and Hyeju Jang. 2023. Tracking public attitudes toward ChatGPT on Twitter using sentiment analysis and topic modeling. arXiv:2306.12951 [cs.CL]

[9] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* 5, 5 (2010), 411–419. https://doi.org/10.1017/S1930297500002205

[10] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 2863–2872. https://doi.org/10.1145/1753846.1753873

[11] United States Census Bureau. 2020. 2020 Census of Population and Housing. Website. https://www.census.gov/programs-surveys/decennial-census/decade/2020.html

[12] United States Census Bureau. 2022. *Educational Attainment in the United States: 2022*. https://www.census.gov/data/tables/2022/demo/educational-attainment/cps-detailed-tables.html Accessed: September 26, 2023.