Uneven Trust in LLMs: Beliefs About Accuracy Vary Across 11 Countries

SANCHAITA HAZRA, University of Utah, USA
MARTA SERRA-GARCIA, University of California San Diego, USA

Large Language Models (LLMs) can efficiently provide users with information, but this information may not always be accurate. This paper provides novel evidence on cross-cultural trust in an LLM (GPT-4o) as an information source. In a large-scale experiment with over 2,900 participants across 11 countries, participants were incentivized to evaluate the accuracy of LLM-generated answers to 100 policy-relevant, factual questions. Across all countries, individuals have a limited ability to assess the accuracy of the LLM. In low- and middle-income countries, participants are overly trusting and overconfident in their assessment abilities. By contrast, in high-income countries, especially the U.S., participants exhibit limited trust in the LLM and are skeptical of their ability to detect LLM inaccuracy. These findings highlight the importance of understanding cross-cultural differences in human interactions with globally-available LLMs.

Contents

Abstract 0
Contents 0
1 Introduction 1
2 Experimental Design 2
2.1 The LLM: Questions and Answers 2
2.2 Participants 5
2.3 Design Outline 5
2.4 Procedures 5
3 Hypotheses 6
4 Results 7
4.1 Cross-country Differences in Beliefs and Accuracy 8
4.2 Determinants of Beliefs About Accuracy 13
5 Discussion and Conclusion 17
References 19
A Inclusion and Ethics 21
B Data, Materials, and Software Availability 21
C Additional Results: Human Beliefs and Accuracy 21
C.1 Demographics 21
C.2 Beliefs about LLM Accuracy 22
C.3 Human Accuracy 23
C.4 Confidence 24
C.5 Checking Outside Sources 25
D Experimental Design: Additional Information 26
D.1 Participant Instructions 26
D.2 Example from the Task 27
D.3 Additional Questions 28
E Coding Instructions: LLM Accuracy 30

Corresponding author: *sanchaita.hazra@utah.edu.* This paper has benefited from highly insightful comments from seminar and conference participants. Wanting Zhou, and Sky Chen provided excellent research assistance.

1 Introduction

Generative AI has transformed societies worldwide, showing the potential to increase productivity and complement human intelligence [Bubeck et al., 2023, Jones, 2024]. Large Language Models (LLMs), a form of generative AI, assist humans in a variety of tasks, such as writing [Noy and Zhang, 2023], idea-generation [Doshi and Hauser, 2024], and programming [Peng et al., 2023]. LLMs are also used as information sources, an ability that has received little empirical attention thus far. Rather than web searching, individuals can quickly and efficiently obtain answers to questions on any topic by asking the LLM [del Rio-Chanona et al., 2023, Reddit, 2023a,b]. Such use of LLMs can have a large impact on societal knowledge. If people turn to LLMs for information rather than posting (and receiving answers to) questions on public sources, publicly available knowledge could decrease.

Human interaction with LLMs is complex, and it depends on the context studied. Existing research has focused on political and health misinformation, showing that individuals in the United States often fail to detect LLM-generated misinformation [Chen and Shu, 2023, Kreps et al., 2022, Spitale et al., 2023] and that they can be persuaded by messages generated by LLMs, as much as by those written generated by humans [Goldstein et al., 2024]. In this paper, we propose to take a step back and study LLM-generated content that attempts to answer global, non-political (non-partisan and not news-related) questions accurately, although it unintentionally makes mistakes. Since LLMs are globally available, we must understand, across societies worldwide [Arechar et al., 2023, Cohn et al., 2019, Gelfand et al., 2011, Globig et al., 2024, Kimbrough et al., 2024], how accurately individuals evaluate well-intentioned LLM-generated content related to a variety of topics, their trust in the LLM [Dimant and Kimbrough, 2024], and their confidence in their ability to spot mistakes when using LLMs as information sources.

This paper provides evidence from over 2,900 individuals across 11 countries that measures beliefs about the accuracy of LLM (GPT-4o) responses to 100 different policy-relevant questions with factual answers. We pre-registered a large-scale online experiment that included participants from the following countries: Australia, Canada, Chile, India, Kenya, Mexico, New Zealand, Spain, South Africa, the UK, and the US. Nine of these countries were among the top 50 countries worldwide in online traffic to the LLM (ChatGPT), and three of them were in the top 3 (India, the US, and the UK), at the time of the study [SimilarWeb, 2024]. Each participant ($N = 2,932$) was presented with 20 out of 100 policy-relevant questions and was incentivized to accurately evaluate the LLM's answer (Figure 1).

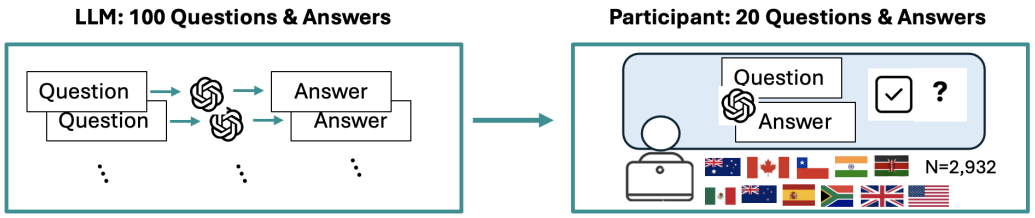


Fig. 1. Experimental Design: Outline

Notes: 100 factual questions about policy-relevant topics affecting the US and other countries worldwide were asked to the LLM (GPT-4o). The answers were recorded and presented to participants. Each participant read a random subset of 20 questions and evaluated whether the answer by the LLM is accurate or not, in an incentivized experiment with participants from 11 countries (flags represent each country), $N = 2,932$.

The questions concerned facts such as the gender wage gap, racial differences in financial markets, and statistics related to the economy of a country (trade, employment, and other topics),

that individuals might look up through a search engine or via an LLM. They were US-specific in half of the cases and related to international topics in the other half of the cases, including questions specific to each country. Although the LLM provided correct responses in 74% of the cases, it made mistakes (see also, [Ji et al., 2023, Narayanan and Kapoor, 2024]) and generated incorrect responses in close to a quarter of the cases (26%). An important question is, therefore, whether individuals can spot those unintentional mistakes and whether they anticipate the potential limitations in their own assessments.

Our findings reveal that, across all countries, the ability to correctly identify mistakes by the LLM is very limited. But, there is substantial heterogeneity in how much they trust the LLM and how confident they are in their ability to assess LLM-generated answers. In low- and middle-income countries, participants overly trust the LLM's accuracy and are overconfident. They overestimate their absolute ability to accurately assess LLM-generated answers, and they overestimate their ability relative to others. By contrast, individuals in high-income countries display significantly less confidence and lower trust in the LLM. A remarkable case is the US, as Americans overly *distrust* the LLM and exhibit the most pessimistic beliefs about their own ability to spot mistakes. These findings cannot be explained by a lack of familiarity with LLMs since participants in low- and middle-income countries report using LLMs *more* often than individuals in high-income countries.

If this study had been solely conducted in the US, we might have concluded that individuals are adept at recognizing their limited ability and maintain a healthy skepticism towards LLMs when asking them for information. However, we find the US to be an exception rather than the rule. Focusing solely on one or a small group of (high-income) countries would thus risk painting an incomplete picture of how LLMs are understood and received worldwide. Such a conclusion could be harmful for individuals in countries where knowledge about LLM-related risks and safety is limited and further perpetuates inequities worldwide. This paper, therefore, showcases the importance of considering cross-cultural heterogeneity as we build further knowledge regarding how humans interact with generative AI.

2 Experimental Design

The design of the experiment is divided into two parts: the questions asked to the LLM and the questions asked to participants, who were provided with the LLM's answers.

2.1 The LLM: Questions and Answers

We posed 100 policy-relevant questions that had a factual answer to an LLM (GPT-4o).[1] The questions are evenly divided between topics pertinent to the United States (50 questions) and international topics (50 questions). We record the zero-shot response for each question.

We systematically selected questions relating to policy-relevant topics that have an objectively correct answer. We first identified the top discussed issues for a respective country on the internet (majorly from the most recent articles from top media outlets such as the New York Times, Washington Post, and Forbes). To obtain factually correct answers, we chose questions that have factual data available from major government websites and leading research institutes (such as Pew Research).

The topics of discussion include the economy, migration, health, crime, and politics. For example, in the context of the United States, we selected statements relating to the Black homeownership rates and the gender wage gap among different age and racial groups. Similarly, for several other countries, such as India and Spain, we included statements pertaining to employment.

[1]Throughout, we used GPT-4o, the most recent version available in July 2024. See https://platform.openai.com/docs/models/gpt-4o.

We introduced variation in how the questions were asked to the LLM by including a source to use in the search for answers or not providing a source, and therefore giving the LLM no guidance in the generation of answers. This variation allows us to examine whether the accuracy of the LLM changes with the type of prompt and whether individuals' perceptions about the LLM accuracy also change. When the LLM is provided a source, it receives a specific source link that contains the answer, following the procedure of retrieval-augmented generation [Lewis et al., 2020, Shi et al., 2023]. The LLM was told it could consult the source, but it was not restricted to only using the provided source. When we do not provide a source, we ask the same question, but we do not provide any source link. In both cases, we ask the LLM to cite appropriate sources that it used to generate its responses.

Table 1 shows the fraction of questions that were asked with and without source, the fraction that concerned the US relative to other geographical locations, and the distribution of topics across the questions. By design, half of the questions contained a source, and half of the questions concerned the US. Across questions, the topics varied, with Economics (40%) and Politics (36%) being the most frequent.

Table 1. LLM Questions Asked: Descriptive Statistics

	Fraction of Questions
Question included source	0.50
Question concerns US	0.50
Topic=Climate	0.05
Topic=Crime	0.03
Topic=Economics	0.40
Topic=Education	0.03
Topic=Health	0.13
Topic=Politics	0.36
Total	200

Since providing a source did not affect LLM accuracy (70% accuracy without a source, and 78% accuracy with a source, $p-$value= 0.202), we focus on testing whether individuals believe such prompt differences affect LLM accuracy. To rate the accuracy of LLM-generated responses, we employ three raters. Each rater first independently evaluated whether the response provided by the LLM was correct. Then, the ratings were compared, and the rating differences were discussed to reach unanimity. The instructions followed by raters are provided in Online Appendix E.

Overall, the LLM answered 74% of the questions correctly. Its accuracy did not vary by topic, by region, or by whether it was provided the source to consult or not, as shown in Table 2. Table 2 examines the determinants of accuracy for the LLM. None of the covariates significantly affect the accuracy of the LLM. The LLM was provided the source to consult (78% accuracy), and when it was not provided a specific source (70% accuracy, t-test, p-value= 0.202). The geographical region to which the question refers or the topic of the question does not correlate with the accuracy of the LLM either.

In the context of our questions, the source of incorrect responses (also known as hallucinations) of the LLM can be attributed to failures in knowledge retrieval and comprehension [Ji et al., 2024].[2] Consistent with this type of incorrect responses, mistakes by the LLM are more common when the

[2]Generally, the source of hallucinations can be several: confabulations, which are unrelated features of text generation such as random seeds [Farquhar et al., 2024], mechanistic failures in knowledge retrieval and comprehension [Ji et al., 2024], a consequence of training on erroneous knowledge [Lin et al., 2022], or being strategically driven by bad actors to sample propaganda-driven wrong information [Evans et al., 2021]. Since the questions we ask are factual and the sources we use

Table 2. Determinants of LLM Accuracy

	(1)	(2)
	LLM provides accurate response= 1	
Question included source	0.080	0.080
	(0.062)	(0.062)
Question concerns US		0.005
		(0.066)
Crime		-0.033
		(0.228)
Economics		0.059
		(0.153)
Education		-0.202
		(0.229)
Health		-0.049
		(0.168)
Politics		0.075
		(0.153)
Constant	0.700***	0.660***
	(0.044)	(0.143)
Observations	200	200

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the LLM answers the question it was asked correctly, according to three independent raters. The variable Question included source is an indicator for whether the LLM was given a link to the source to consult for the question asked, and the Question concerns US is an indicator for whether the question asked concerned the US. Crime, Economics, Education, Health, and Politics are indicator variables for topics of the questions asked, where Climate is the omitted variable. Robust standard errors, clustered at the participant level, in parentheses.
* p<.10; ** p<.05; *** p<.01

answer is provided in a figure, rather than in text or table (in one or more online sources). When they occur, the LLM either used a wrong statistic from the same source or provided a number that was not presented anywhere in the sources it checked.

Exploring the LLM's responses, we find that in most cases (78%), the LLM is rated the same, as correct or incorrect, irrespective of the presence of a source (78 cases out of 100, of which 63 cases are correct and 15 cases incorrect). In the remaining 22 cases, there are 15 cases in which LLM with source is correct, while the LLM without source is incorrect. There are only 7 cases in which the LLM without source is correct, but the LLM with source is incorrect.

The LLM listed the source or sources it used to generate the response (under searched sites). When given a source, it used the source given in 84% of the cases and did not mention an explicit source in the remaining cases. When not given a source, the LLM provided, on average, 3.41 sources and provided at least one source in 99% of the cases. When not given a source link, in 38% of the cases, one of the sources provided by itself coincided with the source given to the LLM (in the source cases).

are reputable, we minimize the likelihood that the LLM error is based training on erroneous knowledge or a strategically driven wish to use propaganda. Since we record one-shot responses, we do not capture randomness in response generation.

Several mistakes from the LLM occurred when the answer to the question was provided in a figure, instead of text. In 25% of the cases, the factual response was presented in a figure in the source link (or in at least one of the sources if two or more sources were used by the LLM to find the answer). The average accuracy when the factual answer is in a figure is 62%, compared to 78% when the answer is presented in the text or a table.

2.2 Participants

We recruit participants on Prolific Academic. Each participant read 20 randomly selected questions from the set of 100. The participant was shown the question asked and the LLM response for each question individually (on separate pages). The main task of the participant was to evaluate whether the answers provided by the LLM were correct or incorrect. Participants were informed that one of the 20 answers they evaluated would be randomly selected for bonus payment, and they would be paid $2 if they guessed correctly whether the LLM provided a correct or incorrect answer. Across participants, we varied whether they saw questions with a source link or without out, and the corresponding LLM answers. In addition to tracking their responses, we measured the time participants took to make their decisions.[3]

After providing their evaluation of LLM answers, we elicited two measures of participants' confidence. First, we elicited participants' absolute confidence in their own performance. We asked participants how many of the 20 questions and answers by the LLM they thought they evaluated correctly. They received a bonus of $1 if they correctly estimated their performance. Next, we elicited participants' relative confidence in their ability using the question, "Relative to others who completed the same study as you, how do you think you did?" They were asked to choose a quartile in the distribution of performance, and they earned $1 if their answer was correct.

At the end of the experiment, we collected information on the participants' experience with GPT by asking whether they use GPT regularly, rarely, or never, whether they have paid subscriptions, and asking them how they typically use GPT (in an open-ended text question). We also asked participants to report their gender, age, education, social media use, and political orientation. Since participants from multiple countries participated in the study, we asked the question used in the World Values Survey [Haerpfer et al., 2020], which elicits, on a scale from 1 to 10, whether the participant views their views about political matters more on the left (1) or on the right (10).

2.3 Design Outline

Figure 2 shows a graphical summary of the experimental design. The first step (column (1)) was to select a topic and fact, finding a source in which the fact was presented. This fact was used to formulate a question that was then asked to GPT-4, as shown in column (2). The questions asked, and the answers provided by the LLM were then presented to the subjects. The subject then selected whether the answer was correct or incorrect, as shown in column (3).

2.4 Procedures

We pre-registered to collect a total of 3,000 observations from 11 countries. The pre-registration details are provided in Appendix B. We aimed to gather 300 observations each, or as many as possible up to this number, from the following countries: the United States, the United Kingdom,

[3]The survey was programmed on Qualtrics with a setting that prevented copy-pasting the text, as it was an image, to consult other sources online easily. We also asked participants about their checking of other sources, and 29% of them reported having checked at least some answers. We examine whether this behavior relates to their accuracy in assessing the LLM, and we do not find evidence that their checking is correlated with their accuracy ($t-$test, $p-$value= 0.237). Descriptive statistics for how often participants checked and their accuracy, depending on this behavior by country, are provided in Table 13.

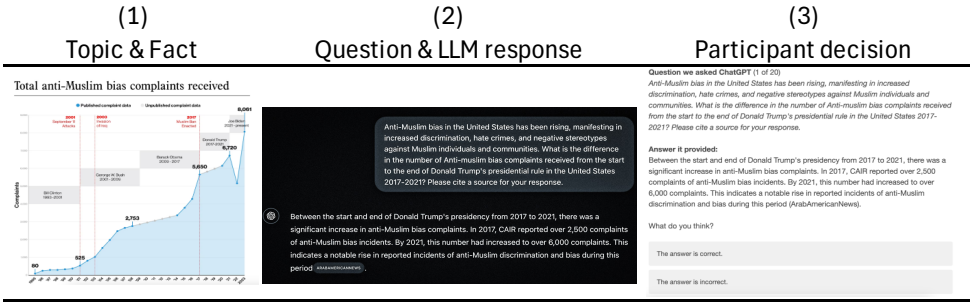


Fig. 2. Experimental Design

Notes: This figure provides a graphical summary of the experimental design by presenting screenshots for one fact/question. The first screenshot presents the answer to the question in column (1), then the question and the LLM's answer are presented in column (2), and the screenshot for the participant is presented in column (3).

Canada, Australia, New Zealand, South Africa, Spain, and Mexico. Additionally, we intended to collect 200 observations each, or as many as possible up to this number, from India, Chile, and Kenya. [4] Out of the 11 countries, 9 are on the list of the top 50 countries with the most online traffic to GPT between July and September 2024. Their rank is shown in parentheses, for the top-50: India (1), US (2), UK (3), Canada (6), Mexico (8), Australia (14), Spain (15), South Africa (20), and Chile (34).[5]

Overall, the data includes 2,932 participants. The fraction of female participants ranged from 30 to 71%, with average ages ranging from 27 to 40 years old across countries. Their socio-demographic characteristics are summarized in the Appendix, Table 9.[6]

In countries where English is the main or one of the main languages (US, UK, Canada, Australia, New Zealand, South Africa, India, and Kenya), all the instructions were presented in English. In Spanish-speaking countries (Chile, Mexico, and Spain), all the instructions were presented in Spanish. The questions to the LLM were asked in English, and its responses were in English, and they were translated into Spanish and shown in that language. Participants knew about this translation.

3 Hypotheses

The first hypotheses focus on individuals' beliefs about LLM accuracy and their ability to detect LLM errors generally, and our later hypotheses focus on differences across groups or types of questions asked (following our pre-registration).

The first hypothesis concerns individuals' beliefs about the accuracy of the LLM. We hypothesize that individuals are unaware of the actual accuracy of the LLM and believe it has a higher accuracy than it actually does. While there has been some discussion about the potential of LLMs to hallucinate [Li et al., 2023], there has also been significant "hype" in the capabilities of LLMs to enhance human knowledge and productivity [Narayanan and Kapoor, 2024], leading individuals to overestimate their capabilities.

[4]Participants had to be located in the respective country, which was checked via their IP geolocation as well as through Prolific. They also had to have reported to Prolific that they held that country's nationality.

[5]Data source is Similarweb, https://www.similarweb.com/website/chat.openai.com/ #demographics, retrieved October 16, 2024.

[6]We had two control questions for the participants to answer before starting the experiment. As pre-registered, we excluded subjects who failed to answer either one of the control questions correctly.

Hypothesis 1. Individuals believe the LLM is more often correct than it actually is.

The second hypothesis concerns the ability of individuals to detect whether an LLM response is correct or incorrect. Existing research has shown that individuals cannot detect whether others are lying to them [Bond Jr and DePaulo, 2006]. The LLM can also mislead individuals, providing them with factually incorrect information [Li et al., 2023]. We hypothesize that detecting these hallucinations is challenging for most individuals. We also evaluate their accuracy conditional on whether the LLM is correct or incorrect, expecting individuals to have a limited ability to correctly classify LLM responses in both cases.

Hypothesis 2. Individuals will not be better than chance at evaluating whether the LLM is correct or incorrect, both when it is actually correct and when it is incorrect.

The third hypothesis concerns individuals' confidence in their ability to correctly evaluate the LLM's answers. Individuals are frequently overconfident in their ability to detect false information [Serra-Garcia and Gneezy, 2021]. We hypothesize that individuals are overconfident in detecting false content generated by AI, in absolute terms and in relative terms [Burks et al., 2013, Moore and Healy, 2008].

Hypothesis 3. Individuals will be overconfident in their ability to evaluate the LLM, in absolute and relative terms.

We also test whether individuals' ability to evaluate the LLM is related to their confidence and whether men are more overconfident than women in this type of task [Bordalo et al., 2019, Coffman, 2014]. Our second set of hypotheses examines individuals' beliefs across different topics the LLM is asked about, across different ways in which the LLM can be asked, and for individuals across different countries. Most LLMs have been developed in the US and are in English by default. Because of how an LLM like GPT is trained, we hypothesize that participants will believe that the LLM is more accurate when asked about matters that concern the US, compared to international ones. Because LLMs are applied by online platforms, affecting people's online experience worldwide, and they are broadly discussed in traditional and social media, we hypothesize that participants in high-income countries (US, UK, Canada, Australia, New Zealand, and Spain, in our study) will exhibit similar ability in detecting LLM accuracy than participants in other countries (middle- and low-income countries, Chile, Mexico, India, South Africa, and Kenya, in our study). But, we hypothesize participants in high-income countries will be more confident and more likely to have experience with it.

Hypothesis 4.

4A. Individuals believe the accuracy of the LLM does not vary by whether it is provided a source or not, and by whether the matter it is asked about is a US or international matter.
4B. Individuals in high-income countries exhibit a higher use of LLMs and higher confidence in their ability to evaluate them, although their ability will not differ from that in low-income countries.

4 Results

Individuals' beliefs about LLM accuracy, and their accuracy conditional on the LLM being correct and conditional on it being incorrect, did not vary depending on whether the source was provided

or not (Table 3). We hence pool the observations for these treatments for the first in Section 4.1 and separate them later in Section 4.2.

Table 3. Human Beliefs and Accuracy

	(1)	(2)	(3)
	Human Beliefs and Accuracy		
	Belief LLM is correct = 1	If LLM is correct, human is accurate= 1	If LLM is incorrect, human is accurate= 1
Question included source	-0.008	0.010	-0.009
	(0.006)	(0.009)	(0.006)
Constant	0.730***	0.288***	0.737***
	(0.017)	(0.024)	(0.018)
Observations	58,640	15,075	43,565
Individual participants	2,932	2,928	2,932

Notes: This table displays the estimated coefficients from linear regressions on likelihood that the human beliefs the LLM answered correctly (column (1)), on the likelihood that, if the LLM is correct, the human accurately detects it (column(2)), and on the likelihood that, if the LLM is incorrect, the human accurately detects it (column (3)). The regressions include country-fixed effects and demographic characteristics. Robust standard errors, clustered at the participant level, in parentheses. * p<.10; ** p<.05; *** p<.01

4.1 Cross-country Differences in Beliefs and Accuracy

We first examine how much individuals trust LLM-generated answers by measuring the percentage of cases (out of 20) in which participants rated the LLM as correct. Considering all countries, consistent with the wisdom of the crowds [Lee and Lee, 2017, Surowiecki, 2005], individuals are equally likely to overestimate or underestimate LLM accuracy (Figure 3, Panel A; $t-$test, $p-$value=0.28). Hence, contrary to Hypothesis 1, in a majority of countries, individuals do not overestimate the accuracy of the LLM. While this result is found in the aggregate and in eight of the eleven countries, it is important to highlight that three countries show a different pattern. In Kenya and Mexico, participants overly trust the LLM, as they overestimate its accuracy, while in the US, participants overly distrust the LLM, as they significantly underestimate its accuracy.

Additionally, at the individual level, there is significant heterogeneity in how much individuals trust the LLM (Figure 3, Panel B). Considering all countries, a majority of the individuals (64%) over- or underestimate the accuracy of the LLM by more than 10 percentage points, revealing that many individuals misperceive how accurate LLM-generated answers are. Comparing low- and middle-income countries (Chile, Kenya, India, Mexico, and South Africa) with high-income countries (Australia, Canada, New Zealand, Spain, UK, and US) we observe significantly higher trust in the LLM's accuracy in low- and middle-income countries, with a believed accuracy that is too optimistic (of 76.2%), compared to high-income countries, with a believed accuracy of 73.7% ($t-$test, $p-$value$<$ 0.001), which is only marginally below the actual accuracy.

Result 1. *A majority of participants hold incorrect beliefs about the accuracy of the LLM. In the US, individuals overly distrust the LLM, while individuals from Kenya and Mexico overly trust it.*

Next, we examine individuals' ability to spot mistakes by the LLM. We answer the question when confronted with a specific accurate or an inaccurate LLM answer, are they able to distinguish correct and incorrect answers? We measure how often the participant accurately (or correctly) rates the LLM output as correct or incorrect.

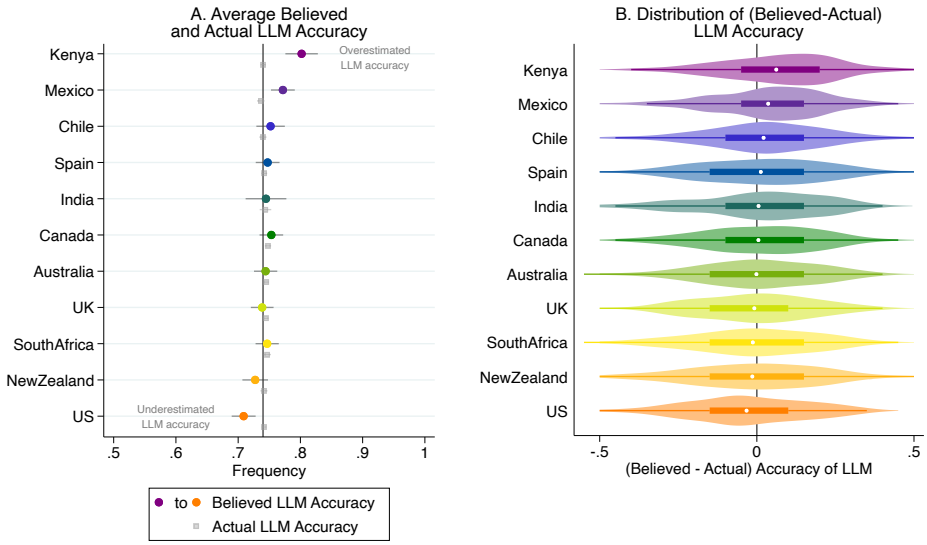


Fig. 3. Human Beliefs about the Accuracy of LLM Output

Notes: Panel A shows the average believed LLM accuracy and the actual LLM accuracy, based on regression-adjusted estimates for country-level beliefs, including fixed effects for question (100 questions in total) and presence of a source in it. The vertical gray line indicates the average accuracy of the LLM over all questions (74%). The bars underlying each symbol are 95% confidence intervals from covariate-adjusted means that included socio-demographic characteristics and fixed effects for each question-answer and for whether the LLM was provided a source. Panel B shows the distribution of the difference between individual-level believed accuracy and actual accuracy for the 20 LLM responses that each individual evaluated. The white circle in each row indicates the mean difference, while the box inside the density plot indicates the 25th and 75th percentiles. The vertical line indicates the observations for which there is no difference between believed and actual LLM accuracy (difference of zero). See Tables 4 and 10 for detailed results.

Individuals' ability to accurately rate the LLM's answers is limited. Humans correctly rate the LLM as correct or incorrect in 62.7% of the cases. While this rate is higher than 50%, the appropriate benchmark is how often an individual would spot mistakes, if it knew the accuracy of the LLM (74%) and it would randomly rate an answer as correct with a probability of 0.74. We simulate the accuracy of such a random data generator, with 58,640 draws that follow a probability of 0.74 of rating an LLM statement as correct. In that case, the accuracy (by chance) would be 61.6%. Although the accuracy by chance is statistically significantly lower than that of participants, 62.7% ($t-$test, $p-$value< 0.001), it is only 1.1 percentage points lower in magnitude. As Panel A of Figure 4 shows, the distribution of accuracy based on human responses, in purple, is similar to the distribution based on chance.

When examining participants' ability to evaluate answers provided by the LLM, it is important to distinguish between their ability to correctly identify correct answers and their ability to correctly identify mistakes by the LLM. The data show that individuals make mistakes both when confronted with correct and incorrect LLM-generated responses, in line with Hypothesis 2. The rate of mistakes, in fact, parallels their beliefs about the LLM's accuracy (i.e., their trust in the LLM on average). Conditional on the LLM being correct (blue squares), humans correctly rate the LLM 75.1% of the time, which is close to their belief about how often the LLM is accurate in general. Consistent with this finding, when the LLM is incorrect (green triangles), which occurs in a small fraction of cases (26%), humans correctly rate the LLM in 26.7% of the cases.

Table 4. Human Beliefs and Actual LLM Accuracy

	(1)	(2)
	Human Beliefs and Actual LLM Accuracy	
	Belief LLM is accurate = 1	LLM is accurate= 1
Canada	0.009	0.003
	(0.013)	(0.003)
Chile	0.008	-0.006*
	(0.015)	(0.003)
India	0.001	-0.001
	(0.019)	(0.005)
Kenya	0.058***	-0.005*
	(0.017)	(0.003)
Mexico	0.028**	-0.009**
	(0.014)	(0.004)
NewZealand	-0.017	-0.004
	(0.014)	(0.003)
SouthAfrica	0.002	0.001
	(0.014)	(0.004)
Spain	0.003	-0.003
	(0.014)	(0.003)
UK	-0.006	-0.001
	(0.013)	(0.003)
US	-0.035***	-0.004
	(0.013)	(0.003)
Female	-0.018***	
	(0.006)	
Age	0.001**	
	(0.000)	
Education is less 4yr college	-0.015**	
	(0.007)	
Number of social media platforms used	0.004	
	(0.003)	
Politically leaning left (standardized)	-0.000	
	(0.003)	
Lives in urban area	-0.006	
	(0.007)	
Constant	0.693***	0.962***
	(0.024)	(0.023)
Observations	58,640	58,640
Clusters	2,932	100

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the human beliefs the LLM answered correctly (column (1)), on the likelihood that the LLM is correct (column(2)). The regressions include fixed effects for questions (100 questions in total) and the presence of a source in it. Each country name represents an indicator variable for that country, where the omitted country is Australia. Robust standard errors, clustered at the participant level in column (1) and at the question level in column (2), in parentheses. * p<.10; ** p<.05; *** p<.01

The differences across countries mirror differences observed in the trust of the LLM. Participants in Kenya who overly trust the LLM exhibit a significantly lower ability to rate the LLM accurately when it is incorrect, and participants in the US who overly distrust the LLM exhibit a lower ability to rate the LLM accurately when it is correct. These findings lead to Result 2.

Result 2. *Individuals' ability to accurately assess the LLM is limited, only marginally better than chance, both if the LLM is correct and if it is incorrect.*

An important question that follows from these findings is whether individuals are aware of their limited ability to assess LLM-generated answers. In a majority of the cases, individuals are not.

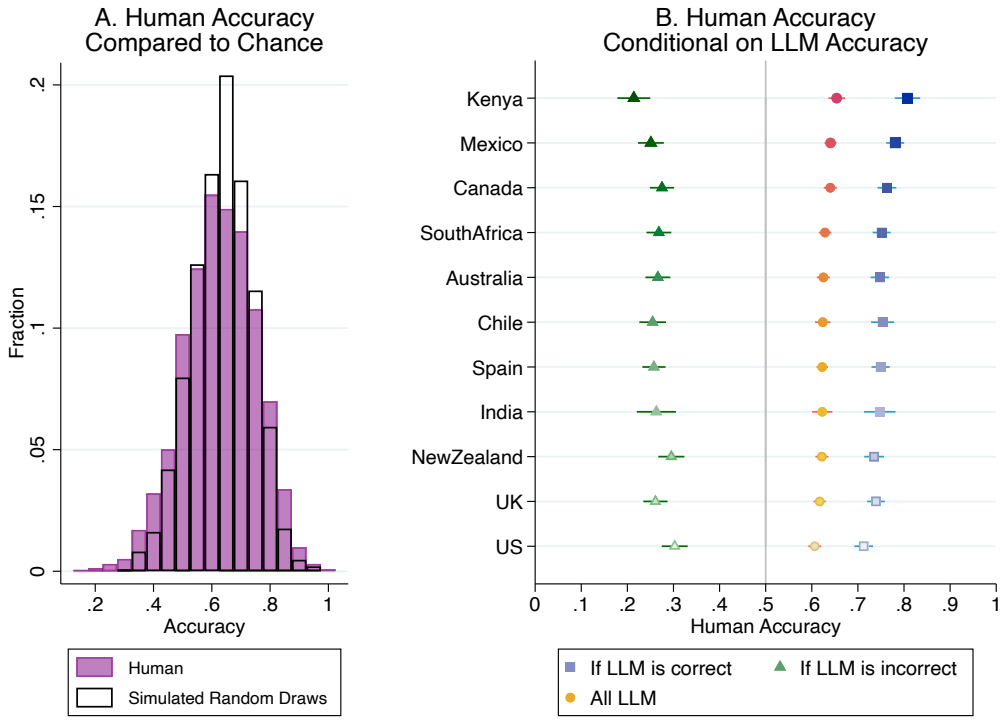


Fig. 4. Human Accuracy in Evaluating LLM Output

Notes: Panel A shows the distribution of human accuracy (in purple). It also shows the distribution of accuracy from 58,640 random draws, which evaluate an answer as correct with a 70% chance and incorrect with a 30% chance (white). Panel B shows the average accuracy of humans by country, (1) if the LLM is correct (blue squares), (2) if the LLM is incorrect (green triangles), and (3) shows the average accuracy by country (orange circles). The bars underlying each symbol are 95% confidence intervals, from covariate-adjusted means that included socio-demographic characteristics and fixed effects for each question-answer and for whether the LLM was provided a source. See Tables 5 and 11 for detailed results.

We first consider individuals' absolute confidence, based on how often they believe they correctly rated the LLM's answers. Individuals are overly confident: They believe they can accurately rate 70% of the LLM-generated answers, which is significantly larger than their actual average ability of 62.7% ($t-$test, $p-$value< 0.001), as shown in Panel A of Figure 5. The difference between actual and believed ability varies by country, with Kenya and India showing the highest confidence (78% and 75%, respectively), and the US and the UK showing the lowest confidence (63% and 66%, respectively).

Although individuals overestimate their ability, their beliefs are not simply noisy. There is a significant correlation between participants' believed and actual performance – the Spearman rank correlation coefficient is 0.32 ($p-$value< 0.001). This finding shows that their confidence provides some information about their actual ability, although this information is biased (a finding that is consistent in all countries, as shown in Figure 6).

Individuals perceived ability relative to others (relative overconfidence) is also biased (Panel B of Figure 5). Participants overestimate how well they perform. Their average reported quartile in the distribution of ability is 1.91, which is lower than the expected average of 2.5 ($t-$test,

Table 5. Human Accuracy

	(1)	(2)	(3)
	Human Accuracy		
	Human Accurate if LLM is correct = 1	Human Accurate if LLM is incorrect = 1	Human Accurate All
Canada	0.015	0.009	0.015
	(0.014)	(0.019)	(0.010)
Chile	0.006	-0.011	-0.001
	(0.016)	(0.020)	(0.011)
India	-0.000	-0.003	-0.002
	(0.020)	(0.026)	(0.013)
Kenya	0.060***	-0.052**	0.029**
	(0.018)	(0.023)	(0.012)
Mexico	0.033**	-0.015	0.016
	(0.015)	(0.020)	(0.010)
NewZealand	-0.012	0.029	-0.003
	(0.015)	(0.020)	(0.010)
SouthAfrica	0.004	0.002	0.004
	(0.014)	(0.020)	(0.010)
Spain	0.002	-0.009	-0.002
	(0.014)	(0.019)	(0.010)
UK	-0.008	-0.005	-0.008
	(0.014)	(0.019)	(0.010)
US	-0.035**	0.036*	-0.019*
	(0.014)	(0.020)	(0.010)
Female	-0.016**	0.023**	-0.005
	(0.007)	(0.009)	(0.005)
Age	0.001*	-0.001***	0.000
	(0.000)	(0.000)	(0.000)
Education is less 4yr college	-0.012*	0.024**	-0.002
	(0.007)	(0.009)	(0.005)
Number of social media platforms used	0.004	-0.004	0.002
	(0.003)	(0.004)	(0.002)
Politically leaning left (standardized)	-0.001	-0.002	-0.002
	(0.003)	(0.004)	(0.002)
Lives in urban area	-0.003	0.014	0.001
	(0.007)	(0.009)	(0.005)
Constant	0.692***	0.315***	0.686***
	(0.025)	(0.031)	(0.021)
Observations	43,565	15,075	58,640
Clusters	2,932	2,928	2,932

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the human correctly rates the LLM if the LLM answered correctly (column (1)), the human correctly rates the LLM if the LLM answered incorrectly (column(2)) and the human correctly rates the LLM in all cases (column (3)). The regressions include fixed effects for questions (100 questions in total) and the presence of a source in it. Each country name represents an indicator variable for that country, where the omitted country is Australia. Robust standard errors, clustered at the participant level, in parentheses. * p<.10; ** p<.05; *** p<.01

$p-$value<0.001). Comparing participants' beliefs to their actual placements, only 19 percent of participants in the third quartile of ability place themselves in that quartile, while 79.2 percent of these participants place themselves in the first and second quartiles of the distribution. Similarly, only 3.3 percent of participants in the bottom quartile of ability place themselves in the correct quartile. These beliefs violate the condition that the largest group of subjects placing themselves in each quartile must belong to that quartile, consistent with overconfidence [Burks et al., 2013, Moore and Healy, 2008].

Result 3. *Individuals overestimate their ability to accurately assess the LLM, in absolute and relative terms.*

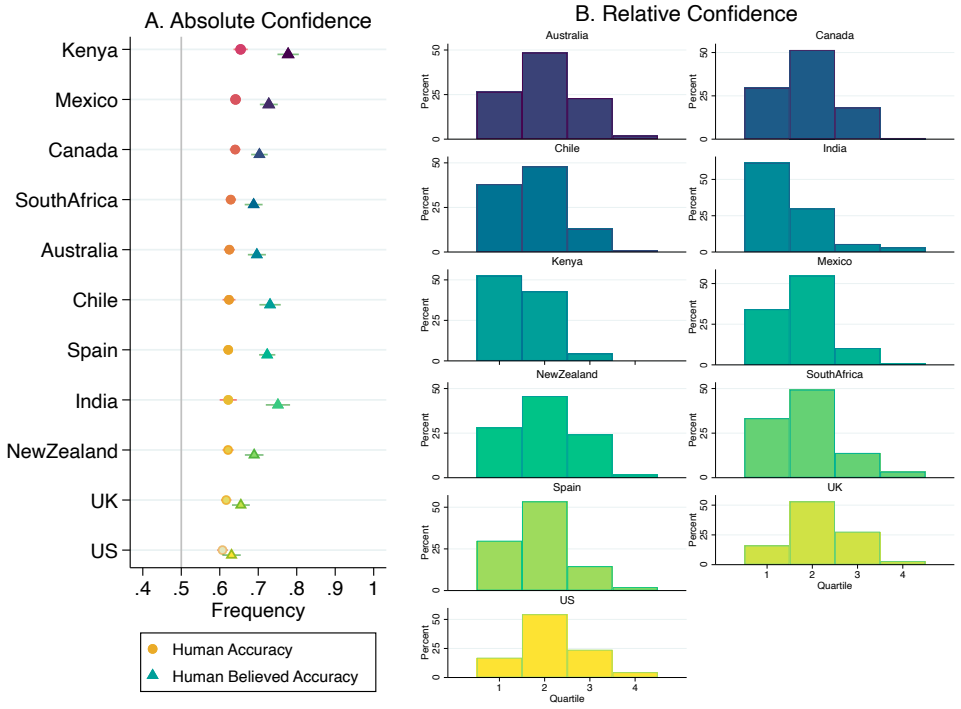


Fig. 5. Human Confidence in Evaluating LLM Output

Notes: Panel A shows average beliefs about the absolute ability to accurately evaluate the LLM (green triangles), comparing them to actual accuracy (orange circles). Panel B shows the distribution of relative confidence, plotting how frequently participants in each country reported believing their quartile of ability was the 1st, 2nd, 3rd, or 4th. See Tables 6 and 12 for detailed results.

Gender differences in confidence are also present in the data, consistent with previous evidence [Bordalo et al., 2019, Coffman, 2014]. Women are significantly less confident in both absolute and relative terms (as shown in Table 6). In terms of absolute confidence, women believe they rate 5.5 percentage points fewer answers correctly – out of 20 questions, this number implies that they believe they correctly evaluate one LLM answer less than men, on average. In relative terms, they also display significantly less confidence: women's belief about the quartile in which they performed in the distribution of ability is 0.21 higher (2.01 on average, compared to 1.80 for men).

4.2 Determinants of Beliefs About Accuracy

The nature of the question asked to the LLM could affect individuals' beliefs about accuracy and, in turn, suggest types of questions about which individuals hold more or less accurate beliefs. Following Hypothesis 4, we examine whether beliefs about the accuracy of the LLM depend on whether it is asked about a matter concerning the US or an international topic and whether it is given a source to use to answer the question. Although the accuracy of the LLM does not vary significantly, the left panel of Figure 7 shows that participants believe that the LLM is marginally more accurate when asked about an international matter (76.2%) than a US matter (73.0%, $t-$test,$p-$value< 0.001),

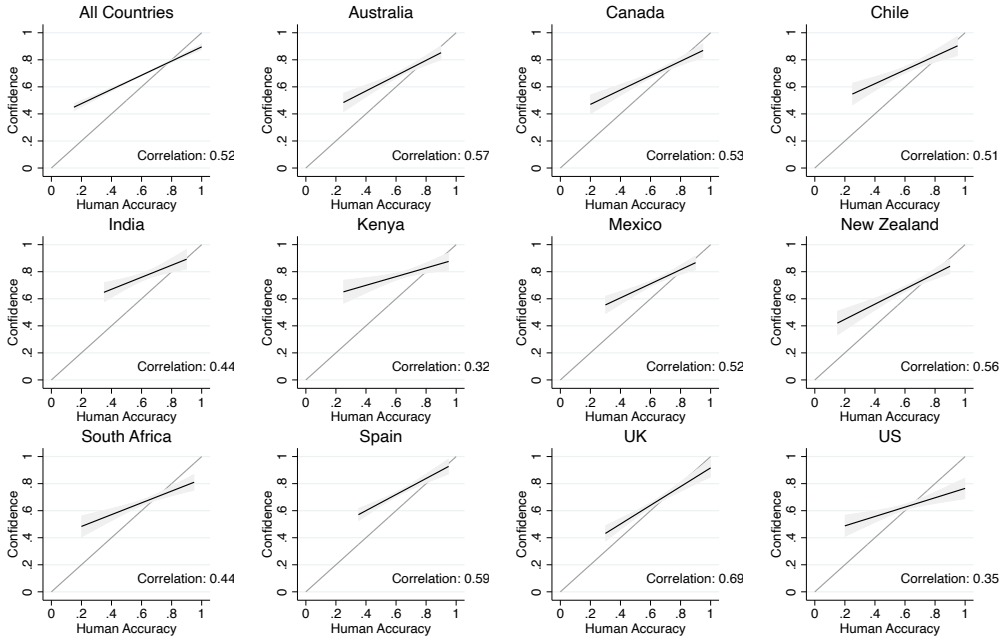


Fig. 6. Correlation between Accuracy and Believed Accuracy (Absolute Confidence)

Notes: Linear relationship between human accuracy and their believed accuracy in evaluating LLM output. Confidence intervals are shown as gray areas underlying each black line. The 45-degree line is shown in light gray in each plot. Each plot shows the linear correlation coefficient, corresponding to the slope shown in the figure.

but equally likely to be accurate when it is provided a source (74.3%) and when it is not (75.1%, $t-$test,$p-$value= 0.213).[7]

To explore other determinants of accuracy, we test whether beliefs differ depending on the topic of the question and find significant differences between questions about climate, where believed accuracy is high (79.3%), while for other topics, excluding crime, it is significantly lower. For example, education and politics are topics that individuals believe the LLM provides accurate responses in 73.2% and 73.7% of the cases.

The heterogeneity across countries raises the question of whether individuals in high-income countries display systematically different beliefs regarding the LLM. We first consider how often the individuals use the LLM and, contrary to Hypothesis 4A, we find that almost 50% of participants report using GPT regularly (49.7%) in low- and middle-income countries, while 36.7% of individuals report using it with the same frequency in high-income countries (see Figure 8). Although individuals from low- and middle-income countries use GPT more, their ability does not differ by a large magnitude from that of individuals in high-income countries (63.3% compared to 62.3% in high-income countries, $t-$test, $p-$value= 0.044).[8]

[7]The country of the participant and the country to which the question asked to the LLM relates is the same in 6.9% of the cases. Comparing the accuracy of participants in these cases to the accuracy in other questions, we do not find significant differences ($t-$test, $p-$value= 0.842).

[8]An important concern in comparing different countries could be that their behavior during the experiment and, in particular, whether they check outside sources, differed. Consistent with previous results regarding this behavior, these differences by country income remain robust when we include controls for such behavior.

Table 6. Human Confidence

	(1)	(2)	(3)
	Human Accuracy and Confidence		
	Human Accurate	Human Absolute Conf. (% out of 20)	Human Relative Conf. (1-4 Quartile)
Canada	0.015	0.007	-0.104*
	(0.010)	(0.016)	(0.059)
Chile	-0.007	0.034*	-0.213***
	(0.012)	(0.019)	(0.066)
India	-0.004	0.055***	-0.436***
	(0.013)	(0.020)	(0.079)
Kenya	0.028**	0.081***	-0.466***
	(0.012)	(0.019)	(0.063)
Mexico	0.014	0.031*	-0.216***
	(0.010)	(0.017)	(0.059)
NewZealand	-0.003	-0.007	-0.029
	(0.010)	(0.017)	(0.062)
SouthAfrica	0.007	-0.008	-0.186***
	(0.010)	(0.017)	(0.061)
Spain	-0.008	0.027*	-0.092
	(0.010)	(0.016)	(0.059)
UK	-0.004	-0.042**	0.153**
	(0.010)	(0.017)	(0.060)
US	-0.019*	-0.066***	0.139**
	(0.010)	(0.017)	(0.060)
Female	-0.004	-0.055***	0.207***
	(0.005)	(0.008)	(0.027)
Age	0.000	0.001***	-0.003**
	(0.000)	(0.000)	(0.001)
Education is less 4yr college	-0.001	-0.037***	0.134***
	(0.005)	(0.008)	(0.029)
Number of social media platforms used	0.001	0.006**	-0.015
	(0.002)	(0.003)	(0.011)
Politically leaning left (standardized)	-0.002	0.011***	-0.048***
	(0.002)	(0.004)	(0.013)
Lives in urban area	0.002	0.006	-0.003
	(0.005)	(0.008)	(0.029)
Constant	0.622***	0.669***	1.998***
	(0.012)	(0.020)	(0.072)
Observations	58,640	2,932	2,932
Clusters	2,932		

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the human correctly rates the LLM (column (1)), their belief in their absolute ability to correctly rate the LLM, measured as the fraction of answers they believe to have correctly rated out of 20 (column(2)) and their belief in their relative ability to rate the LLM, measured as their believed quartile of ability relative to others (column (3)). Each country name represents an indicator variable for that country, where the omitted country is Australia. Robust standard errors, clustered at the participant level, in parentheses. * p<.10; ** p<.05; *** p<.01

There are, however, significant differences in the perceived ability to evaluate the LLM depending on the country's income level. Individuals from low- and middle-income countries overestimate their ability significantly more: They believe their answers are correct 72.4% of the time, while individuals in high-income countries believe their evaluation of the LLM is correct 68.5% of the time ($t-$test, $p-$value< 0.001). These findings suggest that it may be important to raise awareness worldwide, especially in countries where LLMs are used more often, about the limitations and potential mistakes of LLMs as information sources.

Result 4. *In low- and middle-income countries, individuals use LLMs more often and believe they are better at evaluating their accuracy, although they are not.*

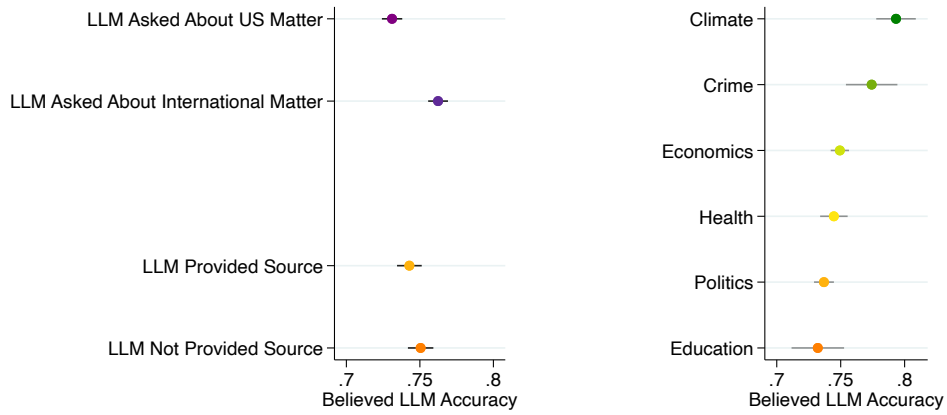


Fig. 7. Determinants of Beliefs about the Accuracy of the LLM

Notes: Average believed LLM accuracy by whether the topic was related to the US or International, by whether the LLM was provided with a source, and by topic. The average is based on regression-adjusted estimates, including country fixed effects and socio-demographic characteristics (gender, age, education, urban location, social media use, and politics). Confidence intervals are shown as gray horizontal lines and are calculated using the Delta method for standard errors clustered at the participant level. See Table 8 for details.

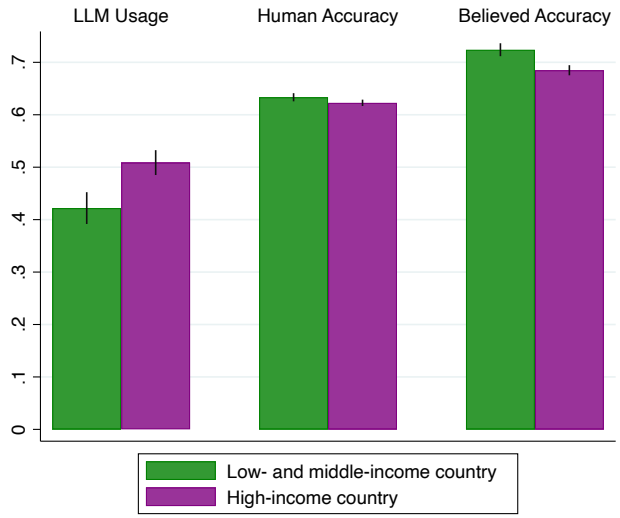


Fig. 8. Usage and Beliefs about LLM Accuracy by Country Income

Notes: Results are shown for low and middle-income countries (Chile, Kenya, India, Mexico, and South Africa) and for high-income countries (Australia, Canada, New Zealand, Spain, UK, and US). The first two columns show the fraction of individuals who report using LLM regularly when asked about their frequency of use with three options "never," "sometimes," and "regularly." The middle pair of columns shows average accuracy in evaluating the LLM answers. The right-most pair of columns shows average belief — the ability to accurately evaluate the LLM's answers. Confidence intervals are shown as dark gray lines and are calculated using the Delta method for standard errors clustered at the participant level. See Table 7 for details.

Table 7. Usage and Beliefs about LLM Accuracy: Country Income Comparison

	(1) Use LLM Regularly = 1	(2) Human Accuracy Average	(3) Believed Accuracy
High-income country	-0.131***	-0.011**	-0.039***
	(0.020)	(0.005)	(0.008)
Female	-0.103***	-0.004	-0.066***
	(0.017)	(0.005)	(0.007)
Age	-0.006***	-0.000	0.001***
	(0.001)	(0.000)	(0.000)
Education is less 4yr college	-0.092***	-0.001	-0.042***
	(0.018)	(0.005)	(0.008)
Number of social media platforms used	0.050***	0.002	0.006**
	(0.007)	(0.002)	(0.003)
Politically leaning left (standardized)	0.029***	-0.002	0.011***
	(0.009)	(0.002)	(0.004)
Lives in urban area	0.053***	0.003	0.015**
	(0.018)	(0.005)	(0.008)
Constant	0.602***	0.630***	0.713***
	(0.041)	(0.011)	(0.018)
Observations	2,932	2,932	2,932

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the individual uses GPT regularly, their accuracy in evaluating the LLM over 20 questions, and their believed accuracy in evaluating the LLM. Robust standard errors, clustered at the participant level, in parentheses. * p<.10; ** p<.05; *** p<.01

5 Discussion and Conclusion

LLMs are increasingly used as an information source that provides efficient summaries to answers that individuals may raise about various policy-relevant topics. Over time, LLMs may come to replace traditional search engines. This change could save users search costs but may increase inaccuracies in beliefs. As this functionality of LLMs expands and users integrate LLMs as an information source in their daily lives, it is important to examine how much they trust the LLM and how able they are to understand its potential mistakes.

Existing research has shown that when it comes to lies and misinformation, these can be generated by LLMs and be hard to detect by humans [Angelucci and Prat, 2024, Belot and Van de Ven, 2017, Bhattacharya et al., 2024, Bond Jr and DePaulo, 2006, Capraro et al., 2024, Evans et al., 2021, Feuerriegel et al., 2023, Greevink et al., 2023, Pennycook and Rand, 2019a,b, Serra-Garcia and Gneezy, 2021]. Proponents of misinformation spreading in society may use this to their benefit, and individuals switching to decisions with help from LLM agents may face harmful consequences [Bhattacharya et al., 2024, Evans et al., 2021]. An important open question that this paper addresses is how humans assess the LLM when it is trying to provide accurate information, considering its worldwide usage and potential cross-cultural differences in how LLMs are perceived. This paper documents significant differences in beliefs about the LLM's ability to be a reliable information source across countries, and in individuals' perceptions of their ability to assess the LLM's information. This finding provides novel evidence in human cognition and highlights the heterogeneity in human beliefs about their own abilities across societies.

While in low- and middle-income countries, individuals are overconfident in their ability, this overconfidence disappears in the US, where participants are skeptical of the LLM's accuracy. The differential skepticism about LLM accuracy raises concerns about potential inequities arising over time across countries, notably in time-sensitive contexts such as public health [De Angelis et al.,

Table 8. Determinants of Believed LLM Accuracy

	(1)	(2)	(3)	(4)
	Human Believes LLM is Accurate = 1			
LLM provided source	-0.008			-0.008
	(0.006)			(0.006)
LLM asked about US matter		-0.031***		-0.026***
		(0.003)		(0.004)
Crime			-0.019	-0.019
			(0.012)	(0.012)
Economics			-0.059***	-0.044***
			(0.008)	(0.008)
Education			-0.070***	-0.061***
			(0.013)	(0.013)
Health			-0.063***	-0.049***
			(0.009)	(0.009)
Politics			-0.069***	-0.056***
			(0.008)	(0.008)
Female	-0.018***	-0.018***	-0.018***	-0.018***
	(0.006)	(0.006)	(0.006)	(0.006)
Age	0.001**	0.001**	0.001**	0.001**
	(0.000)	(0.000)	(0.000)	(0.000)
Education is less 4yr college	-0.014**	-0.014**	-0.015**	-0.015**
	(0.007)	(0.007)	(0.007)	(0.007)
Lives in urban area	-0.005	-0.005	-0.005	-0.005
	(0.007)	(0.007)	(0.007)	(0.007)
Number of social media platforms used	0.003	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)	(0.003)
Politically leaning left (standardized)	-0.000	-0.000	-0.000	-0.000
	(0.003)	(0.003)	(0.003)	(0.003)
Constant	0.730***	0.742***	0.785***	0.789***
	(0.017)	(0.017)	(0.018)	(0.018)
Observations	58,640	58,640	58,640	58,640
Clusters	2,932	2,932	2,932	2,932

Notes: This table displays the estimated coefficients from linear regressions on the likelihood that the human believes the LLM answered correctly. The regressions include country fixed effects. The variables Crime, Economics, Education, Health, and Politics are indicators for the topic of the question asked to the LLM, where Climate is the omitted category. The regressions include sociodemographic characteristics as controls. Robust standard errors, clustered at the participant level, in parentheses. * p<.10; ** p<.05; *** p<.01

2023, Meyrowitsch et al., 2023, Nan et al., 2022], voting [Time-Magazine, 2024], national security [Horowitz and Kahn, 2024]. If individuals differentially and, often incorrectly, asses the abilities of LLMs as information sources, the use of LLMs could be more harmful in low and middle-income countries than in high-income ones.

Research focusing on how LLMs are adopted should consider these potential global differences. It will be important to educate citizens on how much trust to extend to LLMs, and further examine what measures they can take (such as checking the LLM's answers) to ensure the correct use of LLM-provided information, taking country-specific differences into account, as new regulations are rolled out [Gibney, 2024, Warren, 2024].

References

Charles Angelucci and Andrea Prat. 2024. Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News. *American Economic Review* 114, 4 (2024), 887–925.

Antonio A Arechar, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, et al. 2023. Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour* 7, 9 (2023), 1502–1513.

Michele Belot and Jeroen Van de Ven. 2017. How private is private information? The ability to spot deception in an economic game. *Experimental economics* 20 (2017), 19–43.

Haimanti Bhattacharya, Subhasish Dugar, Sanchaita Hazra, and Bodhisattwa Prasad Majumder. 2024. The Good, the Bad, and the Ugly: The Role of AI Quality Disclosure in Lie Detection. *arXiv preprint arXiv:2410.23143* (2024).

Charles F Bond Jr and Bella M DePaulo. 2006. Accuracy of deception judgments. *Personality and social psychology Review* 10, 3 (2006), 214–234.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. Beliefs about gender. *American Economic Review* 109, 3 (2019), 739–773.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

Stephen V Burks, Jeffrey P Carpenter, Lorenz Goette, and Aldo Rustichini. 2013. Overconfidence and social signalling. *Review of Economic Studies* 80, 3 (2013), 949–983.

Valerio Capraro, Austin Lentsch, Daron Acemoglu, Selin Akgun, Aisel Akhmedova, Ennio Bilancini, Jean-François Bonnefon, Pablo Brañas-Garza, Luigi Butera, Karen M Douglas, et al. 2024. The impact of generative artificial intelligence on socioeconomic inequalities and policy making. *PNAS nexus* 3, 6 (2024).

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).

Katherine Baldiga Coffman. 2014. Evidence on self-stereotyping and the contribution of ideas. *The Quarterly Journal of Economics* 129, 4 (2014), 1625–1660.

Alain Cohn, Michel André Maréchal, David Tannenbaum, and Christian Lukas Zünd. 2019. Civic honesty around the globe. *Science* 365, 6448 (2019), 70–73.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in public health* 11 (2023), 1166120.

Maria del Rio-Chanona, Nadzeya Laurentsyeva, and Johannes Wachs. 2023. Are large language models a threat to digital public goods? evidence from activity on stack overflow. *arXiv preprint arXiv:2307.07367* (2023).

Eugen Dimant and Erik O Kimbrough. 2024. Polarization in multidisciplinary perspective. *PNAS Nexus* 3, 10 (2024), pgae425.

Anil R Doshi and Oliver P Hauser. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances* 10, 28 (2024), eadn5290.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674* (2021).

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature* 630, 8017 (2024), 625–630.

Stefan Feuerriegel, Renée DiResta, Josh A Goldstein, Srijan Kumar, Philipp Lorenz-Spreen, Michael Tomz, and Nicolas Pröllochs. 2023. Research can help to tackle AI-generated disinformation. *Nature Human Behaviour* 7, 11 (2023), 1818–1821.

Michele J Gelfand, Jana L Raver, Lisa Nishii, Lisa M Leslie, Janetta Lun, Beng Chong Lim, Lili Duan, Assaf Almaliach, Soon Ang, Jakobina Arnadottir, et al. 2011. Differences between tight and loose cultures: A 33-nation study. *science* 332, 6033 (2011), 1100–1104.

Elizabeth Gibney. 2024. What the EU's tough AI law means for research and ChatGPT. *Nature* (2024).

Laura K Globig, Rachel Xu, Steve Rathje, and Jay J Van Bavel. 2024. Perceived (Mis) alignment in Generative Artificial Intelligence Varies Across Cultures. *Preprint. DOI* 10 (2024).

Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is AI-generated propaganda? *PNAS nexus* 3, 2 (2024), pgae034.

Ivo Greevink, Theo Offerman, and Giorgia Romagnoli. 2023. *AI-Powered Promises: The Influence of ChatGPT on Trust and Trustworthiness.* Technical Report. Technical report, AEA RCT Registry, https://doi. org/10.1257/rct. 11511-1.0.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bi Puranen, et al. 2020. World values survey: Round seven–country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat* 7 (2020), 2021.

Michael C Horowitz and Lauren Kahn. 2024. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly* 68, 2 (2024), sqae020.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. *arXiv preprint arXiv:2407.03282* (2024).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

Charles I Jones. 2024. The ai dilemma: Growth versus existential risk. *American Economic Review: Insights* 6, 4 (2024), 575–590.

Erik O Kimbrough, Erin L Krupka, Rajnish Kumar, Jennifer M Murray, Abhijit Ramalingam, Sharon Sánchez-Franco, Olga L Sarmiento, Frank Kee, and Ruth F Hunter. 2024. On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. *Experimental Economics* (2024), 1–28.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science* 9, 1 (2022), 104–117.

Michael D Lee and Megan N Lee. 2017. The relationship between crowd majority and accuracy for binary decisions. *Judgment and Decision Making* 12, 4 (2017), 328–343.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747* (2023).

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334* (2022).

Dan W Meyrowitsch, Andreas K Jensen, Jane B Sørensen, and Tibor V Varga. 2023. AI chatbots and (mis) information in public health: impact on vulnerable communities. *Frontiers in Public Health* 11 (2023), 1226776.

Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.

Xiaoli Nan, Yuan Wang, and Kathryn Thier. 2022. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine* 314 (2022), 115398.

Arvind Narayanan and Sayash Kapoor. 2024. *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference.* Princeton University Press.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590* (2023).

Gordon Pennycook and David G Rand. 2019a. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

Gordon Pennycook and David G Rand. 2019b. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.

Reddit. 2023a. *Anyone else basically done with Google Search in 2023?* https://www.reddit.com/r/ChatGPT/comments/13ik8wh/anyone_else_basically_done_with_google_search_in/ Accessed: 18-Oct-2024.

Reddit. 2023b. *Anyone else basically done with Google Search in 2023?* https://www.reddit.com/r/ChatGPT/comments/13ik8wh/anyone_else_basically_done_with_google_search_in/ Accessed: 18-Oct-2024.

Marta Serra-Garcia and Uri Gneezy. 2021. Mistakes, overconfidence, and the effect of sharing on detecting lies. *American Economic Review* 111, 10 (2021), 3160–3183.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652* (2023).

SimilarWeb. 2024. *chat.openai.com Traffic Analytics & Market Share.* https://www.similarweb.com/website/chat.openai.com/ Accessed: 16-Oct-2024.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.

J Surowiecki. 2005. The Wisdom of the Crowds.

Time-Magazine. 2024. How to Protect Yourself from AI Election Misinformation. *Time* (2024). https://time.com/7022120/ai-election-misinformation-2024/ Accessed: 2025-02-10.

Zach Warren. 2024. Legalweek 2024: Current US AI Regulation Means Adopting a Strategic — and Communicative — Approach. *Thomson Reuters Institute* (2024). https://www.thomsonreuters.com/en-us/posts/legal/legalweek-2024-ai-regulation Accessed: 2025-02-10.

A Inclusion and Ethics

This research was conducted under IRB #810990. The authors provided equal payment to all participants of the experiment, which was conducted on the online platform Prolific Academic.

B Data, Materials, and Software Availability

The pre-registration can be found at https://aspredicted.org/9b8r-t5wd.pdf. All the original data, coding, and STATA scripts to perform the analysis and visualizations will be deposited on the OSF depository (https://osf.io/g5rp3/?viewonly=6a704f2e1f654064bcdd073ae137d2ef). Currently, the depository includes additional experimental materials (all 100 questions asked and answered by the LLM).

C Additional Results: Human Beliefs and Accuracy

C.1 Demographics

Table 9 shows the average values of each socio-demographic characteristic across countries.

Table 9. Socio-demographics by Country

	Australia	Canada	Chile	Kenya	India	Mexico
Female	0.49	0.49	0.34	0.30	0.39	0.49
Age	34.98	35.05	29.05	32.15	27.73	29.59
Education is less 4yr college	0.33	0.35	0.35	0.15	0.33	0.26
Lives in urban area	0.32	0.50	0.60	0.62	0.45	0.64
Number of social media platforms used	2.43	2.47	3.01	2.92	3.63	3.69
Observations	305	307	200	127	175	301

	NewZealand	Spain	SouthAfrica	UK	US	Total
Female	0.56	0.71	0.37	0.62	0.56	0.50
Age	35.92	29.47	32.53	40.19	39.08	33.67
Education is less 4yr college	0.40	0.41	0.34	0.41	0.44	0.35
Lives in urban area	0.35	0.27	0.44	0.22	0.24	0.40
Number of social media platforms used	2.36	3.57	2.93	2.72	2.34	2.89
Observations	284	324	301	303	305	2932

C.2 Beliefs about LLM Accuracy

Table 10 shows regression-adjusted, country-specific averages of LLM believed accuracy and actual accuracy.

Table 10. Human Beliefs and Actual LLM Accuracy: Country Level Averages

	(1)	(2)
	Human Beliefs and Actual LLM Accuracy	
	Belief LLM is accurate = 1	LLM is accurate= 1
Australia	0.744	0.745
	(0.010)	(0.002)
Canada	0.753	0.748
	(0.010)	(0.003)
Chile	0.752	0.740
	(0.012)	(0.003)
India	0.745	0.744
	(0.017)	(0.005)
Kenya	0.802	0.740
	(0.013)	(0.002)
Mexico	0.772	0.736
	(0.010)	(0.003)
NewZealand	0.727	0.742
	(0.010)	(0.003)
SouthAfrica	0.747	0.747
	(0.010)	(0.003)
Spain	0.747	0.742
	(0.010)	(0.002)
UK	0.739	0.745
	(0.009)	(0.002)
US	0.709	0.742
	(0.010)	(0.002)
Observations	58640	58640

Notes: This table displays regression-adjusted averages of LLM believed accuracy (column (1)) and actual accuracy (column (2)) for each country, based on the estimation in Table 4. Robust standard errors, using the Delta method, in parentheses.

C.3 Human Accuracy

Throughout the analyses in the paper, to measure accuracy, we focus on how often participants correctly evaluate an LLM answer, conditional on the answer of the LLM being correct or incorrect. These measures are closely related to Type I and II errors. Although we expected to use them initially, the fact that LLM accuracy is significantly better than 50% (74%), means that the frequency of Type I and II errors is highly dependent on the LLM accuracy rate. In the extreme case that the LLM was always correct, the frequency of Type I errors, participants believing it is correct when it is not, would always be 0%. Hence, we measure accuracy conditional on whether the LLM is correct or incorrect.

Table 11. Human Accuracy: Country Level Averages

	(1)	(2)	(3)
	Human Accuracy		
	Human Accurate if LLM is correct = 1	Human Accurate if LLM is incorrect = 1	Human Accurate All
Australia	0.748	0.266	0.625
	(0.010)	(0.014)	(0.007)
Canada	0.763	0.275	0.640
	(0.010)	(0.013)	(0.007)
Chile	0.754	0.255	0.624
	(0.013)	(0.015)	(0.009)
India	0.747	0.263	0.623
	(0.017)	(0.022)	(0.011)
Kenya	0.807	0.214	0.654
	(0.014)	(0.018)	(0.009)
Mexico	0.781	0.251	0.641
	(0.010)	(0.014)	(0.007)
NewZealand	0.735	0.295	0.622
	(0.011)	(0.014)	(0.007)
SouthAfrica	0.752	0.268	0.629
	(0.010)	(0.014)	(0.007)
Spain	0.749	0.258	0.623
	(0.010)	(0.013)	(0.006)
UK	0.739	0.261	0.617
	(0.010)	(0.013)	(0.007)
US	0.712	0.303	0.606
	(0.010)	(0.014)	(0.007)
Observations	43565	15075	58640

Notes: This table displays regression-adjusted averages of human accuracy if the LLM is correct (column (1)), if the LLM is incorrect (column (2)), and for all answers (column(3)), for each country, based on the estimation in Table 5. Robust standard errors, using the Delta method, in parentheses.

C.4 Confidence

Table 12 highlights country-wide averages of human accuracy, absolute confidence, and relative confidence.

Table 12. Human Confidence: Country Level Averages

	(1)	(2)	(3)
		Human Accuracy and Confidence	
	Human Accurate	Human Absolute Confidence (% out of 20)	Human Relative Confidence (1-4 Quartile)
Australia	0.626	0.696	2.006
	(0.007)	(0.012)	(0.043)
Canada	0.640	0.703	1.902
	(0.007)	(0.011)	(0.040)
Chile	0.619	0.731	1.794
	(0.009)	(0.014)	(0.050)
India	0.621	0.751	1.570
	(0.011)	(0.016)	(0.066)
Kenya	0.653	0.778	1.541
	(0.010)	(0.014)	(0.045)
Mexico	0.640	0.727	1.790
	(0.007)	(0.012)	(0.039)
NewZealand	0.623	0.689	1.978
	(0.008)	(0.012)	(0.046)
SouthAfrica	0.633	0.688	1.820
	(0.007)	(0.012)	(0.043)
Spain	0.618	0.723	1.915
	(0.007)	(0.011)	(0.041)
UK	0.621	0.655	2.160
	(0.007)	(0.012)	(0.043)
US	0.606	0.631	2.145
	(0.007)	(0.012)	(0.043)
Observations	58640	2932	2932

Notes: This table displays regression-adjusted averages of human accuracy (column (1)), absolute confidence (column (2)), and relative confidence (column(3)), for each country, based on the estimation in Table 6. Robust standard errors, estimated using the Delta-method, in parentheses.

C.5 Checking Outside Sources

This table presents regression-adjusted mean frequencies of participants reporting that they checked at least one answer provided by the LLM online.

Table 13. Outside Consult: Country Level Averages

	(1)	(2)	(3)
	Individual Behavior & Beliefs		
	Checked Online LLM Output	Human Accuracy if did NOT check	Human Accuracy if checked
Australia	0.288	0.664	0.751
	(0.026)	(0.033)	(0.048)
Canada	0.298	0.669	0.658
	(0.026)	(0.032)	(0.052)
Chile	0.327	0.662	0.699
	(0.034)	(0.043)	(0.057)
India	0.536	0.762	0.554
	(0.044)	(0.058)	(0.061)
Kenya	0.377	0.692	0.650
	(0.038)	(0.046)	(0.058)
Mexico	0.341	0.609	0.625
	(0.028)	(0.036)	(0.049)
NewZealand	0.312	0.580	0.667
	(0.028)	(0.036)	(0.052)
SouthAfrica	0.315	0.669	0.586
	(0.027)	(0.033)	(0.050)
Spain	0.296	0.643	0.679
	(0.026)	(0.033)	(0.050)
UK	0.139	0.605	0.736
	(0.020)	(0.031)	(0.078)
US	0.189	0.574	0.521
	(0.023)	(0.032)	(0.073)
Observations	2932	2071	861

Notes: This table displays regression-adjusted averages of how frequently participants report to have checked at least one answer provided by the LLM online (column (1)), their accuracy if they report not to have checked (column (2)), and their accuracy if the report to have checked (column(3)), for each country. The regressions include individual socio-demographic characteristics. Robust standard errors, estimated using the Delta method, are in parentheses.

D Experimental Design: Additional Information

D.1 Participant Instructions

Welcome

This Study - You will make several decisions in this study.
Because any decision may affect your BONUS payment, please read closely all the information you are shown. Thank you!

Task Instructions

This Study
In this study, we will show you 20 questions we asked to ChatGPT (version 4).
We will also show you the answers that ChatGPT provided. If ChatGPT listed any sources, they will appear in the answer as sources or in brackets.

Your Task
Your task is to evaluate whether the answers provided by ChatGPT are correct. Each answer is either correct or incorrect.

Your Bonus
We will select one question at random.
If your evaluation is correct (your selected option matches whether GPT is correct or not), you will receive a bonus of $2.

Understanding Questions Below are two understanding questions that you are asked to answer before we proceed with the study. You will have two attempts to answer them correctly.
Understanding Question 1. What is your task?

- Answer questions that ChatGPT also answered.
- Evaluate the answers provided by ChatGPT as correct or incorrect.
- Evaluate the answers provided by ChatGPT on a scale between 0 and 10, where 0 is incorrect, and 10 is correct.

Understanding Question 2. How is your bonus payment determined?

- I receive $1 no matter what my answers are.
- I receive $2 for completing the study and answering all the questions.
- I receive $2 if in a randomly selected question my evaluation for the GPT answer is correct.

D.2 Example from the Task

Question we asked ChatGPT (1 of 20)
Anti-Muslim bias in the United States has been rising, manifesting in increased discrimination, hate crimes, and negative stereotypes against Muslim individuals and communities. What is the difference in the number of Anti-muslim bias complaints received from the start to the end of Donald Trump's presidential rule in the United States 2017-2021? Please cite a source for your response.

Answer it provided:
Between the start and end of Donald Trump's presidency from 2017 to 2021, there was a significant increase in anti-Muslim bias complaints. In 2017, CAIR reported over 2,500 complaints of anti-Muslim bias incidents. By 2021, this number had increased to over 6,000 complaints. This indicates a notable rise in reported incidents of anti-Muslim discrimination and bias during this period (ArabAmericanNews).

What do you think?

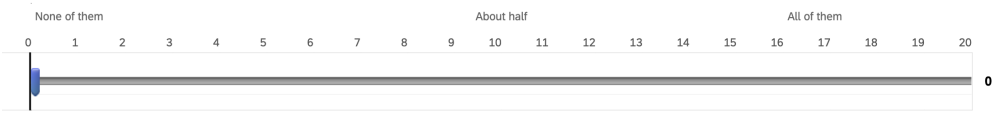
The answer is correct.

The answer is incorrect.

D.3 Additional Questions

Thank you for completing the task. In what follows, we will ask you a few more questions and then conclude the study.

1. How many of the 20 questions and answers by GPT that you evaluated do you think you evaluated correctly? If your answer is correct, you will receive an additional $1 bonus.



2. Compared with previous participants in this experiment, how well do you think you did? We ask you to choose a quartile. If your answer is correct, you will receive an additional $1 bonus. Relative to the other participants, my number of correct answers is in the following quartile:

- **Quartile 4:** 75th-100th percentile (better than at least 75% of participants).
- **Quartile 3:** 50th-75th percentile
- **Quartile 2:** 25th-50th percentile
- **Quartile 1:** 0th-25th percentile (worse than at least 75% of participants)

3. What is your experience with ChatGPT?

- Never used it.
- Used it only a few times
- Use it regularly

4. Do you have a paid subscription to ChatGPT?

- No
- Yes

5. In what ways, if any, is ChatGPT useful for you? Please explain briefly. [Text Box Entry]

6. To answer the questions on whether the GPT-provided answer was correct, did you check yourself what the correct answer was elsewhere (on ChatGPT, or online, or else)? Please respond truthfully, you will be paid your bonus regardless of your answer.

- I did not check elsewhere whether the GPT-provided answers were correct.
- I checked elsewhere whether (some of) the GPT-provided answers were correct.

About you

1. How old are you (in years)? [select number from list below]
2. What is your gender?

- Male
- Female
- Other [Text Box Entry]

3. In political matters, people talk of "the left" and "the right." How would you place your views on this scale, considering 1 is left and 10 is right, generally speaking? [scale presented]

4. Which of the following best describes the place where you currently live?

- A large city

- A suburb near a large city
- A small city or town
- A rural area

5. Please select the level of education that you have completed.

- Elementary School
- Middle School
- High School or equivalent
- Some college
- College Graduate with Associate's Degree (2 year)
- College Graduate with Bachelor's Degree (4 year)
- Master's Degree (MS)
- Doctoral Degree (PhD)
- Professional Degree (MD, JD, etc.)
- Other [Text Box Entry]

6. Which social media platforms do you regularly use? [multiple options could be selected]

- Facebook
- Instagram
- Tiktok
- Twitter
- Whatsapp
- Other, please list here [Text Box Entry]
- None, I do not use social media

E Coding Instructions: LLM Accuracy

Each response given by GPT has to be coded into "Correct" or "Incorrect." Participants in an experiment saw "screenshots" of the question and answer, and provided their own rating (correct or incorrect). We will have 3 coders classify each statement from US and international, and 2 responses by GPT, depending on whether it was given the source or not. We will then meet and reach an agreement if there's a discrepancy between us.

What is CORRECT?

An answer is correct if it coincides with the "correct answer" column in the google sheet. [*Coders were provided with a sheet that contained all statements and correct answers according to the source consulted (and provided as a link). They also had access to the questions, answers and screenshots for the correct answers in separate documents.*]. Then it's from the source used in the source treatment. Or from a different source that republished/requoted the source from the source treatment.

If the answer does not "coincide" with the correct answer in the google sheet:

- It does not give ANY specific answer. This could be when it reports it's increased/decreased, but does not provide a %, although we asked for it.
  - Then, we classify it as CORRECT.
  - Because it's not explicitly incorrect, and from the perspective of the reader, they're not getting the information requested, but they're not getting incorrect information either.
- If it gives a "specific" or "numerical" answer, that differs from the source:
  - First, evaluate the sources (searched sites) for their reliability or the sources that they themselves use.
    - Official government websites
    - UN, WB, and other international organizations
    - Mainstream media
  - Case 1: The numbers reported by GPT in the answer correspond correctly to those listed in that source and the source is reliable, then the answer is CORRECT.
  - Case 2: The numbers reported by GPT in the answer do not correspond to those in the source. They could be on that source/site/page but apply to a different outcome/variable than the one we asked about. Regardless of whether the source is reliable or not, the answer is INCORRECT.
  - Case 3: If the numbers are similar in magnitude to those in the source listed (and ideally somewhat close to the original source), then the answer is CORRECT. Example would be: approximately 40% is the original source, GPT reports with a different source 36%, that's CORRECT.