nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00976-7

What large language models know and what people think they know

Received: 2 May 2024

Accepted: 13 December 2024

Published online: 21 January 2025

Check for updates

Mark Steyvers[1], **Heliodoro Tejeda**[1], **Aakriti Kumar**[1], **Catarina Belem**[2], **Sheer Karny**[1], **Xinyue Hu**[1], **Lukas W. Mayer**[1] **& Padhraic Smyth**[2]

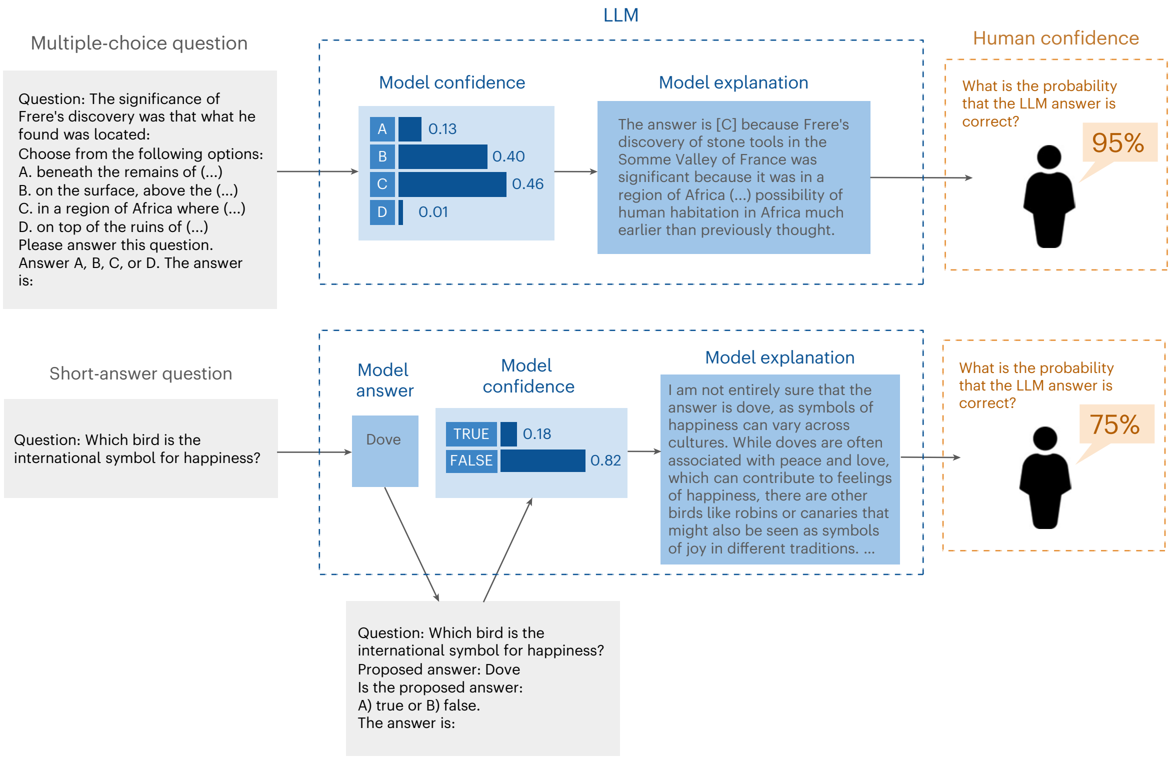
As artificial intelligence systems, particularly large language models (LLMs), become increasingly integrated into decision-making processes, the ability to trust their outputs is crucial. To earn human trust, LLMs must be well calibrated such that they can accurately assess and communicate the likelihood of their predictions being correct. Whereas recent work has focused on LLMs' internal confidence, less is understood about how effectively they convey uncertainty to users. Here we explore the calibration gap, which refers to the difference between human confidence in LLM-generated answers and the models' actual confidence, and the discrimination gap, which reflects how well humans and models can distinguish between correct and incorrect answers. Our experiments with multiple-choice and short-answer questions reveal that users tend to overestimate the accuracy of LLM responses when provided with default explanations. Moreover, longer explanations increased user confidence, even when the extra length did not improve answer accuracy. By adjusting LLM explanations to better reflect the models' internal confidence, both the calibration gap and the discrimination gap narrowed, significantly improving user perception of LLM accuracy. These findings underscore the importance of accurate uncertainty communication and highlight the effect of explanation length in influencing user trust in artificial-intelligence-assisted decision-making environments.

Uncertainty communication plays a critical role in decision-making and policy development. Uncertainties are often expressed verbally to help stakeholders understand risks and make informed choices across a wide range of domains, including climate policy, law, medicine and intelligence forecasting. Psychology research has investigated perceptions of verbally expressed uncertainty (for example, phrases such as 'very unlikely' or 'almost certain') in these domains[1–5]. Despite their lack of precision in communicating probabilities, verbal probability phrases provide a simple and effective way to communicate uncertainty in natural language contexts. The emergence of large language models (LLMs) introduces new complexities in the area of uncertainty communication. These models are increasingly integrated into areas such as public health[6], coding[7] and education[8]. However, the question of how effectively LLMs communicate uncertainty is unexplored. As the primary mode of communication with LLMs is through natural language, it is critical to understand whether LLMs are able to accurately convey through verbal means what they know or do not know.

Recent research raises doubts about the reliability of the information that LLMs generate. One notable issue is the possibility of generating responses that, while convincing, may be inaccurate or nonsensical[9,10]. The unreliability of LLMs has led developers of LLMs to caution against the uncritical acceptance of model outputs[11], suggesting that it is not always clear when the models are or are not confident in the knowledge communicated to the user.

---

[1]Department of Cognitive Sciences, University of California, Irvine, CA, USA. [2]Department of Computer Science, University of California, Irvine, CA, USA. e-mail: mark.steyvers@uci.edu



Question: The significance of Frere's discovery was that what he found was located:
Choose from the following options:
A. beneath the remains of (...)
B. on the surface, above the (...)
C. in a region of Africa where (...)
D. on top of the ruins of (...)
Please answer this question.
Answer A, B, C, or D. The answer is:

The answer is [C] because Frere's discovery of stone tools in the Somme Valley of France was significant because it was in a region of Africa (...) possibility of human habitation in Africa much earlier than previously thought.

What is the probability that the LLM answer is correct?

Question: Which bird is the international symbol for happiness?

I am not entirely sure that the answer is dove, as symbols of happiness can vary across cultures. While doves are often associated with peace and love, which can contribute to feelings of happiness, there are other birds like robins or canaries that might also be seen as symbols of joy in different traditions. ...

What is the probability that the LLM answer is correct?

Question: Which bird is the international symbol for happiness?
Proposed answer: Dove
Is the proposed answer:
A) true or B) false.
The answer is:

Fig. 1 | Overview of the evaluation methodology for assessing the calibration gap between model confidence and human confidence in the model. The multiple-choice questions, the approach works as follows: (1) prompt the LLM with a question to obtain the model's internal confidence for each answer choice; (2) select the most likely answer and prompt the model a second time to generate an explanation for the given answer; (3) obtain the human confidence by showing users the question and the LLM's explanation and asking users to indicate the probability that the model is correct. In this toy example, the model confidence for the multiple-choice question is 0.46 for answer C, whereas the human confidence is 0.95. For short-answer questions, the approach is similar except that internal model confidence is obtained by an additional step where the LLM is prompted to evaluate whether the previously provided answer to the question is true or false[13]. In the short-answer question example, the LLM model explanation was modified with uncertainty language to convey the low model confidence (0.18). For the two toy examples, the correct answers are 'A' and 'blue bird'.

At the same time, recent research has also indicated that LLMs have the ability, to a certain degree, to accurately discern their own knowledge boundaries. LLMs in particular can exhibit a reasonable level of calibration for multiple-choice questions such that the probability the model assigns to a selected answer tracks with the probability that this answer is correct[12–14]. In addition, recent studies show that LLMs can distinguish between answerable and unanswerable questions[13,15] and the internal state of an LLM can distinguish between truthful statements and lies[16] and truthful statements and confabulations[17]. These findings suggest that LLMs may possess an internal mechanism that is reflective of self-knowledge.

In the specific context of question-answering, an LLM's model confidence is typically equated to the probability assigned by the LLM to the selected answer relative to other possible answers (for example, refs. 18,19). However, from the perspective of a human interacting with the LLM, this internal model confidence is not usually displayed to human users as part of LLM output. Instead, in current practice, humans rely solely on the language produced by the LLM to assess LLM confidence. To contrast with model confidence, in this article we use the term human confidence to refer to a human's assessment (expressed as a probability) of how likely it is that the LLM's answer is correct based only on the language produced by the LLM without any knowledge of the LLM's internal model confidence.

Surprisingly, studies focused on investigating human confidence in LLM outputs are lacking. In this article, we take a step in addressing this issue and investigate what we term the calibration gap, namely the difference in the reliability of LLM model confidence and human confidence. In effect, the calibration gap represents the gap between an LLM's own internal confidence of what it knows and human perception of this confidence. In addition, we investigate the discrimination gap, which relates to the difference in the ability to discriminate between probably correct and incorrect answers. Any discrimination gap shows that whatever internal LLM representation is used to tell the difference between probably correct and incorrect answers is not conveyed effectively to humans. We address two specific research questions in this context. First, how large are the calibration and discrimination gaps? That is, is there a notable gap between LLM model confidence and human confidence in terms of how each assesses the true accuracy of the LLM? Second, can the calibration and discrimination gaps be reduced? Can the quality of human confidence in an LLM be improved by adapting the textual output of the LLM to internal model confidence? These questions have important implications for the design of reliable LLM assistants. By aligning the LLM's internal confidence with human perception of this confidence, we can bridge the gap between what LLMs know and what people think they know, which is crucial for the development of effective and trustworthy assistants

Our contributions in this context are twofold. First, we present a set of experimental studies and a dataset that directly captures human assessment of LLM confidence in a question-answering context, providing insight into human perceptions of LLM textual responses. Second, we test and suggest ways of generating LLM responses that improve the calibration quality of human confidence relative to the LLM assistant's model confidence and the LLM's true accuracy.

LLMs

We use three publicly available LLMs in our studies: GPT-3.5 (ref. 20), PaLM2 (ref. 21) and GPT-4o. We apply the GPT-3.5 and PaLM2 models to a subset of multiple-choice questions from the Massive Multitask

Table 1 | Overview of experiments

Experiment	Question type	LLM	Explanation type	Number of participants
1a	Multiple choice	GPT-3.5	Default explanations	41
1b	Multiple choice	PaLM2	Default explanations	39
1c	Short answer	GPT-4o	Default explanations	42
2a	Multiple choice	GPT-3.5	Modified explanations	60
2b	Multiple choice	PaLM2	Modified explanations	60
2c	Short answer	GPT-4o	Modified explanations	59

Language Understanding (MMLU) dataset, a comprehensive dataset that contains multiple-choice questions from various knowledge domains, such as science, technology, engineering and mathematics (STEM), humanities, social sciences and more[19]. We apply the GPT-4o model to a subset of short-answer questions from the Trivia QA dataset[22]. For each multiple-choice and short-answer question, we assess the model confidence by computing the token likelihoods (see the Methods for details). This method for reading out model confidence allows a direct computation of the relative probabilities of different possible answers in multiple-choice questions[12,13,18,19,23] and the probability that the answer to an open-ended question is correct[13,17]. We investigate the relationship between model confidence and accuracy to determine whether the LLM is reasonably well calibrated, independent of the LLM's ability to elicit well-calibrated confidence from humans who use the LLM.

Behavioural experiments

We designed behavioural experiments to evaluate human perceptions of LLM confidence. In these experiments, participants estimate the probability that the LLM's answer to a multiple-choice or short-answer question is correct based on the explanation that the LLM provided (Fig. 1). Participants are not provided any direct access to the LLM's numerical model confidence, allowing us to make inferences about participants' perceptions of the confidence of the LLM based on model explanations alone. In addition, for the multiple-choice questions part of the experiment only, with the assistance of the LLM, participants provided answers to the questions. Previous research has demonstrated that the MMLU multiple-choice questions are difficult for participants who lack domain expertise, resulting in near-chance accuracy[19]. We anticipate that these questions will be difficult to answer without the assistance of the LLM because the majority of the participants in our experiments lack domain expertise, and their perception of the explanation's content will influence their evaluation more than their own knowledge.

We conducted two experiments each involving the three types of LLMs and two types of questions (see Table 1 for an overview). Experiment 1 assesses human perceptions of LLM accuracy using the LLM's default explanations for either multiple-choice or short-answer questions. The results from this experiment allow us to address the first research question regarding the size of the calibration and discrimination gap between model and human confidence. Experiment 2 manipulates the prompts to produce three levels of uncertainty language (low, medium and high confidence) and three levels of explanation length, resulting in nine different types of explanations presented to participants. The prompts are designed to include uncertainty language corresponding to model confidence at the start of the explanation. Table 2 illustrates explanations from a particular multiple-choice question used in the experiments (see Supplementary Table 2 for the full model explanations). The results from this experiment serve two purposes. First, we establish that human confidence varies with the uncertainty language and the length of the explanation. Next, we use the results from experiment 2 to answer the second research question, which is to understand how the calibration and discrimination gap can be reduced by aligning the uncertainty language with model confidence—showing a low-, medium- and high-confidence explanation when the model has low, medium and high confidence, respectively. The Supplementary Information ('experiment 3') reports the results from an additional experiment with a different prompting approach that alters the default explanations from experiment 1. We use the two metrics to assess the relationship between human and model confidence and model accuracy. See Methods for details.

Table 2 | Example explanations for the answer by GPT-3.5 to a multiple-choice question used in experiments 1a and 2a

Prompt style	Explanation
Experiment 1a	
Baseline	The answer is [A] Velocity. This is because in an inelastic collision, the two objects involved will stick together and move with a common velocity after the collision. […] In this case, the initial momentum of the 1.5-kg cart is transferred to the combined mass of the two carts after the collision, resulting in the same velocity for both carts. Therefore, the correct answer is [A] Velocity.
Experiment 2a	
Low confidence and long	I am not completely sure that the answer is [A] because velocity is a vector quantity that includes both magnitude and direction. In this scenario, the carts may have the same magnitude of velocity, but they may not necessarily have the same direction. Therefore, I am not completely convinced that the answer is [A]. […] In this case, the carts would indeed have the same velocity as each other.
Medium confidence and long	I am somewhat sure that the answer is [A] because velocity is a vector quantity that takes into account both the speed and direction of an object's motion. In this scenario, the two carts are colliding and sticking together, […] Therefore, I am somewhat sure that the answer is [A] because the velocity of the two carts will be the same after the collision due to the conservation of momentum principle.
High confidence and long	I am sure that the answer is [A] because velocity is a vector quantity that describes the speed and direction of an object's motion. In this scenario, the two carts are colliding and sticking together, meaning they will have the same final velocity after the collision. […] This is also supported by the fact that the carts have different masses, so their kinetic energies and linear momenta will not be the same. Thus, the correct answer is [A] velocity.
Low confidence and short	I am not sure that the answer is [A] because the carts may have different velocities before and after the collision due to the conservation of momentum.
Medium confidence and short	I am somewhat sure that the answer is [A] because velocity is a vector quantity that takes into account both the speed and direction of an object's motion. In this scenario, the carts will have the same velocity after the collision because they are stuck together and moving in the same direction.
High confidence and short	I am sure that the answer is [A] because velocity is a vector quantity that is conserved in collisions, meaning that the final velocity of the two carts will be the same.
Low confidence and uncertainty only	I am not sure that the answer is [A].
Medium confidence and uncertainty only	I am somewhat sure that the answer is [A].
High confidence and uncertainty only	I am sure that the answer is [A].

The particular question is: "A moving 1.5 kg cart collides with and sticks to a 0.5 kg cart which was initially at rest. Immediately after the collision, the carts each have the same _____ as each other. Choose from the following options: A) Velocity B) Kinetic energy C) Mass D) Linear momentum." The correct answer to this question is A.

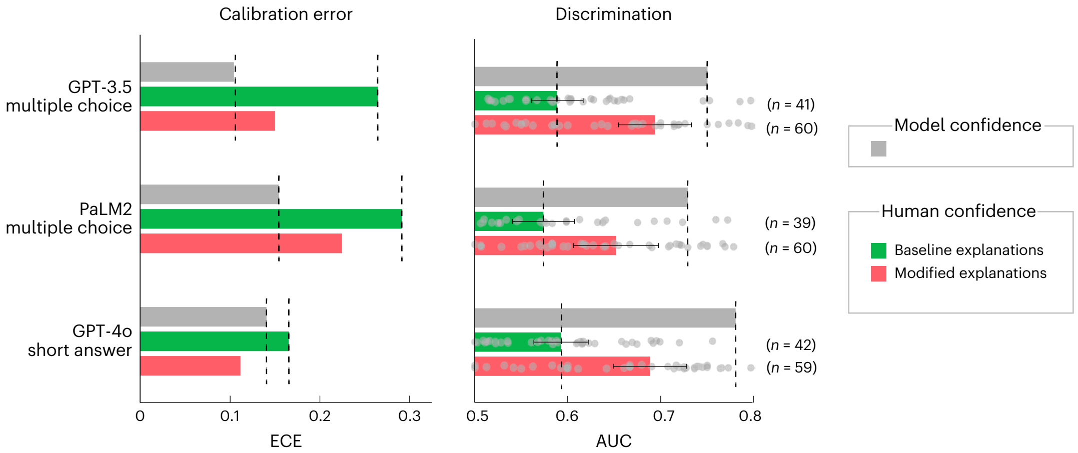


Fig. 2 | Calibration error and discrimination for model confidence and human confidence across the behavioural experiments and LLMs. Calibration error is assessed by ECE (lower is better), while discrimination is assessed by AUC (higher is better). The vertical dashed lines represent the calibration and discrimination gap between model confidence and human confidence for unmodified explanations (experiments 1a, 1b and 1c). For human confidence, the data points represent the AUC values computed separately for each participant (*n* shown in figure) and the error bars represent the 95% confidence interval of the mean across participants. Because of data sparsity, the ECE values were computed at the group level.

Results

We start by examining the results from experiment 1 and compare human and model confidence in the case where LLMs generate default explanations for participants. We present the results for two different metrics: (1) expected calibration error (ECE), which assesses the degree to which confidence scores from the model or the human reflect the true accuracy of the LLM, and (2) area under the curve (AUC), which assesses the degree to which confidence scores discriminate between correct and incorrect responses (see the Methods for details). The findings indicate that there is a significant gap, as measured by calibration and discrimination, between what LLMs know and what humans believe they know based on default explanations.

Calibration gap

Figure 2 (left) shows the ECE for both model and human confidence. The results show a calibration gap; across the different types of LLMs and types of questions (multiple choice and short answer), the ECE metric is much lower for model confidence (in grey) than for human confidence with baseline explanations (in green). This gap demonstrates that standard explanations provided by the LLM do not enable participants to judge the likelihood of correctness of the LLM's answers, leading to a misalignment between perceived accuracy and actual LLM accuracy.

Figure 3 expands on the calibration results in Fig. 2 to show detailed calibration results for each LLM and each experimental condition. The diagrams show how well model confidence (left column) and human confidence (right two columns) are calibrated. The ideal calibration (that is, ECE of 0) would yield results along the diagonal. For multiple-choice questions, both LLMs have a tendency to be overconfident, resulting in calibration lines below the diagonal. For the short-answer questions, the LLM is somewhat underconfident. Comparing the LLM with the human calibration in experiment 1 (middle column), the results show that, for the multiple-choice questions, human miscalibration is primarily due to overconfidence, indicating that people generally believe that LLMs are more accurate than they actually are. The histograms (inset panels) demonstrate that a substantial portion of the calibration error is due to participants' propensity to produce high-confidence scores, even though the model accuracy for the associated questions is much lower than expected based on confidence.

Discrimination gap

Participants are not very good, relative to the LLM, at discriminating between which answers are probably correct or incorrect based on the default explanation. We assess discrimination using the AUC metric applied to the human confidence ratings. Figure 2 (right) shows the AUC for both model and human confidence. The results show a gap between how well model and human confidence discriminate between correct and incorrect answers. The LLM model confidence discriminates between correct and incorrect answers well above chance (GPT-3.5 AUC 0.751 and PaLM2 AUC 0.746 for the multiple-choice questions and GPT-4o AUC 0.781 for the short-answer questions). By contrast, participants who view the default explanations in experiment 1 were only slightly better than random guessing (AUC 0.589 and AUC 0.602 for the multiple-choice explanations by GPT-3.5 and PaLM2, respectively, and AUC 0.592 for the short-answer explanations by GPT-4o). Therefore, default explanations lead to a discrimination gap as well.

Explanation style and length affect human confidence

Experiment 2 evaluates how human confidence is affected by the degree of uncertainty expressed in LLM explanations (across three levels of confidence) as well as the overall length of the LLM explanation (across three levels of length).

Figure 4 shows that the type of uncertainty language used in the explanations has a strong influence on human confidence regardless of the type of LLM that produced the explanation or the type of question. Low-confidence explanations ('I am not sure') produced significantly lower human confidence than medium-confidence explanations ('I am somewhat sure'); Bayes factor (BF) >100 across experiments 2a, 2b and 2c. Similarly, medium-confidence explanations produced lower human confidence than high-confidence explanations; BF >100 across experiments 2a, 2b and 2c. The Supplementary Information ('human confidence agreement') shows an analysis of the reliability of the confidence ratings across participants.

In addition, the length of the explanations also affected the human confidence in the LLM answers. Long explanations led to significantly higher confidence than the short explanations (BF of 25 with data combined across experiments 2a, 2b and 2c), and short explanations led to significantly higher confidence than the responses that contained only the uncertainty expression (BF >100 with data combined across experiments 2a, 2b and 2c). The additional information presented in longer explanations did not enable participants to better discriminate between probably correct and incorrect answers for longer explanations. Across experiments 2a, 2b and 2c, the mean participant AUC is 0.54 and 0.57 for long and uncertainty-only explanations, respectively (BF of 0.23). Therefore, the length of the answer led to an increase in

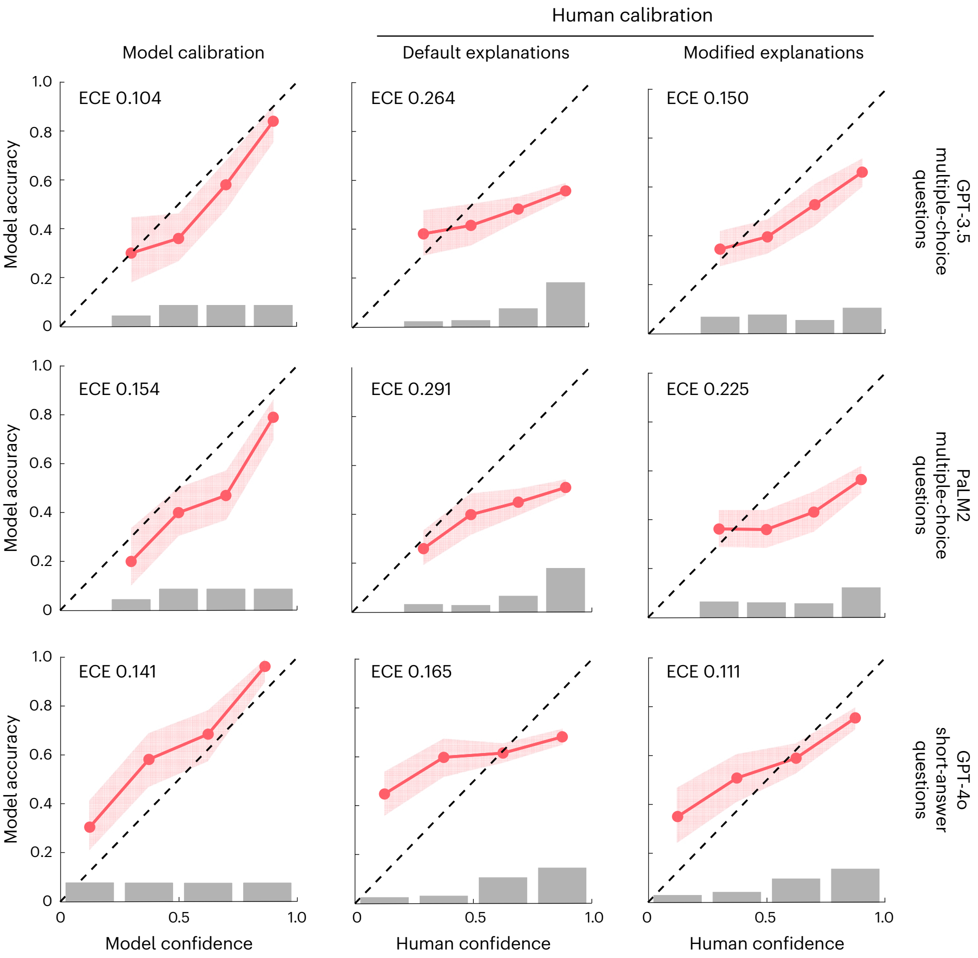


Fig. 3 | Calibration diagrams for model confidence and human confidence across experiments 1 and 2. The top and middle rows show results for multiple-choice questions with the GPT-3.5 and PaLM2 models, respectively. The bottom row shows results for short-answer questions with the GPT-4o model. The histograms at the bottom of each plot show the proportion of observations in each confidence bin (values are scaled by 30% for visual clarity). The shaded regions represent the 95% confidence interval of the mean computed across participants and questions.

human confidence without any corresponding change in sensitivity to discriminating between correct and incorrect answers.

The results confirm that people can appropriately interpret verbal cues about uncertainty and that manipulating the length of the explanation can directly impact human confidence.

Reducing the calibration and discrimination gap

Having established in experiment 2 that the uncertainty language in the LLM explanation can modify human confidence, we now evaluate whether linking the type of uncertainty language to the LLM model confidence (that is, showing a low-, medium- and high-confidence explanation when the model confidence is low, medium and high, respectively) can reduce the calibration and discrimination gap.

Selecting explanations on the basis of model confidence. We simulated the effect of aligning the explanation style to model confidence by a simple decision rule. With this rule, we select the type of explanation $s \in$ {low confidence, medium confidence, high confidence} on the basis of the LLM model confidence score p

$$s = \begin{cases} \text{low confidence} & \text{if } p \le \theta_1 \\ \text{medium confidence} & \text{if } \theta_1 < p \le \theta_2 \\ \text{high confidence} & \text{if } \theta_2 < p \end{cases} . \qquad (1)$$

The parameters θ_1 and θ_2 determine the ranges where low-, medium- and high-confidence explanations are chosen. The application of this rule to a given parameter setting leads to a participant's estimates being filtered out if the explanation style used for a specific question does not match the selected style. This allowed us to simulate the effect of participants receiving different types of explanations based on model confidence (that is, lower-confidence explanations for low model confidence and higher-confidence explanations for high model confidence). The Supplementary Information ('optimization procedure') provides details on the optimization procedure and also a demonstration that the results are not particularly sensitive to the parameter settings.

Calibration and discrimination results. Figure 2 shows the calibration and discrimination results when the selection rule is applied to the results from experiment 2. The results in Fig. 2 (left, red bars) show that the calibration gap has narrowed substantially. While there is still generally a higher calibration error for human confidence relative to model confidence, the calibration gap has decreased for all three LLMs relative to the baseline explanations in experiment 1. Furthermore, Fig. 2 (right) shows that the discrimination gap (as measured by AUC) has also been narrowed relative to the baseline explanations across LLMs and question types (BF >100, BF 6.48 and BF >100 for experiments 2a, 2b and 2c, respectively). Therefore, the results show that selecting the type of explanation on the basis of LLM model confidence improves both

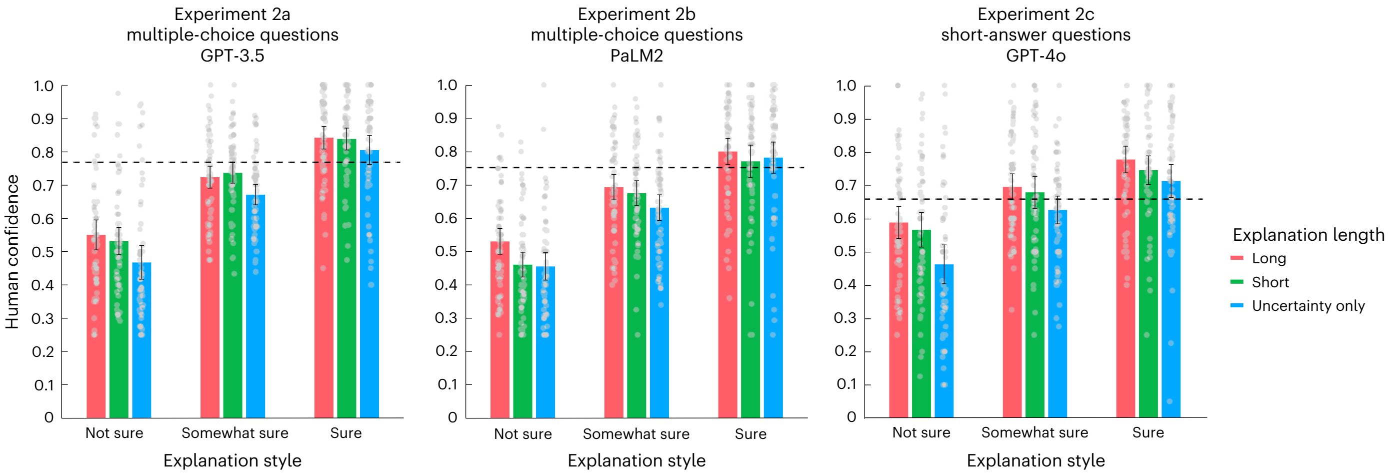


Fig. 4 | Mean human confidence across LLM explanation styles varying in uncertainty language and length. Data are presented as mean values of participant confidence in experiments 2a ($n = 60$), 2b ($n = 60$) and 2c ($n = 59$). For reference, the dashed lines show the average human confidence for the baseline explanations in experiments 1a, 1b, and 1c. The error bars represent the 95% confidence interval of the mean across participants.

calibration and discrimination performance, as human confidence in the LLM becomes more closely aligned with the LLM's actual accuracy.

Participants lack specialized knowledge

For the experiments with multiple-choice questions (1a, 1b, 2a and 2b), participants provided their own answer after seeing the answer from the LLM. This allowed us to analyse whether participants have any independent knowledge from the LLM that allowed them to improve on LLM accuracy. In experiments 1a and 2a with GPT-3.5, participants' average answer accuracy was 51%, closely aligning with LLM's 52% accuracy rate. Similarly, for the multiple-choice experiments 1b and 2b with PaLM2, average participant accuracy was 45%, similar to the 47% accuracy rate for the LLM. In the majority (82%) of responses across all multiple-choice experiments, participants selected the response that agreed with the LLM's explanation.

When participants chose to alter the answer, the average accuracy was 33% which is lower than the LLM's accuracy of 39% for these particular questions. These findings suggest limited success in participants' ability to accurately answer the questions independent of the LLM's explanation. This is consistent with findings from ref. 19, showing that Mechanical Turk workers without specialized knowledge (akin to our participant pool) scored 35% accuracy on similar questions.

When we applied the selection rule and the explanation type was aligned with model confidence, human decision accuracy in experiments 2a and 2b did not improve for the selected questions (even though discrimination and calibration improved). This shows that accurate uncertainty communication by the LLM allowed participants to recognize when the LLM was providing a probably correct or incorrect answer, but the lack of accurate human knowledge independent from the LLM prevented participants from improving on the LLM answer.

At the end of the experiment, participants estimated the performance they would achieve on similar questions for each of the ten topics in the sample of MMLU questions. The median of these self-assessed expertise estimates did not substantially vary between topics: from 30% (for example, high school physics) to 45% (for example, high school world history). Examining the impact of perceived expertise on accuracy estimation, we divided participants into two groups based on whether their self-rated expertise was above or below 50% separately for each of the ten topics. For the experiments with GPT-3.5, the higher-expertise groups generally had better discrimination (AUC 0.600 versus AUC 0.579), but there was no evidence that this difference was significant (BF <1). In addition, the calibration error was comparable between the two groups (ECE 0.289 versus 0.292). Similarly, no effect of expertise was found for the experiments with PaLM2. Therefore, participants who considered themselves more knowledgeable about a topic were not more adept at estimating the LLM's performance in that area.

Discussion

Our research focused on bridging the gap between what an LLM knows and what users perceive it knows. This gap is critical, especially as reliance on LLMs for decision-making across various domains is rapidly increasing.

Research on LLMs has begun to address these challenges, with a focus on improving uncertainty communication and the quality of explanations. Several studies have explored LLM confidence in answering multiple-choice questions, focusing on how well the models' self-reported confidence aligns with their actual accuracy[12,13,19,24] and whether users can accurately assess the reliability of the explanations provided[25]. The work by ref. 26 investigates how users respond to verbal phrases of uncertainty in a simulated trivia task but does not use actual LLM outputs. Overall, there has been little research examining user confidence in LLM output. Our work uses actual LLM outputs and its confidence in an attempt to quantify the calibration and discrimination gap. As a result, we directly address the issue of miscommunication of uncertainty from LLMs to humans.

Our results showed that users consistently overestimated how accurate LLM outputs were, especially when they relied on the models' default explanations. This was true for three different LLMs and two different types of questions (multiple choice and short answer). This tendency towards overconfidence in LLM capabilities is an important concern, particularly in scenarios where critical decisions rely on LLM-generated information. The inability of users to discern the reliability of LLM responses not only undermines the utility of these models but also poses risks in situations where user understanding of model accuracy is critical.

In addition, the results also showed a length bias where longer explanations led to higher human confidence levels even though they did not contain any additional information to help users to better discriminate between probably correct and incorrect answers. This suggests that users were processing the explanations at a shallow level, relying on simple textual cues such as overall length to predict LLM accuracy. This result is consistent with studies in social psychology and communication research that suggest that longer answers or explanations may be perceived as more persuasive or credible, even when they do not contain more meaningful information[27,28]. This length bias has also been found in domains such as peer reviews, where longer reviews are perceived as more persuasive and informative even if the information content remains the same[29].

Although default LLM explanations do not enable users to perceive what the models truly know, this research shows that a simple approach based on tailored explanations can bridge this perception gap. This was achieved by altering the prompts used to generate explanations based on model confidence, allowing better control over how uncertainty was expressed in the responses. Specifically, we designed these prompts to induce varying degrees of certainty in the explanations, ranging from expressions of low confidence (for example, 'I am not sure the answer is [B] because') to high confidence (for example, 'I am confident the answer is [B] because'). By modifying the language of the LLM's responses to better reflect model confidence, users showed improved calibration in their assessment of the LLM's reliability and were better able to discriminate between correct and incorrect answers. This improvement underscores the importance of transparent communication from LLMs, suggesting a need for researchers to investigate how model explanations affect user perception.

One limitation of the current study is the focus on a specific type of question involving a small number of response alternatives (multiple choice) and short answers to open-ended questions. The extent to which these results apply to longer open-ended questions remains an open question. Further research could investigate the applicability of our findings across a broader range of scenarios. Another limitation of this study is that our approach to modifying the prompt on the basis of internal uncertainty required the LLM to be prompted twice: once to read out the answer and model confidence, and again to produce an explanation modified by the model confidence. Future research could investigate how to produce confidence-modified explanations in a single step.

Another important area for future research is to understand the fundamental causes of the miscommunication of uncertainty. Why do LLMs generate calibrated model confidences while also producing explanations that are not consistent with those confidences? One hypothesis is that current LLMs are aligned to human preferences using reinforcement learning from human feedback (RLHF)[30], which produces some built-in biases. In these RLHF procedures, various types of explanations are presented to human participants, who can then choose their preferred explanations. LLMs are then fine-tuned on the basis of human preference data, making them more likely to produce explanations that people prefer. While RLHF encourages human-aligned output, it inevitably reproduces any human preference biases. For example, people prefer detailed and generally longer explanations[31,32]. As a result, LLMs trained on these human preferences may produce explanations that are overly convincing, potentially misleading users about the reliability of the information.

An alternative hypothesis to the production of overconfident explanations lies in the autoregressive nature of well-established LLMs. In particular, we conjecture that, after committing to an answer (encoded as a sequence of tokens), the model will generate a sequence of tokens (explanation) that maximizes the likelihood of the previous answer, effectively resulting in an assertive answer. A similar hypothesis was also presented in ref. 16. Interestingly, the possibility that the LLM's choice of a particular answer inflates the rationale for that answer resembles the phenomenon of choice-supportive biases in psychology [33]. After making a decision, people tend to overestimate the desirability of the chosen option while underestimating the desirability of rejected alternatives. This can make them feel more confident in their decision than they were when they first made it.

Our work shares some parallels with prior studies on the human perception and evaluation of artificial intelligence (AI)-generated explanations in the domain of machine learning classifiers (see ref. 34 for an overview). These studies frequently use feature highlighting to explain what areas of the image[35] or what fragments of documents[36] can support the suggested classification. Studies have found mixed evidence for the effectiveness of these types of AI explanations in human decision-making[37–40]. These results highlight the challenge of ensuring that AI-generated explanations align with human expectations and allow humans to distinguish between correct and incorrect answers.

In conclusion, our research highlights the critical role of clear and accurate communication in the interaction between users and LLMs. Enhancing the alignment between model confidence and the user's perception of model confidence can lead to more responsible and trustworthy use of LLMs, particularly in areas where the accuracy of AI-generated information is critical.

Methods

Question datasets

MMLU dataset for multiple-choice questions. The MMLU dataset is a comprehensive multitask dataset that contains multiple-choice questions from various knowledge domains, such as STEM, humanities, social sciences and more[19]. In total, there are 14,042 test set questions from 57 categories curated by undergraduate and graduate students from freely available online resources such as the Graduate Record Examinations and United States Medical Licensing Examination. These questions range in difficulty from high school to the professional level. The MMLU dataset is widely used to measure a text model's multitask accuracy, as it challenges models on their real-world text understanding beyond mere linguistic comprehension, thus making it a robust benchmark for model evaluation[19,41,42]. For this research, we sampled a subset of 350 questions from a range of model confidence levels in ten select categories from the full dataset to comprehensively evaluate people's assessment of LLM model confidence.

Trivia QA dataset for short-answer questions. Trivia QA is a dataset of trivia questions that can be answered in short answers[22]. Similar to methodology by ref. 17, contextual information was excluded to make the question answering more challenging for LLMs and more suitable for our behavioural experiments. For this research, we assessed model confidence for 5,000 questions from the original 650,000-question dataset before selecting a final sample of 336 questions from a range of model confidence levels. The final set of questions was categorized into seven different topics (culture and society, entertainment, geography, history, politics, science and technology, and sports).

Assessing model confidence and creating question subsets

Several approaches have been developed to elicit confidence in LLMs and to assess the degree to which the elicited confidence scores are calibrated (see ref. 43 for an overview). In this research, we use an approach commonly used to access internal model information based on token likelihoods, allowing the direct computation of relative probabilities of different possible answers in multiple-choice questions[12,13,18,19,23]. In addition, the token-likelihood approach can be extended to short-answer questions such that the token-likelihood reflects the model confidence that the LLM answer is correct[13].

Methods that do not require access to internal model representations have used prompting strategies designed to elicit verbal expressions of uncertainty[24,44]. Confidence is expressed in natural language as numeric strings (for example, '80%')[24,45] or more qualitative expressions of confidence (for example, 'I am not confident the answer is X'). Prompts can be designed to emphasize step-by-step reasoning about the correctness of individual steps and clarify the space of possible answers to lead to better calibration than simple prompts that simply ask for a confidence rating[24]. For short-form question answering, prompting strategies can lead to calibrated confidence levels[46]. However, prompting approaches have been found to be less accurate compared with methods that read out model confidence[24].

Multiple-choice questions. For the multiple-choice questions, we followed the procedures based on reading out the internal token likelihoods as described in the GPT-4 Technical Report[12]. We used a zero-shot prompting approach, in which the model was prompted only with the target question and its associated answer options (Extended Data Fig. 1). We first assessed the LLM model confidence of GPT-3.5 and PaLM2 language models to 14,042 MMLU multiple-choice questions. This allowed us to then select questions with (somewhat) evenly distributed confidence levels. We read out the log-probabilities for the top five tokens completed by the model using the APIs for the GPT3.5 (gpt-3.5-turbo-instruct) and the PaLM2 (text-bison@002) models. The temperature parameter was set to 0. The answer was considered complete if the tokens included the single letters A, B, C and D. The log scores were then converted and normalized to probabilities across the four answer options (so that the sum of the scores equalled one). In this research, internal uncertainties, referred to in this Article as model confidence, were represented by these probabilities in all experiments, a common technique in calibration assessment with LLMs[12,13,18,19,23].

Based on the model confidence levels of each LLM for all MMLU questions, we created a subset separately for each LLM. In total, 35 questions were sampled for each of 10 topics, for a total of 350 questions. For each topic, the 35 questions were sampled to approximately create a uniform distribution over model confidence using the confidence bins 0.2–0.4, 0.4–0.6, 0.6–0.8 and 0.8–1.0. However, owing to the small number of questions that lead to model confidence in the lowest confidence bin, fewer questions were sampled for the 0.2–0.4 confidence range. Supplementary Fig. 1 shows the distribution over model confidence levels for the entire MMLU dataset as well as the question subset sampled for our study. Model accuracy across the 350 questions is 55% and 50% for GPT-3.5 and PaLM2, respectively.

Short-answer questions. For the short-answer questions, we used a procedure based on the pTrue method[13] to assess internal model confidence. All experiments with short-answer questions were performed with the API for the GPT-4o model (gpt-4o-mini) with the temperature parameter set to 0.7 (similar to refs. 13,17). The model was first prompted to generate the answer to each of the 5,000 trivia questions in the sample. To ensure that the model response was restricted to short answers, we used a ten-shot prompting approach where the prompt contained the target question preceded by a random sample of ten trivia question with the reference answers. The median answer length was two words.

To assess the model confidence for short-answer questions, as shown in Fig. 1 (bottom), we prompted the model with the question and the proposed answer and asked it to determine whether the proposed answer is true or false (see Extended Data Fig. 1 for an example of the exact prompt). The log scores for the true and false answer options were then converted and normalized to probabilities across the two answer options. The model confidence in our experiments corresponds to the probability for the true answer option.

For the behavioural experiments, we created a subset of 336 questions to ensure a uniform distribution across four confidence bins: 0–0.25, 0.25–0.50, 0.50–0.75 and 0.75–1.0. Supplementary Figs. 1 and 2 show the distribution of model confidence levels for the 5,000 sample and the 336 subset used in our behavioural experiments. The model accuracy across the 336 questions is 63%.

We used both automatic and human scoring methods to assess model accuracy. For the 5,000-question sample, we prompted an LLM (GPT-4o) to determine whether the reference answer from the Trivia QA had the same meaning as the LLM answer within the context of the question. For the 336-question sample, we also applied human scoring. For 97% of questions, automatic and human scoring agreed. The model accuracy for the 336-question subset was based on human evaluation.

Behavioural experiments

This section describes the methodology we used for our behavioural experiments. Experiment 1 presented default explanations from LLMs to participants, whereas experiment 2 presented explanations that were altered by different types of uncertainty language and overall length (see Table 1 for an overview of all experiments). Within each experiment, across different groups of participants, we varied the type of question as well as the type of LLM. Experiments 1a and 2a used explanations from GPT-3.5 for the MMLU multiple-choice questions. Experiments 1b and 2b used explanations from PaLM2 for the MMLU multiple questions. Finally, experiments 1c and 2c used explanations from GPT-4o for the Trivia QA short-answer questions. The Supplementary Information ('experiment 3') describes the results from an additional experiment 3, which was conducted to verify that our results generalize to different ways to vary the type of uncertainty language in the explanations.

Participants. A total of 301 participants completed the study across experiments 1 and 2 (Table 1 presents the breakdown by experiment). The participants were native English speakers residing in the USA, recruited through Prolific (www.prolific.com). Demographic data were obtained for 284 participants. There were 146 female and 138 male participants. The median age was 34 years (age range from 18 to 79). Supplementary Table 1 presents the breakdown of demographic information by experiment. The University of California, Irvine institutional review board approved the experimental protocol. The participants who completed experiments 1a, 1b, 2a or 2b were paid US$8 for their participation. The participants in experiments 1c and 2c required less time to complete the study and were paid US$5. The payments across experiments corresponded to a rate of approximately US$12 per hour. Before participating, all individuals provided informed consent and received detailed instructions outlining the experimental procedure as well as how to understand and interact with the user interface. Participants were also asked to sign an integrity pledge after reading all of the instructions, stating that they would complete the experiment to the best of their abilities. After submitting their integrity pledge, the participants were granted access to the experiment.

Experimental procedure. Across all experiments, participants were randomly assigned 40 questions (from the pool of 350 multiple-choice questions or the pool of 336 short-answer questions). The questions were sampled to balance across model confidence bins ensuring that all participants were exposed to questions at all levels of difficulty.

Furthermore, in experiments 2a, 2b and 2c, we balanced the types of explanation style across questions so that each question was presented approximately the same number of times with each style. It should be noted that, for each subject, each question was presented only once, and each question received only one explanation style. The counterbalancing ensured that the same question had (roughly) an equal number of observations for each explanation style (across participants).

For the multiple-choice questions, the participant's task was divided into two phases for each question (Extended Data Fig. 2). In the first stage, the participants had to provide a probability estimate that the LLM's answer was correct. In the second phase, the participants had to answer the question with the assistance of the LLM. The participants were instructed to use their own knowledge as well as the LLM's response when making their own answer selection for this phase. For the short-answer questions, the participants only had to provide a probability estimate that the LLM's answer was correct. They were instructed not to look up the answer. For the short-answer questions, the question-answering phase was omitted to prevent participants from looking up the answers.

At the end of the experiments, we administered a brief survey in which the participants self-assessed their knowledge of topics associated with the multiple-choice and short-answer questions. The participants were asked to estimate the expected accuracy for each topic if they were presented with questions similar to those they encountered during the experiment.

Creating explanation styles with varying degrees of uncertainty. This section describes how we constructed prompts to elicit explanations with varying levels of uncertainty language in the experiments. Table 2 contains examples of explanations from the two experiments and styles. For additional details on the exact prompts used, see Supplementary Table 3.

Experiment 1. To generate the baseline explanations for experiment 1, we used a simple prompt that asked for an explanation, which would also repeat the answer.

Experiment 2. In experiment 2, explanations were manipulated in terms of the level of confidence expressed in the answer as well as the length of the answer. In total, the experiment included nine types of explanations (three levels of uncertainty × three levels of length). The three levels of confidence (low, medium and high) were generated by prompts that instructed the LLM to 'mention you are not sure/somewhat sure/sure' in the explanations, respectively. The prompts elicited responses in which the beginning of each explanation indicated the level of uncertainty (for example, 'I am not sure the answer is [B] because' for the low-confidence prompt). Note that expressions of uncertainty were not limited to the start of the explanation. Answers often contained additional explanations for why the LLM lacked confidence (for example, 'further research may be required to confirm this' and 'it is not possible to definitively state that…'). Experiment 2 also varied the length of the explanation across three levels: long, short and uncertainty only. The long explanations were generated by not including any instruction regarding the length of the answer. The short explanations were generated by adding an instruction to use as few words as possible in the explanation. The uncertainty-only explanations were generated by removing the rationale for the answer and including only the expression of uncertainty and the answer (for example, 'I am not sure the answer is [B]').

For experiment 2, the median lengths of the long and short explanations were as follows: 115 and 34 words (GPT-3.5, multiple choice), 64 and 24 words (PaLM2, multiple choice) and 95 and 24 words (GPT-4o, short answer). By comparison, the uncertainty-only responses contained a median of nine words across all variants of experiment.

Metrics

To investigate the relationship between the accuracy of answers to the multiple-choice and short-answer questions and the confidence (either human confidence or model confidence) associated with them, we utilize a range of metrics to evaluate this association. The primary focus is on understanding how well confidence levels correlate with the correctness of answers. To achieve this, we use both ECE and the AUC metric. These metrics assess the extent of overconfidence in predictions as well as the diagnostic effectiveness of confidence scores in distinguishing between correct and incorrect answers[13,18,23,24,46]. The use of AUC in this context parallels various metrics in psychology for metacognitive discrimination or sensitivity, which similarly aim to evaluate the effectiveness of confidence scores in distinguishing between correct and incorrect answers[47]. In addition, in the Supplementary Information ('Overconfidence Error'), we also show results for the additional metric of Overconfidence Error (OE).

ECE. We evaluate miscalibration using the ECE, as detailed in refs. 48,49. The ECE is calculated by averaging the absolute differences between accuracy and confidence across M equal-width probability bins

$$\mathrm{ECE}=\sum_{m=1}^{M}\frac{|B_m|}{N}|\mathrm{conf}(B_m)-\mathrm{acc}(B_m)|, \tag{2}$$

where N represents the total sample count, B_m represents the mth confidence bin, and acc(B_m) and conf(B_m) respectively denote the accuracy and average confidence for samples in the mth bin. ECE does not account for the direction of deviations between the accuracy and the confidence per bin, respectively, so a non-zero ECE can indicate a mix of over- and underconfidence. Although recent work[50,51] has shown that ECE can underestimate the true calibration error, the potential for underestimation should not be an issue given that we are interested in analysing differences in ECE rather than unbiased estimates of the error itself.

AUC. The AUC metric is used to assess the diagnostic ability of confidence scores in distinguishing between correct and incorrect answers. Utilizing the Mann–Whitney *U* statistic approach, the AUC represents the probability that a randomly chosen correct answer has a higher-confidence score compared with a randomly chosen incorrect answer

$$\mathrm{AUC}=\frac{1}{N_{\mathrm{pos}}\times N_{\mathrm{neg}}}\sum_{i=1}^{N_{\mathrm{pos}}}\sum_{j=1}^{N_{\mathrm{neg}}}I(C_i>C_j), \tag{3}$$

where N_{pos} and N_{neg} denote the counts of correct (positive) and incorrect (negative) answers, respectively. C_i and C_j represent the confidence scores of the ith and jth correct and incorrect answers, respectively. I is an indicator function, which equals 1 if $C_i > C_j$ and 0 otherwise. This method evaluates each pair of correct and incorrect answers to determine if the confidence score for the correct answer surpasses that of the incorrect one. The AUC is then the fraction of these pairs satisfying this criterion, measuring the capability of confidence scores to differentiate between correct and incorrect responses, with AUC values ranging from 0.5 (indicating no better than chance discrimination) to 1 (signifying perfect discrimination).

Statistical analysis

To assess statistical significance, we utilize BFs to determine the extent to which the observed data adjust our belief in the alternative and null hypotheses. Values of $3 < \mathrm{BF} < 10$ and $\mathrm{BF} > 10$ indicate moderate and strong evidence against the null hypothesis, respectively. Similarly, values of $1/10 < \mathrm{BF} < 1/3$ and $\mathrm{BF} < 1/10$ indicate moderate and strong evidence in favour of the null hypothesis, respectively. We report BFs for Bayesian *t*-tests using the default priors as recommended by ref. 52.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All behavioural data as well as data produced by the LLMs used in this study are publicly available via the Open Science Framework (OSF) at https://osf.io/y7pr6/ (ref. 53). Source data are provided with this paper.

Code availability

The code used for data analysis and extracting LLM model confidence is available via OSF at https://osf.io/y7pr6/ (ref. 53).

References

1. Budescu, D. V., Por, H.-H., Broomell, S. B. & Smithson, M. The interpretation of IPCC probabilistic statements around the world. *Nat. Clim. Change* **4**, 508–512 (2014).
2. Ho, E. H., Budescu, D. V., Dhami, M. K. & Mandel, D. R. Improving the communication of uncertainty in climate science and intelligence analysis. *Behav. Sci. Policy* **1**, 43–55 (2015).
3. Karelitz, T. M., Dhami, M. K., Budescu, D. V. & Wallsten, T. S. Toward a universal translator of verbal probabilities. In *Proc. 15th International Florida Artificial Intelligence Research Society Conference* (eds Haller, M. S. & Simmons, G.) 498–502 (AAAI Press, 2002).
4. Wallsten, T. S., Shlomi, Y. & Ting, H. *Final Report for Research Contract 'Expressing Probability in Intelligence Analysis'* (2008).

5. O'Brien, B. J. Words or numbers? The evaluation of probability expressions in general practice. *J. R. Coll. Gen. Pract.* **39**, 98–100 (1989).
6. Ali, S. R., Dobbs, T. D., Hutchings, H. A. & Whitaker, I. S. Using ChatGPT to write patient clinic letters. *Lancet Digit. Health* **5**, 179–181 (2023).
7. Zambrano, A. F. et al. From nCoder to ChatGPT: from automated coding to refining human coding. In *Proc. International Conference on Quantitative Ethnography* (eds Arastoopour Irgens, G. & Knight, S.) 470–485 (Springer, 2023).
8. Whalen, J. et al. ChatGPT: challenges, opportunities, and implications for teacher education. *Contemp. Iss. Technol. Teach. Educ.* **23**, 1–23 (2023).
9. Jo, A. The promise and peril of generative AI. *Nature* **614**, 214–216 (2023).
10. Huang, L. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. Preprint at https://arxiv.org/abs/2311.05232 (2024).
11. *Introducing ChatGPT* (OpenAI, 2022).
12. Achiam, J. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
13. Kadavath, S. et al. Language models (mostly) know what they know. Preprint at https://arxiv.org/abs/2207.05221 (2022).
14. Srivastava, A. et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.* (2023).
15. Yin, Z. et al. Do large language models know what they don't know? In *Proc. Findings of the Association for Computational Linguistics* (eds Rogers, A. et al.) 8653–8665 (ACL, 2023).
16. Azaria, A. & Mitchell, T. in *Findings of the Association for Computational Linguistics: EMNLP 2023* (eds Bouamor, H. et al.) 967–976 (ACL, 2023).
17. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024).
18. Jiang, Z., Araki, J., Ding, H. & Neubig, G. How can we know when language models know? On the calibration of language models for question answering. *Trans. Assoc. Comput. Linguist.* **9**, 962–977 (2021).
19. Hendrycks, D. et al. Measuring massive multitask language understanding. In *Proc. International Conference on Learning Representations* (2021).
20. *GPT-3.5* (OpenAI, 2022).
21. Anil, R. et al. *Palm 2 Technical Report* (2023).
22. Joshi, M., Choi, E., Weld, D. S. & Zettlemoyer, L. Triviaqa: a large scale distantly supervised challenge dataset for reading comprehension. In *Proc. 55th Annual Meeting of the Association for Computational Linguistics* Vol. 1 (eds Barzilay, R. & Kan, M.-Y.) 1601–1611 (ACL, 2017).
23. Xiao, Y. et al. in *Findings of the Association for Computational Linguistics: EMNLP 2022* (eds Goldberg, Y. et al.) 7273–7284 (ACL, 2022).
24. Xiong, M. et al. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations* (2024).
25. Tanneru, S. H., Agarwal, C. & Lakkaraju, H. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics* 1072–1080 (PMLR, 2024).
26. Zhou, K., Hwang, J., Ren, X. & Sap, M. in *Relying on the Unreliable: The Impact of Language Models' Reluctance to Express Uncertainty* 3623–3643 (Association for Computational Linguistics, 2024).
27. Petty, R. E. & Cacioppo, J. T. The effects of involvement on responses to argument quantity and quality: central and peripheral routes to persuasion. *J. Person. Soc. Psychol.* **46**, 69 (1984).
28. Oppenheimer, D. M. Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Appl. Cogn. Psychol.* **20**, 139–156 (2006).
29. Goldberg, A. et al. Peer reviews of peer reviews: a randomized controlled trial and other experiments. Preprint at https://arxiv.org/abs/2311.09497 (2023).
30. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
31. Bower, A. H., Han, N., Soni, A., Eckstein, M. P. & Steyvers, M. How experts and novices judge other people's knowledgeability from language use. *Psychonom. Bull. Rev.* 1–11 (2024).
32. Saito, K., Wachi, A., Wataoka, K. & Akimoto, Y. Verbosity bias in preference labeling by large language models. Preprint at https://arxiv.org/abs/2310.10076 (2023).
33. Mather, M., Shafir, E. & Johnson, M. K. Misremembrance of options past: source monitoring and choice. *Psychol. Sci.* **11**, 132–138 (2000).
34. Rong, Y. et al. Towards human-centered explainable AI: a survey of user studies for model explanations. In *Proc. IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 46 1–20 (IEEE, 2023).
35. Smith-Renner, A. et al. No explainability without accountability: an empirical study of explanations and feedback in interactive ML. In *Proc. 2020 CHI Conference on Human Factors in Computing Systems* 1–13 (Association for Computing Machinery, 2020).
36. Feng, S. & Boyd-Graber, J. What can AI do for me? Evaluating machine learning interpretations in cooperative play. In *Proc. 24th International Conference on Intelligent User Interfaces* (eds Fu, W.-T. & Pan, S.) 229–239 (ACL, 2019).
37. Steyvers, M. & Kumar, A. Three challenges for AI-assisted decision-making. Perspect. *Psychol. Sci.* **19**, 722–734 (2023).
38. Bansal, G. et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proc. 2021 CHI Conference on Human Factors in Computing Systems* (eds Kitamura, Y. & Quigley, A.) 1–16 (ACL, 2021).
39. Buçinca, Z., Malaya, M. B. & Gajos, K. Z. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc. ACM Hum. Comput. Interact.* **5**, 1–21 (2021).
40. Wang, X. & Yin, M. Effects of explanations in AI-assisted decision making: principles and comparisons. *ACM Trans. Interact. Intell. Syst.* **12**, 1–36 (2022).
41. Hoffmann, J. et al. Training compute-optimal large language models. Preprint at https://arxiv.org/abs/2203.15556 (2022).
42. Rae, J. W. et al. Scaling language models: methods, analysis & insights from training gopher. Preprint at https://arxiv.org/abs/2112.11446 (2021).
43. Geng, J. et al. A survey of language model confidence estimation and calibration. Preprint at https://arxiv.org/abs/2311.08298 (2023).
44. Zhou, K., Jurafsky, D. & Hashimoto, T. Navigating the grey area: how expressions of uncertainty and overconfidence affect language models. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5506–5524 (Association for Computational Linguistics, 2023).
45. Lin, S., Hilton, J. & Evans, O. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.* (2022).
46. Tian, K. et al. Just ask for calibration: strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proc. 2023 Conference on Empirical Methods in Natural Language Processing* 5433–5442 (Association for Computational Linguistics, 2023).

47. Fleming, S. M. & Lau, H. C. How to measure metacognition. *Front. Hum. Neurosci.* **8**, 443 (2014).
48. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) vol. 70 of *Proceedings of Machine Learning Research* 1321–1330 (PMLR, 2017).
49. Naeini, M. P., Cooper, G. & Hauskrecht, M. Obtaining well calibrated probabilities using Bayesian binning. In *Proc. AAAI Conference on Artificial Intelligence* Vol. 29 2901–2907 (AAAI, 2015).
50. Kumar, A., Liang, P. S. & Ma, T. Verified uncertainty calibration. *Adv. Neural Inf. Process. Syst.* **32**, (2019).
51. Gruber, S. & Buettner, F. Better uncertainty calibration via proper scores for classification and beyond. *Adv. Neural Inf. Process. Syst.* **35**, 8618–8632 (2022).
52. Rouder, J. N., Morey, R. D., Speckman, P. L. & Province, J. M. Default Bayes factors for ANOVA designs. *J. Math. Psychol.* **56**, 356–374 (2012).
53. Steyvers, M., Tejeda, H. & Belem, C. What large language models know and what people think they know. *OSF* https://doi.org/10.17605/OSF.IO/Y7PR6 (2024).

Acknowledgements

This research was supported by the National Science Foundation under award number 1900644 (P.S. and M.S.)

Author contributions

M.S. and P.S. conceptualized and designed the study. M.S., A.K., C.B., S.K., L.W.M. and X.H. developed methodology for the behavioural experiments. H.T. implemented and conducted the behavioural experiments. M.S. analysed the data. M.S. and A.K. wrote the first draft. H.T., C.B., S.K., L.W.M. and P.S. reviewed and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00976-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00976-7.

Correspondence and requests for materials should be addressed to Mark Steyvers.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

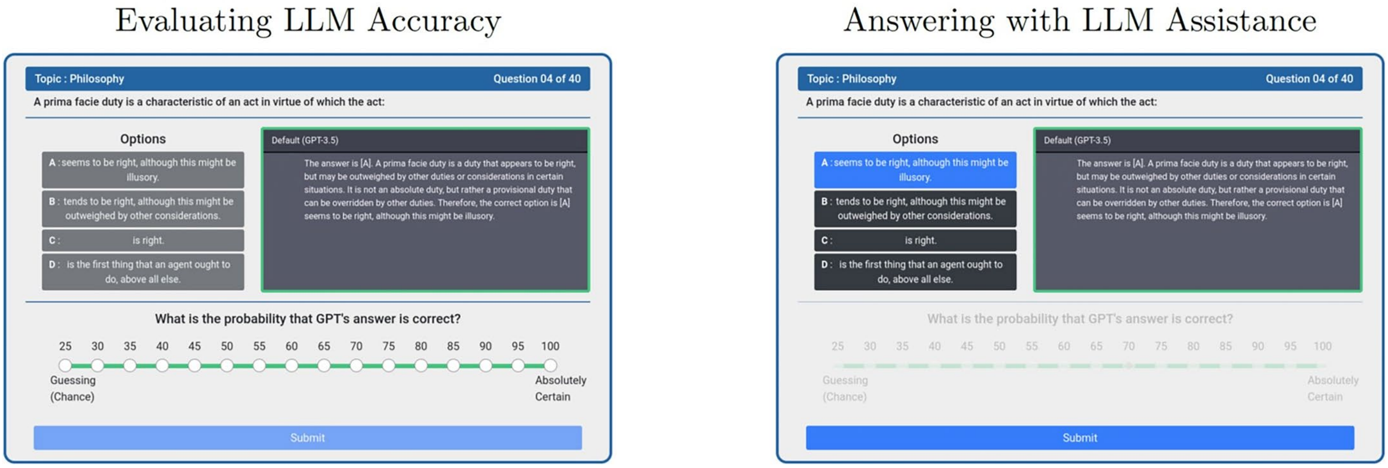
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

```
A horizontal spring-block oscillator is in motion when a second block of
equal mass falls on top of the oscillating block.  If the two blocks
stick together, what effect will this have on the amplitude of the
oscillations?
Choose from the following options:
[A]  It will increase the amplitude.
[B]  It will decrease the amplitude.
[C]  It will have no effect.
[D]  It would depend on where the original block was in its cycle when
the second block fell.
Please answer this question.  Answer [A], [B], [C], or [D].
The answer therefore is:  [
```

```
Question:  which bird is the international symbol for happiness?
Proposed answer:  dove
Is the proposed answer:
[A]  True
[B]  False
The proposed answer is:  [
```

Extended Data Fig. 1 | Example prompts for multiple choice and short-answer questions. Example prompt to elicit model confidence for a multiple-choice question (top) and short-answer question (bottom). For the multiple-choice question, the prompt elicits the answer and model confidence across answer options. Note that for the short-answer question, the LLM answer (dove) was elicited through another prompt.



Evaluating LLM Accuracy

Topic : Philosophy
Question 04 of 40

A prima facie duty is a characteristic of an act in virtue of which the act:

Options

A : seems to be right, although this might be illusory.
B : tends to be right, although this might be outweighed by other considerations.
C : is right.
D : is the first thing that an agent ought to do, above all else.

Default (GPT-3.5)

The answer is [A]. A prima facie duty is a duty that appears to be right, but may be outweighed by other duties or considerations in certain situations. It is not an absolute duty, but rather a provisional duty that can be overridden by other duties. Therefore, the correct option is [A] seems to be right, although this might be illusory.

What is the probability that GPT's answer is correct?

Answering with LLM Assistance

Topic : Philosophy
Question 04 of 40

A prima facie duty is a characteristic of an act in virtue of which the act:

Options

A : seems to be right, although this might be illusory.
B : tends to be right, although this might be outweighed by other considerations.
C : is right.
D : is the first thing that an agent ought to do, above all else.

Default (GPT-3.5)

The answer is [A]. A prima facie duty is a duty that appears to be right, but may be outweighed by other duties or considerations in certain situations. It is not an absolute duty, but rather a provisional duty that can be overridden by other duties. Therefore, the correct option is [A] seems to be right, although this might be illusory.

What is the probability that GPT's answer is correct?

Extended Data Fig. 2 | Illustration of the behavioral experiment interface. Illustration of the behavioral experiment interface used to conduct the multiple-choice experiments. The left panel displays the experimental interface during phase 1 of the task, where participants evaluate the probability that the LLM is correct based on the explanation. The right panel displays the experimental interface during phase 2 of the task, where participants answer the multiple-choice question with the assistance of the LLM.

nature portfolio

Corresponding author(s): Mark Steyvers

Last updated by author(s): 11/20/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used APIs for OpenAI and PaLM2 to query the Large Language Models
Data analysis	We used Matlab to analyze the behavioral data. The analysis code used for this research is available at: https://osf.io/y7pr6/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All behavioral data as well as data produced by the Large Language Models used in this study are publicly available from the following OSF repository: https://osf.io/y7pr6/

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Demographic data was available for 284 out of the 300 participants in the experiments. The recruitment platform we used (Prolific) relies on self-reports for Sex. There were 146 Female and 138 Male participants.
Population characteristics	Participants were native English speakers residing in the United States. The median age was 34 (age range from 18 to 79)
Recruitment	Participants who completed Experiments 1a, 1b, 2a, or 2b were paid $8 USD for their participation. Participants in Experiments 1c and 2c required less time to complete the study and were paid $5. The payments across experiments corresponded to a rate of approximately $12/hr
Ethics oversight	The University of California, Irvine Institutional Review Board (IRB) approved the experimental protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We designed a series of behavioral experiments to evaluate human perceptions of LLM confidence. In these experiments, participants estimate the probability that the LLM's answer to a multiple-choice or short-answer question is correct based on the explanation that the LLM provided. The data collected is quantitative in nature and consists of confidence judgments as well as answers to multiple choice questions.
Research sample	A total of 301 participants completed the study across Experiments 1a,1b,1c, 2a,2b, and 2c. Participants were native English speakers residing in the United States, recruited through Prolific (www.prolific.com). Demographic data was available for 284 out of the 300 participants in the experiments. The recruitment platform we used (Prolific) relies on self-reports for Sex. There were 146 Female and 138 Male participants.
Sampling strategy	The sampling procedure for participants was based on a random sample from the recruitment platform Prolific. To assess statistical significance, we utilize Bayes factors to determine the extent to which the observed data adjust our belief in the alternative and null hypotheses. Relative to null-hypothesis testing, Bayes factors are less sensitive to sample size strategies.
Data collection	Our behavioral studies were conducted on a web browser. The data collected from the user interface was stored on a Google real-time database (a cloud-based server).
Timing	Data collection took place between Dec 7, 2023 and Sept 18, 2024
Data exclusions	No data from participants who completed the experiment was excluded
Non-participation	12 participants did not complete the experiment. No reasons were provided for their non-participation.
Randomization	Across all experiments, participants were randomly assigned 40 questions (from the pool of 350 multiple-choice questions or the pool of 336 short-answer questions). The questions were sampled to balance across model confidence bins ensuring that all participants were exposed to questions at all levels of difficulty. Furthermore, in Experiments 2a, 2b, and 2c, we balanced the types of explanation styles across questions so that each question was presented approximately the same number of times with each style. It should be noted that for each subject, each question was presented only once, and each question received only one explanation style. The counterbalancing, on the other hand, ensured that the same question had (roughly) an equal number of observations for each explanation style (across participants).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	☐	Antibodies
☒	☐	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☒	☐	Animals and other organisms
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☒	☐	Flow cytometry
☒	☐	MRI-based neuroimaging