scientific reports

OPEN

A comparison of human and GPT-4 use of probabilistic phrases in a coordination game

Laurence T. Maloney[1,2✉], Maria F. Dal Martello[1,3], Vivian Fei[1] & Valerie Ma[1]

English speakers use probabilistic phrases such as *likely* to communicate information about the probability or likelihood of events. Communication is successful to the extent that the listener grasps what the speaker means to convey and, if communication is successful, individuals can potentially coordinate their actions based on shared knowledge about uncertainty. We first assessed human ability to estimate the probability and the ambiguity (imprecision) of twenty-three probabilistic phrases in a coordination game in two different contexts, investment advice and medical advice. We then had GPT-4 (OpenAI), a Large Language Model, complete the same tasks as the human participants. We found that GPT-4's estimates of probability both in the Investment and Medical Contexts were as close or closer to that of the human participants as the human participants' estimates were to one another. However, further analyses of residuals disclosed small but significant differences between human and GPT-4 performance. Human probability estimates were compressed relative to those of GPT-4. Estimates of probability for both the human participants and GPT-4 were little affected by context. We propose that evaluation methods based on coordination games provide a systematic way to assess what GPT-4 and similar programs can and cannot do.

English has a rich vocabulary of *probabilistic phrases* used to communicate the relative frequency or likelihood of events (We will use the term *probability* as a catch-all for relative frequency, likelihood, degree of uncertainty and probability. Following O'Brien[1] we use the term *ambiguity* as a synonym for imprecision or ambiguity). Your doctor might tell you that you will **probably** feel better after a night's sleep. What probability or range of probabilities do you think he intends to convey? What if instead he used the probabilistic phrase **possibly**? In the latter case the doctor is likely signaling a lower probability, but did he intentionally choose a more ambiguous term? The doctor may have reasons for introducing ambiguity. Perhaps he is not sure what effect a good night's sleep will have. Perhaps he wants to manage the patient's expectations.

The coordination game. In using these phrases, you, your doctor, and other English speakers you encounter are playing a simple *coordination game*[2–7] in which both players win to the extent that the probabilities and ambiguities the second player estimates are close to those the first player intended.

Human language comprises many different language games[8–10] and the successful speaker must play them all well. Our goal here is to compare human and machine performance in one of many language coordination games. Performance in the game measures how well any given player—doctor or patient—uses language as a tool[8–10]. Success in playing the game would suggest that the two players in effect share an understanding of the language of probability and ambiguity that allows them to coordinate and work together well.

A coordination game is an objective task and we can compare human and machine performance trial by trial. Failure would pinpoint shortcomings in the performance of the machine. Can a computer play the game as well as humans do? Do successes and failures in playing the game mimic those of humans?

There is disagreement on how to assess the performance of Large Language Models such as GPT-4[11,12]. We follow the recommendations of Burnell et al.[13] in our assessment: (1) We use multiple related tasks (probability and ambiguity). (2) The tasks are objective with degree of failure measured quantitatively. (3) We report exactly one run of GPT-4—the first—in each experimental condition, not an aggregate of multiple runs. (4) We look at how GPT-4 performs in the two contexts, giving medical advice and giving investment advice.

[1]Department of Psychology, New York University, 6 Washington Place, Room 574, New York, NY 10012, USA. [2]Center for Neural Science, New York University, 6 Washington Place, New York, NY 10012, USA. [3]Dipartmento di Psicologia Generale, Università di Padova, Via Venezia 8, Padua, Italy. ✉email: ltm1@ nyu.edu

Lastly, (5) the tasks we consider—communicating information about probability and ambiguity—are intrinsically important. There is a large literature concerning human error in decision making[14], failures in probabilistic reasoning[15], and the consequences of these errors and failures[16,17].

The capabilities of GPT-4 and other candidate Artificial General Intelligences have been compared to human with mixed results. GPT-4 fails simple intelligence tests[18]. On the other hand, Webb, Holyoak, and Lu[19] report that GPT-4's ability to engage in analogical reasoning and abstract pattern induction is comparable to human. Gurnee & Tegmark[20] find that it can reason about spatial and temporal structure. GPT-4 can do more than chat it can write simple computer code for applications specified in natural language[21].

The probabilistic phrases we consider (Table 1) have been used in previous research with human participants[1]. We modify any probabilistic phrase as needed so that its use in context is grammatical (e.g. **possible** can become **possibility**). In Fig. 1 we illustrate one turn of the coordination game as a communication channel[22,23]. For simplicity we assume that, in the coordination game, the first player has only one probability and one ambiguity to signal and the second player is constrained to report a single estimate of probability and one of ambiguity.

In the full coordination game (Fig. 1), Player 1 is given a target probability (for example, 63%) and must **encode** it as one of the probabilistic phrases in Table 1. Perhaps she picks **likely**. This probabilistic phrase is transmitted to Player 2 who must **decode** it and estimate the target probability. In Fig. 1 she estimates 70%. The difference (7%) in absolute value between Player 1 's target probability, 63%, and Player 2 's estimate, %70, is the **error**, a measure of failure of coordination in the coordination game.

We will focus on the second stage (DECODE) of the coordination game (outlined in red in Fig. 1), evaluating GPT-4's performance as Player 2 and comparing GPT-4's performance to that of human participants also playing as Player 2. That is, GPT-4 and its human counterpart will be asked to DECODE probability phrases and estimate corresponding probabilities. We will also ask GPT-4 and human players to rate the ambiguity (imprecision) on a scale of 0–100 of the 23 probabilistic phrases they decode. All ambiguity estimates were done after all probability estimates by both human participants and GPT-4. Probabilistic phrases were presented in randomized order to both human participants and GPT-4.

We emphasize that, from the viewpoint of Player 2 (the human participant or GPT-4) the game is the coordination game illustrated in Fig. 1 in which Player 2 believes he is trying to coordinate a choice of probability with Player 1. In actuality the role of Player 1 is played by the experimenter who provides Player 2 with probabilistic phrases in a predetermined randomized order but Player 2 does not know that. We collect and analyze data from only Player 2, the part of Fig. 1 marked with a dashed red contour. By having the researcher play the role of Player

A never	B very rare	C almost never	D low risk
E low probability	F small chance	G unlikely	H there is a chance
I possible	J perhaps	K sometimes	L could be
M moderate risk	N not certain	O significant chance	P reasonable chance
Q reasonable to assume	R likely	S probable	T most likely
U expected	V almost certain	W certain	

Table 1. Probabilistic phrases taken from[1]. Each phrase has a code letter used in plotting results. Each phrase was embedded in a Medical Context and an Investment Context. Half the participants chosen at random completed the Medical Context, the remaining participants completed the Investment Context.
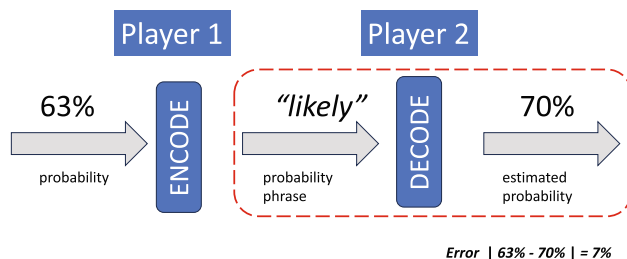
The Coordination Game

Error | 63% - 70% | = 7%

Figure 1. The Coordination Game. On each turn in the coordination game, the First Player is given a probability p (unknown to the second player) and asked to encode it as a *probabilistic phrase*. Table 1 lists the probabilistic phrases that the First Player could use to encode the probability. The Second Player (either a human participant or GPT) is then given only the selected phrase and asked to estimate the original probability. The game is a model of transmission through a communication channel, We focused on only the Decoding phase of the game (enclosed by a dashed red contour). Either GPT-4 or a human participant played the role of Player 2 while the experimenter played the role of Player 1. In a variant of the game we asked GPT-4 or the human participant to estimate not the probability but instead the ambiguity of the probability phrase.

1, we ensure comparability of input between human and GPT-4: Player 2 saw probabilistic phrases in an order determined by the experimenter alone, not the experimenter and an uncontrolled Player 1.

The essence of a 2-person coordination game[2] is that the two players must each anticipate what the other player is thinking, and the anticipation is recursive: "I know that you know that I know ….". Our game captures that essence. Player 2—human or GPT-4—is asked to see himself through the eyes of a doctor or financial advisor who is trying to communicate information about uncertainty to him.Can an LLM do this as well as a human?

There are previous studies whose participants were asked to assign explicit probabilities to probabilistic phrases[1,24–31]. See[32] or[33] for review.These studies assess the extent to which humans—for the most part without any special training—agree with each other in their use of probabilistic phrases to signal probability. If all the speakers in a language community assign the same probabilities to probabilistic phrases, then the players would do very well at the coordination game.

Questions

1. Can GPT-4 play the coordination game as well as humans? Are there patterned deviations between human and machine? We needed a criterion to judge whether GPT-4, playing as Player 2, is doing what a human player would. We developed two criteria, the first based on linear model fits, the second on performance.

The first criterion is based on fitting a linear model to bivariate scatterplot data as we explain below. They give us clues about shortcomings of GPT-4 even when GPT-4 overall plays the game well as measured by the second criterion. We do not claim that the linear model is an adequate model of the mapping from GPT-4's estimates of probability or ambiguity to the corresponding human estimates. The fitted parameters serve as summary statistics intended to aid in interpreting the data.

Second, we evaluate how well human and GPT-4 coordinate. We develop a measure—*discordance*—of the extent to which human players disagree with one another in playing the game and compare this measure to the discordance between GPT-4 and human players. Does GPT-4 perform as well in the coordination game as the median human player?

There is more to language competence than assigning probability estimates and ambiguity ratings, but systematic failure to coordinate with human participants in our game would weaken any claim that GPT-4's abilities are human-like. We could not trust an Artificial General Intelligence to give medical advice if the probability phrases it uses were not correctly understood by human patients.

2. Is GPT-4 correctly sensitive to context? The meanings of words can depend on the context in which they occur. If your doctor and your financial consultant both use the phrase ***not certain,*** does it signal the same probability? Ambiguity? We will include two contexts in the experiment, medical and financial, and ask participants, including GPT-4, to rate the 23 probabilistic phrases in Table 1 for probability and for ambiguity in each context.

.Humans distort small and large probabilities[33]. A doctor may well avoid probabilistic phrases near the extremes of the probability scale (e.g. "almost certain") precisely because he knows his patients will distort them. Or he might bias his choice of words to counteract the expected bias of his patients. Similarly, if the Player suspects the motives of a financial adviser he might try to "debias" his estimates.

Keep in mind that the issue is not whether GPT-4's assessment of probability and ambiguity is invariant under context but whether GPT-4 exhibits the same changes or lack of change in probability and ambiguity ratings across context as do the human participants. In a coordination game it doesn't matter whether you are right, only whether you agree with everyone else.

3. Is GPT-4 stable? Lastly, we briefly investigate the *stability* of GPT-4 in this game. If we rerun the estimates by GPT-4, do we get series of similar estimates or a series of similar estimates with the occasional highly discrepant estimate? Is GPT-4 stable? We might hesitate to permit an Artificial General Intelligence to give medical advice if 1 time out of 100 it produced markedly discrepant estimates of probability or ambiguity. The motivation for testing stability will become clear when we examine the data.

Results

The results are presented in three numbered sections corresponding to the numbered questions above. We split the first question in two, one part (1A) concerned with probability, one with ambiguity (1B).

Human vs. GPT-4: probability

In Fig. 2 we plot the median probability ratings assigned to each of the 23 probabilistic phrases by the 25 human participants against the GPT-4 ratings of each of the 23 probabilistic phrases. Figure 2a shows results in the Investment Context while Fig. 2b shows results for the Medical Context. The ratings range from 0 to 100%. If the median human participant agreed with GPT-4 in rating probability the plotted points would fall on the dashed blue identity line. The letter codes correspond to the letter codes assigned to each probabilistic phrase in Table 1. An intercept significantly different from 0 or a slope significantly different from 1 would indicate a patterned discrepancy between GPT-4 and the median human participant. We test for both possibilities.

There is an evident outlier in Fig. 2a for the probability phrase "low risk" plotted in red. There is a similar outlier in Fig. 2b for the probabilistic phrase "not certain". The outliers represent probabilistic phrases where GPT-4 and the median human participant assigned markedly different probabilities to the same probabilistic phrase. In the main text we report statistical analyses for this and later figures without these outliers. All results of hypothesis tests—with and without outliers—are included in a Supplement. We discuss outliers further in the section *Stability*.

We refer to tests with *p*-values less than 0.05 as "significant" for convenience in presenting the data. We report exact *p*-values for all tests in the main text and report exact *p*-values for all tests (including those with and without outliers removed) in the Supplement.
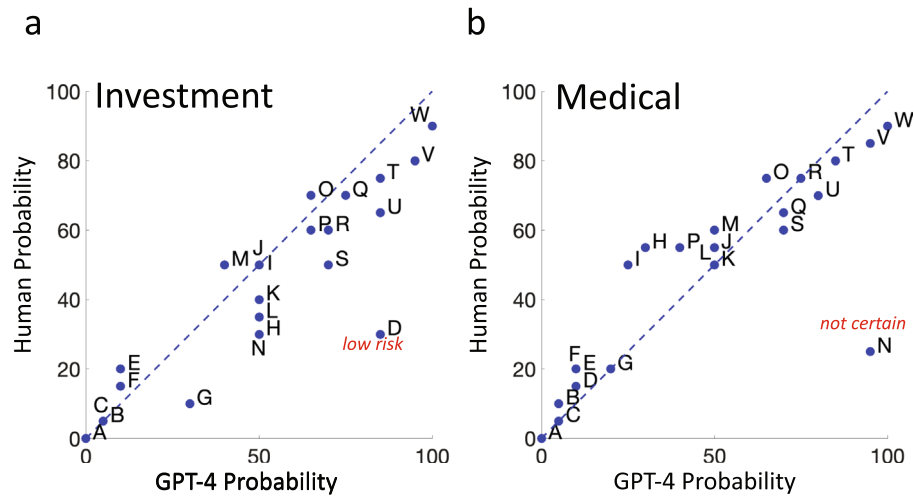
Figure 2. Human versus GPT-4. (**a**) Median human estimates of probability in the Investment Context are plotted versus GPT-4 estimates of probability in the same context (blue filled circles). A letter code adjacent to the blue filled circle identifies the probabilistic phrase associated with each circle. See Table 1. One outlier (D *low risk*) is marked with its probabilistic phrase in red. See text. (**b**) Median human estimates of probability in the Medical Context are plotted versus GPT-4 estimates of probability. One outlier (N *not certain*) is marked with its probabilistic phrase in red. See text.

Intercept The Intercept estimate in Fig. 2a is 1.59, not significantly different from 0 [t (22) = 0.421, $p = 0.339$]. The Intercept estimate in Fig. 2b is 11.35, significantly different from human 0 [t (22) = 3.458, $p = 0.0011$].

Slope The Slope estimate in Fig. 2a is 0.833, significantly different from 1 [t (22) = −2.577, $p = 0.009$]. The Slope estimate in Fig. 2b is 0.825, also significantly different from 1 [t (22) = −2.933, $p = 0.0038$].

Summary There are significant patterned differences between median human probability estimates and those of GPT-4. In both contexts median human estimates of probability are compressed by a factor of 0.8 relative to the estimates by GPT-4. In the Medical Context but not in the Investment, human estimates of probabilities are also offset vertically by roughly 10%. Human use of probability and relative frequency are typically distorted[34,35] and the deviations we detect may be connected to probability distortion.

Discordance. Both the human participants and GPT-4 are engaged in a coordination game and the second criterion of human and machine is trial by trial winnings in the game. Did GPT-4 disagree with the other human human players more than they disagreed with one another?

We define a measure of the disagreement between the probability or ambiguity estimates of the human participants. Let $p_i = [p_{i1}, \dots p_{im}]$ be the vector containing the $m = 23$ probability estimates of the ith human participant in the order of Table 1. We define the *discordance* of the ith human participant to be.

$$D_i = \sum_{j \sim = i} \left\| p_i - p_j \right\| \tag{1}$$

where $\left\| p_i - p_j \right\| = \sqrt{\sum_{k=1}^{m} \left| p_{ik} - p_{jk} \right|^2}$ denotes the Euclidean distance between p_i and p_j. The discordance of a participant is just the sum of the squared distances between the vector corresponding to the participant and each of the vectors corresponding to the remaining participants. It can be zero only if all the participants give identical estimates for all probability phrases.

Let p_{GPT} denote the vector of probability estimates made by GPT-4 and define the discordance of GPT-4 to be.

$$D_{GPT} = \left[\frac{m-1}{m} \right] \sum_{j=1}^{m} \left\| p_{GPT} - p_j \right\| \tag{2}$$

There are $m - 1$ summands in Eq. (1) and m in Eq. (2). The multiplicative term $_{m-1/m}$ in Eq. (2) corrects for the difference in the number of summands in the two equations.

Were GPT-4's judgments of probability more discrepant from those of the human participants than those of the human participants were from one another?

Figure 3a is a boxplot of all the discordance values for the Investment Context, one blue dot per human participant. The discordance values are plotted vertically, and the red horizontal line marks the median discordance in each context. The lower and upper edges of the box mark the 25th-percentile and the 75th-percentile, respectively. Figure 3b is the corresponding plot for the Medical Context. The red diamonds mark the discordances of GPT-4 in the two contexts. The discordance of GPT-4 is below the median of the discordances for the humans for both Contexts. *GPT-4 agreed with the human participants as as least as well as they agreed with one another.*
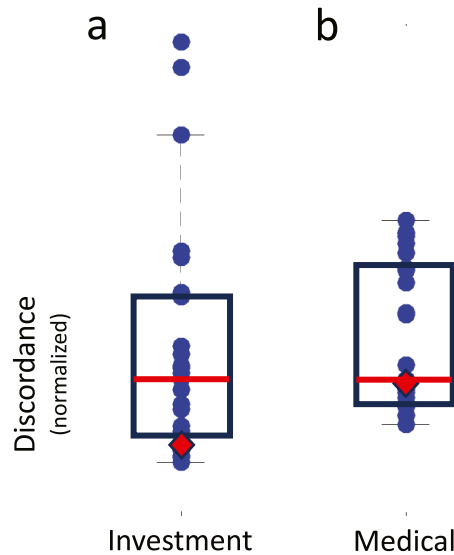
Figure 3. Discordance. We computed discordance, a measure of disagreement among each human observer and the remaining human observers and between GPT-4 and the human observers. See text. The left and right panels are boxplots of discordance values for the Investment Context and for the Medical Context, respectively. The top and bottom of the boxes mark the 75th and 25th percentiles for each context. The discordance for GPT-4 is marked by a solid red diamond in each context. The discordance for GPT-4 is below the median discordance (the solid red line segment) for the human participants in both contexts.

Human vs. GPT-4: ambiguity

In Fig. 4 we plot the median ambiguity ratings of the 25 human participants to the GPT-4 ratings of each of the 23 probabilistic phrases. Figure 4A shows results in the Investment Context while Fig. 4B shows results for the Medical Context. The letter codes once again correspond to the letter codes assigned to each probabilistic phrase in Table 1.

Intercept The Intercept estimate in Fig. 4A is 4.17, not significantly different from 0 [t (23) = 0.729, $p = 0.237$]. The Intercept estimate in Fig. 4B is 5.82, not significantly different from 0 [t (23) = 1.286, $p = 0.106$].

Slope The Slope estimate in Fig. 4A is 0.530, significantly different from 1 [t (23) = − 4.83, $p < 0.0001$]. The Slope estimate in Fig. 4B is 0.710, also significantly different from 1 [t (23) = − 3.458, $p = 0.001$].
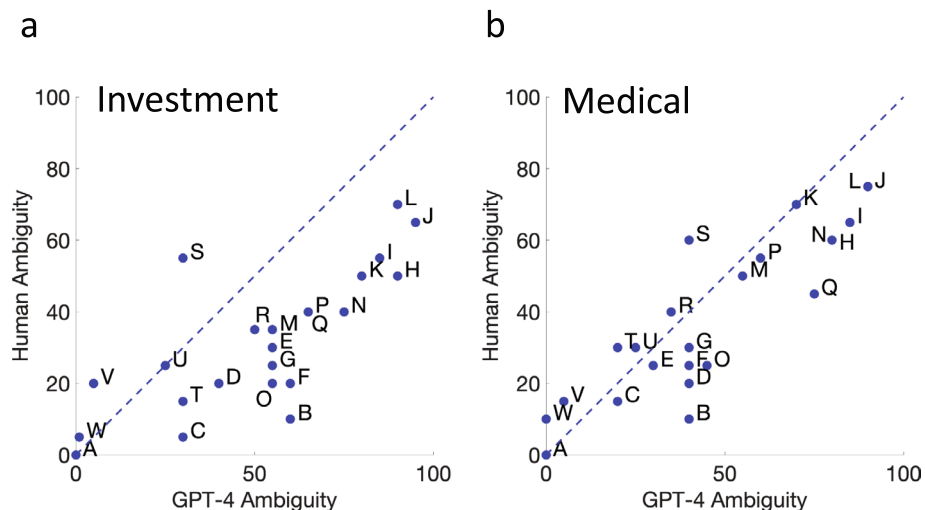

Figure 4. Human versus GPT-4. (**a**) Median human estimates of ambiguity in the Investment Context are plotted versus GPT-4 estimates of ambiguity in the same context as blue filled circles. The format is analogous to that of Fig. 2a,b. Median human estimates of ambiguity in the medical context are plotted versus GPT-4 estimates of ambiguity in the same context. The format is analogous to that of Fig. 2b.

Summary There are significant patterned differences between median human confidence estimates and those of GPT-4. In both contexts median human estimates are compressed, by a factor of 0.5 to 0.7 relative to the estimates by GPT-4.

We did not analyze discordance for ambiguity since there are evident large differences in estimation of ambiguity by GPT-4 and the median human participant.

Comparisons across context

We next evaluate the extent to which human judgments of probability and ambiguity are invariant across context. In Fig. 5A we plot the median probability ratings of the human participants in the Investment Context versus the median probability ratings for each of the 23 probabilistic phrases of a different group of human participants in the Medical Context. A similar plot for GPT-4 is shown in Fig. 5B. Figure 6A,B show corresponding plots for ambiguity.
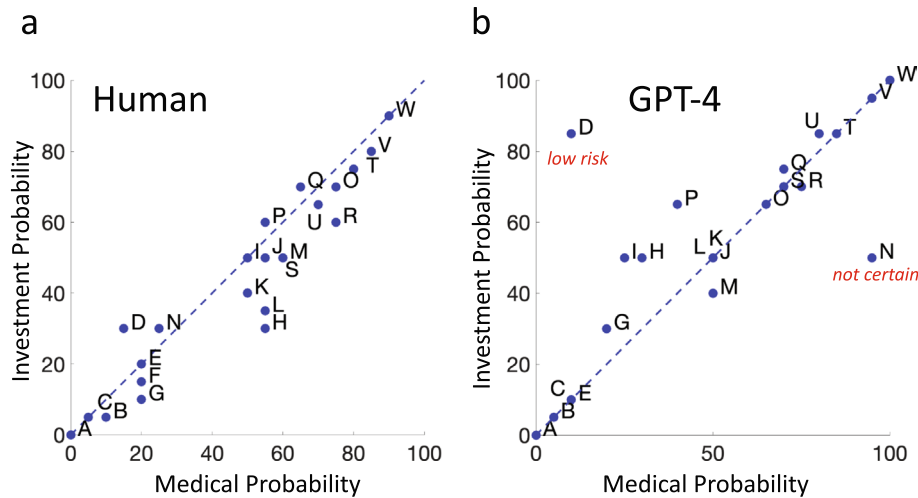

Figure 5. The effect of context. (**a**) Comparison of probability estimates of the median human observer across contexts. The format is analogous to that of the corresponding panels in Fig. 2a. All data fall roughly along the identity line. Human participants select lower probabilities for the same probabilistic phrase in the Medical Context. (**b**) Comparison of GPT-4 probability estimates across contexts. The format is analogous to that of Fig. 2b. All data fall roughly along the identity line. The same two outliers appear in Fig. 4b as in Fig. 2a,b.
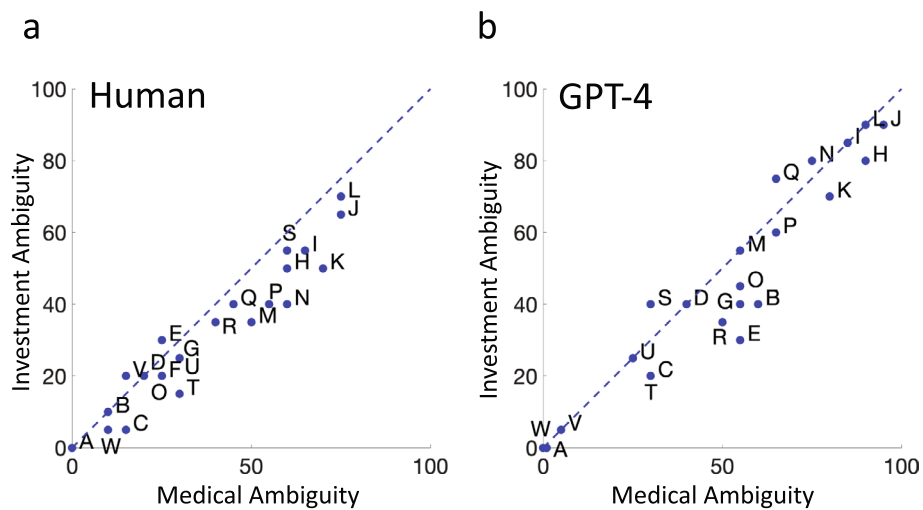

Figure 6. The effect of context. (**b**) The median of human estimates of ambiguity in the Investment Context are plotted versus the median of human estimates of ambiguity in the Medical Context. The format is analogous to that of Fig. 5a. Human estimates are significantly higher in the medical context. See text. (**b**) The median of GPT-4 estimates of ambiguity in the Investment Context are plotted versus the median of GPT-4 estimates of ambiguity in the Medical Context. GPT-4 estimates are significantly higher in the medical context. See text.

Intercept The Intercept estimate in Fig. 5A is −0.147, not significantly different from 0 [t (23) = −0.041, p = 0.484]. The Intercept estimate in Fig. 5B is −2.41, not significantly different from 0 [t (23) = −0.608, p = 0.275].

Slope The Slope estimate in Fig. 5A is 0.907, is not significantly different from 1 [t (23) = −1.41, p = 0.0865]. The Slope estimate in Fig. 5B is 0.977, not significantly different from 1 [t (21) = −0.341, p = 0.368].

Summary There are no significant patterned differences between median human probability estimates in the Investment Context and the Medical Context. There are no significant patterned differences between estimates by GPT-4 in the Investment Context and the Medical Context.

Intercept The Intercept estimate in Fig. 6A is 0.048, not significantly different from 0 [t (23) = 0.021, p = 0.492]. The Intercept estimate in Fig. 6B is 8.42, significantly different from 0 [t (23) = 2.214, p = 0.019].

Slope The Slope estimate in Fig. 6A is 0.819, significantly different from 1 [t (23) = −3.46 p = 0.0010]. The Slope estimate in Fig. 6B is 0.941, not significantly different from 1 [t (23) = −0.341, p = 0.207].

Summary There are significant patterned differences between ambiguity estimates by GPT-4 in the Investment Context and the Medical Context.

Stability

The two outliers in GPT-4's performance raise issues concerning the stability of GPT-4. We chose to examine the outlier in Fig. 2B (the probabilistic phrase *not certain*) to determine whether it reliably recurs (representing a large but reliable discrepancy between human and GPT-4 estimates) or whether it is evidence of instability. If it reliably reoccurs then it is effectively a difference of opinion between human and machine as to the meaning of a particular probabilistic phrase. If not, it would suggest that GPT-4 is unstable.

GPT-4 included explanations for its responses. We tabulate these explanations for the outlier in Fig. 2A (***low risk***) in the Investment Context and the outlier in Fig. 2B (***not certain)*** in the Medical Context. The reader may agree with GPT-4 or not, but GPT-4's response acknowledges that the probabilistic phrases can be interpreted in more than one way and perhaps human and machine are simply in disagreement.

Will the outlier recur if we rerun the trial? The GPT-4 interface that we have access to limits the number of runs that we can carry out in a fixed period of time, precluding analyses that require large numbers of repetitions of trials. We redid the GPT-4 estimates in the Medical Context four times, plotting the estimates as four blue contours in Fig. 7. The original estimates are plotted in red with a red solid circle marking the outlier. In brief, we did not reproduce the anomalous outlier we initially encountered nor did other outliers emerge for any of the other probabilistic phrases. The four new estimates of probability are in good agreement with those of the human participants and with one another but not with the original GPT-4 estimate.

Discussion

In coordination games we share information to coordinate actions[5,6]. The specific coordination game we consider here concerns correct use of probabilistic phrases signaling probability and ambiguity. There were two versions of the game, one with probabilistic phrases used to give investment advice, one with these same phrases used to give medical advice. Half the human participants ran in the Investment Context, half in the Medical.
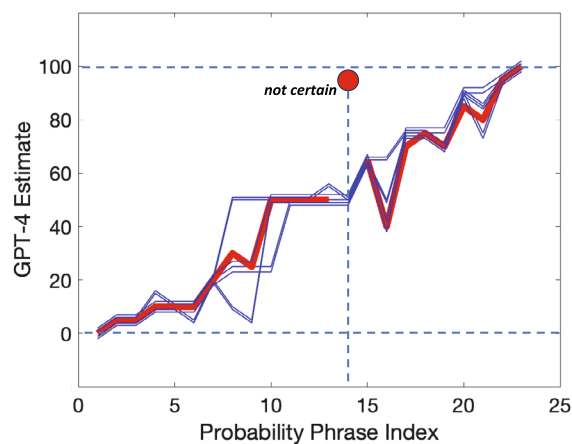

Figure 7. Stability of GPT-4 estimates in the Medical Condition. Figure 2b (GPT-4 estimates of probability in the Medical Context) contains an evident outlier *not certain* whose estimate in the Investment Condition is 50%, in the Medical Context, 95%. We plot the GPT-4 estimates from the Medical Context in red with the outlier marked by a red disk. We repeat these GPT-4 estimates four times and plot them as well, in blue. The contours are slightly offset vertically so that they are visible. The outlier does not reoccur, and the four repetitions are in good agreement with each other and the original estimates except for the outlier. See text.

Conclusions: Estimates of Probability

- As measured by discordance, GPT-4 agreed with human participants as least as well as the median human participant agreed with the other participants. Based on overall performance we cannot distinguish GPT-4 and human (Fig. 3).
- Examined in detail (linear model fits), we found significant patterned discrepancies between GPT-4's estimates of probability and those of human participants that could be captured by fits to a linear model (Fig. 2). In both contexts, human estimates of probabilities tended to be compressed relative to those of GPT-4.
- Use of probabilistic phrases by both GPT-4 and humans signaling probability transferred well across contexts for the two contexts we considered (Fig. 5). A doctor's use of *likely* conveys the same information about probability as that of an investment consultant.

Conclusions: Estimates of Ambiguity

Human estimates of ambiguity were compressed relative to those of GPT-4 by roughly a factor of 2. However, unlike probability, there is no standard scale of ambiguity. We can only conclude that GPT-4 did not anticipate human use of the ambiguity scale, a failure to coordinate.

There is some indication that GPT-4 is unstable, producing occasional outliers. Further research is needed to evaluate this apparent instability (Table 2).

We focused on one coordination game and compared human and machine. Similar games could be based on color terms or sets of dimensional adjectives[36], for example, the dimensional adjectives describing size: *small*, *big*, *large,* etc[36,37]. Gurnee & Tegmark[20] look at representation of space and time.

But human use of probability phrases is a particularly rich source of possible coordination games that we could use to compare human and machine. We can challenge GPT-4 to play each game we develop, comparing human and machine as we did here.

When, for example, do humans use probability phrases and when do they use numerical probability? Dhami & Mandel[38] in their review article argue that the choice between the use of numerical or verbal probabilities by senders is influenced by several factors. For example, Juanchich and Sirota[39] find that, in the medical context, senders prefer to use numerical values when uncertainty is about very consequential events, as, for example, the serious side effects of a drug. Would GPT-4 have similar preferences?

Wallsten et al.[40] found that most people preferred to receive information about the probability of a chance event in numerical form but preferred to transmit this information as a probabilistic phrase. Erev & Cohen[41] referred to this pattern of preference as the Communication Mode Preference Paradox. Whatever justification we offer for transmitting a probabilistic phrase instead of a numerical probability would seem to apply to receiving it in the same form, an apparent paradox. Would GPT-4 exhibit the same paradox?

Senders' use of verbal probabilities has several effects other than conveying an estimate of uncertainty[38]. Honda[42] found that the use of verbal probabilities, for example using positive rather than negative terms, can introduce bias into the decison making process. Verbal probabilities can also be used as receiver's and sender's face-saving strategy[43–45]. Does GPT-4 exhibit similar biases?

Comparing GPT-4 to human in these coordination games provides a systematic way to assess what GPT-4 (or any other LLM) can and cannot do, where its performance matches, exceeds, or falls short of, human. There is more to language than a series of coordination games but such games provide a scaffolding allowing us to describe what GPT-4 does in a principled way.

Context	Investment	Medical
"Low risk"	Estimate: 85%	Estimate:10%
	Justification: When you describe something as "low risk" in the context of investments increasing in value, you're conveying that there's a high likelihood the investments will perform positively or, at the very least, not decline significantly. Given the options provided, and interpreting "low risk" as a high probability of the investments increasing in value, I would estimate the communicated probability to be: 85%. This reflects that there's a high likelihood of positive performance, but as with all investments, there's never a guarantee. Always remember that while terms like "low risk" are commonly understood in finance, their interpretation can still vary somewhat based on individual perception and context."	**Justification**: 'If I use the phrase "low risk" to describe the probability of a headache occurring as a side effect, I'm suggesting that it's not a common outcome, but there's still some possibility. Given the list of percentages you provided, "10%" seems to be a good representation of what "low risk" might imply in this context."
"Not certain"	Estimate: 50%	95%
	The phrase "not certain" is somewhat ambiguous. While it indicates that there isn't a guarantee, it doesn't provide a clear indication of how likely or unlikely the event is. Given its vagueness, but still leaning toward the idea that there's less confidence than "certain," a neutral probability might be represented. Therefore, I would estimate the communicated probability for "not certain" to be: 50%. However, it's essential to understand that this phrase leaves a lot to interpretation, and different people might infer different probabilities from it	The phrase "not certain" is ambiguous and indicates that there's uncertainty, but it doesn't necessarily provide a clear indication of how probable or improbable the event is. It could be slightly probable or highly probable, but just not 100% certain. Given the options provided and the context, "95%" might be the best choice to convey that there's a high likelihood, but it's not certain. However, it's important to note that the phrase "not certain" doesn't inherently align with any specific percentage; the chosen value is more of an educated guess based on the given options

Table 2. GPT-4's Justifications of Estimated Probabilities. GPT-4's probability estimates for the two different contexts were discrepant for probabilistic phrases "low risk " and "not certain ". GPT-4's output includes an explanation of each choice which we include here.

One puzzle we are left with is the compressions of probability and ambiguity we found. Despite these failures, GPT-4's overall performance measured by discordance is better than that of the median human participant. Yet a simple scaling of output would "fix" the compression problem and presumably improve its performance. The methods used to develop GPT-4 did not result in an LLM that included appropriate scalings. Why?

Methods

GPT-4

Typical inputs for GPT-4 are shown in Table 3 for both contexts and for estimates of probability and ratings of ambiguity. Input consisted of a single context phrase followed by a request for a rating of either the probability or ambiguity of a specified probabilistic phrase. The phrases were presented in randomized order. GPT-4 was constrained to respond with a single percentage (probability) or number (ambiguity) and a justification. The values permitted were multiples of 5% from 0 to 100% (0%, 5%, …, 100%) and multiples of 5 from 0 to 100 for ambiguity ratings. We analyzed only the first run of GPT-4 in Figs. 1, 2, 3, 4, 5, 6.

Human participants

Fifty participants drawn from the New York University SONA Subject Pool agreed to participate in the experiment. All completed the experiment. We report the demographics of the pool. Almost all the potential participants in the pool were between the ages of 18–35 (93.34%), only a small portion of the participants were over 35 (6.38%) or under 18 (0.27%). 30.9% of the potential participants were male at birth and 69.0% were female at birth.

The research protocol and methods were approved by the Institutional Review Board for that Faculty of Arts and Science at New York University (IRB-FY2023-7544). Informed consent was obtained from all subjects and/or their legal guardian(s). Alll methods were carried out in accordance with relevant guidelines and regulations.

Procedure

Each participant was taken to a laboratory room and seated in front of a computer screen. Each participant was assigned at random to one of two contexts, investment or medical, 25 participants in each. Stimuli were presented and responses recorded using Google forms.

Payment

Participants were paid $12 / hour for the time spent in the experiment (typically 15 min or less).

Participants assigned to the medical context answered questions about probability phrases and ambiguity of probability phrases related to medical advice (See Table 3). The questions were prefaced by a single sentence signaling context as were those posed to GPT. Participant assigned to the investment context answered questions related to investment advice (See Table 3).

Questionnaires in both contexts consisted of same two sections: probability and ambiguity. Section "1A. Human vs. GPT-4: Probability" asked subjects to rate the probability of 23 words and phrases on a percentage scale from 0 to 100% (with 5% increments). Section "Comparisons across context" asked participants to rate the ambiguity of the same 23 phrases using a scale ranging from 0 to 100 in steps of 5. Questions within each section were randomly re-ordered for each participant.

Hypothesis tests

In each panel of Figs. 2, 4, 5, 6 we plotted corresponding data (e.g. estimates from human participants and corresponding estimates from GPT-4) as a scatterplot, allowing the reader to assess the relations between variables visually. We fit the data in each scatterplot by a univariate linear model. The null hypothesis for each test was that the points fell on the identity line with slope 1 and intercept 0 with added iid Gaussian error. We used hypothesis

A Investment probability	B Medical probability
Imagine you are a financial consultant trying to convey to a client the probability that their investments will increase in value over the next year. If you use the phrase '***almost certain***' what probability are you trying to communicate? *[Scale from 0 to 100% in steps of 5%.]*	Imagine you are a doctor trying to convey to a patient the probability of a headache occurring as a side effect from a drug you had prescribed. If you use the phrase '***almost certain***' what probability are you trying to communicate? *[Scale from 0 to 100% in steps of 5%.]*
C Investment ambiguity	D Medical ambiguity
Imagine you are a financial consultant trying to convey to a client the probability that their investments will increase in value over the next year. You use the phrase '***almost certain***'. How ambiguous is that term? *[Scale from 0 to 100 in steps of 5.]*	Imagine you are a doctor trying to convey to a patient the probability of a headache occurring as a side effect from a drug you had prescribed. You use the phrase '***almost certain***'. How ambiguous is that term? *[Scale from 0 to 100 in steps of 5.]*

Table 3. Sample input for GPT-4 for the probabilistic phrase "almost certain" as presented in four conditions. For each condition we constructed 23 stimuli for the 23 probabilistic phrases in Table 1. Participants rated the probability (0%, 5%, 10%, …, 100%) and the ambiguity (0, 5, 10, …, 100) of each phrase (0 = least ambiguous, 100 = most ambiguous). Ambiguity ratings were collected after probability ratings. The scale values were enumerated explicitly, 0, 5, … 100. Human participants completed A and then C or B and then D but not both. GPT-4 ran A and then C and was then restarted (erasing knowledge of trials completed) before completing B and D. See text.

tests to detect deviations of slope from 1 or intercept from 0. We discuss any significant trend captured by the estimated slope and intercept values (compression or offset).

All tests were two-tailed with size 0.05. We report the t-statistic, the degrees of freedom and the exact *p*-value for each test and refer to outcomes with *p*-value less than 0.05 as "significant" in discussing the data. We classified two data points as outliers, and label them where they appear in the scatterplots by their probabilistic phrases in red. We report results with outliers excluded in the main text and report the analyses of all tests with outliers excluded and with outliers included in the Supplement.

Data availability
All data available from the corresponding author.

References
1. O'Brien, B. J. Words or numbers? The evaluation of probability expressions in general practice. *J. R. Coll. Gen. Pract.* **39**(320), 98–100 (1989).
2. Schelling, T. C. *The Strategy of Conflict.* Harvard University Press (1960).
3. Lewis, D. *Convention.* Blackwell (2002).
4. Franke, M. Game-theoretic pragmatics. *Philos. ompass* **8**(3), 269–284 (2013).
5. Benz, A. *et al.* (eds) *Language, Games, and Evolution* (Springer, 2011).
6. Benz, A. *et al.* (eds) *Game Theory and Pragmatics* (Springer, 2014).
7. Benz, A. & Stevens, J. Game-theoretic approaches to pragmatics. *Annu. Rev. Ling.* **4**, 173–191 (2018).
8. Wittgenstein, L. *Philosophical Investigations.* Translation of *Philosophische Untersuchungen*, G. E. Anscombe [translator]. New York: Macmillan (1953).
9. Austin, J. L. *How to do Things with Words* (Oxford University Press, 1955).
10. Grice, P. *Studies in the Way of Words.* Harvard University Press (1991).
11. Mitchell, M. How do we know how smart AI systems are? *Science*, **381**(6654) (2023).
12. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's Large Language Models. arXiv:2210.13966v3 [cs. LG] (2023).
13. Burnell, R. *et al.* (13 more authors) Rethink reporting of evaluation results: Aggregate metrics and lack of access to results limit understanding. *Science* **380**(6641), 136–138 (2023).
14. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. **47**(2), 263–292 (1979).
15. Tversky, A. & Kahneman, D. Belief in the law of small numbers. *Psychol. Bull.* **76**(2), 105–110 (1971).
16. Gilovich, T. *How We Know What Isn't So.* Free Press (1993).
17. Gilovich, D., Griffin, T. & Kahneman, D. *Heuristics and Biases.* Cambridge University Press (2002).
18. Biever, C. The easy intelligence tests that AI chatbots fail. *Nature* **619**, 686–689 (2023).
19. Webb, T., Holyoak, K. J. & Lu, H. Emergent analogical reasoning in large language models. arXiv:2212.09196v2 (2023).
20. Gurnee, W. & Tegmark, M. (2023), Language models represent space and time. arXiv:2310.02207v1
21. Poldrack, RA., Lu, T. & Beguš, G. AI assisted coding: Experiments with GPT-4. April 27, 2023, arXiv:2304.13187v1 [cs.AI] (2023).
22. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Techn. J.,* **27** (3)*,* 379–423*,* 623–656 (1948).
23. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication.* Illinois (1949).
24. Lichtenstein, S. & Newman, J. R. Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychon. Sci.* **9**, 563–564 (1967).
25. Beyth-Marom, R. How probable is probable? A numerical translation of verbal probability expressions. *J. Forecast.* **1**, 257–269 (1982).
26. Bryant, G. D. & Norman, G. R. Expressions of probability: Words and numbers. *N. Engl. J. Med.* **302**, 411 (1980).
27. Budescu, D. V. & Wallsten, T. S. Consistency in interpretation of probabilistic phrases. *Organ. Behav. Hum. Decis. Process.* **36**, 391–405 (1985).
28. Kong, A., Barnett, G. O., Mosteller, F. & Youtz, C. How medical professionals evaluate expressions of probability. *N. Engl. J. Med.* **315**, 740–744 (1986).
29. Mapes, R. E. A. Verbal and numerical estimates of probability in therapeutic contexts. *Soc. Sci. Med.* **13A**, 277–282 (1979).
30. Sawant, R. & Sansgiry, S. Communicating risk of medication side-effects: Role of communication format on risk perception. *Pharm. Pract.* **16**, 1174 (2018).
31. Mellers, B. A. *et al.* How generalizable is a good judgement? A multi-task, multi-benchmark study. *Judgm. Decis. Mak.* **12**, 369–381 (2017).
32. Mosteller, F. & Youtz, C. Quantifying probabilistic expressions. *Stat. Sci.* **5**(1), 2–12 (1990).
33. Zhang, H. & Maloney, L. T. Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neurosci.* **6**, 1–14 (2012).
34. Tversky, A. & Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* **5**(4), 297–323 (1992).
35. Faller, M. Dimensional adjectives and measure phrases in vector space semantics. In *Formalizing the Dynamics of Information*. In Faller, M., Kaufmann, S. & Pauly, M. [Eds.] CSLI Publications (1990).
36. Bierwisch, M. Some semantic universals of German adjectivals. *Found. Lang.* **3**, 1–36 (1967).
37. Maloney, L. T. & Gelman, S. A. Measuring the influence of context: The interpretation of dimensional adjectives. *Lang. Cogn. Process.* **2**(3–4), 205–215 (1987).
38. Dhami, M. K. & Mandel, D. R. Communicating uncertainty using words and numbers. *Trends Cogn. Sci.* **26**(6), 514–526 (2022).
39. Juanchich, M. & Sirota, M. Most family physicians report communicating the risks of adverse drug reactions in words (vs, numbers). *Appl. Cogn. Psychol.* **34**, 526–534 (2020).
40. Wallsten, T. S., Budescu, D. V., Zwick, R. & Kemp, S. M. Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bull. Psychon. Soc.* **31**(2), 135–138 (1993).
41. Erev, I. & Cohen, B. L. Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organ. Behav. Hum. Decis. Process.* **45**(1), 1–18 (1990).
42. Honda, H. *et al.* Decisions based on verbal probabilities: Decision bias or decision by belief sampling? In *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (Gunzelmann, G. et al., Eds.), 557–562, Cognitive Science Society (2017).
43. Bonnefon, J.-F. & Villejoubert, G. Tactful or doubtful? Expectations of politeness explain the severity bias in the interpretation of probability phrases. *Psychol. Sci.* **17**, 747–751 (2006).
44. Juanchich, M. *et al.* The perceived functions of linguistic risk quantifiers and their effect on risk, negativity perception and decision making. *Organ. Behav. Hum. Decis. Process.* **118**, 72–81 (2012).

45. Jenkins, S. C. & Harris, A. J. L. Maintaining credibility when communicating uncertainty: the role of directionality. *Think. Reason.* **27**, 97–12 (2020).

Acknowledgements

Author contributions

L.T.M., M.D.M., V.F. and V.M. designed the experiment and analyses. V.F. and V.M. implemented and ran the design as a computer-controlled experiment with human participants and M.D.M. implemented and ran the GPT-4 version. L.T.M., M.D.M., V.F. and V.M. analyzed data and prepared plots and tables. L.T.M., M.D.M., V.F., V.M. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-56740-9.

Correspondence and requests for materials should be addressed to L.T.M.

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.