Incoherent Probability Judgments in Large Language Models

Jian-Qiao Zhu (jz5204@princeton.edu)
Department of Computer Science
Princeton University

Thomas L. Griffiths (tomg@princeton.edu)
Department of Psychology and Computer Science
Princeton University

Abstract

Autoregressive Large Language Models (LLMs) trained for next-word prediction have demonstrated remarkable proficiency at producing coherent text. But are they equally adept at forming coherent probability judgments? We use probabilistic identities and repeated judgments to assess the coherence of probability judgments made by LLMs. Our results show that the judgments produced by these models are often incoherent, displaying human-like systematic deviations from the rules of probability theory. Moreover, when prompted to judge the same event, the mean-variance relationship of probability judgments produced by LLMs shows an inverted-U-shaped like that seen in humans. We propose that these deviations from rationality can be explained by linking autoregressive LLMs to implicit Bayesian inference and drawing parallels with the Bayesian Sampler model of human probability judgments.

Keywords: Large Language Model, Probability Judgments, Coherence, Bayesian model.

Introduction

While Large Language Models (LLMs) are primarily trained to predict the next word in text, they have demonstrated a remarkable breadth in their capabilities across various cognitive tasks and domains (Bubeck et al., 2023). The success of LLMs presents new opportunities to deepen our understanding of both machine and human intelligence. Notably, LLMs can generate long coherent passages of texts, although they occasionally produce fabricated information, a phenomenon referred to as "hallucination" (Radford et al., 2019). While hallucination predominantly affects the factual accuracy of the output, it generally does not compromise its coherence. This raises an interesting question: are the judgments made by LLMs also coherent, even when they are factually incorrect?

In this paper, we focus on evaluating probability judgments – a domain where coherence can be rigorously assessed against the rules of probability theory. We tasked LLMs with estimating the probability of various future events set in the year 2025 (see Table 1 for details). To quantitatively assess the incoherence of these probability judgments, we employed a series of probabilistic identities, comprehensively listed in Figure 1. According to these identities, coherent judgments should invariably yield a value of zero. Any deviation from this benchmark serves as a measure of incoherence in the LLMs' probability judgments. This approach aligns with the framework proposed by Costello and Watts (2014), who have previously used these probabilistic identities to evaluate human probabilistic reasoning. Moreover, as LLMs generate a probability distribution over probability judgments, we can also examine the stochasticity in these judgments (i.e., consistency over time/information) by repeatedly resubmitting the exact same prompt to evaluate the variability in responses.

To preview our results, we found that four state-of-the-art LLMs varying in training dataset size, computational resources, and the number of model parameters consistently exhibit biases in probabilistic identities that mirror those found in human cognition (Costello & Watts, 2014; Zhu et al., 2020). Furthermore, when queries about the same event are made repeatedly at temperature setting 1.0 (i.e., when the outputs of LLMs are based on the model's learned probabilities), we find that the mean and the variance of probability judgments calculated over repetitions display a pattern resembling an inverted-U shape, akin to the mean-variance relationships derived from human probability judgments (Sundh et al., 2023; Zhu et al., 2023).

We propose that the human-like patterns in probability judgments produced by the LLMs may stem from the use of the autoregressive training objectives (cf. McCoy et al., 2023). To understand this phenomenon, we examined two prominent computational models that explain human probability judgments: the Probability Theory plus Noise (PT+N) model (Costello & Watts, 2014) and the Bayesian Sampler model (Zhu et al., 2020). Although the LLMs' responses do not qualitatively differentiate between these two models, we argue that the responses align more closely with the Bayesian Sampler model. This model, which suggests responses are guided by Bayesian inference, appears to better capture the neural networks' behavior than the PT+N model's characterization based on the probability theory distorted by noise. This hypothesis is bolstered by drawing parallels between the autoregressive processes in LLMs and implicit Bayesian inference mechanisms, as recently argued for in machine learning (Griffiths et al., 2023; Xie et al., 2021; Zhang et al., 2023).

Background

There is a rich history of exploring biases in human probabilistic reasoning, a line of inquiry that, arguably, dates back to the early 19th century with the work of Laplace (Miller & Gelman, 2020) and become influential via the "heuristics and biases" research program (Gigerenzer & Gaissmaier, 2011; Kahneman & Tversky, 1972). Here, rather than testing specific biases or fallacies in LLMs, as has been explored in Binz and Schulz (2023) and Horton (2023), we choose to focus on analyzing the statistical properties of LLMs' probability judgments for a range of events. This can be achieved

Table 1: Twelve weather-related and twelve political pairs of events posed to the LLMs.

Weather A	Weather B	Politics A	Politics B
Rainy	Cold	Britain left the EU	Greece left the EU
Cloudy	Stormy	American cars increase	Petrol price increases
Chilly	Thundery	Climate change impacts American weather	World greenhouse gas emissions are reduced
Hot	Windy	The US is at war in the Middle East	Major terrorist attack occurs in the US
Sunny	Misty	Europe grows poorer	Unemployment in Europe rises above 20%
Foggy	Wet	Hurricanes and typhoons are more frequent	Average world temperature increases
Freezing	Humid	Generative AI is a trillion-dollar market	AI-designed antibiotics are available on prescription
Drizzly	Breezy	All television become Internet-based	AI has made full-length movies
Hazy	Warm	Tech unemployment has risen	The divide in income levels has expanded
Icy	Snowy	Cities ban fossil-fuel vehicles	One-third of new cars are electric
Breezy	Dry	Depression becomes the No.1 disease burden	Manufacturing jobs disappear in the West
Normal	Typical	The majority of UK homes are rented	Married couples are a minority in the UK

through (i) the combination of individual probability judgments into probabilistic identities, and (ii) the repeated elicitation of probability judgments for identical events.

Probabilities in Large Language Models

Autoregressive LLMs are trained to predict the next word, where the training distribution is typically based on a large sample of Internet text. These models typically use a neural network architecture known as the Transformer (Vaswani et al., 2017). To predict the next word, LLMs compute a conditional probability given the preceding sequence of n words: $p(w_{n+1}|w_{1:n})$. The precision of this distribution is controlled by the temperature parameter. When the temperature is set low (i.e., close to 0), the model is more likely to pick the most probable next word based on its training. A higher temperature setting makes the model's word choices more equal in probability, leading to more diverse outputs.

As AI systems based on LLMs are becoming more capable of interacting with people in real-world applications, AI researchers have been studying the probabilities that these systems predict for different events (e.g., Jiang et al., 2012; Liang et al., 2022). The primary objective within AI and machine learning is to calibrate these AI-generated probability estimates with true frequencies (Kumar et al., 2019; Shrivastava et al., 2023; Zadrozny & Elkan, 2001). While it is vital to ensure that the probability judgments produced by AI are well-calibrated, the coherence of probability judgments is equally important. In fact, there are strong theoretical reasons linking coherence and accuracy in boundedly rational agents (Zhu et al., 2022).

It is true that accurate probability judgments (i.e., those that perfectly reflect true frequencies) are also coherent. However, the reverse is not always true: coherent probability judgments are not necessarily more accurate. It has been proved that, for any set of incoherent probability judgments, there exists at least one set of coherent probability judgments that better correspond to the true frequencies (Leitgeb & Pettigrew, 2010).

Assessing Coherence via Probabilistic Identities

Coherent probability judgments should be consistent when applied to events that are logically related. This type of coherence can be studied using probabilistic identities, which are formulated by combining individual probability judgments from a pair of binary events, A and B (Costello & Watts, 2014; Zhu et al., 2020). In typical experimental setups, participants are asked about either single events (e.g., $P(A)$), conjunctions of the events (e.g., $P(A \cap B)$), or disjunctions of the events (e.g., $P(A \cup B)$). These individual probability judgments are asked independently and subsequently combined to construct probabilistic identities. For example, the identity Z_1 is composed of four distinct probability judgments: $Z_1 = P(A) + P(B) - P(A \cap B) - P(A \cup B)$. A key feature of these identities is that they should all equal to zero, according to probability theory.

These identities serve as instrumental tools for quantifying the degree of incoherence in an individual's probability judgments. An individual strictly following the rules of probability theory should make all such identities zero. Moreover, when an agent uses an unbiased estimate of the underlying probabilities (e.g., by drawing i.i.d. samples and using relative frequency for probability estimation), the identities derived from her responses should also equal zero, on average.

When these probabilistic identities are tested on human participants (see Figure 1, left panel), the results consistently show that the extent of the deviations from zero appears to be influenced by the balance of positive and negative terms within each probabilistic identity (Chater et al., 2020; Costello & Watts, 2014; Zhu et al., 2020). For instance, identities Z_1 and Z_2, which each contain an equal number of positive and negative terms (two of each), typically yield mean responses that approximate zero. In contrast, identities Z_3 through Z_6, which have one more positive term than negative, demonstrate a different pattern in mean responses. This disparity becomes more pronounced in identities Z_7 and Z_8, where there are two more positive terms than negative. Consequently, the mean responses for Z_7 and Z_8 are observed to be twice as large as those for Z_3 through Z_6.
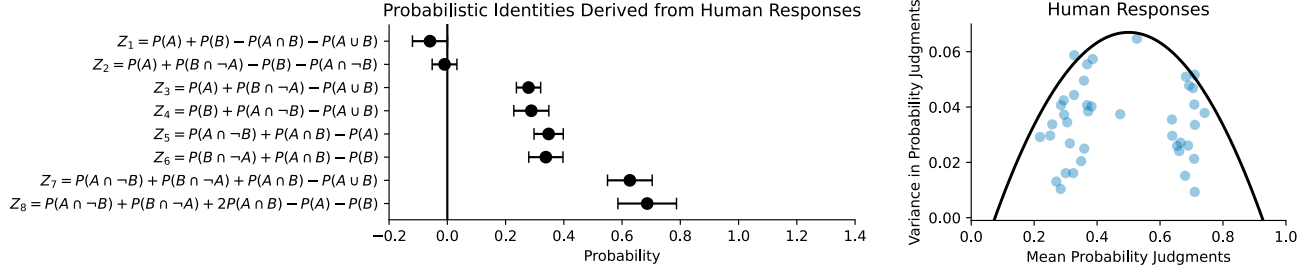
Figure 1: Bias and variability in human probability judgments as revealed by **(left)** probabilistic identities and **(right)** mean-variance relationship. Error bars are 95% CI. Solid line represents the best-fitting linear regression. Data adapted from Zhu et al. (2020) Experiment 1.

It is important to note that if an individual's probability judgments remain constant across various queries (for instance, consistently responding with a probability of 0.5 for all queries), their probabilistic identities would deviate in a manner similar to that observed in human judgments. However, it is rarely the case that individuals respond to all queries with a fixed probability value. In reality, even when presented with the same query on different occasions, individuals often exhibit variability in their probability judgments (Sundh et al., 2023). This is also true of LLMs, which generate stochastic responses even when presented with identical prompts.

Repeated Judgments

Coherent probability judgments should also remain consistent as long as the underlying information or evidence hasn't changed. If one judges the probability of an event at one moment, a coherent reassessment at a later time without new information should yield the same probability. This kind of self-consistency is studied using repeated judgments, as variability in probability judgments is important for understanding the underlying mechanisms that people use to formulate these judgments.

In economics and psychology, there have been numerous experiments where individuals are asked to repeat the same task at different times, sometimes even when these occasions are close together in time (e.g., Mosteller and Nogee, 1951; Sundh et al., 2023; Zhu et al., 2020). This has led to the observation of significant fluctuations in behavior. For instance, preferences between different monetary gambles have been noted to display marked variability within short time frames (Loomes & Sugden, 1998; Mosteller & Nogee, 1951). Moreover, people's probability judgments of the same event have been found to differ from one instance to another, even in the absence of any new information influencing these judgments (Sundh et al., 2023; Zhu et al., 2020).

In addition to this observed variability, certain structures have been identified using repeated probability judgments. Notably, it has been observed that probability judgments closer to 0.5 exhibit greater variability compared to those near the extremities of 0 or 1 (Sundh et al., 2023). This phenomenon renders an inverted-U-shaped relationship between the mean and variance of repeated judgments (see Figure 1, right panel). This inverted-U-shaped curve also demonstrates variation across different participants and events (Sundh et al., 2023). Generally, a downward shift in the curve indicates a reduction in the variability of repeated judgments; an inward shift of the curve suggests a 'shrinkage effect,' which has been theorized to occur when employing a 'stronger' prior to moderate all judgments (Zhu et al., 2023).

Evaluating Coherence in LLMs

The probabilistic identities and repeated judgments used to study the coherence of human judgments provide an opportunity to study coherence in LLMs. To this end, we evaluated four distinct LLMs with varying levels of sophistication and model transparency. These included OpenAI's GPT-4 and GPT-3.5-turbo (Achiam et al., 2023), as well as Meta's LLaMA-2-70b and LLaMA-2-7b (Touvron et al., 2023).

Methods

Each of these LLMs is equipped with a system message feature, which facilitates the application of instructions and ensures the consistency of constraints across different scenarios. For our experiment, we used a uniform system message for all probability queries, directed to each of the LLMs. The message instructed: "*You will estimate probabilities of some real-world events. Respond only with a number corresponding to a probability between 0 and 1. Do not respond with any other text. We are interested in your subjective evaluation of the probability so just respond what you think.*"

We presented LLMs with 24 pairs of constituent events, denoted A and B (detailed in Table 1). The LLMs were tasked with assessing the probability of various combinations of these events, specifically $P(A)$, $P(B)$, $P(A \cap B)$, $P(A \cap \neg B)$, $P(B \cap \neg A)$, $P(A \cup B)$. Of the 24 event pairs, 12 were related to weather conditions and the remaining 12 pertained to political scenarios. For the weather-related events, the prompts posed to the LLMs were framed as follows: "*What is the probability that the weather will be X on a randomly-selected day in England during the year 2025?*". In contrast, for the political events, the prompts were structured as: "*What is the probability that X by the year 2025?*". Appropriate logical
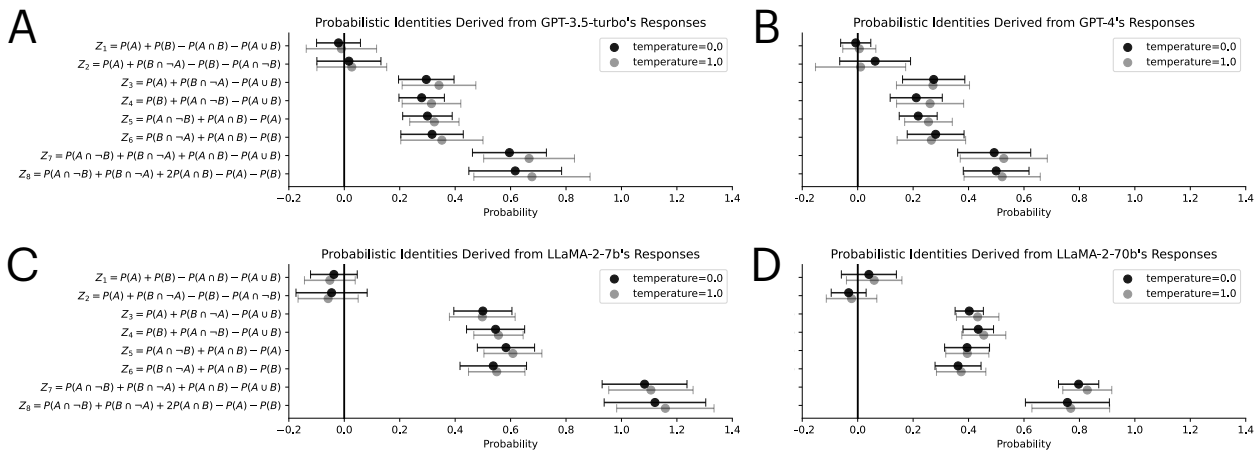
Figure 2: Probabilistic identities based on LLM responses. For coherent judgments, all probabilistic identities should be zero. **(A)** GPT-3.5-turbo model. **(B)** GPT-4 model. **(C)** LLaMA-2 model with 7b parameters. **(D)** LLaMA-2 model with 70b parameters. Error bars represent 95% CI across the 24 event pairs. Increasing the configured temperatures within each LLM, thereby introducing more stochastic responses, slightly elevates the average level of incoherence observed.

operators, including negation, conjunction, and disjunction, were applied to the pairs of events.

We configured the temperature parameter for all four LLMs to both 0 and 1. A zero temperature setting was selected to prompt each model to generate its most probable response. Conversely, a temperature setting of 1 was employed to make the LLMs to produce responses based on their learned probability distributions. To assess the variability of the responses, we submitted identical prompts to each LLM five times, with the temperature parameter maintained at 1 during these repetitions.

Results

As shown in Figure 2, the responses of LLMs exhibit characteristics akin to human reasoning biases. This similarity is primarily due to the way positive and negative terms within each identity collectively influence the extent of deviations from zero. The relationship between the balance of positive and negative terms and the magnitude of deviations from zero remains consistent across different types and variants of models. Moreover, the mean-variance relationships of repeated probability judgments produced by LLMs also resemble the human pattern (see Figure 3). The consistency of reproducing human patterns across LLMs strongly implies that these statistical properties of LLM responses is an inherent characteristic of autoregressive models.

On average, model variants characterized by either an increased number of model parameters or a lower temperature setting exhibit reduced deviations from zero in probabilistic identities, a trend observed across both the GPT and LLaMA series. Additionally, variants with a greater number of model parameters display reduced variability and a less pronounced 'shrinkage effect,' as evidenced by the mean-variance curves shifting concurrently downward and outward.

These observed patterns suggest a correlation between increased model size and enhanced coherence, although it is important to note that perfect coherence is not achieved in any model.

Theories of Human Probability Judgments

Before we delve into possible accounts for the observed patterns in probability judgments produced by LLMs, it is instructive to revisit theories of human probability judgments due to the behavioral parallels observed between LLMs and people.

Human biases in probability judgments have been explained using two leading models: PT+N (Costello & Watts, 2014) and the Bayesian Sampler model (Zhu et al., 2020). Both models posit the existence of an underlying event probability, denoted θ, which, while coherent, may not necessarily align with actual probabilities. For example, when judging $P(\text{head})$ and $P(\text{tail})$ for a fair coin, θ can manifest as both coherent and accurate, exemplified by $\theta_{\text{head}} = \theta_{\text{tail}} = 0.5$. Alternatively, θ can be coherent yet inaccurate, such as $\theta_{\text{head}} = 0.6, \theta_{\text{tail}} = 0.4$, where the underlying probabilities deviate from the actual probabilities of 0.5 in a fair coin. The key to coherence, in this example, lies in the condition that $\theta_{\text{head}} = 1 - \theta_{\text{tail}}$, ensuring the internal consistency of logically related events.

Both PT+N and Bayesian Sampler models incorporate the concept of an approximation process, which involves drawing samples from the underlying event probability θ. A crucial distinction between the two theories lies in the handling of these samples: the PT+N model posits that these samples are subject to corruption by independent noise factors (Costello & Watts, 2014). Therefore, the PT+N's interpretation of the observed patterns in probabilistic identities hinges on the presence of additive noise within individual probability
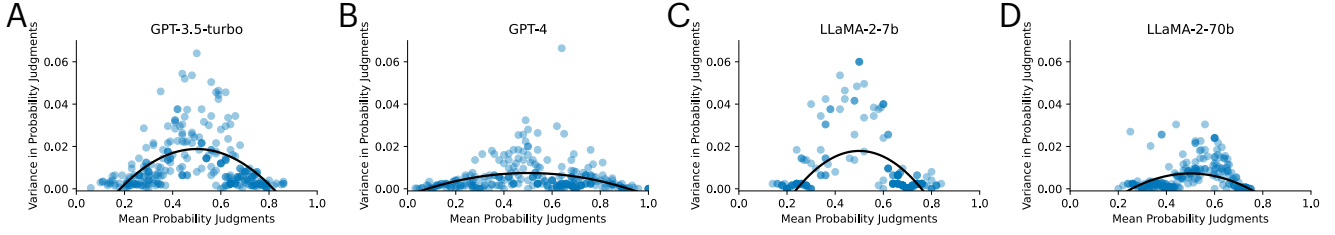
Figure 3: The relationship between mean and variance in repeated probability judgments produced by LLMs exhibits an inverted-U shape. Solid lines represent the best-fitting regression models. Variants with more parameters tend to shift the curve outward and downward. **(A)** GPT-3.5-turbo model. **(B)** GPT-4 model. **(C)** LLaMA-2 model with 7b parameters. **(D)** LLaMA-2 model with 70b parameters. The temperature parameters were fixed at 1.0 for all models.

judgments. This model suggests that an increase in the additive noise is likely whenever there is an imbalance between positive and negative terms.

In contrast, the Bayesian Sampler model postulates that these samples are integrated with a prior, $p(\theta)$, reflecting the prior knowledge of the underlying event probability (Zhu et al., 2020):

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \qquad (1)$$

where $p(x|\theta)$ represents an approximation of the underlying event probability. Furthermore, the mean probability judgment as predicted by the Bayesian Sampler model corresponds to the mean of the posterior distribution in Equation 1:

$$\mathbb{E}[\theta] = \int \theta p(\theta|x)d\theta = \frac{\int \theta p(x|\theta)p(\theta)d\theta}{\int p(x|\theta)p(\theta)d\theta} \qquad (2)$$

Thus, the interpretation offered by the Bayesian Sampler model is grounded in the repeated use of the prior, $p(\theta)$, in each instance of probability judgment. According to this model, the imbalance between positive and negative terms in an probabilistic identity serves to magnify the influence exerted by the prior.

To further illustrate the mechanisms of the Bayesian Sampler model, let's consider a symmetric Beta prior for $p(\theta)$, denoted as $\text{Beta}(\beta,\beta)$, and a binomial distribution for $p(x|\theta)$, denoted as $\text{Bin}(N,\theta)$. That is, the agent is assumed to draw N i.i.d.-samples from the underlying event probability θ. The posterior as predicted by the Bayesian Sampler model is in the form of $\text{Beta}(\beta+x,\beta+N-x)$ and $x \sim \text{Bin}(N,\theta)$, which has mean and variance:

$$\mathbb{E}[\theta] = \frac{N}{N+2\beta}\theta + \frac{\beta}{N+2\beta} \qquad (3)$$

$$\mathbb{V}[\theta] = \frac{N}{(N+2\beta)^2}\theta(1-\theta) \qquad (4)$$

Considering Z_3 as an example, the predicted mean values

for the identity can be articulated as follows:

$$\mathbb{E}[Z_3] = \mathbb{E}[\theta_A] + \mathbb{E}[\theta_{B\cap\neg A}] - \mathbb{E}[\theta_{A\cup B}] \qquad (5)$$

$$= \frac{N}{N+2\beta}(\theta_A + \theta_{B\cap\neg A} - \theta_{A\cup B}) + \frac{\beta}{N+2\beta} \qquad (6)$$

$$= \frac{\beta}{N+2\beta} \qquad (7)$$

Given that the underlying event probabilities θ are coherent, it follows that $\theta_A + \theta_{B\cap\neg A} - \theta_{A\cup B} = 0$. Therefore, it is the number of terms $\frac{\beta}{N+2\beta}$ that prompts the identities to deviate from zero, which is related to the imbalance of positive and negative terms within an identity.

In scenarios where the agent does not draw any samples for estimating the underlying event probability (namely, $N = 0$), and the prior distribution of this probability is symmetric, the expected deviation from zero is 0.5 for identities $Z_{3:6}$ and 1 for identities $Z_{7:8}$. These values can be considered as the maximal limits of incoherence within probabilistic identities. This analysis has two key implications: (1) An agent demonstrating incoherence levels behind these extremes is indicative of a more accurate approximation of the underlying event probabilities; (2) Conversely, any factor impeding the accurate approximation of the underlying event probability will result in identities exhibiting deviations that converge towards these bounds.

The Bayesian Sampler model also predicts the inverted-U-shaped relationship between the mean and variance of repeated probability judgments (Sundh et al., 2023). By rearranging the predicted mean and variance, we derive the following relationship for the aforementioned example:

$$\mathbb{V}[\theta] = \frac{1}{N}\mathbb{E}[\theta]\left(1 - \mathbb{E}[\theta]\right) - \frac{\beta(N+\beta)}{N(N+2\beta)^2} \qquad (8)$$

In this equation, the inverted-U shape is a result of the quadratic relationship between mean and variance. As a corollary, an increase in the sample size N (i.e., obtaining a more accurate approximation of the underlying event probability) is predicted to shift the curve downward. Furthermore, the shrinkage effect that shifts the curve inward is driven by

the incorporation of the Beta(β, β) prior. An increase in the value of β (i.e., a stronger prior) is expected to cause an inward shift of the mean-variance curve.

Connecting LLMs to Bayesian Inference

We now consider a possible explanation for the systematic deviations from rationality observed in LLMs. Our approach begins with a formalization of the autoregressive process employed by LLMs in the generation of probability judgments.

Let w_{n+1} be the model-predicted probability judgment and $w_{1:n}$ the query (e.g., "what is the probability that Greece has left the European Union by the year 2025?"). We note that the process for generating probability judgments in autoregressive LLMs can be understood as $p(w_{n+1}|w_{1:n})$. Under the exchangeablility assumption[1], we can rewrite the autoregressive objective as an implicit Bayesian inference process using de Finetti's theorem (Korshunova et al., 2018; Zhang et al., 2023):

$$p(w_{n+1}|w_{1:n}) = \int p(w_{n+1}|\theta)p(\theta|w_{1:n})d\theta \qquad (9)$$

where θ represents the latent generative process underlying the data or, in our study, the underlying probability of the event in the query. Equation 9 links the concept of learning to fit the autoregressive distribution (i.e., the left hand side) with conducting Bayesian inference on the latent generative process underlying the data (i.e., the right hand side) (Zhang et al., 2023). In our context, it is specifically the underlying event probabilities on which the Bayesian inference is executed.

This decomposition of the autoregressive distribution resembles the Bayesian inference process applied to samples in the Bayesian Sampler model:

$$w_{n+1} = \arg\max_{w_{n+1}} p(w_{n+1}|w_{1:n}) \qquad (10)$$

$$= \arg\max_{w_{n+1}} \int p(w_{n+1}|\theta)p(\theta|w_{1:n})d\theta \qquad (11)$$

$$= \arg\max_{\theta} p(\theta|w_{1:n}) \qquad (12)$$

$$\approx \int \theta p(\theta|w_{1:n})d\theta = \frac{\int \theta p(w_{1:n}|\theta)p(\theta)d\theta}{\int p(w_{1:n}|\theta)p(\theta)d\theta} \qquad (13)$$

where we assume in Equation 12 that $p(w_{n+1} = \theta|\theta) = 1$ (i.e., $p(w_{n+1} \neq \theta|\theta) = 0$) and in Equation 13 that the mode is approximated by the mean of the integral. By contrasting Equation 13 with 2, we postulate a possible link between the probability judgments produced by LLMs and the Bayesian Sampler model. In this framework, the prompt $w_{1:n}$ acts as the target for approximating the underlying event probability in LLMs.

---

[1]While this assumption might not be rigorously applicable at the token level in LLMs, it can nonetheless function as a valuable proxy within the semantic space to explain LLMs' responses (Xie et al., 2021; Zhang et al., 2023).

Discussion

We have presented empirical evidence demonstrating the incoherence of probability judgments generated by LLMs. Furthermore, through the analysis of probabilistic identities and mean-variance relationship, we identified shared patterns of incoherence in LLMs and humans. These structures offer insights into the underlying mechanisms employed by LLMs in the formation of probability judgments. We conjecture that this process originates from the implementation of autoregression for the four LLMs. Specifically, we postulated a possible connection between the autoregressive training objective and the Bayesian Sampler model, which has been previously employed to account for similar patterns of incoherence observed in human judgments (Sundh et al., 2023; Zhu et al., 2020, 2023).

Our results suggest a novel approach for enhancing the accuracy of probability outputs generated by AI systems. This improvement might be achieved not through calibration with true frequencies, but rather by adjusting the degree of incoherence in the output. The established relationship between coherence and accuracy in probability estimation, particularly in the context of agents with bounded rationality (Leitgeb & Pettigrew, 2010; Zhu et al., 2022), provides a theoretical foundation for this method. Applying this relationship to recalibrate incoherent judgments offers a promising avenue for future research. We see this approach as having the potential to help refine AI-based probability judgments, making the resulting models more reliable and effective in practical applications.

Our work also exemplifies the idea that Bayesian and neural network models are complementary approaches in understanding machine and human intelligence (Griffiths et al., 2023). The Bayesian model serves as a valuable tool for defining the computational-level objectives that a neural network model aims to achieve. This synergy between Bayesian and neural network models provides a more holistic understanding of intelligence, offering insights into the interplay between computational-level objectives and algorithmic-level execution in artificial and natural cognitive processes.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), e2218523120.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Chater, N., Zhu, J.-Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020). Probabilistic biases meet the Bayesian brain. *Current Directions in Psychological Science*, *29*(5), 506–512.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*, 451–482.

Griffiths, T. L., Zhu, J.-Q., Grant, E., & McCoy, R. T. (2023). Bayes in the age of intelligent machines. *arXiv preprint arXiv:2311.10206*.

Horton, J. J. (2023). *Large language models as simulated economic agents: What can we learn from homo silicus?* (Tech. rep.). National Bureau of Economic Research.

Jiang, X., Osl, M., Kim, J., & Ohno-Machado, L. (2012). Calibrating predictive model estimates to support personalized medicine. *Journal of the American Medical Informatics Association*, *19*(2), 263–274.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454.

Korshunova, I., Degrave, J., Huszár, F., Gal, Y., Gretton, A., & Dambre, J. (2018). Bruno: A deep recurrent model for exchangeable data. *Advances in Neural Information Processing Systems*, *31*.

Kumar, A., Liang, P. S., & Ma, T. (2019). Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, *32*.

Leitgeb, H., & Pettigrew, R. (2010). An objective justification of Bayesianism i: Measuring inaccuracy. *Philosophy of Science*, *77*(2), 201–235.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. (2022). Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Loomes, G., & Sugden, R. (1998). Testing different stochastic specificationsof risky choice. *Economica*, *65*(260), 581–598.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M., & Griffiths, T. L. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. *arXiv preprint arXiv:2309.13638*.

Miller, J. B., & Gelman, A. (2020). Laplace's theories of cognitive illusions, heuristics and biases. *Statistical Science*, *35*(2), 159–170.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy*, *59*(5), 371–404.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Shrivastava, V., Liang, P., & Kumar, A. (2023). Llamas know what gpts don't show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*.

Sundh, J., Zhu, J.-Q., Chater, N., & Sanborn, A. (2023). A unified explanation of variability and bias in human probability judgments: How computational noise explains the mean–variance signature. *Journal of Experimental Psychology. General*, 2842–2860.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing systems*, *30*.

Xie, S. M., Raghunathan, A., Liang, P., & Ma, T. (2021). An explanation of in-context learning as implicit Bayesian inference. *arXiv preprint arXiv:2111.02080*.

Zadrozny, B., & Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *Icml*, *1*, 609–616.

Zhang, L., McCoy, R. T., Sumers, T. R., Zhu, J.-Q., & Griffiths, T. L. (2023). Deep de Finetti: Recovering topic distributions from large language models. *arXiv preprint arXiv:2312.14226*.

Zhu, J.-Q., Newall, P. W., Sundh, J., Chater, N., & Sanborn, A. N. (2022). Clarifying the relationship between coherence and accuracy in probability judgments. *Cognition*, *223*, 105022.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The Bayesian sampler: Generic Bayesian inference causes incoherence in human probability judgments. *Psychological Review*, *127*(5), 719.

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated Bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*.