A Taxonomy for Understanding and Identifying Uncertainty in AI-Generated Responses

Snehal Prabhudesai, Daniel G. Goldstein, Jake M. Hofman, David M. Rothschild

May 7, 2024

Abstract

People search for information to meet their personal, business, and civic goals and increasingly do so with AI-based tools. We present a taxonomy of uncertainties in responses generated by Large Language Models (LLMs). It identifies three main types of uncertainty: outcome variability, model uncertainty, and prompt ambiguity. Outcome variability encompasses the unpredictability in data, world as aleatory uncertainty. Model uncertainty arises from insufficient knowledge available to the model, termed epistemic uncertainty. Prompt ambiguity involves unclear user inputs leading to multiple valid interpretations. The study explores detection methods for these uncertainties, employing strategies such as token probability analysis and temperature sampling. This taxonomy aims to enable researchers and regulators to identify, track, and remediate uncertainty from LLM-based tools that may bias or otherwise impair decision making.

Decision making has long been supported by various computational tools, ranging from general-purpose search engines (Broder, 2002) to more specialized algorithmic decision aids (e.g., Bell and Carcello, 2000; Hunt et al., 1998). Search engines, while broad in scope, serve primarily as indirect aids—they excel at helping users locate resources across a wide range of domains but until recently have fallen short of providing direct answers. In contrast, specialized algorithmic decision aids have a narrower focus but deliver direct assistance. These tools may furnish decision-makers with pertinent quantitative forecasts or recommendations, though the utility of each is limited to the narrow task it is designed to handle.

Recently, Large Language Models (LLMs) have raised the possibility of merging the expansive domain coverage of search engines with the direct-response capability of algorithmic decision aids. LLMs allow people to ask questions in natural language and provide detailed, direct answers across a wide array of topics (Jiang et al., 2021).However, despite these advantages, LLMs face significant challenges, primarily because direct answers to questions are often uncertain (Lu et al., 2022; Arora et al., 2022; Wachowiak and Gromann, 2023; Jiang et al., 2021). In this work, we aim to describe and address the various types of uncertainty that arise when people turn to conversational AI for answers.

This question of how to reason about and communicate uncertain information is far from new. Not every resource we consult, nor every person we communicate with is completely reliable. People have ways of coping with this everyday uncertainty, whether it is comparing information from multiple sources returned by a search engine, asking a person for clarification, updating beliefs (Griffiths et al., 2008), or reasoning about the quantitative uncertainty in a decision aid or scientific result (Zhang et al., 2023; Prabhudesai et al., 2023). Yet, LLMs pose a distinct challenge, as there are no well-established norms or practices for systematically evaluating or conveying the accuracy or certainty of the information they report. The lack of a common taxonomy makes it difficult to measure, track, and ultimately remediate concerns around uncertainty in AI-generated responses.

The goal of this work is to delineate the different types of uncertainty that can arise when users seek answers from LLM-based systems and to suggest how they can be detected and addressed. We propose three main sources of uncertainty: outcome variability, model uncertainty, and prompt ambiguity. Outcome variability and model uncertainty have analogs in the philosophy (Hacking, 2006), psychology (Ülkümen et al., 2016; Tannenbaum et al., 2017) and engineering (Der Kiureghian and Ditlevsen, 2009) literatures. Prompt ambiguity and model uncertainty have roots in the total survey error framework (Groves and Lyberg,

	Prompt Ambiguity	Model Uncertainty	Outcome Variability
Description	Ambiguity in the question that is asked	Uncertainty due to lack of knowledge or missing information	Inherent variability in the answer, even with full knowledge
Example prompt	"What is the average mpg of the 2020 Toyota Corolla?"	"What is the length of the 2024 Honda Accord?"	"How long would it take to drive from South Ferry to Columbus Circle in Manhattan right now?"
Remediation	Ask the user for clarification or produce an exhaustive list of possible answers	Acknowledge missing information, consult outside sources, or respond with an estimate and subjective probability	Communicate a range of calibrated responses
Example responses	"Could you be more specific" "Do you mean ___ or ___" "There are a few different answers"	"I don't have access to that" "Let me search the web" "Here's my best estimate"	"It could range from" "It's hard to predict, but"

Figure 1: A taxonomy for sources of uncertainty in LLM-based responses.

2010). These three types of uncertainty are most easily demonstrated by considering different examples of a user issuing a prompt to a model, a model determining the answer to the prompt, and the LLM-based tool reporting a version of the model's answer back to the user, as outlined in Figure 1.

Outcome Variability: Perhaps the most straightforward case occurs when a person asks for information that is fully known to a model, but where the answer varies due to inherent randomness or unpredictability, known in the literature as *aleatory uncertainty*. For instance, imagine that someone asks an LLM how long it would take to drive from South Ferry terminal to Columbus Circle in Manhattan. Even with knowledge of how far apart these two locations are, time of day, and approximate traffic, there is variability in the travel time. In such settings, a reliable decision aid should not respond with a single estimate (e.g., 20 minutes), but with a range that indicates how much it might vary (20-45 minutes). Ideally this would be a calibrated interval along with information about the range of outcomes covered by the interval.

Model Uncertainty: In addition to outcome variability, there can also be uncertainty due to a lack of knowledge on the part of the model, known as *epistemic uncertainty*. For instance, someone might ask an LLM that was trained in 2023 for details about the 2024 model of a specific car. Here an appropriate response might be for a decision aid to acknowledge that this is a "known unknown" due to lack of training data for 2024, or to return an estimate based on historical data that includes some confidence range or subjective probability, for example based on how the requested car model has evolved over the years. Importantly, in contrast to aleatory uncertainty, epistemic uncertainty can be reduced by acquiring more information—in this case by obtaining updated data on the latest vehicle models, perhaps by consulting outside sources via retrieval augmented generation. Besides examples where training data does not exist (Bender et al., 2021; Gabriel and Ghazavi, 2022), model uncertainty may arise when training data is blocked (e.g., as various content creators negotiate for access to their data) (Milmo, 2023; Bogle, 2023), or when there is sampling bias in the data that the model is trained on (Baeza-Yates, 2016).

Prompt Ambiguity: Both outcome variablility and model uncertainty assume that a specific question, with a perhaps unknowable or simply unknown answer, exists, at least in principle. However, in the context of LLMs there is the possibility that the question a user has asked does not even have a clear answer because it is ambiguous or ill-posed. This leads a third category: prompt ambiguity. An illustrative example of

prompt ambiguity occurs when a person requests the documented average fuel economy of a vehicle without specifying the context of city versus highway driving, which would yield significantly different answers. In instances of specification ambiguity, an ideal decision aid would recognize the ambiguous prompt and either present all plausible responses or solicit additional information from the user to refine the query. Prompt ambiguity can arise for several reasons, ranging from people simply minimizing effort, to having incomplete thoughts, to using LLM-based tools to learn about a topic while exploring it. (Marchionini, 2006).

Above, we discussed three types of uncertainty as if they occur in isolation. However, it is quite likely that they occur in combination. Consider the scenario where a user asks for travel time from New York to London by train. This initial query introduces prompt ambiguity: Does the user mean New York to London, Ontario, by train? Or do they mean New York to London, UK, mistakenly typing train instead of plane? Ideally, the system should ask the user to clarify. Suppose the user specifies London, UK by plane. The model might then face model uncertainty because it hasn't been trained on flight times. In response, it could estimate travel time based on the approximate distance between the two cities and average plane speeds, thus communicating this uncertainty. Lastly, the model should account for outcome variability by providing a range of possible flight times, reflecting factors like wind speeds and airport delays.

Having described these three different sources of uncertainty, each of which requires its own unique remediation, we are faced with the question of how each source might be detected and appropriately communicated to users, and recorded and tracked by researchers. To investigate this, we designed different scenarios that isolate each of the three sources of uncertainty and explored various methods suggested in the literature for detecting uncertainty in an LLM response to them. Specifically, motivated by prior work on consumer decision making with LLM-based search (Spatharioti et al., 2023), we constructed four reusable templates that simulate a consumer researching the specifications of various vehicle and driving scenarios.

The scenarios are as follows. For prompt ambiguity, the scenario asks about the cargo space for an existing SUV without clarifying whether the rear seats are up or down, creating ambiguity in the question. For model uncertainty, the scenario queries the cargo space of a fictional vehicle, emphasizing the lack of available information. For outcome variability, the scenario involves estimating delivery times between two locations, which naturally varies due to unpredictable external factors. Lastly, an uncertainty-free scenario asks for the length of a vehicle that is well-documented within the LLM's training data.

We programatically provided prompts for each of these scenarios to GPT-3.5 and used four different methods based on prior literature (Vasconcelos et al., 2023; Lee et al., 2022; Chung et al., 2022; Ziems et al., 2023) to explore whether the amount and type of uncertainty was discernible from each.[1]

Token Probability: First, we explore simply making use of the internal token probability that the LLM assigned to each response it gave, information that is accessible through the OpenAI API. Token probabilities correspond to the probability that the LLM assigns to the key word (in our case, a number) in the response.[2] The intuition behind this method is that tokens generated with higher probabilities are more likely to be correct.

Multiple Guesses: Second, we prompt the LLM to generate multiple (25) responses (with replacement). This requires just one API call, but result in a longer and thus costlier response compared to recovering the token probability. The intuition here is that the distribution of the 25 values may demonstrate the model's uncertainty about the response.

Top 5 Token Probabilities: Third, we recover the probabilities associated with each of the top 5 most likely tokens according to the model.[3] Similar to the multiple guess strategy, the intuition behind this approach is that the distribution of token probabilities contains information about the uncertainty of the response. Narrow, peaked distributions might indicate high confidence, broader ones might indicate outcome variability or model uncertainty, and multi-modal distributions might reasonably indicate prompt ambiguity.

[1] Unless otherwise noted, experiments were done at a zero temperature, so that each query resulted in a fixed response

[2] For decimal numbers, we use only the probability for the token corresponding to the part of the number before the decimal point (e.g., for "57.3", we use the token probability associated with "57")

[3] At the time of writing, these token probabilities were available via APIs for GPT-3.5, but not for GPT-4, which was the latest model available

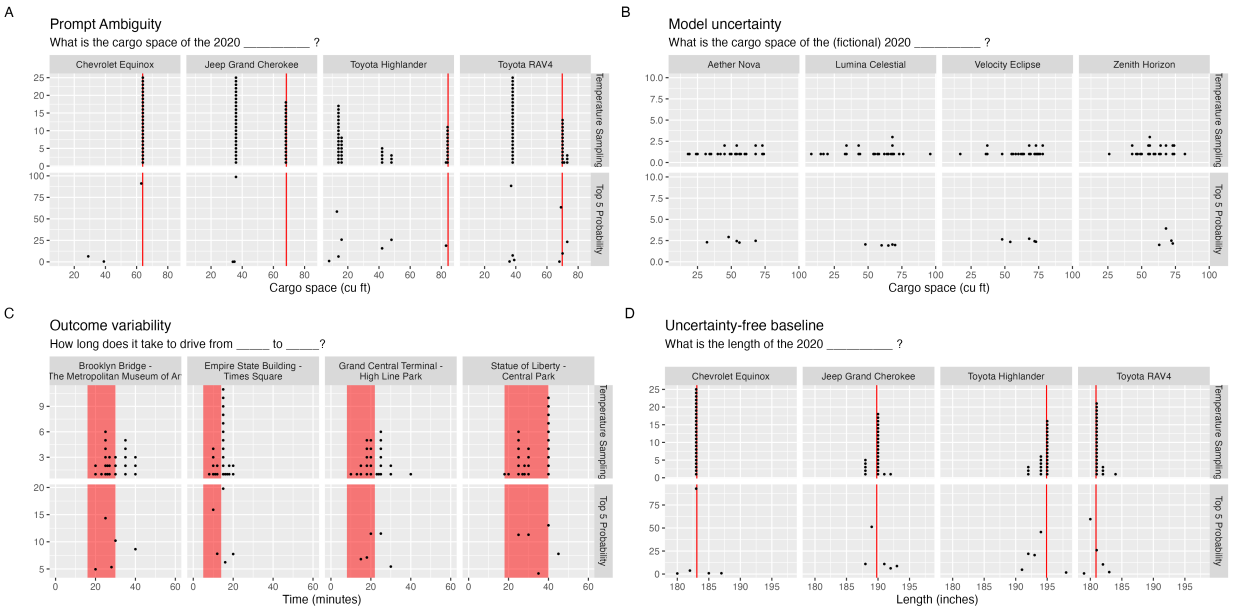

Figure 2: Four examples of all four scenarios (prompt ambiguity, model uncertainty, outcome variability, and uncertainty-free) for Generating Multiple Guesses and Temperature Sampling.

Temperature Sampling: Finally, instead of using a zero temperature response, we raised the temperature parameter and sampled an individual response many (25) times. The idea behind this approach is similar to the top 5 token probability method, but uses the model to directly sample values that explore more of the distribution than just the top 5 most likely responses. The downside of this approach, however, is that it is much more expensive, requiring 25 calls to the model instead of just one.

The approaches vary in their ability to represent different types of uncertainty. The token probability method, for instance, showed limited promise. This is partly because token probabilities often aren't well-calibrated: a correct response might have a low probability, while an incorrect one might appear highly probable. This miscalibration could stem from the use of preference-based reinforcement learning to align models (Achiam et al., 2023). Another limitation is that, even with well-calibrated generation probabilities, the token probability method cannot differentiate between aleatory, epistemic, and prompt-based uncertainty.

The second approach, generating multiple guesses, turned out not to discriminate well. The prompt ambiguity scenario produced a mixture of peaked, bell-shaped, and uniform distributions (but no multi-model distributions). The model uncertainty scenario produced a mix of uniform and normal distributions. The outcome variability scenario produced uniform distributions.The uncertainty-free scenario comprised a few peaked distributions, but also several relatively uniform ones.

The approach of using the top 5 token probabilities, however, did provide useful information about what type of uncertainty may have existed. Examples from the top-5 and temperature sampling approaches are shown in Figure 2. In the prompt ambiguity scenario, the model often seemed to detect the ambiguity, providing multiple answers approximately half of the time and showing two prominent peaks in probability. When the prompt was less ambiguous, it typically resulted in a single, distinct peak. In the model uncertainty scenario, the output was characterized by a low, uniform distribution. This spread indicated a high degree of uncertainty, as the model displayed little confidence in any single response. For the output variability scenario, the results comprised a blend of distributions that were roughly normal and centered around the expected answer range. This pattern suggested that while the model generally identified the correct area, its responses varied significantly. In the uncertainty-free scenario, the probability distribution was sharply peaked around the correct answer, indicating high confidence and little to no uncertainty in the model's response.

Temperature sampling was comparable to the top 5 token probability approach but tended to produce multiple peaks more consistently, especially with ambiguous prompts. It naturally extended beyond the top 5 tokens, capturing a longer tail of possibilities and including sample error, which introduced variability not present in the deterministic top 5 method. Unlike the top 5 approach, which might not always be available, temperature sampling offered a reliably accessible alternative.

Methods for identifying states of uncertainty are not limited to the four types explored in our experiment. Building on the insights gained from our taxonomy and empirical research, we can develop pre-prompts that explicitly instruct LLM-based tools to identify potential issues across a series of prompts. Furthermore, uncertainty does not need to be discrete as it was in our simulations; often a blend of types of uncertainties will be present.

We anticipate a continuous co-evolution of LLM-based tools and their users, which will alter the types of uncertainty encountered, the methods used for identifying them, and the optimal strategies for their remediation. During the writing of this paper, both the tools and the underlying LLMs evolved significantly, reducing the occurrence of certain issues such as the brittleness of models—which had previously produced different responses to very similar prompts. At the same time, users have become more skilled at crafting prompts, likely gaining a better understanding of which outputs to trust.

Conversational LLMs, with their ability to provide direct answers to a broad range of questions, merge the features of search engines and decision aids. This integration holds considerable promise for enhancing human decision-making. However, significant improvements are still needed. For LLMs to be truly effective, users must understand how reliable these tools are, which necessitates better methods for assessing uncertainty, suggesting remediations, and effectively communicating uncertainty back to users. We hope that the classification of uncertainty types detailed in this essay will aid in creating such improved methods.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. arXiv:2210.02441 [cs.CL]

Ricardo Baeza-Yates. 2016. Data and Algorithmic Bias in the Web. In *Proceedings of the 8th ACM Conference on Web Science* (Hannover, Germany) *(WebSci '16)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/2908131.2908135

Timothy B Bell and Joseph V Carcello. 2000. A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory* 19, 1 (2000), 169–184.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Ariel Bogle. 2023. *New York Times, CNN and Australia's ABC block OpenAI's GPTBot web crawler from accessing content.* https://www.theguardian.com/technology/2023/aug/25/new-york-times-cnn-and-abc-block-openais-gptbot-web-crawler-from-scraping-content Published on August 24, 2023.

Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.

John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA,

USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 209, 19 pages. https://doi.org/10.1145/3491102.3501819

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.

Iason Gabriel and Vafa Ghazavi. 2022. The Challenge of Value Alignment: From Fairer Algorithms to AI Safety. In *The Oxford Handbook of Digital Ethics*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.18

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition. (2008).

Robert M. Groves and Lars Lyberg. 2010. Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly* 74, 5 (01 2010), 849–879. https://doi.org/10.1093/poq/nfq065 arXiv:https://academic.oup.com/poq/article-pdf/74/5/849/5144458/nfq065.pdf

Ian Hacking. 2006. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Dereck L Hunt, R Brian Haynes, Steven E Hanna, and Kristina Smith. 1998. Effects of computer-based clinical decision support systems on physician performance and patient outcomes: a systematic review. *Jama* 280, 15 (1998), 1339–1346.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics* 9 (2021), 962–977.

Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. https://doi.org/10.1145/3491102.3502030

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8086–8098. https://doi.org/10.18653/v1/2022.acl-long.556

Gary Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (apr 2006), 41–46. https://doi.org/10.1145/1121949.1121979

Dan Milmo. 2023. *The Guardian blocks ChatGPT owner OpenAI from trawling its content*. https://www.theguardian.com/technology/2023/sep/01/the-guardian-blocks-chatgpt-owner-openai-from-trawling-its-content Published on September 1, 2023.

Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing Traditional and LLM-based Search for Consumer Choice: A Randomized Experiment. *ArXiv* abs/2307.03744 (2023).

David Tannenbaum, Craig R Fox, and Gülden Ülkümen. 2017. Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Science* 63, 2 (2017), 497–518.

Gülden Ülkümen, Craig R Fox, and Bertram F Malle. 2016. Two dimensions of subjective uncertainty: Clues from natural language. *Journal of experimental psychology: General* 145, 10 (2016), 1280.

Helena Vasconcelos, Gagan Bansal, Adam Fourney, Q. Vera Liao, and Jennifer Wortman Vaughan. 2023. Generation Probabilities Are Not Enough: Exploring the Effectiveness of Uncertainty Highlighting in AI-Powered Code Completions. http://arxiv.org/abs/2302.07248 arXiv:2302.07248 [cs].

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 Grasp Metaphors? Identifying Metaphor Mappings with Generative Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 1018–1032. https://aclanthology.org/2023.acl-long.58

Sam Zhang, Patrick R Heck, Michelle N Meyer, Christopher F Chabris, Daniel G Goldstein, and Jake M Hofman. 2023. An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proceedings of the National Academy of Sciences* 120, 33 (2023), e2302491120.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023. Large Language Models are Built-in Autoregressive Search Engines. arXiv:2305.09612 [cs.CL]