ARL-TR-10041 ● DEC 2024

DEVCOM ARMY RESEARCH LABORATORY

Developing a Framework to Evaluate Credibility Tracking in Large Language Models

by Avvai Chandrasekaran, Erin Zaroukian, Justine Rawal, Mark Mittrick, and Adrienne Raglin

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10041 ● DEC 2024

Developing a Framework to Evaluate Credibility Tracking in Large Language Models

Avvai Chandrasekaran
Georgia State University

Erin Zaroukian, Justine Rawal, Mark Mittrick, and Adrienne Raglin
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
December 2024	Technical Report	2 June 2024	8 August 2024

4. TITLE AND SUBTITLE
Developing a Framework to Evaluate Credibility Tracking in Large Language Models

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Avvai Chandrasekaran, Erin Zaroukian, Justine Rawal, Mark Mittrick, and Adrienne Raglin

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-IC Aberdeen Proving Ground, MD 21005	ARL-TR-10041

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCIDs: Avvai Chandrasekaran, 0009-0006-3550-5561; Erin Zaroukian, 0000-0002-1381-085X; Justine Rawal, 0000-0002-5232-1135; Adrienne Raglin, 0000-0003-2147-8938

14. ABSTRACT
Large language models (LLMs) are increasingly being used in human problem-solving tasks such as standardized tests, debugging code, and providing customer support. However, little is known about the performance of these models when reasoning over longitudinal data and in detecting changes concerning the credibility of an information source over time. Recent work has explored this gap by asking LLMs to make predictions based on different patterns of historical data, but it was limited in the prompt variations used. Because LLMs are known to be very sensitive to seemingly minor changes in their prompts, the work presented in this report addresses these shortcomings by employing a variety of modifications to how the model is asked to make a prediction based on the historical data and by changing the historical data itself. A method was developed to evaluate the results of these modifications, finding that the LLM's ability to reason over longitudinal data and assess changes in credibility is improved by including the linguistic hedges "likely" or "probably" in the template for the model's prediction, likely because this helps avoid hyperconservatism. While other manipulations inhibited performance, adding "most likely" was the strongest inhibitor, and we hypothesize that adding the superlative "most" exacerbated hyperconservatism. Future work will explore further manipulations and improvements in methodology.

15. SUBJECT TERMS
large language models, explainable AI, machine psychology, decision-making, prompt engineering, Military Information Sciences

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	28

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Erin Zaroukian	(410) 278-3203

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

The ability to detect change in the environment is important for both humans and AI. While humans can be impressively blind to change and struggle to notice large visual changes in a scene if their attention is not properly aligned (Resnik et al. 1997), they can also be quite sensitive to it—e.g., catastrophically losing established trust in otherwise-reliable automation as soon as a failure is perceived (Zafari et al. 2024). Work in decision-making has shown that humans generally have an ability to track subtle changes in the credibility of an advisor over time (Diaconescu 2014), and while large language models' (LLMs') recent history of success with human problem-solving tasks (e.g., Kosinski 2023) suggests they may have a similar ability, Zaroukian (2024) is among few research projects that have explored LLMs' reasoning capabilities over longitudinal data.

Zaroukian (2024) compared the performance of two LLMs (i.e., BigScience Large Open-science Open-access Multilingual Language Model [BLOOM] and Generative Pretrained Transformer [GPT] 3.5) that were provided 30 days of fictional historic weather information from a fictional weatherman. Each day, the weatherman predicted either sunny or rainy weather, and each prediction was then labeled as correct or incorrect. Following this data, the LLM was prompted to generate information about day 31: "Finish this sentence: On day 31, the weatherman said it would be. . .," as shown in Figure 1.

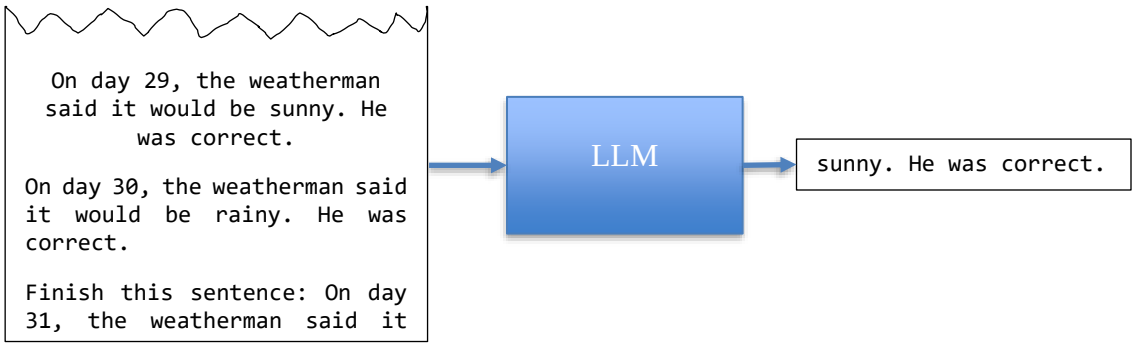

Figure 1. Example of an excerpted LLM input/output from Zaroukian (2024).

The 30 days of data followed one of six patterns (Table 1). The data was uniform, conditional, or probabilistic. "Uniform" indicates that the weather was sunny every day, and the weatherman made a correct prediction every day; "conditional" indicates that the weather was sunny every day, and the weatherman made a correct prediction 50% of the days; and "probabilistic" indicates that the weather was sunny 50% of the days, and the weatherman made a correct prediction 67% of the days. Additionally, these patterns were either "consistent," holding across all

30 days (i.e., the weatherman's credibility remained stable), or they were "inconsistent," changing on day 16 (i.e., the weatherman's credibility changed), as shown in Table 1. Specifically, for "uniform inconsistent," the weather remained sunny every day, but the weatherman switched to making an incorrect prediction every day; for "conditional inconsistent," the weather remained rainy every day, and the weatherman continued to make a correct prediction 50% of the days; and for "probabilistic inconsistent," the weather continued to be sunny 50% of the days, but the weatherman switched to making correct predictions only 33% of the days. As shown in Table 1, the change in the inconsistent patterns resulted in a lowering of the weatherman's accuracy in the uniform and probabilistic conditions.

Table 1. Summary of data presented to LLMs as percentage of sunny days and percentage of correct predictions in each condition. Adapted from Zaroukian (2024).

Type	Consistency	Days	Actual weather	Predictions
Uniform	Consistent	1–30	100% sunny	100% correct
	Inconsistent	1–15	Same as consistent	Same as consistent
		16–30	100% sunny	0% correct
Conditional	Consistent	1–30	100% sunny	50% correct
	Inconsistent	1–15	Same as consistent	Same as consistent
		16–30	0% sunny	50% correct
Probabilistic	Consistent	1–30	50% sunny	67% correct
	Inconsistent	1–15	Same as consistent	Same as consistent
		16–30	50% sunny	33% correct

To evaluate an LLM's performance, its responses were compared to human-like responses, where it was assumed that a human given the same task would continue the most recent pattern (days 16–30). The results show these LLMs were able to make human-like responses for day 31 for the simpler patterns (particularly uniform consistent) but the responses were less intuitive for more complex patterns. While this sheds some light on LLMs' ability to track a source's reliability across longitudinal data, these results are limited in a variety of ways, such as including minimal variation in how the LLM is prompted to complete the sentence about day 31. Previous work has shown that even small variations in prompts can significantly affect model outputs, and so systematic manipulation of prompts is recommended to get a clearer understanding of an LLM's behavior (e.g., Hagendorff et al. 2024; Sclar et al. 2023). Therefore, the current study provides a battery of prompt manipulations to more rigorously determine the extent to which LLMs can detect changes in the credibility of an information source over time and what factors have the greatest effect.[*]

[*] The code and data used here are available at https://osf.io/kqvpe/

2. Methodology

A total of nine manipulations were performed on the inputs from Zaroukian (2024), as summarized in Table 2. The first family of manipulations, "likelihood," primarily changes how the LLM is asked to make a prediction about day 31. The first way this is done is not by asking about what the weatherman said, but about what he likely said. Four variations of this were tested, asking the LLM to complete a sentence about what the weatherman "likely said," "most likely said," "most likely will say," and "probably said." These manipulations aim to reduce hyperconservatism, which is an LLM's tendency to avoid committing to a singular answer, even when it can generate the correct answer. This method of asking for the likely answer has been shown to improve LLM performance of theory-of-mind tasks (Strachan et al. 2024). For the second manipulation in this family, "multiple choice," the LLM was not asked to complete a sentence about day 31, but rather to select a completion from a list of options. This was done both using a set order of four options, as well as by randomly shuffling the order of the options in each prompt. Other work with LLMs has found that large enough models tend to be well-calibrated (i.e., the model's confidence estimates closely match the probability of its output being correct) on well-formatted, multiple-choice questions (Kadavath et al. 2022). Additionally, including multiple choices has been suggested to improve reasoning, particularly if answer choices are shuffled to avoid a position bias, a tendency to prefer options in a certain position (e.g., last) (Hagendorff et al. 2023).[*] For the third manipulation in this family, "generalization," the LLM is asked to predict the winner of a baseball game. This manipulation allows us to see whether the LLM is willing to extend the weatherman's credibility into a new domain.

[*] Note that additional challenges with multiple-choice questions include "token bias," a bias toward a particular label (e.g., "A" when options are labeled [Zheng et al. 2023]), and model performance has also been seen to decrease overall when (unlabeled) options are provided (Srivastava et al. 2023).

Table 2. Summary of manipulations, with the LLM's input broken down into three parts (introduction, data, and question).

Manipulation	Prompt		
	Introduction	**Data**	**Question**
Likelihood: "likely"	None	Unchanged	"Finish this sentence: On Day 31, the weatherman likely said it would be"
Likelihood: "most likely"	None	Unchanged	"Finish this sentence: On Day 31, the weatherman most likely said it would be"
Likelihood: "most likely" (future)	None	Unchanged	"Finish this sentence: On Day 31, the weatherman most likely will say it will be"
Likelihood: "probably"	None	Unchanged	"Finish this sentence: On Day 31, the weatherman probably said it would be"
Multiple choice: single order	"The following is a history of a weatherman's predictions:"	Unchanged	"Using the information above, choose one of the following: 'On day 31, the weatherman said it would be sunny. He was correct.,' 'On day 31, the weatherman said it would be sunny. He was incorrect.,' 'On day 31, the weatherman said it would be rainy. He was correct.,' 'On day 31, the weatherman said it would be rainy. He was incorrect.'"
Multiple choice: shuffled	"The following is a history of a weatherman's predictions:"	Unchanged	Same as multiple choice: fixed order, but order of choices is randomized
Generalization	"The following is the history of a weatherman's predictions:"	Unchanged	"On Day 31, the weatherman said that the winner of a baseball game would be Team A. He was"
Inversion	None	In uniform and conditional, all instances of "sunny" and "rainy" are swapped. In probabilistic, "correct" and "incorrect" probabilities are switched (33% correct for consistent and for inconsistent days 1–15, 67% correct for inconsistent days 16–30).	Unchanged
Comparison	None	Add a second weatherman who is always wrong, "The following is a history of predictions for Weatherman A and Weatherman B: On day X, Weatherman A and B said it would be sunny/rainy and sunny/rainy, respectively. They were correct/incorrect and incorrect, respectively."	"On day 31, Weatherman A and B said it would be"

The second family of manipulations primarily changes the historical data the LLM is given to reason over. The first manipulation in this family, "inversion," inverts selected terms in the input: in the uniform and conditional conditions, all instances of "sunny" and "rainy" are swapped (e.g., in the uniform consistent condition, every day is rainy and the weatherman is always correct); and in the probabilistic condition, all instances for "correct" and "incorrect" are swapped, meaning that the weatherman is now 33% correct in the consistent condition and for inconsistent days 1–15, and he is 67% correct for inconsistent days 16–30. This manipulation explores how sensitive the LLM is to the specific vocabulary used in the prompt and was inspired by the disproportionate number of sunny predictions the models provided in Zaroukian (2024). For the second manipulation in this family, "comparison," an additional weatherman is added to the data whose predictions are always wrong, and the LLM is asked to complete a sentence about each weatherman's prediction for day 31. This explores how well the LLM tracks multiple sources of information, beginning with a very simple case where one source is always incorrect. Additionally, the contrast between the two weathermen may help the LLM identify their respective patterns.

For each manipulation, as in Zaroukian (2024), the historical data was generated 20 times for each condition and each string of Introduction + Data + Question was provided to the LLM BLOOM via the HuggingFace application programming interface (BigScience Workshop, 2022). The LLMs' outputs were compared to the most human-like or "intuitive" answers, assuming that humans would continue the most recent pattern (days 16–30) in the input.

3. Evaluation

Model outputs were evaluated by comparing the proportion of correct predictions in the output with the input under the assumption that humans would provide correct predictions at the same proportion as they say in their recent input. The percentage of correct predictions in the input for each condition can be seen in Tables 1 and 3. For each condition, the absolute value of the difference between the percentage of correct predictions in the input and output across all 20 trials is computed. To correct for the fact that different conditions have different maximum possible differences (see Table 3), the absolute difference values are normalized to the range of 0–1 using min–max normalization. Example computations using the BLOOM results from Zaroukian (2024) are given in Table 3.

Table 3. Summary of values for computing ANAD score using BLOOM results of Zaroukian (2024).

Condition	Correct in input (days 16–30; %)	Max possible difference	Correct in output from Zaroukian (2024) BLOOM (%)	Absolute difference (|input–output|)	NAD (|input–output|/max)
Uniform consistent	100	100 (if output is 0%)	100	|100–100| = 0	0/100 = 0
Uniform inconsistent	0	100 (if output is 100%)	0	|0–0| = 0	0/100 = 0
Conditional consistent	50	50 (if output is 0% or 100%)	85	|50–85| = 35	35/50 = 0.7
Conditional inconsistent	50	50 (if output is 0% or 100%)	10	|50–10| = 40	40/50 = 0.8
Probabilistic consistent	67	67 (if output is 0%)	35	|67–35| = 32	32/67 = 0.48
Probabilistic Inconsistent	33	67 (if output is 100%)	20	|33–20| = 13	13/67 = 0.19

These normalized absolute differences (NADs) are then averaged across all conditions to provide a score representing how much the LLM struggled with tracking changes in the credibility, referred to as an average normalized absolute difference (ANAD) score. The lower this ANAD score, the better the LLM performed.

The results from Zaroukian (2024) using BLOOM are shown in Figure 2. Figure 3 is a chart of the NADs calculated from these results, with an ANAD of 0.36. This ANAD will be the baseline score used as a classification threshold for determining whether a manipulation improves the output of this LLM relative to Zaroukian (2024). If the ANAD resulting from the output of a manipulation is lower than the baseline, it will be classified as a "facilitator;" if it is an increase from this original score, then it is an "inhibitor" of the LLM's human-like reasoning skills.

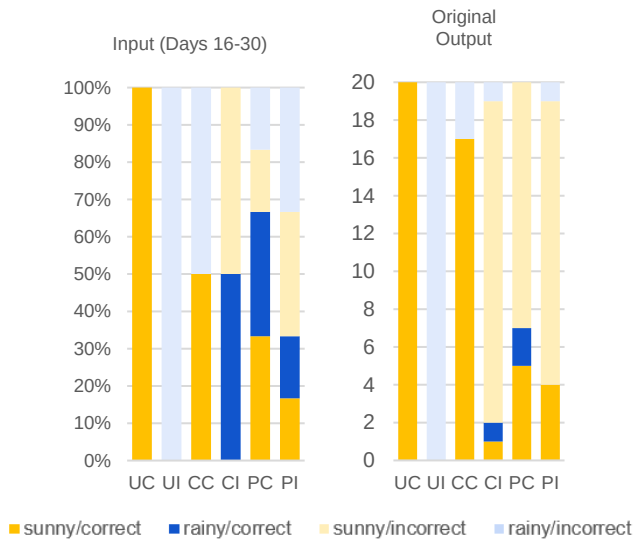

Figure 2. BLOOM input percentages and output histogram from data in Zaroukian (2024). Here and below, "sunny/correct" represents LLM continuations of "sunny. He was correct," "rainy/correct" represents LLM continuations of "rainy. He was correct," "sunny/incorrect" represents LLM continuations of "sunny. He was incorrect," and "rainy/incorrect" represents LLM continuations of "rainy. He was incorrect." UC = uniform consistent, UI = uniform inconsistent, CC = conditional consistent, CI = conditional inconsistent, PC = probabilistic consistent, and PI = probabilistic inconsistent.

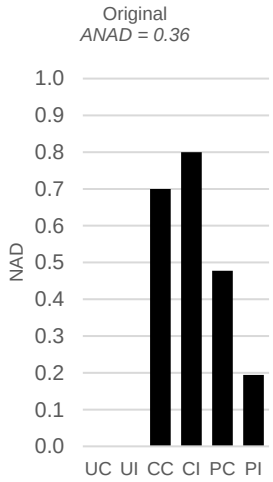

Figure 3. NAD and ANAD scores calculated from data in Zaroukian (2024).

4. Results

4.1 Replication

Results of a replication of Zaroukian (2024), with no modification to the prompts, is presented in Figures 4 and 5. These are similar to the original results shown in Figures 2 and 3, although while the original study found a boosting of sunny responses (compare the yellow shift from input to output in Figure 2), the replication appears to have even stronger boosting of these responses and correspondingly generally higher NAD and ANAD scores. Note that the LLM* and its inputs are identical here and in Zaroukian (2024), highlighting how relatively small the sample size of 20 is.

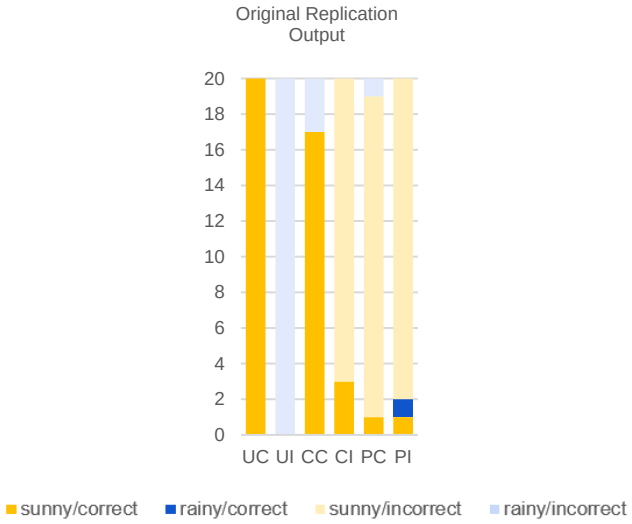

Figure 4. Histogram of LLM responses using the original prompts from Zaroukian (2024).

* BLOOM version 1.3; dated 6 July 2022.

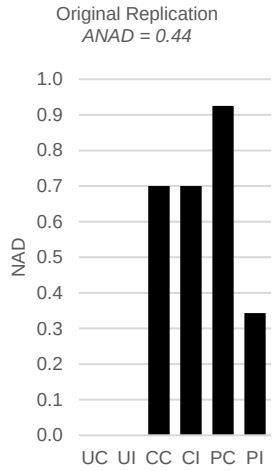

Figure 5. NAD and ANAD scores from the replication of Zaroukian (2024) in Figure 4.

4.2 Likelihood

Likelihood manipulations were included with the aim of reducing hyperconservatism, but results were mixed (Figures 6 and 7).

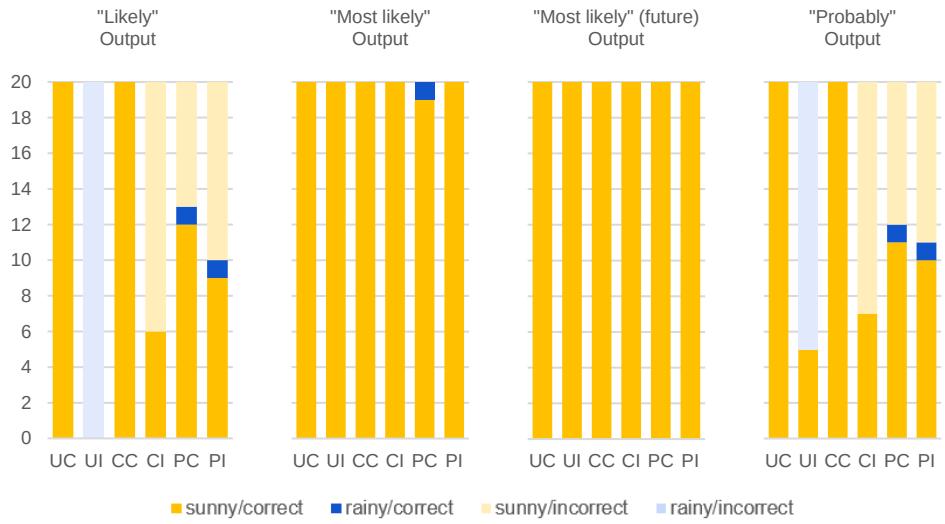

Figure 6. Histogram of LLM responses for likelihood manipulations.

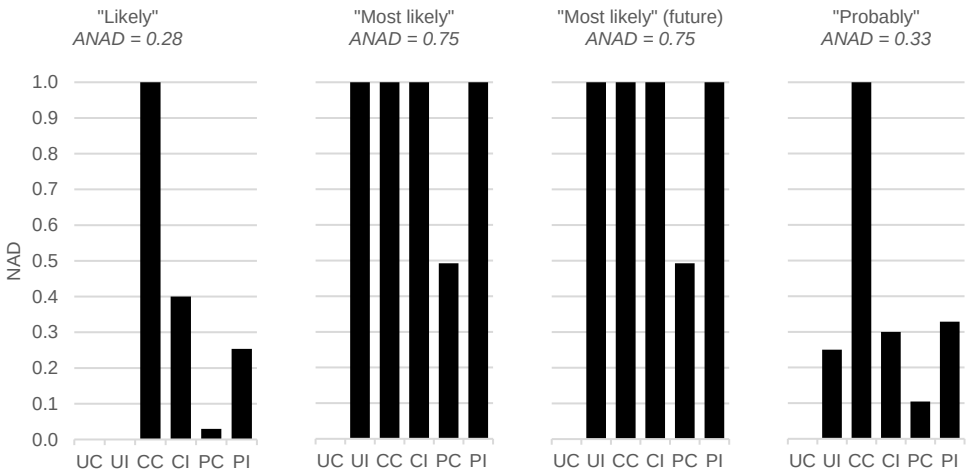

Figure 7. NAD and ANAD scores for likelihood manipulations.

Manipulations for "most likely" and "most likely" (future) result in all correct responses and overwhelmingly sunny/correct responses. This occurs even in the uniform inconsistent condition, where the last 15 days of the input are rainy/correct (compare to the input shown in Figure 2). This results in high ANAD scores of 0.75 (>0.36), indicating that these manipulations were inhibitors. "Likely" and "probably" manipulations have a more varied output distribution and lower ANAD scores (0.28 and 0.33 < 0.36), which classifies them as facilitators although they continue to boost the sunny/correct responses compared with both the input and original output (Figure 2). This may be evidence of the LLM overcoming its hyperconservatism, as it does reflect findings that a hedging belief likelihood improves an LLM's output (Strachan et al. 2024). However, the inclusion of the superlative "most" in the "most likely" manipulations may have been interpreted by the LLM as a strengthening of belief, exacerbating hyperconservatism.

4.3 Multiple Choice

Multiple-choice manipulations were included because LLMs often show good performance when prompts are formatted as multiple-choice questions, though results here emphasize that this is not a straightforward path to success (Figures 8 and 9).

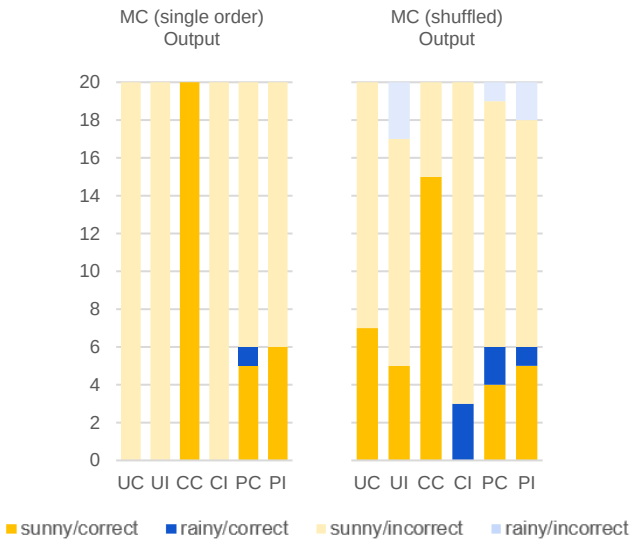

Figure 8. Histogram of LLM responses for multiple-choice manipulations.

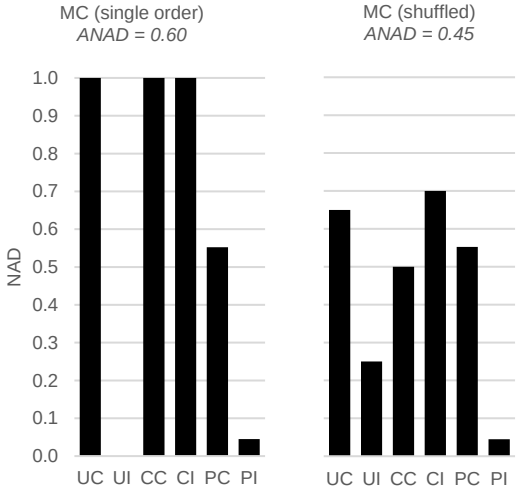

Figure 9. NAD and ANAD scores for multiple-choice manipulations.

Again, the outputs heavily favor sunny responses, and sunny/incorrect (the second option in the single-order manipulation) is the most frequent selection overall, followed by sunny/correct (the first option in the single-order manipulation). Shuffling answer choices appears to somewhat mitigate this, but both multiple choice manipulations act as inhibitors with ANAD scores of 0.60 and 0.45 (<0.36), respectively. It may be that offering explicit options that include "sunny" exacerbates the model's bias toward "sunny" responses.

4.4 Inversion

The inversion manipulation serves as a form of control and tested whether seemingly minor modifications to the original study would affect the output. Here, the instances of "sunny" and "rainy" were swapped in the uniform and conditional conditions, and the instances of "correct" and "incorrect" were swapped in the probabilistic conditions (Figures 10 and 11).

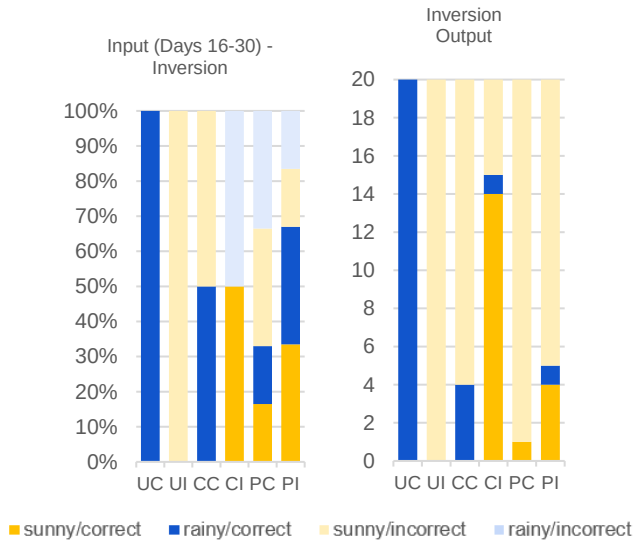

Figure 10. LLM input percentages and output histogram for the inversion manipulation.

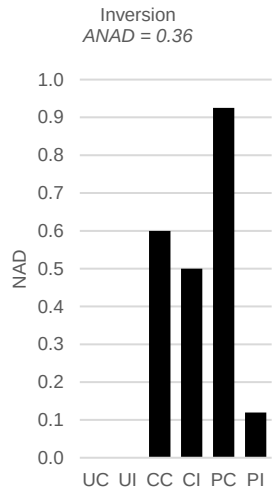

Figure 11. NAD and ANAD scores for the inversion manipulation.

Indeed, this manipulation shows a pattern of results very similar to what was seen with the original prompts where the uniform conditions matched the input perfectly (Figures 2–5), while the others showed a boosted proportion of sunny responses on par with what was seen in the original study (overall the responses in the original

study were 76% sunny, and the inversion manipulation responses were 78% sunny; Zaroukian 2024).

4.5 Generalization

The generalization manipulation probes the LLM's willingness to generalize a source's credibility from one domain (weather) to another (baseball) (Figures 12 and 13).

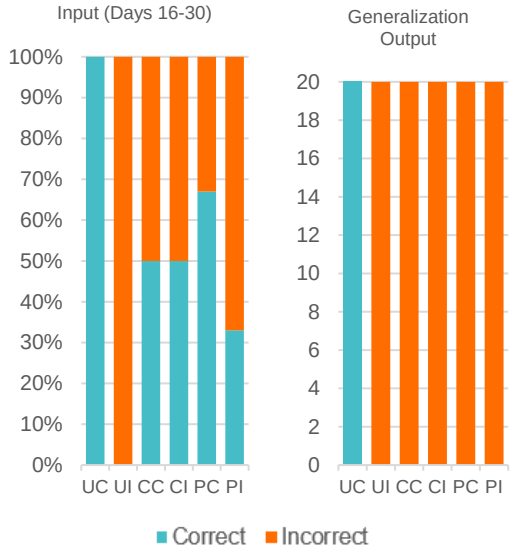

Figure 12. LLM input percentages and output histogram for the generalization manipulation. Responses are coded simply as correct or incorrect.

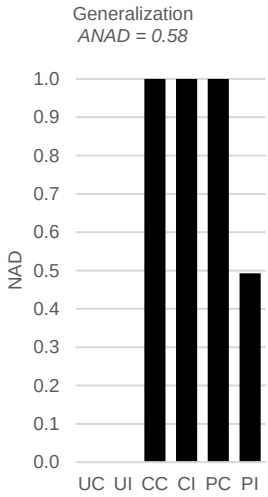

Figure 13. NAD and ANAD scores for the generalization manipulation.

The LLM appeared to find the source credible on the new topic in the uniform consistent condition (i.e., when the source was always credible in the old topic).

However, in all other conditions (conditions where the source was ever incorrect), the source was considered entirely incorrect in the new topic. This is interesting, as the source was not predicted to be at chance for the new topic, but was reliably incorrect, which seems like a very unhuman-like assessment.

4.6 Comparison

For the comparison manipulation, the LLM was presented with predictions from an additional weatherman (i.e., Weatherman B) who was always incorrect, every day, in every condition. This was done to determine whether the LLM could track two separate information sources when one was entirely predictable in its accuracy and generally exists in strong contrast to the other source (Figures 14 and 15).

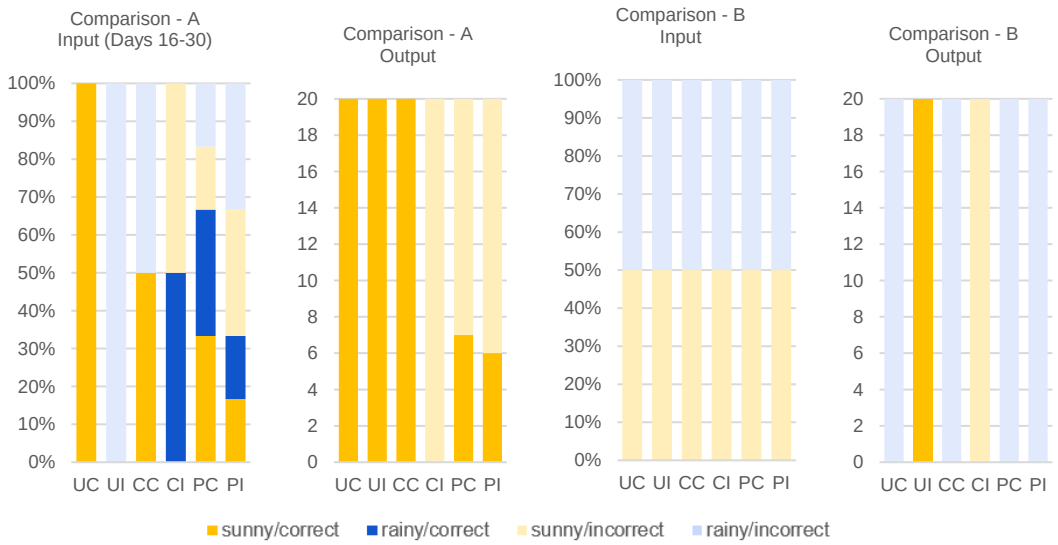

Figure 14. LLM input percentages and output histograms for Weathermen A and B for the comparison manipulation (the input for Weatherman A is the same as that in Figure 2).

The presence of a second information source seems to decrease the accuracy of the LLM for Weatherman A, which has an ANAD of 0.59 (>0.36). The responses for Weatherman A are entirely "sunny," and the proportion of correct and incorrect responses seem to have little relation to the input. The LLM performs better for Weatherman B's very simple pattern, with an ANAD of 0.17 (the NAD and ANAD scores for Weatherman B were computed relative to the always-incorrect input). Surprisingly, the responses for Weatherman B in each condition were either 100% sunny or 100% rainy, despite being 50% sunny and 50% rainy in the input; and in the uniform inconsistent condition (where Weatherman A also makes 100% incorrect predictions for days 16–30), the responses for Weatherman B were 100% correct (the responses for Weatherman A are also 100% correct, suggesting that there could have been some difficultly distinguishing the sources).

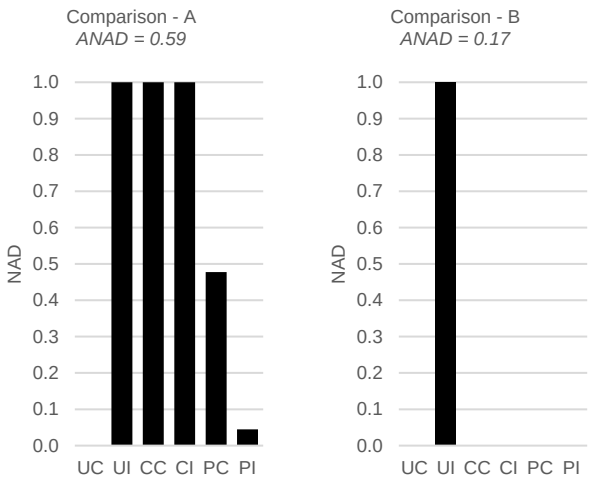

Figure 15. NAD and ANAD scores for Weathermen A and B responses in the comparison manipulation.

4.7 Overall

The ANAD scores presented above are summarized in Figure 16 relative to the 0.36 baseline calculated from Zaroukian (2024).

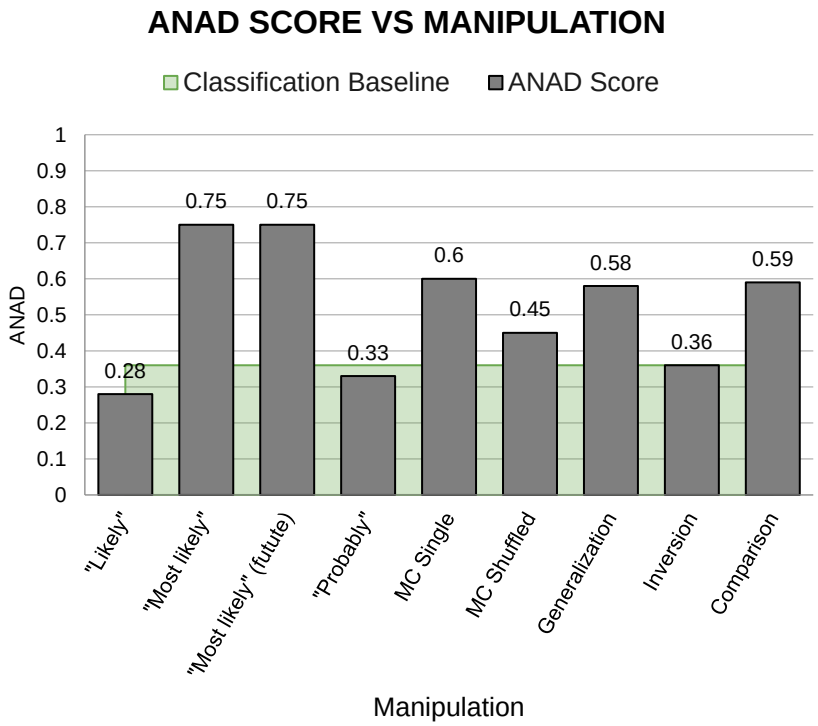

Figure 16. Summary of ANAD scores and manipulations. MC = multiple choice.

Only two manipulations were facilitators, improving performance over baseline. These were the manipulations that added "likely" or "probably" to the prompt, and this addition is believed to have improved performance by reducing the model's hyperconservatism. The other likelihood manipulations, which included the superlative "most," appear instead to have exacerbated hyperconservatism. Similarly, the inclusion of multiple-choice options seems to have decreased performance, possibly due to a bias that lead the model to overwhelmingly select the first (sunny/correct) and second (sunny/incorrect) options, which appears to have compounded with the overall bias in the model to select "sunny" responses in every condition. High ANAD scores were also seen when the model was asked to generalize from weather accuracy to sports accuracy and when a second information source was added.

5. Conclusion

This study provided nine prompt manipulations to more rigorously determine the extent to which LLMs can detect changes in the credibility of an information source over time, finding that only the nonsuperlative hedges "likely" and "probably" facilitated the LLM's performance, while superlative hedges like "most likely" and the inclusion of multiple-choice options inhibited performance. Additionally, the relatively superficial inverting sunny and rainy frequencies had no effect on the LLM's ability to match input correct and incorrect frequencies, but asking the LLM to generalize to a new domain or including a contrasting source of information inhibited performance.

These conclusions were drawn using the novel method of comparing input and output correctness frequencies using ANAD scores that were used to classify manipulations as facilitators or inhibitors relative to the results in Zaroukian (2024). In this report, however, these scores only represent one dimension of the LLM's output: whether the weatherman's prediction was labeled as correct or incorrect. Additionally, this method assumes that a human given the same data would match the most recent input frequencies, but this has not been tested, and ANAD score may not ultimately correlate strongly with desired human-like LLM outputs. While this report centers around the idea of detecting changes in credibility, ANAD scores do not directly compare outputs from consistent conditions (no change in credibility) with outputs from inconsistent conditions (change in credibility) and instead lean on the average score across all conditions. Furthermore, the ANAD metric currently asks a very simple question: how well do the model outputs match the last 15 model inputs? That is, an LLM could receive perfect ANAD scores of 0 by possessing no ability to detect the change that occurred in the data after day 15 and by simply ignoring the first 15 days of data. Future work will address this by

comparing consistent and inconsistent outputs for various days that will be omitted from the input (e.g., days 7 and 22). Future assessments and methodology adjustments may include 1) the use of multimodal inputs, as "embodying" the LLM by giving it a sense of perspective has shown to improve reasoning capabilities (Huang et al. 2022); 2) inputs that simultaneously compare the credibility of one information source on two topics, to determine how well an LLM can keep track of the changes in the credibility of an information source on two topics at the same time; 3) manipulating the model, modifying the temperature of the LLM or comparing two different models, such as GPT 3.5 versus GPT 4.0; and 4) increasing sample size. Additionally, future assessments and methodology adjustments can include improving the ANAD classification threshold calculation by adding an estimation of error (directly assessing the relation between the input position and output frequency when using multiple-choice prompts) and developing a method of calculating a single ANAD score when there are two information sources.

Initial testing shows that this manipulation-evaluation framework is a promising step towards reducing subjectivity and variance in the analysis of LLM reasoning capabilities. The results support the premise that question-answering with LLM can be improved by engineering prompts to reduce hyperconservatism, with the strong caveat that linguistic hedges (including superlatives) may instead increase hyperconservatism. Future work will expand the battery of manipulations tested and improve the evaluation methodology.

6. References

BigScience Workshop. BLOOM: A 176B-parameter open-access multilingual language model. arXiv; 2022 Nov 9. arXiv:2211.05100. https://doi.org/10.48550/arXiv.2211.05100

Diaconescu AO et al. Inferring on the intentions of others by hierarchical Bayesian learning. PLoS Computational Biology. 2014;10(9):e1003810. https://doi.org/10.1371/journal.pcbi.1003810

Hagendorff T et al. Machine psychology. arXiv; 2023 Mar 24. arXiv:2303.13988. https://doi.org/10.48550/arXiv.2303.13988

Huang W et al. Inner monologue: embodied reasoning through planning with language models. arXiv; 2022 July 12. arXiv:2207.05608. https://doi.org/10.48550/arXiv.2207.05608

Kadavath S et al. Language models (mostly) know what they know. arXiv; 2022 July 11. arXiv:2207.05221. https://doi.org/10.48550/arXiv.2207.05221

Kosinski M. Theory of mind might have spontaneously emerged in large language models. arXiv; 2023. https://arxiv.org/vc/arxiv/papers/2302/2302.02083v1.pdf

Rensink RA, O'Regan JK, Clark JJ. To see or not to see: The need for attention to perceive changes in scenes. Psychological Science. 1997;8(5):368–373. https://doi.org/10.1111/j.1467-9280.1997.tb00427.x

Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying language models' sensitivity to spurious features in prompt design or: how I learned to start worrying about prompt formatting. arXiv; 2023 Oct 17. arXiv:2310.11324. https://doi.org/10.48550/arXiv.2310.11324

Srivastava A et al. Beyond the imitation game: quantifying and extrapolating the capabilities of language models. arXiv; 2022 June 9. arXiv:2206.04615. https://doi.org/10.48550/arXiv.2206.04615

Strachan JWA et al. Testing theory of mind in large language models and humans. Nature Human Behaviour. 2024;8(7);1285–1295. https://doi.org/10.1038/s41562-024-01882-z

Zafari S et al. Trust development and explainability: a longitudinal study with a personalized assistive system. Multimodal Technologies and Interaction. 2024;8(3):20. https://doi.org/10.3390/mti8030020

Zaroukian E. Large language models for tracking reliability of information sources. In: Degen H, Ntoa S, editors. Artificial Intelligence in HCI: Fifth International Conference, AI-HCI 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Proceedings, Part III; 2024 June 29–July 4; Washington, DC. Springer Nature; 2024. p. 158–169. https://doi.org/10.1007/978-3-031-60615-1_11

Zheng C, Zhou H, Meng F, Zhou J, Huang M. Large language models are not robust multiple choice selectors. arXiv; 2023 Sep 7. arXiv:2309.03882. https://doi.org/10.48550/arXiv.2309.03882

Zhong L, Wang Z, Shang J. LDB: A large language model debugger via verifying runtime execution step-by-step. arXiv; 2024 Mar 2. arXiv:2402.16906v2. https://arxiv.org/html/2402.16906v2

List of Symbols, Abbreviations, and Acronyms

AI	artificial intelligence
ANAD	average normalized absolute difference
BLOOM	BigScience Large Open-science Open-access Multilingual Language Model
CC	conditional consistent
CI	conditional inconsistent
GPT	Generative Pretrained Transformer
LLM	large language model
NAD	normalized absolute difference
PC	probabilistic consistent
PI	probabilistic inconsistent
UC	uniform consistent
UI	uniform inconsistent

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB