Understanding the Limitations of Large Language Models in Credibility-Tracking Tasks

Avvai Chandrasekaran[1][0009-0006-3550-5561] and Erin Zaroukian[2][000-0002-1381-085X]

[1] Georgia State University, Atlanta GA 30302, USA
[2] DEVCOM Army Research Laboratory, Aberdeen Proving Ground, MD, USA
lncs@springer.com
avvai.chandrasekaran@gmail.com

Abstract. Previous research has modeled humans' ability to track changes in the credibility of an information source 20(Diaconescu et al., 2014), but this has yet to be replicated in Large Language Models (LLMs). Recent studies have shown that LLMs generally exhibit poor abilities to reason about longitudinal data (Chandrasekaran et al., 2024; Zaroukian, 2024), but the prompting method used in these studies may result in the LLM referring only to the most recent data provided. In this study, we evaluate an LLM's longitudinal reasoning capabilities by expanding upon this pervious work to test the LLM's ability to reason about two data points, both before and after a change occurs in the information source's reliability. We find that the LLM performs consistently worse when asked to reason about data points occurring earlier within the pattern and reveal the limitations of previous studies.

Keywords: Large Language Models, Prompt Engineering, Artificial Reasoning, Theory of Mind.

1 Introduction

Large Language Models (LLMs) have become promising tools for artificial reasoning, though they are not without their limitations (e.g., Hawkins et al., 2024; Amirizaniana et al., 2024). While previous research has modeled humans' ability to track changes in the credibility of an information source (Diaconescu et al., 2014), LLMs have not yet been shown to track information in the same way.

A recent study found that the LLMs tested (i.e., BigScience Large Open-science Open-access Multilingual Language Model [BLOOM] and Generative Pretrained Transformer [GPT] 3.5) perform poorly on tasks that require reasoning about longitudinal data (Zaroukian, 2024). This longitudinal data consisted of 30 days of a weatherman's predictions of sunny and rainy weather, as well as whether each prediction was correct or incorrect. The LLM was asked to extrapolate to provide the weatherman's prediction for day 31. As the patterns of predictions in the input became increasingly complex, the LLM's responses became less human-like, not only missing the patterns in the input but also showing an unexpected bias for sunny and incorrect predictions.

Chandrasekaran et al. (2024) expanded on this research by evaluating the efficacy of various prompting methods to improve BLOOM's ability to identify changes in the weatherman's credibility relative to results from Zaroukian (2024). The study established a scoring system referred to as "ANAD" (Average Normalized Absolute Difference), with lower ANADs indicating greater credibility tracking skills. Moreover, prompts were classified into "Facilitators" and "Inhibitors" relative to the ANAD calculated from Zaroukian (2024). In both studies, however, the prompting method did not account for the possibility of the LLM only referring to the most recent data points to generate a response: as the ANAD score is presented in Chandrasekaran et al. (2024), the LLM can achieve an ANAD score of 0 by using only the last 15 days from the 30 day input. Some prompts included a pattern change between days 15 and 16 to assess the LLM's ability detect a change in the input and continue the most recent pattern, but detecting this change and then creating a continuation of the most recent pattern looks identical to simply ignoring the first 15 or more days of input. The ANAD score, then, does not necessarily represent the LLM's ability to detect the changes that were systematically introduced into the input data between days 15 and 16.

This research will address whether the LLM is detecting patterns both before and after the 15th day by prompting BLOOM to guess the weatherman's predictions for days both before and after the change, then comparing the resulting ANAD scores to the ANAD scores from Chandrasekaran et al. (2024). These results should provide a more accurate evaluation of BLOOM's credibility tracking skills.

2 Methodology

The longitudinal weather forecast data provided to the LLM, henceforth referred to as a "prediction history", was based on the prediction history from the original study Zaroukian (2024). BLOOM was presented with 30 days of a weatherman's forecasts, where, for each day, the weatherman predicted that the weather would be "Sunny" or "Rainy". Moreover, for each of the 30 predictions, the weatherman's accuracy was labeled as either "Correct" or "Incorrect".

The prediction history was also provided in one of 6 patterns from a 3x2 design; the pattern was Uniform, Conditional, and Probabilistic, and each of these was either Consistent or Inconsistent. In the Uniform condition, the weather was "Sunny" every day, and the weatherman was "Correct" in his prediction every day; in the Conditional condition, the weather was "Sunny" every day, and the weatherman's prediction was "Correct" for 50% of the days (i.e., "Sunny" days) and "Incorrect" for the rest (i.e., "Rainy" days); in the Probabilistic condition, the weather was "Sunny" for 50% of the days, and the weatherman's prediction was "Correct" for 67% of the days and "Incorrect" for the rest. These patterns were then further divided into two types: Consistent (no change in pattern for all 30 days) and Inconsistent (pattern changes for days 15-30). All patterns are summarized in **Table 1**. Continuing with the methodology used in the original study, each pattern was provided to BLOOM twenty times along with a prompt specifying the day for which to provide the weatherman's prediction (sunny or rainy) and accuracy (correct or incorrect) (Zaroukian, 2024).

The battery of prompt manipulations from Chandrasekaran et al. (2024) was used as well; these prompts, which altered the way the questions' wording or frame, aimed to improve the LLM's reasoning skills. Previous research showed that these tools resulted in improved reasoning skills for LLMs (Hagendorff, 2023). The Likelihood manipulation, which included the addition of "likely" and "probably", aimed to reduce the LLM's hyperconservatism, or the tendency of an LLM to avoid committing to a singular answer even when it can generate a correct one (Strachan, 2024). Similarly, providing the LLM with randomized multiple-choice answers has improved LLM reasoning (Hagendorff, 2024).

Unlike in the previous studies, Day 7 or Day 22 predictions were omitted from the inputs provided to the LLM. BLOOM predicted the weatherman's forecast for the omitted day, to assess BLOOM's ability to predict the score both before and after the pattern shift on Day 16 in the Inconsistent patterns. For the Day 22 predictions, ANAD scores are calculated relative to the last 15 days, exactly done in Chandrasekaran et al. (2024) for the original Day 31 predictions. For the Day 7 predictions, they were calculated relative to the first 15 days. The ANAD scores resulting from this prompting method were compared to the scores calculated from Day 31 predictions in the original study (Zaroukian (2024), see **Table 3**).

Moreover, the LLM responses were compared to what would generally be considered the "human" intuitive response for such a task. For example, when presented with 30 days of the weatherman predicting "Sunny" correctly, a human subject would reasonably assume this pattern would continue onto the 31st day. We used this as the baseline for determining whether an LLM generated the expected "logical" response.

Table 1. Summary of data presented to BLOOM as percentage of sunny days and percentage of correct predictions for 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

Type	Consistency	Days	Actual Weather	Predictions
	Inconsistent	1-15	*Same as Consistent*	*Same as Consistent*
		16-30	50% Sunny	33% Correct

Table 2. Summary of Manipulations. LLM Prompt = Introduction + Data + Question.

Manipulation	**Introduction**	**Data**	**Question**
Day 7	“Day 7 has been skipped. Fill it in.”	“...On day 6, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect]. On day 8, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect] ...”	“On day 7, the weatherman said it would be”
Day 7 – No Introduction	None – no explicit call to complete prediction.	*Same as Day 7*	“On day 7, the weatherman said it would be”
Day 22	“Day 22 has been skipped. Fill it in.”	“...On day 21, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect]. On day 23, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect]...”	“On day 22, the weatherman said it would be”
Day 22 – No Introduction	None – no explicit call to complete prediction.	*Same as Day 22*	“On day 22, the weatherman said it would be”
Likelihood – “Likely”	“Day [7/22] has been skipped. Fill it in.”	*Same as Day 7 or Day 22, respectively*	“On day [7/22], the weatherman **likely** said it would be”
Likelihood – “Probably”	“Day [7/22] has been skipped. Fill it in.”	*Same as Day 7 or Day 22, respectively*	“On day [7/22], the weatherman

			probably said it would be”
Multiple Choice – Fixed Order	“The following is the history of a weatherman's predictions. Day [7/22] has been skipped.”	*Same as Day 7 or Day 22, respectively*	“Choose one of the following: 'On day 22, the weatherman said it would be sunny. He was correct.', 'On day 22, the weatherman said it would be sunny. He was incorrect.', 'On day 22, the weatherman said it would be rainy. He was correct.', 'On day 22, the weatherman said it would be rainy. He was incorrect.”
Multiple Choice – Shuffled	“The following is the history of a weatherman's predictions. Day [7/22] has been skipped.”	*Same as Day 7 or Day 22, respectively*	Same as Multiple Choice – Fixed Order, but the answer choice order is randomized
Day 7 & 22	“The following is the history of a weatherman's predictions. Day 7 and Day 22 have been skipped. Fill them in following this format: On day 7, he said it would be sunny/rainy and he was correct/incorrect. On day 22, he said it would be sunny/rainy and he was correct/incorrect.”	“...On day 6, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect]. On day 8, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect] … On day 21, the weatherman said it would be [sunny/rainy]. He was [correct/incorrect]. On day 23, the weatherman said it would be [sunny/rainy]. He was	“ Answer now following the format. On Day 7, the weatherman said it would be…”

		[correct/incorrect]...”	
Day 7/22 - Un-omitted From History	Unchanged	Includes Day 7 and Day 22 prediction within the forecast history; unchanged from Zaroukian (2024)	“On day [7/22], the weatherman said it would be…
Day 7/22 - Un-omitted From History, No Intro	None – no explicit call to complete prediction.	Includes Day 7 and Day 22 prediction within the forecast history; unchanged from Zaroukian (2024)	“On day [7/22], the weatherman said it would be…

Table 3. Computing ANAD Score using BLOOM results from Zaroukian (2024). Adapted from Chandrasekaran, et al. (2024).

Condition	**Correct in Input (Days 16-30)**	**Max Possible Difference**	**Correct in Output from Zaroukian (2024) (%)**	**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

3 Results

The results are shown for each manipulation with the left charts showing raw outputs, and the right charts showing the NAD (Normalized Absolute % Difference) for each pattern. The left charts' Y axis shows the number of trials, while the right charts' Y axis shows the NAD. The patterns are displayed on the X axis for both charts: "uc" for Uniform Consistent, "ui" for Uniform Inconsistent, "cc" for Conditional Consistent, "ci" for Conditional Inconsistent, "pc" for Probabilistic Consistent, and "pi" for Probabilistic Inconsistent.

3.1 Day 7

Results from requesting the LLM to fill in the Day 7 predictions are shown in **Fig. 1**. Again, this day was chosen because it is before the pattern switch in the inconsistent conditions and requires that the LLM separates this initial pattern from the final pattern. All responses were "Sunny" responses and a mix of "Correct" and "Incorrect" responses. The normalized absolute difference scores per condition again show worse (i.e., higher) scores as patterns become more complex, with a resulting ANAD score of 0.53. This ANAD score is greater than the baseline ANAD score of 0.36 from Day 31 requests in Zaroukian (2024). This indicates that this Day 7 method of assessing the LLM's awareness of the pattern appears to show poorer awareness relative to the baseline.

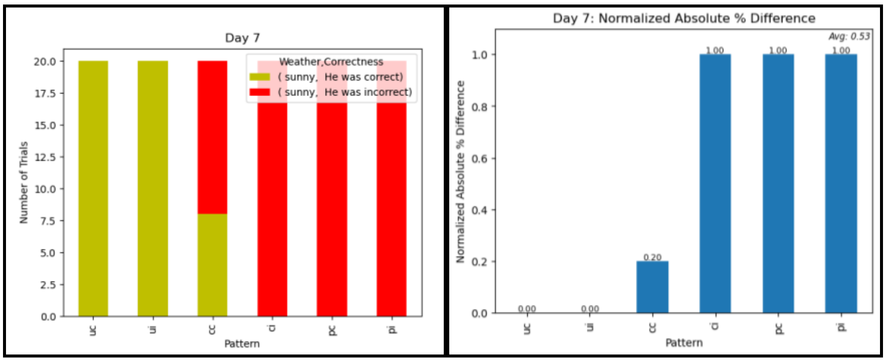

Fig. 1. Results from Day 7 prompts. Left: Responses of "The weather man said it would be sunny. He was correct." and "The weatherman said it would be sunny. He was incorrect." across conditions. Right: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

3.2 Day 7 – No Introduction

Results from requesting the LLM to fill in the Day 7 predictions without providing an introduction ("Day 7 has been skipped. Fill it in.") are shown in **Fig. 2**. All responses were "Sunny" responses, similar to results for Day 7 with introduction. However, this prompt also resulted in increased "Incorrect" responses. The resulting ANAD score was 0.82, showing reduced performance compared to the 0.36 threshold established in Zaroukian (2024); this may be due to the introduction highlighting the location of the missing data point in the first part of the sequence, which contains more "correct" predictions). the LLM not 'noticing' the fact that Day 7 was skipped in the provided history, because the inclusion of an introductory statement acknowledging this fact resulted in significant improvement, as shown in 3.1.

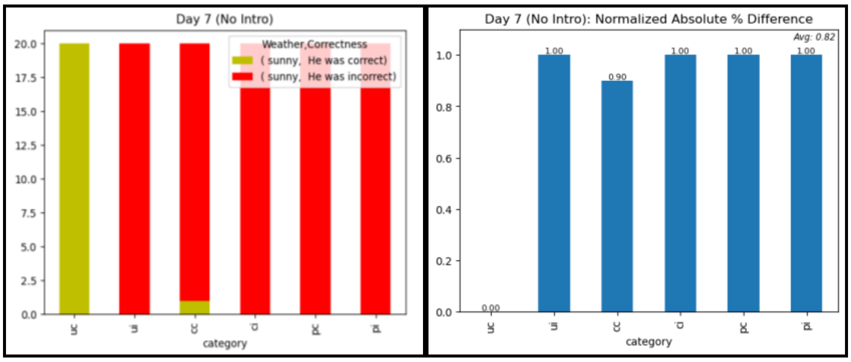

Fig. 2. Results from Day 7 – No Introduction prompts. Left: Responses of "The weather man said it would be Sunny. He was correct." and "The weatherman said it would be sunny. He was incorrect." across conditions. Right: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

3.3 Day 22

Results from requesting the LLM to fill in the Day 22 predictions are shown in **Fig. 3**. Again, this day is beyond the pattern switch in the Inconsistent conditions and so Day 22 responses should differ from Day 7 responses in the Inconsistent conditions.

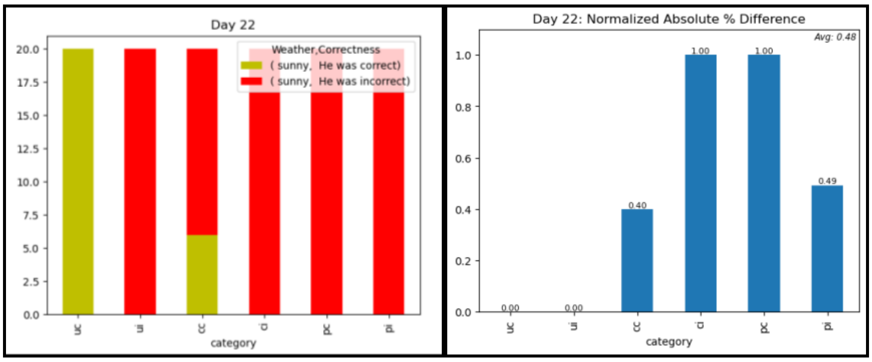

Fig. 3. Results from Day 22 prompts. Left: Responses of "The weather man said it would be sunny. He was correct." and "The weatherman said it would be sunny. He was incorrect." across conditions. Right: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

However, the Day 22 responses were all “Sunny”, even in the Uniform Inconsistent condition, where all predictions for days 16-30 are “Rainy”. The ANAD score is 0.48 (>=0.36), indicating that this prompt appears to show a deterioration in BLOOM’s credibility tracking skills relative to its performance in Zaroukian (2024).

3.4 Day 7 - Likelihood

Results from requesting the LLM to fill in Day 7 predictions with and without “likely” and “probably” are shown in **Fig. 4**. As was done in Chandrasekaran, et al. (2024), this manipulation aimed to reduce hyperconservatism (Strachan et al. 2024).

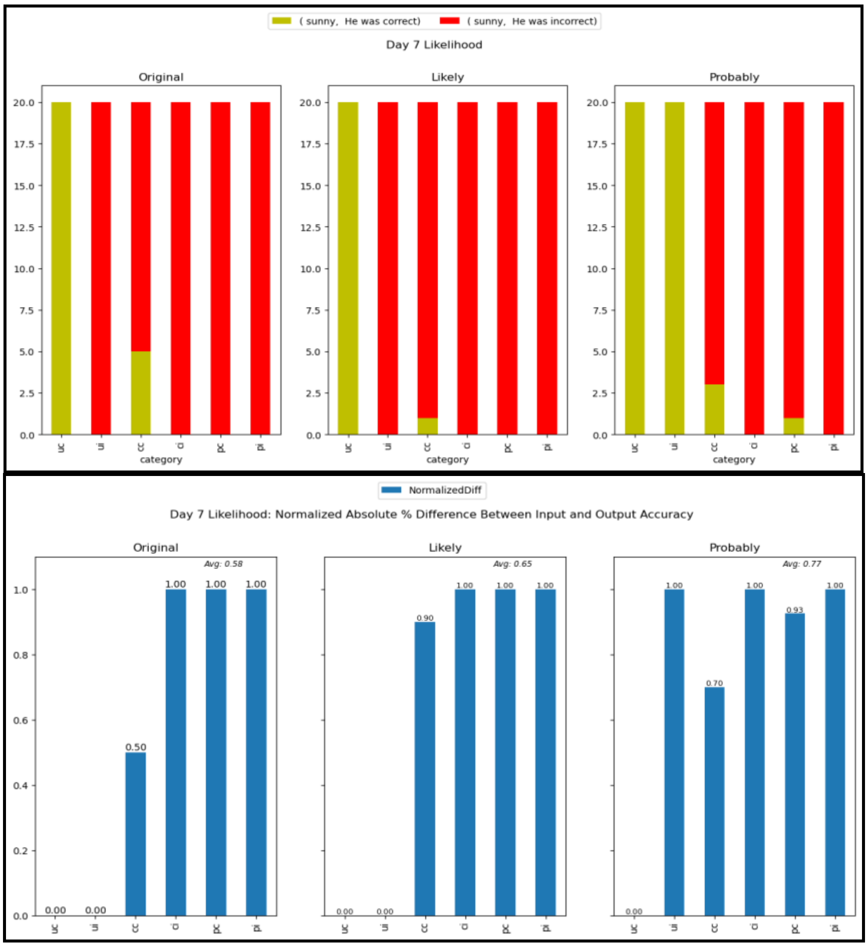

Fig. 4. Results from Day 7 prompts with and without “likely” or “probably” included. Top: Responses of “The weather man said it would be sunny. He was correct.” and “The weatherman said it would be sunny. He was incorrect.” across conditions. Bottom: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

All responses were “Sunny”. This manipulation was a facilitator in Chandrasekaran, et. al (2024), but here, it increased ANAD scores from the unaltered Day 7 prompt of 0.65 and 0.77 (>=0.36); the LLM performed worse here than with Day-31 requests

(Zaroukian, 2024). Also, this manipulation inhibited the LLM's reasoning compared to the unaltered Day-7 prompts. In Chandrasekaran et al. (2024), however, it was a facilitator.

3.5 Day 22 – Likelihood

Results from requesting the LLM to fill in Day 22 predictions with and without “likely” and “probably” are shown in **Fig. 5**.

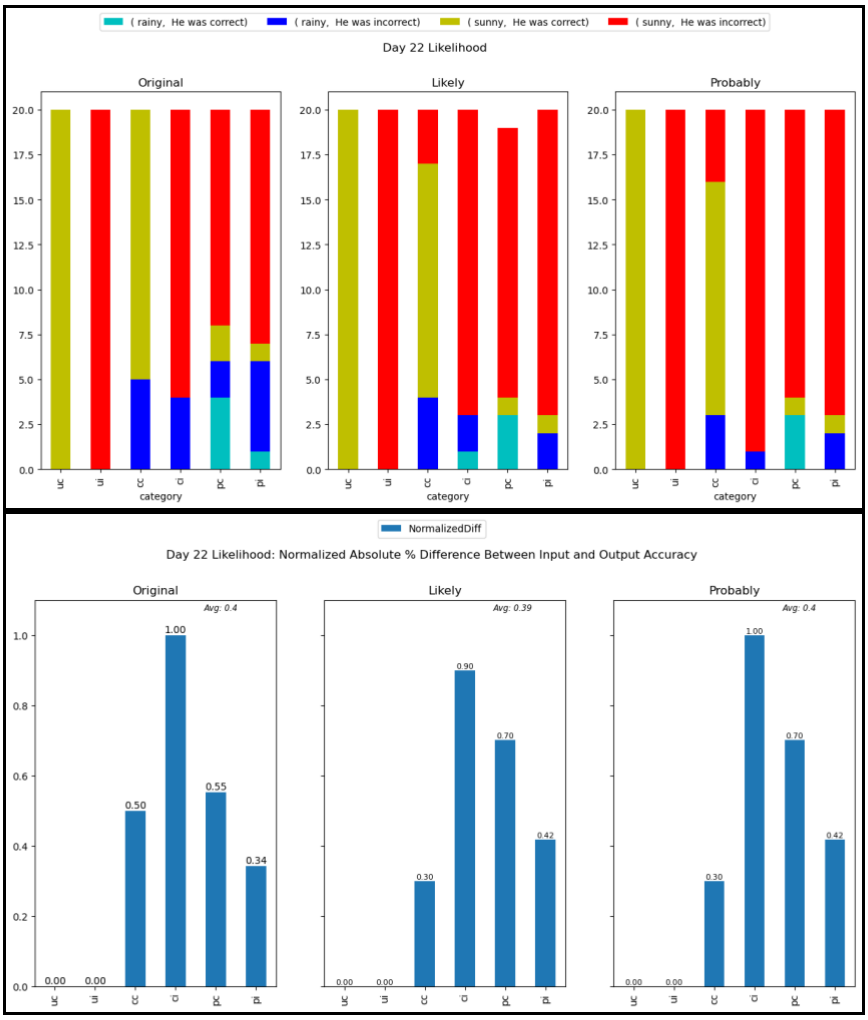

Fig. 5. Results from Day 22 prompts with and without “likely” or “probably” included. Top: Responses of “The weather man said it would be rainy. He was correct.”, “The weatherman said it would be rainy. He was incorrect.”, “The weather man said it would be sunny. He was correct.”, and “The weatherman said it would be sunny. He was incorrect.” across conditions. Bottom: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

Including a “Likelihood” manipulation on the Day 22 prompts resulted in more varied responses, including “Rainy” predictions. Moreover, the inclusion of the likelihood wording also resulted in increased ANAD scores from Zaroukian (2024).

3.6 Day 7: Multiple Choice, Single order and shuffled

Results from requesting the LLM to fill in Day 7 predictions using a multiple-choice format are shown in **Fig. 6**, with either a fixed order of options or by shuffling the ordering of options across trials.

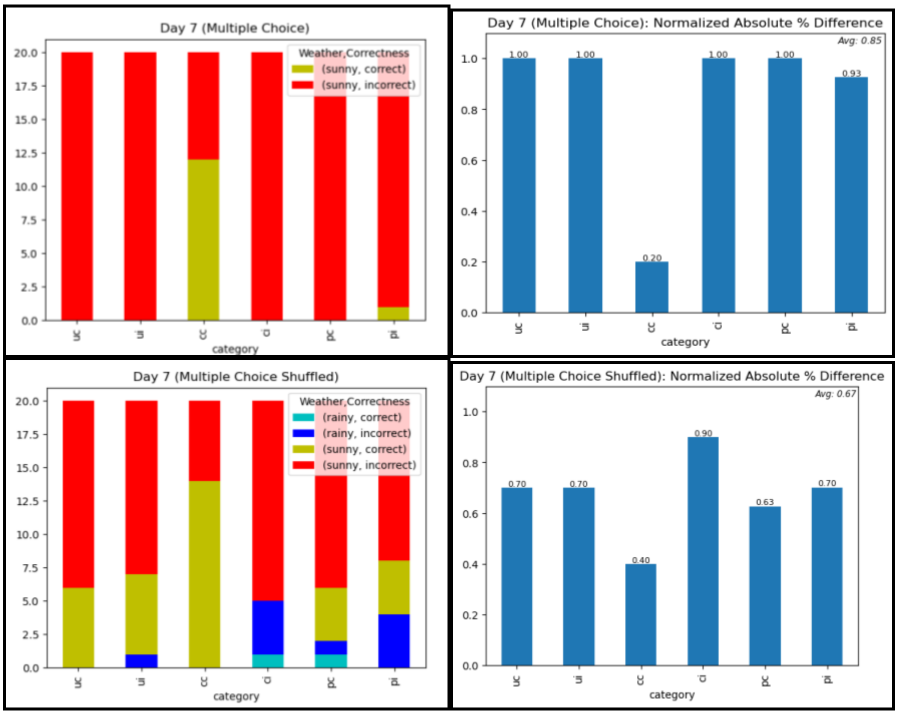

Fig. 6. Results from Day 7 prompts with fixed (top) or shuffled (bottom) order of multiple-choice options. Left: Responses of “The weather man said it would be rainy. He was correct.”, “The weatherman said it would be rainy. He was incorrect.”, “The weather man said it would be sunny. He was correct.”, and “The weatherman said it would be sunny. He was incorrect.” across conditions. Bottom: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

The fixed order multiple-choice manipulation resulted in all “Sunny” options. However, it resulted in an increase in “Incorrect” responses, even for Uniform patterns, where all predictions in the first 15 days are “Correct”. This may be due to the order in which the answer choices are presented; notably, the presence of varied answer choices resulted in the LLM providing “Rainy” responses. The ANAD score for single order was higher at 0.85 (>=0.36) than for shuffled order at 0.67 (>=0.36), making it an inhibiting manipulation. This occurred in Chandrasekaran, et al. (2024) as well – while the Multiple Choice variation was an overall inhibitor of the LLM’s skills, the shuffled order variation resulted in better performance compared to the single order version.

3.7 Day 22: Multiple Choice, Single Order and Shuffled

Results from requesting the LLM to fill in Day 22 predictions using a multiple-choice format are shown in **Fig. 7**, with either a fixed order of options or by shuffling the order of options across trails.

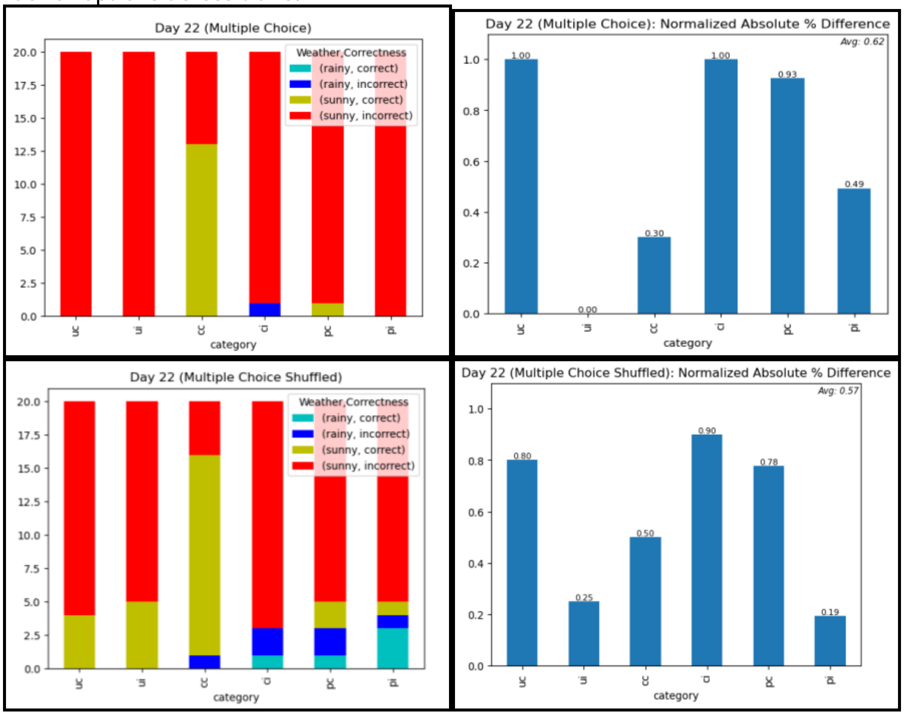

Fig. 7. Results from Day 22 prompts with fixed (top) or shuffled (bottom) order of multiple-choice options. Left: Responses of "The weather man said it would be rainy. He was correct.", "The weatherman said it would be rainy. He was incorrect.", "The weather man said it would be sunny. He was correct.", and "The weatherman said it would be sunny. He was incorrect." across conditions. Bottom: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

The introduction of multiple choice prompts resulted in a far more varied set of responses compared to the unaltered Day 22-request. However, it resulted in an increased number of "Incorrect" responses, even for the Uniform Correct pattern, where all the weatherman was "Correct" for all 30 days. The ANAD score for the shuffled method (0.57), was lower than for the single order version (0.62). Both of these scores were lower than the 0.36 ANAD score found in Zaroukian (2024) meaning that the LLM failed to perform as it did for Day 31-requests. However, this manipulation resulted in ANAD scores that were higher than the unaltered Day 22-requests (0.48), meaning that this was an inhibiting manipulation. This elevated ANAD score is largely due to its inability to correctly continue the pattern for the Uniform Correct response.

3.8 Day 7: Without Omission from Data

Results from requesting the LLM to provide Day 7 predictions when Day 7 predictions were present in the input and shown in **Fig. 8**, both with and without introduction.

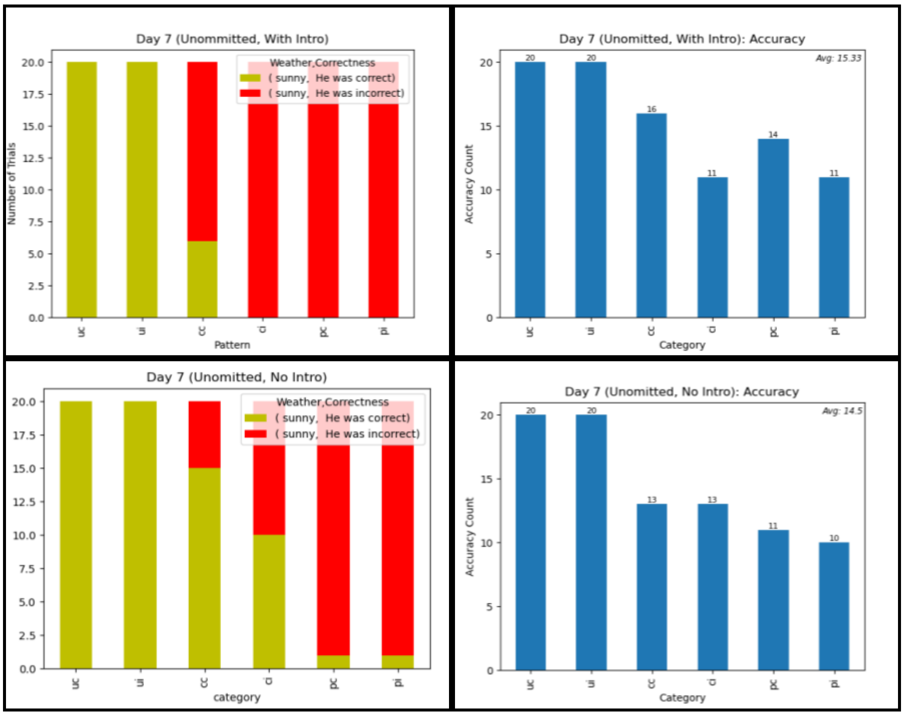

Fig. 8. Results from Day 7 prompts when Day is not omitted from input, both with (Top) and without (bottom) the introduction ("Day 7 has been skipped. Fill it in."). Left: Responses of "The weather man said it would be sunny. He was correct." and "The weatherman said it would be sunny. He was incorrect." across conditions. Right: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

The LLM was provided with all 30 days of the weatherman's forecast prediction, instead of omitting Day 22 from the history. We speculated that the inclusion of all 30 days of history would result in improved performance, but this manipulation resulted in less accurate performance from the unaltered Day 7 prompt, which omitted the Day 7 data. Because there is only one correct answer for the Day 7 prediction (present in the input data), we utilized "Accuracy" instead of the ANAD score here to compare performance. Overall, the LLM struggled more and generated less accurate outputs without an introduction; this applied for all patterns except for Uniform Consistent and Uniform Inconsistent, where the LLM had perfect accuracy.

3.9 Day 22: Without Omission from Data

Results from requesting the LLM to provide Day 7 predictions when Day 7 predictions were present in the input and shown in **Fig. 9**, both with and without introduction.

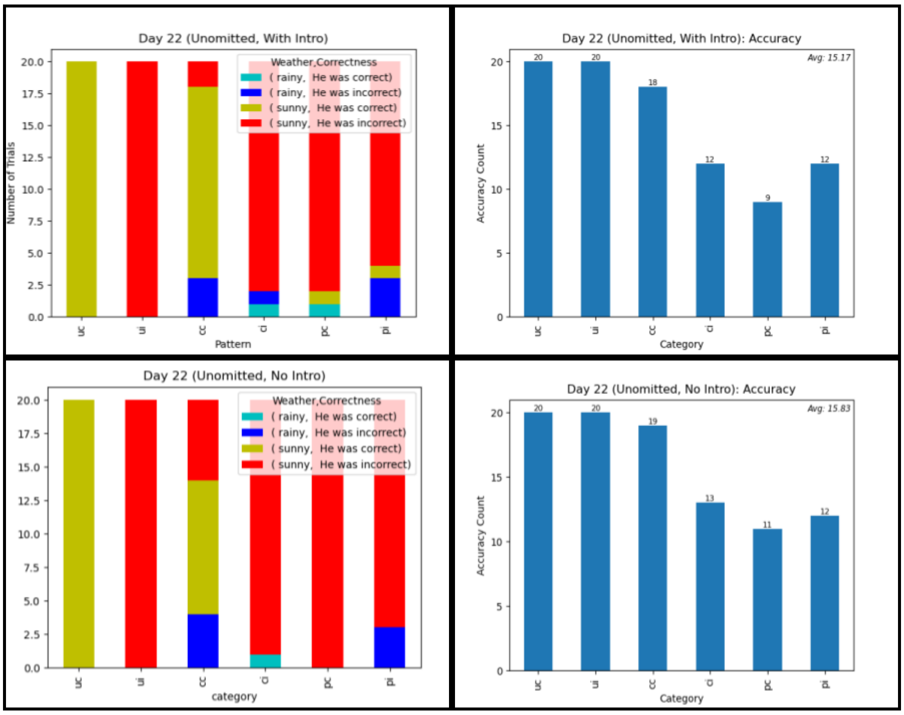

Fig. 9. Results from Day 22 prompts when Day 22 is not omitted from input, both with (Top) and without (bottom) the introduction ("Day 22 has been skipped. Fill it in."). Left: Responses of "The weather man said it would be rainy. He was correct.", "The weatherman said it would be rainy. He was incorrect.", "The weather man said it would be sunny. He was correct.", and "The weatherman said it would be sunny. He was incorrect." across conditions. Right: Normalized absolute difference scores for each condition relative to Zaroukian (2024).

The LLM was provided with all 30 days of the weatherman's forecast, instead of omitting Day 22 from the history. We speculated that the inclusion of all 30 days of history would result in less accurate performance, similar to the Day 7 request results. However, this manipulation resulted in overall a more accurate performance from the unaltered Day 22 prompt, which omitted the Day 22 data. Because there is only one correct answer for the Day 22 prediction (present in the input data), we utilized

"Accuracy" instead of the ANAD score here to compare performance. Overall, the LLM generated more accurate outputs when an introduction is included.

4 Conclusion

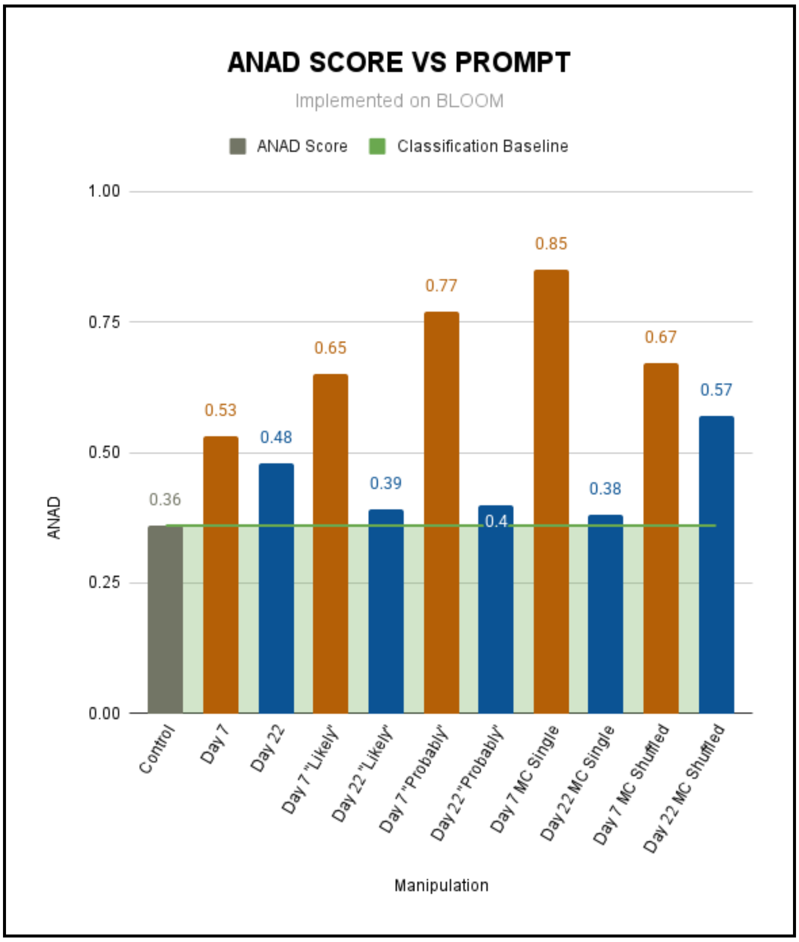

Fig. 10. ANAD score vs Prompt; Green line indicates original ANAD score from Zaroukain (2024).

This study found that BLOOM (BigScience Workshop, 2022) generally exhibits "worse" performance (elevated ANAD scores) when asked to complete the predictions for days occurring within the 30-day history as opposed to predicting the pattern occurring on Day 31. We hypothesize that this may be due to a "Lost in the Middle" effect – because the Day 7 and Day 22 forecasts occur towards the 'middle' of the data, the LLM may have struggled to access these data points more than it did for the data points relevant for Day 31 predictions, which occurred towards the very end (Nelson, 2023). Day 22 predictions generally resulted in lower ANAD scores (better performance), but overall, none of the manipulations were clear facilitators of BLOOM's credibility tracking skills.

Notably, the LLM rarely responded with "rainy" predictions despite multiple attempts, even for a Day 22 request based on the "Uniform Inconsistent" pattern, where the weatherman was consistently accurate in predicting rainy weather for Days 16-30. This may be attributed to "Majority Bias", or the tendency of LLMs to rely on information that is most frequently mentioned within a prompt (Hagendorff, 2023). In other words, because the LLM is asked to continue the prompt from Day 22 onwards, it may have only referred to the previous 21 days of predictions within the weatherman's history, meaning 70% of the predictions present in the data were "Sunny".

Moreover, the ANAD scores for both the Uniform Inconsistent and Uniform Consistent data were a perfect 0 despite the error in the LLM's predictions about the "Sunny" vs "Rainy" forecast. This fact reveals a weakness in the ANAD score's ability to truly quantify an LLM's ability to detect changes in an information source's reliability, especially when asked to take note of two variables, a task that we speculate would be easily achieved by human decision-makers.

Finally, alternative prompting methods as used in Chandrasekaran, et al. (2024) were also tested. For Day 7, utilizing a "Likelihood" variation resulted in worsened performance, while it improved performance for Day 22. This may be because Day 7-requests did not provide the model with enough data occurring before the prediction point, and this prompting method might be better suited for longer input sequences/greater amounts of context. The Multiple Choice prompt variation also worsened model performance for Day 7-requests, but shuffling the answer choices did result in a relatively better performance. In fact, additional manipulations to the prompt for Day 7-requests seems to result in increasingly worse performance. However, Day 22 prompts performed significantly better overall, with the Likelihood and Multiple Choice variations improving performance

Ultimately, this study reveals the limitations of the LLM BLOOM's credibility tracking skills about multiple data points occurring within a 30-day prediction history of an information source. We hypothesize that these limitations were due to the size of the input sequence, and that longer sequences will result in improved credibility-tracking performance. This is corroborated by the fact that Day 22 prompts seemed to consistently result in better LLM performance than Day 7 prompts, as Figure 1 shows. This phenomenon may also be explained by the LLM's recency bias (Hagendorff, 2023), meaning that the LLM tends to refer to the most recent data points within the history, causing it to generate incorrect responses for Day 7 prompts.

The LLM also struggled with correctly responding to prompting methods that combined both Day 7 and Day 22 requests, frequently generating nonsensical responses or instead choosing to continue the prediction history pattern onto the 32nd day and so on.

Future research will include testing longer sequences of data to determine whether larger temporal contexts improve LLM's credibility tracking skills. For example, the LLM could be provided with 100 days of forecast history, with Inconsistent inputs shifting patterns on Day 50, and the LLM will be asked to predict forecasts on Day 40 vs Day 80. The additional amount of context may result in an improvement compared to Day 7 vs Day 22 requests. The evaluation methodology, including the ANAD score will also be improved by including the LLM's predictions of "Sunny" and "Rainy" in the calculations, to properly account for its ability to track credibility through more than one variable.

Acknowledgments. We thank Mark Mittrick, Adrienne Raglin, and Justine Rawal for their support in this research program. Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amirizaniani, M., Martin, E., Sivachenko, M., Mashhadi, A., Shah, C.: Can LLMs Reason Like Humans? Assessing Theory of Mind Reasoning in LLMs for Open-Ended Questions. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, pp. 34-44. ACM, New York, NY (2024)
2. BigScience Workshop: BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 (2022)
3. Chandrasekaran, Avvai, et al. 'Developing a Framework to Evaluate Credibility Tracking in Large Language Models'. DEVCOM Army Research Laboratory Technical Report ARL-TR-10041. 2024
4. Diaconescu, Andreea O., et al. 'Inferring on the Intentions of Others by Hierarchical Bayesian Learning'. PLoS Computational Biology, vol. 10, no. 9, Public Library of Science (PLoS), Sept. 2014, p. e1003810, https://doi.org/10.1371/journal.pcbi.1003810.
5. Hagendorff, Thilo. 'Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods'. arXiv [Cs.CL], 24 Mar. 2023, http://arxiv.org/abs/2303.13988. arXiv.
6. Hawkins, T., Zaroukian, E., Raglin, E.: AI's not to Reason Why (because we don't know if it can). Modern War Institute, 1 Nov. 2024, https://mwi.westpoint.edu/ais-not-to-reason-why-because -we-dont-know-if-it-can/
7. Nelson, L.F., et al. 'Lost in the Middle: How Large Language Models Use Long Contexts'. arXiv[Cs.CL], 6 Jul. 2023, https://arxiv.org/abs/2307.03172
8. Strachan, J.A., Albergo, D., Borghini, G., Pansardi, O., Scaliti, E., Gupta, S., Saxena, K., Rufo, A., Panzeri, S., Manzi, G., Graziano, M.S.A., Becchio, C.: Testing theory of mind in large language models and humans. Nature Human Behaviour **8**(7), 1285–1295 (2024)

9. Zaroukian, Erin. ‘Large Language Models for Tracking Reliability of Information Sources’. *Artificial Intelligence in HCI*, Springer Nature Switzerland, 2024, pp. 158–169, https://doi.org10.1007/978-3-031-60615-1_11. Lecture Notes in Computer Science.
10. LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25