Investigating LLMs Confidence Through Doubt Creation

Or Lichter
204169775

Nir Endy
205686397

Ron Tsibulsky
204737779

Abstract

This study investigates the impact of iterative prompting on Large Language Model (LLM) responses, focusing on how expressions of doubt influence the consistency and reliability of answers across various models of comparable size. Using a methodology that presents LLMs with binary-choice questions followed by expressions of doubt in varying settings, we analyze how different prompting strategies impact their behavior. By comparing the responses of different LLM architectures under the tested prompting strategies, we aim to uncover patterns in how these models react to various forms of iterative prompting. Our research contributes to the growing body of knowledge on LLM behavior, offering insights into the malleability of AI-generated content and the potential for guiding or manipulating LLM outputs.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing, demonstrating remarkable capabilities in tasks ranging from text generation to complex problem-solving(Bubeck et al., 2023; Bommasani et al., 2022). However, questions persist regarding the consistency, reliability, and potential for manipulation of their outputs. This research explores the effects of iterative prompting on LLM responses, comparing models of similar size under different feedback conditions to gain insights into their behavior and decision-making processes.

Our approach involves presenting LLMs with binary-choice questions, followed by an expression of uncertainty about their initial answer, and a request to reconsider (e.g., "I am not sure about the answer. Can you try again?"). The outcomes of this interaction are explored in Experiments 1 and 2, detailed in Section 3. We then extend our investigation in Experiment 3 by introducing performance feedback and iteratively posing new questions, examining how feedback and iterative questioning interact with expressions of doubt to influence model responses. Finally, in Experiment 4, we shift our focus to model confidence, as measured by output logits, and evaluate how induced doubt impacts this confidence metric.

Through this methodology, we aim to investigate several key questions:

1. How do different types of feedback impact response consistency across multiple iterations?
2. To what extent does positive reinforcement through performance feedback lead to more confident or stable answers?
3. How do these effects vary across different LLM architectures of comparable size, and what does this reveal about their underlying mechanisms?
4. What are the implications of these findings for the development of more robust and reliable AI systems?

To facilitate these experiments, we use the factual dataset *CounterFact-Tracing*, adapted from (Meng et al., 2022). This dataset consists of 21,919 questions, each paired with both a correct and an incorrect answer, serving as the binary-choice options in our study.

In the following sections, we will detail our experimental setup, present our findings, and discuss their implications for the field of artificial intelligence and natural language processing. **Or:** Add things we did and what are our results are here in the introduction

2 Related Work

LLMs consistency is a widely researched topic. (Elazar et al., 2021) investigated the consistency of pretrained language models when prompts are

phrased differently. They proposed a method to enhance model consistency by modifying the loss function during training. However, this approach requires retraining the models, which may not be practical for large-scale LLMs.

(Xu et al., 2023) explored the robustness of LLM confidence when exposed to repetitive misinformation. Their results showed that repeated exposure to misleading information can diminish the model's confidence in correct answers, even leading to incorrect responses. However, this study primarily focused on confidence robustness rather than providing a practical framework for assessing the model's original confidence in its answers.

(Krishna et al., 2024) investigated strategies to enhance the truthfulness of LLM outputs through iterative prompting, while (Salinas and Morstatter, 2024) demonstrated how varying prompt phrasing can significantly impact LLM performance. (Wei et al., 2022) specifically shows that a technique called "Chain Of Thought" prompting improve reasoning and performance in LLMs.

Works have been done also on the ability of a model to asses its own confidence, and calibrate this assessment to improve performance. (Xiong et al., 2024) Develop black-box methods to estimate the confidence in a model's answer, relying on the model assessing its own confidence. (Mielke et al., 2022) took this further by training a calibrator model to predict the likelihood of correctness, then adjusting the responses to reduce overconfidence and improve calibration. (Guo et al., 2017) also conducted extensive research on confidence calibration in LLMs, suggesting practical improvements during training to enhance confidence calibration.

(Perez et al., 2022) shows LLMs present a behavior called "sycophancy", where models tend to generate responses that echo user's preferred answers. This highlights the importance of careful prompt engineering to avoid unintended biases when seeking to improve model performance.

Finally, (Liu et al., 2023) conducted a comprehensive survey of research on LLM truthfulness and reliability, consolidating various findings and methodologies for evaluating and enhancing these aspects.

3 Experiments

To investigate the effects of iterative prompting and doubt expressions, we design a series of experiments that focused on analyzing model responses to factual questions, examining response changes, studying the impact of answer positioning, measuring model confidence through logit differences, and evaluating the effectiveness of repeated doubt with feedback. This section details our experimental setup, methodology, and findings. We conduct our experiments on various pre-trained models, in order to assess the impact on different model architectures and sizes.

3.1 Impact of Introducing Doubt on Factual Questions

In this experiment, we presented each model with a series of factual questions, each having two possible answers. After the model provided its initial response, we introduced an element of doubt as presented in Figure 1. We then recorded the model's subsequent response.

Experimental Setup

- Number of questions: 22,000
- Evaluation metric: Accuracy (percentage of correct responses)

Model	Size	Before Doubt	After Doubt
Llama 3.2	1B	52.2%	49.3%
Llama 3.2	3B	64.3%	44.1%
Phi 3.5	3.82B	86.2%	86.7%
Llama 3.1	8B	71.9%	80%
Mixtral	8x7B	73.4 %	76.1%
Nemo	12.2B	81.5%	83.9%

Table 1: Experiment 1 results: Accuracy comparison before and after adding doubt

Results and Discussion Our results show a nuanced impact of expressing doubt on model performance, strongly correlated with model size:

- Smaller models (Llama 3.2 1B and 3B): Expressing doubt led to a decrease in accuracy for both models, with a more pronounced effect on the 3B model (20.2 percentage point decrease) compared to the 1B model (2.9 percentage point decrease).
- Larger models (Llama 3.1 8B, Mixtral 8x7B): These models demonstrated improved accuracy after the expression of doubt, with the most substantial improvement observed in the Llama 3.1 8B model (8.1 percentage point increase).

```
User: I have a question with two possible answers. Please provide the
    correct one, either a or b.
Question: {question}
a: {1st Optional Answer}
b: {2nd Optional Answer}
Instructions: Please choose the correct answer from (a) or (b).

Answer: {1st Model Answer}      ← [Point A]

User: I am not sure about the answer. Can you try again? Just answer
   with a or b.

Answer: {2nd Model Answer}      ← [Point B]
```

Figure 1: Baseline template for question-answer interaction.

- Medium-sized model (Phi 3.5 mini instruct 3.82B): This model showed a slight improvement in accuracy (0.5 percentage point increase), suggesting a transition point in model behavior.

These findings suggest that:

1. Model size plays a crucial role in how LLMs respond to expressed doubt.
2. Larger models (8B and above) appear more capable of using the doubt prompt as an opportunity for reassessment and improvement.
3. Smaller models (3B and below) are more susceptible to uncertainty, leading to decreased performance when doubt is expressed.
4. There may be a transitional size range (around 3-4B parameters) where models begin to show resilience to doubt and potentially benefit from it.

These results highlight the complex relationship between model size, confidence, and the ability to process and benefit from user feedback. The clear divide in behavior between smaller and larger models suggests that as models grow in size, they develop more robust internal representations and decision-making processes that allow them to leverage uncertainty productively.

3.2 Examining Response Changes

In this second experiment, we took a closer look at how the expression of doubt impacted the models' responses. Specifically, we categorized the changes in responses as follows:

1. **Correct to Incorrect (V $\rightarrow$ X)**: The model had an initially correct answer, but expressing doubt caused it to switch to an incorrect answer. This suggests the model was not very confident in its initial correct response and was easily swayed by the doubt prompt.
2. **Incorrect to Correct (X $\rightarrow$ V)**: The model had an initially incorrect answer, but expressing doubt led it to correct that answer. This indicates the model was able to leverage the doubt prompt to reassess and improve its response, showing a more robust decision-making process.
3. **Correct to Correct (V $\rightarrow$ V)**: The model maintained its initially correct answer even after the doubt prompt was introduced. This implies the model was very confident in its initial correct response and was not significantly affected by the expression of doubt, demonstrating a stable and resilient decision-making strategy.
4. **Incorrect to Incorrect (X $\rightarrow$ X)**: The model had an initially incorrect answer and maintained that incorrect answer even after the doubt prompt was introduced. This suggests the model was not able to use the doubt prompt to improve its response, indicating potential limitations in its understanding or decision-making capabilities.

By analyzing the distribution of these response changes, we aimed to gain a more nuanced understanding of how doubt affects the models' decision-making processes.

Results and Discussion Table 2 presents the distribution of response shifts for each model. While the initial experiment suggested a decrease in accuracy among the smaller models, closer analysis reveals that this change predominantly reflects a shift in response type, with approximately 90% of the answers simply change when expressing doubt to the model.

In contrast, larger models demonstrate a higher incidence of incorrect-to-correct response transitions compared to correct-to-incorrect shifts, with the ratio of X→V transitions consistently exceeding V→X by more than double. This pattern suggests that expressions of doubt are associated with improved accuracy.

In subsequent experiments, we will explore the extent to which this trend holds under different conditions.

Model	Size	V→V	V→X	X→V	X→X
Llama 3.2	1B	6.3%	45.9%	42.9%	4.9%
Llama 3.2	3B	8.2%	56.3%	35.2%	0.3%
Phi 3.5	3.82B	86.1%	0%	0.5%	13.4%
Llama 3.1	8B	65.1%	6.1%	14.6%	14.2%
Mixtral	8x7B	70.9%	2.2%	5.1%	21.8%
Nemo	12.2B	81.8%	0.5%	2.71%	15%

Table 2: Experiment 2 results: How adding doubt actually affects the correctness of

3.3 Impact of Answer Position and Prompt Variations

During the design of the previous experiments, we had to make several decisions regarding the structure of the prompts and the positioning of the correct answer. In order to understand the impact of these decisions on the results, we designed an experiment that present the same factual questions to the models, but with different prompt variations and answer positions.

Experimental Setup

- Number of questions: 1500
- **Evaluation Metrics**: We defined several metrics to be able to quantify the model-prompt interaction.
  - **Accuracy**: Percentage of correct responses.
  - **Positional Robustness (PR)**: $1 - \left| Acc^{CorrectFirst}_{BeforeDoubt} - Acc^{CorrectSecond}_{BeforeDoubt} \right|$ The model's level of robustness to the position of the correct answer.
  - **Correctness Certainty (CC)**: $\begin{cases} \frac{V \to V}{V \to V + V \to X}, & \text{if } V \to V + V \to X > 0 \\ 0, & \text{otherwise} \end{cases}$ The model's ability to maintain correct answers even after doubt is introduced.
  - **Incorrectness Improvement (II)**: $\begin{cases} \frac{X \to V}{X \to V + X \to X}, & \text{if } X \to V + X \to X > 0 \\ 1, & \text{otherwise} \end{cases}$ The model's ability to correct its wrong answers after doubt.
  - **Average Metric (AM)**: $\frac{PR+CC+II}{3}$ Averages the three metrics above.
- **Prompt Variations**: We introduced variations of the baseline prompt from the first experiment, as follows:
  1. **baseline plus**: Similar to basic but with explicitly added "Assistant:"
  2. **basic with system message**: Includes system context for the assistant's role
  3. **encouraging**: Uses positive reinforcement and reward motivation
  4. **discouraging mild**: Added doubt with "That's completely wrong." feedback.
  5. **discouraging harsh**: Added doubg with "Wow, that's such a stupid answer." feedback
  6. **example_a**: Includes one example with option 'a' as correct
  7. **example_b**: Includes one example with option 'b' as correct
  8. **example_ab**: Includes two examples with 'a' then 'b' as correct
  9. **example_ba**: Includes two examples with 'b' then 'a' as correct

Results and Discussion Table 3 and Figure 3 summarizes the accuracy of each model under different prompt variations and answer positions.

Our findings indicate:

- **Positional Bias**: We observed a significant positional bias in the models' responses, that was hidden by just looking at the overall accuracy. This may indicate that the models did not entierly understand the prompt. The only model that seems to be robust to the answer positioning is phi-3.5-mini. But as we can see in Figure 3, it comes with an expense of low Incorrectness Improvement.

Table 3: Models Accuracy conditioned by answer positioning on baseline prompt

Model	V→V	V→X	X→V	X→X	Before Doubt	After Doubt
Results when correct answer was presented first						
llama3_2_1B	0.00	82.67	0.00	17.33	82.67	0.00
llama3_2_3B	0.07	98.60	0.40	0.93	98.67	0.47
phi_3.5-mini	82.00	0.00	0.00	18.00	82.00	82.00
llama3_1_8B	67.73	29.87	0.40	2.00	97.60	68.13
mistral_8x7B	91.39	8.01	0.00	0.60	99.40	91.39
mistral_Nemo	97.27	0.87	0.13	1.73	98.14	97.40
Results when correct answer was presented second						
llama3_2_1B	12.87	0.00	87.13	0.00	12.87	100.00
llama3_2_3B	13.00	17.60	69.40	0.00	30.60	82.40
phi_3.5-mini	91.80	0.00	0.00	8.20	91.80	91.80
llama3_1_8B	43.27	1.73	27.47	27.53	45.00	70.74
mistral_8x7B	53.84	2.67	11.94	31.55	56.51	65.78
mistral_Nemo	64.87	0.53	6.40	28.20	65.40	71.27
Combined results						
llama3_2_1B	6.43	41.33	43.57	8.67	47.76	50.00
llama3_2_3B	6.53	58.10	34.90	0.47	64.63	41.43
phi_3.5-mini	86.90	0.00	0.00	13.10	86.90	86.90
llama3_1_8B	55.50	15.80	13.93	14.77	71.30	69.43
mistral_8x7B	72.61	5.34	5.97	16.08	77.95	78.58
mistral_Nemo	81.07	0.70	3.27	14.97	81.77	84.34

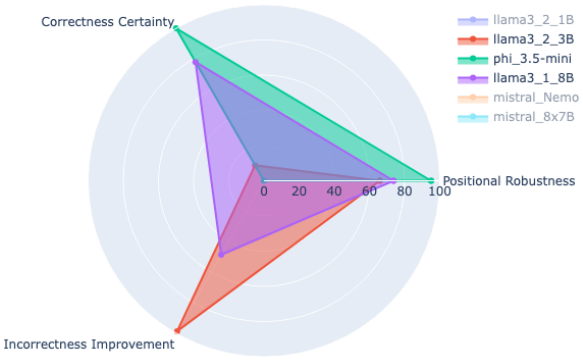

Figure 2: Illustraion of model's metrics on the baseline prompt

- **Effect of Prompt Variations**: We can see that the choise of the prompt has an effect on the model's performance. But we have not found a prompt that is significantly better than the others. The encouraging prompt seems to be the best one, but the difference is not significant. But we see that some models are more sensitive to the prompt variations than others.

- **Model Strengths and Weaknesses**: We can see that the models have different strengths and weaknesses. and non of the models is the best in all metrics.

Conclusion Although we found deviations in the models' performance based on prompt variations and answer positioning, we are not concerned about our choice of a baseline prompt, as it seems to be working fairly well. But we understand that this may require further investigation in future experiments.

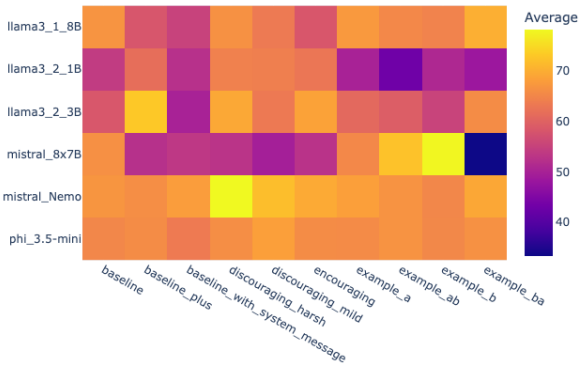

Figure 3: Models performance on all prompts

3.4 Analyzing Model Confidence through Logit Differences

In the previous experiments, we focused on accuracy based metrics to evaluate the models' performance. However, the way that models produces their answers is not binary, and using only accuracy as a metric may not provide a full picture of the models' decision-making process. In this experi-

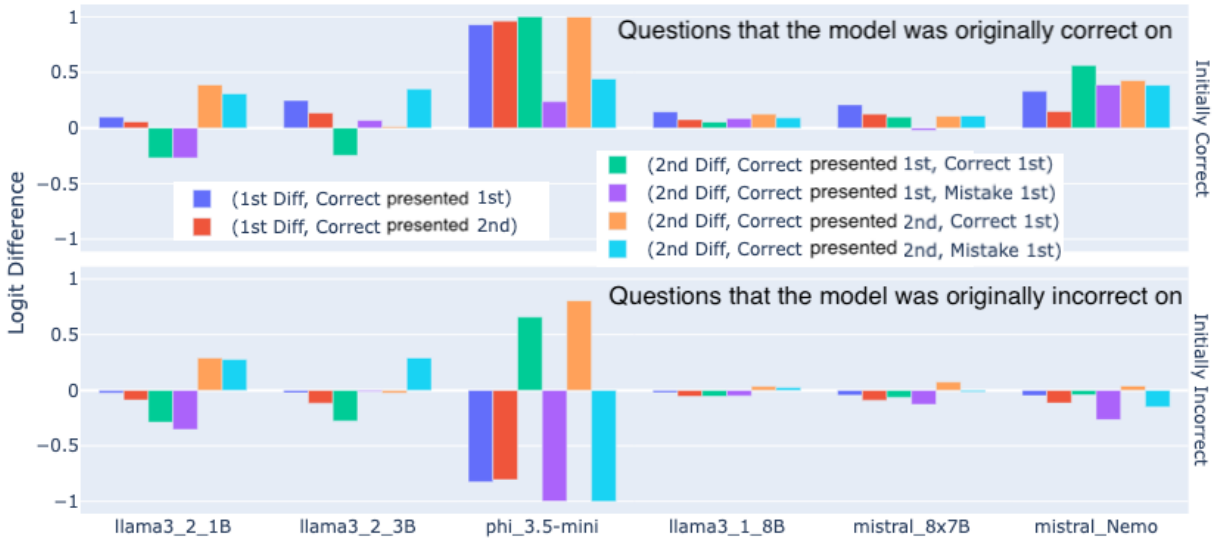

Figure 4: Model's confidence in each of the 6 confidence samples, on the baseline prompt. The top row represents the models' were correct in their initial response, while the bottom row represents the models' were incorrect in their initial response.

ment, we aim to analyze the models' confidence in their answers by examining the difference in logits between the correct and incorrect answer tokens. By analyzing the confidence shifts after expressing doubt, we assess the models' ability to adjust their internal certainty and correct their answers.

For each question, we sampled the confidence (logit difference between the correct and incorrect answer tokens) at 2 points, as marked in the baseline template 1, **Point A** (1st response / before doubt) and **Point B** (2nd response / after doubt).

In **Point A**, models can either answer **a** or **b** as their initial response. While models can either **correct** or **incorrect**, we wanted to control this factor, thus we simulated both cases, regardless of the natual model's response. so in fact, for each question, we sampled 3 logit differences, one in Point A and two in Point B, one for each case of the model's initial response. In addional, based on the results of the previous experiment, we found out that the answer positioning has a significant impact on the models' performance. Therefore, we decided to control this factor as well, so we ended up with **6** logit differences for each question.

Experimental Setup

- Number of questions: 1,500
- **Experiment Scenarios**:
  - **Correct Answer Position**: Correct answer positioned at **a** or **b**.
  - **First Response**: Initial response is correct or mistaken.
- **Evaluation Metrics**:
  - **Baseline Confidence**: Confidence at Point A (initial response).
  - **Adjusted Confidence**: Confidence at Point B (after expressing doubt).
  - **Change in Confidence** (Δ Confidence): Difference between adjusted and baseline confidence.

Results and Discussion Figure 4 illustrates the average confidence for each model across the 6 confidence samples, separated by the initial response correctness. Figure 5 illustrate the distribution of the confidence in the second response, in relation to the initial response.

- **Size do matter**: We can see that that llama3.2 3B is almost always the better than the 1B model. And that high a postive sign for the correctness of the response.
- **Confidence Shifts**: We observed that models generally exhibited a change in confidence after expressing doubt.

3.5 Repeated Doubt with Feedback

Building on the results of previous experiments, which showed no consistent or significant improvement in model accuracy when doubt was introduced, this experiment investigates whether adding

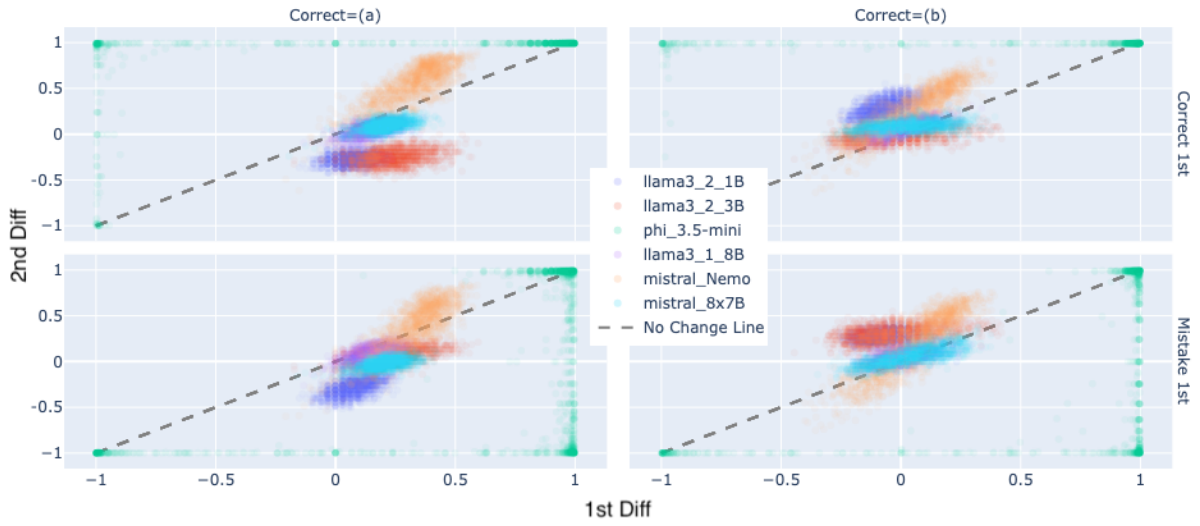

Figure 5: Model Confidence distribution on the baseline prompt, columns represent whether the correct answer was positioned first or second, and rows represent the models' initial response correctness.

feedback on the model's performance after expressing doubt, combined with iterative repetition, can enhance accuracy.

Experimental Setup The experimental setup was consistent with the first experiment, using the same set of models and evaluation methodology. The procedure was as follows:

1. The model was presented with factual questions, each with two possible answers. After selecting an answer, doubt was expressed regarding the model's choice.

2. After the model refined its answer in response to the expressed doubt, feedback was provided indicating whether its answer was correct both before and after the doubt stage.

3. This process was repeated over five iterations to observe whether performance improved over time.

To manage computational constraints, each model underwent 1,000 repetitions of this iterative process. To assess whether feedback influenced accuracy, we compared these results to a similar iterative process without performance feedback.

Results and Discussion The results of this experiment are presented in figure 6. In addition, an ANOVA test was conducted to asses the statistical significance of the effects of doubt, feedback, and iteration on accuracy. Table 4 summarizes the p-values for each factor across the tested models. Significant effects (p-value < 0.05) are highlighted in bold.

Model	Size	Doubt	Feedback	Iteration
Llama 3.2	1B	0.21	0.81	0.95
Llama 3.2	3B	**0.0001**	0.1	0.46
Phi 3.5	3.82B	0.72	0.17	**0.01**
Llama 3.1	8B	**0.0001**	**0.002**	**0.0001**
Mixtral	8x7B			
Nemo	12.2B	**0.02**	**0.03**	**0.01**

Table 4: P-values from ANOVA tests for the effects of doubt, feedback, and iteration on accuracy for each model. Bold values indicate significance (p-value < 0.05).

These results reveal varied responses to doubt, feedback, and iteration across different LLM architectures:

1. **Larger Llama Models (3B and 8B):**
   - Doubt negatively impacted accuracy. A possible reason for that may be that the doubt reduces model's confidence in its answers, thus confusing it.
   - Iterative questioning led to performance improvements at the pre-doubt stage (i.e before the doubt was induced), suggesting that a preliminary "warm-up" phase could be beneficial.
   - To test the necessity of doubt during warm-up, we conducted an additional experiment with Llama-8B, iteratively questioning it without inducing doubt. We chose to focus on Llama-8b because

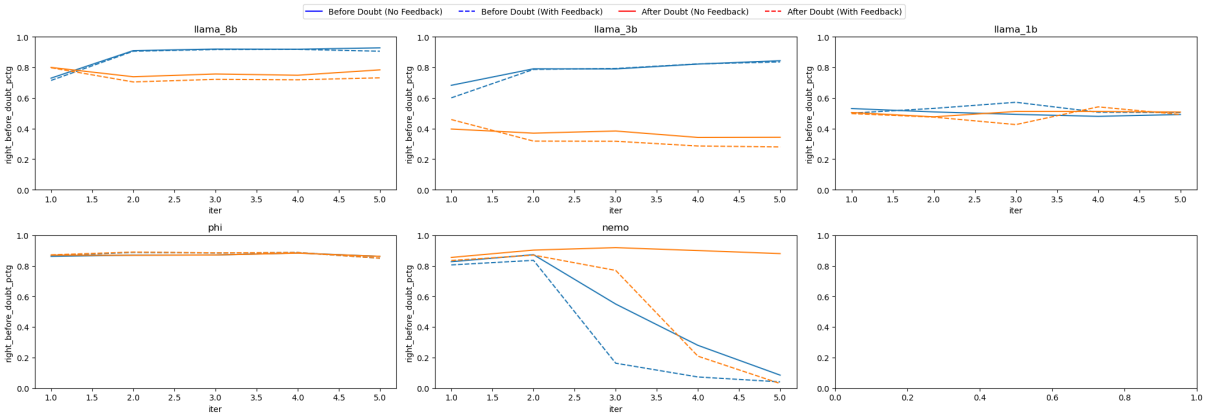

Figure 6: Model accuracy across iterations, separated by conditions: before/after doubt and with/without feedback.

table 4 shows that the effect of iterative questioning on accuracy is statisticaly significant for this model. The results (Figure 7) indicate that iterative questioning alone achieves similar improvements, showing that induced doubt is unnecessary in the suggested warm-up step.

2. **Stable Performance (Phi, Llama 1B):**
   - These models exhibited stable performance, unaffected by doubt, feedback, or iterative processes.
3. **Nemo Model:**
   - Doubt had no significant impact during the first iteration, but subsequent iterations revealed an intriguing pattern: accuracy decreased in the pre-doubt stage but improved post-doubt.
   - A possible explanation may be that the model anticipate the doubt prompt and intentionally adjust its answers. However, when feedback was added, performance degraded also after the doubt is induced.

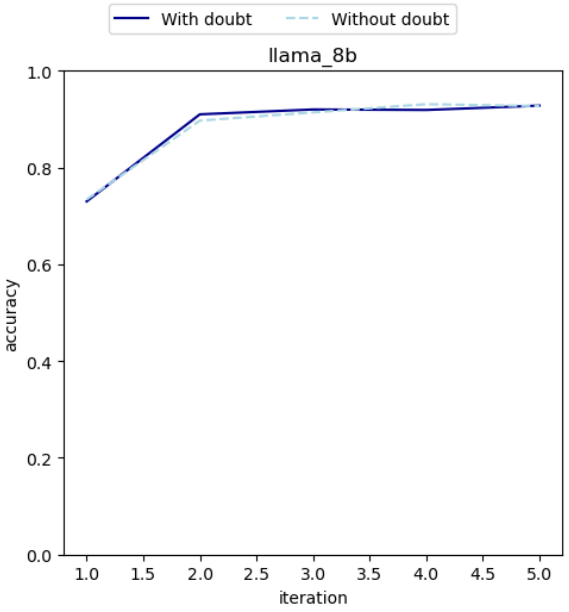

Figure 7: Llama-8B accuracy across iterations with and without induced doubt. For the experiment with doubt, accuracy before the doubt stage is reported, consistent with the blue line in Figure 6.

4 Discussion and Conclusion

4.1 Future Work

References

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Porte-

lance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Understanding the effects of iterative prompting on truthfulness. *arXiv preprint arXiv:2402.06625*.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, et al. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *arXiv preprint arXiv:2401.03729*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Rongwu Xu, Brian S Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2023. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*.