Always So Sure: Can LLM's Confidence be Trusted?

Anonymous authors
Paper under double-blind review

Abstract

Confidence estimation techniques are often used to better gauge the answers given by a Large Language Model (LLM). One such technique is *verbalized confidence*. This prompting setup produces confidence scores alongside the actual answers, but the mechanisms behind these self-reported confidence values remain poorly understood. This paper presents a comprehensive analysis of verbalized confidence across multiple datasets spanning factual questions, multiple-choice QA, and causal reasoning using four different LLMs. Our investigation reveals that verbalized confidence scores are *highly quantized*, clustering around specific values (e.g., 0, 90, 100) with minimal differentiation between correct and incorrect answers. Through causal mediation analysis and targeted input perturbations, we demonstrate that confidence score generation is primarily influenced by structural prompt elements like the word "confidence" and the specified scale range rather than the actual question's content. These findings provide valuable insights into the behavior of verbalized confidence and underscore the importance of developing more reliable self-evaluation mechanisms for LLMs.

1 Introduction

Imagine a person is asked what the capital city of a country is, but the country name is not spoken clearly but mumbled. The natural reaction is to either give an unsure or caveated answer or ask for clarification right away. Similar scenarios can happen when using large language models (LLMs), like a typo or out-of-distribution word in the user query. As LLMs are being increasingly deployed across a wide variety of tasks, including critical applications such as mental health (Xu et al., 2024; Stade et al., 2024) or legal advice (Cheong et al., 2024), it is important to be able to trust the outputs of these models, especially in situations like the ones described above. To gauge and potentially reinforce trustworthiness, one option is to provide the user with a confidence score for the generated answer, indicating how certain the model is about it. Such confidence scores could be obtained in different ways - analyzing token probabilities (Kumar et al., 2024), training proxy models to predict confidence (Stengel-Eskin et al., 2024; Tsai et al., 2024), evaluating semantic coherence between answers for paraphrased questions (Yang et al., 2024), or simply prompting the model to supply confidence scores. The latter — *verbalized confidence* — is a comparably lightweight and model-agnostic approach, which works without direct access to the model (Lin et al., 2022; Tian et al., 2023; Yang et al., 2024; Mahaut et al., 2024). To use verbalized confidence, the prompt can be amended to directly ask the model for its confidence; for example, "What is the capital of Italy?" becomes "What is the capital of Italy? Provide your confidence.". With this adaptation, while the model provides a confidence score, it is unclear if this self-reported score is meaningful or if it is trustworthy.

Prior work has already shown that verbalized confidence scores come with several inconsistencies (Xiong et al., 2024; Zhao et al., 2025), which are exacerbated by the sycophancy of LLMs (Sun & Wang, 2025). We briefly reexamine verbalized confidence across multiple datasets and tasks (factual questions, multiple-choice QA, and causal reasoning) with LLMs of different sizes and model families. We then analyze the relationship between the expressed confidence and the given question. For verbalized confidences of the model to be a reliable uncertainty estimation, the confidence must be closely intertwined with the question, or the understanding thereof. We build on mechanistic interpretability methods (Vig et al., 2020; Meng et al., 2022) and study *how* the model produces confidence scores. We utilize ideas from causal mediation

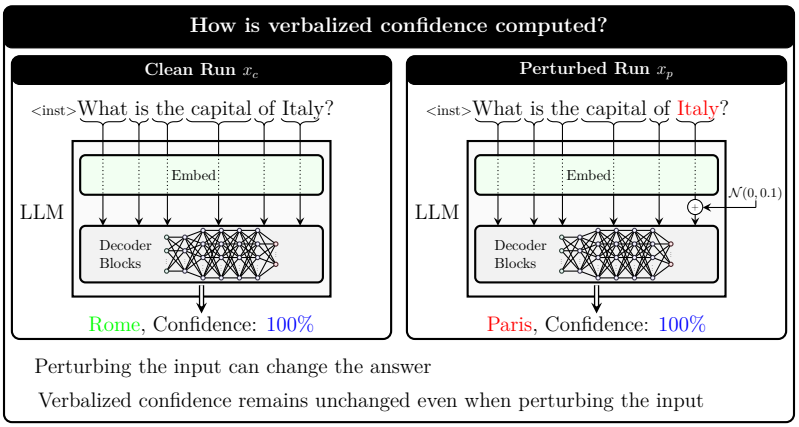


Figure 1: **Overview of our investigation into verbalized confidence. Left is a (simplified) illustration of the default setup and right is our intervention. In Section 4, we study the phenomenology of verbalized confidence scores, finding that the predicted scores are highly quantized. In Section 5, we perturb different parts of the input (e.g., in the prompt shown in the picture, the subject "Italy"). We do this to study how different parts of the input affect the predicted answers and confidence scores. `<inst>` is the instruction given in the prompt, passed through the model prepended to the question.**

analysis (Pearl, 2001) to intervene on the input representations.We analyze two methods of intervention. First, replacing words on the token level, and second, adding noise to the embedding level. For both our intervention methods, the tested models show a similar behavior, i.e., similar response formatting and overconfidence in their responses. Both intervention methods represent real-world scenarios, like typos, OCR artifacts, or out-of-distribution inputs. These scenarios do not cause a model collapse and may slip by the user unnoticed. However, a trustworthy model should be able to detect them. Through this analysis, we aim to better understand verbalized confidence and decide whether to trust the predicted scores. An illustrative overview of our proposed approach is presented in Figure 1.

Note that our study does not attempt to force the expressed confidence scores to match a ground-truth value (i.e., calibrate the model). This target is inherently elusive since the model does not have direct access to an external "correctness" signal. Instead, we investigate the process by which LLMs produce their verbalized confidence scores. Focusing on the production process rather than post-hoc alignment (Kumar et al., 2024), our work investigates the trustworthiness of these default linguistic expressions of uncertainty.

Contributions. Our analysis shows that verbalized confidence values are highly discrete, with a small range of predicted confidence scores such as 0, 90, or 100 (also observed in (Xiong et al., 2024) albeit for commercial models), irrespective of the correctness of the given answers. Finally, we perturb parts of the input to obfuscate the meaning for the model and find that the actual question is largely irrelevant for producing confidence scores. Solely relevant for verbalized confidence scores is the word "confidence" as well as the scale on which they should be predicted (e.g., "0-100"). These results show the pitfalls of this expression of confidence and call into question the suitability and trustworthiness of verbalized confidence as a form of self-evaluation.

2 Related Work

Measuring Confidence. A considerable amount of research has focused on how to estimate and improve confidence in language models (Wei et al., 2022; Yao et al., 2023; Wiegreffe et al., 2021; Marasovic et al., 2022; Zhou et al., 2024; Li et al., 2024; Shorinwa et al., 2026; Geng et al., 2024). Early investigations, e.g., Jiang et al. (2021), examined calibration for question answering. Subsequent studies have extended this work to conversational agents and multi-answer settings (Mielke et al., 2022; Li et al., 2024). The main methods to quantify model confidence are: (i) Token Probabilities: Relying on log or softmax probabilities

as a baseline for confidence estimation (Kumar et al., 2024; Liu et al., 2025). (ii) Semantic Coherence: Employing heuristics such as semantic alignment between the prompt and the answer to gauge certainty (see, e.g., Section 2 in Yang et al. (2024)). (iii) Verbalized Confidence: Directly eliciting uncertainty via natural language expressions—a line of work initiated by Lin et al. (2022) and further refined in subsequent studies (Tian et al., 2023). Verbalized confidence has been found to imitate human patterns of confidence and also depends on the authority of the response (e.g., response by a student vs. the model itself) (Xiong et al., 2024; Zhao et al., 2025). These findings are also applicable to reasoning models (Fu et al., 2025). (iv) Proxy Models: Utilizing auxiliary models to predict or adjust confidence scores, as demonstrated by Stengel-Eskin et al. (2024) and Tsai et al. (2024).

Calibration and Ground-Truth Alignment. Calibration or alignment aims to adjust the confidence of the model to its correctness (Müller et al., 2019). This alignment can focus on internal alignment, i.e., aligning the entropy of output distributions (a proxy for confidence) to the correctness of the model (Xie et al., 2024; Tripathi et al., 2025; Li et al., 2025). Another line of research aims to align the expressed confidence scores with external measures of correctness. For instance, Mielke et al. (2022), Kadavath et al. (2022) and Tian et al. (2023) focus on calibrating the output confidence — ensuring that the model's verbalized uncertainty corresponds more closely with ground-truth accuracy. Such calibration efforts are crucial for enhancing the reliability and interpretability of LLM outputs. Relatedly, Kumar et al. (2024); Yona et al. (2024); Ghafouri et al. (2024) analyze the link between the linguistically expressed confidence (e.g., "I'm sure of X") and the internal model confidence, measured by token probabilities. Another work (Zhao et al., 2024) focuses on prompt engineering to elicit more calibrated confidence score expressions.

Mechanistic Interpretability and Causal Mediation Analysis in LLMs. Parallel to the line of work on calibration, there has been significant progress in mechanistic interpretability. Recent studies have endeavored to reverse-engineer internal model processes or "circuits" to better understand how language models function. For example, recent studies have shed light on how GPT-2 handles greater-than computations (Hanna et al., 2023) or the function of feed-forward layers as key-value stores in language models (Geva et al., 2021). For a general overview of mechanistic interpretability, we point to the survey by Bereska & Gavves (2024).

Most closely related to our investigation are those studies that use causal mediation analysis (Pearl, 2001) to pinpoint where and how certain information is processed within the model. By intervening on gender-biased input tokens, Vig et al. (2020) study the effect of gender bias on language modeling. Similarly, Meng et al. (2022) perturb and restore input tokens with the goal of localizing the factual recall in language models. This perturbation approach has been used and extended in several follow-up works (Meng et al., 2023; Hase et al., 2023; Alpay & Alpay, 2025; Guo et al., 2025). We study which parts of the input ultimately contribute to the final verbalized confidence.

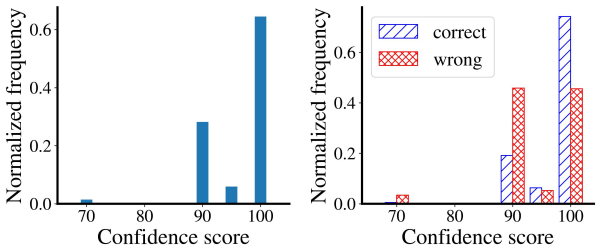
Present work. Most of the existing work focuses on aligning the confidence scores with the external ground-truth correctness – ensuring that an answer with a high confidence score is also highly likely to be correct (Stengel-Eskin et al., 2024; Mielke et al., 2022; Tian et al., 2023). Instead, our work is situated at the intersection of confidence estimation and mechanistic interpretability, emphasizing our unique focus on the production of verbalized confidence scores. Moreover, instead of a general overview, we take a close look into a specific setting (i.e., prompting the model to provide its confidence directly), allowing in-depth analysis of the inner workings of verbalized confidence computation.

3 Experimental Setup

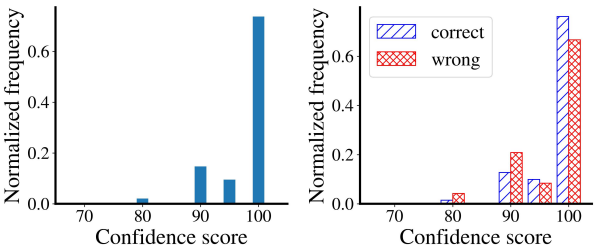
For our analysis, we consider a diverse range of datasets as well as LLMs of different sizes.

3.1 Datasets

We consider different types of QA datasets, including factual and knowledge-intensive questions and a text classification dataset. While the factual questions can be answered from the LLM's parametric knowledge, others would require the model to utilize its reasoning ability.



(a) The confidence distribution for Phi-4 on factual questions in total (left) and segregated by whether the generated response is correct or not (right).



(b) The confidence distribution for Llama3.3 70B on MMLU anatomy questions in total (left) and segregated by whether the generated response is correct or not (right).

Figure 2: **Overview of verbalized confidence for Phi-4 (Figure 2a) and Llama3.3 70B (Figure 2b).**

Factual questions. We construct factual questions using a counterfactual dataset created by Meng et al. (2022). The entries in the original dataset[1] consist of prompts, subjects, and objects, and they are tagged with Wikidata relations (e.g., "works for"). We use these relations to rephrase the prompts as questions (e.g., "Who does X work for?"). We have 20,942 questions in total, of which we sample a subset of 1000 questions. The LLM is supposed to generate an answer to the question and the confidence for that answer.
MMLU. We apply our approach to multiple-choice questions from MMLU (Hendrycks et al., 2021a;b). We use the test split of the subtasks high school geography (198 questions), astronomy (152 questions), anatomy (135 questions), and college medicine (173 questions).
BIG-bench. Lastly, we also use the empirical judgments from BIG-bench (bench authors, 2023) with 99 examples. The task is to determine whether two events in a given sentence have a causal or correlative relationship, or neither. This can be considered a 3-class text classification task, with the model expected to predict the correct class.

3.2 Models

We consider autoregressive LLMs of different sizes and model families. Specifically, we consider Llama3.2 3B-Instruct, Llama3.1 8B-Instruct, and Llama3.3 70B-Instruct, instruction fine-tuned Llama3 (Dubey et al., 2024) models with 3, 8, and 70 billion parameters, respectively. We also use Phi-4 14B, a 14 billion parameter model released by Microsoft (Abdin et al., 2024). We use default hyperparameters for all models.

3.3 Prompts

We use a short and simple prompt for question answering, similar to prompts used in previous works (Xiong et al., 2024; Tian et al., 2023). We use different prompts for each dataset to best comply with the task. The prompts all specify how the model should answer and that the model should provide a confidence score between 0 and 100. The answer that the model gives depends on the task. See Appendix A for more details.

4 Characterizing Verbal Confidence

In this section, we perform an empirical analysis of verbal confidence on various tasks and encompassing LLMs of different sizes.

[1] `https://rome.baulab.info/data/dsets/`

4.1 Performance

As a preliminary analysis, we report the performance of different models on the tasks in Table 1. The performances are noted for zero-shot settings. Overall, larger models do better than smaller ones, at least for the MMLU tasks. All models perform poorly for the empirical judgments dataset.

	Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
Factual Qs	51%	62%	63%	70%
MMLU geo	70%	56%	90%	83%
MMLU ast	66%	76%	92%	89%
MMLU ana	55%	62%	73%	71%
MMLU med	55%	62%	79%	74%
BIG-bench	49%	34%	41%	43%

Table 1: **Accuracy of the LLMs for different tasks in a zero-shot setting.**

4.2 Confidence score distribution

We probe the confidence score distribution across different tasks. In Figure 2a (Left) and Figure 2b (Left), we plot the frequency distribution for confidence scores generated by the Phi-4 and Llama3.3 70B models, respectively, for the factual and MMLU anatomy questions.

We note that the confidence scores are quantized, i.e., the scores are only 70, 80, 90, 95, or 100 in this case. This is reminiscent of how humans generally assign confidence values to future events, tending to assign confidence scores in steps of 5 or 10 (highly effective forecasters may assign more fine-grained scores (e.g., 87%)) (Mellers et al., 2015). This behavior is consistent across models of different sizes.

Additionally, we find the confidence scores to always be high, which points to its overconfidence regarding content generated by itself. This is also in line with findings reported for commercial models in Xiong et al. (2024). However, we further notice the confidence scores to be high irrespective of whether the generated answer is correct or wrong. In Figure 2a (Right) and Figure 2b (Right), we present the frequency distribution of the confidence scores segregated based on whether the generated answer is correct or wrong. It can be observed that the distributions are similar, with the model providing high confidence scores even for the cases where it generates wrong answers.

Lastly, the behavior is also consistent across LLMs and tasks. The factual questions can only be answered from a model's parametric knowledge (i.e., either the model gets the correct answer or it does not). However, we still observe the LLMs generating different confidence scores for different questions (i.e., topics), which might indicate the model being more confident about certain topics than others. In the case of other datasets (e.g., BIG-bench empirical judgments), the model needs to utilize its reasoning ability to answer the question, and the answer cannot be directly retrieved from parametric knowledge. Irrespective of this difference, LLMs tend to exhibit similar behavior regarding their confidence in the generated answer. We further find that the verbalized confidence does not change when the model is prompted multiple times with the same input. We provide further details in Appendix D.

When answering a question, humans interpret confidence scores as a measure of our uncertainty about the question or the topic in general. Confidence is directly related to the question itself. Given that LLMs generate overly high confidence scores irrespective of whether the generated answer is right or wrong, we ask, can these confidences be trusted, and is the generated confidence score at all associated with the question asked?

5 Confidence in what?

As we have seen above, LLMs are overconfident, and it is unclear how these confidence scores are generated by the LLMs. This ambiguity makes it difficult to trust these generated confidence scores. Hence, we proceed

to perform a more in-depth analysis to gain further insights into the origin of the expressed confidence scores. In particular, we hypothesize that the generated confidence scores should be highly correlated with the question/topic, which also reflects how humans interpret confidence.

We take inspiration from mechanistic interpretability methods and adapt them to our investigation. We closely align with causal mediation analysis (Pearl, 2001), which investigates the effect of an independent variable on a dependent variable via an intermediate variable. In the context of LLMs, causal mediation analysis is used to quantify the influence the input tokens have on the output through the flow of information in internal model states. The key idea is to perturb the input, either by adding noise (Meng et al., 2022) or replacing the tokens in the input directly (Vig et al., 2020) and measure its impact on the output. We elaborate on the exact method in the following.

5.1 Method

We intend to investigate whether the confidence score generated by an LLM for a given question is related to the question itself.

We consider perturbing the embeddings of the tokens corresponding to the question subject and observe the patching effect. In particular, for a given question, we have a clean forward pass x_c, without any intervention, and a perturbed forward pass x_p where we intervene and perturb the subject token embeddings. This procedure is similar to activation patching, as discussed by Meng et al. (2022); Zhang & Nanda (2024), although one important distinction is that we do not aim to recover the answer but directly investigate the effect of our intervention on the model output. Furthermore, we extend this approach to question answering instead of single next token prediction. In our setup, we can assume that both x_c and x_p return an answer and a confidence value, as all other cases are filtered out (see below for more discussion on this point). The patching effect is measured as the change in confidence scores and accuracy of the model between x_c and x_p, denoted as $x_c \rightarrow x_p$. For a given question, we have three possible outcomes for $x_c \rightarrow x_p$: (i) **both** the answer and the confidence score change, (ii) **either** the answer or the confidence changes, and (iii) **neither** the answer nor the confidence changes. If it is the case that the model's confidence is indeed related to the given question, we would ideally see only outcomes of types (i) and (iii). For case (i), the perturbation changes the model's understanding and the given answer and the confidence is adjusted accordingly. For case (iii), the answer remains unchanged, and thus the confidence does not change either. Case (ii), on the other hand, means the confidence is unrelated to the confidence, as the answer and confidence change independently of one another. The proposed method is also illustrated with an example of type (ii) in Figure 1.

5.1.1 Selecting perturbation tokens

We intend to minimally corrupt the model's notion of the question topic to provoke a change of the answer in x_p. For this, we identify subject tokens in the prompt. For the example in Figure 1, "Italy" is the subject token. To automatically extract the subject from a given question, we use GLiNER (Zaratiana et al., 2024), a bidirectional transformer trained to identify entities. We remove any question where no entities are extracted or the quality score of the extraction is low (< 0.8). This removes all questions with no clear subject (e.g., "Which of the following statements is true?" for multiple-choice questions).

5.1.2 Perturbation method

There are two possible ways of perturbing the model's input: either replacing the subject tokens with a counterfactual or adding noise to the embeddings of the subject tokens. While replacing subject tokens is a direct and traceable approach, it is only possible for factual questions. Multiple-choice question counterfactuals require adapting the answer possibilities, dramatically increasing the complexity. On the other hand, perturbing subjects by adding noise is a straightforward and easy approach that can be applied to all question types. However, this approach is not as traceable as replacing tokens.

As adding noise is a broader approach, we investigate how comparable this approach is to replacing the entities with counterfactuals. For replacing the subject, we select a random entity sampled from the factual

questions dataset Meng et al. (2022) and replace the word before tokenization. While sampling, we ensure that the entity belongs to a different relation so that the question is nonsensical. For example, given the question "What country is Feng Tianwei a citizen of?", with "Feng Tianwei" being the subject, we sample the entity "aloha" and replace it to form the (nonsensical) question "What country is aloha a citizen of?". For noise perturbation, we add random noise sampled from a normal distribution ($\mathcal{N}(0, 0.1)$) to the subject. Specifically, given an input question $q = q_1 q_2 ... q_T$ with a token length of T, we obtain the embedding of q:

$$E_q = [E_{q_0}, E_{q_1}, \ldots, E_{q_T}] \tag{1}$$

We set $E_{q_j} = E_{q_j} + \epsilon; \epsilon \sim \mathcal{N}(0, 0.1)$ for all tokens q_j that correspond to the subject in q (see Appendix B for more details).

After perturbing the input with either strategy, we pass it through the model, obtaining x_p. With our intervention, the question is nonsensical and/or the model can no longer ascertain what the question is about, i.e., we move from x_c to x_p. As a consequence, the confidence distribution should shift greatly towards the lower end of the spectrum, if the reported confidence is truly related to the question and model's knowledge.

We now compare the two perturbation methods. First, we provide the average confidence scores of answers (regardless of correctness) the model provides for both perturbation types, see Table 2. The average confidence is high, being close to or above 90%, regardless of perturbation type and model. The at-times high standard deviation is explained by occasional confidences of 0 that skew the otherwise high confidence values.

Model	None	Replace	Noise
Llama3.2 3B	92.23%$_{(\pm 10.42)}$	90.08%$_{(\pm 13.84)}$	89.63%$_{(\pm 14.08)}$
Llama3.1 8B	90.49%$_{(\pm 9.63)}$	86.82%$_{(\pm 14.32)}$	86.96%$_{(\pm 11.28)}$
Phi-4	96.13%$_{(\pm 6.25)}$	93.98%$_{(\pm 13.02)}$	94.49%$_{(\pm 12.09)}$
Llama3.3 70B	94.94%$_{(\pm 10.98)}$	88.50%$_{(\pm 22.49)}$	97.36%$_{(\pm 8.88)}$

Table 2: **Verbalized confidence with different perturbation strategies averaged over** $1,000$ **examples.**

We further count how often the model refuses to answer or asks for clarification, shown in Table 3. For both replacement and adding noise, the number of refusals increases in comparison to non-perturbed response. We expect a good model to refuse the answer in the perturbed setting. The high refusal count when adding noise, while having an overly high confidence in non-refused questions, is a good setting for our experiments.

Model	None	Replace	Noise
Llama3.2 3B	40	139	146
Llama3.1 8B	6	54	6
Phi-4	25	74	374
Llama3.3 70B	1	14	1

Table 3: **Refusal count with different perturbation strategies across** $1,000$ **examples.**

Lastly, we analyze the similarity between the responses given when replacing the subjects versus adding noise to them. Qualitatively, we find that answers are very similarly formatted for each model. The answers are usually formatted as "`<answer>, <number>`" (e.g., "Hindi, 90") or "`<answer>. Confidence: <number>`" (e.g., "Polish. Confidence: 100"). Especially for Phi-4 and Llama3.3 70B this formatting is very consistent. To quantitatively analyze this similarity, we calculate the similarity score between the answers given when adding noise and the answers when replacing the subject tokens. To compute the similarity, we use the Python `difflib` library, specifically `SequenceMatcher`[2]. The similarity is calculated based on matching trigrams and and is 1.0 for identical sequences and 0.0 for no overlap. We report the similarity in Table 4. Note that the "`<answer>`" part usually is different, as the replacing of tokens entirely changes the subject of the question. With this consideration, the similarity of around or above 0.5 for 3 of the 4 models is rather high, meaning the formatting of the answers for both perturbation methods is rather similar. We further see

[2]`https://docs.python.org/2/library/difflib.html#difflib.SequenceMatcher.ratio`

that the formatting of the answer when adding noise is closer to the original answer than when replacing the tokens. This shows that adding noise is a smaller change than replacing tokens that still provokes a change in the answer.

Model	O vs. R	O vs. N	R vs. N
Llama3.2 3B	0.33	0.34	0.33
Llama3.1 8B	0.46	0.46	0.44
Phi-4	0.59	0.69	0.56
Llama3.3 70B	0.60	0.60	0.61

Table 4: **Similarity of answers between two strategies across $1,000$ examples, where O is the original, i.e., no perturbation, R is replacing the tokens, and N is adding noise.**

Nevertheless, adding noise is not as traceable as replacing tokens. Adding noise may shift the token embeddings into an entirely different part of the embedding space. To ensure that our change is not drastic, we check how often the original embedding E_o of entity e is in the k-nearest neighborhood of the perturbed embedding E_p. For this, we calculate the Euclidean distance between E_p and all embeddings in the embedding matrix of the respective model and check if E_o is in the k token embeddings with the lowest distance. Besides Euclidean distance, we also compute similarity using cosine similarity. For all models, E_p is most similar to E_o using cosine similarity in all cases. For Euclidean distance, the frequency of the occurrences is shown in Table 5. We see that in the large majority of cases, the original embedding E_o is the most similar embedding, signifying that our change is in fact minimal. There is some discrepancy for Llama3.3 70B. Most likely, this is caused by the larger dimensionality of the embeddings in Llama3.3 70B. Llama3.3 70B has a embedding dimension of $8,192$, while the other models have dimensions of $3,072$ (3B), $4,096$ (8B), and $5,120$ (14B). Nevertheless, two-thirds of the time E_o is most similar.

k	Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
1	100.00%	97.88%	100.00%	67.29%
5	100.00%	99.57%	100.00%	76.64%
10	100.00%	99.57%	100.00%	80.03%
20	100.00%	100.00%	100.00%	88.72%

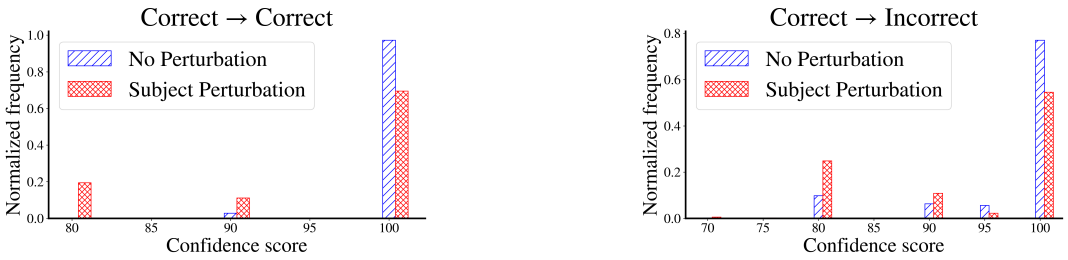
Table 5: **Percentage of original entity embeddings that are in the k-nearest neighborhood to the perturbed embeddings.**

Naturally, this is not always the case, showing that these embeddings are, if slightly, out-of-distribution. We argue that this is a realistic case, as LLMs are in practice also used with slightly out-of-distribution inputs that they should handle appropriately. For example, models might get inputs with typos, OCR artifacts, or adversarial perturbations. A good model should be able to handle these cases correctly. We also explicitly prompt the model with random letters in place of the entity tokens, e.g., we replace "Italy" with "qWbJa". In these cases the model either, correctly, does not respond and requests further information or, incorrectly, collapses and generates gibberish. Adding noise does not cause such a collapse, as the model provides an answer and high confidence. Noise perturbation is thus not a dramatic change in comparison.

All in all, we can say that perturbing by adding noise is on par with replacing entity tokens with counterfactuals. Adding noise is a small intervention that is easy to apply in all cases but is sufficient to influence the model's answer without causing a collapse. Further, this minor intervention reflects real-world scenarios in which a model should appropriately respond. Therefore, we use added noise as the perturbation of choice for further experiments.

5.1.3 Procedure

For each question, we prompt the model twice. First, we do not perturb anything (x_c) and get the default response. Secondly, we perturb the subject on the embedding level (x_p). In both runs, we ask the model to provide the answer and confidence in the response (see Section 3.3). We split the model's responses for x_c into three categories: correct answer given, incorrect answer given, and refusal to answer. We further differentiate between answers that remain correct, change to incorrect, or lead to a refusal to answer in x_p.



(a) (b)

Figure 3: **The confidence distribution of verbalized confidence given by Llama3.3 70B on factual questions. We show the difference between the confidence with and without subject perturbation, and between questions where the answer remains correct or changes after perturbation. Figure 3a shows the shift of verbalized confidence between no and added perturbation on questions where the answer does not change. Figure 3b shows the same shift, but for questions where the answer does change.**

We remove questions from further analysis that are answered incorrectly or are refused to answer in an unperturbed filtering run, as the model cannot provide an accurate confidence for these questions to begin with. See Appendix C for more details on the filtering. Questions that are no longer answered after perturbation are removed as well. We end up with two sets of questions: questions that are answered correctly with and without perturbation and questions that are answered correctly without perturbation but are answered incorrectly with perturbation.

5.2 Results

In this described setup, we analyze how the confidence changes when perturbing the subject for 1000 questions sampled from the factual question dataset. Phi-4 provides a correct answer for all of those questions and has an average confidence of 98.32%. When perturbing the subject, Phi-4 answers 64 questions correctly and has an average confidence of 98.75%. Questions that are no longer answered correctly (737 questions) have an average confidence of 97.54%. The remaining responses are a refusal to answer or do not provide a confidence score. For Llama3.3 70B, 946 out of 1000 questions are answered correctly with an average confidence of 97.04%. With subject perturbation, 36 questions are still answered correctly with an average confidence of 95%. 592 questions are answered incorrectly and have an average confidence of 94.66%. We provide an overview of $x_c \rightarrow x_p$ for the confidence distribution in Figure 3. Again, the remaining responses are a refusal to answer or do not provide a confidence score. We provide an overview of all models on factual questions and the average confidence in Table 6. Additionally, we apply a subset of questions on a reasoning model (Appendix E) and find the same discrete confidence values as with non-reasoning models.

The same observations hold for MMLU. For example, in MMLU astronomy, Phi-4 answers 83 questions correctly with an average confidence of 96.93%. Of these 83 questions, 69 remain correctly answered after perturbation, with an average confidence of 96.81%. 14 questions are answered incorrectly with an average confidence of 93.93%. We provide an overview of $x_c \rightarrow x_p$ for the confidence distribution in Appendix F. Llama3.3 70B answers 94 questions correctly with an average confidence of 98.56%. Adding noise to the subject, 48 answers are still correct with an average confidence of 95.83%. 28 questions are not answered correctly with an average confidence of 98.21%. Results for all models on MMLU astronomy and all other datasets are presented in Appendix F.

We also confirm our analysis using a chain-of-thought (Wei et al., 2022) prompt on a subset of 100 factual questions. We observe a very similar distribution of confidence: a high confidence overall and little to no shift between the confidences given with and without subject perturbation. We provide plots of the confidence distributions in Appendix F.

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	832	942	1000	946
	Avg. confidence	$94.66\%_{\pm 9.46}$	$93.15\%_{\pm 8.57}$	$98.32\%_{\pm 4.15}$	$97.04\%_{\pm 6.80}$
Subject perturbation	# correct answers	27	47	64	36
	Avg. confidence, correct	$88.74\%_{\pm 18.66}$	$87.66\%_{\pm 9.78}$	$98.75\%_{\pm 5.45}$	$95.00\%_{\pm 7.99}$
	# incorrect answers	438	600	737	592
	Avg. confidence, incorrect	$82.34\%_{\pm 25.90}$	$86.89\%_{\pm 16.94}$	$97.54\%_{\pm 10.59}$	$94.66\%_{\pm 15.81}$

Table 6: **Results of all models on factual questions, showing the number of correct/incorrect answers and average confidence with and without perturbation. "Avg. confidence, correct" is the average confidence given in responses that also contain the correct answer ("Avg. confidence, incorrect" is the opposite case). Note that we do not include questions that the model refuses to answer in the average confidence calculation. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.**

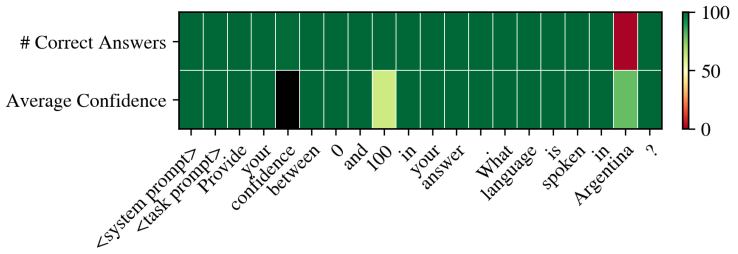


Figure 4: **Overview of the number of correct answers across 100 runs (top) and the average verbalized confidence in these runs (bottom) when perturbing each token in the input separately with Phi-4. The color black signifies that no confidence scores were reported in all 100 runs for that token.**

5.3 What leads to a change in confidence?

The results above indicate that *the generated confidence score is generally unrelated to the prompt question/topic*, which may indicate that the verbalized confidence is untrustworthy and it does not truly represent the underlying uncertainty of the model.

As a further analysis, we investigate if there are *any* tokens that, when perturbed, lead to a change in confidence. To this end, we separately perturb all the tokens in the prompt and measure their impact. We illustrate with an example question, "What language is spoken in Argentina?" (we make similar observations for others). We perturb each token individually and also prompt the model 100 times for each such perturbation. We report the number of correct answers and the average confidence for these answers in Figure 4. We can see that no tokens have much influence on the number of *correct answers* except the subject, in this case, "Argentina." For the average confidence, the only tokens that have significant influence are "confidence" and "100". When perturbing the token "confidence", the model does not report any confidence scores. When perturbing the token "100", the model does report confidence scores, but instead of a score between 0 and 100, it uses a different scale (e.g., 0 to 1 or 0 to 10). These different scales decrease the average confidence. When shifting the differing scales to fit the 0 to 100 scale, the confidence when perturbing the token "100" increases to 81.78%. Lastly, perturbing the token "Argentina" results in only one correct response with a confidence score of 80%. For the incorrect responses, the average confidence remains high at 96%.

To summarize, we find that *most of the input tokens have little influence on the reported confidence.* When perturbing the subject, however, the model can no longer consistently respond correctly, but even though the given answers are incorrect, the model remains confident.

6 Conclusion

In this paper, we investigated the notion of verbalized confidence in LLMs, whereby we directly prompt the LLM to generate a confidence score for a generated answer. We observed the generated confidence scores to be highly quantized and the scores to be high, irrespective of whether the generated answer is correct or wrong. To gain deeper insights, we deployed a causal mediation analysis-inspired approach and observed that the generated scores are also not associated with the question or the topic, which raises questions about the trustworthiness of such verbalized confidence scores.

Broader Impact Statement

We study the brittle behavior of verbalized confidence for LLMs. We do not believe this is a suitable attack vector to achieve harmful behavior. While not directly derivable from this work, it might be possible to use adversarial attacks to intentionally bias the confidence estimation of LLMs, given the brittle behavior showcased in this study. If successful, this could have downstream effects on LLM reasoning and the perceived utility of such models.

References

Marah I Abdin, Jyoti Aneja, Harkirat S. Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *CoRR*, abs/2412.08905, 2024. doi: 10.48550/ARXIV.2412.08905. URL https://doi.org/10.48550/arXiv.2412.08905.

Faruk Alpay and Taylan Alpay. Manipulating transformer-based models: Controllability, steerability, and robust interventions. *CoRR*, abs/2509.04549, 2025. doi: 10.48550/ARXIV.2509.04549. URL https://doi.org/10.48550/arXiv.2509.04549.

BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for AI safety - A review. *CoRR*, abs/2404.14082, 2024. doi: 10.48550/ARXIV.2404.14082. URL https://doi.org/10.48550/arXiv.2404.14082.

Inyoung Cheong, King Xia, K. J. Kevin Feng, Quan Ze Chen, and Amy X. Zhang. (a)i am not a lawyer, but...: Engaging legal experts towards responsible llm policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, pp. 2454–2469, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3659048. URL https://doi.org/10.1145/3630106.3659048.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Tairan Fu, Javier Conde, Gonzalo Martínez, María Grandury, and Pedro Reviriego. Multiple choice questions: Reasoning makes large language models (llms) more self-confident even when they are wrong. *CoRR*, abs/2501.09775, 2025. doi: 10.48550/ARXIV.2501.09775. URL https://doi.org/10.48550/arXiv.2501.09775.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 6577–6595. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.366. URL https://doi.org/10.18653/v1/2024.naacl-long.366.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.EMNLP-MAIN.446. URL https://doi.org/10.18653/v1/2021.emnlp-main.446.

Bijean Ghafouri, Shahrad Mohammadzadeh, James Zhou, Pratheeksha Nair, Jacob-Junqi Tian, Mayank Goel, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. Epistemic integrity in large language models. *CoRR*, abs/2411.06528, 2024. doi: 10.48550/ARXIV.2411.06528. URL https://doi.org/10.48550/arXiv.2411.06528.

Phillip Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/guo25k.html.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/3927bbdcf0e8d1fa8aa23c26f358a281-Abstract-Conference.html.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 1049–1065. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.67. URL https://doi.org/10.18653/v1/2023.findings-acl.67.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know *When* language models know? on the calibration of language models for question answering. *Trans. Assoc. Comput. Linguistics*, 9:962–977, 2021. doi: 10.1162/TACL_A_00407. URL https://doi.org/10.1162/tacl_a_00407.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *CoRR*, abs/2207.05221, 2022. doi: 10.48550/ARXIV.2207.05221. URL https://doi.org/10.48550/arXiv.2207.05221.

Abhishek Kumar, Robert Morabito, Sanzhar Umbet, Jad Kabbara, and Ali Emami. Confidence under the hood: An investigation into the confidence-probability alignment in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 315–334. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.20. URL https://doi.org/10.18653/v1/2024.acl-long.20.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 11858–11875. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-emnlp.693.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *CoRR*, abs/2508.18847, 2025. doi: 10.48550/ARXIV.2508.18847. URL https://doi.org/10.48550/arXiv.2508.18847.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pp. 6107–6117. ACM, 2025. doi: 10.1145/3711896.3736569. URL https://doi.org/10.1145/3711896.3736569.

Matéo Mahaut, Laura Aina, Paula Czarnowska, Momchil Hardalov, Thomas Müller, and Lluís Màrquez. Factual confidence of llms: on reliability and robustness of current estimators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 4554–4570. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.250. URL https://doi.org/10.18653/v1/2024.acl-long.250.

Ana Marasovic, Iz Beltagy, Doug Downey, and Matthew E. Peters. Few-shot self-rationalization with natural language prompts. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 410–424. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-NAACL.31. URL https://doi.org/10.18653/v1/2022.findings-naacl.31.

Barbara Mellers, Eric Stone, Terry Murray, Angela Minster, Nick Rohrbaugh, Michael Bishop, Eva Chen, Joshua Baker, Yuan Hou, Michael Horowitz, Lyle Ungar, and Philip Tetlock. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015. doi: 10.1177/1745691615577794. URL https://doi.org/10.1177/1745691615577794. PMID: 25987508.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho,

and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex J. Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=MkbcAHIYgyS.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Trans. Assoc. Comput. Linguistics*, 10:857–872, 2022. doi: 10.1162/TACL_A_00494. URL https://doi.org/10.1162/tacl_a_00494.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 4696–4705, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

Judea Pearl. Direct and indirect effects. In Jack S. Breese and Daphne Koller (eds.), *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, pp. 411–420. Morgan Kaufmann, 2001.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck. Reasoning with large language models, a survey. *CoRR*, abs/2407.11511, 2024. doi: 10.48550/ARXIV.2407.11511. URL https://doi.org/10.48550/arXiv.2407.11511.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Comput. Surv.*, 58(3):63:1–63:38, 2026. doi: 10.1145/3744238. URL https://doi.org/10.1145/3744238.

Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12, 2024.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: listener-aware finetuning for confidence calibration in large language models. *CoRR*, abs/2405.21028, 2024. doi: 10.48550/ARXIV.2405.21028. URL https://doi.org/10.48550/arXiv.2405.21028.

Yuan Sun and Ting Wang. Be friendly, not friends: How LLM sycophancy shapes user trust. *CoRR*, abs/2502.10844, 2025. doi: 10.48550/ARXIV.2502.10844. URL https://doi.org/10.48550/arXiv.2502.10844.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 5433–5442. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.330. URL https://doi.org/10.18653/v1/2023.emnlp-main.330.

Sahil Tripathi, Md Tabrez Nafis, Imran Hussain, and Jiechao Gao. The confidence paradox: Can LLM know when it's wrong. *CoRR*, abs/2506.23464, 2025. doi: 10.48550/ARXIV.2506.23464. URL https://doi.org/10.48550/arXiv.2506.23464.

Yao-Hung Hubert Tsai, Walter Talbott, and Jian Zhang. Efficient non-parametric uncertainty quantification for black-box large language models and decision planning. *CoRR*, abs/2402.00251, 2024. doi: 10.48550/ARXIV.2402.00251. URL https://doi.org/10.48550/arXiv.2402.00251.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. Causal mediation analysis for interpreting neural NLP: the case of gender bias. *CoRR*, abs/2004.12265, 2020. URL https://arxiv.org/abs/2004.12265.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. Measuring association between labels and free-text rationales. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10266–10284, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.804. URL https://aclanthology.org/2021.emnlp-main.804/.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 18128–18138. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1007. URL https://doi.org/10.18653/v1/2024.emnlp-main.1007.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gjeQKFxFpZ.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James A. Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1):31:1–31:32, 2024. doi: 10.1145/3643540. URL https://doi.org/10.1145/3643540.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. *CoRR*, abs/2412.14737, 2024. doi: 10.48550/ARXIV.2412.14737. URL https://doi.org/10.48550/arXiv.2412.14737.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 7752–7764. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.emnlp-main.443.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. Gliner: Generalist model for named entity recognition using bidirectional transformer. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 5364–5376. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.300. URL https://doi.org/10.18653/v1/2024.naacl-long.300.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew C. Yao. Cumulative reasoning with large language models. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=grW15p4eq2.

Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. Evaluating large language models through role-guide and self-reflection: A comparative study. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=E36NHwe7Zc.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. Fact-and-reflection (far) improves confidence calibration of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 8702–8718. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.515. URL https://doi.org/10.18653/v1/2024.findings-acl.515.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. SELF-DISCOVER: large language models self-compose reasoning structures. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/e41efb03e20ca3c231940a3c6917ef6f-Abstract-Conference.html.

Appendix

A Prompts

We use separate prompts for each task. See the prompt templates for each task below. The parts in curly brackets vary and are inserted for each dataset element.

Factual questions
"You are a helpful assistant. Answer the question provided below. Answer in a word or a phrase and be as precise as possible.
Only provide the correct answer without any explanation. Provide your confidence between 0 and 100 in your answer.
{question}"

Multiple-Choice Questions
"You are a helpful assistant. Choose the correct option between A, B, C and D for the question given below. Only provide the correct answer without any explanation. Provide your confidence between 0 and 100 for your answer.
{question}
A. {option A}
B. {option B}
C. {option C}
D. {option D} "

Empirical Judgement
"You are a helpful assistant. Determine whether the sentence provided below asserts a causal, correlative, or neutral relation between two events. If the sentence asserts a causal relation respond causal, if the sentence asserts a correlative relation respond correlative, if the sentence asserts neither a causal nor a correlative relation between two events respond neutral.
Only provide the correct option without any explanation. Provide a confidence between 0 and 100 for your answer.
{statement}"

B Perturbation

We perturb the subjects using noise from a normal distribution with $\sigma = 0.1$, i.e., $\mathcal{N}(0, 0.1)$. We tested several different settings for this noise. We found that using higher noise values leads to a deterioration of the response, i.e., the model outputs nothing useful, and even lower noise values lead to no change in the response. For example, using $\sigma = 0.05$ results in a high accuracy (90 of 100 sampled questions correct), while using $\sigma = 1$ the model only answers 2 questions correctly, with an average answer length 10 times longer than with $\sigma = 0.1$ as the model begins to generate elaborate but unfitting responses, e.g., "I am an AI language model developed by OpenAI." (generated by Phi-4, which is not developed by OpenAI!). The chosen normal distribution with $\sigma = 0.1$ strikes the best balance between removing the notion of the question without leading to significant deterioration of the output.

The noise is added by using a hook in the model for the token or tokens of interest. This hook is activated in a forward pass, adds the sampled noise to the embedding of the token and returns the perturbed embedding for further processing in the forward pass.

C Pre-processing

We only want to analyze questions that we know the model can answer correctly, as only these questions are reasonable candidates for the expressed high confidence. We do note, however, that runs where the model

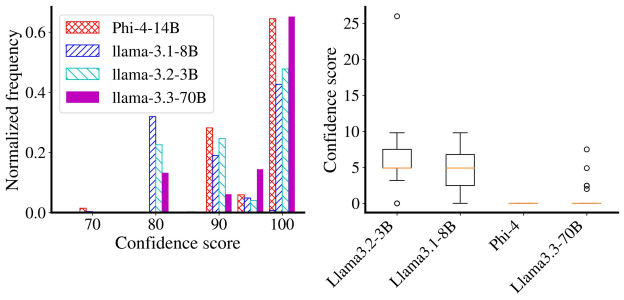


Figure 5: **Overview of confidence distributions (left) across all the LLMs and standard deviation of confidence values across different runs (right) for all models on the BIG-bench dataset. The quantized behavior is observed for all the models, irrespective of size. Smaller models also tend to generate slightly different confidence scores between runs. Larger models are more consistent.**

incorrectly answers the question but provides a confidence are rare ($< .5\%$ of questions) and these confidence scores collapse to the same values (i.e., 0, 90, 95, 100). To ensure that the model knows the answer, we run all questions through a pre-processing step by prompting the respective model and only keeping questions the model answers correctly. Note that for the MMLU and BIG-Bench questions, this reduces the number of questions for each model individually. For the factual questions, we sample questions until we have 1000 known questions for each model.

It must be noted that despite this pre-processing step, during the experiment, the models still sometimes give an incorrect answer or responses that do not contain a confidence value or answer (e.g., the response provides further instructions instead of answering).

D Sensitivity of Verbalized Confidence

To determine whether the confidence score generated for a particular prompt is consistent, each prompt is provided as input to the model multiple times. For a given dataset, we prompt the language model with the same question 5 times and obtain the answer and the confidence score generated by the model. In Figure 5 (Left), we present the results for the BIG-bench empirical judgments dataset, noting the generated answer and the confidence score. The generated confidence score's standard deviation is computed for each question. In Figure 5 (Right), we present the box plot of the standard deviation values. Ideally, we expect the models to have a standard deviation of 0 across all the questions, which is the case for Phi-4 and Llama3.3-70B (non-zero standard deviation was obtained for 4% of the questions). However, the smaller models Llama3.1-8B and Llama3.2-3B tend to generate different confidence scores across runs, although not too large (mean standard deviation of 5). Lastly, we prompt the models with just the word *confidence*, which leads to the model generating outputs such as "0.8", "8/10", etc., and when prompted to generate random numbers between 0 and 100, the model resorts to generating the number "42" frequently, which shows that the confidence score is not generated in the same manner as a "random" number.

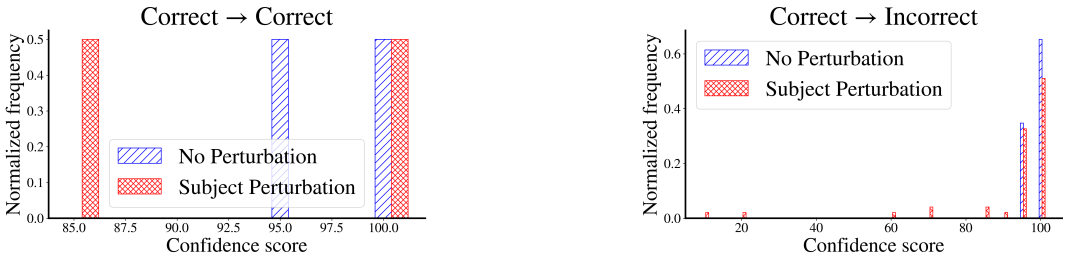
E Verbalized Confidence with Reasoning Models

Reasoning models are inspired by work from Wei et al. (2022) and have been applied to many different topics and tasks (Plaat et al., 2024; Huang & Chang, 2023; Zhang et al., 2025; Zhou et al., 2023). Reasoning models seemingly "think" through the given prompt more and have been shown to perform better, even with simple reasoning processes (Wei et al., 2022). This raises the question, if the verbalized confidence may be more accurate as well. Although not the main focus of our study, we provide results for a state-of-the-art reasoning model (Qwen3-30B-A3B-Thinking[3]) (Yang et al., 2025) on a subset ($n = 100$) of the factual questions.

[3]https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507

Pert.		Factual Qs	MMLU ast	MMLU geo	MMLU ana	MMLU med	BIG-bench
None	# correct	81	76	77	71	50	67
	Avg. conf.	$98.27\%_{\pm 2.38}$	$96.13\%_{\pm 2.48}$	$94.77\%_{\pm 4.90}$	$96.11\%_{\pm 5.31}$	$96.80\%_{\pm 2.60}$	$96.19\%_{\pm 2.30}$
Subject	# correct	5	59	53	39	36	53
	Avg. conf., correct	$95.00\%_{\pm 5.48}$	$95.76\%_{\pm 4.27}$	$93.49\%_{\pm 6.34}$	$95.08\%_{\pm 5.33}$	$94.44\%_{\pm 6.54}$	$95.28\%_{\pm 3.42}$
	# incorrect	61	11	14	22	12	14
	Avg. conf., incorrect	$90.57\%_{\pm 19.71}$	$92.27\%_{\pm 6.17}$	$92.36\%_{\pm 5.74}$	$93.41\%_{\pm 4.86}$	$91.25\%_{\pm 7.40}$	$93.21\%_{\pm 4.47}$

Table 7: Results of Qwen3 30B on the different tasks, showing the number of correct/incorrect answers and average confidence with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.



(a) (b)

Figure 6: **The confidence distribution of verbalized confidence given by Qwen3 30B, a reasoning model, on factual questions. We show the difference between the confidence with and without subject perturbation and between questions where the answer remains correct or changes after perturbation. Figure 6a shows the shift of verbalized confidence between no and added perturbation on questions where the answer does not change. Figure 6b shows the same shift, but for questions where the answer does change.**

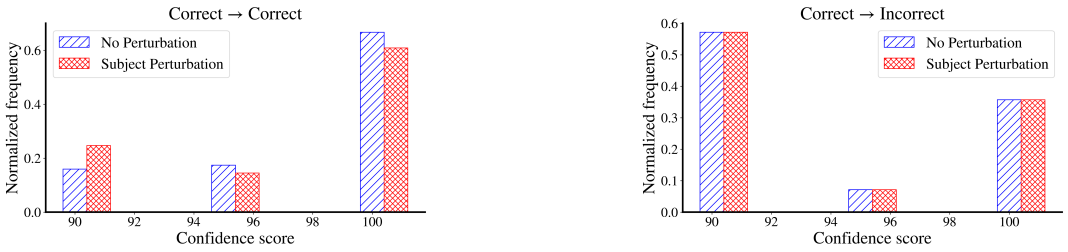
We find that the studied reasoning model exhibits the same behavior as the non-reasoning models. We provide a table of correct answers before and after perturbation in Table 7 and the distribution of verbalized confidence in Figure 6.

F Subject Perturbation Results

We show $x_c \to x_p$ for the confidence distribution of Phi-4 on MMLU astronomy questions in Figure 7.

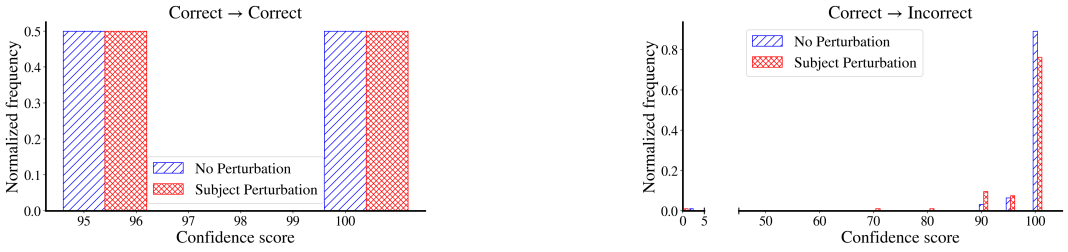
We report additional results for MMLU astronomy (Table 8), MMLU geography (Table 9), MMLU anatomy (Table 10), MMLU college medicine (Table 11 and BIG-bench (Table 12).

In Figure 8, we show the confidence distribution of verbalized confidence given by Phi-4 on a subset of factual questions when using a chain-of-thought (Wei et al., 2022) prompt.



(a) (b)

Figure 7: **The confidence distribution of verbalized confidence given by Phi-4 on MMLU astronomy questions. We show the difference between the confidence with and without subject perturbation and between questions where the answer remains correct or changes after perturbation. Figure 7a shows the shift of verbalized confidence between no and added perturbation on questions where the answer does not change. Figure 7b shows the same shift, but for questions where the answer does change.**



(a) (b)

Figure 8: **The confidence distribution of verbalized confidence given by Phi-4 on factual questions using a *chain-of-thought* prompt. We show the difference between the confidence with and without subject perturbation and between questions where the answer remains correct or changes after perturbation. Figure 8a shows the shift of verbalized confidence between no and added perturbation on questions where the answer does not change. Figure 8b shows the same shift, but for questions where the answer does change.**

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	66	69	83	94
	Avg. confidence	$91.89\%_{\pm 8.34}$	$89.35\%_{\pm 7.75}$	$96.93\%_{\pm 4.17}$	$98.56\%_{\pm 4.29}$
Subject perturbation	# correct answers	53	49	69	48
	Avg. confidence, correct	$91.70\%_{\pm 9.06}$	$88.65\%_{\pm 8.10}$	$96.81\%_{\pm 4.25}$	$95.83\%_{\pm 19.98}$
	# incorrect answers	11	17	14	28
	Avg. confidence, incorrect	$83.64\%_{\pm 28.05}$	$89.12\%_{\pm 8.44}$	$93.93\%_{\pm 4.70}$	$98.21\%_{\pm 9.28}$

Table 8: Results of all models on MMLU astronomy, showing the number of correct/incorrect answers and average confidence with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	67	66	95	91
	Avg. confidence	$91.94\%_{\pm 9.14}$	$88.48\%_{\pm 8.39}$	$94.95\%_{\pm 6.00}$	$93.49\%_{\pm 14.53}$
Subject perturbation	# correct answers	50	45	80	54
	Avg. confidence, correct	$89.50\%_{\pm 8.62}$	$90.22\%_{\pm 8.50}$	$94.69\%_{\pm 5.66}$	$93.89\%_{\pm 19.48}$
	# incorrect answers	15	20	15	27
	Avg. confidence, incorrect	$89.33\%_{\pm 8.54}$	$86.00\%_{\pm 10.20}$	$95.00\%_{\pm 4.83}$	$96.30\%_{\pm 18.89}$

Table 9: The number of correct/incorrect answers and average confidence on MMLU geography for all models with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	58	57	80	82
	Avg. confidence	$96.21\%_{\pm 7.09}$	$90.79\%_{\pm 8.05}$	$98.56\%_{\pm 2.98}$	$99.94\%_{\pm 0.55}$
Subject perturbation	# correct answers	42	35	56	41
	Avg. confidence, correct	$92.62\%_{\pm 9.01}$	$90.00\%_{\pm 10.69}$	$96.16\%_{\pm 13.53}$	$97.56\%_{\pm 15.43}$
	# incorrect answers	13	21	22	22
	Avg. confidence, incorrect	$94.62\%_{\pm 8.43}$	$86.19\%_{\pm 7.06}$	$98.18\%_{\pm 3.55}$	$100.0\%_{\pm 0.00}$

Table 10: The number of correct/incorrect answers and average confidence on MMLU anatomy for all models with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	43	49	57	56
	Avg. confidence	$96.74\%_{\pm 6.28}$	$91.53\%_{\pm 7.44}$	$97.63\%_{\pm 3.87}$	$99.29\%_{\pm 3.19}$
Subject perturbation	# correct answers	25	37	45	24
	Avg. confidence, correct	$92.00\%_{\pm 8.94}$	$87.43\%_{\pm 8.02}$	$97.00\%_{\pm 4.52}$	$95.83\%_{\pm 19.98}$
	# incorrect answers	16	12	12	18
	Avg. confidence, incorrect	$91.25\%_{\pm 9.27}$	$79.33\%_{\pm 24.46}$	$94.58\%_{\pm 4.77}$	$100.0\%_{\pm 0.00}$

Table 11: The number of correct/incorrect answers and average confidence on MMLU college medicine for all models with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.

		Llama3.2 3B	Llama3.1 8B	Phi-4	Llama3.3 70B
No perturbation	# correct answers	27	35	51	57
	Avg. confidence	$87.78\%_{\pm 9.16}$	$94.57\%_{\pm 8.31}$	$92.84\%_{\pm 13.15}$	$99.39\%_{\pm 2.31}$
Subject perturbation	# correct answers	20	27	41	36
	Avg. confidence, correct	$86.0\%_{\pm 8.00}$	$93.89\%_{\pm 8.64}$	$93.05\%_{\pm 6.24}$	$99.72\%_{\pm 1.64}$
	# incorrect answers	7	8	10	21
	Avg. confidence, incorrect	$81.43\%_{\pm 11.25}$	$81.25\%_{\pm 3.31}$	$92.5\%_{\pm 5.12}$	$79.52\%_{\pm 38.85}$

Table 12: The number of correct/incorrect answers and average confidence on BIG-bench for all models with and without perturbation. Note that we do not include questions in the average confidence calculation that the model refuses to answer, or questions without a clear subject. Further, we provide the standard deviation for the average confidences as a subscript of the respective values. Note that these deviations should be taken with a grain of salt, as usually the confidence is either very high (> 90) or 0, leading to a high standard deviation.