# Does ChatGPT Know that It Does Not Know? Evaluating the Black-Box Calibration of ChatGPT

**Youliang Yuan**[1,2], **Wenxuan Wang**[3], **Qingshuo Guo**[1],
**Yiming Xiong**[1], **Chihao Shen**[1], **Pinjia He**[1,2]
[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), China
[2]Shenzhen Research Institute of Big Data, China
[3]The Chinese University of Hong Kong
[1]{youliangyuan, 121090151, yimingxiong, 121020163}@link.cuhk.edu.cn,hepinjia@cuhk.edu.cn
[3] wxwang@cse.cuhk.edu.hk

**Abstract**

Recently, ChatGPT has demonstrated remarkable performance in various downstream tasks such as open-domain question answering, machine translation, and code generation. As a general-purpose task solver, an intriguing inquiry arises: Does ChatGPT itself know that it does not know, without any access to internal states? In response to this query, we present an initial evaluation of ChatGPT for black-box calibration (Ye and Durrett, 2022). We designed three types of proxy confidence, from three perspectives to assess its performance. Experiments are conducted on five datasets, spanning four tasks, and the results show that ChatGPT has a degree of capability for black-box calibration. Specifically, proxy confidence displayed a significantly positive Pearson correlation (95.16%) with accuracy in the TruthfulQA dataset, while revealing a negative correlation in the ModAr dataset. We delved deeper into ChatGPT's black-box calibration ability by examining failure cases in the ModAr dataset. Our analysis revealed that ChatGPT's tendency to exhibit overconfidence may stem from its reliance on semantic priors. Furthermore, we investigated why ChatGPT performs relatively well in TruthfulQA. The findings suggest that ChatGPT might implicitly acquire calibration skills during the reinforcement learning process, rather than relying solely on simplistic heuristics.

**Keywords:** Calibration, ChatGPT, Black-box

## 1. Introduction

Large language models (LLMs) (OpenAI, 2023a,b; Anthropic, 2023; Chiang et al., 2023; Bubeck et al., 2023) have taken off rapidly in natural language processing (NLP) recently and showcased remarkable performance on a variety of NLP tasks (Chowdhery et al., 2022; Brown et al., 2020; Ouyang et al., 2022; Schick et al., 2024; Liang et al., 2023). ChatGPT is a representative such that it not only performs well on NLP tasks (Jiao et al., 2023; Bang et al., 2023b; Qin et al., 2023; Park et al., 2023) but also exhibits excellent instruction-following capabilities to produce informative and coherent responses, which is attributed to the reinforcement learning from human feedback (RLHF, Christiano et al., 2017). However, these LLMs still suffer from the uncertainty issue such that they remain prone to confidently hallucinated predictions that appear plausible but are actually wrong (Ji et al., 2023; Zhang et al., 2023a; Li et al., 2024).

Existing studies focus on the uncertainty issue of LLMs in a white-box manner. For example, (Kadavath et al., 2022) demonstrates that LLMs (mostly) know their uncertainty by examining the softmax probability. (Lin et al., 2022a) shows that it is possible to teach LLMs to express their uncertainty through words by fine-tuning the model. Different from these works, we direct our attention toward a more nuanced question:

> *Does ChatGPT itself know that it does not know under the black-box setting?*

To answer the above question, this paper conducts a comprehensive study on the black-box calibration (Guo et al., 2017; Ye and Durrett, 2022) capability of ChatGPT. Specifically, we investigate to what extent the confidence of ChatGPT is correlated with the accuracy of its responses. To estimate the black-box confidence of ChatGPT, we propose three types of proxy confidence as outlined in Table 1, which are designed in distinct perspectives:

- **Qualitative Confidence**: We directly ask ChatGPT to output a confidence percentage (from 0% to 100%) accompanied by its response to the query, which essentially explores the self-awareness of ChatGPT. We are curious whether responses with higher confidence are more likely to be accurate than those with lower confidence.
- **Quantitative Confidence**: We ask ChatGPT to first provide an initial response and then determine the correctness of it by answering a follow-up question: "Is the answer True or False?" This binary response serves as an

Pinjia He is the corresponding author.



*LREC-COLING 2024*, pages 5191–5201

Table 1: Illustration of three distinct proxy confidence. [hist] represents the previous dialogue turn between the user and ChatGPT ("User:[query]; ChatGPT:[ans]). "consistency" refers to the percentage of subsequent answers that match the first answer provided by ChatGPT.

| Proxy Type | Prompt | Output | Confidence |
|---|---|---|---|
| `Quantitative\|` | answer question and give your confidence (%): [query] | [ans], confidence is $c$% | $c$% |
| `Qualitative\|` | [hist]; is the answer true or false? | true / false | high / low |
| `Consistent` | [query$_1$]; $\cdots$ ; [query$_n$]<br>[hist]; please think again. | [ans$_1$]; $\cdots$ ; [ans$_n$]<br>[ans$_{new}$] | consistency |

indicator of ChatGPT's confidence. Typically, a response of "True" suggests a higher confidence while that of a "False" indicates a lower confidence.

- **Consistency**: Generally, we suppose that a model exhibits higher confidence when its responses remain consistent despite perturbations. We empirically involve two specific perturbations here: (1) repeatedly asking the same question to ChatGPT multiple times; (2) questioning ChatGPT's initial response and requesting it to answer again. By leveraging the resulting consistency of the obtained responses, we may quantify the confidence of ChatGPT.

ChatGPT is evaluated on multiple datasets, including TruthfulQA (Lin et al., 2022b), MMLU (Hendrycks et al., 2021), and three datasets from BIG-bench (Srivastava et al., 2022), namely Modified Arithmetic (ModAr), Analytic Entailment (AnaEnt), and Language Identification (LangId). Results indicate that **ChatGPT possesses some level of black box calibration.**

- In particular, a positive correlation exists between quantitative confidence and accuracy in the TruthfulQA dataset (see Figure 1), but this is not observed in all datasets.
- ChatGPT exhibits a significant degree of overconfidence, as indicated by an average confidence level of 93.64% and an accuracy level of 49.99%.
- Furthermore, the accuracy of responses in the high consistency subset surpasses that of the low consistency subset, with values of 87.64% and 48.41% respectively. It is important to note, however, that there is an imbalance in the number of samples between these subsets, which undermines the black-box calibration ability of ChatGPT.
- Further analysis indicates that one of the reasons for ChatGPT's tendency to exhibit overconfidence in providing answers is its strong prior knowledge base.
- Additionally, the black-box calibration of ChatGPT cannot be explained solely by simple heuristics such as option numbers. It is speculated that ChatGPT implicitly learns this calibration during the process of reinforcement learning.

## 2. Experiments

In this section, we present the datasets employed in the experiments (Section 2.1). Subsequently, we assess ChatGPT's black-box calibration ability by utilizing three proxy confidence (Section 2.2), and consistency (Section 4). All experiments are conducted in the zero-shot setting. The temperature of ChatGPT is set to 0.7.

### 2.1. Dataset

ChatGPT was comprehensively evaluated by conducting experiments on 5 carefully selected datasets, which measure the model's performance in different aspects, including truthfulness, knowledge, reasoning, and multilingual capability. The datasets used are as follows:

- **TruthfulQA** (Lin et al., 2022b) (Multi-choice QA) is a benchmark comprising questions specifically curated to detect imitative falsehoods that could be perpetuated by AI systems. The dataset consists of a total of 817 questions, organized into 38 distinct categories, such as Health, Law, and Conspiracies.
- **MMLU** (Lin et al., 2022b) (Multi-choice QA) is a large multitask test dataset that contains questions from various fields of knowledge, such as mathematics, social sciences, and business. It includes 57 tasks and questions from various difficulty levels, like "Elementary," "High School", "College", and "Professional".
- **Modified Arithmetic** (Srivastava et al., 2022) (ModAr) involves a type of math problem that requires identifying patterns same or similar to standard arithmetic but with subtle differences.

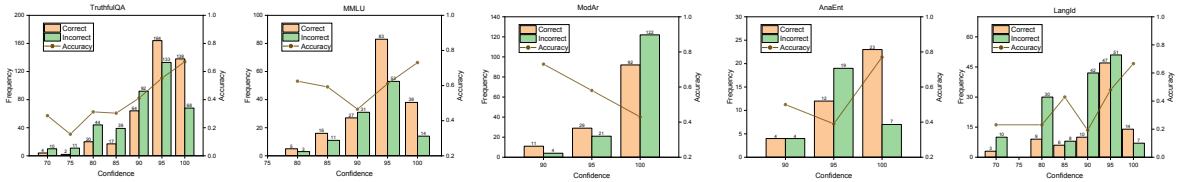


Figure 1: The distribution of correct and incorrect samples across various confidence intervals on the five datasets, along with their respective accuracies.

For instance, a model may be asked to add one to the result of an operation performed on two three-digit numbers.

- **Analytic Entailment** (Srivastava et al., 2022) (NLI) is designed to measure a model's ability to recognize the logical inference between sentences within a piece of reasoning by analyzing their meanings.
- **Language Identification** (Srivastava et al., 2022) (Text Classification) involves the task of classifying an input text into one of eleven possible languages, with the objective of accurately identifying the language among the given options.

We randomly selected 817 samples from TruQA, 250 samples from LangID, 285 samples from MMLU, 70 samples from AnaEnt, and 300 samples from ModAr. We referred to the samples where the model provided accurate answers as "correct samples" and the samples where the model provided incorrect answers as "incorrect samples".

## 2.2. Quantitative & Qualitative Confidence

In this subsection, we evaluate ChatGPT's ability to determine the accuracy of provided answers based on quantitative or qualitative confidence.

**Metric** To evaluate ChatGPT's performance, we employ Expected Calibration Error (ECE) and Pearson Correlation Coefficient (PCC) as metrics. ECE quantifies the degree of correspondence between confidence and accuracy, whereas PCC measures the correlation between confidence and accuracy. Additionally, we introduce the Monotonicity Score (MS), which quantifies the degree to which the accuracy increases as the confidence increases:

$$\text{MS} = \frac{1}{Z}\sum_{i=1}^{N-1}\frac{n_{i+1}+n_i}{2}\,\text{sign}\left(acc_{i+1}-acc_i\right),$$

where $N$ is the number of the confidence intervals and $acc_i$ the average accuracy for confidence interval $i$ (large $i$ refer high confidence). $Z = \sum_{i=1}^{N-1}\frac{n_{i+1}+n_i}{2}$ is the normalization factor. sign is the sign function. The value of MS ranges from -1 to 1, with values of -1 and 1 indicating strictly decreasing and increasing monotonicity, respectively.

Table 2: Expected calibration error ($\times 10^{-2}$) on various datasets. The terms "UnderConf," "OverConf," and "ECE" refer to underconfidence, overconfidence, and overall Expected Calibration Error, respectively. Underconfident ECE denotes the calibration error when the confidence is lower than the corresponding accuracy, while overconfident ECE represents the calibration error when the confidence is higher than the accuracy.

| Dataset | UnderConf | OverConf | ECE |
|---|---|---|---|
| TruQA | 0.10 | 41.95 | 42.05 |
| LangId | 0.00 | 52.70 | 52.70 |
| MMLU | 0.11 | 33.39 | 33.50 |
| AnaEnt | 0.29 | 36.14 | 36.43 |
| ModAr | 0.75 | 50.12 | 50.87 |
| Average | 0.25 | 42.86 | 43.11 |

Table 3: The Pearson correlation coefficient, monotonicity score, average confidence, and overall accuracy of ChatGPT in 5 datasets.

| Dataset | PCC | MS | Avg.Con | Acc |
|---|---|---|---|---|
| TruQA | 95.16 | 78.91 | 92.49 | 50.49 |
| LangId | 74.97 | 63.73 | 89.18 | 36.48 |
| MMLU | 54.47 | 52.44 | 93.36 | 60.07 |
| AnaEnt | 87.95 | 35.51 | 96.42 | 55.65 |
| ModAr | -33.45 | -86.96 | 96.73 | 47.28 |
| Average | 55.82 | 28.73 | 93.64 | 49.99 |

**ChatGPT Appears To Know It Is Uncertainty But Not Always.** Based on Table 2 and 3, we observe that ChatGPT demonstrates a significant tendency towards overconfidence, with an average confidence of 93.19 and an average accuracy of 47.02. On the other hand, ChatGPT displays a positive correlation with TruthfulQA, MMLU, AnaEnt, and LangId. ChatGPT appears to have the ability to recognize uncertainty in TruthfulQA, where confidence is strongly correlated with accuracy (PCC

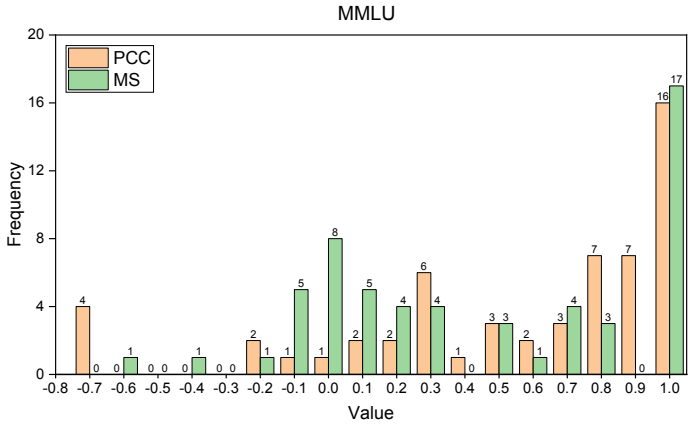


Figure 2: The distribution of 57 tasks in the MMLU across various PCC and MS intervals.

= 95.16%, MS = 78.91%). This suggests that ChatGPT has the capacity to acknowledge when its answers may be incorrect. However, this ability is not consistent across ModAr. The differences in task nature and input/output space between these datasets may contribute to this discrepancy. For instance, ModAr poses questions that are similar to standard simple arithmetic but have different answers, potentially leading the model to give a wrong answer with high confidence. We will discuss this in Section 3.1.

We performed additional experiments on MMLU by sampling 57 sub-datasets, each containing 100 samples, from 57 different tasks. The results, presented in Figure 2, indicate that out of the 57 tasks, 35 tasks have a PCC greater than 0.50, and 16 tasks have a PCC greater than 0.90. However, as previously observed, some datasets exhibit weak or even negative correlations. We further discuss this in Section 3.2.

In the experiments before, we directly applied ChatGPT's confidence in its response as our quantitative confidence. However, the confidence is not deterministic, which may undermine the reliability of our findings. Thus, we repeatedly prompt ChatGPT to generate the response for each question 5 times, after which we obtain a total of 5 confidence values. The mean confidence standard deviations for each dataset are shown in Figure 5, which indicates the confidence fluctuates in an acceptable range from 3.31% (MMLU) to 6.14% (ModAr).

**ChatGPT Always Think Its Answer Is Correct.** According to Table 4, ChatGPT deems its response as "True" in the majority of cases (95.36%). This suggests that ChatGPT faces difficulty in conveying its uncertainty in this way. Consequently, utilizing this proxy confidence could lead to a high rate of false negatives (incorrect with "True"). Surprisingly, incorrect samples have a higher likelihood of being considered "True" than correct samples (97.83% vs. 94.97%).

Table 4: The "True" response proportion across correct, incorrect, and overall subsets.

| Dataset | Correct | Incorrect | Overall |
|---|---|---|---|
| TruQA | 97.24 | 97.98 | 97.54 |
| MMLU | 98.85 | 99.07 | 92.76 |
| ModAr | 90.37 | 90.00 | 90.18 |
| AnaEnt | 100.00 | 100.00 | 100.00 |
| LangId | 91.75 | 99.33 | 96.34 |
| Average | 94.97 | 97.83 | 95.36 |

## 2.3. Consistency

Consistency refers to the proportion of subsequent answers that are the same as the first answer. It can serve as an indicator of confidence, assuming that a confident model should produce stable answers. We offer two methods: (1) **Repetition Consistency**, repeating the same question multiple times (e.g., 5 times in our experiments), and (2) **Regeneration Consistency**, asking ChatGPT to re-generate its answer after questioning its initial response. If the subsequent answer differs from the original, we classify it as a "flip". In this section, we show the results of four datasets except for ModAr, because the prediction of ModAr is always consistent, we will analyze this in Section 3.1.

### 2.3.1. Repetition Consistency

**Higher Consistency Means Higher Accuracy.** We present the relationship between accuracy and consistency in Figure 4. The result indicates that as ChatGPT becomes more confident (higher consistency), its accuracy also improves. This is intuitive, because the model may provide different answers when uncertain, but only outputs the correct answer when it is certain.

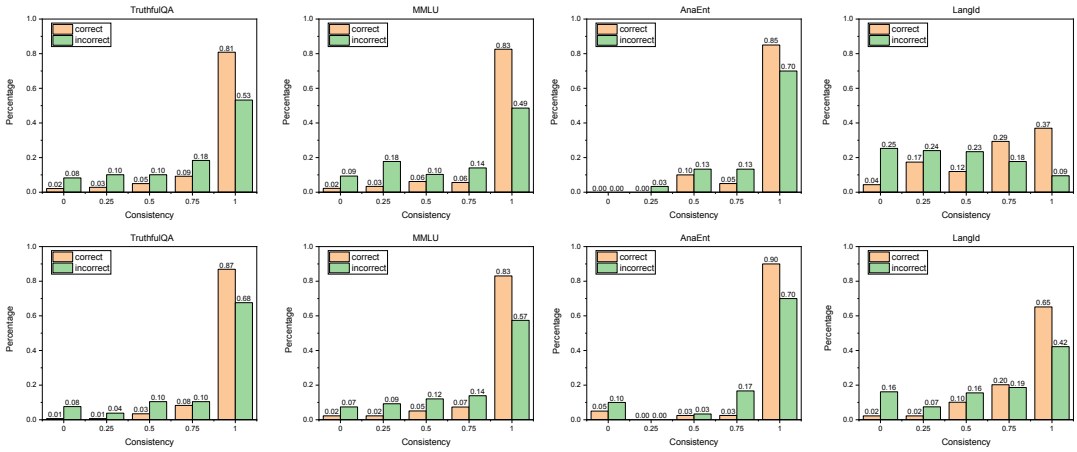


Figure 3: The data distribution on consistency. The first row is the prompt-change setting, where we use the same prompt each time. The second row is the prompt-unchanged setting, where we use different prompts each time. The consistency in the prompt-change setting is generally lower than that in the prompt-unchanged setting.

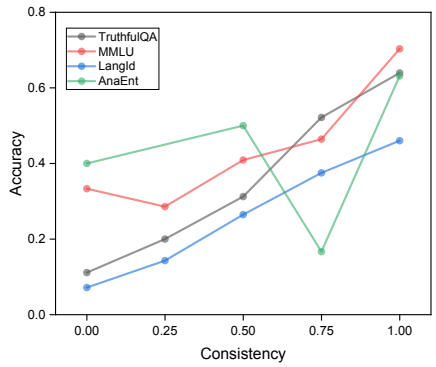


Figure 4: The relationship between accuracy and answer consistency. We repeatedly ask ChatGPT 5 times for each sample. Consistency is calculated by comparing the first answer with subsequent answers.

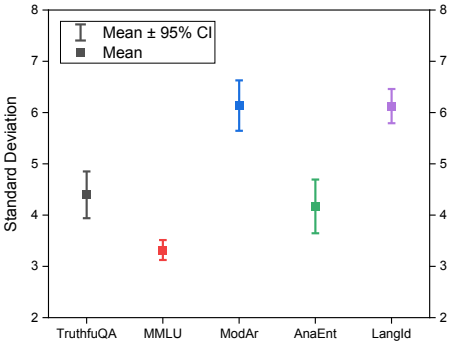


Figure 5: The standard deviation of confidence (repeat 5 times for each query).

Table 5 shows that the proportion with consistent responses in correct samples is substantially higher than that in incorrect samples in these datasets. The proportion in correct samples is 81%, while it is only 59% in incorrect samples. If we use different prompts each time, the gap will further increase from 22% to 26%. This is reasonable because it introduces more randomness by using different prompts, which distinguishes answer stability better (see Figure 3). Overall, these results suggest ChatGPT is more confident in its answers which are correct. It should be mentioned that this conclusion is not held on ModAr, we will discuss this in Section 3.1.

Table 5: Percentage of responses that do not change during repeated asking. "T/F" represents correct and incorrect samples. "Prompt" means we use different prompts each time. The difference between correct and incorrect samples is denoted by $\Delta$.

| Dataset | Base(T/F) | Base($\Delta$) | Prompt(T/F) | Prompt($\Delta$) |
|---|---|---|---|---|
| TruQA | 0.87 / 0.68 | 0.19 | 0.81 / 0.53 | 0.28 |
| MMLU | 0.83 / 0.57 | 0.26 | 0.83 / 0.49 | 0.34 |
| AnaEnt | 0.90 / 0.70 | 0.20 | 0.85 / 0.70 | 0.15 |
| LangId | 0.65 / 0.42 | 0.23 | 0.37 / 0.09 | 0.28 |
| Average | 0.81 / 0.59 | 0.22 | 0.72 / 0.45 | 0.26 |

### 2.3.2. Regeneration Consistency

**Non-flip Means Higher Accuracy.** As shown in Table 6, the accuracy of the non-flip subset is significantly higher than that of the flip subset (87.64% vs 48.41%), supporting the notion that stable answers have higher accuracy. However, the average flip rate is high (79.48%), indicating that ChatGPT is prone to changing its answer. This may lead to many false positive errors (correct but flip). To further analyze this, we present the flip rate of the

Table 6: The accuracy in flip and non-flip subsets. The Flip rate represents the percentage of instances where ChatGPT alters its response upon questioning.

| Dataset | Acc (Flip) | Acc (Non-flip) | Flip Rate |
|---|---|---|---|
| TruQA | 47.21 | 95.12 | 82.77 |
| MMLU | 55.64 | 96.00 | 91.13 |
| AnaEnt | 62.32 | 100.00 | 98.57 |
| LangId | 44.29 | 100.00 | 85.37 |
| Average | 48.41 | 87.64 | 79.48 |

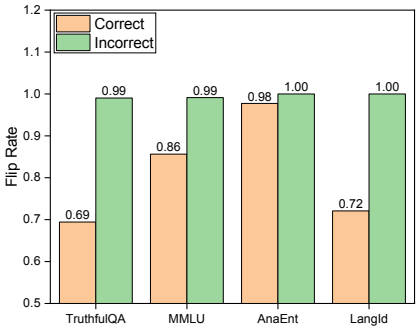


Figure 6: The flip rate of correct and incorrect subsets.

correct and incorrect subsets in Figure 6. In these 4 datasets, almost all incorrect answers will flip after prompting "Please think again". Regarding the flip rate of the correct subset, it is high too while significantly lower than that of the incorrect subset. In AnaEnt, there are 98% correct answers and 100% incorrect answers flips when required to regenerate answers. We consider this extremely high flip rate may be because the task is a binary classification task, and ChatGPT is strongly biased to the remaining answer if its answer is argued by users.

## 3. Results and Discussion

As mentioned before, Table 3 illustrates a positive correlation between confidence and accuracy in TruthfulQA, MMLU, Analytic Entailment, and Language Identification, whereas a negative correlation is observed in Modified Arithmetic. In the following section, we delve into the underlying reasons for this discrepancy.

### 3.1. ModAr

We divided Modar samples into two distinct categories: Normal and Overriding types. The Normal type involves standard mathematical operations such as addition, subtraction, or multiplication. In contrast, the Overriding type presents similar mathematical operations but with different input-label mappings, requiring the model to understand the operations and override the semantic prior (refer to Table 7 for details). Table 9 displays the results obtained from these two types of data. The model achieves nearly 100% accuracy on normal data, but only 0% accuracy on overriding data. Despite the remarkably low accuracy on the latter, the model exhibits high confidence, surpassing that of normal data. This finding suggests that the model tends to be excessively confident in its answers when it possesses strong prior knowledge, which can significantly impair its calibration. Furthermore, we employed Chain-of-thought (CoT) and an explanation prompt on the overriding data, but the model still demonstrates extremely high confidence. The explanation prompt is *You should understand the meaning of the operation '->' before giving the answer. For example, if a + b -> (a+b+1), '->' represents the sum of two numbers plus one.*

### 3.2. TruthfulQA

In this subsection, we provide preliminary explanations for ChatGPT's black-box calibration capability to enhance our understanding. We specifically focus on the TruthfulQA dataset, where confidence and accuracy exhibit a high correlation, and use quantitative confidence as the proxy. Our analysis considers two aspects:

- **Input Data.** Does ChatGPT solely utilize some simple heuristics in questions to generate confidence? We know that some heuristics can have a strong correlation with accuracy. For instance, in TruthfulQA, the number of options in a question ranges from 2 to 11. If ChatGPT assigns lower or higher confidence to questions with fewer or more options, respectively, this may result in a high correlation between confidence and accuracy. In our experiments, we consider heuristics such as the number of options, query length, perplexity, difficulty level, and superficial semantic features.

- **Model.** Can ChatGPT's capability be attributed to techniques in the training process? For example, ChatGPT might implicitly learn to know its uncertainty from the designed reward function used in the reinforcement learning process. We explore this by comparing the performance of models with varying sizes and training strategies, including Alpaca (7B) (Taori et al., 2023), text-davinci-002, ChatGPT.

**The Calibration Capability Can Not Be Explained By Heuristics We Provide.** Table 8 presents the average option count, question length,

Table 7: Normal and overriding samples from ModAr.

| ModAr | Normal | Overriding |
|---|---|---|
| Query\| | In the following lines, the symbol -> represents a simple mathematical operation.<br>355 + 927 -> 1282<br>694 + 206 -> 900<br>767 + 234 -> 1001<br>441 + 354 -> 795<br>514 + 207 -> 721<br>708 + 852 -> | In the following lines, the symbol -> represents a simple mathematical operation.<br>980 + 605 -> 1586<br>660 + 623 -> 1284<br>566 + 848 -> 1415<br>935 + 466 -> 1402<br>627 + 142 -> 770<br>200 + 179 -> |
| Response\| | ANSWER: 1560. The operation 708 + 852 results in 1560. This is obtained by adding the two numbers together. Percentage of confidence: 100% | ANSWER: 379 . The last mathematical operation was 200 + 179, which equals 379. My confidence in this answer is 100%. (ground-truth: 380) |

Table 8: The number of options, length, and perplexity in 5 different confidence groups. "H.a" and "P.m" represent the human aging test and professional medicine test from MMLU respectively.

| Heuristic / Group | Conf=100 | Conf=95 | Conf=90 | Conf=85 | Conf=80 |
|---|---|---|---|---|---|
| Option Number | 4.67 | 5.21 | 5.28 | 5.66 | 5.27 |
| Perplexity | 10.72 | 8.57 | 9.11 | 7.92 | 8.71 |
| Length (TruQA) | 274 | 338 | 320 | 355 | 318 |
| Length (H.a) | 155 | 193 | 171 | 204 | 167 |
| Length (P.m) | - | 733 | 790 | 731 | 705 |

Table 9: The number of correct/incorrect predictions on normal data and overriding data. "with CoT" means we use CoT on overriding data. "with Expl" means we used a prompt with more explanation for the operation.

| Right / Wrong | Conf=90 | Conf=95 | Conf=100 |
|---|---|---|---|
| Normal | 11 / 0 | 29 / 0 | 91 / 0 |
| Overriding | 0 / 4 | 0 / 21 | 0 / 113 |
| with CoT | 0 / 2 | 0 / 6 | 0 / 120 |
| with Expl | 0 / 3 | 2 / 4 | 20 / 109 |
| ModAr | 11 / 4 | 29 / 21 | 91 / 113 |

and perplexity[1] for various confidence intervals. Despite the most confident group having fewer options and shorter question lengths than other groups, no apparent correlation exists. Figure 7 displays t-SNE visualization (Van der Maaten and Hinton, 2008) of the superficial semantic features of each sample. We employ SentenceBERT (Reimers and Gurevych, 2019) to encode each TruthfulQA sample into an embedding and subsequently use t-SNE to project these features onto a two-dimensional space with different confidence levels represented by colors. It is observed that points of different confidence are mixed together. As to difficulty, we select tasks with varying levels of difficulty from MMLU, including Elementary, High School, and College Mathematics Tests. The results, as shown in Table 10, demonstrate that confidence is roughly unchanged as difficulty levels

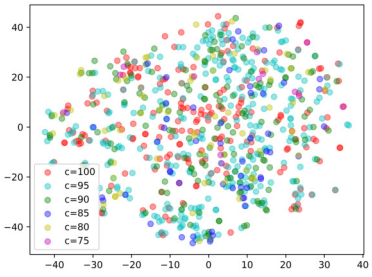


Figure 7: t-SNE visualization of superficial semantic features. We employ SentenceBERT to encode each TruthfulQA sample into an embedding and subsequently use t-SNE to project these features onto a two-dimensional space with different confidence levels represented by colors.

increase. For instance, the average confidence in elementary, high school, and college mathematics tests are 96.60%, 95.73%, and 94.90% respectively. In psychology tests, the professional test demonstrates only slightly lower confidence than the high school test (91.17% vs 92.95%). In summary, the above results suggest that the heuristics used cannot explain the confidence expressed by ChatGPT.

**This Capability May Be Implicitly Learned In Reinforcement Learning.** We compare different models in Table 11. Alpaca[2] is fine-tuned from the

[1]we calculate perplexity using a pre-trained GPT-2-large model as a proxy language model.

[2]here we use a replicate version from https://huggingface.co/chavinlo/alpaca-native

Table 10: Average confidence and accuracy in different tasks from MMLU. The difficulty level includes Elementary, High School, College, and Professional. There might be an accuracy gap between tasks in different difficulty levels, while the confidence is very close.

| **Task & Difficulty** (Conf/Acc) | **Elementary** | **High School** | **College** | **Professional** |
|---|---|---|---|---|
| Biology | - | 94.75/81.82 | 94.25/71.00 | - |
| Chemistry | - | 93.47/58.16 | 93.35/46.39 | - |
| Computer Science | - | 94.10/66.00 | 93.93/41.84 | - |
| Mathematics | 96.60/45.00 | 95.73/34.38 | 94.90/35.35 | - |
| Medicine | - | - | 92.95/65.00 | 90.20/73.74 |
| Physics | - | 93.60/36.00 | 93.43/43.43 | - |
| Psychology | - | 92.95/83.00 | - | 91.17/68.37 |
| Average | 96.60/45.00 | 94.10/59.89 | 93.80/50.50 | 90.69/71.06 |

Table 11: The black-box calibration of different models. “Avg con” and "Acc" mean average confidence and accuracy respectively.

| **Model** | **PCC** | **MS** | **Avg.Con** | **Acc** |
|---|---|---|---|---|
| Alpaca(7B) | -16.37 | 15.89 | 82.32 | 18.89 |
| Davinci(175B) | -20.69 | 25.10 | 87.97 | 28.98 |
| ChatGPT | 95.16 | 78.91 | 92.49 | 50.49 |

LLaMA, utilizing instruction-following demonstrations from text-davinci-003. Davinci (text-davinci-002) is trained with supervised fine-tuning instead of reinforcement learning. However, these models exhibit a negative PCC (-16.37% in Alpaca and -20.69% in Davinci), while ChatGPT demonstrates significantly better black-box calibration performance. We speculate this observed difference may stem from the reinforcement learning employed in the training of ChatGPT, whereby some reward functions are incorporated to discourage model responses. As a result, it is possible for the models to implicitly learn the black-box calibration.

## 4. Related Works

### 4.1. Evaluation of LLMs

Large language models (LLMs) are gaining increasing popularity in both academia and industry. As LLMs continue to play a vital role, their evaluation becomes increasingly critical (Chang et al., 2023). Previous works have evaluated LLMs from different perspectives, such as correctness (Zhang et al., 2023b; Bang et al., 2023a; Wang et al., 2024; Liu et al., 2023), fairness (Gallegos et al., 2023; Li et al., 2023; Wan et al., 2023; Wang et al., 2023a), and safety (Wei et al., 2023; Kumar et al., 2023; Tian et al., 2023; Yuan et al., 2023; Wang et al., 2023b). Different from the previous works mentioned above, this paper evaluates LLMs from a calibration perspective and aims to answer the following question: does ChatGPT itself know that it does not know?

### 4.2. Uncertainty of LLMs

Due to the popularity of large language model studies, several concurrence works are coming out during our submission. For example, (Ye et al., 2024) introduces a new benchmarking approach for LLMs that integrates uncertainty quantification. However, the method is white-box and can not assess the widely deployed closed-source LLMs, such as ChatGPT and GPT-4. (Xiong et al., 2023) explores black-box methods for confidence elicitation for LLMs on five types of tasks and finds that LLMs tend to be over-confident. Our paper is a complementary work that verifies this finding on different tasks and datasets.

## 5. Conclusion

In this paper, we examined ChatGPT’s black-box calibration capability and proposed three types of proxy confidence: quantitative confidence, qualitative confidence, and answer consistency. Our experiments on five datasets revealed that ChatGPT has a certain level of black-box calibration ability. Although the quantitative confidence metric showed a strong positive correlation with accuracy in some datasets, this was not the case for all datasets. Our analysis indicates that ChatGPT does not learn this ability from simple heuristics but may learn it implicitly through reinforcement learning. Overall, our findings contribute to a better understanding of the capabilities of ChatGPT and provide insights for future research.

## 6. Acknowledgement

This paper was supported by the National Natural Science Foundation of China (No. 62102340) and Shenzhen Science and Technology Program.

Anthropic. 2023. Model card and evaluations for claude models, https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Yu-Chu Chang, Xu Wang, Jindong Wang, Yuanyi Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Weirong Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qian Yang, and Xingxu Xie. 2023. A survey on evaluation of large language models. *ArXiv*, abs/2307.03109.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md. Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen Ahmed. 2023. Bias and fairness in large language models: A survey. *ArXiv*, abs/2309.00770.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, An-

drea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt A good translator? A preliminary study. *CoRR*, abs/2301.08745.

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. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Himabindu Lakkaraju. 2023. Certifying llm safety against adversarial prompting. *ArXiv*, abs/2309.02705.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Y. Wang. 2023. A survey on fairness in large language models. *ArXiv*, abs/2308.10149.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching models to express their uncertainty in words. *CoRR*, abs/2205.14334.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

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

OpenAI. 2023a. ChatGPT, https://openai.com/chatgpt.

OpenAI. 2023b. GPT-4 technical report, https://cdn.openai.com/papers/gpt-4.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew K. Lampinen, Andy Zou,

Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, and et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *CoRR*, abs/2206.04615.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. *GitHub repository*.

Yu Tian, Xiao Yang, Jingyuan Zhang, Yinpeng Dong, and Hang Su. 2023. Evil geniuses: Delving into the safety of llm-based agents. *ArXiv*, abs/2311.11855.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. 2023. Biasasker: Measuring the bias in conversational ai system. *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen tse Huang, Zhaopeng Tu, and Michael R. Lyu. 2023a. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *ArXiv*, abs/2310.12481.

Wenxuan Wang, Juluan Shi, Zhaopeng Tu, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. The earth is flat? unveiling factual errors in large language models. *ArXiv*, abs/2401.00761.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023b. All languages matter: On the multilingual safety of large language models. *ArXiv*, abs/2310.00905.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *ArXiv*, abs/2307.02483.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ICLR*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *ArXiv*, abs/2401.12794.

Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212.

Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Pinjia He, Shuming Shi, and Zhaopeng Tu. 2023. Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher. *ArXiv*, abs/2308.06463.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.