Assessing Factual Reliability of Large Language Model Knowledge

Anonymous ACL submission

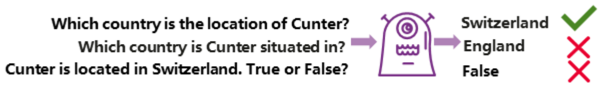
Abstract

The factual knowledge of LLMs is typically evaluated using accuracy, yet this metric does not capture the vulnerability of LLMs to hallucination-inducing factors like prompt and context variability. How do we evaluate the capabilities of LLMs to consistently produce factually correct answers? In this paper, we propose **MOdel kNowledge relIabiliTy scORe (MONITOR)**, a novel metric designed to directly measure LLMs' factual reliability. MONITOR is designed to compute the distance between the probability distributions of a valid output and its counterparts produced by the same LLM probing the same fact using different styles of prompts and contexts. Experiments on a comprehensive range of 12 LLMs demonstrate the effectiveness of MONITOR in evaluating the factual reliability of LLMs while maintaining a low computational overhead. In addition, we will release the **FKTC** (Factual Knowledge Test Corpus) to foster research along this line.

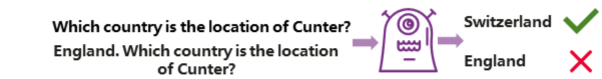
1 Introduction

Recently, large pre-trained language models (LLMs) have been used as de facto storage for factual knowledge (Petroni et al., 2019). However, applying LLMs to real-world scenarios inevitably leads to language generation deviating from known facts (aka "factual hallucination" (Chang et al., 2023)) due to multiple causes. For example, Cao et al. (2021) argued that the performance of an LLM is over-estimated due to biased prompts overfitting datasets (also referred to as the framing effect in Jones and Steinhardt (2022)) and in-context information leakage.

Given the variability of LLMs' performance under different prompts and contexts, it becomes evident that relying solely on accuracy as an evaluation metric is insufficient. We also need to gauge how robust LLMs are to variations in prompting. In Figure 1 we show examples of factual probes where either the framing of the prompt, or the context to the prompt, is varied, leading to the issue of "accuracy instability".



(a) Prompt framing effect



(b) Effect of in-context interference

Figure 1: "Accuracy instability" during language generation under various prompts.

Prompt framing effect: An LLM generates different predictions depending on how prompts are framed. Predictions are associated with prompts instead of factual knowledge learned in LLMs. As shown in Figure 1(a), for a fact represented in a triplet <*Cunter, is located in, Switzerland*>, the generated predictions for re-framed prompts *"Which country is Cunter situated?"* and *"Cunter is located in Switzerland. True or False?"* are non-factual.

Effect of in-context interference: An LLM leverages in-context information during its decoding stage, but this information may negatively affect an LLM's prediction during knowledge probing. As shown in Figure 1(b), for the same fact, when presented with a context *"England."* concatenated with the prompting question *"Which country is the location of Cunter?"*, an LLM generates a non-factual prediction *"England"*.

How do we assess the reliability of factual knowledge of LLMs under the effects of these hallucination-inducing factors? Investigations into the behaviors of language models during knowledge probing (Petroni et al., 2019; Kassner and

Schütze, 2020; Gupta, 2023) have mainly used metrics like precision and accuracy to quantify errors under a specified factor like prompt framing (Jones and Steinhardt, 2022) or mis-primed information (Kassner and Schütze, 2020). Despite the insights gained by showing the instability of LLMs during knowledge probing, these studies are subject to two limitations:

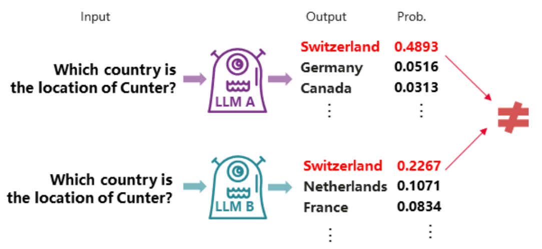


Figure 2: The same top-1 answer with different output probabilities from two LLMs.

No Exploration of Uncertainty. Metrics like top-one accuracy may capture the ordering of predictions in the output space, but they lack the resolution to reflect on the degree of factual knowledge being learned by LLMs. Figure 2 depicts an example where two LLMs (Models A and B) may produce the same result even though their output probabilities vary. By equating the performance of Model A with that of Model B, one introduces a level of approximation in representation, which can be regarded as a source of uncertainty. In this paper, we directly use the output probabilities and construct a high-resolution metric to perform knowledge assessment.

Limited Scope. Previous works focus on understanding the effect of variability of a specific type. We design experiments to investigate the combined effects of multiple causes of accuracy instability: prompt framing and in-context interference during knowledge assessment. In addition, few studies have experimented on LLMs with billions of parameters. In contrast, we investigate the knowledge reliability of 12 freely downloadable LLMs with a range of parameter sizes and origins (with and without instruction fine-tuning).[1]

In this paper, we propose a novel distance-based approach **MOdel kNowledge relIabiliTy scORe (MONITOR)** which captures the deviation of output probability distributions under contexts of prompting variance, interference from mispriming (Kassner and Schütze, 2020) and positively-primed prompts.

We perform experiments on a comprehensive set of knowledge probing tasks and investigate the effectiveness of MONITOR in assessing LLMs' factual reliability. Through experiments with a large variety of different facts, we show that a lower-MONITOR LLM is less likely to suffer from **"accuracy instability"** issue. Computing MONITOR takes only one-third GPU hours of those consumed by a comprehensive accuracy reliability study, making MONITOR a low-cost metric for assessing factual knowledge reliability of LLMs. **Our contributions are:**

1. We propose a novel method to assess the factual reliability of LLMs in the presence of the prompt framing effect and in-context interference. The proposed metric, MONITOR, can be used in conjunction with an end-to-end metric (i.e., accuracy) as part of a multi-dimensional approach to LLM knowledge evaluation.

2. We construct the **FKTC** (Factual Knowledge Test Corpus) by developing question answering probing prompts (210,171 prompts in total) based on 16,167 triplets of 20 fact datasets from T-REx corpus (Elsahar et al., 2018). We will release **FKTC** to the public to foster research works along this line.

2 Related Work

Petroni et al. (2019) demonstrated that factual knowledge can be directly extracted from language models without needing an external knowledge source. However, extracting knowledge (aka knowledge probing) from language models is error-prone due to various biases. For example, Elazar et al. (2021) showed that the consistency of knowledge extracted is generally low when the same fact is queried with different prompts. Many works in prompt engineering attempt to automatically construct prompts outperforming manual prompts (Shin et al., 2020; Jiang et al., 2020; Zhou et al., 2023; Kojima et al., 2022). Cao et al. (2021) argued that the decent performance of a language model is ascribed mainly to the application of these biased prompts, in other words "better" prompts are found to over-fit the answer distribution of the test set instead of reflecting on LLMs' generalization ability to predict factual knowledge.

LLMs are sensitive to in-context information. Kassner and Schütze (2020); Gupta (2023) showed

[1]Only freely downloadable LLMs are used as we need to access to the output probability distributions.

that language models fail on most negated probes and are easily misled by misprimes added to the probing context. On the other hand, Zhao et al. (2021); Si et al. (2023); Webson and Pavlick (2022) found the presence of context biases in few-shot probing results. The works mentioned above focused on pinpointing issues affecting LLMs' factual prediction. Few studies were motivated to develop evaluation approaches insensitive to the hallucination-inducing causes. Recently, Raj et al. (2023) presented a framework for evaluating the consistency of LLMs based on accuracy. Zhu et al. (2023) designed a benchmark for assessing the robustness of LLMs to adversarial instruction attacks, measuring the corresponding end-to-end performance drops. Dong et al. (2023) proposed a new metric to measure factual knowledge capability under the bias caused by aliases (alternative names for entities or relations) by reducing the effect of entity and relation aliases in the factual probing. Without tackling other factors like the prompt framing effect and in-context interference (and their interactions), the scope of the study is limited.

3 LLMs in Hallucination

In this section, we investigate LLMs' accuracy under the influence of various hallucination-inducing causes mentioned above. We design five formats of prompts to demonstrate two categories of hallucination-inducing causes during knowledge probing (Table 1). Twelve LLMs with a wide range of parameter size (from 560 million to 30 billion parameters) are covered in this study and experiments (in Section 5), including foundation language models of OPT (Zhang et al., 2022), Galactica (Taylor et al., 2022), and instruction finetuned language model of BLOOMZ (Muennighoff et al., 2023), Vicuna (Zheng et al., 2023), Flan-T5 (Chung et al., 2022), WizardLM (Xu et al., 2023), Flan-UL2 (Tay, 2023; Tay et al., 2023), LLaMa-30b-instruct-2048 (upstage, 2023).

3.1 Effect of Prompt Framing on Accuracy

We design three probing templates to show the effect of prompt framing on LLMs, depicted below, and for each task, we use seven paraphrased prompts to ensure diversity:

Word Prediction (WP) Template: Given the "subject" and the prompt template, LLMs perform word prediction to complete the sentence, e.g., the template (1) in Table 1.

Prompt frames
(1) WP: [X] is located in _
(2) QA: Which country is [X] situated in?
(3) FC: Statement: [X] is located in [Y]. The statement is True of False?
In-context interference
(4) [Y]. Which country is the location of [X]?
(5) [Y_]. Which country is the location of [X]?

Table 1: Examples of designed probing task templates extending the P17 (a fact dataset containing 931 subject-object pairs with the "country" relation from T-REx (Elsahar et al., 2018)). [Y] is the object wrt the subject [X], [Y_] is an entity weakly related to [X].

Question-Answer (QA) Template: In the QA template, question prompts are constructed from paraphrasing templates in T-REx (Elsahar et al., 2018) targeting each fact. For example, a template "[X] is located in [Y]." for a triplet <[X], is located in, [Y]> can be paraphrased to "Which country is [X] situated in?".

Fact Checking (FC) Template: An FC prompt is designed as a verification statement based on a template in T-REx, e.g., "Statement: [X] is located in [Y]. The statement is True or False?". We build the positive checking probe (**FC-pos**) and negative checking probe (**FC-neg**) corresponding to whether the statement is factual or not. For a negative fact-checking prompt, we average the prediction accuracy for five random entities chosen from the same category.

The probing results are shown in Table 2 as accuracy in predicting P17 factual knowledge for each involved LLM under prompting biases presented in terms of WP, QA, and FC templates. The performances of LLMs in predicting the fact test data vary significantly under prompt variability. Abnormal performances of LLMs between QA and WP template-based probes (bold numbers of Vicuna-7b) and between the FC probes for positive and negative interference (bold numbers of BLOOMZ-1b1) are strong evidences of the prompt framing effect. The fluctuation under WP, QA, and FC templates shown as box plots in Figure 7 (Appendix A.1) further demonstrates the effect of prompt framing on the performances of LLMs.

3.2 Effect of In-context Interference

To explore the effect of in-context interference bias, we add probes with misprimed (Kassner and Schütze, 2020) interference by concatenating contexts in terms of factual/non-factual information preceding the associated QA prompt (template (2) in Table 1). Table 3 captures the accuracy of LLMs

LLMs	Size	WP	QA	FC-pos	FC-neg
BLOOMZ-560m	0.56	14.73	26.09	28.77	73.78
BLOOMZ-1b1	1.1	14.96	28.29	**0.11**	**99.89**
Galactica-1b3	1.3	2.36	46.43	86.05	12.29
OPT-2b7	2.7	28.27	55.67	75.80	22.07
BLOOMZ-3b	3	20.46	30.69	58.29	81.95
Vicuna-7b	7	**34.89**	**73.25**	91.19	85.67
BLOOMZ-7b1	7.1	26.26	33.72	88.32	64.98
Flan-T5-XXL	11	51.47	31.01	88.05	78.78
Vicuna-13b	13	38.96	78.15	90.87	89.68
WizardLM-13b	13	34.66	78.55	87.71	93.89
Flan-UL2	20	21.57	46.44	79.51	73.58
LLaMa-30b-ins.	30	67.94	87.72	96.99	86.69

Table 2: Accuracy of various LLMs in predicting P17 fact dataset. The performances of LLMs have undergone significant variations for different prompting templates. The unit of "size" is billion.

LLMs	×	[Y]	[Y_]
BLOOMZ-560m	25.91	66.17 (+40.26)	14.50 (-11.41)
BLOOMZ-1b1	27.74	64.02 (+36.28)	16.99 (-10.75)
Galactica-1b3	53.81	56.39 (+2.58)	10.42 (-43.39)
OPT-2b7	58.00	77.23 (+19.23)	19.83 (-38.17)
BLOOMZ-3b	35.38	**79.05 (+43.67)**	24.30 (-11.08)
Vicuna-7b	82.71	99.67 (+16.96)	**16.71 (-66.00)**
BLOOMZ-7b1	39.03	70.57 (+31.54)	26.40 (-12.63)
Flan-T5-XXL	37.85	42.53 (+4.68)	29.77 (-8.08)
Vicuna-13b	84.21	90.76 (+6.55)	44.58 (-39.63)
WizardLM-13b	85.61	55.75 (-29.86)	47.09 (-38.52)
Flan-UL2	33.44	47.58 (+14.14)	33.19 (-0.25)
LLaMa-30b-ins.	90.76	99.46 (+8.70)	47.78 (-42.98)

Table 3: The effect of probing the P17 fact dataset with QA templates (4) and (5) in Table 1, where "×" means experimental results with the original QA templates, "[Y]" means results using the factual information as in-context information, and "[Y_]" refers to results using non-factual in-context information of entities weakly related to "[X]".

in a comparative study using factual entity probes and misprimes consisting of weakly associated entities. We observe a strong interference effect from nonfactual antecedents for all 12 LLMs. A factual entity (positive interference) can improve the accuracy by up to +43.67 while a weakly related entity (negative interference) reduces the accuracy by -66.00 at most.

4 Methodology

In this section, we introduce MONITOR, a distance-based score, to assess how the factual knowledge of LLMs is affected by the previously mentioned prompt framing and in-context interference.

Firstly, we introduce a new variable (i) to represent hallucination-inducing in-context information into the initial knowledge representation triplet <*subject, relation, object*>. The newly formed knowledge representation quadruple can be expressed as $< s, r, o, i >$. The information i can be further categorized into two variables: we use a

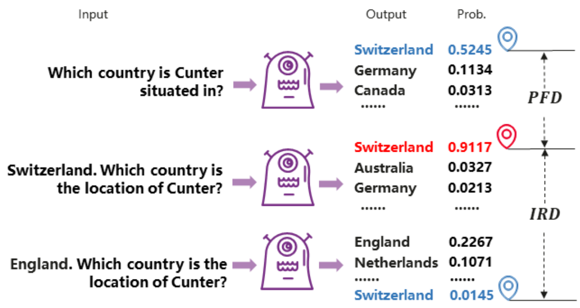


Figure 3: A primary anchor (in red font) corresponds to its multiple foreign anchors with different output probabilities (blue fonts) when an LLM is exposed to different prompts and context interference. "PFD" and "IRD" refer to the two distance measurements defined as the prompt-framing degree and interference-relevance degree.

factual object entity to implement a positive information i^+; and the negative information i^- represents interference when predicting o. For example, "England" is considered as an i^- when acting as a noisy condition to negatively affect an LLM in predicting a desirable outcome <*Switzerland*> for a fact <*Cunter, is located in, Switzerland*>. Corresponding to an object, $P(o|s, r, i)$ is the probability of the model generating the object o with the conditions of subject s, prompt framing expression r, and the in-context information i.

To quantify the effect of i on LLMs, we establish **"anchor"** as a reference point, which is the gold answer with its probability in the output space. A "primary anchor" (shown as the red font "Switzerland 0.9117" in Figure 3) is defined as an enforced-accurate answer with its probability produced by an LLM in response to a knowledge probe. A primary anchor is produced by prompting an LLM with a QA template prefixed with positive information i^+ (i.e. template (4) in Table 1). A primary anchor has multiple **foreign anchors** with various output probabilities (i.e., "Switzerland" in blue fonts in Figure 3) when an LLM is exposed to different prompts and in-context interference. Foreign anchors are generated using paraphrased Templates (2)[2] and (5)[3] presented in Table 1. By calculating the distance (using the probability changes) between a primary anchor and its corresponding foreign anchors in the influenced output space, we can measure how reliable an LLM is in predicting facts in the test set.

MONITOR consists of two distance-based mea-

[2] QA template without in-context information

[3] QA template with negative in-context interference

surement components: Prompt-framing Degree (PFD) and Interference-relevance Degree (IRD).

4.1 Prompt-framing Degree

The prompt-framing degree (PFD) is the mean distance between the output probability distributions of a primary anchor ($P(o|s, r, i^+)$) and those produced by the same LLM using prompting frames r_j probing the same fact without any add-on context (foreign anchors $P(o|s, r_j)$). PFD evaluates the similarity of two output probabilities between prompting frame relation expressions r (the basic prompt framing) and r_j. It is defined as:

$$PFD = \frac{1}{R}\sum_{j=1}^{R}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c, r, i^+)_l - P(o_c|s_c, r_j)_l| \quad (1)$$

where R is the count of prompt framing expressions for a subject, and the count of subject and object in a fact dataset is S, $c \in \{1, ..., S\}$. L_c is the length of the anchor in terms of the number of subwords in the c-th object. PFD is a cumulative metric for assessing an LLM's capability in producing output probability distributions sharing the same characteristics under various prompting frames. PFD has a value between 0 and 1. The smaller the value is, the more robust an LLM is under the effect of prompt framing.

4.2 Interference-relevance Degree

Interference-relevance Degree (IRD) is the distance between the output probability distributions of a primary anchor ($P(o|s, r, i^+)$) and the probability distributions generated by the same LLM under the influence of in-context interference (foreign anchors $P(o|s, r, i^-)$). IRD measures an LLM's capability to predict factual knowledge under the effect of in-context interference.

$$IRD = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{L_c}\sum_{l=1}^{L_c}|P(o_c|s_c, r, i^+)_l - P(o_c|s_c, r, i_m^-)_l| \quad (2)$$

We define the count of positive and negative information as one and M, respectively, corresponding to an object. IRD has a value between 0 and 1. As positive contextual information likely leads to factual knowledge generation, a smaller value of IRD indicates a lower level of effect from in-context interference biases.

4.3 MONITOR

The prompt-framing degree PFD and interference-relevance degree IRD are integrated to produce the proposed model knowledge reliability score (MONITOR). MONITOR captures the quadratic interaction of PFD and IRD, as illustrated in Eq 3 for a specified number of quadruples $< s, r, o, i >$, where the count of subject and object is S. A set of coefficients (α_{1-3}) is introduced to quantify the contributions from PFD, IRD, and their interaction on MONITOR. In this experiment, we consider an equal contribution scenario ($\alpha_1 = \alpha_2 = \alpha_3 = 0.33$). The smaller the value of MONITOR, the less degree an LLM is influenced by hallucination-induced factors when producing factual outputs. Taking the average output probabilities of primary anchors for an LLM as the denominator, MONITOR captures the degree of knowledge learned by an LLM when assessing its factual knowledge. MONITOR measures the effect of prompt framing and interference per unit of average primary anchor probability, demonstrating the strength of anchor representations.

LLMs are resource-hungry even during their inference phases. It is essential to ensure that an assessment metric is computation-efficient. Combining PFD, IRD, and their interaction in one metric can reduce the computation cost when evaluating factual reliability. Considering a fact dataset with R prompt frames, M negative interference, and one positive interference, there are $R*M$ combinations required to compute the average accuracy (and accuracy range). In comparison, we only require $R + (1 + M)$ combinations to compute MONITOR. The computation complexity for calculating MONITOR ($O(R+M)$) is considerably lower than that of accuracy ($O(R * M)$).

$$MONITOR = \frac{\sum_c^S \sqrt{\alpha_1 PFD^2 + \alpha_2 IRD^2 + \alpha_3 PFD * IRD}}{\sum_c^S \frac{1}{L_c}\sum_{l=1}^{L_c} P(o_c|s_c, r, i^+)_l} \quad (3)$$

5 Experiments and Results

In this section, we describe how to apply MONITOR to assess the factual knowledge of the 12 LLMs as mentioned above.

5.1 Data Setting

In this section, we describe how we develop a test corpus to accommodate prompts with various styles and in-context interference.

Expanding Probing Prompt: Based on 16,167 <subject, relation, object> triplets from T-REx (Elsahar et al., 2018), we develop QA probing prompts. We expand the probing prompt dataset by paraphrasing using GPT-4 (OpenAI, 2023) to

LLMs	MONITOR ↓	avg↑	max ↑	min ↑	probs ↑
BLOOMZ-560m	0.701	27.770	40.411	15.062	0.467
BLOOMZ-1b1	0.692	30.055	43.369	16.654	0.501
Galactica-1b3	0.747	22.936	39.414	9.427	0.637
OPT-2b7	0.637	25.599	37.117	11.347	0.360
BLOOMZ-3b	0.686	30.638	44.760	16.760	0.610
Vicuna-7b	0.504	38.194	59.727	18.361	0.884
BLOOMZ-7b1	0.632	36.232	49.328	22.870	0.613
Flan-T5-XXL	0.630	32.968	48.864	19.868	0.798
Vicuna-13b	0.484	44.882	65.499	26.967	0.862
WizardLM-13b	0.560	**51.477**	66.036	**33.076**	0.774
Flan-UL2	0.684	32.723	51.442	16.319	0.711
LLaMa-30b-ins.	**0.479**	50.798	**71.188**	30.516	**0.909**
Correlation	Pearson			p-value	
r(MONITOR,avg acc)	**-0.846**			**0.001**	

Table 4: Results are evaluated on FKTC with "bold" numbers indicating the best measurement over the same column category. The "avg", "max", and "min" mean the average, maximum, and minimum accuracy across the 20 fact datasets. The "probs." depicts the probabilities of primary anchors. "↓" means a smaller measurement wins.

create seven prompt frames for each triplet. In order to maintain diversity of prompts, we choose prompts with a similarity score (BLEU) below a threshold (0.7). Moreover, we manually check the paraphrased prompts to ensure validity.

Adding In-context Interference: Based on the QA prompts constructed above, we create a test dataset to explore the effectiveness of MONITOR with in-context interference. The corpus FKTC stands for "Factual Knowledge Test Corpus". Following the template patterns (Templates 4 and 5) in Table 1, we concatenate interference information (in terms of positive and negative in-context information) with the probing question for each subject. The negative information is entities from the same category weakly related to the corresponding subject, sampled from all objects that share the same relation. This process is applied to all expanded templates presented in Table 9 (Appendix A.2).

After applying these two processes (expanding the probing prompts and adding in-context interference) we produce 210,171 prompts focusing on 20 fact datasets.

5.2 Results and Analysis

5.2.1 Results on FKTC

The results evaluated on FKTC are shown in Table 4, and the results of each fact dataset are shown in Table 10 (Appendix A.3), where MONITOR and the average accuracy (avg acc) are recorded for each LLM across the 20 fact datasets in our experiments. Each LLM's minimal and maximal accuracy are also recorded to show the accuracy variability.

As shown in Table 4, LLaMa-30b-ins. stands out as the most capable (with the smallest MONITOR 0.479) LLM, followed by Vicuna-13b (0.484) and Vicuna-7b (0.504). Even though MONITOR is a fundamentally different from an end-to-end metric (like accuracy), it correlates significantly with the average accuracy (0.846 Pearson coefficient). MONITOR adds a dimension to a point-measured metric (like accuracy) to show factual reliability of LLMs under prompt and context variability.

As shown in Table 5 (bold italic fonts), MONITOR can differentiate LLMs, for example, BLOOMZ-3b and Vicuna-7b, with a similar average accuracy on P37, by considering distance and probability information. We further discuss this in Subsection 5.2.3.

We present a detailed view of the knowledge assessment of LLMs by drilling down into specific facts. Unlike the results mentioned above, showing a general trend, the results disclosed here show more detailed insights. As shown in Table 5, the overall winning LLM (i.e., LLaMa-30b-ins.) can lose its edge in predicting a particular fact (P37).

5.2.2 Accuracy Instability

We analyze the LLMs' "accuracy instability" when predicting P1412[4] with the results captured in Table 6 and Figure 4. A variety of statistics, including the base accuracy ("base acc") and standard deviation ("std") of an LLM's accuracy, are recorded for comparisons. A significant correlation is observed between accuracy standard deviation and MONITOR (0.754), demonstrating that a lower-MONITOR LLM is less likely to suffer from "accuracy instability" (Figure 8 in Appendix A.5). Furthermore, as shown in Figure 4, an LLM with a lower MONITOR has a smaller value of accuracy standard deviation when two LLMs with equivalent base accuracy are evaluated (bold fonts in Table 8). From an accuracy stability viewpoint, one may choose an LLM with a lower MONITOR. For example, we prefer Vicuna-13b over WizardLM-13b, as the MONITOR of Vicuna-13b is lower even though they have similar accuracy.

5.2.3 Resolution Characteristics

It can be observed in Table 4 that the correlation between MONITOR and average accuracy is significant. How should one use MONITOR when assessing the reliability of LLM knowledge?

[4] P1412: the fact dataset describing a relation of "languages spoken, written, and signed"

LLMs	P178			P108			P37		
	MONITOR ↓	avg acc ↑	probs. ↑	MONITOR ↓	avg acc ↑	probs. ↑	MONITOR ↓	avg acc ↑	probs. ↑
BLOOMZ-560m	0.594	53.260	0.471	0.947	2.634	0.313	0.669	33.142	0.679
BLOOMZ-1b1	0.492	56.752	0.684	0.853	7.454	0.191	0.662	39.679	0.751
Galactica-1b3	0.595	27.763	0.543	0.876	0.686	0.393	0.639	42.444	0.703
OPT-2b7	0.470	64.119	0.348	0.739	12.420	0.343	0.471	52.866	0.419
BLOOMZ-3b	0.624	50.460	0.863	0.858	17.639	0.436	***0.570***	***51.242***	0.797
Vicuna-7b	0.339	64.575	0.969	0.620	32.756	0.969	***0.432***	***51.384***	0.931
BLOOMZ-7b1	0.492	60.865	0.865	0.770	31.340	0.443	0.462	61.114	0.827
FLAN-T5-XXL	0.368	67.065	0.852	0.676	29.968	0.855	0.650	34.773	0.865
Vicuna-13b	0.327	77.787	0.955	0.632	39.951	0.899	**0.311**	69.590	0.942
WizardLM-13b	0.411	84.878	0.850	0.626	54.735	0.769	0.467	**69.907**	0.856
Flan-UL2	0.613	49.968	0.792	0.844	23.942	0.836	0.575	56.731	0.738
LLaMa-30b-ins.	**0.180**	**87.461**	**0.983**	**0.522**	**60.493**	**0.972**	0.411	63.109	**0.950**

Table 5: Performance of various LLMs in predicting factual knowledge captured in the P178, P108, and P37 fact datasets with "bold" numbers indicating the winning measurement over the same column category. P178, P108, and P37 are fact datasets representing relations of "developer", "employer" and "official language", respectively. The "bold and italic" fonts on P37 show how MONITOR can differentiate two LLMs (BLOOMZ-3b and Vicuna-7b) with similar average accuracy.

LLMs	MONITOR ↓	base acc ↑	std ↓
Flan-T5-XXL	0.772	51.713	31.023
OPT-2b7	0.536	64.027	12.087
Flan-UL2	0.706	67.029	33.981
BLOOMZ-560m	**0.490**	**70.888**	**17.253**
BLOOMZ-1b1	**0.426**	**71.932**	**11.891**
Galactica-1b3	0.659	74.086	26.576
BLOOMZ-7b	**0.472**	**78.922**	**19.252**
BLOOMZ-3b	**0.456**	**79.143**	**18.016**
Vicuna-7b	0.427	82.086	27.585
LLaMa-30b-ins.	0.543	85.340	34.131
WizardLM-13b	**0.425**	**91.960**	**8.978**
Vicuna-13b	**0.190**	**93.099**	**5.768**
Correlation	Pearson		p-value
r(MONITOR,std)	**0.754**		**0.001**

Table 6: LLMs with lower MONITOR are strongly correlated with smaller values of accuracy standard deviation, indicating less influence from prompt and context variability. "base acc" is the accuracy associated with the base prompt evaluated on the P1412 fact dataset.

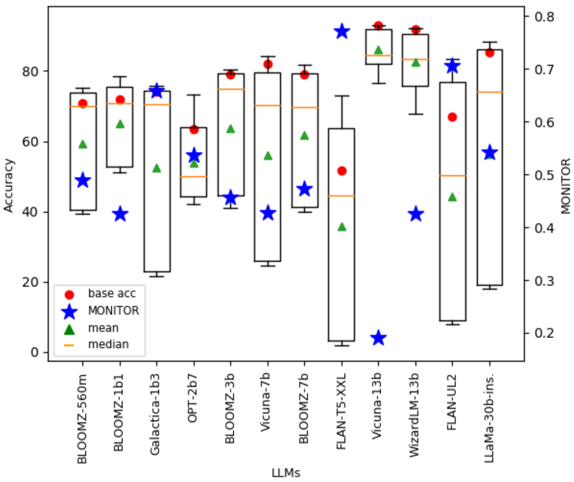


Figure 4: MONITOR can be used to differentiate LLMs' factual knowledge reliability when models with an equivalent base accuracy are evaluated. The box plots show the related distributions of accuracy when testing on P1412 fact dataset.

Base Prompt	What language is the official language of Haiti?		
effect	input	output BLOOMZ/Vicuna	prob. BLOOMZ/Vicuna
pos. context	French.{base}	French/French	0.761/**0.928**
neg. context	Irish.{base}	French/French	0.411/**0.622**
framing	{base}	French/French	0.527/**0.849**

Table 7: Vicuna-7b outperforms BLOOMZ-3b in MONITOR when evaluated on the P37 fact dataset by producing correct answers with higher output probabilities in response to positive, negative in-context interference and prompt framing effect. {base} refers to the base prompt.

We regard MONITOR as a high-resolution metric because it directly uses output probabilities and their changes (in terms of anchored distance) induced by hallucination factors. MONITOR considers both the output (nominal or qualitative data) and the probability of the output (quantitative information). Comparatively, assessing LLMs' knowledge with an end-to-end metric, such as accuracy, is purely reliant on a nominal output from the softmax layer of a transformer. It is shown in Table 5 that two LLMs (BLOOMZ-3b vs. Vicuna-7b) with almost identical average accuracy on P37 fact dataset have two distinctive values of MONITOR (0.570 vs 0.432). Delving into the log file of the inference task, we gain in-depth insights into why Vicuna-7b outperforms BLOOMZ-3b in the reliability score. As shown in Table 7, despite their similarities in the accuracy measurement, Vicuna-7b has much higher output probabilities than those of BLOOMZ-3b, contributing to the discrepancies in MONITOR.

Additionally, we plot out the probability distribution of the above two LLMs with almost identical average accuracy but very distinctive MONITOR (Figure 9 in Appendix A.6). It can be observed that a more reliable LLM based on MONITOR, Vicuna-7b, has a much higher percentage of solid output probability (i.e., ≥ 0.8) than those of a volatile LLM (BLOOMZ-3b). It is recommended to adopt MONITOR when using accuracy alone cannot differentiate LLMs' knowledge reliability.

Cost	MONITOR	Average Accuracy	MONITOR-saved
GPU hours	14.4	42.7	2.97X

Table 8: GPU hours consumed calculating MONITOR and average accuracy on P1412 fact dataset for LLaMa-30b-ins."MONITOR-saved" denotes that GPU hours saved from using MONITOR compared to accuracy.

5.2.4 Lower Computation Cost

We compare the GPU hours consumed by LLaMa-30b-ins. in producing MONITOR and a full-scale accuracy reliability score (average accuracy). The experiment is to test the model on a specific fact dataset (P1412) using 8 NVIDIA V100 GPUs. It can be observed in Table 8 that using MONITOR leads to a 2.97-fold resource saving in GPU hours compared to applying an accuracy metric to a factual reliability evaluation. MONITOR is an economical method to add a dimension to LLM knowledge assessment when performing a full-scale reliability study on accuracy is not an option.

6 Discussion

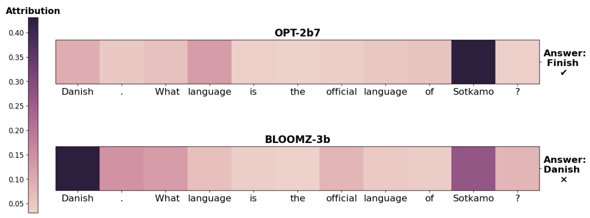


Figure 5: Visualizing model behaviors of BLOOMZ-3b and OPT-2b7 under the influence of an input with misprimed in-context interference. The input is *"Danish. What language is the official language of Sotkamo?"*.

6.1 Attribution of In-Context Interference

To demonstrate the resilience of LLMs with different MONITOR, we conduct an additional experiment by applying the Integrated Gradients (Sundararajan et al., 2017) technique implemented in Sarti et al. (2023). By examining and visualizing the attribution of input features to the model's outputs, we can infer the reliability of LLMs with different MONITOR. We study the behaviors of two LLMs (OPT-2b7 vs. BLOOMZ-3b) with distinctive values of MONITOR (0.471 vs. 0.570). The heat map shown in Figure 5 illustrates that a more reliable model with a lower MONITOR, OPT-2b7, is less influenced by in-context interference.

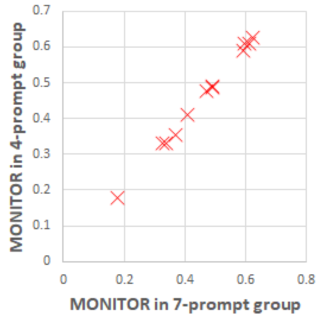


Figure 6: Significant correlation of MONITOR between the 7-prompt group and the 4-prompt group when assessing the reliability of 12 LLMs in the P178 fact dataset.

6.2 Prompt Ablation

We design an ablation study to investigate the consistency of MONITOR across different prompt settings by analyzing the MONITOR in the P178 fact dataset. The MONITOR from an expanded prompts group setting (consisting of seven prompts) and a sub-sampled group with four prompts are captured in Figure 6. We observe a strong linear correlation between MONITOR of the expanded group and those from the sub-sampled group, indicating the scalability of MONITOR across prompt settings. Additionally, it is noted that MONITOR ranks LLMs in a consistent order for different prompt settings as show in Figure 10 (Appendix A.7).

7 Conclusion

In this paper, we show that large language models are subject to the influence of various hallucination-inducing causes. We propose a novel distance-based metric, directly computing the output probabilities and their changes to address "accuracy instability" caused by the prompt framing effect and in-context interference. A comprehensive set of experiments demonstrates that the proposed MONITOR is a high-resolution economic method suitable for evaluating the reliability of large language model knowledge. MONITOR can be used in conjunction with an end-to-end metric (i.e., accuracy) as part of a multi-dimensional approach to LLM knowledge evaluation. The constructed FKTC, consisting of 210,171 question answering prompts on 20 fact datasets, will be made available to the public to foster research along this line.

Limitations

We focus on proposing MONITOR to assess the reliability of factual knowledge of LLMs during knowledge probing. Whether MONITOR can be generalized to a wider scope of tasks (e.g., summarization) warrants a future study. Additionally, the initial setup of contribution coefficients of PFD, IRD, and their interaction on MONITOR should be further investigated to establish an empirical benchmark. Currently MONITOR applies exact matching to obtain anchors to measure the reliability of LLM knowledge. Extending the automatic evaluation to anchors consisting of sentences is challenging. Our approach needs to access to the output probability distributions of an LLM, therefore is not applicable to SOTA commercialized LLMs such as GPT4. Additionally, FKTC is developed based on the latest version of T-REx benchmark dataset. The quality of the factual knowledge contents in FKTC is reliant on the alignment accuracy of T-REx. Even though we could argue that FKTC has already accommodated over 210 thousands prompts in the gold dataset to successfully support MONITOR in assessing LLMs behaviors under prompt and context variability. It can still be extended to host more knowledge categories.

Licensing and Intended Use

FKTC is based on a widely adopted T-REx benchmark dataset, which is publicly available under a Creative Commons Attribution-ShareAlike 4.0 International License. FKTC will be released to the public under the same license, consistent with the original intended use.

References

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1860–1874. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *CoRR*, abs/2307.03109.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for generative language models. *CoRR*, abs/2305.10519.

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

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Akshat Gupta. 2023. Probing quantifier comprehension in large language models. *CoRR*, abs/2306.07384.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. In *NeurIPS*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7811–7818. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *NeurIPS*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15991–16111. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2023. Semantic consistency for assuring reliability of large language models. *CoRR*, abs/2308.09138.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar van der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pages 421–435. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4222–4235. Association for Computational Linguistics.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep 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 3319–3328. PMLR.

Yi Tay. 2023. A New Open Source Flan 20B with UL2. `https://www.yitay.net/blog/flan-ul2-20b`.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *CoRR*, abs/2211.09085.

upstage. 2023. LLaMa-30b-instruct-2048. `https://huggingface.co/upstage/llama-30b-instruct-2048`.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2300–2344. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, and Xing Xie. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *CoRR*, abs/2306.04528.

A Appendix

A.1 Prompt Framing Effect

We paraphrase each fact dataset in three prompting templates (WP, QA, and FC) so that each template can be used to produce seven prompts. For example, the template "Which country is the location of [X]?" could be paraphrased as: "Which country is [X] situated in?", "Which country can [X] be found?", "Which country is the geographical position of [X]?", "Which country is the site of [X]?", "In Which country is [X] situated?", "Whereabouts is [X] located?". In this way, context diversity and semantic invariance are guaranteed. Figure 7 shows the "accuracy instability" of LLMs under the effect of prompt framing in predicting P17 facts based on three tasks (WP, QA, and FC).

A.2 Templates Examples

Table 9 shows all templates and corresponding prompts on 20 fact datasets.

A.3 MONITOR for All LLMs Experimented on FKTC

Table 10 shows the results of various LLMs evaluated on each fact dataset from FKTC.

A.4 Correlation between MONITOR and Accuracy

Table 11 shows the Pearson correlation between MONITOR and average accuracy, evaluated on the 20 fact datasets from FKTC corpus.

A.5 Correlation between MONITOR and Accuracy Standard Deviation

Figure 8 shows a lower-MONITOR LLM is less likely to suffer from "accuracy instability".

A.6 Probability Distribution

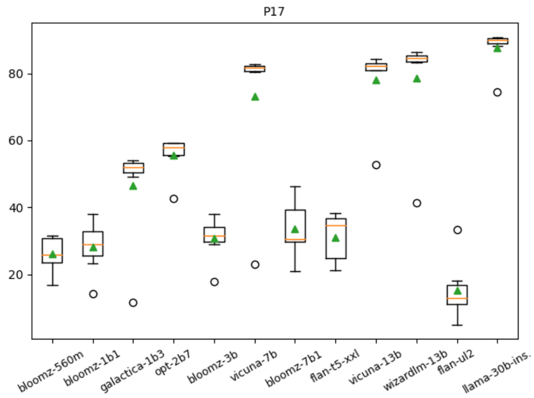
Figure 9 shows the probability distribution of two LLMs (BLOOMZ-3b and Vicuna-7b) with almost identical average accuracy but very distinctive MONITOR.

A.7 Consistency across Different Prompts Settings

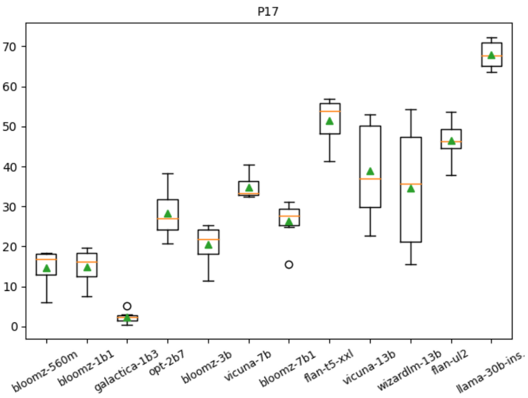
Figure 10 shows a consistent order in ranking LLMs across different prompt settings (7-prompts group VS. 4-prompts group).

A.8 Analysis on LLMs Scale

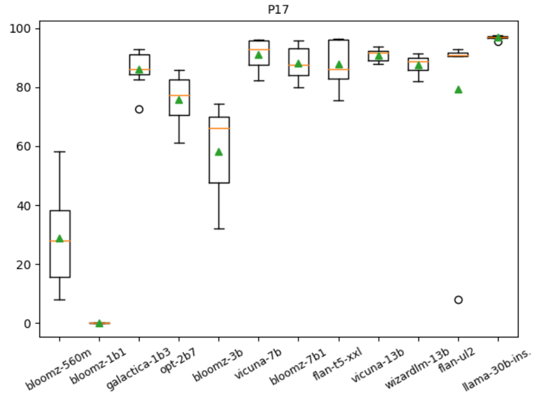
To further verify if MONITOR of LLMs follows the law of scaling, where larger LLMs are more knowledge-reliable, we present how MONITOR changes across BLOOMZ series for each specific fact dataset (shown in Figure 11). While MONITOR of LLMs may not conform to the scaling law at the granularity of each fact, their aggregated values in a comprehensive scope of experiments do follow the rule of scale (shown in Figures 11-12).



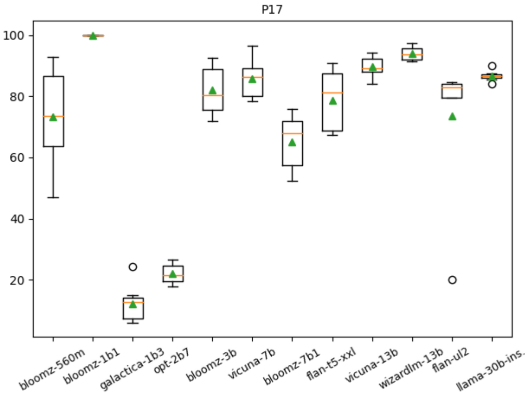
(a) QA



(b) WP



(c) FC-pos



(d) FC-neg

Figure 7: Box plots show the "accuracy instability" of LLMs under the effect of prompt framing in predicting P17 based on three tasks (WP, QA, and FC).

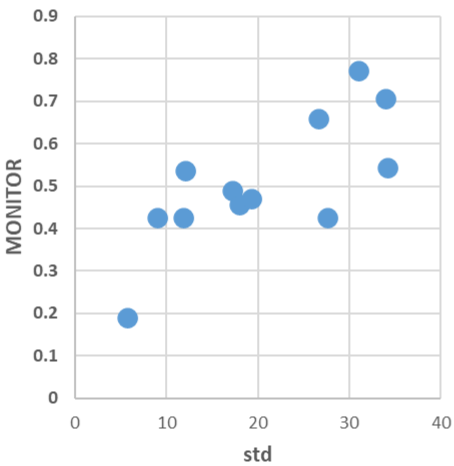


Figure 8: A significant correlation between MONITOR and accuracy standard deviation when testing the 12 LLMs on P1412 fact dataset, indicating lower-MONITOR models are less likely to suffer from the "accuracy instability" issue.

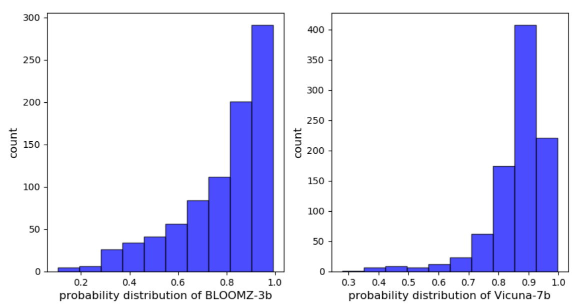


Figure 9: A comparison of the probability distribution of anchors between BLOOMZ-3b and Vicuna-7b on P37. The population percentages with a solid probability (i.e., ≥ 0.8) are **59%** and **85%** for BLOOMZ-3b and Vicuna-7b, respectively.

Fact	Relation	Object Type	Template	Prompt example	Count
P17	country	sovereign state	[X] is located in [Y].	Which country is the location of [X]?	12,103
P19	place of birth	city	[X] was born in [Y].	Where was [X] born?	12,272
P20	place of death	city	[X] died in [Y].	In what place did [X] pass away?	12,389
P27	country of citizenship	sovereign state	[X] is [Y] citizen.	What country is [X] a citizen of?	12,558
P30	continent	continent	[X] is located in [Y].	Which continent is [X] located in?	12,675
P37	official language	language	The official language of [X] is [Y].	What language is the official language of [X]?	12,558
P101	field of work	organization	[X] works in the field of [Y].	What is [X]'s area of expertise?	9,048
P103	native language	Indo-European languages	The native language of [X] is [Y].	What is the native language of [X]?	12,701
P108	employer	business	[X] works for [Y].	Which organization does [X] work for?	4,979
P127	owned by	company	[X] is owned by[Y].	Which company is the owner of [X]?	7,059
P159	headquarters location	sovereign state	The headquarter of [X] is in [Y] .	In what city is [X] headquartered?	12,571
P176	manufacturer	manufacturer or producer	[X] is produced by [Y].	What is the manufacturer of [X]?	12,766
P178	developer	organisation	[X] is developed by [Y]	Which company is the creator of [X]?	7,696
P264	record label	record label	[X] is represented by music label [Y].	What is the record label for [X]?	5,577
P276	location	sovereign state	[X] is located in [Y].	What is the location of[X]?	12,467
P364	original language of film or TV show	Nostratic languages	The original language of [X] is [Y].	What is the native language of [X]?	11,128
P495	country of origin	sovereign state	[X] was created in [Y].	Which country was [X] created in?	11,817
P740	location of formation	sovereign state	[X] was founded in [Y].	Which city was [X] founded in?	12,168
P1376	capital of	country	[X] is the capital of [Y].	Which country's capital is [X]?	3,042
P1412	languages spoken, written or signed	Indo-European languages	[X] used to communicate in [Y].	What language did [X] previously speak to communicate?	12,597

Table 9: Examples of template for different fact datasets and the corresponding prompts we build in this work.

Fact Dataset	BLOOMZ -560m	BLOOMZ -1b1	Galactica -1b3	OPT -2b7	BLOOMZ -3b	Vicuna -7b	BLOOMZ -7b1	Flan-T5 -XXL	Vicuna -13b	WizardLM -13b	Flan -UL2	LLaMa- 30b-ins.
P17	0.782	0.780	0.852	0.541	0.785	0.523	0.714	0.690	0.544	0.602	0.788	0.395
P19	0.866	0.927	0.914	0.858	0.898	0.719	0.873	0.882	0.629	0.752	0.918	0.817
P20	0.810	0.926	0.942	0.849	0.921	0.671	0.873	0.888	0.667	0.725	0.893	0.803
P27	0.704	0.746	0.868	0.597	0.706	0.460	0.724	0.674	0.489	0.573	0.786	0.490
P30	0.809	0.839	0.801	0.748	0.887	0.652	0.546	0.670	0.611	0.680	0.815	0.617
P37	0.669	0.662	0.639	0.471	0.570	0.432	0.462	0.650	0.311	0.467	0.575	0.411
P101	0.899	0.822	0.919	0.888	0.877	0.816	0.838	0.879	0.823	0.927	0.858	0.857
P103	0.512	0.515	0.671	0.468	0.457	0.424	0.451	0.599	0.296	0.506	0.561	0.410
P108	0.947	0.853	0.876	0.739	0.858	0.620	0.770	0.676	0.632	0.626	0.844	0.522
P127	0.522	0.613	0.676	0.627	0.712	0.547	0.545	0.437	0.382	0.438	0.621	0.346
P159	0.829	0.851	0.858	0.755	0.800	0.523	0.751	0.731	0.478	0.479	0.758	0.454
P176	0.684	0.461	0.457	0.527	0.609	0.244	0.632	0.290	0.437	0.467	0.518	0.322
P178	0.594	0.492	0.595	0.470	0.624	0.339	0.492	0.368	0.327	0.411	0.613	0.180
P264	0.887	0.923	0.916	0.863	0.748	0.678	0.887	0.883	0.606	0.661	0.799	0.560
P276	0.707	0.699	0.751	0.650	0.737	0.535	0.674	0.639	0.489	0.557	0.664	0.515
P364	0.756	0.762	0.850	0.662	0.780	0.576	0.751	0.786	0.619	0.714	0.774	0.599
P495	0.802	0.834	0.868	0.661	0.695	0.413	0.715	0.716	0.476	0.530	0.790	0.499
P740	0.941	0.961	0.961	0.858	0.931	0.689	0.905	0.837	0.646	0.669	0.882	0.647
P1376	0.505	0.451	0.606	0.602	0.352	0.299	0.202	0.158	0.501	0.555	0.202	0.079
P1412	0.490	0.426	0.659	0.536	0.456	0.427	0.472	0.772	0.190	0.425	0.706	0.543

Table 10: MONITOR for all involved LLMs experimented on FKTC corpus.

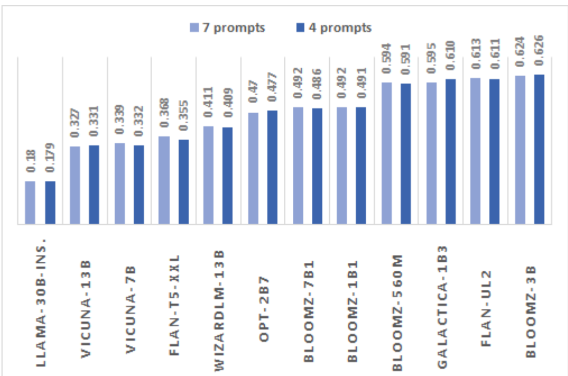


Figure 10: The consistency of MONITOR when assessing LLM's factual reliability in predicting P178 facts across different prompts settings.

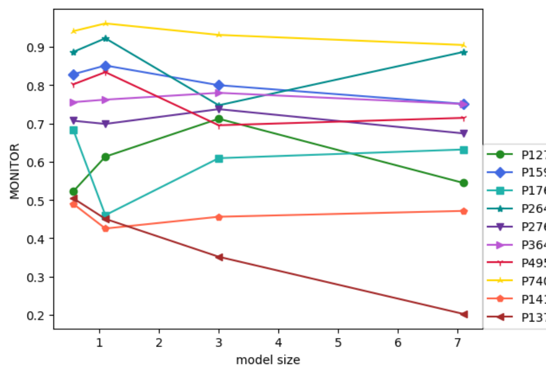


Figure 11: The BLOOMZ series adheres to the scale law for the specific facts with smaller MONITOR for bigger models. The horizontal axis represents the model's size in billions, and the vertical axis represents the results of MONITOR.

Pearson	P17	P19	P20	P27	P30	P37	P101	P103	P108	P127
correlation	-0.579	-0.709	-0.685	-0.826	-0.648	-0.867	-0.474	-0.767	-0.889	-0.926
p-value	0.048	0.009	0.013	0.001	0.023	0.001	0.119	0.004	0.001	0.001
	P159	P176	P178	P264	P276	P364	P495	P740	P1376	P1412
correlation	-0.941	-0.941	-0.828	-0.950	-0.703	-0.740	-0.899	-0.919	-0.872	-0.900
p-value	0.001	0.001	0.001	0.001	0.011	0.006	0.001	0.001	0.001	0.001

Table 11: Pearson correlation between MONITOR and the average accuracy, evaluated on FKTC corpus.

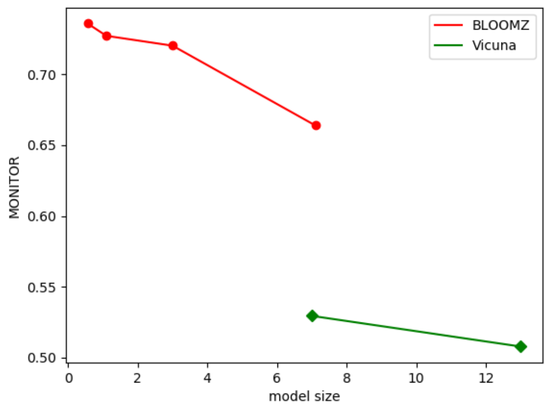


Figure 12: The BLOOMZ and Vicuna series adhere to the scale law based on the overall MONITOR results obtained from experiments on 20 fact datasets. The horizontal axis represents the size of a model in billions, and the vertical axis represents the results of MONITOR.