Estimating LLM Consistency: A User Baseline vs Surrogate Metrics

Xiaoyuan Wu[1], Weiran Lin[1], Omer Akgul[2, 1], Lujo Bauer[1]
[1]Carnegie Mellon University [2]RSAC Labs
Correspondence: wxyowen@cmu.edu

Abstract

Large language models (LLMs) are prone to hallucinations and sensitive to prompt perturbations, often resulting in inconsistent or unreliable generated text. Different methods have been proposed to mitigate such hallucinations and fragility, one of which is to measure the consistency of LLM responses—the model's confidence in the response or likelihood of generating a similar response when resampled. In previous work, measuring LLM response consistency often relied on calculating the probability of a response appearing within a pool of resampled responses, analyzing internal states, or evaluating logits of resopnses. However, it was not clear how well these approaches approximated users' perceptions of consistency of LLM responses. To find out, we performed a user study ($n = 2,976$) demonstrating that current methods for measuring LLM response consistency typically do not align well with humans' perceptions of LLM consistency. We propose a logit-based ensemble method for estimating LLM consistency and show that our method matches the performance of the best-performing existing metric in estimating human ratings of LLM consistency. Our results suggest that methods for estimating LLM consistency without human evaluation are sufficiently imperfect to warrant broader use of evaluation with human input; this would avoid misjudging the adequacy of models because of the imperfections of automated consistency metrics.

1 Introduction

Large language models (LLMs) have seen rapid adoption across a multitude of domains despite numerous inherent limitations, such as hallucinating responses and being fragile to adversarial inputs. Hallucinations can have disastrous consequences in high-stakes fields like healthcare and law (Merken, 2025), prompting concern even as adoption continues (Tiffin and Fraser, 2025; Metnick et al., 2024). Further, the fragility of LLMs enables a range of misuses that could harm users (Kumar and Lakkaraju, 2024; Lin et al., 2025). For instance, Lin et al. (2025) showed that minor, unnoticed changes to suggested prompts can result in outputs with biases controlled by an adversary.

Researchers have suggested that some of these issues correlate with the inconsistency of LLMs, which is generally defined as their tendency to generate low-confidence responses or conflicting responses when the same prompt is resampled (Manakul et al., 2023). Accurately estimating LLM consistency is important because it can support multiple critical applications: predicting whether an answer is factual, thereby improving reliability in high-stakes domains (Duan et al., 2024); detecting fragile or malicious prompts (Lin et al., 2025); informing membership-inference attacks (Mattern et al., 2023); or signaling to users the level of trust they should place in LLM outputs (Kapoor et al., 2024; Ruggieri and Pugnana, 2025).

Consequently, many attempts have been made to define and measure the consistency of LLMs. This body of work can roughly be divided into two categories: (1) estimating consistency based on LLMs' internal states or logits and (2) estimating consistency based on resampling LLMs' responses. While the former is more computationally efficient, the latter is empirically well-grounded (Kuhn et al., 2023; Qiu and Miikkulainen, 2024) and applicable even without white-box access to a model (Lin et al., 2024).

Sampling-based estimation methods fundamentally rely on sampling from an LLM–prompt pair and comparing the outputs using a comparison function. This function differs between works—e.g., Duan et al. (2024) vs Manakul et al. (2023)—but partially boils down to estimating semantic similarity of responses. However, to our knowledge, none have based their estimation of LLM model consistency on user-based comparisons, the ground truth for semantic similarity (Bowman et al., 2015;

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30530–30544
November 4-9, 2025

Agirre et al., 2014). Further, existing uncertainty estimation methods have not yet demonstrated reliability or alignment with human judgments of LLMs' consistency. Without a baseline, consistency metrics are typically evaluated by their ability to predict whether a model output is factual (Kuhn et al., 2023; Qiu and Miikkulainen, 2024; Duan et al., 2024; Zhang et al., 2024; Kapoor et al., 2024). Joining recent work (Novikova et al., 2025), we argue that consistency might be able to predict accuracy but is fundamentally an independent property of the model-prompt pair. There should be a metric that conveys to what degree a model, when provided a prompt, will repeatedly produce responses with equivalent meaning. We further posit that the baseline for measuring consistency is reliant on the comparison of semantics, the ground truth of which is defined by human judgement (Bowman et al., 2015; Agirre et al., 2014; Nguyen et al., 2014; Raj et al., 2025).

We aim to fill this gap by using user-based comparisons to estimate the consistency of LLMs, and by investigating whether an automated metric can come close to simulating users' judgments. Specifically, we conduct a user study with 2,976 participants to collect semantic similarity ratings between a sample of 10 responses to each of 100 prompts, totaling 14,880 comparisons. We then calculate response-level (one score per response) and prompt-level (one score per prompt) consistency, establishing a user-based LLM consistency baseline. Through a series of experiments, we show that existing metrics of measuring LLM response uncertainty do not align well with human judgements collected in this study. We further show that an ensemble of logit-based scores is as similar to human judgement as the best-performing of the other methods we tested. We find that the discrepancy between existing metrics and human judgements fluctuates between models and between datasets, with previous metrics being less similar to human judgements on real-world prompts then on synthetic prompts. Based on our results, we advocate for more human-based LLM response consistency evaluation in future work.

We structure the remainder of the paper in four sections: Section 2 details background work; Section 3 describes our model- and prompt-selection strategy, user study, and consistency calculation; Section 4 reports our experiments and comparisons; finally, we summarize and discuss implications of our work in Section 5.

2 Related Work

Here, we review work in consistency estimation of LLMs with resampling and internal model states. (2.1). We also cover work that investigated logit-based metrics to estimate LLM consistency (2.2). Finally, we reiterate our study motivation (2.3).

2.1 LLM Consistency from Sampling and Internal States

Prior definitions of model consistency (Lakshminarayanan et al., 2017) do not apply to the near-infinite output space, auto-regressive nature of LLMs (Kuhn et al., 2023). As a result, a line of work has emerged to define uncertainty of LLM responses (sometimes defined as the confidence of the model in a response, or how likely a response is to be generated when resampled).

Sampling When estimating uncertainty, researchers often resample multiple responses from a model-prompt pair. A chosen response is then compared to the set of responses, ultimately calculating an uncertainty score. For instance, (Kuhn et al., 2023) establishes a set of semantically equivalent responses within the sampled set using entailment. The probabilities of any one of these semantically equivalent groups are fed into a predictive entropy based formulation to obtain an uncertainty score. Follow-up work has suggested similar sampling-based metrics (Qiu and Miikkulainen, 2024; Duan et al., 2024).

Internal state representation Some prior work has sidestepped the definition of uncertainty to directly predict if a given answer is right or wrong. This body of work (sometimes also referred to as uncertainty estimation), uses techniques such as asking follow-ups (Kadavath et al., 2022; Sam et al., 2025) and training auxiliary prediction models based on internal states (Kapoor et al., 2024; Kossen et al., 2024).

2.2 Probability- and Logit-Based Uncertainty Metrics

Existing work suggests that the confidence of individual token generation can suggest the consistency of LLM responses (Manakul et al., 2023). Borrowing definitions from Manakul et al. (2023) , we denote the probability distribution of an individual token generated as p, and the entropy of individual token generation H, $H = -\sum_i p_i log(p_i)$. Existing metrics include the average of the minus log

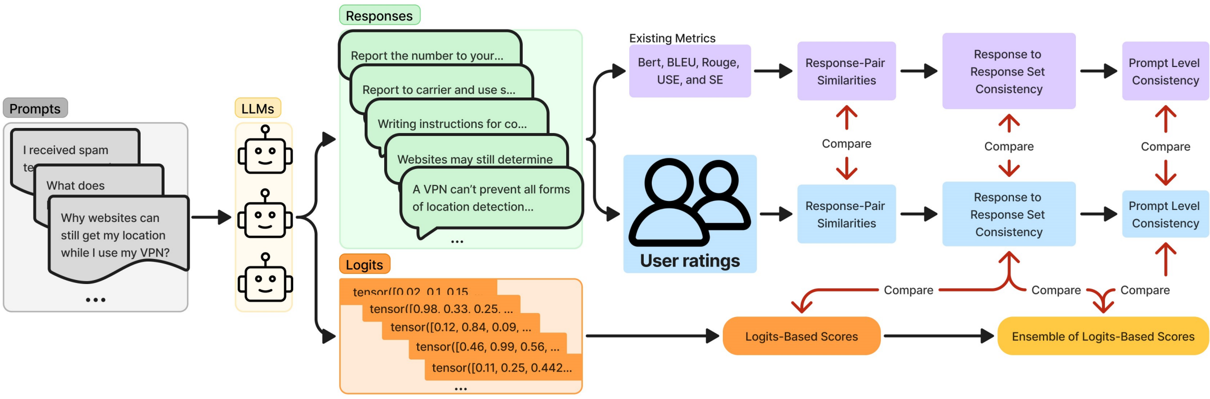


Figure 1: Overview of our study design

probability $Avg(-log(p))$, maximum of the minus log probability $Max(-log(p))$, average of entropy $Avg(H)$ and maximum of entropy $Max(H)$. In this paper, we ensemble these metrics (and variants) in 3.3.

Difference of Logits Ratio Loss Besides probability-based (i.e., p-based) uncertainty metrics, we also find logit-based uncertainty metrics to suggest LLM response consistency. We denote the logits of individual token generation as l, $p = softmax(l)$. Specifically, we denote the largest, second largest, third largest, and fourth largest logit of a token generation as l_1, l_2, l_3, and l_4 correspondingly. The Difference of Logits Ratio Loss (DLR loss), formally $\frac{l_1 - l_2}{[l_1 - (l_3 + l_4)]/2}$, is a loss function commonly used by adversarial attacks to suggest how confident a model is (Croce and Hein, 2020). In this paper, we borrow the DLR loss as a logit-based uncertainty of a token generation to estimate the consistency of LLM responses.

Since probability-based uncertainty estimation is ultimately derived from token logits, for the rest of the paper, we call this group of uncertainty estimation logit-based methods.

2.3 Importance of User Evaluation in NLP Tasks

Human evaluations are important in improving and assessing the quality of natural language processing (NLP) (Gritta et al., 2024; Boyd-Graber et al., 2022; Blodgett et al., 2024). Researchers have investigated how well machines are able to translate sentences into another language, compare sentences for semantic similarity, and many other NLP tasks with human evaluations (Chatzikoumi, 2020; Graham et al., 2017). This trend has continued in the LLM era (Ouyang et al., 2022; Bai et al., 2022)

While prior work has found ways to estimate the consistency of responses produced by LLMs, we argue that consistency is defined by what degree humans see responses as similar to each other. Using this definition, we establish a baseline through a user study.

3 Methods

In this study, we seek to understand how accurately existing metrics can approximate users' perceptions of LLM consistency. We explore ways to create novel and more efficient ways of estimating LLM consistency without the need for expensive user studies or generating multiple responses per prompt. We provide an overview of our study design in Figure 1.

In this section, we first describe the prompts, models, and responses used in the study (3.1). We then describe in detail how we conducted the user study with 2,976 participants (3.2). Finally, we explain how we calculate consistency, both with existing metrics and our user-based method (3.3).

3.1 Prompts, Models, and Responses

Prompts Following prior work (Geng et al., 2024), we selected two prompt sets: (1) to ensure comparability with prior research, we use a dataset of open-ended questions from *CoQA* (Reddy et al., 2019), which is commonly adopted in the literature; and (2) to better represent real-world use cases, we include a sample of 50 prompts from LMSYS-Chat-1M (Zheng et al., 2023).

Each entry in the *CoQA* dataset contains a story with a series of questions about the story. For our study, we only used the first question in each series, as many following questions are dependent on the previous question or answer for context. We randomly sampled 50 story–question pairs from the *CoQA* dataset.

Real-world prompts from the *LMSYS* dataset are more diverse than those in *CoQA*. Hence, we used labels in the data to filter out prompts that are were not in English or *flagged*[1] (e.g., harassment, violence). Next, we randomly sampled 100 prompts from the filtered *LMSYS* dataset and manually removed 48 prompts: 14 coding questions, 9 about the LLM itself, 9 that were inappropriate (e.g., sexual), 6 that contain nonsensical sentences, 5 that we expected to have non-English responses (e.g., translations), 3 asking for answers over 1,500 words, and 2 that contain time-sensitive information. We excluded coding questions since participants might not have the background to rate coding-related responses. We randomly removed 2 prompts from the remaining 52 to obtain a final sample of 50.

Models A generalizable definition of consistency should be model-agnostic. Thus, we used three open-weights models from competing institutions to generate responses: Llama-3.2-3B-Instruct (Llama), Gemma-2-9B-it (Gemma), and Mistral-7B-Instruct-v0.3 (Mistral). Each was the most recent publicly available version of their repective families, was widely used in previous work, and met our hardware memory constraints (Nvidia RTX A6000). We used open-weights models as they provided us access to logits, which were necessary for our calculation described in Section 3.3, and enabled reproducibility of our study (Ma et al., 2024). We left temperature settings unchanged from the default configuration because there is no consensus from pervious work on optimal temperature for uncertainty estimation (Cecere et al., 2025; Du et al., 2025; Wang et al., 2023; Renze and Guven, 2024; Zou et al., 2023)

Responses The 100 prompts were randomly assigned to the three LLMs; Gemma received 34, Llama 33, and Mistral 33. Using the assigned model, we then generated 10 responses per prompt, totaling 1,000 responses. We took this approach as previous work showed that a set of 10 responses adequately captures the semantic diversity of the response space for uncertainty estimation purposes (Kuhn et al., 2023; Qiu and Miikkulainen, 2024). Response generation took 30 minutes of GPU time.

[1]by OpenAI Moderation (Zheng et al., 2023)

Model Name	Num. Prompts	Num. Responses	Num. Pairs of Unique Responses
Gemma	34	340	793
Llama	33	330	826
Mistral	33	330	1,319

Table 1: Breakdown of number of prompts, responses, and pairs of unique responses per model.

3.2 User Study

While many metrics have been created to approximate LLM consistency, a reporting of LLM consistency based on users' perceptions has been largely missing from the literature. Our user-rating based approach establishes a baseline against which other metrics can be evaluated. We recruited users from Prolific and required that participants be 18 years or older, reside in the United States, read and type in English fluently, and have at least a 95% approval rate on the platform. We use Prolific over MTurk because recent work has shown its superior data quality (Tang et al., 2022; Moss and Litman, 2018).

Each participant was given an introduction to the study before seeing instructions for, and examples of, how to rate semantic similarity between a pair of sentences using a 6-point scale (Appendix A.1). We adopted this commonly-used scale, with explanations and examples, from Agirre et al. (2014). After the instructions, each participant rated the semantic similarity of five pairs of responses from five different prompts. Participants were compensated $1; the median survey completion time was 5 minutes and 13 seconds. We collected participants' Prolific IDs for compensating those who completed our survey. We removed Prolific IDs before doing any analysis. Lastly, we collected participants' demographics. The sentence pairs and the order in which they were shown to participants were randomized. Our study was approved by our institution's ethics review board.

For each prompt, we generated 10 responses, yielding $\binom{10}{2} = 45$ pairs of responses. With 100 prompts, we obtained a total of 4,500 pairs of responses, of which 2,938 were unique. More details about the number of prompts, responses, and pairs of unique responses for each model are described in Table 1.

3.3 Calculating and Comparing Consistency

Here, we outline how we calculated consistency scores (we calculated similarities between pairs of responses, consistency for each response, and consistency for each prompt) and compared user-based scores to previous work.

User ratings of similarities We first aggregate user ratings of similarity between a pair of responses r_a and r_b by averaging ratings of $n \geq 5$ participants after removing the highest and lowest scores (Curran, 2016). Formally,

$$s(r_a, r_b) = \frac{\sum_{k=1}^{n-2} h_k}{n-2} \quad (1)$$

Response-pair level comparison To understand how existing metrics used for LLM consistency compare to human ratings, we compared them at the response-pair level. For each prompt (e.g., "What is Wh in batteries?"), 10 responses yielded 45 response-pairs, totaling 4,500 response-pairs for the 100 prompts in our study. For these response-pairs, we calculated the Spearman correlation coefficient ρ between user ratings of similarity (s_{user}) and each of the four existing metrics: *Bert* (s_{Bert}), *BLEU* (s_{BLEU}), *Rouge* (s_{Rouge}), and *USE* (s_{USE}). We used the Spearman correlation coefficient ρ because human ratings were on a 6-point Likert scale, and we did not assume linear correlation between human ratings and existing metrics. We share our findings on how each of these compare to human ratings in Section 4.2.

We calculated s_{Bert} using the *BERTScore* python package[2], s_{BLEU} using the *NLTK* python package[3], s_{Rouge} using the *rouge-score* python package[4], and s_{USE} using the *universal-sentence-encoder* model on Kaggle[5].

Response to response-set consistency We used the similarity scores between pairs of responses to calculate the consistency between a specific response r_i and all the remaining $m-1$ responses. We call this *response to response-set* consistency, and compute it by averaging the response-to-response similarities between the specific response and each of the other responses. Formally,

$$C(\text{prompt}, r_i) = \frac{\sum_{\substack{j=1 \\ j \neq i}}^{m} s(r_j, r_i)}{m-1} \quad (2)$$

[2]https://pypi.org/project/bert-score/0.3.0/

[3]https://www.nltk.org/_modules/nltk/translate/bleu_score.html

[4]https://pypi.org/project/rouge-score/

[5]https://www.kaggle.com/models/google/universal-sentence-encoder

In addition to the four metrics used in response-pair level comparison, we add Semantic Entropy (*SE*) (Kuhn et al., 2023) at this level of comparison, as their definition of consistency is measured at the response to response-set level. To understand how well these five metrics perform at this level, we compute the Spearman correlation coefficient and the mean squared error between each of the metrics and human ratings in Section 4.3.

Additionally, motivated by reducing the overhead of resampling, we compare logits-based scores to human ratings and propose an ensembling method that uses a linear combination of logits in Section 4.3.

To calculate *SE*, we used the *roberta-large-mnli* model to check entailment between pairs of responses.[6] We provide more details on logit-based scores and our ensembling method in section 3.3.

Prompt level consistency Building on response to response-set consistency, we define *prompt consistency* as the average of the m response to response-set consistency values of all responses to the prompt; formally,

$$C(\text{prompt}) = \frac{\sum_{i=1}^{m} C(\text{prompt}, r_i)}{m} \quad (3)$$

We use the Spearman correlation and mean squared error to determine how well each of the existing metrics and our ensemble approximate human ratings of prompt level consistency. We share our findings in Section 4.4.

Estimating consistency with logit-based metrics We estimate the consistency of LLM responses using token-based uncertainty metrics used in previous work (Manakul et al., 2023). Specifically, we use four uncertainty metrics: the probability, minus log probability, entropy, and the DLR loss (introduced in 2.2). For each response, we measure the maximum, sum, minimum, and average of these four token-based metrics, totaling 16 values. The maximum and minimum show the ranges on individual tokens, while the average is the expected value across all tokens. The sum suggests a cumulative estimate over the whole response, as different responses may have different lengths. Unlike existing work that detected hallucination based producing a single response (Manakul et al., 2023), we

[6]https://huggingface.co/FacebookAI/roberta-large-mnli

collect several responses to estimate LLM consistency for a prompt, corresponding to the definition in Equation (3).

Ensembling uncertainty metrics In addition to individual uncertainty-based metrics described in the previous paragraph, we create an ensemble of these in an attemp to better approximate human ratings of LLMs' consistency. We used *Sequential Feature Selection* (Rückstieß et al.) to determine the most important metrics before composing them into an ensemble.

4 Results

Here, we first summarize the user-study data collection (4.1), the results of which were used to establish our baseline for consistency. To compare previous consistency methods to our user-rating-based method, we followed a bottom-up approach. First, we looked at how methods for estimating similarity between sentences (or LLM responses, in our case) compare to user ratings (4.2). Next, we used these similarity scores (including user-based scores) to compute consistency scores for individual responses (4.3). We also compared logit-based uncertainty-estimation methods to our user baseline. Finally, we aggregated the consistency scores within a set of responses to obtain a prompt-level consistency score (4.4). For each step of the hierarchy, we compared automated methods for estimating consistency to our user-rating-based baseline, which we believe is the closest to representing ground truth. We found that while some methods outperform others, none are very close to the user-rating-based baseline. We further found that our ensemble logit-based scores approximates the best-performing sampling method with the benefit of not needing pools of resampled responses.

4.1 User Study Overview and Demographics

We recruited 2,976 participants from Prolific, asking each to rate five unique response pairs from different prompts. About 52% of participants were female and 47% male. Participants' ages ranged from 18 to 88 with various education, ethnicity, and income backgrounds. More details are provided in Table 6.

For the user study, we removed duplicate response-pairs (including pairs with identical responses, see Section A.2), and were left with 756 unique response-pairs, each of which was rated by five or more participants.

Model	*CoQA*		*LMSYS*	
	Prompt	Reponse	Prompt	Reponse
Gemma	14.00	18.96	11.00	37.08
Llama	14.44	18.54	11.59	46.29
Mistral	14.69	49.18	13.85	133.83

Table 2: Average prompt and response lengths for each model and dataset.

LLM	Dataset	*Bert*	*BLEU*	*Rouge*	*USE*
Gemma	*CoQA*	0.82	0.82	**0.83**	0.82
	LMSYS	0.58	0.63	0.59	**0.68**
Llama	*CoQA*	0.84	0.84	0.89	**0.91**
	LMSYS	0.60	0.60	0.64	**0.70**
Mistral	*CoQA*	**0.66**	0.55	0.65	0.53
	LMSYS	**0.57**	0.57	0.52	0.51
All	**Both**	0.71	0.74	0.73	**0.75**

Table 3: At the response-pair level (not consistency), *USE* have the highest Spearman ρ correlation coefficient with human ratings overall. Additionally, existing metrics better correlate with human evaluations for prompts from the *CoQA* dataset than *LMSYS* dataset.

The prompts were on average 13.33 tokens long and the responses were 52.80 tokens long. We break down the lengths for prompts drawn from the two datasets and responses generated by the three models in Table 2.

4.2 Response-Pair Similarity

We first compare user ratings of the similarity between LLM response-pairs to semantic similarity metrics used in prior work. Semantic similarity metrics are a core component of sampling-based consistency metrics. As defined in Section 3.3, we refer to such comparison as *response-pair similarities*.

We evaluated the semantic similarities in a response-pair using previously published methods and computed their correlation with human ratings. Each response-pair was rated by at least five participants. We calculated Krippendorff's Alpha after removing the highest and lowest score (see Section 3.3) and found moderate agreements ($\alpha = 0.72$) across raters (Krippendorff, 2011; Marzi et al., 2024). Comparing each of the four metrics for evaluating the similarity of response-pairs—*Bert*, *BLEU*, *Rouge*, and *USE* (Zhang et al., 2020; Papineni et al., 2001; Lin, 2004; Cer et al., 2018)—to participants' ratings, we found that no single metric correlated best with how participants rated response-pairs' similarities across all datasets and models. Notably, *USE* had the highest correlation coefficient with response-pairs produced by Llama, while *Bert* performed the best for Mistral.

Interestingly, we found the correlation between participants' ratings and the five metrics from previous work to be higher for prompts from the *CoQA* dataset than the *LMSYS* dataset. Participants found response-pairs created from prompts from the *CoQA* dataset to be more similar to each other than response-pairs created from *LMSYS* prompts, by one level in the 6-point scale. To the best of our knowledge, *Bert*, *BLEU*, *Rouge*, and *USE* have not been used to evaluate LLM consistency with the more open-ended *LMSYS* dataset, which could have contributed to their lower correlation with human ratings.

Result 1: Among four existing metrics for evaluating the consistency of LLM responses based on similarity between pairs of sentences, *USE* best approximates human ratings in most cases. Additionally, existing metrics better approximate human ratings on *CoQA* than on *LMSYS*.

4.3 Response to Response-Set Consistency

In Section 4.2, we evaluated methods for measuring the similarity between pairs of responses. Consistency metrics, however, typically involve computing a score for each response that represents the extent to which that response is similar to or different from other responses to the same prompt. Here, we compare how human-rating-based consistency metrics perform in relation to those calculated with response-pair similarity metrics, as well as one additional metric that requires sampling multiple responses, *SE* (Kuhn et al., 2023). As defined in Section 3.3, we call this *response to response-set consistency*.

Motivated by reducing the cost of generating multiple responses to a prompt for consistency estimation, we further explore whether logit-based consistency-estimations methods (which have negligible cost compared to response generation) can match human-based consistency scores. For this investigation, we used 16 different logits based scores—mean, minimum, maximum, and sum for each of DLR, Entropy, Probability, and LogProbability (explained in Section 3.3)—and found none of them individually approximates the consistency of LLMs as well as the *USE* score (see Table 4).

As described in Section 3.3, we used an ensemble of the 16 logit-based scores to attempt to approximate human ratings. To determine which combination of the 16 logit-based scores can create the best ensemble, we used *Sequential Feature Selection* (SFS) [7]. We ran SFS 1,600 times, performing 100 repetitions at each selection size (from using 1 logit-based score up to using all 16 scores). Across the 1,600 runs, we found maximum entropy to be most frequently selected (1,569 times) followed by sum of LogProbability (1,384 times). We provide the full list in Table 5. Within each of the 1,600 runs, we used 10-fold cross validation and evaluated the performance (i.e., Spearman ρ and MSE) of our ensemble relative to human ratings. We found that using all 16 logit-based scores resulted in the highest Spearman correlation coefficient and lowest mean squared error when compared to human ratings (see Figure 5). We found our ensemble method with 16 logit-based scores had a higher correlation with human ratings than *Bert*, *BLEU*, *Rouge*, and *SE*. Our ensemble, when compared to the human ratings, performed as well as *USE*, with a 0.002 better MSE (Figure 2). As in Section 4.2, we found that at the response to response-set level, existing metrics correlate with human ratings better on the *CoQA* dataset than on *LMSYS* (see Figure 6).

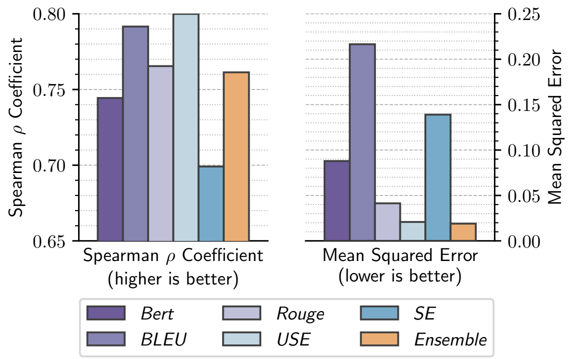


Figure 2: At response to response-set level, our ensemble of 16 logit-based scores is as close of an approximation of human ratings as *USE*.

In addition to matching the performance of the best existing metrics, our ensemble method benefits from reduced computational cost. With the same GPU, it took *Bert* 2,649 seconds, *USE* 23 seconds, *Rouge* 14 seconds, *SE* 47 seconds, and our ensembling method 0.0044 seconds to evaluate the consistency of 1,000 responses to 100 prompts. Training the ensemble with 10-fold cross-validation on 90%

[7] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html

Stats	DLR Mean	DLR Min.	Entropy Mean	Entropy Max.	Prob Mean	Prob Min.	LogProb Mean	LogProb Max.	*USE*
Spearman ρ	0.37	0.22	0.71	0.75	0.69	0.70	0.70	0.70	**0.80**
MSE	0.10	0.05	0.41	3.88	0.04	0.14	0.33	0.67	**0.02**

Table 4: At response to response-set level, logit-based scores (showing 2 best ones per type) did not perform as well as *USE* in approximating human ratings.

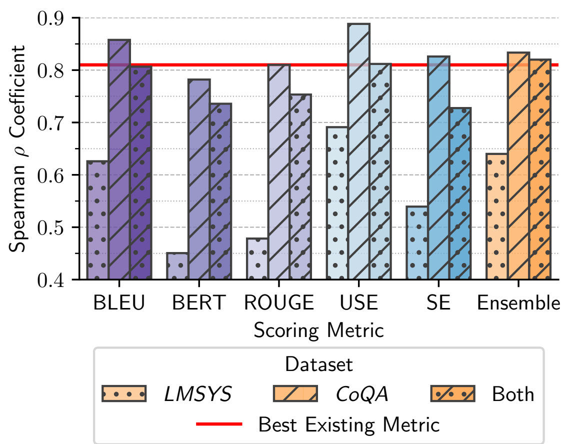


Figure 3: At per-prompt level, among existing metrics, our logit-based ensemble method and *USE* have the highest Spearman correlation coefficient with human evaluation of model-prompt consistency.

of the responses required an average of 0.037 seconds across 100 runs using all 16 features.

Result 2: Using an ensemble of 16 logit-based scores can produce a consistency estimate that approximates human ratings as accurately as the best existing metrics. In addition to matching *USE*, our method greatly reduces computational cost by not requiring the generation of multiple responses.

4.4 Prompt Level Consistency

As evidenced by the difference in how existing metrics perform on the two different prompt datasets (Table 3), prompt choice is a major contributor to consistency.

For example, responses to the factual question "Where was the first modern Olympic Game?" were objectively more consistent than responses to the open-ended question "How can LLMs help users in their daily life?" As such, we next investigate how to best measure consistency for a given prompt. Specifically, we aggregate response-level consistency from Section 4.3 to obtain a single consistency score per prompt.

Existing metrics Using the definition from Section 3.3, we calculate the consistency score of each prompt using participants' ratings and metrics from previous work. We found that *USE*-based consistency scores best correlate with human-based consistency scores across all prompts (Figure 3). The consistency scores on the *CoQA* dataset correlates better to human ratings than the consistency scores on *LMSYS*. This is similar to the bias we observed with pair-wise comparisons in Section 4.2.

Logit-based ensemble method Our logit-based ensemble method, described in Sections 3.3 and 4.3, performs as well as *USE*, when aggregated to the prompt level, in approximating human ratings of per-prompt consistency. While it did not outperform *USE*, our method benefits from not needing to generate a set of responses and compare responses within the set to determine LLMs' responses consistency at the prompt level.

We further investigated the number of responses needed to train an ensembled model with logit-based scores to predict the consistency of LLM responses based on generating a single response. However, due to the relatively small number of prompts and responses in our sample, we were not able to conclude how many responses are needed for such model to accurately predict LLMs' response consistency at the prompt level.

Result 3: Our ensemble method using logit-based scores performs as well as existing metrics in approximating human ratings of LLM prompt-level consistency. This approach enables estimation of LLM consistency without the need to generate multiple responses to a prompt.

5 Conclusion and Discussion

Using the human evaluation data, obtained via our user study, and subsequent experiments, we show that existing (automated) formulations of LLMs' consistency are meaningfully different from our human-judgment baseline. The best existing response to response-set consistency method achieves Spearman's $\rho = 0.80$, indicating correlation but not representation. Further, prior work has suggested that correlations considered "strong" in less precise contexts (Schober et al., 2018) are not as meaningful for measuring success on NLP tasks (Deutsch, 2022; Bavaresco et al., 2024; Shen et al., 2023). This result holds regardless of whether we examine consistency at the per-response level (Section 4.3) or at the prompt level (Section 4.4).

Though sampling-based methods come closest to the human baseline, we show that an ensemble of logit-based methods can approximate this performance ($\rho_{ensemble} = 0.82$ vs $\rho_{USE} = 0.81$), creating an opportunity to avoid the sampling overhead.

We also find (Section 4.2) that the difference in human judgement-based uncertainty vs. existing metrics for LLM consistency approximation is greater for real-world prompt datasets (*LMSYS*) than for artificial ones (*CoQA*). Determining why this is the case is out of scope for our work, but we speculate that the research community's focus on artificial testing datasets might be a contributing factor.

Future directions The discrepancy between the human baseline and existing automated methods of measuring LLM consistency raises important questions: How do imperfect consistency estimation methods affect downstream tasks? How would end-users be affected if shown consistency metrics (Kapoor et al., 2024)? How are models affected when consistency measurements are part of the training cycle (Liu et al., 2024)? The answers are unclear and could be investigated by future work.

Finally, we urge researchers to consider the human evaluation baseline in future research on how to measure LLM consistency, as well as in research that utilizes these metrics for downstream tasks. We further recommend they (also) use real-world prompts for evaluation.

Limitations

Since our investigation required logit-based scores, which are often not accessible with black-box models, we used open-weights models. Such usage limits the generalizability of our findings to black-box models. Additionally, the prompts and responses in our sample are relatively short (see Table 2) despite using realistic prompts from users (Zheng et al., 2023); therefore, our findings may not generalize to prompts and responses of all lengths. Lastly, user studies are expensive to conduct. While we recruited almost 3,000 users, we were only able to evaluate 100 prompts and 10 responses per prompt. As we alluded to in Section 5, more data from users are needed to establish a robust baseline for measuring LLM consistency.

Ethical Considerations

Our study was approved by our institution's ethics review board. For the user study, we first provided an informed consent form to participants explaining the purpose of our survey, expected length, risks and benefits, as well as compensation. Participants who gave consent to participate proceeded to the survey. As explained in Section 3.1, responses shown to participants within the survey were filtered by the authors to remove potentially harmful (e.g., harassment, sexual, violence) content.

Distribution of Data and Artifact

Code developed for this project, anonymized data collected from participants, and analysis results are accessible at `https://doi.org/10.17605/OSF.IO/T9BF5`.

Acknowledgments

This work was supported in part by the National Institute of Standards and Technology (NIST) (ror.org/05xpvk416) and the Carnegie Mellon University (ror.org/05x2bcf33) AI Measurement Science and Engineering Center (AIMSEC); and by the PNC Center for Financial Services Innovation at Carnegie Mellon University. We thank Clement Fung for technical help with the ensemble predictor.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *CoRR*, abs/2406.18403.

Su Lin Blodgett, Jackie Chi Kit Cheung, Vera Liao, and Ziang Xiao. 2024. Human-centered evaluation of language technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 39–43. Association for Computational Linguistics.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Jordan Boyd-Graber, Samuel Carton, Shi Feng, Q. Vera Liao, Tania Lombrozo, Alison Smith-Renner, and Chenhao Tan. 2022. Human-centered evaluation of explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 26–32. Association for Computational Linguistics.

Nicola Cecere, Andrea Bacciu, Ignacio Fernández Tobías, and Amin Mantrach. 2025. Monte carlo temperature: a robust sampling strategy for LLM's uncertainty quantification methods. *Preprint*, arxiv:2502.18389 [cs].

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *Preprint*, arxiv:1803.11175 [cs].

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pages 2206–2216. JMLR.org.

Paul G. Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66:4–19.

Daniel Deutsch. 2022. *Methods for Text Summarization Evaluation*. Ph.D. thesis, University of Pennsylvania.

Weihua Du, Yiming Yang, and Sean Welleck. 2025. Optimizing temperature for language models with multi-sample inference. *Preprint*, arxiv:2502.05234 [cs].

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. *Preprint*, arxiv:2307.01379.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2024. HumanRankEval: Automatic evaluation of LMs as conversational assistants. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8237–8249. Association for Computational Linguistics.

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, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arxiv:2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. *Preprint*, arxiv:2406.08391.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. (43).

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arxiv:2302.09664.

Aounon Kumar and Himabindu Lakkaraju. 2024. Manipulating large language models to increase product visibility. *arXiv preprint arXiv:2404.07981*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Weiran Lin, Anna Gerchanovsky, Omer Akgul, Lujo Bauer, Matt Fredrikson, and Zifan Wang. 2025. LLM whisperer: An inconspicuous attack to bias LLM responses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arxiv:2305.19187.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek Wong, Lidia Chao, and Min Zhang. 2024. Can llms learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21635–21645.

Zilin Ma, Yiyang Mei, Krzysztof Z Gajos, and Ian Arawjo. 2024. Schrödinger's update: User perceptions of uncertainties in proprietary large language model updates. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arxiv:2303.08896 [cs].

Giacomo Marzi, Marco Balzano, and Davide Marchiori. 2024. K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. *MethodsX*, 12:102545.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schoelkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. 2023. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11330–11343, Toronto, Canada. Association for Computational Linguistics.

Sara Merken. 2025. Ai 'hallucinations' in court papers spell trouble for lawyers. Accessed: 2025-05-16.

Carolyn V. Metnick, Lynsey Mitchel, and Michael D. Sutton. 2024. California limits health plan use of ai in utilization management. Accessed: 2025-05-16.

Aaron Moss and Leib Litman. 2018. After the bot scare: Understanding what's been happening with data collection on MTurk and how to stop it.

Dong Nguyen, Dolf Trieschnigg, and Mariët Theune. 2014. Using crowdsourcing to investigate perception of narrative similarity. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 321–330. ACM.

Jekaterina Novikova, Carol Anderson, Borhane Blili-Hamelin, and Subhabrata Majumdar. 2025. Consistency in language models: Current landscape, challenges, and future directions. *arXiv preprint arXiv:2505.00268*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311. Association for Computational Linguistics.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Preprint*, arxiv:2405.13845.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. 2025. Semantic consistency for assuring reliability of large language models. *Preprint*, arxiv:2308.09138 [cs].

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Preprint*, arxiv:1808.07042 [cs].

Matthew Renze and Erhan Guven. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356.

Salvatore Ruggieri and Andrea Pugnana. 2025. Things machine learning models know that they don't know. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28684–28693.

Thomas Rückstieß, Christian Osendorfer, and Patrick van der Smagt. Sequential feature selection for classification. In *AI 2011: Advances in Artificial Intelligence*, pages 132–141. Springer.

Dylan Sam, Marc Finzi, and J Zico Kolter. 2025. Predicting the performance of black-box llms through self-queries. *arXiv preprint arXiv:2501.01558*.

Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. *arXiv preprint arXiv:2305.13091*.

Jenny Tang, Eleanor Birrell, and Ada Lerner. 2022. Replication: How well do my results generalize now? the external validity of online privacy and security surveys. In *Proceedings of the 18th USENIX Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 367–385.

Angus Tiffin and Graham Fraser. 2025. Law firm restricts ai after 'significant' staff use. Accessed: 2025-05-16.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arxiv:2203.11171 [cs].

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. *Preprint*, arxiv:1904.09675 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *Proceedings of the 12th International Conference on Learning Representations*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arxiv:2307.15043 [cs].

A Appendix

A.1 Instructions for Survey Participants

Risks The primary risk is a breach of confidentiality since we use a third-party (Qualtrics) to design our survey and collect survey responses. Additionally, we utilize third-party vendors such as Prolific to recruit participants, and Google Drive to store and process survey responses. This risk is similar to what you encounter anytime you provide identifiable and private information online. The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life or other online activities. Participants might encounter boredom or fatigue.

Instructions A screenshot of the instructions given to participants is provided (Figure 4).

Logit-Based Score Name	Frequency
Max. Entropy	1569
Sum. LogProb.	1384
Mean Prob.	1309
Sum. DLR	1220
Min. DLR	1123
Mean Entropy	1044
Mean DLR	984
Sum. Prob.	856
Sum. Entropy	691
Min. Entropy	678
Max. DLR	662
Min. Prob.	555
Max. LogProb.	533
Max. Prob.	414
Mean LogProb.	344
Min. LogProb.	234

Table 5: Across 1,600 SFS runs (16 features $\times$ 100 repetitions), the most frequently selected features were Mac. Entropy, Sum. LogProb., and Mean Prob.

A.2 Experiemnt asking users to rate identical responses

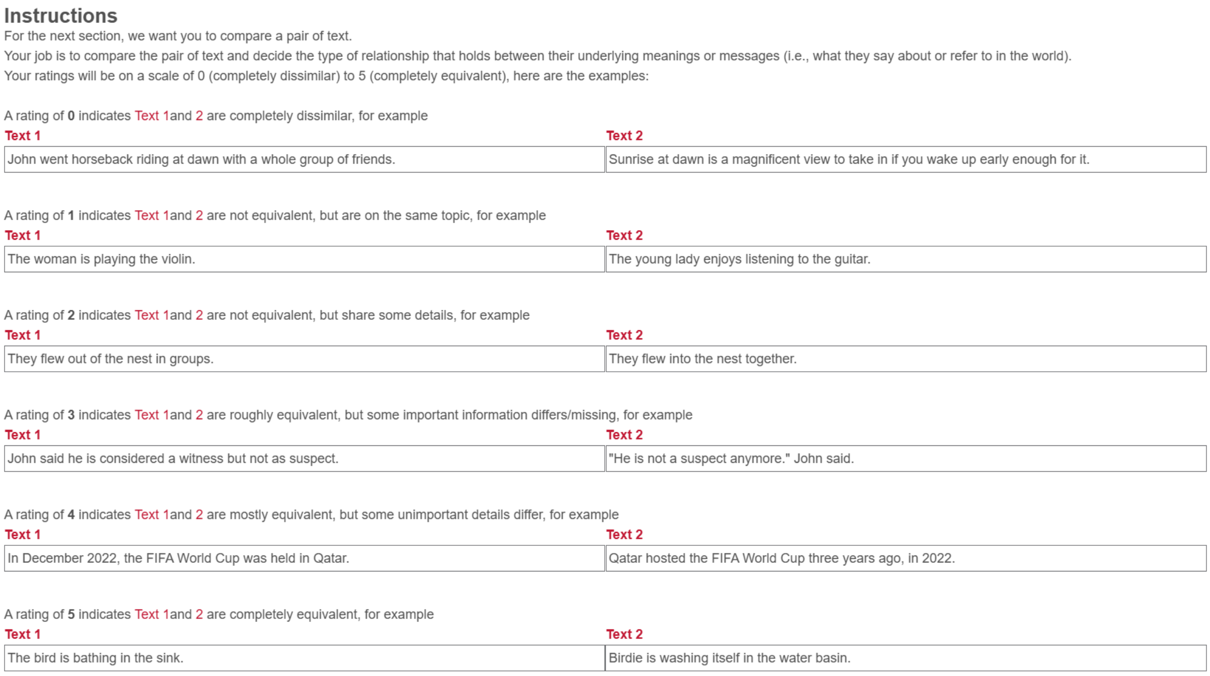
For each prompt, the 10 sampled responses described in Section 3.1 are not all unique. We ran a 20 participant experiment on Profific with 5 pairs of unique responses for 5 different prompts. We found 17 participants gave a *5 (completely equivalent)* for the identical responses. Further, three participants each gave a *4*, *3*, and *0* to three different pairs of identical responses. This result show participants are highly likely to rate identical pairs of responses as a *5 (completely equivalent)* so we excluded all identical pairs of responses from further study.

A.3 Supplemental Tables and Figures

Table 6 shows participants' demographics information.

Figure 5 shows our procedure of selecting all 16 logit-based scores in our ensemble method.

Figure 6 shows the response to response-set level comparison between existing metrics and our ensemble method across different datasets.



Instructions

For the next section, we want you to compare a pair of text.
Your job is to compare the pair of text and decide the type of relationship that holds between their underlying meanings or messages (i.e., what they say about or refer to in the world).
Your ratings will be on a scale of 0 (completely dissimilar) to 5 (completely equivalent), here are the examples:

A rating of **0** indicates Text 1and 2 are completely dissimilar, for example

Text 1	Text 2
John went horseback riding at dawn with a whole group of friends.	Sunrise at dawn is a magnificent view to take in if you wake up early enough for it.

A rating of **1** indicates Text 1and 2 are not equivalent, but are on the same topic, for example

Text 1	Text 2
The woman is playing the violin.	The young lady enjoys listening to the guitar.

A rating of **2** indicates Text 1and 2 are not equivalent, but share some details, for example

Text 1	Text 2
They flew out of the nest in groups.	They flew into the nest together.

A rating of **3** indicates Text 1and 2 are roughly equivalent, but some important information differs/missing, for example

Text 1	Text 2
John said he is considered a witness but not as suspect.	"He is not a suspect anymore." John said.

A rating of **4** indicates Text 1and 2 are mostly equivalent, but some unimportant details differ, for example

Text 1	Text 2
In December 2022, the FIFA World Cup was held in Qatar.	Qatar hosted the FIFA World Cup three years ago, in 2022.

A rating of **5** indicates Text 1and 2 are completely equivalent, for example

Text 1	Text 2
The bird is bathing in the sink.	Birdie is washing itself in the water basin.

Figure 4: Instructions provided to participants for comparing pairs of sentences.

Age	**Num.**	**%**
18-24	290	9.74
25-34	917	30.81
35-44	708	23.79
45-54	556	18.68
55-64	324	10.89
65+	159	5.34
Prefer not to say	22	0.74
Gender	**Num.**	**%**
Woman	1538	51.68
Man	1398	46.98
Other	27	0.91
Prefer not to say	13	0.44
Education	**Num.**	**%**
Associate's degree	221	7.43
Bachelor's degree	1225	41.16
Doctorate degree	137	4.60
High school graduate	298	10.01
Master's degree	532	17.88
No high school degree	11	0.37
Professional degree	59	1.98
Some college credit, no degree	420	14.11
Trade, technical, vocational training	67	2.25
Other	1	0.03
Prefer not to say	5	0.17
Income	**Num.**	**%**
Under \$25K	311	10.45
\$25K to \$50K	582	19.56
\$50K to \$75K	636	21.37
\$75K to \$100K	470	15.79
\$100K or more	915	30.75
Prefer not to say	62	2.08

Table 6: Demographics data of 2,976 participants.

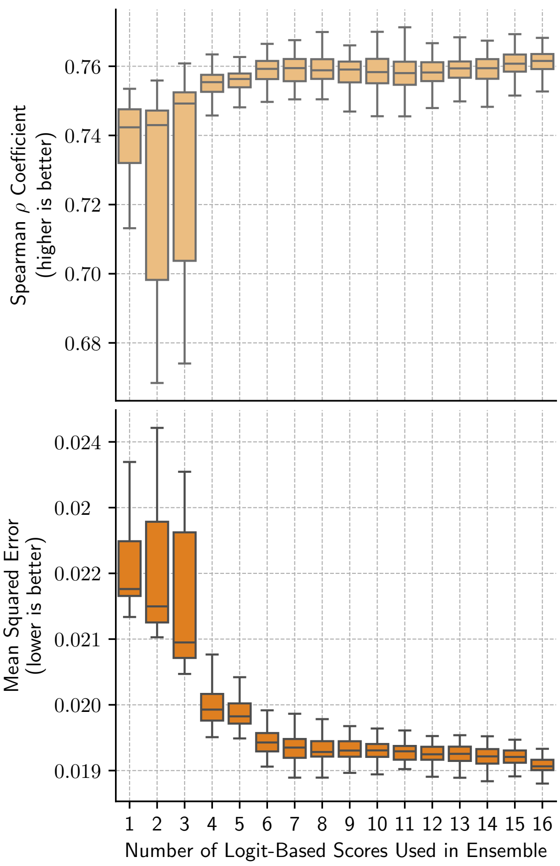


Figure 5: Running 100 10-fold cross validation shows using an ensemble of all 16 logits-based scores yield the lowest Mean Squared Error and highest Spearman ρ coefficient when compared to human ratings.

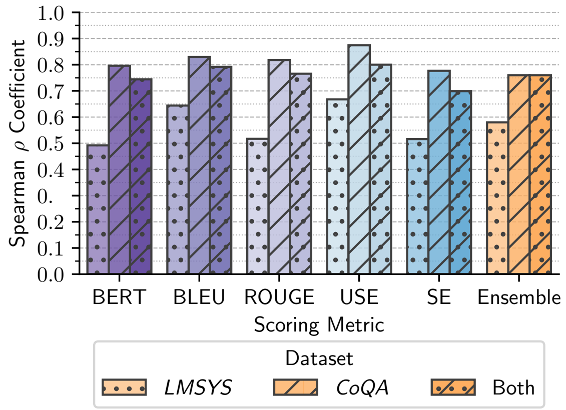


Figure 6: At response to response-set level, existing metrics and our ensemble method correlates better with human ratings on *CoQA* than *LMSYS*.

A.4 Example Prompts

Table 7 shows prompts that had large discrepancies between human rated consistency and consistency calculated with existing metric (i.e., *USE*).

Prompt	Dataset	LLM Used	User Rated Consistency	*USE* Consistency
What has Williams become with this win?	CoQA	Gemma	0.69	0.42
Please resume the book "a libertarian walks into a bear"	LMSYS	Gemma	0.50	0.28
Write to me something about Introversion and how I feel peace with small circle of people.	LMSYS	Llama	0.77	0.51
Write a recipe for a unique dessert.	LMSYS	Mistral	0.30	0.63

Table 7: Sample prompts with large differences between humans' ratings and the best surrogate metric (i.e., *USE*).