Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation

Zhen Lin[1], **Shubhendu Trivedi**, **Jimeng Sun**[1,2]
[1]University of Illinois at Urbana-Champaign
[2]Carle's Illinois College of Medicine, University of Illinois at Urbana-Champaign,
Correspondence: zhenlin4@illinois.edu

Abstract

The advent of large language models (LLMs) has dramatically advanced the state-of-the-art in numerous natural language generation tasks. For LLMs to be applied reliably, it is essential to have an accurate measure of their confidence. Currently, the most commonly used confidence score function is the likelihood of the generated sequence, which, however, conflates semantic and syntactic components. For instance, in question-answering (QA) tasks, an awkward phrasing of the correct answer might result in a lower probability prediction. Additionally, different tokens should be weighted differently depending on the context. In this work, we propose enhancing the predicted sequence probability by assigning different weights to various tokens using attention values elicited from the base LLM. By employing a validation set, we can identify the relevant attention heads, thereby significantly improving the reliability of the vanilla sequence probability confidence measure. We refer to this new score as the Contextualized Sequence Likelihood (CSL). CSL is easy to implement, fast to compute, and offers considerable potential for further improvement with task-specific prompts. Across several QA datasets and a diverse array of LLMs, CSL has demonstrated significantly higher reliability than state-of-the-art baselines in predicting generation quality, as measured by the AUROC or AUARC. The code to replicate our experiments is available at `https://github.com/zlin7/ContextSL`.

1 Introduction

The development of large language models (LLMs) has afforded tremendous advancements in natural language generation (NLG). Recently, LLMs have been widely applied across various natural language domains (Zhang et al., 2023; Wang et al., 2023a; Alves et al., 2023; Zhang et al., 2024), even extending to tasks and domains traditionally dominated by other machine learning algorithms, such as graph data (Fatemi et al., 2024), tabular data (Borisov et al., 2023; Hegselmann et al., 2023), time series (Gruver et al., 2023; Rasul et al., 2024), predictive chemistry (Jablonka et al., 2024; Shi et al., 2023), computer vision (Wang et al., 2023b), amongst others. As LLMs continue to demonstrate outstanding performance, their reliability is increasingly scrutinized. Uncertainty quantification, an area of research that can provide some guidance on reliability, has recently gained much attention. Despite its long history in other machine learning tasks (Gawlikowski et al., 2023), our understanding of uncertainty quantification in NLG remains relatively limited.

During pre-training, (autoregressive) LLMs are optimized to predict high logits for the target token to minimize (variants of) the negative log likelihood. Consequently, one of the most natural and widely used confidence scores in selective NLG, conformal NLG, or uncertainty quantification is the (sometimes normalized) likelihood of the sequence, equivalent to the sum/mean of token logits. At first glance, sequence likelihood appears to be the most faithful reflection of a model's confidence, as it represents the (log of) the predicted probability of the output sequence, or $\log \hat{p}(\mathbf{s}|x)$. However, this measure often lacks proper contextualization and disregards the specific nature of the task at hand. $\hat{p}(\mathbf{s}|x)$ conflates syntactic and semantic likelihoods, though we typically prioritize the semantic aspect, to a varying degree depending on the task. For example, depending on whether the question is "Which country won the World Cup in 2022" or "When did Messi win the World Cup," the answer "Messi emerged victorious in the 2022 World Cup" could be considered correct or incorrect. Further, the somewhat unusual expression here could adversely impact $\hat{p}(\mathbf{s}|x)$.

Despite its limitations, relatively limited attention has been paid to improving the vanilla sequence likelihood. A recent study by Duan et al.

(2023) proposed assessing token relevance using an external natural language inference (NLI) model. The relevance score of a token is negatively proportional to the similarity between the original sequence and the sequence with this token removed, which is then applied to weight the token logits. However, since modern LMs rely on sub-word tokenizers, removing one sub-word token at a time results in non-words, is computationally expensive for longer texts, and remains context-unaware.

In this paper, we investigate the potential of utilizing the LLM's own attention mechanism to develop a weighted sequence likelihood as an enhanced confidence score. Specifically, the LLM is prompted to concentrate on the relevant tokens in its own generation. The most appropriate attention heads are then employed to re-weight the original token logits. The main contributions of this paper are summarized as thus:

- We introduce a straightforward yet effective method to reweight token logits, resulting in Contextualized Sequence Likelihood (CSL), a more reliable confidence measure.
- We improve current automatic evaluation methods for confidence measures on Question-Answering (QA) datasets and manually verify their effectiveness.
- In popular free-form QA datasets, and on a variety of LLMs, we verify that CSL significantly outperforms baselines. Case studies suggest the attention weights are meaningful.

2 Related Works

With the rapid proliferation of LLMs and their swift adoption across various domains, uncertainty quantification (UQ) for natural language generation (NLG) is a fast-growing area of research (see Baan et al. (2023) and references therein). Adopting the language used by Lin et al. (2023), uncertainty measures the predictive distribution, while confidence (our focus) further depends on the specific generation. A common approach involves reducing NLG to a de facto classification problem and utilizing or enhancing classical UQ methods for classification (Desai and Durrett, 2020; Jiang et al., 2021; Kamath et al., 2020; Wang et al., 2022; Xiong et al., 2023). Recognizing the unique challenges posed by the inherently high (potentially infinite) dimensionality of NLG, recent research increasingly considers UQ for NLG from a sequence perspective (Hou et al., 2023; Kuhn et al., 2023; Malinin and Gales, 2021; Lin et al., 2023).

Relatively less attention has been devoted to confidence measures. Sequence likelihood, or the log-probability of the generated sequence, remains one of the most popular proxies for assessing the quality of individual answers (Quach et al., 2023; Kuhn et al., 2023; Cole et al., 2023). Another natural approach, given the versatility and strong performance of LLMs, involves using prompts to elicit the LLM's own confidence level (Kadavath et al., 2022; Lin et al., 2022a; Mielke et al., 2020; Chen and Mueller, 2023; He et al., 2023; Li et al., 2024; Wightman et al., 2023). For free-form generation datasets, this is typically done by arranging answers as options and extracting the model's logits for each option. While our method also employs prompts, they primarily guide the model to focus on tokens relevant to the context of the question. Another approach involves sampling additional generations and comparing their similarities (Lin et al., 2023; Cole et al., 2023), which demonstrates good discriminative capability but can be rather expensive. Although this approach has the advantage that it can work for black-box LLMs. Ensemble methods have also been proposed (Chen and Mueller, 2023).

Recently, Duan et al. (2023) proposed an improvement to sequence likelihood by weighting the tokens using their importance, which is computed by removing one token at a time from the original sequence and computing the NLI dissimilarity between the new and original sequences. The removal of sub-word tokens, however, introduces drastically grammatically incorrect sentences, which could confuse the NLI model. In addition, such a method could incur high computation overhead by making n NLI comparisons, where n is the length of the generation. In contrast, our method incurs no artificial grammatical errors and miniscule overhead.

An important downstream application of confidence measures is **abstention in LLMs**. Classification with rejection (Corbière et al., 2019; Fumera et al., 2000; Geifman and El-Yaniv, 2017; Jiang et al., 2018; Lin et al., 2022b) could be considered the direct antecedent of selective NLG—both aiming to determine when to trust a model. Naturally, approaches similar to those in the classification with rejection literature have been applied to NLP applications (Varshney et al., 2022a,b). More recently, Cole et al. (2023); Yadkori et al. (2024); Lin et al. (2023); Quach et al. (2023) started investigating NLG with abstention from UQ or risk control perspectives.

Calibration is another crucial and relevant topic that has been extensively studied (Mielke et al., 2022; Si et al., 2022; Xiong et al., 2023; Zhu et al., 2023). Although calibration is not directly related to our primary focus—our main interest lies in ranking the confidence of different measures—we include results on the calibrated performance in the Appendix. Conformal prediction (Vovk et al., 2005) could also be considered a form of calibration at the distribution level, and has been extended to NLG to bound variants of error rates (Quach et al., 2023; Yadkori et al., 2024).

3 Contextualized Sequence Likelihood

3.1 Background: Sequence Likelihood

In this section we describe our approach: contextualized sequence likelihood (CSL). We first fix notation and introduce relevant definitions. We will focus on auto-regressive LMs, the current dominant paradigm. Denoting the model as $\mathcal{M}$, for any given input prompt denoted as x, the response $\mathbf{s}$ is a sequence of tokens $[s_1, \ldots, s_n]$ sampled from the predictive distribution $P(S; x, \mathcal{M})$. We will denote $\mathbf{s}_{<i}$ as the truncated sequence $[s_1, \ldots, s_{i-1}]$. Given the auto-regressive assumption, the (log-softmax'd) logit for the i-th token represent $\mathcal{M}$'s prediction of the log probability of token s_i at this location. Consequently, the sum of all the logits for the sequence represents the log of the model's predicted probability of the output sequence:

$$\mathrm{C}_{\mathsf{SL}} = \sum_{i=1}^{n} l_i = \log \prod_{i=1}^{n} \hat{p}(s_i|\mathbf{s}_{<i}, x). \quad (1)$$

Eq. (1) is commonly used as a confidence score (Quach et al., 2023; Cole et al., 2023), and often normalized by the length n, as otherwise longer sequences tend to receive lower confidence (Kuhn et al., 2023; Malinin and Gales, 2021):

$$\mathrm{C}_{\mathsf{SL}(norm)} = \frac{1}{n} \sum_{i=1}^{n} l_i. \quad (2)$$

As pointed out by Cole et al. (2023), in practice C_{LL} is far from the actual log-probability of the sequence $\mathbf{s}$, $\log \hat{p}(\mathbf{s}|x)$. Techniques like nucleus sampling or top-k sampling will reduce the sum of the predicted probability of all tokens below 1. However, if we view this new distribution that $\mathbf{s}$ is effectively sampled from as $P'(S|x, \mathcal{M})$, then we at least have:

$$\forall \mathbf{s}, P'(\mathbf{s}|x) \propto \prod_{i=1}^{n} \hat{p}(s_i|\mathbf{s}_{<i}, x). \quad (3)$$

Thus, C_{SL} still faithfully preserves the ranking of LM's predicted probability of all possible generated sequences, which is all we care about for a good (pre-calibrated) confidence measure. However, as we shall see next, this does not imply sequence likelihood is a good confidence measure.

3.2 Contextualized Likelihood via Attention

While intuitively natural as confidence measures, Eqs. (1) and (2) sweep a crucial consideration under the proverbial rug: While sequence-likelihood reflects the model's predicted probability of the sequence $\mathbf{s}$, what does this probability actually *mean*? Unfortunately, there is an inherent ambiguity here. For instance, let's say we are classifying an image x from ImageNet (Deng et al., 2009), and take the first-choice confidence score $\hat{p}(y|x)$ as predicted by a ResNet (He et al., 2016). However, $\hat{p}(y = \text{cat}|x)$ could mean the (model's predicted) probability of "the input image is that of a cat", "there is a cat in the input image", or probably more precisely speaking, "this image should be labeled as a cat by ImageNet standards"[1]. Similarly, strictly speaking, $\hat{p}(\mathbf{s}|x)$ only reflects the LM's prediction of the probability of "$\mathbf{s}$ follows x, according to the training data".

To illustrate further, consider the question-answer pair where x is "Q: What did Neil Armstrong do on July 20, 1969?". A response $\mathbf{s}$ which goes "A: On July 20, 1969, Armstrong and Buzz Aldrin landed on the Moon for the first time in human history." appears appropriate and highly probable. However, the confidence of "whether $\mathbf{s}$ correctly answers x" is minimally influenced by the redundant mention of the date or Armstrong's fellow astronaut, Buzz Aldrin. Generally speaking, the confidence that we are concerned with in a response $\mathbf{s}$ depends on the context, with some tokens being significantly more relevant than others.

How can we systematically identify the most relevant tokens? To address this, we propose using a prompt (Fig. 1) to elicit the LM's attention on its own response $\mathbf{s}$. Essentially, we prompt the LM to assess whether its generated response correctly answers the question. Unlike previous approaches that rely on similar prompts, we disregard the actual judgment and instead extract the attention values of the LM on $\mathbf{s}$ during this process to reweight the token logits. Assuming the $\mathbf{s}_i$ becomes the i'-th token in the attention-eliciting prompt, the new

[1] See a similar discussion in Beyer et al. (2020).

confidence score could be written as:

$$\mathrm{C}_{CSL} = \sum_{i=1}^{n} w_i l_i \text{ where } w_i = \frac{a_i}{\sum_{i'=1}^{n} a_{i'}} \quad (4)$$

where $a_{i'}$ is the attention of the last token of the attention-eliciting prompt on the $\mathbf{s}_i$.

```
Read the following question with optional context and decide if the answer correctly answer the question. Focus on the answer, and reply Y or N.
...
Context: Harry is a good witcher.
Question: How old is Harry?
Answer: Harry practices witchcraft.
 Decision: N. (The answer does not mention Harry's age.)
...
[$optional_context]
Question: [$question]
Answer: [$response]
 Decision:
```

Figure 1: The attention-eliciting prompt used in this paper (full version deferred to the Appendix due to space constraints). $optional_context, $question and $response are replaced with the corresponding values of a sample. In our experiments, $optional_context refers to the story and conversation history that accompanies each question in CoQA (Reddy et al., 2019).

Fig. 2 illustrates how the attention-eliciting prompt (Fig. 1) induces varying emphases on the same response for different questions. For instance, when the question focuses on "when", the section of the response detailing the event date receives greater attention weight across all heads. This indicates that the weighting scheme introduced in Eq. (4) effectively overweights relevant tokens and underweights irrelevant ones, resulting in a more contextualized version of sequence likelihood.

Besides using a prompt, a more straightforward source of attention weights is the original generating process. Specifically, when the LM completes generating $\mathbf{s}$, we retrieve the attention for the next token. We refer to this variant as `CSL-Next`, in contrast to `CSL` with the prompt. Our hypothesis is that the LM induces some internal attention on the critical words of the generation even without an explicit prompt asking for it, and we will explain how to identify such attention in Section 3.3.

3.3 The Choice of Heads

While the prompt can enhance overall attention on relevant tokens, averaging attention across heads—even with the prompt—is unwise. Many attention heads likely focus, for example, on ensuring grammatical correctness, with only a fraction dedicated to the response. Selecting only the useful heads from the multitude of heads (e.g., out of 1,600 in LLaMA2-13B) remains a challenging task.

We present a systematic approach to identify the appropriate heads. Let w^h denote the attention weights from head h, and C^h_{CSL} the associated confidence score computed via Eq. (4). For each head h and a subset of the samples, we use the associated confidence C^h_{CSL} to predict the accuracy of responses, and compute an AUROC (similar to Kuhn et al., 2023), denoted as AUROC_h. Notably, we found that the "functionality" of the heads appears relatively stable: That is, if we compare the AUROC_h on two subsets of the population, the ranking of these heads is highly consistent across the two subsets, as shown in Fig. 3. As a result, we propose to pick the top $k = 10$ heads on the validation dataset and average the attention weights of only these heads. The reason why we pick more than one head is because picking only the best head is likely affected by noise due to the size of the validation set. In Section 4, we show that leveraging the attention from about $k = 10$ top heads performs better than either the average attention of all heads, or the top head.

4 Experiments

4.1 Datasets

We use the following standard benchmark datasets, largely following the practices in Kuhn et al. (2023); Lin et al. (2023):

- CoQA (`coqa`) (Reddy et al., 2019), an open-book conversational question answering dataset. We use the development split of `coqa` with 7,983 questions.
- TriviaQA (`trivia`) (Joshi et al., 2017), a closed-book QA dataset. We use the validation split of the `rc.nocontext` subset of `trivia` with 9,960 (de-duplicated) questions.
- Natural Questions (`nq`) (Kwiatkowski et al., 2019), a closed-book QA dataset. We use the validation split of `nq` with 3,610 questions.

Following (Lin et al., 2023), for each experiment we use a random subset of 1,000 questions as the validation set, and the remaining as the test set. We report the mean and standard deviation of all evaluation metrics (see Section 4.4) on the test set, calculated from 10 random data splittings.

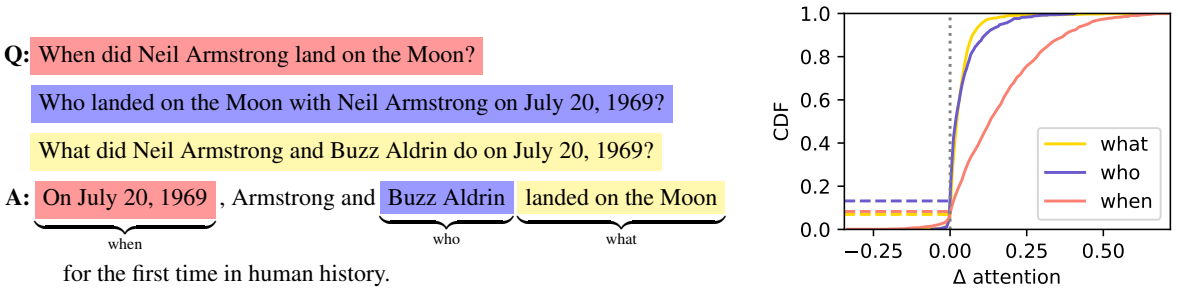

Figure 2: Depending on the question, the attention-eliciting prompt introduced in Fig. 1 induces attention focusing on different parts of the same response ("when", "who" and "what"). In the plot on the right, we show the CDF of Δ_{attn}, the change of attention weight on the corresponding concept when asked the relevant question, on all 1,024 heads of Mistral-7B. For example, for "when", we compute Δ_{attn} as the attention weight on "On July 20, 1969" when asked the "when" question minus the average of the cases where the other two questions were asked. In all cases, the attention significantly increases on the relevant tokens (p-value from one-sided t-test is at most 9e-90).

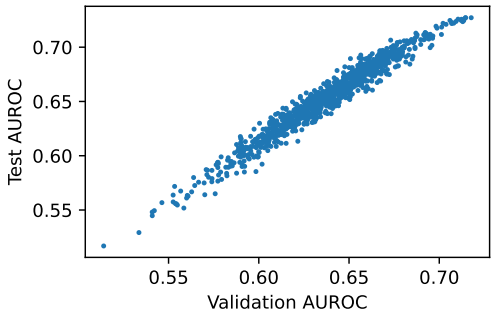

Figure 3: Scatter plot of test vs validation AUROC for confidence measures computed via Eq. (4) with different heads' attention weights, on Natural Questions (nq) with Mistral-7B model. The ranking is highly consistent—the best heads on the validation set continue to perform well on the test set. In this case, the Spearman correlation (Spearman, 1961) is $> 97\%$. We can thus pick only a small subset of the 1024 heads (or more for other LMs) to construct the final confidence measure C_{CSL}.

4.2 Baselines

We compare CSL (and CSL-Next) with several recent confidence measures:

- Deg, a confidence score based on the degree of similarity graph from Lin et al. (2023). Note that this method requires sampling multiple responses, and we set the number of additional generations to 5. We use the "Entailment" version as suggested by Lin et al. (2023).
- P(true) (Kadavath et al., 2022), which elicits the confidence by asking the LM itself whether its response is correct. We use the prompt from Kadavath et al. (2022); Lin et al. (2023).
- Sequence Likelihood (SL): This is the sequence likelihood measure discussed in Eq. (1), and widely in literature as a confidence score (Lin et al., 2023; Quach et al., 2023; Huang et al., 2024) or as a building block for predictive entropy (Kuhn et al., 2023; Kadavath et al., 2022; Malinin and Gales, 2021). We include the length-normalized version in Eq. (2), SL(norm), as well.
- TokenSAR (Duan et al., 2023): It proposes to estimate the relevance of each token as w_i in Eq. (4), with $w_i \propto 1 - sim(\mathbf{s}, \mathbf{s}_{-i})$ where sim is similarity measured by an NLI model and $\mathbf{s}_{-i}$ is the response of interest $\mathbf{s}$ without token i.

In addition, we also replace the sequence likelihood used in Semantic Entropy (SE,Kuhn et al., 2023), which is an *uncertainty*[2] measure, with CSL, and report results in Section 4.6

4.3 Language Model and Generation

For the base LLMs, we include the most popular open LLMs: LLaMA2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023) and Gemma (Team et al., 2024). We use the 13B version for LLaMA, and the 7B version for Mistral and Gemma due to their improved performance. Response generation largely follows Kuhn et al. (2023) with some improvements: The original pipeline often removes content after periods in abbreviations (such as "Dr."), so we modified the prompt to ensure each response $\mathbf{s}$ ends with a newline character but keeps contents after punctuations. Like Kuhn et al. (2023), we focus on the greedily-decoded generation $\mathbf{s}$ for each question and use a temperature of 0.5 for baselines that require additional response sampling.

4.4 Evaluation

For an effective confidence measure, low confidence should correlate with a higher probability of incorrect generation. To assess the quality of confi-

[2]See Lin et al. (2023) for a discussion distinguishing between confidence and uncertainty.

Table 1: Agreement with human annotation, on 720 sampled question-response pairs in total.

Agreement with Human (%)	coqa	nq	trivia
acc_{agree}	98.2	91.8	98.2
acc_{llama2}	97.0	86.7	95.0
acc_{gpt}	94.1	85.8	96.6

dence measures, we adhere to established methodologies (Kuhn et al., 2023; Band et al., 2022). This involves utilizing them to predict the correctness of a generation and calculating the Area Under the Receiver Operating Characteristic curve (**AUROC**) for this prediction task. Formally, let acc_i represent the indicator function $\mathbb{1}\{\mathbf{s}_i \text{ correctly answers } x_i\}$. We compute the AUROC using the confidence measure $\mathrm{C}(x., \mathbf{s}.)$ to predict $acc.$. This approach allows us to systematically evaluate how well the confidence measures distinguish between correct and incorrect responses across multiple samples.

In addition to AUROC, following Lin et al. (2023), we also report Area Under Accuracy-Rejection Curve (**AUARC**) (Nadeem et al., 2009). The Accuracy-Rejection Curve (ARC) computes the average the accuracy when a subset of samples is rejected based on C. As we exclude more low-confidence samples, the accuracy of the remaining samples should increase. The *upper bound* is achieved by predicting only the correct $\mathbf{s}$ (i.e. set C to accuracy), and the AUARC of a *random* predictor is equal to the base accuracy without rejection.

Correctness of Generations: A critical requirement for computing AUROC or AUARC is a reliable acc_i, the accuracy of each response. This is a unique challenge in UQ for NLG, and deserves separate research. Prior work typically relies on lexical similarity measures such as ROUGE (Kuhn et al., 2023; Quach et al., 2023) or BLEU (Huang et al., 2024). Lin et al. (2023) uses `gpt-3.5` to evaluate the correctness of each response given a reference answer[3], and considers anything with a rating above 70% as correct. Inspired by (Lin et al., 2023), we use the agreement of both LLaMA2-70B and `gpt-3.5-turbo-0125`'s evaluations as acc (More details are in Appendix B). As shown in Table 1, this notably improves the correctness evaluation and thus the reliability of AUROC/AUARC. We include the results using acc_{llama2} in the Appendix D since acc_{agree} and acc_{gpt} may not remain reproducible in the future.

[3]The original paper uses `gpt-3.5-turbo-0301` which is no longer accessible.

4.5 Results

Table 2 presents the AUROC of different confidence measures. Clearly, all the confidence baselines consistently detect good $\mathbf{s}$ over bad ones, but CSL outperforms baselines. Similar results are observed for AUARC in Table 3 as well: As the LMs have good base accuracy (from the "Random" column) for `coqa` and `trivia`, the gap between different confidence measures is relatively small, but generally significant. In particular, among likelihood-based methods, it is sometimes confusing whether normalized or unnormalized likelihood should be used (Kuhn et al., 2023; Malinin and Gales, 2021), and `TokenSAR` does not always outperform the better of the two. However, CSL consistently outperforms all three. Thanks to the additional sampling, `Deg` performs quite well, especially if we further increase the temperature of the base LM (see Appendix), but in practice it could significantly increase the computation cost as it requires m additional generations and $O(m^2)$ similarity comparisons. Finally, after the head selection process, the difference between `CSL` and `CSL-Next` is small, but still extremely significant if we perform a pooled test. The observed similar performance is because chosen attention heads exhibit highly correlated patterns (see Section 4.7). We recommend `CSL` over `CSL-Next` as they share similar overhead (close to none) but the attention-eliciting prompt is more structured compared with the more arbitrary prompt used to generate $\mathbf{s}$, and therefore the "good" heads are likely to be more stable.

4.6 Improving Uncertainty Measures

As noted earlier, sequence likelihood is widely used in entropy computation, which is used as an uncertainty measure for NLG. Semantic Entropy (Kuhn et al., 2023) is a state-of-the-art uncertainty measure that groups sampled generations into semantic sets and computes the entropy over these sets. In doing so, it uses sequence likelihood (sometimes normalized). We simply replace it with `CSL` to create a new uncertainty measure, `SE+CSL`. The comparison is shown in Table 4. `SE+CSL` consistently outperforms either the normalized or the unnormalized version of `SE`, showing potential in replacing sequence likelihood in other domains such as conformal NLG (Quach et al., 2023).

4.7 Is the Improvement a Fluke?

Despite not using an explicit attention-eliciting prompt, `CSL-Next` performs close to `CSL`, outper-

Table 2: AUROC of using confidence measures C to predict the accuracy of responses. Methods not significantly different from the best are in **bold**.

	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
trivia (llama2)	81.74±0.25	64.82±0.18	88.19±0.12	87.86±0.12	87.91±0.13	**89.70±0.19**	**89.61±0.18**
trivia (gemma)	84.00±0.19	81.82±0.18	88.72±0.11	88.11±0.08	88.09±0.08	**89.71±0.14**	89.42±0.11
trivia (mistral)	81.85±0.30	68.78±0.28	88.81±0.15	88.64±0.14	88.74±0.13	**90.76±0.18**	**90.73±0.15**
coqa (llama2)	69.93±0.57	53.64±0.41	69.50±0.40	72.59±0.44	72.78±0.47	**73.34±0.74**	**73.36±0.57**
coqa (gemma)	70.03±0.68	55.99±0.42	70.83±0.46	71.96±0.57	72.38±0.55	**73.30±0.57**	**73.64±0.63**
coqa (mistral)	69.84±0.52	52.33±0.42	68.97±0.38	70.60±0.38	70.87±0.41	**71.79±0.75**	**71.91±0.63**
nq (llama2)	71.61±0.51	52.51±0.47	66.57±0.33	69.48±0.50	70.43±0.46	**73.73±0.49**	**73.54±0.46**
nq (gemma)	73.32±0.68	63.66±0.46	72.09±0.65	75.81±0.65	75.88±0.68	**77.95±0.58**	77.17±0.65
nq (mistral)	73.03±0.52	54.77±0.51	69.22±0.53	71.06±0.54	72.61±0.48	**76.65±0.43**	75.73±0.68

Table 3: AUARC of using confidence measures C to predict the accuracy of responses. Methods not significantly different from the best are in **bold**.

	Random	Upper Bound	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
trivia (llama2)	82.60±0.14	98.39±0.03	92.84±0.28	87.50±0.12	95.99±0.06	95.95±0.05	95.96±0.05	**96.32±0.08**	**96.29±0.06**
trivia (gemma)	78.14±0.13	97.41±0.03	91.48±0.19	91.83±0.11	94.61±0.05	94.38±0.04	94.38±0.04	**94.79±0.04**	94.65±0.04
trivia (mistral)	79.94±0.12	97.84±0.03	91.91±0.31	87.55±0.13	95.27±0.06	95.21±0.05	95.22±0.05	**95.68±0.09**	**95.68±0.07**
coqa (llama2)	91.36±0.17	99.62±0.01	94.71±0.22	92.24±0.18	95.69±0.13	96.09±0.14	96.17±0.14	**96.26±0.18**	**96.26±0.16**
coqa (gemma)	92.64±0.14	99.72±0.01	95.46±0.20	94.13±0.15	96.54±0.10	96.63±0.11	96.68±0.11	**96.85±0.11**	**96.90±0.13**
coqa (mistral)	92.04±0.14	99.67±0.01	95.09±0.33	92.62±0.16	95.92±0.10	96.13±0.11	96.22±0.10	**96.37±0.13**	**96.38±0.12**
nq (llama2)	56.49±0.68	88.74±0.39	70.59±1.27	57.68±0.68	70.51±0.76	71.01±0.81	71.97±0.80	**73.46±0.66**	**73.31±0.87**
nq (gemma)	47.16±0.65	82.59±0.49	62.96±0.94	57.30±0.67	65.78±0.95	66.41±0.92	67.02±0.93	**67.78±1.07**	66.89±1.03
nq (mistral)	52.90±0.74	86.57±0.47	67.48±1.00	56.47±0.93	69.15±0.82	69.27±0.72	70.62±0.65	**72.25±0.74**	71.85±0.69

Table 4: AUROC of using variants of Semantic Entropy to predict the accuracy of responses. SE+CSL significantly improves the original version based on vanilla sequence likelihoods.

	SE(norm)	SE	SE+CSL
trivia (llama2)	89.88±0.11	89.33±0.12	**90.50±0.12**
trivia (gemma)	90.33±0.10	90.02±0.12	**90.75±0.13**
trivia (mistral)	90.35±0.12	89.78±0.15	**91.13±0.13**
coqa (llama2)	**75.26±0.33**	72.50±0.34	**75.58±0.51**
coqa (gemma)	**74.81±0.48**	72.95±0.48	**75.08±0.50**
coqa (mistral)	74.06±0.41	71.54±0.32	**74.56±0.51**
nq (llama2)	74.13±0.51	69.62±0.38	**76.33±0.50**
nq (gemma)	77.93±0.58	73.67±0.77	**79.50±0.52**
nq (mistral)	75.71±0.40	71.85±0.49	**78.66±0.35**

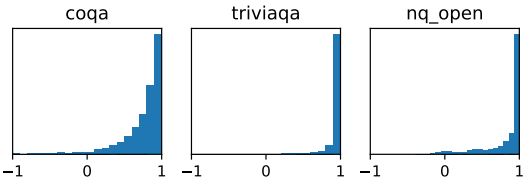

Figure 4: Histogram of the correlation between attentions from CSL and CSL-Next (top 10 heads' average). We keep only generations with more than 2 tokens. For most responses, the chosen heads' attentions are highly correlated, suggesting that both methods focus on the same tokens, as exemplified in Fig. 5.

forming baselines significantly. It is reasonable to then worry that the improvement is just the result of black-box data-mining by the head-selection step in Section 3.3. This section presents evidence that CSL most likely identifies meaningful concepts.

Fig. 4 shows the correlation between the w_i vectors used for CSL and CSL-Next—which is almost always positive and usually close to 1. As the prompts used to induce these attention weights are quite different, such agreement could be taken as preliminary evidence that these weights are "meaningful" and less likely to be purely the results of two independent black-box "fitting" processes.

Fig. 5 shows a few examples of the induced attention weights, illustrating how CSL makes the confidence measure more focused on important tokens. Given the nature of soft attention, despite the fact that w_i chosen by Section 3 generally identifies the important tokens and improves upon vanilla sequence likelihood, it is still sometimes difficult to interpret why each token is under/over-weighted. For interpretability reasons, a future research direction might be to directly use natural language to identify such tokens. For example, one might directly ask the LM to list the important entities with respect to a question. The challenge, then, transfers to identifying the actual tokens implied by the natural language output listing important entities in the original $\mathbf{s}$.

Question 1: how early did he want to get there?
Response 1: an hour before the time
Question 2: What does Barwell think of him?
Response 2: he is not fit to be his guardian
Question 3: who is Susan Boyle?
Response 3: Susan Boyle is a Scottish singer who became famous after appearing on the TV show "Britain's Got Talent" in 2009.

Figure 5: Tokens whose attention is increased are marked (others decreased). As expected, such re-weighting are not always interpretable, but help locating the more relevant tokens in general.

Table 5: AUROC using heads picked from coqa. Despite the big distribution shift from coqa to nq/trivia, the heads chosen on coqa still provides attention weights that significantly improves the AUROC.

	SL(norm)	CSL	CSL-Next
trivia (llama2)	87.86±0.12	**88.59±0.85**	88.37±0.87
trivia (gemma)	88.11±0.08	**89.03±0.61**	87.43±0.94
trivia (mistral)	88.64±0.14	**89.70±1.43**	88.46±1.93
nq (llama2)	69.48±0.50	**70.94±1.66**	70.10±2.09
nq (gemma)	75.81±0.65	**77.21±1.06**	74.99±1.30
nq (mistral)	**71.06±0.54**	**72.71±3.32**	**72.27±2.72**

Finally, if the attention weights are actually focusing on the important concepts as intended, one might expect that they should transfer *between datasets* and be more robust to distribution shifts. In Table 5, we select heads based on validation sets from coqa and apply them on the other two datasets, which are quite different from coqa (e.g. in the format of questions). CSL still provides consistent performance boost, while CSL-Next sometimes lags behind SL(norm). This suggests both the prompt and the head selection step increase the weights on the more relevant tokens.

4.8 Ablation: Choice of Attention Heads

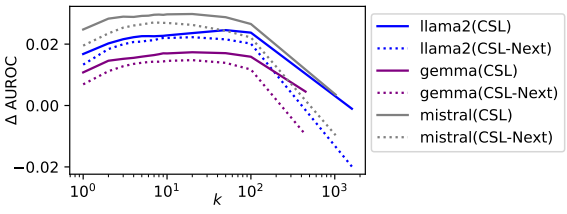

Figure 6: AUROC gain compared with SL(norm) for different k (number of attention heads to keep), from 1 to all heads. The performance peaks around $k = 10$ and is stable. CSL is also consistently better than CSL-Next- notably, when we average the attention from all heads, CSL still outperforms SL(norm), but CSL-Next is significantly worse.

In Fig. 6, we compute the *gain* in AUROC compared with SL(norm) (i.e. equal weighting), keeping different number of attention heads. The solid lines represent CSL, and the dotted lines denote CSL-Next. Note that using only one head is also significantly better than the baseline, but performance increases and peaks around 10 heads (sometimes more). We believe using a small number of "good" heads can reduce the noise introduced by using a small validation set, making CSL more contextualized to the question.

5 Discussion and Conclusion

In this paper, we explore enhancements to the widely used confidence measure, sequence likelihood, which serves as a quality metric for model generations in selective generation and risk control for natural language generation. We introduce Contextualized Sequence Likelihood, or CSL, a novel approach that utilizes attention weights on generated tokens to re-weight the logits in sequence likelihood computation. This new confidence measure surpasses existing methods across several popular datasets and large language models by more accurately predicting the accuracy of each response.

Despite these improvements, there are limitations to the current approach. First, the interpretability of attention weights is often obscured by the nature of the self-attention mechanism. While attention re-weighting generally enhances the confidence measure, there are individual cases where selected heads do not align with tokens that humans would typically consider important in the context of the question. As discussed in Section 4, a possible solution involves identifying key tokens through a language model via natural language, though this introduces the additional challenge of matching these words to the original response.

Another limitation is the applicability of the current prompt, which is tailored for question-answering and may require modifications for other tasks. Additionally, like many baseline methods, the current approach cannot leverage external information for "fact-checking." Integrating confidence from multiple models could potentially bridge the gap between a language model's perceived confidence and the actual correctness of a response. We hope that future research will address these issues and expand the toolkit available to practitioners for assessing the reliability of large language models.

References

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *Preprint*, arXiv:2307.15703.

Neil Band, Tim G. J. Rudner, Qixuan Feng, Angelos Filos, Zachary Nado, Michael W. Dusenberry, Ghassen Jerfel, Dustin Tran, and Yarin Gal. 2022. Benchmarking bayesian deep learning on diabetic retinopathy detection tasks. *ArXiv*, abs/2211.12717.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. 2020. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175*.

Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.

Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *Preprint*, arXiv:2307.01379.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *The Twelfth International Conference on Learning Representations*.

Giorgio Fumera, Fabio Roli, and Giorgio Giacinto. 2000. Reject option with multiple thresholds. *Pattern Recognition*.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero shot time series forecasters. In *Advances in Neural Information Processing Systems*.

Guande He, Peng Cui, Jianfei Chen, Wenbo Hu, and Jun Zhu. 2023. Investigating uncertainty calibration of aligned language models under the multiple-choice setting. *Preprint*, arXiv:2310.11732.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *Preprint*, arXiv:2311.08718.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. *arXiv preprint arXiv:2404.03163*.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Heinrich Jiang, Been Kim, Maya Gupta, and Melody Y. Guan. 2018. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. 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, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before assure: Confidence estimation for large language models through reflection on multiple answers. *arXiv preprint arXiv:2403.09972*.

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

Zhen Lin, Lucas Glass, M Brandon Westover, Cao Xiao, and Jimeng Sun. 2022b. Scrib: set-classifier with class-specific risk bounds for blackbox models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7497–7505.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *Preprint*, arXiv:2305.19187.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Sabrina J. Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *CoRR*, abs/2012.14983.

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

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. In *Proceedings of the third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pages 65–81, Ljubljana, Slovenia. PMLR.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. 2023. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. 2024. Lag-llama: Towards foundation models for probabilistic time series forecasting. *Preprint*, arXiv:2310.08278.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Yaorui Shi, An Zhang, Enzhi Zhang, Zhiyuan Liu, and Xiang Wang. 2023. ReLM: Leveraging language models for enhanced chemical reaction prediction. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. 2022. Re-examining calibration: The case of question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2814–2829, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Charles Spearman. 1961. The proof and measurement of association between two things.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022a. Investigating selective prediction approaches across several tasks in IID, OOD, and adversarial settings. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1995–2002, Dublin, Ireland. Association for Computational Linguistics.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. 2022b. Towards improving selective prediction ability of NLP systems. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 221–226, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*, volume 29. Springer.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and Jifeng Dai. 2023b. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuxia Wang, Daniel Beck, Timothy Baldwin, and Karin Verspoor. 2022. Uncertainty estimation and reduction of pre-trained models for text regression. *Transactions of the Association for Computational Linguistics*, 10:680–696.

Gwenyth Portillo Wightman, Alexandra DeLucia, and Mark Dredze. 2023. Strength in numbers: Estimating confidence of large language models by prompt agreement. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 326–362.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Jing Zhang, Hui Gao, Peng Zhang, Boda Feng, Wenmin Deng, and Yuexian Hou. 2024. LA-UCL: LLM-augmented unsupervised contrastive learning framework for few-shot text classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10198–10207, Torino, Italia. ELRA and ICCL.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

A Prompts

In this section, we present the full prompts used in various aspects of our experiments.

A.1 Question-Answering Generation Prompts

We use the following prompts when generating responses for the question-answering datasets.

CoQA:

```
Read the context and answer the questions below.

*Context*: [$context]
[additional question-answer pairs]
*Question*: [$question]
*Answer*:
```

where additional question-answer pairs are preceding turns of the conversation about the paragraph consisting of questions and reference answers.

TriviaQA:

```
Answer these questions:

*Question*: In Scotland a bothy/bothie is a?
*Answer*: House
*Question*: [$question]
*Answer*:
```

Natural Questions is a much harder dataset than TriviaQA, so we use the same 5-shot prompt version of the prompt in (Touvron et al., 2023a) (with 5 questions randomly picked from the training set).

```
Answer these questions:

*Question*: who makes up the state council in russia
*Answer*: governors and presidents
*Question*: when does real time with bill maher come back
*Answer*: November 9, 2018
*Question*: where did the phrase american dream come from
*Answer*: the mystique regarding frontier life
*Question*: what do you call a group of eels
*Answer*: bed
*Question*: who wrote the score for mission impossible fallout
*Answer*: Lorne Balfe
*Question*: [$question]
*Answer*:
```

A.2 Attention-eliciting Prompt

In the following, we provide the full prompt previewed in Fig. 1.

```
Read the following question with optional context and decide if the answer correctly answer the question. Focus on the answer, and reply Y or N.


Context: Luxor International Airport is a airport near Luxor in Egypt (EG). It is 353km away from the nearest seaport (Duba). The offical IATA for this airport is LXR.
Question: Luxor international airport is in which country?
Answer: It is in the United States, and its IATA is LXR.
Decision: N. (The airport is in Egypt, not the United States.)


Context: Harry is a good witcher.
Question: How old is Harry?
Answer: Harry practices witchcraft.
Decision: N. (The answer does not mention Harry's age.)


Question: What is the capital of Kenya?
Answer: Nairobi is the capital of Kenya.
Decision: Y.


Question: Who has won the most Premier League titles since 2015?
Answer: Manchester City have win the most Premier League title after 2015.
Decision: Y. (Grammar errors are ignored.)


[$optional_context]
Question: [$question]
Answer: [$response]
Decision:
```

A.3 Automatic Accuracy Evaluation

For the prompts used to elicit judgment from `gpt-3.5-turbo-0125` and LLaMA2-70B, we use the same ones from the Appendix of Lin et al. (2023).

B Automatic Accuracy Evaluation

To verify the efficacy of automatic accuracy evaluation by `gpt-3.5-turbo-0125` and LLaMA2-70B, we compare their judgements' alignment with human annotations. Specifically, we first sample 80 (question, response) pairs for each (model, dataset), resulting in 720 samples in total. We then manually compare each sample's response with the reference answer, and decide if the generated response is correct (given the context if any). Then, we retrieve the ratings on these 720 samples from `gpt-3.5-turbo-0125` and LLaMA2-70B, and find the thresholds that result in the highest agreement. The resulting accuracy-threshold relation is illustrated in Fig. 7. From the results, we chose 0.2 as the threshold for LLaMA2-70B and 0.6 for `gpt-3.5-turbo-0125`. In other words, denote the rating from LLaMA2-70B on a gener-

ated response $\mathbf{s}$ as $r_{llama2}(\mathbf{s})$, then $\text{acc}_{llama2}(\mathbf{s}) = \mathbb{1}\{r_{llama2}(\mathbf{s}) \geq 0.2\}$. Note that we found 20 out of the 720 samples hard to decide during our manual annotation process, usually due to incorrect reference answers or intrinsic ambiguity in the questions. We ignore them during the computation of Fig. 7.

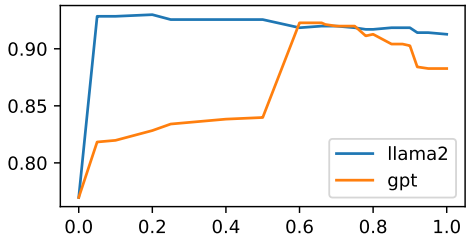

Figure 7: Agreement with human annotation with different thresholds over all 720 samples.

C Temperature of Base Language Model

We include results using a high temperature of 1.0 in Tables 6 and 7. Higher temperature increases the sampling diversity of `Deg` and `P(true)`, which increases the performance of `Deg` significantly, but brings mixed results to `P(true)`. We also note that with 5 additional generations (6 in total) `Deg` is sometimes better than `CSL` or `CSL-Next`. We repeat the uncertainty experiments with higher temperature as well, in Table 8, and found that `SE+CSL` is the most predictive of the generations' quality at this temperature.

The experiments here suggest that sampling from an appropriate temperature really helps quantifying the reliability of the generations, but it should be noted that either sampling based confidence measures (`Deg`,`P(true)`) or sampling-based uncertainty measures (`SE(norm)`,`SE`,`SE+CSL`) bear a significantly higher computational overhead - $(m+1)\times$ the generation cost and up to (m^2-m) pairwise NLI inference, with m being the additional generations. On the other hand, CSL requires one inference (not generation) call to the base LM which is only a fraction of the overhead. In practice, one might choose the best practice basing on cost and latency tolerance as well as difference in performance for the particular data distribution, and as shown in `SE+CSL`, CSL could be combined with sampling to yield better results as well.

D Results Using acc_{llama2}

We include results using acc_{llama2} (accuracy as judged by LLaMA2-70B) for reproducibility purposes in Tables 9 and 10. Conclusions stay the same as in the main text.

E Calibration Quality

We perform post-hoc calibration using the default `sklearn.calibration.CalibratedClassifierCV`. Fig. 8 shows the reliability diagrams (similar to those in Kull et al. (2019)) after calibration for `CSL`. The resulting probabilities are overall quite calibrated.

Table 6: Similar to Table 2, but with different number of generations at temperature of 1.0.

	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
3 generations							
trivia (llama2)	86.49±0.14	61.63±0.27	88.16±0.13	87.84±0.13	87.89±0.13	**89.65±0.20**	**89.58±0.17**
trivia (gemma)	87.00±0.12	77.87±0.17	88.83±0.11	88.10±0.09	88.07±0.09	**89.73±0.16**	89.43±0.10
trivia (mistral)	87.01±0.21	73.75±0.16	88.85±0.15	88.66±0.14	88.75±0.13	**90.77±0.17**	**90.75±0.13**
coqa (llama2)	72.51±0.29	53.83±0.40	69.48±0.39	72.67±0.43	**72.88±0.46**	**73.46±0.69**	**73.45±0.56**
coqa (gemma)	**73.27±0.44**	56.82±0.57	70.78±0.47	72.07±0.52	72.46±0.49	73.07±0.52	**73.60±0.52**
coqa (mistral)	**71.86±0.57**	55.04±0.47	68.98±0.38	70.61±0.38	70.89±0.41	**71.79±0.74**	**71.91±0.63**
nq (llama2)	72.47±0.50	50.33±0.55	66.64±0.35	69.48±0.50	70.41±0.46	**73.83±0.52**	**73.59±0.48**
nq (gemma)	75.78±0.68	62.57±0.41	72.09±0.65	75.81±0.65	75.88±0.68	**77.96±0.57**	77.19±0.63
nq (mistral)	73.76±0.70	58.65±0.51	69.22±0.53	71.06±0.54	72.61±0.48	**76.65±0.43**	75.73±0.68
5 generations							
trivia (llama2)	88.79±0.12	61.63±0.27	88.16±0.13	87.84±0.13	87.89±0.13	**89.65±0.20**	**89.58±0.17**
trivia (gemma)	89.19±0.11	77.87±0.17	88.83±0.11	88.10±0.09	88.07±0.09	**89.73±0.16**	89.43±0.10
trivia (mistral)	89.39±0.21	73.75±0.16	88.85±0.15	88.66±0.14	88.75±0.13	**90.77±0.17**	**90.75±0.13**
coqa (llama2)	**75.73±0.30**	53.83±0.40	69.48±0.39	72.67±0.43	72.88±0.46	73.46±0.69	73.45±0.56
coqa (gemma)	**76.40±0.50**	56.82±0.57	70.78±0.47	72.07±0.52	72.46±0.49	73.07±0.52	73.60±0.52
coqa (mistral)	**74.41±0.51**	55.04±0.47	68.98±0.38	70.61±0.38	70.89±0.41	71.79±0.74	71.91±0.63
nq (llama2)	**74.09±0.53**	50.33±0.55	66.64±0.35	69.48±0.50	70.41±0.46	**73.83±0.52**	73.59±0.48
nq (gemma)	77.40±0.74	62.57±0.41	72.09±0.65	75.81±0.65	75.88±0.68	**77.96±0.57**	77.19±0.63
nq (mistral)	**76.31±0.65**	58.65±0.51	69.22±0.53	71.06±0.54	72.61±0.48	**76.65±0.43**	75.73±0.68

Table 7: Similar to Table 3, but with different number of generations at temperature of 1.0.

	Random	Upper Bound	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
3 generations									
trivia (llama2)	82.61±0.14	98.39±0.03	95.04±0.18	87.06±0.16	95.99±0.06	95.95±0.05	95.96±0.05	**96.33±0.08**	**96.30±0.07**
trivia (gemma)	78.11±0.12	97.41±0.03	93.52±0.12	91.01±0.11	94.62±0.05	94.37±0.04	94.37±0.04	**94.79±0.04**	94.66±0.04
trivia (mistral)	79.90±0.12	97.83±0.03	94.12±0.16	90.17±0.10	95.27±0.06	95.20±0.05	95.21±0.05	**95.67±0.09**	**95.67±0.07**
coqa (llama2)	91.36±0.17	99.62±0.02	95.52±0.20	92.20±0.21	95.69±0.13	96.11±0.14	**96.19±0.14**	**96.28±0.17**	**96.27±0.15**
coqa (gemma)	92.63±0.14	99.72±0.01	96.14±0.19	94.33±0.17	96.53±0.11	96.66±0.10	96.70±0.10	96.82±0.09	**96.90±0.11**
coqa (mistral)	92.04±0.14	99.67±0.01	95.80±0.26	93.06±0.19	95.92±0.10	96.13±0.11	96.22±0.10	**96.37±0.14**	**96.39±0.11**
nq (llama2)	56.48±0.68	88.74±0.39	72.37±0.89	56.84±0.70	70.50±0.77	71.00±0.81	71.95±0.80	**73.49±0.67**	**73.28±0.86**
nq (gemma)	47.16±0.65	82.59±0.49	66.59±0.87	56.54±0.78	65.78±0.95	66.41±0.92	67.02±0.93	**67.78±1.07**	66.89±1.03
nq (mistral)	52.90±0.74	86.57±0.47	70.09±0.84	59.77±0.52	69.15±0.82	69.27±0.72	70.62±0.65	**72.24±0.74**	71.83±0.67
5 generations									
trivia (llama2)	82.61±0.14	98.39±0.03	95.80±0.09	87.06±0.16	95.99±0.06	95.95±0.05	95.96±0.05	**96.33±0.08**	**96.30±0.07**
trivia (gemma)	78.11±0.12	97.41±0.03	94.51±0.08	91.01±0.11	94.62±0.05	94.37±0.04	94.37±0.04	**94.79±0.04**	94.66±0.04
trivia (mistral)	79.90±0.12	97.83±0.03	95.15±0.11	90.17±0.10	95.27±0.06	95.20±0.05	95.21±0.05	**95.67±0.09**	**95.67±0.07**
coqa (llama2)	91.36±0.17	99.62±0.02	96.13±0.15	92.20±0.21	95.69±0.13	96.11±0.14	**96.19±0.14**	**96.28±0.17**	**96.27±0.15**
coqa (gemma)	92.63±0.14	99.72±0.01	96.74±0.20	94.33±0.17	96.53±0.11	96.66±0.10	96.70±0.10	96.82±0.09	**96.90±0.11**
coqa (mistral)	92.04±0.14	99.67±0.01	96.22±0.16	93.06±0.19	95.92±0.10	96.13±0.11	96.22±0.10	**96.37±0.14**	**96.39±0.11**
nq (llama2)	56.48±0.68	88.74±0.39	**73.77±0.89**	56.84±0.70	70.50±0.77	71.00±0.81	71.95±0.80	**73.49±0.67**	73.28±0.86
nq (gemma)	47.16±0.65	82.59±0.49	**67.47±0.98**	56.54±0.78	65.78±0.95	66.41±0.92	67.02±0.93	**67.78±1.07**	66.89±1.03
nq (mistral)	52.90±0.74	86.57±0.47	**72.07±0.84**	59.77±0.52	69.15±0.82	69.27±0.72	70.62±0.65	**72.24±0.74**	71.83±0.67

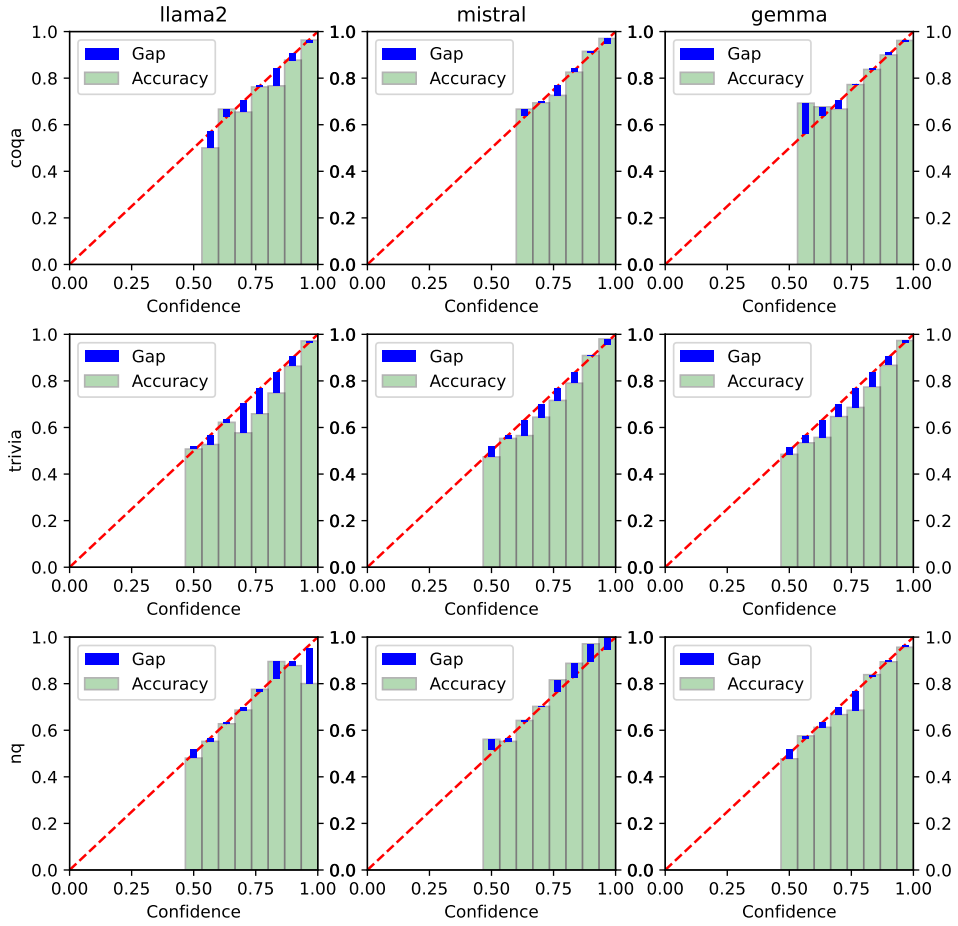

Figure 8: Reliability diagrams of CSL. Bins with fewer than 10 samples are ignored due to noise.

Table 8: Similar to Table 4, but with different number of generations at temperature of 1.0.

	SE(norm)	SE	SE+CSL
3 generations			
trivia (llama2)	88.18±0.11	87.90±0.08	**88.99±0.13**
trivia (gemma)	88.05±0.13	88.08±0.15	**88.73±0.15**
trivia (mistral)	88.90±0.16	88.50±0.18	**89.84±0.16**
coqa (llama2)	**75.26±0.46**	72.11±0.32	**75.27±0.57**
coqa (gemma)	**75.42±0.38**	72.81±0.36	**75.25±0.42**
coqa (mistral)	73.53±0.35	70.49±0.37	**73.89±0.50**
nq (llama2)	72.91±0.42	68.44±0.34	**74.05±0.47**
nq (gemma)	77.36±0.73	72.73±0.80	**78.33±0.69**
nq (mistral)	75.65±0.66	71.15±0.72	**77.37±0.59**
5 generations			
trivia (llama2)	89.40±0.11	89.17±0.10	**89.95±0.11**
trivia (gemma)	89.32±0.09	89.31±0.12	**89.82±0.12**
trivia (mistral)	90.38±0.13	89.86±0.16	**90.91±0.14**
coqa (llama2)	**77.08±0.40**	73.87±0.33	**77.25±0.47**
coqa (gemma)	**77.24±0.51**	74.39±0.46	**77.11±0.46**
coqa (mistral)	75.65±0.31	72.58±0.29	**76.04±0.50**
nq (llama2)	74.45±0.37	70.33±0.35	**75.90±0.41**
nq (gemma)	78.31±0.54	74.17±0.69	**79.33±0.57**
nq (mistral)	77.02±0.68	73.17±0.65	**78.82±0.55**

Table 9: Like Table 2, but using accuracy from LLaMA2-70B.

	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
trivia (llama2)	80.30±0.30	63.90±0.17	86.33±0.11	85.83±0.12	85.86±0.13	**87.72±0.24**	**87.61±0.22**
trivia (gemma)	81.92±0.16	79.94±0.18	86.21±0.10	85.56±0.09	85.51±0.08	**87.14±0.14**	86.89±0.11
trivia (mistral)	79.95±0.25	67.67±0.27	86.33±0.13	86.14±0.13	86.23±0.11	**88.31±0.18**	**88.22±0.13**
coqa (llama2)	68.32±0.62	53.41±0.41	68.17±0.33	71.04±0.53	71.29±0.54	**71.93±0.68**	**71.70±0.63**
coqa (gemma)	69.19±0.62	54.90±0.36	69.67±0.42	70.39±0.63	70.73±0.59	**71.65±0.63**	**71.96±0.64**
coqa (mistral)	68.52±0.52	52.66±0.37	67.72±0.30	69.07±0.39	69.30±0.40	**70.31±0.79**	**70.18±0.66**
nq (llama2)	68.69±0.42	51.70±0.38	64.47±0.44	66.07±0.42	66.64±0.36	**69.86±0.52**	69.53±0.45
nq (gemma)	69.73±0.57	60.51±0.55	69.34±0.47	70.84±0.58	70.60±0.62	**73.51±0.52**	72.27±0.59
nq (mistral)	69.45±0.55	52.60±0.39	66.70±0.49	67.14±0.42	68.07±0.44	**72.34±0.45**	71.19±0.58

Table 10: Like Table 3, but using accuracy from LLaMA2-70B.

	Random	Upper Bound	Deg (E)	P(true)	SL	SL(norm)	TokenSAR	CSL	CSL-Next
trivia (llama2)	82.71±0.13	98.41±0.03	92.48±0.20	87.25±0.13	95.60±0.06	95.51±0.05	95.49±0.05	**95.91±0.06**	95.86±0.07
trivia (gemma)	78.58±0.12	97.52±0.03	91.02±0.23	91.35±0.14	94.12±0.05	93.85±0.04	93.85±0.05	**94.29±0.04**	94.14±0.05
trivia (mistral)	80.23±0.11	97.90±0.02	91.48±0.33	87.38±0.13	94.75±0.06	94.67±0.05	94.66±0.05	**95.16±0.06**	**95.15±0.04**
coqa (llama2)	91.00±0.17	99.58±0.02	94.19±0.16	91.86±0.17	95.30±0.14	95.65±0.16	95.73±0.16	**95.88±0.17**	95.81±0.17
coqa (gemma)	92.14±0.15	99.68±0.01	95.10±0.26	93.39±0.15	96.09±0.11	96.12±0.14	96.18±0.13	**96.35±0.12**	**96.39±0.14**
coqa (mistral)	91.61±0.16	99.64±0.01	94.81±0.34	92.28±0.18	95.47±0.11	95.63±0.13	95.71±0.12	**95.86±0.18**	**95.86±0.15**
nq (llama2)	56.74±0.60	88.89±0.34	68.96±1.00	57.57±0.60	68.96±0.70	69.01±0.73	69.72±0.71	**71.10±0.64**	70.80±0.78
nq (gemma)	48.50±0.60	83.59±0.44	61.92±1.07	56.26±0.62	64.78±0.73	64.47±0.85	64.75±0.85	**65.96±0.85**	64.95±0.99
nq (mistral)	53.65±0.59	87.05±0.37	65.99±0.99	56.07±0.86	67.63±0.69	67.29±0.57	68.31±0.54	**69.88±0.64**	69.46±0.54