FLUE: Streamlined Uncertainty Estimation for Large Language Models

Shiqi Gao, [1] **Tianxiang Gong,** [1] **Zijie Lin,** [3] **Runhua Xu,** [1] **Haoyi Zhou,** [2, 4*] **Jianxin Li** [1, 4]

[1] SKLCCSE, School of Computer Science and Engineering, Beihang University
[2] School of Software, Beihang University
[3] National University of Singapore
[4] Zhongguancun Laboratory, Beijing
{gaoshiqi, gongtianxiang, runhua, haoyi, lijx}@buaa.edu.cn, lin.zijie@u.nus.edu

Abstract

Uncertainty estimation is essential for practical applications such as decision-making, risk assessment, and human-AI collaboration. However, uncertainty estimation in open-ended question-answering (QA) tasks presents unique challenges. The output space for open-ended QA is vast and discrete, and the autoregressive nature of LLMs, combined with the rapid increase in model parameters, makes inference sampling significantly costly. An ideal uncertainty estimation for LLMs should meet two criteria: 1) incur no additional inference cost and 2) capture the semantic dependencies of token-level uncertainty within sequences. We propose a promising solution that converts redundancy into randomness in the extensive parameters of LLMs to quantify knowledge uncertainty. We can obtain token-level Monte Carlo samples without multiple inferences by introducing randomness during a single forward pass. We theoretically analyze the FLUE sampling method and employ a post-processing method to learn the state transitions from token uncertainty to sequence uncertainty. In open-ended QA tasks, we demonstrate that FLUE can achieve competitive performance in estimating the uncertainty of generated sentences *without extra inference overhead*.

1 Introduction

The increasing capabilities of Large Language Models (LLMs) (Achiam et al. 2023; Touvron et al. 2023a) for a wide range of applications have sparked growing interest in uncertainty estimation for tasks such as decision-making, risk assessment, and human-AI collaboration in real-world scenarios (Amodei et al. 2016; Xiao and Wang 2019; Ye et al. 2024). However, uncertainty estimation for open-ended question-answering (QA) tasks, a unique advantage of LLMs, faces distinct challenges compared to well-studied tasks such as classification and regression (Malinin and Gales 2021). The output space for open-ended question answering is discrete and enormous (Malinin and Gales 2021) (i.e., for each question, there are $L^{|\mathcal{V}|}$ possible outputs, where L is the output length and $\mathcal{V}$ is the vocabulary size). Moreover, the autoregressive generative structure and the rapid growth of model parameters make inference sampling significantly expensive.

*Corresponding author.

Several research efforts have focused on providing targeted optimizations for uncertainty estimation in open-ended QA with LLMs to address these challenges. Most of these uncertainty estimation approaches are sample-based methods (Malinin and Gales 2021; Kuhn, Gal, and Farquhar 2023), which essentially measure the consistency of multiple sampling outputs by introducing randomness into the model output process. There are two main approaches: 1) measuring token-level uncertainty (Malinin and Gales 2021; Balabanov and Linander 2024) and approximating sequence uncertainty through entropy chain propagation, where the uncertainty arises from the inconsistency of the conditional prediction posterior of ensemble models under the same prefix condition, ignoring the potential nonlinear relationship between the uncertainties of continuous tokens; and 2) measuring sequence-level uncertainty (Farquhar et al. 2024; Manakul, Liusie, and Gales 2023), which leverages the randomness of decoding (e.g., controlling Temperature or Top-K) to obtain diverse sequence semantics and measure the inconsistency in the semantic space, inherently requiring additional inference costs.

An effective uncertainty estimation for LLMs should possess two key characteristics: 1) no additional inference cost and 2) the ability to capture the semantic features of uncertainty in the sequence. One potential solution is to utilize the large number of parameters in LLMs to introduce randomness in the output and measure knowledge uncertainty. Recent work has revealed a clear distinction between the layers in LLMs, with knowledge storage typically located in the higher-level FFN layers (Chuang et al. 2023; Geva et al. 2020), and these layers usually exhibit similar characteristics (Sun et al. 2024). Introducing randomness during a single forward pass enables the acquisition of token-level Monte Carlo samples without requiring multiple inferences.

In this work, we initially demonstrated that the entropy of hidden states within a specific layer interval of LLMs can serve as an approximate upper bound for the entropy of predictive posterior, providing a theoretical foundation for subsequent sampling strategies. We then introduced the MC-FLUE sampling method, transforming the multi-layer hidden states during forward propagation into a samplable output space. Sampling hidden states within layer intervals provides an approximation of empirical entropy, which aligns with the empirical entropy obtained through MC-Dropout

(Gal and Ghahramani 2016). Although this approximation is effective only within specified layers, it can still be considered as an approximate upper bound for standard uncertainty estimation. Finally, we designed a post-processing approach for token uncertainty, updating sequence uncertainty through a state transition model using new observations (token uncertainties). Importantly, the state transition model can utilize more powerful architectures to capture potential temporal dependencies and non-linear relationships in the token uncertainty sequence, mitigating the difficulty of token-based uncertainty methods in accounting for semantic relationships in generated content.

We validate the effectiveness of FLUE through open-ended question-answering tasks. We demonstrated that it can achieve competitive performance in estimating the uncertainty of generated sentences without incurring extra inference overhead. This efficiency positions FLUE as a promising solution for enhancing the safety and reliability of LLM-powered applications, enabling their deployment in real-time and resource-constrained environments.

Our contributions are as follows:

- We establish a theoretical analysis demonstrating that the entropy of hidden states in intermediate layers of LLMs approximates an upper bound for predictive distribution entropy. Based on this, we introduce MC-FLUE, a novel single-pass sampling strategy that approximates MC-Dropout without multiple inferences.
- We propose a state transition-based post-processing approach to capture temporal dependencies and non-linear relationships in token uncertainty sequences, enabling semantic-aware sequence-level uncertainty estimation.
- Empirical evaluations on open-ended QA tasks, conducted across diverse open-source language models, demonstrate FLUE's competitive performance while maintaining computational parity with standard inference without uncertainty quantification.

2 Background

From a Bayesian perspective, model outputs involve probability distributions rather than single point estimates (MacKay 2003). Bayesian methods combine prior knowledge with observed data to update beliefs about model parameters, resulting in a posterior distribution. This posterior distribution reflects the likelihood of parameters given the data, thus quantifying the uncertainty in the predictions.

2.1 Token Uncertainty of LLMs

LLMs aim to learn the conditional probability distribution $P(y_l|y_1,\ldots,y_{l-1})$ over sequences from a vocabulary $\mathcal{V}$. Consider models $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}^{(m)})$, $\boldsymbol{\theta}^{(m)} \sim q(\boldsymbol{\theta})$, where $q(\boldsymbol{\theta})$ is the implicit approximation of posterior $p(\boldsymbol{\theta}|\mathcal{D})$. Each model captures the mapping between an input sequence $\boldsymbol{x} = x_1,\ldots,x_T \in \mathcal{X}$ and a target sequence $\boldsymbol{y} = y_1,\ldots,y_L \in \mathcal{Y}$, with $x_t, y_l \in \{\omega_1,\ldots,\omega_{|\mathcal{V}|}\}$. The token-level predictive posterior $P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})$ of an LLM can be obtained from following:

$$P(y_l|\boldsymbol{y}<l,\boldsymbol{x},\mathcal{D}) = \mathbb{E}_{q(\boldsymbol{\theta})}[P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\theta})], \quad (1)$$

The models in the ensemble can be approximated by MC-Dropout (Gal and Ghahramani 2016) or Lora-ensemble (Balabanov and Linander 2024). The token uncertainty is then estimated by the entropy of the predictive posterior (Malinin and Gales 2021):

$$\mathcal{H}[P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})] = \mathbb{E}_{P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})}[-\ln P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})] \quad (2)$$

2.2 Sequence Uncertainty of LLMs

Length Normalization (Malinin and Gales 2021) employs a normalization approach, which accumulates token-level uncertainties while constraining the uncertainties of sequences of different lengths to the same scale:

$$\mathcal{H}[P(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})] \approx \mathbb{E}_{\boldsymbol{y}\sim P(\boldsymbol{y}|\boldsymbol{x},\mathcal{D})}\left[\frac{1}{L}\sum_{l=1}^{L}\mathcal{H}[P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\mathcal{D})]\right] \quad (3)$$

While length-normalized uncertainty uses the same sequence prefix for sequence-level uncertainty, it still requires multiple inferences as token-level uncertainty estimation needs Monte-Carlo sampling (Equation 2).

Another sequence-level uncertainty estimation method is Semantic Entropy, which generates multiple completions and calculates the entropy of their distribution in semantic space (Farquhar et al. 2024).

3 Methodology

First, Section 3.1 introduces the theoretical foundation for using the entropy of hidden states as an upper bound for the entropy of the predictive posterior. Section 3.2 describes the MC-FLUE sampling method in detail, explaining how to sample from multiple layers of hidden states in a single forward pass. Section 3.3 discusses a post-processing method for converting token-level uncertainty to sequence-level.

3.1 Uncertainty of Hidden States

Let's consider a scenario where we perform forward passes on multiple models in an ensemble and Monte-Carlo sampling on the hidden layer states rather than on the predictive posterior. Consider a language model $P(y_l|\boldsymbol{y}_{<l},\boldsymbol{x},\boldsymbol{\theta})$ with residual connections between Transformer decoder layers. Let $\mathbf{H}_l^{(i)}$ denote the hidden state of the l-th token at the i-th layer, $\tilde{\mathbf{H}}_l^{(a:b)}$ represent the hidden state from layer a to b considering residual connections, and y_l be the output variable for the l-th token. Assume the inference process of the language model satisfies the Markov property, i.e., given $\tilde{\mathbf{H}}_l^{(a:b)}$ and $\boldsymbol{\theta}$, y_l is conditionally independent of other variables. We have the following proposition.

Proposition 1. *When b is sufficiently close to the total number of model layers N, there exists a small positive constant δ such that the entropy of the hidden states approximately upper bounds the entropy of the predictive posterior:*

$$\mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}\left[\mathcal{H}[y_l|\boldsymbol{\theta}]\right] \lesssim \mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}\left[\mathcal{H}[\tilde{\mathbf{H}}_l^{(b)}|\boldsymbol{\theta}]\right] + \delta \quad (4)$$

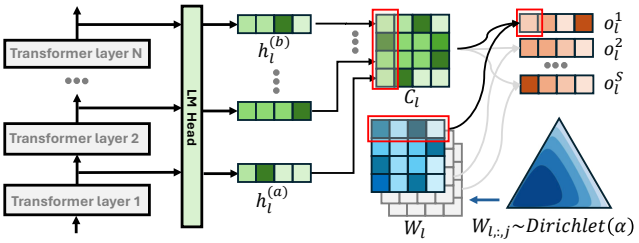


Figure 1: FLUE sampling strategy. During a single forward pass, FLUE selects a consecutive range of hidden layers and applies *layer-wise randomness* to their hidden states, approximating the effect of multiple samples.

where $\boldsymbol{\theta}$ represents the parameters of the model, and $P(\boldsymbol{\theta}|\mathcal{D})$ is the posterior distribution of the parameters given the training data $\mathcal{D}$.

Proposition 1 suggests that the entropy of the hidden states can serve as an approximate upper bound for the entropy of the predictive posterior, with a small additional term δ. The details of the proof are provided in Appendix A.

This proposition relies on the observation that in well-designed and sufficiently trained LLMs, hidden states in higher layers (i.e., as b approaches N) become increasingly informative about the final output, leading us to assume that the conditional entropy $\mathcal{H}[y_l|\tilde{\mathbf{H}}_l^{(b)}]$ becomes very small.

In the following, we describe the details of applying Proposition 1 in practical scenarios. At each position l, we perform S Monte Carlo samples on the hidden state $\mathbf{H}_l^{(b)}$ to obtain S samples $h_l^{(1)}, \ldots, h_l^{(S)}$. Then, we pass these samples through the output layer (i.e., *lm_head*, since it is typically a linear mapping, we will omit this term) to estimate the empirical entropy of the hidden state $\mathbf{H}_l^{(b)}$:

$$\hat{\mathcal{H}}^{(s)}[\mathbf{H}_l^{(b)}] = -\frac{1}{S}\sum_{s=1}^{S} \log P(h_l^{(s)}|\boldsymbol{y}_{<l}, \boldsymbol{x}, \mathcal{D}) \tag{5}$$

where $P(h_l^{(s)}|\boldsymbol{y}_{<l}, \boldsymbol{x}, \mathcal{D})$ is the probability density of the hidden state $h_l^{(s)}$ for the l-th token.

According to Proposition 1, we can use the empirical entropy $\hat{\mathcal{H}}^{(S)}[\mathbf{H}_l]$ to approximate the predictive posterior:

$$\mathcal{H}[P(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}, \mathcal{D})] \approx \mathbb{E}_{q(\boldsymbol{\theta})}\left[\mathcal{H}^{(S)}[\mathbf{H}_l|\boldsymbol{\theta}]\right] \approx \hat{\mathcal{H}^{(S)}}[\mathbf{H}_l] \tag{6}$$

where $\mathcal{H}[P(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}, \mathcal{D})]$ represents the entropy of the output probability distribution $P(y_l|\boldsymbol{y}_{<l}, \boldsymbol{x}, \mathcal{D})$ for l-th token.

In conclusion, by performing Monte-Carlo sampling on the hidden states, we can estimate the uncertainty of the hidden states $\hat{\mathcal{H}}[H_l]$, and this uncertainty can approximately upper bound the uncertainty of the token-level uncertainty.

3.2 FLUE Sampling Strategy

Next, we will estimate the token-level hidden state uncertainty in a single forward pass.

The Deep Ensemble method requires expensive pre-training for each model and inference on multiple models simultaneously during inference, which is computationally unacceptable.

Fortunately, redundancy in the parameters of LLMs (containing multiple structurally identical transformer encoder layers) may provide opportunities for sampling during a single forward pass. Based on this idea, we propose a more efficient sampling strategy called FLUE (Fast LLMs Uncertainty Estimation). The key idea behind the FLUE sampling strategy is to utilize the high similarity between the hidden states of different transformer decoder layers to approximate MC-Dropout (Gal and Ghahramani 2016). Empirically, the hidden layer states in the middle to upper layers are usually highly similar (Sun et al. 2024). When the hidden layer states of these layers are highly similar during inference, it may indicate that the model is certain when processing the current input (Chuang et al. 2023). We will first introduce how to identify a set of consecutive layers with similarity constraints and then describe the method for sampling within these layers.

We use ε to limit the sampling layer range LS. In detail, for any layer i and layer j in LS, the cosine similarity of their output hidden layer states satisfies $c_{i,j} > 1 - \epsilon$. We empirically evaluated the hidden state similarity of multiple open-source models in Figure 2 to illustrate the approximate range of ε selection.

Alternatively, we can adaptively select the layer range with:

$$LS_l = \left\{ i \middle| \frac{1}{N-1}\sum_{j=1,j\neq i}^{N} c_{i,j}^{(l)} > \beta \cdot \bar{c}_l, i = 1, 2, \cdots, N \right\} \tag{7}$$

where N is the number of transformer decoder layers, and $\bar{c}$ is the average cosine similarity of all layer pairs $(i, j), i \neq j$, i.e.:

$$\bar{c} = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{i,j} \tag{8}$$

Next, we describe the process of FLUE sampling. Specifically, in position l of sequence, we concatenate the hidden states of the LS layers into a matrix $\mathbf{C}_l \in \mathbb{R}^{|\mathcal{V}|\times|LS|}$:

$$\mathbf{C}_l = [h_l^{(a)}, h_l^{(a+1)}, \cdots, h_l^{(b)}] \tag{9}$$

where $a = \min(LS), b = \max(LS)$. We introduce a weight matrix $\mathbf{W}_l \in \mathbb{R}^{|LS|\times|\mathcal{V}|}$, where each column $\mathbf{W}_{l,:,j}$ of the matrix $\mathbf{W}_l$ is a random variable following a Dirichlet distribution:

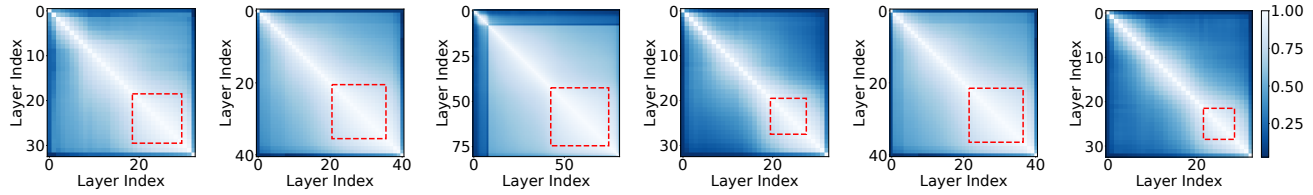
$$\mathbf{W}_{l,:,j} \sim \text{Dirichlet}(\alpha), \quad j = 1, 2, \cdots, |\mathcal{V}| \tag{10}$$

where α is the parameter of the Dirichlet distribution, controlling the concentration of the distribution.

Subsequently, we perform S independent samples on Dirichlet(α) to obtain S matrices $\mathbf{W}_l^{(1)}, \mathbf{W}_l^{(2)}, \cdots, \mathbf{W}_l^{(S)}$, and compute S output vectors $o_l^{(s)} \in \mathbb{R}^{|\mathcal{V}|}$:

$$o_l^{(s)} = \sum_{k=1}^{|LS|}(\mathbf{C}_l[:, k] \odot \mathbf{W}_l^{(s)}[k, :]) \tag{11}$$

Assume that the cosine similarity between the hidden states of different transformer decoder layers satisfies $c_{i,j} \geq$



(a) LLaMA-2-7B (b) LLaMA-2-13B (c) LLaMA-2-70B (d) LLaMA-3-8B (e) Vicuna-13B (f) Mistral-7B

Figure 2: Cosine similarity between different hidden states. To implement *Proposition 2*, we choose $\varepsilon = 0.25$. We highlight in red rectangles the set of transformer decoder layers participating in FLUE sampling, which are the layers where the cosine similarity between the hidden states satisfies $c_{i,j} \geq 1 - \varepsilon$.

$1 - \varepsilon$, where ε is a small positive number. Also assume that $o^{(s)} \sim p_o(x), r^{(s)} \sim p_r(x), s = 1, 2, \ldots, S$, where $r^{(s)}$ are S sampling results obtained through MC-Dropout, and $\frac{1}{S}\sum_{s=1}^{S} r^{(s)} \xrightarrow{P} \mathbb{E}_{p_r}[x]$.

Proposition 2. *Let $\varepsilon > 0$ and $\alpha > 0$. If $c_{i,j} \geq 1 - \varepsilon$ for all i, j, then:*

$$\lim_{\varepsilon \to 0, \alpha \to \infty} D_{KL}(p_r || p_o) = 0 \tag{12}$$

where D_{KL} denotes the Kullback-Leibler divergence, p_r is the output distribution after dropout, and p_o is the output distribution obtained by sampling the hidden layer states.

Proof. Due to space limitations, we provide the detailed proof in Appendix B. □

Given the conclusion of Proposition 2, we can derive the following proposition.

Proposition 3. *If $\lim_{\varepsilon \to 0, \alpha \to \infty} D_{KL}(p_r || p_o) = 0$, then MC-FLUE as a Bayesian Approximation.*

Proof. This is straightforward to demonstrate, as according to the conclusion of Proposition 2, we have:

$$\left| \frac{1}{S}\sum_{s=1}^{S} o^{(s)} - \frac{1}{S}\sum_{s=1}^{S} r^{(s)} \right| < \eta \tag{13}$$

where η is a small positive number. Furthermore, since MC-Dropout approximates Bayesian model averaging (Gal and Ghahramani 2016), i.e.:

$$\frac{1}{S}\sum_{s=1}^{S} r^{(s)} \approx \mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}\left[\mathbb{E}_{P(y_l|\boldsymbol{y}_{<l},\boldsymbol{\theta})}[y_l]\right] \tag{14}$$

Then MC-FLUE is also a Bayesian approximation:

$$\frac{1}{S}\sum_{s=1}^{S} o^{(s)} \approx \mathbb{E}_{P(\boldsymbol{\theta}|\mathcal{D})}\left[\mathbb{E}_{P(y|\boldsymbol{y}_{<l},\boldsymbol{\theta})}[y_l]\right] \tag{15}$$

See Appendix B for more details of this proof. □

The above proposition illustrates that by MC-FLUE, we can approximately obtain the token-level uncertainty in a single-pass inference.

3.3 Uncertainty Propagation from Token to Sequence

To derive sequence-level uncertainty from token-level uncertainties, we propose a post-processing model that satisfies single-pass inference requirements while accounting for temporal dependencies between tokens, which often represent semantic properties within the sequence. Our objective is to determine the most probable sequence of sequence-level uncertainty states given a series of token-level uncertainties.

Formally, let $u_t^{(1:T)} = \{u_t^{(1)}, u_t^{(2)}, ..., u_t^{(T)}\}$ denote the sequence of token-level uncertainties, where $u_t^{(i)}$ represents the uncertainty of the i-th token. We aim to find the optimal sequence of sequence-level uncertainty states $u_s^{(1:T)*} = \{u_s^{(1)*}, u_s^{(2)*}, ..., u_s^{(T)*}\}$ that maximizes the conditional probability:

$$u_s^{(1:T)*} = \arg\max_{u_s^{(1:T)}} P(u_s^{(1:T)} | u_t^{(1:T)}) \tag{16}$$

where $u_s^{(j)}$ denotes the sequence-level uncertainty state corresponding to the j-th token.

Employing a state transition model, we decompose the conditional probability as follows:

$$P(u_s^{(1:T)} | u_t^{(1:T)}) = \prod_{i=1}^{T} P(u_s^{(i)} | u_t^{(1:i)}, u_s^{(1:i-1)}) \tag{17}$$

Here, $P(u_s^{(i)} | u_t^{(1:i)}, u_s^{(1:i-1)})$ represents the probability of the i-th sequence-level uncertainty state, given all preceding token-level uncertainties and sequence-level states.

Implementing the state transition model is not confined to a specific architecture. However, it should possess nonlinear mapping capabilities and the ability to capture temporal dependencies among tokens in the sequence.

4 Evaluation

We evaluate the efficacy of FLUE through comprehensive comparisons with existing white-box and black-box methods, using various LLMs on open-ended QA tasks, assessing both token-level and sequence-level uncertainty.

Methods	Datasets	Models								Average
		LLaMA 2 7B	LLaMA 2 13B	LLaMA 2 70B	Mistral 7B	Vicuna 7B	Vicuna 13B	LLaMA 3 8B	LLaMA 3.1 8B	
Length Normalization	trivia	0.862	0.784	0.759	0.634	0.684	0.740	0.776	**0.850**	0.761
	svamp	0.537	0.538	0.613	0.550	0.588	0.587	0.661	0.813	0.611
	nq	**0.775**	0.771	**0.866**	0.642	0.764	0.754	0.779	**0.756**	0.763
	squad	0.732	0.746	0.725	0.595	0.714	0.655	0.644	0.729	0.693
	bioasq	0.665	0.514	0.581	0.587	0.611	0.551	0.680	0.699	0.611
Semantic Entropy	trivia	0.769	0.755	0.732	0.791	0.765	**0.790**	0.776	0.759	0.767
	svamp	**0.827**	**0.883**	**0.903**	**0.890**	**0.824**	**0.868**	**0.880**	**0.876**	**0.869**
	nq	0.730	0.758	0.757	0.776	0.716	0.778	0.746	0.748	0.751
	squad	0.793	0.734	0.744	0.706	0.766	0.777	0.754	**0.756**	0.754
	bioasq	**0.858**	0.833	0.869	**0.906**	**0.855**	**0.865**	0.865	**0.886**	**0.867**
P(True)	trivia	0.703	0.523	0.558	**0.829**	**0.865**	0.761	0.598	0.849	0.710
	svamp	0.462	0.610	0.513	0.728	0.524	0.526	0.656	0.608	0.578
	nq	0.706	0.476	0.559	0.762	0.712	0.725	0.583	0.750	0.659
	squad	0.624	0.498	0.582	0.684	0.708	0.623	0.641	0.640	0.625
	bioasq	0.663	0.578	0.743	0.735	0.754	0.830	0.584	0.548	0.679
FLUE (ours)	trivia	**0.885**	**0.841**	**0.881**	0.812	0.766	0.768	**0.854**	0.659	**0.808**
	svamp	0.612	0.745	0.775	0.869	0.810	0.842	0.761	0.742	0.770
	nq	0.680	**0.813**	0.769	**0.834**	**0.838**	**0.820**	**0.824**	0.618	**0.775**
	squad	**0.818**	**0.775**	**0.862**	**0.794**	**0.820**	**0.890**	**0.866**	0.579	**0.800**
	bioasq	0.692	**0.929**	**0.925**	0.845	0.772	0.842	**0.873**	0.578	0.807

Table 1: AUROC (⇑) evaluation for sequence-level uncertainty of white-box methods on five datasets across 8 LLMs. **Bold** and underlined values indicate best and second-best performance, respectively.

4.1 Experiment Setup

To validate the feasibility and efficacy of our proposed FLUE method, we conducted experiments under the following experimental settings.

Models. Following (Farquhar et al. 2024), we use LLaMA 2 (Touvron et al. 2023b) models(7B, 13B, 70B), Mistral 7B models (Jiang et al. 2023). Following (Xiong et al. 2024), we also evaluate our experiment on Vicuna (Zheng et al. 2023) models(Vicuan 7B v1.5 and Vicuna 13B v1.5). For diversity, we also test some recently released models, Mixtral 8x7B (Jiang et al. 2024), LLaMA 3 8B (Dubey et al. 2024), and LLaMA 3.1 8B (Dubey et al. 2024).

Datasets. The dataset used in this study aligns with (Farquhar et al. 2024), including SQuAD2.0 (Rajpurkar, Jia, and Liang 2018), SVAMP (Patel, Bhattamishra, and Goyal 2021), NQ-open (Lee, Chang, and Toutanova 2019), TriviaQA (Joshi et al. 2017), BioASQ (Krithara et al. 2023). SQuAD 2.0 is an updated version of the SQuAD dataset (Rajpurkar et al. 2016), which combines questions and contexts, and some questions do not have standard answers. SVAMP is a benchmark mathematical reasoning dataset containing elementary-level math problems described in natural language. NQ-open is a concise open domain question answering benchmark derived from Natural Questions (Kwiatkowski et al. 2019), based on consolidated Google Search queries. TriviaQA is a high-quality supervised reading comprehension question and answer dataset. BioASQ is a manually curated dataset for biomedical question answering, featuring diverse question types and corresponding answers derived from scientific literature and structured biomedical data.

Baselines. We compare our method with Length Normalization (Malinin and Gales 2021), Semantic Entropy (Farquhar et al. 2024; Kuhn, Gal, and Farquhar 2023), P(true) (Kadavath et al. 2022), LLM Self-Expression Uncertainty (Xiong et al. 2024) for sequence-level uncertainty. **Length Normalization(LN)** is an ensemble-based framework that provides both token-level and sequence-level uncertainty estimation in autoregressive structured prediction tasks. At the token level, it estimates total uncertainty and knowledge uncertainty through information-theoretic measures of the predictive posterior distribution. For sequence-level uncertainty, it uses Monte Carlo approximations to obtain asymptotically exact estimates during ensemble inference. Following (Kuhn, Gal, and Farquhar 2023), we implement LN using a single model with multinomial generation. **Semantic Entropy(SE)** generates multiple outputs and uses semantic invariance to cluster the generated set into equivalence classes, obtaining uncertainty estimates by calculating the entropy of the semantic distribution. **P(True)** is the probability a model assigns to the proposition that a specific sample is the correct answer to a question. (Kadavath et al. 2022) **Confidence Elicitation(CE)** leverages LLMs' inherent capabilities by using prompt strategy to encourage the models to express their confidence in their responses.

Performance Evaluation Metrics. Uncertainty estimation in LLMs should reflect the relationship between the response's confidence and accuracy. Our method does not modify the inference or decoding process, so accuracy is a method-agnostic metric. Following prior work (Farquhar

Methods	Models	Dataset									
		bioasq		nq		squad		svamp		trivia	
		AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE
Confidence Elicitation	LLaMA 2 7B	0.484	0.525	0.603	0.625	0.581	0.692	0.577	0.799	0.592	0.377
	LLaMA 2 13B	0.501	0.544	0.451	0.615	0.549	0.661	0.510	0.743	0.486	0.392
	Mistral 7B	0.580	0.537	0.487	0.743	0.597	0.668	0.450	0.699	0.676	0.571
	Vicuna 7B	0.528	0.606	0.560	0.784	0.559	0.740	0.688	0.880	0.568	0.596
	Vicuna 13B	0.516	0.552	0.561	0.662	0.488	0.701	0.527	0.867	0.468	0.457
FLUE (ours)	LLaMA 2 7B	0.592	0.340	0.580	0.384	0.648	0.378	0.612	0.414	0.685	0.387
	LLaMA 2 13B	0.629	0.282	0.613	0.299	0.675	0.312	0.645	0.349	0.641	0.421
	Mistral 7B	0.645	0.397	0.634	0.415	0.694	0.409	0.669	0.315	0.712	0.303
	Vicuna 7B	0.672	0.354	0.638	0.349	0.620	0.369	0.710	0.357	0.666	0.273
	Vicuna 13B	0.642	0.338	0.620	0.324	0.690	0.350	0.742	0.307	0.768	0.245

Table 2: Comparative Evaluation of AUROC (⇑) and ECE (⇓) in Black-box setting

et al. 2024; Lin, Trivedi, and Sun 2024), we adopted the Area Under Receiver Operating Characteristic (AUROC) as a metric for quantifying uncertainty, using it to measure the relationship between uncertainty and answer correctness in open-ended generation scenarios. Consistent with previous work (Xiong et al. 2024), we also use Expected Calibration Error(ECE) to quantify the calibration error between predicted probabilities and observed accuracy.

Choices of ε and layers selection in different LLMs. Figure 2 presents our experimental results for layer selection in FLUE sampling, with rectangular intervals representing parameters a and b from Equation (9).

The cosine similarity is an average of hidden layer states across multiple data points. We created a summary dataset by randomly sampling 100 points from each original dataset. This method can be generalized by a more comprehensive dataset.

4.2 FLUE Token Level Uncertainty Evaluation

Uncertainty for the LLMs can be roughly divided into sequence uncertainty and token uncertainty. Token uncertainty focuses on quantifying the model's uncertainty in predicting specific token positions, providing a real-time measure of uncertainty at the token level. LN approximates the predictive posterior's entropy using the MC entropy estimation chain rule.

To compare our proposed method with LN, we align the inference process of LLMs with LN. We calculate token uncertainty with the same token prefix at each token position. We use Pearson and Spearman correlation coefficients to evaluate the effectiveness of our methods. As shown in Figure 3, FLUE and LN exhibit a strong correlation, evidenced by high Pearson and Spearman coefficients, demonstrating that our method can achieve similar token uncertainty estimates to LN without the computational overhead of using multiple ensemble models and performing multiple forward passes.

4.3 FLUE Sequence Level Uncertainty Evaluation

In an open-ended natural language generation(NLG) scenario, accurately estimating sequence-level uncertainty is

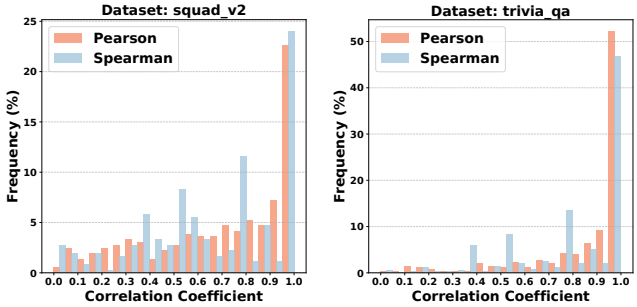


Figure 3: The Pearson and Spearman correlation coefficients of FLUE and length norm methods for the token level uncertainty. (a) squad dataset result. (b) trivia dataset result.

essential for evaluating the reliability of the model-generated text. We utilize a lightweight state transition model (Luo and Wang 2024) to assess sequence uncertainty based on token-level uncertainties computed during the inference process. LN leverages a length-normalized method to conquer the auto-regressive conditional independence assumption that distributions over long sequences can have higher entropy than short ones. SE clusters semantic Equivalence class of multiple generations to represent sequence uncertainty. P(true) needs to calculate the negative log-likelihood loss of LLM to give a specific sequence. We evaluated all these white box uncertainty estimation baselines and FLUE, as shown in Table 1; FLUE achieved the highest AUROC on TriviaQA, NQ-Open, and SQuAD v2.0, and competitive performance on the remaining two datasets across eight LLMs. Although both methods perform admirably, with SE achieving particularly high overall performance, their computational cost is prohibitive, especially for long sequence generation tasks. As for P(true), it performs well on specific models but demonstrates inconsistent results.

Aside from the white-box method that needs to acquire the logits or hidden states, we also compare our approach in the black-box setting. Confidence Elicitation is

Methods	Inference Time (↓)	AUROC(↑)
Semantic Entropy	1h16min	0.774
Length Normalization	1h47min	0.726
Confidence Elicitation	25min	0.537
FLUE(ours)	**15min**	**0.837**

Table 3: Temporal and Performance Evaluation of LLaMA 2 7B Uncertainty Estimation on SQuADv2.0

a prompt-based uncertainty estimation method that encourages LLMs to express their uncertainty through carefully crafted prompts; for detailed implementation specifics of the black box setup, please refer to Appendix E. As shown in Table 2, FLUE achieves higher AUROC and lower ECE overall by emulating black-box model states with a proxy model. CE necessitates well-designed prompts tailored to specific models. Small-scale LLMs typically exhibit weaker in-context learning capabilities than more powerful LLMs, like GPT-4 (Achiam et al. 2023) or Claude-3 (Anthropic 2024). This discrepancy in performance suggests that LLMs' self-expressed confidence is often poor, considering LLMs usually exhibit excessive confidence. (Xiong et al. 2024) Regarding efficiency, CE requires at least a single inference, considering that the effectiveness of prompts is not guaranteed. Moreover, CE demands longer supplementary prompts beyond the queries, increasing computational overhead.

We also explored the impact of selecting various numbers of layers and sampling quantities on uncertainty estimation performance for FLUE. Figure 4 shows that different models exhibit diverse sensitivities to these parameters. More layers and a larger sample size do not necessarily lead to absolute performance improvements.

4.4 Comprehensive Performance Evaluation

We conducted a comparative analysis of four uncertainty estimation methods, evaluating their inference time and AUROC. For the four methods of estimating uncertainty, for the methods that require multiple generations, we fixed the number of generations at 10. The sampling count for FLUE layers is also set to 10. In this setting, we tested the inference time of the Llama 2 7B model on the SQuAD dataset (after random sampling). Table 3 demonstrates that FLUE achieves competitive performance with minimal time. Theoretically, CE and FLUE incur equivalent inference costs, CE requires additional prompts and is susceptible to single-inference failures. For ensemble-based and multiple-sample methods, the time overhead increases significantly. For hardware parameters, please refer to Appendix E.

Compared to the SE method, our approach achieves 5x faster inference while demonstrating better performance in general domains (trivia, nq, squad) and competitive results in specialized domains (svamp, bioasq). Notably, SE's clustering process may be more advantageous in specialized domains, as exemplified by datasets like SVAMP and BioASQ.

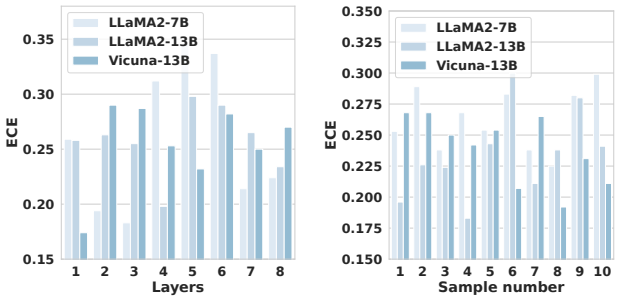


Figure 4: (a) Results of ECE with different selected layer counts. (b) Results of ECE with different sampling counts.

5 Related Works

Bayesian methods have been widely studied in machine learning, but the posterior distribution is often intractable in practice, necessitating the use of approximate methods. Variational inference (Graves 2011) has been a popular approach, with extensions such as Monte-Carlo Dropout (Gal and Ghahramani 2016) and Bayes-by-Backprop (Blundell et al. 2015) being applied to various models.

Recent work has focused on applying these techniques to more complex models and tasks. Zablotskaia et al. (2023) have benchmarked methods like Gaussian Process output layers for improving uncertainty estimation in summarization models. Malinin and Gales (2021) have further formalized information-theoretic measures of parameter uncertainty for structured prediction tasks and their estimators.

Recently, with the rapid development of pre-trained LLMs (Achiam et al. 2023; Touvron et al. 2023a), uncertainty estimation schemes based on ensembles have become increasingly unacceptable due to their expensive training overhead. (Xiong et al. 2024) empirically evaluated the self-awareness of LLMs regarding uncertainty. Another approach is to use semantic equivalence to provide uncertainty estimates (Farquhar et al. 2024; Manakul, Liusie, and Gales 2023), which, although widely recognized, sacrifices the possibility of compressing the sampling space. In this paper, starting from standard entropy estimation, we describe a more efficient sampling scheme and post-processing procedure.

6 Conclusion

This work introduces a single-pass uncertainty estimation method based on hidden state entropy, including the MC-FLUE sampling strategy and a state transition-based post-processing model. This approach captures sequence-level uncertainty without increasing inference costs. In open-ended question-answering tasks, FLUE demonstrated competitive uncertainty estimation performance across multiple open-source LLMs while providing an efficient uncertainty estimation for LLM applications in real-time environments. Future research directions may include further optimization of the sampling strategy and exploration of more complex post-processing models.

Acknowledgments

The corresponding author is Haoyi Zhou. Authors of this paper are supported by the National Natural Science Foundation of China through grants No.62225202, No.62202029, and Young Elite Scientists Spon-sorship Program by CAST (NO.2023QNRC001). We extend our sincere thanks to all authors for their valuable efforts and contributions.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Anthropic. 2024. Claude 3 Haiku: our fastest model yet. Technical report, Anthropic. Available at: https://www.anthropic.com/news/claude-3-haiku.

Balabanov, O.; and Linander, H. 2024. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. *arXiv preprint arXiv:2402.12264*.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, 1613–1622. PMLR.

Chuang, Y.-S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; and He, P. 2023. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Graves, A. 2011. Practical Variational Inference for Neural Networks. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F. C. N.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, 2348–2356.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Hanna, E. B.; Bressand, F.; et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1601–1611. Association for Computational Linguistics.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Krithara, A.; Nentidis, A.; Bougiatiotis, K.; and Paliouras, G. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1): 170.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A. P.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; Toutanova, K.; Jones, L.; Kelcey, M.; Chang, M.; Dai, A. M.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguistics*, 7: 452–466.

Lee, K.; Chang, M.; and Toutanova, K. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 6086–6096. Association for Computational Linguistics.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Trans. Mach. Learn. Res.*, 2024.

Luo, D.; and Wang, X. 2024. ModernTCN: A Modern Pure Convolution Structure for General Time Series Analysis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

MacKay, D. J. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Malinin, A.; and Gales, M. J. F. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Manakul, P.; Liusie, A.; and Gales, M. J. F. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Pro-*

cessing, EMNLP 2023, Singapore, December 6-10, 2023, 9004–9017. Association for Computational Linguistics.

Patel, A.; Bhattamishra, S.; and Goyal, N. 2021. Are NLP Models really able to Solve Simple Math Word Problems? In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tür, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, 2080–2094. Association for Computational Linguistics.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, 784–789. Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. Austin, Texas: Association for Computational Linguistics.

Sun, Q.; Pickett, M.; Nain, A. K.; and Jones, L. 2024. Transformer Layers as Painters. *arXiv preprint arXiv:2407.09298*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023a. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiao, Y.; and Wang, W. Y. 2019. Quantifying Uncertainties in Natural Language Processing Tasks. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 7322–7329. AAAI Press.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ye, F.; Yang, M.; Pang, J.; Wang, L.; Wong, D. F.; Yilmaz, E.; Shi, S.; and Tu, Z. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.

Zablotskaia, P.; Phan, D.; Maynez, J.; Narayan, S.; Ren, J.; and Liu, J. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. *arXiv preprint arXiv:2304.08653*.

Zheng, L.; Chiang, W.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.