PRICE: Preemptive Confidence Estimation in LLMs via Internal Representations

Anonymous ACL submission

Abstract

Large language models have demonstrated strong performance across a wide range of tasks, yet they are prone to generating incorrect information with high confidence. To mitigate this reliability issue, it is necessary to properly assess the model's confidence before generating responses. Existing methods either rely on post-hoc estimation based on generated tokens or sampling, or require additional training through supervised learning, presenting limitations in terms of cost and generalization. This paper proposes **PR**eemptive **I**nternal **C**onfidence **E**stimator (PRICE), a framework that estimates confidence before answer generation using attention and hidden states extractable from a single forward pass. PRICE combines signals that capture pattern agreement across attention heads and semantic clarity from hidden states. Furthermore, we theoretically derive optimal combination weights that minimize the Brier score under a simple error model, providing theoretical justification for the proposed model. Through extensive experiments across various models and benchmarks, PRICE demonstrates superior performance compared to existing methods. By preemptively estimating confidence without additional training, PRICE can contribute to achieving both efficiency in LLMs and safety in high risk domains.

1 Introduction

Large Language Models (LLMs) demonstrate exceptional performance across a wide range of natural language processing tasks, including question answering, summarization, and reasoning, and are expanding into diverse application domains. However, despite these performance improvements, LLMs often exhibit an overconfidence problem, generating responses with very high confidence based on incorrect information (OpenAI et al., 2024; Yang et al.). This can lead to critical errors in domains requiring high accuracy, such as medical and legal fields. Therefore, techniques for accurately interpreting and quantifying the confidence that models form regarding input queries are not merely performance evaluation metrics, but core components for ensuring safety in determining whether model responses can be trusted and utilized.

Research on model uncertainty and confidence estimation broadly falls into two approaches. First, post-hoc methods rely on answer generation or outputs, e.g., computing entropy from output token probabilities (Kadavath et al., 2022) or measuring consistency via multiple sampling (Wang et al., 2023; Kuhn et al.). However, they are only applicable after generation, limiting real time use, and sampling incurs substantial computational overhead due to repeated inference. Second, other work trains separate classifiers or probing models (Azaria and Mitchell, 2023; bur). While often accurate, these methods require additional labeled data and training, and their generalization can degrade under distribution shift to domains different from training data (Zhang et al., 2024).

To overcome these limitations, methods for estimating confidence before generating answers are required. Measuring query confidence before answer generation offers the following advantages. First, efficiency through adaptive computation per input data is possible. Instead of using identical resources for all inputs, fewer resources can be allocated to easy problems where the model is confident, while more inference resources can be dynamically allocated to uncertain difficult problems. Recently, OpenAI's ChatGPT (GPT-5) introduced a similar feature[1]. Such adaptive computation can improve overall computational efficiency while increasing answer reliability for difficult problems through additional inference. Second, risks can be managed proactively. By recognizing its uncer-

[1] https://openai.com/index/introducing-gpt-5/

tainty in advance, the model can prevent errors due to overconfidence by refusing to answer or alerting users instead of forcing answer generation. Finally, inference strategies can be dynamically adopted. When intrinsic instability is detected, optimal prompt strategies appropriate to the situation can be selected, such as selectively inducing Chain of Thought (Wei et al., 2022) or activating Retrieval Augmented Generation (Lewis et al., 2020). Thus, preemptive confidence estimation is a key technology that can improve the safety, efficiency, and reliability of LLMs.

Therefore, this study proposes the **PR**eemptive **I**nternal **C**onfidence **E**stimator (**`PRICE`**) methodology, which utilizes attention and hidden states internal information generated during the input interpretation process of transformer (Vaswani et al., 2017) based LLMs to assess reliability before answer generation. Previous research has suggested that the entropy of attention heads can be a significant indicator of correct answer prediction (Anonymous, 2025), or demonstrated that distributional characteristics of hidden states semantically correspond to the model's interpretation process (Zhang et al., 2025; Cencerrado et al., 2025). These internal signals have the advantage of being extractable with only a single forward pass. Therefore, `PRICE` measures the structural consistency of information flow between tokens through attention and measures the clarity of semantic representations formed through hidden states, interpreting and integrating the reasoning process of LLMs from two complementary perspectives. By combining two signals, `PRICE` offsets biases that a single indicator might have and constructs a robust confidence metric. Specifically, it combines a **consensus signal** that measures the agreement of multi head attention based on Jensen Shannon Divergence (JSD) with a **salience signal** that measures semantic clarity through coherence and sparsity of hidden states. This study demonstrates the effectiveness of `PRICE` through extensive experiments on various model architectures and benchmarks.

In summary, the contributions of this study are as follows:

- We propose `PRICE`, which estimates confidence for queries before answer generation by integrating the structural consistency of attention and semantic clarity of hidden states extracted in a single forward pass without additional training.
- When combining the two signals, we theoretically derive optimal combination weights that improve estimation performance under an error model, rather than relying on simple heuristics, thereby ensuring statistical justification for the methodology.
- Through extensive experiments on various LLM architectures and benchmarks, we demonstrate superior performance compared to existing methodologies and validate the effectiveness of the proposed method through visualization.

2 Related Work

Output-based Uncertainty Estimation Methods Black box approaches estimate confidence from generated text or token probability statistics such as entropy/perplexity (Heo et al., 2025; Penny-Dimri et al., 2025; Shorinwa et al., 2025), but these signals can be miscalibrated with respect to accuracy. To improve reliability, consistency based methods sample multiple outputs for the same input and measure agreement (Kuhn et al.; Wang et al., 2023; Manakul et al., 2023). A key limitation is that they require answer generation and repeated inference, increasing computational cost (Ma et al., 2025).

Supervised Analysis of Internal Representations Research has been conducted to train classifiers to detect truthfulness or knowledge under the hypothesis that hidden states inside models encode such information. Azaria and Mitchell (2023) trained supervised models such as logistic regression on intermediate layer activations. Snyder et al. (2024) trained classifiers to predict the likelihood of hallucination occurrence early based on internal signals at the first token generation time. Ji et al. (2024) developed probes that learn hidden states before LLM answer generation, demonstrating the ability to identify whether the model has seen training data and detect uncertainty. While these methods show high performance, they require separate data collection, labeling, and additional model training costs. Above all, issues of degraded generalization performance on out-of-distribution data continue to be raised.

3 Preliminary Analysis

We empirically analyzed how the internal information of LLMs differs according to the confidence

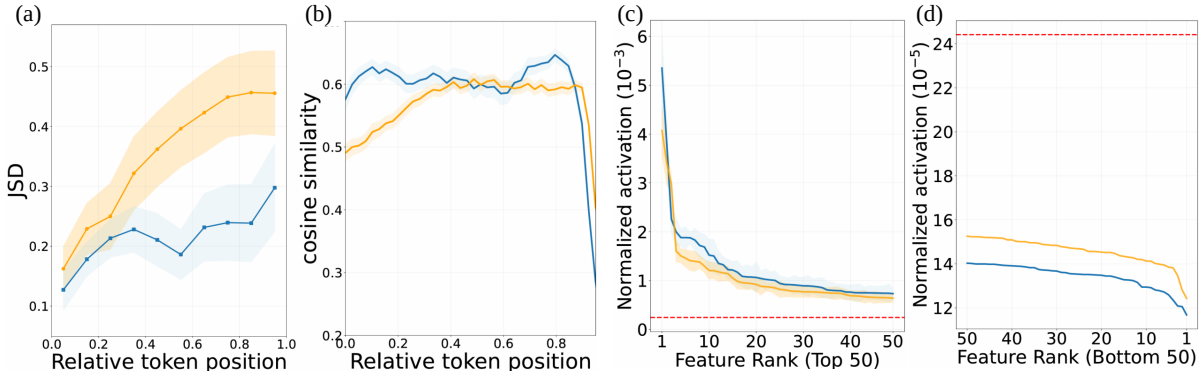

Figure 1: Comparison of internal signal characteristics observed for 500 samples each of consistently correct instances (blue line) and consistently incorrect instances (orange line). (a) Attention pattern dissensus between heads, (b) directional alignment of hidden states within the focus token span, (c) visualization of the 50 features with the largest absolute values in hidden states, (d) visualization of the 50 features with the smallest absolute values in hidden states. Normalized such that the sum of absolute values per feature equals 1, where the red dotted line indicates uniform distribution.

of answers. When solving a problem, LLMs sometimes answer correctly every time, but conversely, there are instances where they provide incorrect answers or give different answers. In this section, before introducing the proposed main methodology, we analyze attention patterns and hidden state representations by contrasting consistently correct and consistently incorrect instances. Specifically, each instance is evaluated independently 30 times. We construct two subsets: 500 instances that are correctly predicted in all 30 runs, and 500 instances that are incorrectly predicted in all 30 runs.

3.1 Extraction of Internal Signals

Attention Let the attention tensor of a transformer based LLM be $\mathbf{A} \in \mathbb{R}^{H \times T \times T}$ for each token position $t \in \{1, \ldots, T\}$ in the input sequence and head $h \in \{1, \ldots, H\}$. At each position t, the mixture distribution averaged over heads is defined as $\mathbf{m}_t = \frac{1}{H} \sum_{h=1}^{H} \mathbf{a}_{h,t}$.[2]

Hidden States According to prior research, due to the autoregressive nature of LLMs, the causal masking structure tends to aggregate information from the entire input in the latter part of the sequence, particularly in the last token (Chen et al., 2024; Azaria and Mitchell, 2023). However, analyzing only the hidden state of a single token risks being sensitive to noise or outliers at that position, so this study comprehensively analyzes the latter region. We denote the hidden state for each token in the input sequence as $\mathbf{h}_t \in \mathbb{R}^d$, and define the focus token span of the last k tokens in the sequence as $I_f = \{T-k+1, \ldots, T\},\ 1 < k < T-1$.

3.2 Consistency of Attention Patterns

Jensen Shannon Divergence To quantify how consistently each head in multi head attention attends to each token of the query, we measure using Jensen Shannon Divergence (JSD) (Lin, 2002).

$$\mathrm{JSD}_t = \mathcal{H}(\mathbf{m}_t) - \frac{1}{H} \sum_{h=1}^{H} \mathcal{H}(\mathbf{a}_{h,t}), \quad (1)$$

where $\mathcal{H}(\cdot)$ is the Shannon entropy. A lower JSD_t value indicates more similar patterns.

Observation Figure 1(a) shows the variation of JSD according to token position. Consistently correct instances generally exhibit low JSD, showing a tendency for different heads to attend to similar key tokens. In contrast, consistently incorrect instances show high JSD. This suggests that the internal evidence formation process in the model for consistently incorrect instance inputs is expressed as an unstable pattern across heads.

3.3 Geometry and Sparsity of Hidden States

Directional Alignment of Hidden States To examine how much hidden states within the focus region align in a single direction, we define the mean direction vector as follows: $\overline{\mathbf{h}} = \frac{1}{k} \sum_{t \in I_f} \mathbf{h}_t$. We calculate the cosine similarity between the computed mean direction vector $\overline{\mathbf{h}}$ and each token's hidden state $\mathbf{h}_t$. As can be seen in Figure 1(b), consistently correct instances generally show higher

[2] Since $\mathbf{a}_{h,t} := \mathbf{A}_h[t,:]$ is a vector that has passed through softmax, each component is non negative and sums to 1, which can be interpreted as a probability distribution.

similarity. The sharp drop at the end is interpreted as a boundary effect that occurs as the eos token takes on a functional role of concluding the sentence (Clark et al., 2019).

Contrastive Feature Activation To examine the distribution of each feature within hidden states, we normalize the hidden state $\mathbf{h}_t = (h_{t,1}, \cdots, h_{t,d}) \in \mathbb{R}^d$ into an absolute value based probability distribution as in Eq. 2.

$$\mathbf{p}_t = \frac{1}{\|\mathbf{h}_t\|_1}(|h_{t,1}|, \cdots, |h_{t,d}|) \in \mathbb{R}^d, \quad (2)$$

where $\|\mathbf{h}_t\|_1$ denotes the L1 norm. Figure 1(c) compares the top 50 activation values of each feature, and Figure 1(d) compares the bottom 50. Consistently correct instances show larger values in top features and lower activation values in bottom features. This means they reduce unnecessary noise and emphasize important features more. In contrast, consistently incorrect instances show a weaker tendency in this regard.

3.4 Discussion

In consistently correct instances, multiple heads show distribution patterns that attend to similar tokens, hidden states within the focus region are relatively more aligned with the mean direction, and the activation distribution of each feature within hidden states tends to be sparse. These observations suggest that the internal signals of the model can be structurally related to the confidence of correct answers. In the next section, we design a preemptive confidence measurement method by combining these signals.

4 PRICE

4.1 Overview of PRICE Framework

As illustrated in Figure 2, PRICE integrates two complementary signals. First, the consensus signal ($S_{\texttt{con}}$) quantifies the dissensus of token attention patterns across heads in multi head attention. Second, the salience signal ($S_{\texttt{sal}}$) quantifies semantic clarity by measuring coherence and sparsity from hidden states. The formulation is shown in Eq. (3).

$$S_{\texttt{PRICE}} = \alpha \cdot S_{\texttt{con}} + (1-\alpha) \cdot S_{\texttt{sal}}, \quad (3)$$

where $\alpha \in [0, 1]$ is a hyperparameter that adjusts the contribution between the two signals. This combination mitigates biases that a single signal might have. For example, even if attention is stable ($S_{\texttt{con}}$ high), the internal representation may be ambiguous ($S_{\texttt{sal}}$ low), and conversely, even if specific neurons are strongly activated ($S_{\texttt{sal}}$ high), if agreement across heads has not been achieved ($S_{\texttt{con}}$ low), it is difficult to trust. Therefore, Eq. (3) is designed to have the highest value when both conditions are satisfied.

4.2 $S_{\texttt{con}}$: Consensus Signal

We calculate the degree of dissensus JSD_t of attention patterns across heads defined in Eq. (1) as an average over all tokens: $\text{JSD}_{\text{all}} = \frac{1}{T}\sum_{t=1}^{T} \text{JSD}_t$. Since JSD for H distributions lies in the range $0 \leq \text{JSD}_t \leq \log H$,[3] we normalize by dividing by $\log H$ to have a value between 0 and 1. We define the consensus signal as in Eq. (4) by subtracting the calculated value from 1 so that a signal value closer to 1 indicates higher agreement:

$$S_{\texttt{con}} = 1 - \frac{\text{JSD}_{\text{all}}}{\log H}. \quad (4)$$

This definition is not simply a heuristic of similarity/distance, but is precisely interpreted as the mutual information between head identity and observed key tokens. We prove this through Theorem 2 in Appendix A. Additionally, while Eq. (4) treats all heads equally, a method for weighting heads individually is described in Appendix B.

4.3 $S_{\texttt{sal}}$: Salience Signal

$S_{\texttt{con}}$ is an indicator that measures agreement across attention heads, but does not directly guarantee the informativeness of attention itself (Jain and Wallace, 2019). For example, if all heads output attention close to a uniform distribution, $S_{\texttt{con}}$ can be high due to $\text{JSD}_t \approx 0$, but this may indicate that no evidence has been formed. To complement this limitation, we aim to quantify and use whether sparse patterns are shown in the most important segment for correct answer generation from hidden states. Therefore, in this study, we interpret the consensus signal as an internal instability signal, but complement such cases by subsequently combining it with the salience signal.

Coherence The normalized directional coherence within the focus token span is as follows:

$$C_{I_f} = \frac{\|\sum_{t \in I_f} \mathbf{h}_t\|_2}{\sum_{t \in I_f} \|\mathbf{h}_t\|_2}. \quad (5)$$

[3] See Proposition 1 in Appendix A.

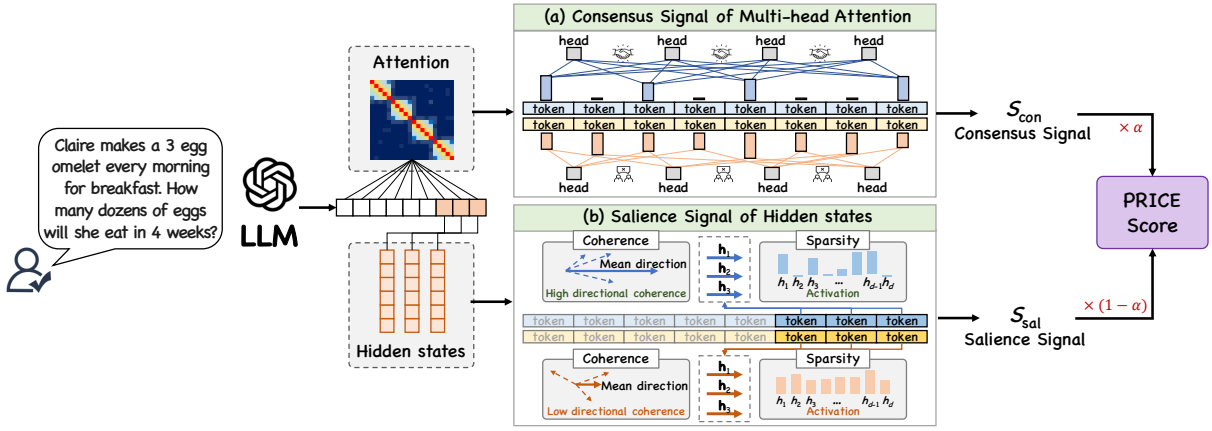

Figure 2: Overview of the PRICE framework for estimating preemptive confidence. (a) Consensus signal is calculated by measuring agreement across heads in multi head attention. Consistently correct instances focus on similar tokens across heads. In contrast, consistently incorrect instances tend to attend to disparate tokens. (b) Salience signal is calculated by measuring coherence and sparsity from hidden states in the focus token span. Consistently correct instances show high directional alignment and sparse activation values. Conversely, consistently incorrect instances show low directional alignment and lack sparse patterns. The final confidence is estimated by a weighted sum of the two signals.

The denominator of C_{I_f} is the sum of individual vector L2 norms, and the numerator is the L2 norm of the vector sum. If vectors point in similar directions, the activation value for each feature of the sum vector will be large, and if they point in different directions, it will be small. Therefore, the value of C_{I_f} is proportional to the degree of directional alignment of the vectors. Using the formula for C_{I_f}, the norm of the mean vector of hidden states can be calculated as follows:

$$\|\overline{\mathbf{h}}\|_2 = \|\sum_{t\in I_f}\mathbf{h}_t\|_2 = C_{I_f}\cdot\frac{1}{k}\sum_{t\in I_f}\|\mathbf{h}_t\|_2. \quad (6)$$

That is, $\|\overline{\mathbf{h}}\|_2$ is a signal that reflects both the degree of alignment in one direction (C_{I_f}) and the average activation ($\frac{1}{k}\sum_{t\in I_f}\|\mathbf{h}_t\|_2$) of token hidden states within the focus token span. Based on this, we use it as the final coherence measurement as in Eq. (7).

$$\text{Coherence} = \phi(\|\overline{\mathbf{h}}\|_2), \quad (7)$$

where the monotonic increasing function $\phi : \mathbb{R} \to [0, 1]$ is a function that normalizes coherence.

Sparsity To measure sparsity, we use the deviation from the uniform distribution $(\frac{1}{d}, \cdots, \frac{1}{d}) \in \mathbb{R}^d$ using the probability distribution $\mathbf{p}_t = (p_{t,1}, \cdots, p_{t,d})$ defined in Eq. (2) (See Eq. (8)).

$$G(\mathbf{p}_t) = \frac{1}{2(1-\frac{1}{d})}\sum_{j=1}^{d}\left|p_{t,j} - \frac{1}{d}\right|. \quad (8)$$

Here, $G(\mathbf{p}_t)$ represents the deviation and is denoted as G inspired by the Gini coefficient. Additionally, the reason for dividing by $2(1-\frac{1}{d})$ is to normalize the value to a maximum of 1. This is because when $G(\mathbf{p}_t)$ is at its maximum value, it is in the sparsest state, which is when $\mathbf{p}_t$ is a vector with only one dimension equal to 1 and all others equal to 0. In this case, $\sum_{j=1}^{d}|p_{t,j} - \frac{1}{d}| = 2(1-\frac{1}{d})$. We calculate Eq. (8) for all hidden states in the focus token span and use it as the final sparsity measurement value.

Salience Signal Using the previously measured indicators, we define and use the salience signal as follows:

$$S_{\texttt{sal}} = \text{Coherence} \times \text{Sparsity}. \quad (9)$$

This is combined in multiplicative form to give a high score only when both coherence and sparsity conditions are simultaneously strong.

5 Experimental Settings

5.1 Datasets

We used four publicly available benchmark datasets spanning diverse domains and difficulty levels: MMLU (Hendrycks et al., 2021) (split into MMLU-STEM and MMLU-Social), GSM8K (Cobbe et al., 2021), ARC-Challenge (Clark et al., 2018), and GPQA (Rein et al., 2023). All experiments were conducted in a zero-shot setting.

5.2 Models

We used four LLM models with different structures and training characteristics. Specifically, we selected `Mistral-7B-v0.3` (Jiang et al., 2023), `Qwen2.5-7B` (Qwen et al., 2025), `Llama-3.1-8B` (Grattafiori et al., 2024), and `Gemma-2-9B` (Team et al., 2024) models. The temperature of all LLMs was fixed at 0.7. All experiments were conducted on eight NVIDIA RTX3090 GPUs.

5.3 Baselines

All baselines take only the problem as input (no answer generation) and output a confidence score. We compare: (i) a prompt based self reported confidence baseline (averaged over three runs), (ii) a perplexity based baseline computed from the model's conditional likelihood of the problem text, and (iii) internal signal baselines, Hidden Score and Attention Score (Sriramanan et al., 2024). Hidden Score summarizes hidden state activation and distribution characteristics, while Attention Score quantifies stability of the attention pattern across heads.

5.4 Evaluation Metrics

To evaluate confidence estimation performance, we used Expected Calibration Error (ECE) (Guo et al., 2017) and Brier score (Glenn et al., 1950). ECE is a calibration metric that measures the discrepancy between model output confidence and actual accuracy, with lower values indicating better alignment. The actual accuracy is calculated as the accuracy rate when the LLM solves each problem in the benchmark 30 times. Next, the Brier score is a metric that measures the accuracy of probabilistic predictions, calculated based on the squared error between predicted probabilities and actual outcomes. Unlike ECE, which examines the average alignment between confidence and accuracy, the Brier score evaluates at the individual sample level how high confidence the model has for correct answers and how low confidence it has for incorrect answers.

5.5 Implementation Details

Representation Extraction Both attention and hidden states are extracted from only a single layer at the 75% point of the LLM. For `Mistral` and `Llama` models with 32 layers, they are extracted from the 24th layer, and for `Qwen` and `Gemma` models with 28 and 42 layers, they are extracted from the 21st and 30th layers, respectively. This is based on prior research (Son et al., 2025) showing that mid to late layers contain richer information than the last layer.

Salience Signal When calculating the salience signal, we use the $\tanh$ function as the nonlinear function, and set the focus token span region as the last 30% tokens of the input sequence.

`PRICE` `PRICE` is determined by a weighted sum of $S_{\texttt{con}}$ and $S_{\texttt{sal}}$ using weight α. When calculating `PRICE`, α is a hyperparameter that can be set differently for each data characteristic, requiring hyperparameter search. However, in this study, we calculate and use a more accurate value using a small amount of data with the following theorem.

Theorem 1. *Consider $S_{\texttt{con}}$ and $S_{\texttt{sal}}$ as unbiased estimators. That is, $S_{\texttt{con}} = p + \epsilon_1$, $S_{\texttt{sal}} = p + \epsilon_2$, $\mathbb{E}[\epsilon_k|p] = 0$, where p is the model's (theoretical) true confidence. The variance of each error is $\sigma_1^2 = \mathbb{E}[\epsilon_1^2]$, $\sigma_2^2 = \mathbb{E}[\epsilon_2^2]$, and the covariance is $C = \mathbb{E}[\epsilon_1\epsilon_2]$. Then, the value α^* that minimizes the Brier Score $\mathrm{Brier}(\alpha) = \mathbb{E}[(S_\alpha - Y)^2]$ for $S_{\texttt{PRICE}}$ is as follows:*

$$\alpha^* = \arg\min_{\alpha} \mathrm{Brier}(\alpha) = \frac{\sigma_2^2 - C}{\sigma_1^2 + \sigma_2^2 - 2C} \quad (10)$$

According to this theorem, if the distributions of $S_{\texttt{con}}$ and $S_{\texttt{sal}}$ exist, the optimal α value (i.e., one that minimizes the Brier score) can be calculated. In this study, we randomly sampled 20% of the entire data to obtain $S_{\texttt{con}}$ and $S_{\texttt{sal}}$, and conducted subsequent experiments on the remaining data. The proof of this theorem is described in detail in proof A of Appendix A.

6 Experimental Results and Analysis

6.1 Main Results

Table 1 shows the confidence estimation performance on four LLMs and five benchmark datasets. The experimental results demonstrate that the proposed `PRICE` surpasses existing methods in most cases, achieving the lowest ECE and Brier score. This suggests that `PRICE` accurately captures model uncertainty before answer generation. In contrast, using prompts to have the model output its confidence shows generally high ECE and Brier score values. This can be seen as supporting previous research findings (Tian et al., 2023; Groot and Valdenegro Toro, 2024) that LLMs tend to be overconfident when expressing their confidence in text.

Model	Method	MMLU-STEM		MMLU-Social		GSM8K		ARC-Challenge		GPQA	
		ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓	ECE ↓	Brier ↓
Mistral7B	Prompt	0.3741	0.3073	0.3924	0.3303	0.3150	0.2044	0.4109	0.3864	0.3520	0.2942
	Perplexity	0.2943	0.2678	0.2435	0.2521	0.3758	0.2475	0.1504	0.0984	0.4965	0.4268
	Hidden Score	0.3234	0.2552	0.2927	0.2585	**0.1116**	0.1097	**0.0938**	**0.0973**	0.5852	0.4840
	Attention Score	0.2490	0.2260	0.2282	0.2361	0.1485	0.1112	0.1221	0.1128	0.3100	0.2568
	PRICE (ours)	**0.0476**	**0.1526**	**0.1027**	**0.1809**	0.1225	**0.1035**	0.1031	0.0998	**0.1285**	**0.1601**
Qwen7B	Prompt	0.1821	0.2599	0.1476	0.2351	0.4077	0.3303	0.2089	0.1603	0.3393	0.2982
	Perplexity	0.1922	0.2295	0.4686	0.2257	0.4645	0.3709	0.1908	0.1053	0.4749	0.4115
	Hidden Score	0.1253	0.2079	0.1042	0.2011	0.2329	0.2198	0.1583	0.1071	0.5297	0.4570
	Attention Score	0.0713	0.2009	0.1110	0.2115	0.2004	0.2047	0.1378	0.0754	0.3625	0.3181
	PRICE (ours)	**0.0430**	**0.1872**	**0.0574**	**0.1818**	**0.1431**	**0.1768**	**0.1189**	**0.0727**	**0.1174**	**0.1778**
Llama8B	Prompt	0.1956	0.2161	0.1293	0.1848	0.4309	0.3538	0.1223	0.1723	0.3595	0.2461
	Perplexity	0.1912	0.2174	0.3650	0.1887	0.5908	0.3911	0.2210	0.1265	0.4466	0.3169
	Hidden Score	0.1576	0.1751	0.1147	**0.1706**	0.2068	0.2166	0.1688	0.1700	0.5848	0.4310
	Attention Score	0.1894	0.2110	0.1371	0.1870	0.2112	0.2270	0.2175	0.1266	0.3593	0.2371
	PRICE (ours)	**0.0359**	**0.1532**	**0.0897**	0.1809	**0.1543**	**0.1074**	**0.1108**	**0.0882**	**0.1594**	**0.1149**
Gemma9B	Prompt	0.1355	0.2552	0.2776	0.2360	0.2349	0.2131	0.2288	0.1191	0.3365	0.3668
	Perplexity	0.1727	0.2360	0.1243	0.2120	0.1570	0.1821	0.1504	0.0987	0.4571	0.3773
	Hidden Score	0.1498	0.2155	0.1159	0.2043	0.1381	0.1726	0.1323	**0.0722**	0.5382	0.4592
	Attention Score	0.0831	0.2036	0.0975	0.1944	**0.0992**	0.1668	0.1891	0.0910	0.4196	0.3566
	PRICE (ours)	**0.0404**	**0.1932**	**0.0786**	**0.1884**	0.1004	**0.1581**	**0.1224**	0.1479	**0.0902**	**0.1844**

Table 1: Main experimental results on confidence estimation performance. We compared the performance of PRICE with baselines (Prompt, Perplexity, Hidden Score, Attention Score) on four LLMs and five benchmark datasets. Evaluation used ECE and Brier Score, with lower values indicating better performance. The best results are marked in **bold**, and the second-best results are underlined. The proposed methodology PRICE shows superior results in most cases.

The perplexity based method shows reasonable performance on the ARC-Challenge dataset but exhibits overall instability, with particularly sharp performance degradation on the high difficulty GPQA dataset. Hidden Score and Attention Score, which use only single internal signals, show overall better performance than prompt or perplexity methods, but with large performance variance. In conclusion, PRICE maximizes overall estimation performance by appropriately combining these two complementary signals, compensating for uncertainty information that a single signal might miss.

6.2 Ablation Studies

6.2.1 Complementarity of Consensus and Salience Signals

Figure 3 visualizes the correlation and distribution between S_{con} and S_{sal} to verify the complementarity of the two proposed signals. The Pearson correlation coefficient between the two metrics is -0.11, suggesting that each score captures uncertainty from independent perspectives rather than redundant information. Particularly in the four quadrant analysis based on the mean value of consistency score (0.73, blue line) and the mean value of activation score (0.36, red line), numerous samples with relatively high accuracy rates were observed even in regions where the two signals' judgments conflict (Q2 and Q4). This shows that individual metrics alone may be insufficient to fully capture the prediction uncertainty of the model, and that considering both signals together can complement these limitations. In other words, this weighted sum approach provides empirical evidence that enables more robust confidence estimation by offsetting biases inherent in individual metrics.

6.2.2 Sensitivity to the Balancing Parameter α

We measured ECE and Brier score by adjusting the weight α required for PRICE calculation from 0 to 1 at 0.05 intervals. As can be seen in Figure 4, all models show that the optimal α records lower error at intermediate values than at the extreme values. This demonstrates that the two sig-

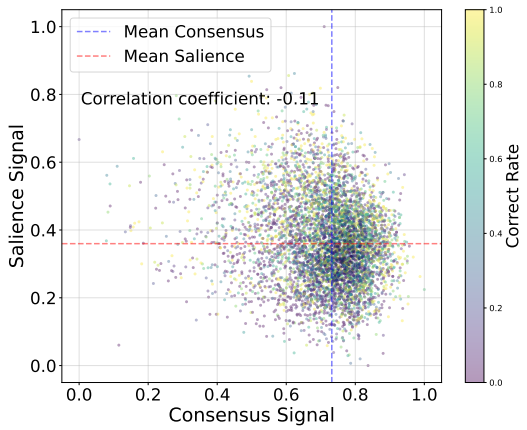

Figure 3: Scatter plot of consensus signal and salience signal. Each point is colored by accuracy rate. The two signals show a weak correlation coefficient (-0.11), indicating that they capture independent information from each other.

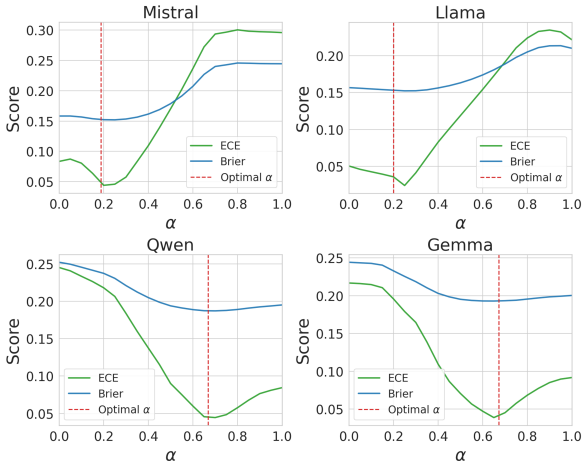

Figure 4: Performance analysis according to the weight α value of PRICE. ECE (green line) and Brier (blue line) were measured by varying α from 0 to 1 at 0.05 intervals. The red dotted line represents the theoretical optimal α value calculated from Theorem 1.

nals work complementarily when combined. We also compared the theoretical optimal value (red dash line) calculated through Theorem 1 with experimentally derived results. As a result, the theoretically derived α^* value shows values very close to the optimal point confirmed through experiments across all models. That is, using Theorem 1, the optimal weight can be found reflecting model specific characteristics. In conclusion, the proposed approach demonstrates that optimal hyperparameters can be effectively estimated without costly search processes (e.g., grid search).

6.2.3 Effect of Focus Token Span Length

We analyzed the impact of focus token span length on performance when calculating $S_{\texttt{sal}}$. Experiments were conducted by varying the ratio of latter part tokens relative to the entire sequence to 0.1, 0.3 (proposed), 0.5, and 1.0, with results shown in Figure 5. For benchmarks like MMLU (left chart), the lowest ECE was recorded when set to 0.3, the default setting of PRICE. A narrow region (0.1) is vulnerable to local noise, and the entire region (1.0) dilutes the confidence signal as unnecessary contextual information is mixed in. On the other hand, for the GPQA dataset (right chart), which requires graduate level high difficulty reasoning, a tendency emerged where viewing a slightly wider range (0.5) showed better performance. This can be interpreted as the model going through a more complex reasoning process to reach conclusions for higher difficulty problems, with clues distributed over a wider range.

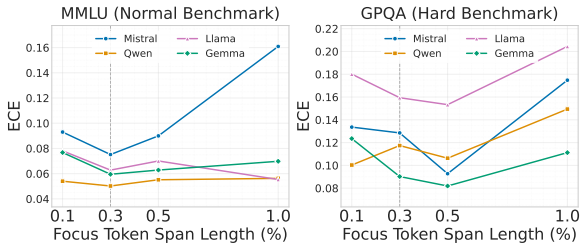

Figure 5: Performance analysis according to focus token span length. The x-axis represents the ratio of how much of the latter part of the input sequence to view. For MMLU, 0.3 shows the lowest error, but for high difficulty tasks like GPQA, a wider region (0.5) tends to be advantageous.

7 Conclusion

In this study, we proposed PRICE, a framework that can measure the confidence of queries before LLMs generate answers. PRICE integrates two signals: the structural consensus of attention patterns and the semantic clarity of hidden states. Through extensive experiments on various benchmarks and models, PRICE demonstrated superior performance by recording lower ECE and Brier scores compared to existing methods. In particular, by confirming that the theoretically derived optimal weight α^* aligns with actual experimental results, we showed that the proposed methodology stands on solid statistical foundations rather than being a simple heuristic. PRICE can overcome the limitations of computational cost and generalization inherent in existing post-hoc approaches or supervised learning methods, and is a methodology that enables achieving both efficiency and safety in LLM systems.

Limitations

This study has several limitations. First, PRICE requires direct access to the model's weights and internal representations. While it can be applied to open source models such as Llama and Mistral, it cannot be applied to closed source models provided in API form. Second, while PRICE can warn of low confidence by capturing the phenomenon of unstable internal signals in the model, it is difficult to specifically identify the reason for low confidence. For example, it is difficult to distinguish whether low confidence is due to the absence of training data or the ambiguity of the input prompt. Third, to measure the structural consistency of the attention mechanism, we analyze attention matrices of size $T \times T$. As the length of the input sequence increases, the cost of attention computation and analysis increases significantly. There is a risk of increased memory usage and computational cost in long context situations.

References

Anonymous. 2025. Attention head entropy of LLMs predicts answer correctness. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Iván Vicente Moreno Cencerrado, Arnau Padrés Masdemont, Anton Gonzalvez Hawthorne, David Demitri Africa, and Lorenzo Pacchiardi. 2025. No answer needed: Predicting llm answer accuracy from question-only linear probes. *Preprint*, arXiv:2509.10625.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

W Brier Glenn and 1 others. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. 2025. Do LLMs estimate uncertainty well in instruction-following? In *The Thirteenth International Conference on Learning Representations*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *CoRR*.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *NeurIPS ML Safety Workshop*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Huan Ma, Jingdong Chen, Guangyu Wang, and Changqing Zhang. 2025. Estimating llm uncertainty with logits. *arXiv e-prints*, pages arXiv–2502.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Frank Nielsen. 2019. On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jahan C Penny-Dimri, Magdalena Bachmann, William R Cooke, Sam Mathewlynn, Samuel Dockree, John Tolladay, Jannik Kossen, Lin Li, Yarin Gal, and Gabriel Davis Jones. 2025. Measuring large language model uncertainty in women's health using semantic entropy and perplexity: a comparative study. *The Lancet Obstetrics, Gynaecology, & Women's Health*, 1(1):e47–e56.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Ben Snyder, Marius Moisescu, and Muhammad Bilal Zafar. 2024. On early detection of hallucinations in factual question answering. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2721–2732.

Minjoo Son, Jonghak Jang, and Misuk Kim. 2025. Lightweight query checkpoint: Classifying faulty user queries to mitigate hallucinations in large language model question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14664–14677, Vienna, Austria. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. On verbalized confidence scores for llms. In *ICLR Workshop: Quantify Uncertainty and Hallucination in Foundation Models: The Next Frontier in Reliable AI*.

Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning models know when they're right: Probing hidden states for self-verification. In *Second Conference on Language Modeling*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8908–8949.

A Theoretical Details

Proposition 1. JSD_t *defined in Eq. 1 satisfies the following.*

$$0 \leq \mathrm{JSD}_t \leq \log H \tag{11}$$

Equality holds in the following cases.

- $\mathrm{JSD}_t = 0$ *when* $\mathbf{a}_{1,t} = \cdots = \mathbf{a}_{H,t}$
- $\mathrm{JSD}_t = \log H$ *when each head attends to completely different tokens*

Proof. First, we prove $0 \leq \mathrm{JSD}_t$. JSD_t can be expressed in the form of KL divergence (Nielsen, 2019).

$$\mathrm{JSD}_t = \frac{1}{H}\sum_{h=1}^{H}\mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t)$$
$$\text{where } \mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t) = \sum_k \mathbf{a}_{h,t}\log\frac{\mathbf{a}_{h,t}}{\mathbf{m}_t}$$

Here, k is the index of the key token being attended to. Since KL divergence is always greater than 0 by definition, $\forall h,\ \mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t) \geq 0$. Therefore, $\mathrm{JSD}_t \geq 0$. In particular, if $\mathbf{a}_{1,t} = \cdots = \mathbf{a}_{H,t}$, then $\mathbf{m}_t = \mathbf{a}_{1,t}$, so $\forall h,\ \mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t) = 0$, and $\mathrm{JSD}_t = 0$.

Next, we prove $\mathrm{JSD}_t \leq \log H$. For each k, the following holds.

$$\mathbf{m}_t(k) = \frac{1}{H}\sum_{h=1}^{H}\mathbf{a}_{h,t}(k) \geq \frac{1}{H}\mathbf{a}_{h,t}(k) \tag{12}$$

Rearranging the above inequality for terms where $\mathbf{a}_{h,t}(k) > 0$ gives the following.

$$\frac{\mathbf{a}_{h,t}(k)}{\mathbf{m}_t(k)} \leq H \ \rightarrow\ \log\frac{\mathbf{a}_{h,t}(k)}{\mathbf{m}_t(k)} \leq \log H$$

Therefore, for each head h, the following holds.

$$\begin{aligned}\mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t) &= \sum_k \mathbf{a}_{h,t}\log\frac{\mathbf{a}_{h,t}}{\mathbf{m}_t}\\ &\leq \sum_k \mathbf{a}_{h,t}\log H\\ &= \log H\sum_k \mathbf{a}_{h,t}\\ &= \log H\end{aligned}$$

Therefore, the following holds.

$$\begin{aligned}\mathrm{JSD}_t &= \frac{1}{H}\sum_{h=1}^{H}\mathrm{KL}(\mathbf{a}_{h,t}\|\mathbf{m}_t)\\ &\leq \frac{1}{H}\sum_{h=1}^{H}\log H\\ &= \log H\end{aligned}$$

Equality holds when $\mathbf{m}_t(k) = \mathbf{a}_{h,t}(k)/H$ for each h, k. This means that if some head h attends to some token position k, the remaining heads assign probability 0 to that position k. □

Theorem 1. *Consider* $S_{\texttt{con}}$ *and* $S_{\texttt{sal}}$ *as unbiased estimators. That is,* $S_{\texttt{con}} = p + \epsilon_1$, $S_{\texttt{sal}} = p + \epsilon_2$, $\mathbb{E}[\epsilon_k|p] = 0$, *where* p *is the (theoretical) true confidence of the model. The variance of each error is* $\sigma_1^2 = \mathbb{E}[\epsilon_1^2]$, $\sigma_2^2 = \mathbb{E}[\epsilon_2^2]$, *and the covariance is* $C = \mathbb{E}[\epsilon_1\epsilon_2]$. *Then, the value* α^* *that minimizes the Brier Score* $\mathrm{Brier}(\alpha) = \mathbb{E}[(S_\alpha - Y)^2]$ *for* $S_{\texttt{PRICE}}$ *is as follows.*

$$\alpha^* = \arg\min_\alpha \mathrm{Brier}(\alpha) = \frac{\sigma_2^2 - C}{\sigma_1^2 + \sigma_2^2 - 2C}$$

Proof. For convenience, let $S_{\texttt{con}} = S_1$, $S_{\texttt{sal}} = S_2$ and $S_{\texttt{PRICE}} = S_\alpha$.

$$\begin{aligned}S_\alpha &= \alpha\cdot S_1 + (1-\alpha)\cdot S_2\\ &= \alpha(p+\epsilon_1) + (1-\alpha)(p+\epsilon_2)\\ &= p + \epsilon_2 + \alpha(\epsilon_1 - \epsilon_2)\end{aligned}$$

Therefore, we can write as follows.

$$S_\alpha - p = \epsilon_2 + \alpha(\epsilon_1 - \epsilon_2) \tag{13}$$

Next, we decompose the Brier score into bias/variance terms.

$$\begin{aligned}\mathrm{BS}(\alpha) &= \mathbb{E}\big[(S_\alpha - Y)^2\big]\\ &= \mathbb{E}\big[\{(S_\alpha - p) + (p - Y)\}^2\big]\\ &= \mathbb{E}\big[(S_\alpha - p)^2\big] + \mathbb{E}\big[(p-Y)^2\big]\\ &\qquad + 2\mathbb{E}\big[(S_\alpha - p)(p - Y)\big]\end{aligned}$$

Here, since $Y|p$ follows $\mathrm{Bernoulli}(p)$, the following holds.

$$Y = p + \eta,\ \mathbb{E}[\eta|p] = 0,\ \mathbb{E}[\eta^2|p] = p(1-p)$$

Therefore, $\mathbb{E}\big[(S_\alpha - p)(p - Y)\big] = \mathbb{E}[(S_\alpha - p)\eta]$, and by definition, if ϵ_1, ϵ_2 are independent of or

uncorrelated with p, η, it equals 0. Therefore, the Brier score is as follows.

$$\mathrm{BS}(\alpha) = \mathbb{E}[(S_\alpha - p)^2] + \underbrace{\mathbb{E}[(Y-p)^2]}_{\text{constant independent of } \alpha}$$

Here, the term $\mathbb{E}[(Y-p)^2]$ is a constant independent of α. Therefore, minimizing the Brier can be reduced to the problem of minimizing $\mathbb{E}[(S_\alpha - p)^2]$.

Let us simplify Eq. 13.

$$\begin{aligned}\mathbb{E}[(S_\alpha - p)^2] &= \mathbb{E}[(\epsilon_2 + \alpha(\epsilon_1 - \epsilon_2))^2] \\ &= \sigma_2^2 + 2\alpha\mathbb{E}[\epsilon_2(\epsilon_1 - \epsilon_2)] \\ &\quad + \alpha^2\mathbb{E}[(\epsilon_1 - \epsilon_2)^2]\end{aligned}$$

Here, the following holds.

$$\begin{aligned}\mathbb{E}[\epsilon_2(\epsilon_1 - \epsilon_2)] &= C - \sigma_2^2 \\ \mathbb{E}[(\epsilon_1 - \epsilon_2)^2] &= \sigma_1^2 + \sigma_2^2 - 2C\end{aligned}$$

Therefore, simplifying $\mathbb{E}[(S_\alpha - p)^2]$ gives the following.

$$f(\alpha) = \alpha^2(\sigma_1^2 + \sigma_2^2 - 2C) + 2\alpha(C - \sigma_2^2) + \sigma_2^2$$

This expression is a quadratic equation in α.

$$f'(\alpha) = 2\alpha(\sigma_1^2 + \sigma_2^2 - 2C) + 2(C - \sigma_2^2)$$

Therefore, the value α^* satisfying $f'(\alpha) = 0$ is as follows.

$$\alpha^* = \frac{\sigma_2^2 - C}{\sigma_1^2 + \sigma_2^2 - 2C}$$

Thus, the theorem is proved. □

Theorem 2. *At query position t, define random variables (Z, K) as follows: $Z \sim \mathrm{Unif}\{1, \cdots, H\}$, $K \mid (Z = h) \sim \mathbf{a}_{h,t}$. Then the following holds.*

$$\mathrm{JSD}_t = I(Z; K) \tag{14}$$

Proof. Before the proof, let us redefine several concepts.

Shannon entropy When the distribution of a discrete random variable X is $p(x)$, entropy is defined as follows.

$$\mathcal{H}(X) = -\sum_x p(x) \log p(x) \tag{15}$$

Conditional entropy For two random variables X, Y, conditional entropy is defined as follows.

$$\mathcal{H}(Y \mid X) = \sum_x p(x)\mathcal{H}(Y \mid X = x) \tag{16}$$

Mutual information Mutual information $I(X; Y)$ quantifies how much the uncertainty about Y decreases when X is known, and is defined as follows.

$$I(X; Y) = \mathcal{H}(Y) - \mathcal{H}(Y \mid X) \tag{17}$$

That is, if $I(X; Y) = 0$, then X and Y are independent, and the larger the value, the more strongly they depend on each other.

Consider the random variables (Z, K) in this lemma. By definition, Z is the head index indicating which head was selected, and K is the token index indicating which key position that head selected.

First, for an arbitrary key position $k \in \{1, \cdots, T\}$, applying the law of total probability gives the following.

$$\begin{aligned}P(K = k) &= \sum_{h=1}^{H} P(Z = h)P(K = k \mid Z = h) \\ &= \sum_{h=1}^{H} \frac{1}{H} a_{h,t}(k)\end{aligned}$$

Therefore, the distribution of K is identical to

$$\mathbf{m}_t(k) = \frac{1}{H}\sum_{h=1}^{H} a_{h,k}(k).$$

That is, the marginal distribution of K is $\mathbf{m}_t$, and the following holds.

$$\mathcal{H}(K) = \mathcal{H}(\mathbf{m}_t) \tag{18}$$

Next, we compute the conditional entropy $\mathcal{H}(K \mid Z)$. Since Z follows a uniform distribution, by the definition of conditional entropy, the following holds.

$$\begin{aligned}\mathcal{H}(K \mid Z) &= \sum_{h=1}^{H} P(Z = h)\mathcal{H}(K \mid Z = h) \\ &= \frac{1}{H}\sum_{h=1}^{H} \mathcal{H}(K \mid Z = h)\end{aligned}$$

Here, since $K \mid (Z = h) \sim \mathbf{a}_{h,t}$, we have $\mathcal{H}(K \mid Z = h) = \mathcal{H}(\mathbf{a}_{h,t})$. Therefore, the following holds.

$$\mathcal{H}(K \mid Z) = \frac{1}{H}\sum_{h=1}^{H} \mathcal{H}(\mathbf{a}_{h,t}) \tag{19}$$

Substituting Eq. 18 and Eq. 19 into the definition of mutual information gives the following.

$$I(Z;K) = \mathcal{H}(K) - \mathcal{H}(K \mid Z)$$
$$= \mathcal{H}(\mathbf{m}_t) - \frac{1}{H}\sum_{h=1}^{H}\mathcal{H}(\mathbf{a}_{h,t})$$

This expression exactly matches the definition of JSD_t defined in Eq. 1. □

In other words, JSD_t represents the amount of information that indicates "how well we can distinguish which head (Z) generated a given key token K when observing it once." If the heads form identical distributions, Z and K become nearly independent, resulting in $I(Z;K) \approx 0$ (high consensus), whereas if each head focuses on different sets of tokens, K carries significant information about Z, leading to an increase in $I(Z;K)$ (large disagreement).

Corollary 1. *The following holds.*

$$\mathcal{H}(K \mid Q) = \mathcal{H}(K \mid H, Q) + I(H; K \mid Q)$$

Therefore, the Consensus Signal can be written as follows.

$$S_{\text{cons}} = 1 - \frac{1}{\log H}\{\mathcal{H}(T|Q) - \mathcal{H}(T|H,Q)\}$$

In the above corollary, each term can be interpreted as follows. $\mathcal{H}(K \mid Q)$ represents the total uncertainty about which key token the model attends to when given query position Q, averaged without distinguishing attention heads. On the other hand, $\mathcal{H}(K \mid H, Q)$ is the conditional uncertainty when the head index is also observed, corresponding to the average entropy of individual attention heads. Therefore, the difference between these two, $I(H; K \mid Q)$, is a quantity that represents how much the uncertainty about key tokens decreases by additionally observing the attention head information. In other words, the $S_{\texttt{con}}$ can be interpreted as a normalized measure of how effectively the internal consensus of the attention structure reduces uncertainty.

B Analysis on head importance weighting

The multi head attention mechanism in transformers is designed to enable the model to simultaneously learn information from different positions and various representational subspaces (Vaswani et al., 2017). Accordingly, each head tends to specialize in different domains, such as capturing syntactic structures or semantic relationships (Clark et al., 2019). However, prior research has pointed out that only a small number of heads make decisive contributions to the inference process, while the remaining heads contain redundant or unimportant information and can be removed (Voita et al., 2019; Michel et al., 2019).

This perspective raises the hypothesis that treating all heads equally when estimating confidence may not be optimal. There is a possibility that selectively focusing on key heads that substantially contribute to predictions could improve the accuracy of confidence estimation. To verify this hypothesis, we conducted additional comparative experiments by introducing weights that reflect the actual contribution of each head.

$$w_{h,t} = \frac{\|\mathbf{o}_{h,t}\|_2}{\sum_{j=1}^{H}\|\mathbf{o}_{j,t}\|_2}, \quad \text{where} \sum_{h=1}^{H} w_{h,t} = 1 \tag{20}$$

In the transformer architecture, the degree to which a specific head h contributes to the residual stream is proportional not to the attention probability values themselves, but to the magnitude of the output vector computed as a result (Kobayashi et al., 2020). Let $\mathbf{o}_{h,t} \in \mathbb{R}^d$ denote the value output vector of head h. We define the importance $w_{h,t}$ of that head at time step t as the ratio of L2 norms as follows.

$$w_{h,t} = \frac{\|\mathbf{o}_{h,t}\|_2}{\sum_{j=1}^{H}\|\mathbf{o}_{j,t}\|_2}, \quad \text{where} \sum_{h=1}^{H} w_{h,t} = 1 \tag{21}$$

This weighting scheme naturally excludes the influence of unnecessary heads that do not substantially contribute to the model's predictions.

Table 2 shows the impact of weighting based on attention head importance on confidence estimation performance. As in the main experiments, we evaluated four models across five benchmarks, and the results showed that the baseline approach (No weight) achieved lower ECE and Brier scores than the head weight approach, demonstrating overall superior performance. In some cases (e.g., `GPQA` for `Mistral`, `GSM8K` and `ARC` for `Qwen`), the head weight setting showed marginally better performance. Particularly for the challenging `GPQA` benchmark, the `Mistral` and `Llama` models performed better with the head weight setting. This suggests

Model	Method	MMLU-STEM		MMLU-Social		GSM8K		ARC-Challenge		GPQA	
		ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier
Mistral7B	Head weight	0.121	0.1724	0.1138	0.1843	0.1438	0.1225	0.1033	0.1022	**0.097**	**0.1446**
	No weight	**0.0476**	**0.1526**	**0.1027**	**0.1809**	**0.1225**	**0.1035**	**0.1031**	**0.0998**	0.1285	0.1601
Qwen7B	Head weight	0.0909	0.2039	0.0662	0.189	**0.1313**	**0.1492**	**0.1058**	**0.0700**	0.1525	0.1789
	No weight	**0.043**	**0.1872**	**0.0574**	**0.1818**	0.1431	0.1768	0.1189	0.0727	**0.1174**	**0.1778**
Llama8B	Head weight	0.0673	0.1692	0.1031	0.1778	**0.1429**	**0.1011**	0.1253	0.1007	**0.1293**	**0.1103**
	No weight	**0.0359**	**0.1532**	**0.0897**	**0.1809**	0.1543	0.1074	**0.1108**	**0.0882**	0.1594	0.1149
Gemma9B	Head weight	0.1195	0.2093	0.2753	0.2593	0.1325	0.1763	0.1398	0.1533	0.1525	0.1869
	No weight	**0.0404**	**0.1932**	**0.0786**	**0.1884**	**0.1004**	**0.1571**	**0.1224**	**0.1479**	**0.0902**	**0.1844**

Table 2: The following are the experimental results comparing the conventional method (no weighting) with a method that assigns weights to attention heads based on their importance. In the majority of scenarios, the unweighted approach demonstrated superior performance. Consequently, we adopted a configuration that assigns equal weights to all heads as our primary setting.

Dataset	Count	Source
MMLU-STEM	5556	https://huggingface.co/datasets/cais/mmlu
MMLU-Social	8590	
GSM8K	1320	https://huggingface.co/datasets/openai/gsm8k
ARC-Challenge	1120	https://huggingface.co/datasets/allenai/ai2_arc
GPQA	448	https://huggingface.co/datasets/Idavidrein/gpqa

Table 3: Dataset statistics

that specific heads may play important roles in complex reasoning processes, but in general situations, uniformly considering all heads proves to provide more robust performance.

These results suggest that for confidence estimation, it is important to comprehensively consider all heads, not just a few key heads. While previous studies assigned importance assuming that specific heads dominate information processing, in the case of confidence estimation, signals from heads with relatively low importance also serve as useful cues. In other words, applying weights risks suppressing some subtle signals, which can hinder overall confidence estimation. Therefore, our PRICE framework adopts an approach that computes the average of JSD across all heads without separate weight calculation. This has lower computational complexity than head weighting and guarantees stable performance across various settings.

However, for tasks like GPQA or specific architectures such as Llama and Qwen, focusing on key heads may be advantageous, so dynamically adjusting weights according to the task could be a future research direction.

C Dataset

Table 3 summarizes the statistics of the datasets used in this study.