VERBALIZED CONFIDENCE CALIBRATION FOR RETRIEVAL-AUGMENTED LARGE LANGUAGE MODELS VIA MULTI-SOURCE EVIDENCE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly deployed in retrieval-augmented generation (RAG) settings, yet they frequently verbalize high confidence when their answers are incorrect. Existing methods for calibrating verbalized confidence typically rely on token-level generation probabilities or heuristic confidence scores, and use the correctness of the predicted answer as supervision, without explicitly modeling the agreement and conflict across the model's parametric knowledge and the retrieved documents. We propose a two-stage framework for verbalized confidence calibration in retrieval-augmented multiple-choice question answering. In the first stage, information from the model's parametric knowledge and from each retrieved document is encoded into source-specific evidence vectors, each of which is used to parameterize a Dirichlet distribution over answer options. Using self-attention, a fusion module captures agreement and conflict among these sources and produces a fused Dirichlet distribution. The expected value of the probability assigned to the predicted answer under this distribution serves as a scalar confidence target. In the second stage, a lightweight confidence estimator takes the question-conditioned answer-option logits as input and learns to predict this scalar target without requiring additional document retrieval during confidence estimation. The predicted scalar confidence is then converted into a Gaussian-smoothed soft target distribution over verbalized confidence levels, and KL divergence regularization is employed to align the model's verbalized confidence distribution with this soft target. Experiments on multiple-choice question-answering benchmarks show that our method improves confidence calibration on the in-distribution dataset and on most out-of-distribution datasets, and, overall, yields more reliable uncertainty estimates than baseline methods.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable success, but hallucination (Huang et al., 2025; Li et al., 2024) remains a major bottleneck for their deployment in real-world applications. Hallucination often manifests as the model confidently generating content about which it is in fact uncertain. Therefore, quantifying model uncertainty and enabling the model to recognize its own uncertainty are key steps toward mitigating hallucination(Geng et al., 2024; Liu et al., 2025).

Several strategies have been proposed to study uncertainty in LLMs. The first relies on output token probabilities, with higher token probabilities indicating lower uncertainty (Jiang et al., 2021; Kadavath et al., 2022). However, token probabilities reflect the predictive probability of generating the next token, rather than the correctness of the response as a whole. The second evaluates the uncertainty of generated responses using metrics such as perplexity or semantic entropy (Farquhar et al., 2024), but these metrics remain implicit in the model's output and are difficult to present directly to users. The third approach is verbalized confidence (Lin et al., 2022a), which refers to the confidence directly expressed by the model through prompting. Users can obtain such confidence statements through simple instructions, making this approach more practical for real-world applications. However, Xiong et al. (2024) found that LLMs such as GPT-4 generally exhibit overconfidence when verbalizing confidence. They proposed a training-free calibration framework that calibrates verbalized confidence by sampling multiple responses. Although this method alleviates overconfidence

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to some extent, its reliance on sampling cannot guarantee a reliable uncertainty estimate for each individual response. Moreover, sampling multiple LLM responses is impractical in specialized domains such as healthcare, where repeated generation incurs prohibitively high computational costs and latency, and unreliable outputs can have serious consequences.

In retrieval-augmented generation (RAG) settings, retrieved documents may be noisy, incomplete, or conflicting, introducing uncertainty from the retrieval side. Meanwhile, the model's parametric knowledge may be insufficient to answer the question reliably, resulting in epistemic uncertainty on the model side. We argue that these two sources of uncertainty, if properly measured and fused, can provide as a principled signal for calibrating the model's verbalized confidence. However, this direction remains largely unexplored in existing research. Therefore, a multi-source evidence fusion framework is proposed for verbalized confidence calibration. Our main contributions are as follows:

- We identify and formalize confidence estimation in RAG as a multi-source evidence fusion problem, where the LLM's parametric knowledge and each retrieved document are treated as separate evidence sources and mapped to Dirichlet distributions that jointly model class probabilities over answer options and epistemic uncertainty.
- We propose an attention-based evidence fusion (AEF) module that produces a fused evidence representation by integrating the LLM's parametric knowledge and retrieved documents. It explicitly captures agreement and conflict among evidence sources, providing a reliable confidence signal for verbalized confidence calibration.
- We introduce a decoupled two-stage calibration framework that avoids the circular dependency of precomputed correctness labels: AEF first estimates confidence from the current LLMs logits, and the resulting soft targets are then distilled into the LLMs verbalized confidence.
- Experiments on multiple-choice question-answering benchmarks show that our method improves confidence calibration on in-distribution and most out-of-distribution datasets, and yields more reliable verbalized uncertainty expressions than strong baselines.

2 Related Works

Traditional approaches to calibrating verbalized confidence can be broadly divided into black-box and white-box methods. Black-box methods do not access internal model signals; instead, they calibrate verbalized confidence through multiple sampled responses or prompt engineering (Xiong et al., 2024; Zong et al., 2025). White-box methods, in contrast, leverage internal model signals and typically fine-tune LLM parameters to adjust the confidence distribution over confidence tokens (Lin et al., 2022a; Xu et al., 2024; Li et al., 2026). Within these categories, some methods require human-annotated confidence data (Lin et al., 2022a), while others elicit confidence based on one or a few inference results via prompting (Xiong et al., 2024; Zong et al., 2025; Xu et al., 2024; Li et al., 2026). However, these methods often either incur high manual annotation costs or fail to capture the model's epistemic uncertainty. Among white-box methods, Li et al. (2026) propose a Brier score-based loss to calibrate verbalized confidence over confidence tokens. Yet this loss is supervised by the pre-update model's binary prediction correctness (100% for correct, 0% for incorrect). Such a binary correctness signal cannot distinguish a confident correct answer from a correct answer obtained by chance, and it provides no estimate of epistemic uncertainty. Moreover, after parameter updates, correctness may change, causing a moving-target problem. This motivates us to estimate epistemic uncertainty from multi-source evidence and use the dynamically estimated confidence as supervision during fine-tuning.

A key challenge in leveraging RAG for confidence calibration is to capture the uncertainty arising from both the retrieved information and the language model itself. Recent work has begun to address this dual-source uncertainty. For example, an axiomatic framework (Soudani et al., 2025) explicitly models retrieval-side and model-side uncertainty, providing a useful foundation for jointly quantifying them. Knowledge partitioning methods (Sun et al., 2025) also touch upon this dual-source uncertainty—retrieval uncertainty and generative uncertainty—by delineating the retrieval knowledge boundary and the model's own knowledge boundary. While these methods address dual-source uncertainty, they primarily aim to improve answer accuracy or uncertainty quantification (Rafailov et al., 2023), rather than explicitly calibrating the model's verbalized confidence. In contrast, our

work focuses on verbalized confidence calibration and extends the dual-source view to a multi-source setting by treating the LLM's parametric knowledge and each retrieved document as distinct evidence sources. To this end, we build on evidential deep learning (Sensoy et al., 2018) and subjective logic (Jsang, 2018), deriving evidence for each source from LLM logits and mapping it to a Dirichlet distribution. This is supported by recent work showing that LLM logits can be interpreted as evidence for Dirichlet-based uncertainty decomposition (Li et al., 2025; Xu et al., 2026).

3 PRELIMINARY

For an LLM f_θ, consider a multiple-choice question-answering instance (x, y), where x is a text sequence containing the question and the set of answer options $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$, and $y \in \{1, 2, \ldots, K\}$ is the index of the correct answer. After feeding x to the model f_θ, the model generates a predicted answer index $\hat{y}$ and a numerical verbalized confidence $\hat{c} \in [0, 1]$, denoted as $\langle \hat{y}, \hat{c} \rangle = f_\theta(x)$. Since the LLM expresses confidence over a discrete set of levels $\mathcal{V}_{\text{conf}} = \{0, \frac{1}{N}, \ldots, 1\}$, we define calibration over these levels. On a multiple-choice question-answering task $T = \{(x, y)\}$, if the model is perfectly calibrated, then for every confidence level $t \in \mathcal{V}_{\text{conf}}$, its conditional accuracy equals t given that the verbalized confidence equals t. Formally,

$$\mathbb{E}_{(x,y)\sim T}\left[\mathbb{I}\left(\hat{y} = y\right) | \, \hat{c} = t\right] = t, \quad t \in \mathcal{V}_{\text{conf}}, \tag{1}$$

where $\mathbb{I}(\hat{y} = y) = 1$ if $\hat{y} = y$, and 0 otherwise.

To calibrate verbalized confidence, we first obtain a confidence estimate via the model's epistemic state, and then use this estimate to guide the LLM toward producing more accurate verbalized confidence. Traditional confidence estimation methods, such as maximum softmax probability (MSP), provide only point estimates of the predicted answer's confidence and cannot characterize the model's epistemic uncertainty.

To this end, we draw on the evidential deep learning framework to transform model logits into non-negative evidence and parameterize a Dirichlet distribution over the probability vector of answer options. The expected value of the probability assigned to each answer option under this distribution serves as the confidence estimate, while the distribution explicitly quantifies epistemic uncertainty.

Let f_θ denote the entire model and f_θ^{logit} denote its logit-producing submodule. Given the input sequence x, the model first produces a K-dimensional logits vector $\boldsymbol{z} = f_\theta^{\text{logit}}(x) = (z_1, z_2, \ldots, z_K)^\top \in \mathbb{R}^K$, whose entries correspond to the K answer options. The predicted answer is obtained from $\boldsymbol{z}$, i.e., $\hat{y} = \arg\max_k z_k$. Meanwhile, we define an evidence vector $\boldsymbol{e} = (e_1, e_2, \ldots, e_K)^\top \in \mathbb{R}^K_{\geq 0}$, where e_k is the evidence assigned by the model f_θ to the answer option A_k given the input sequence x. Specifically, the evidence vector is obtained by applying a non-negative activation function elementwise, such as softplus: $\boldsymbol{e} = \text{softplus}(\boldsymbol{z})$, i.e., $e_k = \text{softplus}(z_k)$.

Unlike the point estimate produced by MSP, we treat the class probability vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)^\top$ over answer options as a random vector and model it with a Dirichlet distribution whose concentration parameter vector is $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K) = \boldsymbol{e} + \mathbf{1}$, in which $\mathbf{1} = (1, 1, \ldots, 1)^\top$. Thus, $\boldsymbol{p} \sim \text{Dir}(\boldsymbol{\alpha})$. Its probability density function is formulated as

$$\text{Dir}(\boldsymbol{p}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}, \tag{2}$$

where $\boldsymbol{p}$ lies in the probability simplex in $\mathbb{R}^K$, i.e., $\Delta^{K-1} = \{\boldsymbol{p} \mid \sum_{k=1}^K p_k = 1, \; p_k \geq 0\}$, and $\Gamma(\phi) = \int_0^{+\infty} t^{\phi-1} e^{-t} dt$. When $\boldsymbol{e} = \mathbf{0}$, the Dirichlet distribution reduces to the uniform distribution $\text{Dir}(\mathbf{1})$, in which case the model's preference among all answer options is completely flat, and the epistemic uncertainty

$$u = \frac{K}{S}, S = \sum_{k=1}^K \alpha_k,$$

reaches its maximum value of 1.

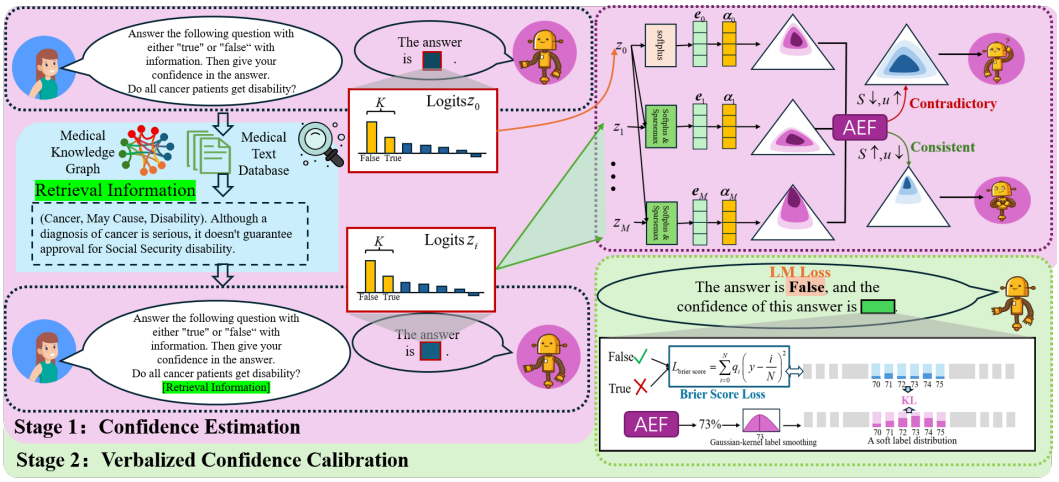


Figure 1: Overview of the proposed multi-source evidence fusion framework for verbalized confidence calibration.

4 Verbalized Confidence Calibration via Multi-Source Evidence Fusion

To estimate the confidence of LLMs, we adopt the evidential deep learning loss to learn a probability distribution over answer options that serves as the confidence estimate. For a training example (x, y), let the one-hot vector corresponding to the true label y be $\boldsymbol{v} = (v_1, v_2, \ldots, v_K)^\top$, where $v_k \in \{0, 1\}$ and $\sum_{k=1}^{K} v_k = 1$. The loss function combines a cross-entropy term with a KL regularization term:

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{v}) = \sum_{k=1}^{K} v_k(\psi(S) - \psi(\alpha_k)) + \lambda_{kl}\mathrm{KL}[\mathrm{Dir}(\tilde{\boldsymbol{\alpha}})||\mathrm{Dir}(\mathbf{1})], \tag{3}$$

where the digamma function $\psi(\cdot)$ is the logarithmic derivative of the gamma function, $\psi(u) = \frac{\Gamma'(u)}{\Gamma(u)}$, and $\tilde{\boldsymbol{\alpha}} = \boldsymbol{v} + (\mathbf{1} - \boldsymbol{v}) \odot \boldsymbol{\alpha}$, $\odot$ denotes element-wise multiplication. The cross-entropy loss term is equivalent to the expected negative log-likelihood of the true answer. When the model makes an incorrect prediction, the KL term suppresses evidence for all non-target answer options. This suppression reduces the total evidence and makes the output distribution more uniform, indicating higher epistemic uncertainty.

4.1 Retrieval-Augmented Evidence Fusion Module

LLMs often produce overconfident answers, and their parametric knowledge may be incomplete or outdated. A single forward pass only reflects the model's parametric knowledge and cannot account for external evidence. To address this, we introduce a retrieval-augmented generation setting, where retrieved documents serve as additional evidence sources. The model's belief over answer options is then jointly determined by multiple sources: if they agree on the same answer, the fused evidence increases, reflecting stronger belief and lower uncertainty; if they conflict, the fused evidence decreases, reflecting contradictory belief and higher uncertainty. We therefore design an evidence fusion module to learn these consistency and conflict relationships, more faithfully reflecting the model's internal knowledge and estimating confidence more reliably. Figure 1 illustrates the overall framework.

Given a question x, we employ the ToG-2 retrieval framework (Ma et al., 2025) to retrieve M relevant documents $\{d_1, d_2, \ldots, d_M\}$ from an external knowledge base (see Appendix A for details). Each document d_m is concatenated with the question x, and the concatenated sequence is fed separately into the LLM to obtain its logits vector $\mathbf{z}_m \in \mathbb{R}^K$, $m = 1, 2, \ldots, M$, over the K answer options. Meanwhile, without any retrieved document, the model outputs the no-retrieval logits vector $\mathbf{z}_0 \in \mathbb{R}^K$. Each evidence source is mapped to a non-negative evidence vector via the softplus

function, and these vectors are stacked into an evidence matrix $\mathbf{E}$. Appendix B.1 provides the detailed construction.

To integrate these sources, the attention-based evidence fusion (AEF) module uses a self-attention layer to model agreement and conflict among them. Further details of the AEF are provided in Appendix B.2. Let $\boldsymbol{\alpha}_{\text{fused}} = \text{AEF}(\mathbf{E}) = (\alpha_{1,\text{fused}}, \alpha_{2,\text{fused}}, \ldots, \alpha_{K,\text{fused}})$ denote the concentration parameters of the resulting fused Dirichlet distribution. Based on Eq. (3), the corresponding loss function is defined as:

$$\mathcal{L}(\boldsymbol{\alpha}_{\text{fused}}, \boldsymbol{v}) = \sum_{k=1}^{K} v_k(\psi(S_{\text{fused}}) - \psi(\alpha_{k,\text{fused}})) + \lambda_{kl}\text{KL}[\text{Dir}(\tilde{\boldsymbol{\alpha}}_{\text{fused}})||\text{Dir}(\mathbf{1})], \tag{4}$$

where $S_{\text{fused}} = \sum_{k=1}^{K} \alpha_{k,\text{fused}}$, $\tilde{\boldsymbol{\alpha}}_{\text{fused}} = \boldsymbol{v} + (\mathbf{1} - \boldsymbol{v}) \odot \boldsymbol{\alpha}_{\text{fused}}$, and $\boldsymbol{v}$ is the one-hot vector of the ground-truth label. After training, the parameters of the AEF module are frozen, and the module is attached to the model's answer-option logits to serve as a confidence estimator.

4.2 Masked Retrieval Strategy for AEF Module

Performing confidence estimation at inference time with M retrieved documents requires an additional forward pass for each document, together with a no-retrieval pass, resulting in $M + 1$ passes in total. However, the confidence estimation module should also remain reliable when the retrieved documents are unavailable or uninformative. To address this issue, we introduce a masked-retrieval strategy during training, which randomly simulates the absence of useful retrieved information.

Specifically, for each sample in a mini-batch, we independently apply masking with probability p_{mask}. When the mask is activated, the logits associated with all retrieved documents are replaced by the no-retrieval logits: $\mathbf{z}_m \leftarrow \mathbf{z}_0$, $m = 1, 2, \ldots, M$. Consequently, for the m-th retrieved document, the incremental logits become a zero vector: $\Delta\mathbf{z}_m = \mathbf{z}_m - \mathbf{z}_0 = \mathbf{0}$. Under our evidence-mapping scheme, the corresponding document evidence vector is $\mathbf{e}_m = \log(2) \cdot \mathbf{1}$, which assigns uniform, non-discriminative evidence to all answer options. With this masked training strategy, the AEF module can output reasonable confidence estimates at inference time even when retrieved-document information is unavailable or uninformative.

During calibration, no retrieved documents are provided for a question x. We therefore employ the AEF module under the same masked-retrieval configuration: the no-retrieval branch receives $\mathbf{z}_0$, while all retrieved-document branches are fixed to their masked state with zero incremental logits.

4.3 Dynamic Verbalized Confidence Calibration

In this subsection, we dynamically estimate the confidence of each sample using the trained confidence estimator and employ the estimate as a soft target to guide the LLM toward better-calibrated verbalized confidence via distillation.

4.3.1 Verbalized Confidence Calibration

Different models use different tokenizers: for example, Llama tokenizes "10" as a single token, whereas Qwen splits it into two tokens, "1" and "0". To make verbalized confidence calibration applicable across different models, we adopt the Brier-score-based calibration strategy proposed by Li et al. (2026), which enables the LLM to express confidence on a discrete scale from 0 to N.

Rather than precomputing predicted answers, predictions are recomputed at each training step. For each question $x^{(i)}$, we feed it into the LLM to obtain the no-retrieval logits vector $\mathbf{z}_0^{(i)}$ over K answer options, yielding the predicted answer index $\hat{y}^{(i)} = \arg\max_k z_{0,k}^{(i)}$, and then construct the prompt template shown in Appendix G. During training, the logits $\mathbf{w}^{(i)} \in \mathbb{R}^{|\mathcal{V}|}$ of the first token after "Confidence:" are extracted, and those corresponding to the predefined confidence token set $\mathcal{T}_N = \{0, 1, \ldots, N\}$ are selected, denoted as $w_0^{(i)}, w_1^{(i)}, \ldots, w_N^{(i)}$. The model's probability distribution over these confidence tokens is then computed as:

$$q_j^{(i)} = \frac{\exp(w_j^{(i)})}{\sum_{t=0}^{N} \exp(w_t^{(i)})}, \quad j = 0, 1, \ldots, N, \tag{5}$$

where $q_j^{(i)}$ denotes the probability that the model assigns to the j-th confidence token for sample i. Using the correctness label $y_{\text{corr}}^{(i)} = \mathbb{I}[\hat{y}^{(i)} = y^{(i)}]$ as a hard target, we adopt the tokenized Brier score to calibrate verbalized confidence (Li et al., 2026):

$$\mathcal{L}_{\text{Brier}} = \sum_{i=1}^{B} \sum_{j=0}^{N} q_j^{(i)} \left(y_{\text{corr}}^{(i)} - \frac{j}{N} \right)^2, \tag{6}$$

where the hard correctness label $y_{\text{corr}}^{(i)} \in \{0, 1\}$ provides only binary supervision.

A KL distillation term is introduced to provide richer supervision, aligning the confidence distribution with a soft target derived from the AEF module. Given the predicted answer index $\hat{y}^{(i)}$, the confidence estimate $\hat{s}_{\text{aef}}^{(i)} \in [0, 1]$ is computed as the Dirichlet expectation at the predicted answer:

$$\hat{s}_{\text{aef}}^{(i)} = \mathbb{E}[p_{\hat{y}^{(i)}}] = \frac{\alpha_{\hat{y}^{(i)},\text{fused}}^{(i)}}{S_{\text{fused}}^{(i)}}, S_{\text{fused}}^{(i)} = \sum_{k=1}^{K} \alpha_{k,\text{fused}}^{(i)}, \tag{7}$$

where $\boldsymbol{\alpha}_{\text{fused}}^{(i)} = (\alpha_{1,\text{fused}}^{(i)}, \alpha_{2,\text{fused}}^{(i)}, \ldots, \alpha_{K,\text{fused}}^{(i)})$ are the fused Dirichlet parameters output by the AEF module.

To obtain a soft target over confidence tokens, we project $\hat{s}_{\text{aef}}^{(i)}$ onto the discrete confidence-token scale and apply a Gaussian kernel, producing $\boldsymbol{\pi}^{(i)} = (\pi_0^{(i)}, \pi_1^{(i)}, \ldots, \pi_N^{(i)})$:

$$\pi_j^{(i)} = \frac{\exp\left(-\dfrac{(\hat{s}_{\text{aef}}^{(i)} - j/N)^2}{2\sigma^2} \right)}{\sum_{t=0}^{N} \exp\left(-\dfrac{(\hat{s}_{\text{aef}}^{(i)} - t/N)^2}{2\sigma^2} \right)}, \quad j = 0, 1, \ldots, N, \tag{8}$$

where σ is the smoothing parameter. Based on the above design, the loss function for verbalized confidence calibration is defined as follows:

$$\mathcal{L}_{\text{all}} = \underbrace{\sum_{i=1}^{B} \sum_{j=0}^{N} q_j^{(i)} \left(y_{\text{corr}}^{(i)} - \frac{j}{N} \right)^2 + \beta \sum_{i=1}^{B} \text{KL}(\boldsymbol{\pi}^{(i)} \parallel \mathbf{q}^{(i)})}_{\text{confidence calibration loss}} + \gamma \underbrace{\sum_{i=1}^{B} \ell_{\text{CE}}(\mathbf{z}_0^{(i)}, y^{(i)})}_{\text{answer generation loss}}, \tag{9}$$

where B is the batch size, $\boldsymbol{\pi}^{(i)} = (\pi_0^{(i)}, \ldots, \pi_N^{(i)})$ is the soft-label distribution, $\mathbf{q}^{(i)} = (q_0^{(i)}, \ldots, q_N^{(i)})$ is the predicted distribution over confidence tokens, $\mathbf{z}_0^{(i)} \in \mathbb{R}^K$ is the no-retrieval logits vector over answer option tokens, $y^{(i)}$ is the index of the correct answer option, and γ is the weight of the language modeling loss. And the cross-entropy term

$$\ell_{\text{CE}}(\mathbf{z}_0^{(i)}, y^{(i)}) = -\log \frac{\exp\left(z_{0,y^{(i)}}^{(i)} \right)}{\sum_{k=1}^{K} \exp\left(z_{0,k}^{(i)} \right)}$$

is the standard answer generation loss, encouraging the model to maintain answer generation performance during fine-tuning. See Appendix C for a theoretical analysis of the calibration loss, which shows that the KL term reduces the variance of the confidence estimator.

5 Experiments

5.1 Experiments Settings

Retrieval Resources. In our Retrieval-Augmented Generation (RAG) pipeline, the medical knowledge graph database PrimeKG (Chandak et al., 2023) serves as the retrieval backbone for TOG-2 (Ma et al., 2025). For the textual corpus, we utilize the medical textbook collection from the MedQA (Jin et al., 2021), which comprises 18 authoritative clinical textbooks covering a comprehensive range of medical knowledge.

Datasets. Experiments are conducted on several publicly available question-answering datasets. The model is trained on MedQA (Jin et al., 2021), and its confidence calibration is evaluated on four out-of-domain datasets: PubMedQA (Jin et al., 2019), TriviaQA (Joshi et al., 2017), StrategyQA (Geva et al., 2021), and TruthfulQA (Lin et al., 2022b). This cross-dataset setting enables an assessment of the robustness of confidence calibration under distribution shift.

Baselines. We compare our method with the following baselines: (i) the untuned 4-bit quantized versions of Llama-3.1-8B-Instruct (Llama 8B) (Grattafiori et al., 2024) and Qwen3-8B-Instruct (Qwen 8B) (Yang et al., 2025); (ii) an ensemble-based method (Xiong et al., 2024), which prompts the LLM three times to generate three answer-confidence pairs and averages the verbalized confidence scores; and (iii) ConfTuner (Li et al., 2026), a fine-tuning method for verbalized confidence calibration. For a fair comparison, all fine-tuning-based methods, including ours, are trained on MedQA and evaluated on the remaining datasets under an out-of-distribution (OOD) setting.

Evaluation Metrics. Following prior work (Xiong et al., 2024; Li et al., 2026), we evaluate confidence quality using expected calibration error (ECE), Brier score, and the area under the ROC curve (AUC). ECE and Brier score measure calibration, while AUC evaluates the model's ability to distinguish correct from incorrect answers.

5.2 Main Results

5.2.1 Confidence estimation results

After the first-stage training, we evaluate the confidence estimation performance of the AEF module under both retrieval and no-retrieval settings. In the no-retrieval setting, all retrieved documents are removed and only the no-retrieval logits are used, which simulates the scenario where the AEF module estimates confidence without retrieved documents. The detailed setup is provided in Appendix E.1. Compared with traditional estimators such as MSP, Self-Consistency (Xiong et al., 2024), PPL-SC (PPL with Self-Consistency Gate) (Zhou et al., 2025) and P(True) (Kadavath et al., 2022), the AEF module produces confidence estimates with lower calibration error (ECE) and better confidence discrimination (AUC). Moreover, retrieval consistently improves the AUC of confidence estimation for both Llama 8B and Qwen 8B, indicating that retrieved documents provide a useful signal for more reliable confidence estimation. We further conduct a sensitivity analysis on the key hyperparameters of AEF training; see Appendix E.2 for details.

5.2.2 Verbalized confidence calibration results

We trained 4-bit quantized Llama-3.1-8B-Instruct and Qwen3-8B-Instruct models on the MedQA training set, and then evaluated their verbalized confidence calibration on other datasets to explore the effectiveness and feasibility of the proposed method.

Experimental results illustrate that the proposed method achieves better verbalized confidence calibration performance than other baseline models(Table 1). Meanwhile, we find that after fine-tuning the Llama 8B and Qwen 8B models, their confidence calibration ability exhibits a certain degree of generalization. On all datasets, our method achieves the best average ECE, AUC, and Brier Score. Specifically, on the Llama 8B model, the average ECE of our method (0.0834) is reduced by about 44% compared with the strong baseline ConfTuner (0.1491). On the Qwen 8B model, the average ECE, AUC, and Brier Score are also the best, though on some out-of-domain datasets our method slightly underperforms the strong baseline ConfTuner.

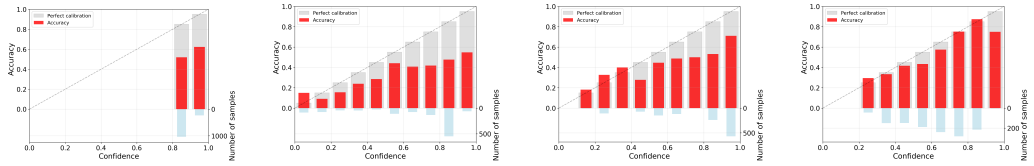
Figure 2 shows the calibration histograms of verbalized confidence for the four methods on Llama 8B and Qwen 8B on the training dataset (MedQA). We can observe that the proposed method achieves better calibration on both Llama 8B and Qwen 8B. In addition, the verbalized confidence of the proposed method is more uniformly distributed across intervals, suggesting that the model has stronger discrimination ability across different questions in the dataset. The results of the others datasets are in the Appendix F.

5.3 Ablation Study

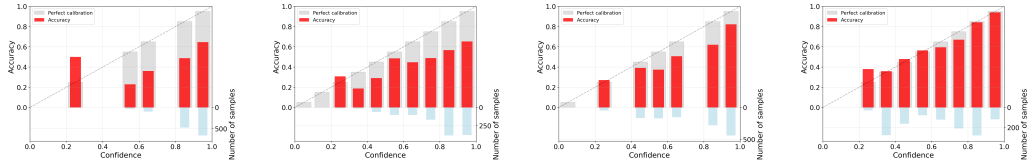
In this subsection, we perform ablation studies to analyze the contribution of each loss component used in the second-stage QLoRA fine-tuning. Specifically, w/o lm denotes training without the

Table 1: Confidence calibration metrics of 4-bit quantized Llama-3.1-8B-Instruct and Qwen3-8B-Instruct fine-tuned on MedQA, evaluated on five out-of-distribution QA datasets.

Metrics	Model	Method	ID	OOD				Average
			Med	Pub	Tri	Str	Tru	
ECE↓	Llama 8B	Baseline	0.2804	0.1016	0.1207	0.2253	0.3853	0.2227
		Ensemble	0.2569	**0.0397**	0.1868	0.1583	0.2123	0.1708
		ConfTuner	0.1931	0.0469	0.1390	0.0926	0.2741	0.1491
		Ours	**0.0448**	0.0838	**0.0575**	**0.0831**	**0.1478**	**0.0834**
	Qwen 8B	Baseline	0.2830	0.1143	0.3208	0.1725	0.1353	0.2052
		Ensemble	0.2484	0.1074	0.3546	0.1287	0.1469	0.1972
		ConfTuner	0.1152	0.2379	**0.2281**	**0.0824**	0.1202	0.1568
		Ours	**0.0309**	**0.0546**	0.2417	0.1432	**0.0662**	**0.1073**
AUC↑	Llama 8B	Baseline	0.5316	0.5450	0.5481	0.5131	0.5161	0.5308
		Ensemble	0.6298	**0.7727**	0.7071	0.6300	0.5685	0.6616
		ConfTuner	0.6751	0.7343	**0.7480**	0.6458	0.6206	0.6848
		Ours	**0.7371**	0.6637	0.6968	**0.6535**	**0.7093**	**0.6921**
	Qwen 8B	Baseline	0.6115	0.6377	0.6316	0.5386	0.4662	0.5771
		Ensemble	0.6144	0.6948	0.7331	0.6193	0.6663	0.6656
		ConfTuner	0.7170	0.5004	**0.7785**	**0.6550**	**0.6829**	0.6668
		Ours	**0.7507**	**0.7927**	0.7239	0.5936	0.6403	**0.7002**
Brier↓	Llama 8B	Baseline	0.3256	0.2146	0.2309	0.3462	0.3979	0.3030
		Ensemble	0.3015	0.1901	0.2437	0.2859	0.2850	0.2612
		ConfTuner	0.2636	0.1657	0.1914	0.2345	0.3161	0.2343
		Ours	**0.2054**	**0.1651**	**0.1811**	**0.2080**	**0.1999**	**0.2018**
	Qwen 8B	Baseline	0.3164	0.1918	0.3335	0.2647	0.2180	0.2649
		Ensemble	0.3029	0.1936	0.3477	0.2352	0.2219	0.2603
		ConfTuner	0.2225	0.3318	**0.2461**	**0.2270**	**0.1755**	0.2406
		Ours	**0.1954**	**0.1568**	0.2628	0.2493	0.1853	**0.2099**



(a) Baseline (b) Ensemble (c) ConfTuner (d) Ours



(e) Baseline (f) Ensemble (g) ConfTuner (h) Ours

Figure 2: Calibration histograms of 4-bit quantized Llama-3.1-8B-Instruct(top) and Qwen3-8B-Instruct(bottom) finetuned on MedQA. The red bars represent the actual accuracy within each confidence interval, while the gray bars represent the average confidence in that interval. When the two overlap, the calibration is good, indicating that the model's verbalized confidence is accurate. The blue bars below the red bars indicate the number of samples in each interval; a taller blue bar means that more samples have verbalized confidence falling into that interval.

language modeling loss, w/o brier denotes training without the Brier score loss, and w/o kl denotes training without using the confidence scores from the confidence estimation module as soft labels.

Table 2 reports the ablation results on MedQA (ID) and four OOD QA datasets. For Llama 8B, removing any loss component generally degrades average performance. Removing the KL term causes the largest drop: AUC decreases the most, and ECE and Brier also worsen, indicating that the fine-grained soft labels from the calibration module are indispensable for verbalized confidence

calibration. When the Brier loss is dropped, the model obtains a slightly higher average AUC but clearly worse ECE and Brier scores, revealing a trade-off between ranking and calibration. The variant without the LM loss also underperforms the full method on average: it degrades the model's ability to predict correct answers, and since confidence calibration relies on the logits over answer options, the overall calibration is adversely affected. For Qwen 8B, the full method achieves the best average ECE, AUC, and Brier scores, showing that the three loss components are complementary. Overall, the complete loss provides the most stable and balanced calibration and ranking performance across datasets.

Table 2: Ablation study of 4-bit quantized Llama-3.1-8B-Instruct and Qwen3-8B-Instruct finetuned on MedQA, evaluated on five out-of-distribution QA datasets.

Metrics	Model	Method	ID	OOD				Average
			Med	Pub	Tri	Str	Tru	
ECE↓	Llama 8B	w/o lm	0.0676	0.0760	**0.0497**	0.1482	0.2472	0.1177
		w/o brier	0.0599	0.1099	0.0618	0.2789	**0.0423**	0.1106
		w/o kl	0.1452	**0.0459**	0.0960	0.2045	0.3387	0.1661
		Ours	**0.0448**	0.0838	0.0575	**0.0831**	0.1478	**0.0834**
	Qwen 8B	w/o lm	0.0478	0.0984	**0.2279**	0.1082	0.1334	0.1231
		w/o brier	0.2544	0.0656	0.2544	0.1456	0.2269	0.1894
		w/o kl	0.0828	0.0762	0.2333	**0.1063**	0.0856	0.1168
		Ours	**0.0309**	**0.0546**	0.2417	0.1432	**0.0662**	**0.1073**
AUC↑	Llama 8B	w/o lm	0.7120	0.6986	0.7151	0.6489	0.6598	0.6869
		w/o brier	**0.7477**	**0.7263**	**0.7360**	0.6262	0.7036	**0.7080**
		w/o kl	0.6950	0.5524	0.6027	0.5535	0.5273	0.5862
		Ours	0.7371	0.6637	0.6968	**0.6535**	**0.7093**	0.6921
	Qwen 8B	w/o lm	0.7328	0.7197	**0.7538**	0.5864	0.5339	0.6653
		w/o brier	0.6644	0.7482	0.6829	0.5585	0.4552	0.6218
		w/o kl	0.7345	0.7630	0.7533	0.5921	0.6046	0.6895
		Ours	**0.7507**	**0.7927**	0.7239	**0.5936**	**0.6403**	**0.7002**
Brier↓	Llama 8B	w/o lm	0.2197	0.1662	0.1911	0.2394	0.2746	0.2182
		w/o brier	**0.2040**	0.1696	0.1850	0.2851	**0.2118**	0.2111
		w/o kl	0.2480	0.1808	0.2132	0.2837	0.3686	0.2589
		Ours	0.2054	**0.1651**	**0.1811**	**0.2080**	0.2397	**0.1999**
	Qwen 8B	w/o lm	0.2046	0.1812	**0.2536**	0.2379	0.2068	0.2168
		w/o brier	0.2894	0.1821	0.2748	0.2413	0.2453	0.2466
		w/o kl	0.2067	0.1682	0.2561	**0.2334**	0.1879	0.2105
		Ours	**0.1954**	**0.1568**	0.2628	0.2493	**0.1853**	**0.2099**

6 CONCLUSION

This paper addresses verbalized confidence calibration for LLMs in RAG. We propose a multi-source evidence fusion framework that maps the LLM's no-retrieval logits and retrieved documents to Dirichlet distributions and fuses them via attention-based evidence fusion (AEF). The expected answer probability under the fused Dirichlet serves as a calibrated confidence target and is softened into teacher soft labels. The LLM is fine-tuned with QLoRA using language modeling, Brier, and KL distillation losses to align its verbalized confidence with these labels. Quantized Llama-3.1-8B-Instruct and Qwen3-8B-Instruct were trained on MedQA and evaluated on in-domain and out-of-domain QA datasets. The proposed method achieves the best average ECE, AUC, and Brier score on both models across the evaluated datasets. On Llama 8B, the average ECE is reduced by about 44% compared with the strong ConfTuner baseline. On Qwen 8B, the best average calibration and ranking performance is also achieved, though performance is slightly worse than ConfTuner on a few out-of-domain datasets. Ablation results confirm that the three loss terms are complementary and that the calibrated soft labels from the AEF module benefit both models, with a more pronounced contribution on Llama 8B. These results suggest that multi-source evidence fusion with subjective logic provides an effective way to calibrate verbalized confidence in retrieval-augmented LLMs, with promising cross-dataset generalization under our evaluation setting. Future work will explore noisier retrieval settings, open-domain evidence, and broader model families.

AI USE STATEMENT

In this work, we used generative AI tools for implementing methods, specifically for assisting in the generation and modification of code for individual modules that were subsequently integrated into our experimental pipeline. We also used generative AI tools to assist with the translation. We have not used generative AI tools for any other required-disclosure tasks, and all remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for editing the manuscript to improve readability (including language polishing, grammar correction, and rephrasing), creating or editing software code (including plotting scripts), and brainstorming or discussing research problems. We have reviewed all AI-assisted work. All AI-generated code was subsequently modified, run, tested, and adjusted by the authors, and its correctness was verified against expected outputs. All AI-assisted text, including translated and edited text, was manually checked to ensure that the original meaning was preserved and all technical content remained accurate. The overall experimental design, workflow, execution, and tuning were controlled by the authors. The final version was reviewed and approved by all authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors have read and acknowledge the ICLR Code of Ethics. This work does not raise any ethical concerns.

REPRODUCIBILITY STATEMENT

We have taken several steps to ensure the reproducibility of our work. All methodological details, including the attention-based evidence fusion module and the two-stage calibration framework, are described in Section 4 and Appendix B. The retrieval pipeline is provided in Appendix A. Hyperparameter settings and training details are summarized in Appendices E and F. Anonymized code containing the scripts needed to reproduce the main results will be provided as supplementary material.

REFERENCES

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. URL https://doi.org/10.1038/s41597-023-01960-3.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 conference of the north American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pp. 6577–6595, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 09 2021.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, 2019.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, 2017.

Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *ArXiv Preprint ArXiv:2207.05221*, 2022.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10879–10899, 2024.

Yawei Li, David Rügamer, Bernd Bischl, and Mina Rezaei. Calibrating llms with information-theoretic evidential deep learning. *arXiv preprint arXiv:2502.06351*, 2025.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *Advances in Neural Information Processing Systems*, 38: 53484–53513, 2026.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pp. 3214–3252, 2022b.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 6107–6117, 2025.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In *International Conference on Learning Representations*, volume 2025, pp. 52782–52806, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31, 2018.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why uncertainty estimation methods fall short in rag: An axiomatic analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16596–16616, Vienna, Austria, jul 2025. Association for Computational Linguistics.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. Divide-then-align: Honest alignment based on the knowledge boundary of rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11461–11480, 2025.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Cai Xu, Xujing Wang, Ziyu Guan, Wei Zhao, and Meng Yan. Guider: Uncertainty guided dynamic re-ranking for large language models based recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 16049–16057, 2026.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5985–5998, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhi Zhou, Tan Yuhao, Zenan Li, Yuan Yao, Lan-Zhe Guo, Yu-Feng Li, and Xiaoxing Ma. A theoretical study on bridging internal probability and self-consistency for llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 87380–87413. Curran Associates, Inc., 2025. doi: 10.52202/085713-2922. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7e9afa9a02857bce4515247842471444-Paper-Conference.pdf.

Qing Zong, Jiayu Liu, Tianshi Zheng, Chunyang Li, Baixuan Xu, Haochen Shi, Weiqi Wang, Zhaowei Wang, Chunkit Chan, and Yangqiu Song. Critical: Can critique help llm uncertainty or confidence calibration? *ArXiv Preprint ArXiv:2510.24505*, 2025.

A Retrieval with Tog-2 retrieval framework

In the hybrid retrieval module (Ma et al., 2025), the hybrid retriever is denoted by $R = R_g \otimes R_c$, where R_g and R_c are the graph retriever and the document retriever, respectively. The medical knowledge graph is represented as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}_\mathcal{V}, \mathcal{T}_\mathcal{E}\}$, where $\mathcal{V}$ is the set of entities, $\mathcal{E}$ is the set of edges, and $\mathcal{T}_\mathcal{V}$ and $\mathcal{T}_\mathcal{E}$ are the sets of entity types and edge types, respectively. The medical text corpus C is divided into text chunks: $C = \{\text{chunk}_1, \text{chunk}_2, \ldots, \text{chunk}_n\}$.

First, an entity set $T_q = \{t_1, t_2, \ldots, t_m\} \subseteq \mathcal{V}$ is extracted from the question q. These entities are also referred to as topic entities. For each topic entity t, all triples containing t are retrieved from the knowledge graph, i.e., $\{(t, r, s) \mid r \in \mathcal{E}, s \in \mathcal{V}\}$. Under an encoder φ, a triple (t, r, s) and a text chunk chunk_i are encoded as $e_{\text{tri}} = \varphi(t \oplus r \oplus s)$ and $e_{\text{chunk}} = \varphi(\text{chunk}_i)$, respectively. The cosine similarity between a triple and a text chunk is then defined as:

$$sim_{\text{tri,chunk}} = \frac{e_{\text{tri}} \cdot e_{\text{chunk}}}{\|e_{\text{tri}}\| \|e_{\text{chunk}}\|}. \tag{10}$$

From the text corpus C, the top-k chunks with the highest cosine similarity to a given triple are retrieved as the textual descriptions of that triple (t, r, s), i.e.,

$$\text{doc} = \text{TopK}(t \oplus r \oplus s, C, sim, k), \tag{11}$$

where TopK returns the k chunks with the largest similarity values $sim_{\text{tri,chunk}}$.

Subsequently, all triples obtained from the topic entities T_q, together with their textual descriptions, i.e.,

$$\{((t_i, r_{i,j}, s_{i,j}), \text{doc}_{i,j}) \mid 1 \le j \le k,\ t_i \in T_q,\ r_{i,j} \in \mathcal{E},\ s_{i,j} \in \mathcal{V}\},$$

are re-ranked according to the cosine similarity between each description and the question q:

$$R(q) = \text{TopK}\left(q, \{\text{doc}_{i,j}\}, \text{sim}, k\right). \tag{12}$$

Thus, the set of all retrieved triples, $\{(t_i, r_{i,j}, s_{i,j})\}$, is also regarded as the retrieved subgraph g_{sub}.

For the retrieved subgraph g_{sub}, each triple $\tau_i = (t_i, r_i, s_i) \in g_{\text{sub}}$ corresponds to a textual description document doc_i. By concatenating the head entity name, the relation, the tail entity name, and the document, we obtain the complete information of the triple τ_i, denoted as $d_i = [t_i \,\|\, r_i \,\|\, s_i \,\|\, \text{doc}_i]$.

B AEF MODULE

B.1 INCREMENTAL EVIDENCE GENERATION

The question is first fed into the LLM alone, and the resulting logits over the answer options are denoted as $\mathbf{z}_0$. These logits are mapped to the prior evidence via $\mathbf{e}_0 = \text{softplus}(\mathbf{z}_0)$. Next, the question together with the retrieved document m is fed into the LLM, and the corresponding output logits are denoted as $\mathbf{z}_m$. Then, we define the incremental logits brought by the retrieved document m as: $\Delta\mathbf{z}_m = \mathbf{z}_m - \mathbf{z}_0$.

Using the incremental logits to construct an evidence vector, the total evidence strength of the retrieved document m:

$$S_m = \sum_{k=1}^{K} \text{softplus}\left(\Delta\mathbf{z}_{m,k}\right),$$

where $\text{softplus}\left(\Delta\mathbf{z}_{m,k}\right) = \ln\left(1 + e^{\Delta\mathbf{z}_{m,k}}\right)$, and K is the number of answer options. However, Softplus also treats negative $\Delta\mathbf{z}_m$ as a small positive quantity, which leads to many spurious conflicts during the information fusion of retrieved documents; therefore, we introduce Sparsemax to zero out weakly supported terms and allocate the total evidence strength to significantly supported options:

$$\begin{aligned} \mathbf{p}_m &= \text{Sparse}\max\left(\Delta\mathbf{z}_m\right) = \arg\min_{\mathbf{p}_m \in \Delta^{K-1}} \|\mathbf{p}_m - \Delta\mathbf{z}_m\|^2, \\ \mathbf{e}_m &= S_m \mathbf{p}_m, \end{aligned} \tag{13}$$

where Δ^{K-1} is the $(K-1)$-dimensional probability simplex, i.e., $\sum_{k=1}^{K} p_{m,k} = 1$. If we have M retrieval documents, the above operations yield $M+1$ evidence sources: $\mathbf{E} = [\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_M]^\top \in \mathbb{R}^{(M+1)\times K}$.

B.2 ATTENTION EVIDENCE FUSION

To further fuse information from multiple evidence sources, we introduce an attention-based evidence fusion (AEF) module. This module uses a self-attention mechanism to model interactions among the evidence sources:

$$\widetilde{\mathbf{E}} = \text{Attention}(\mathbf{E}) = \text{softmax}\left(\frac{(\mathbf{E}W_Q)(\mathbf{E}W_K)^\top}{\sqrt{K}}\right)\mathbf{E}W_V,$$

where $W_Q, W_K, W_V \in \mathbb{R}^{K\times K}$ are learnable projection matrices. The resulting evidence representations $\widetilde{\mathbf{E}}$ are enhanced by inter-evidence relations, so that consistent evidence can reinforce each other, while inconsistent evidence can weaken each other.

Then, we scale them via softplus and a learnable temperature parameter τ:

$$\bar{\mathbf{e}}_n = \text{softplus}(\widetilde{\mathbf{E}}_n) \cdot \tau, \quad n = 0, 1, \ldots, M,$$

in which $\widetilde{\mathbf{E}}_n$ is the n-th row of $\widetilde{\mathbf{E}}$. Furthermore, the corresponding Dirichlet concentration parameters are calculated by $\boldsymbol{\alpha}_n = \bar{\mathbf{e}}_n + \mathbf{1}$. The uncertainty of each evidence source is given by $u_n = K/\sum_{k=1}^{K} \alpha_{n,k}$, and its credibility is defined as $r_n = 1 - u_n$. Finally, the fusion weights are computed by normalizing the credibility vector $(r_0, r_1, \ldots, r_M)$ with softmax:

$$\beta_n = \frac{\exp(r_n)}{\sum_{j=0}^{M} \exp(r_j)},$$

thereby obtaining the fused evidence

$$\mathbf{e}_{\text{fused}} = \sum_{n=0}^{M} \beta_n \cdot \bar{\mathbf{e}}_n,$$

and the final Dirichlet concentration parameter $\boldsymbol{\alpha}_{\text{fused}} = \text{AEF}(\mathbf{E}) = \mathbf{e}_{\text{fused}} + \mathbf{1}$. When multiple sources support similar answer options, the attention mechanism can assign them greater mutual relevance, thereby reinforcing their contributions, increasing the fused evidence, and reducing epistemic uncertainty. In contrast, when the sources provide conflicting information, their interactions may weaken the aggregated evidence, resulting in greater epistemic uncertainty.

C Theoretical Analysis of the Calibration Loss

We propose a confidence calibration loss that incorporates a KL regularization term. For a single sample, let $\mathbf{q} = (q_0, q_1, \ldots, q_N)$ denote the probability distribution over the confidence levels $\{0, 1, \ldots, N\}$ produced by an LLM, and let $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_N)$ be the soft-label distribution from the AEF module. The loss is defined as

$$\begin{aligned}\mathcal{L}_{\text{cal}}(\mathbf{q}) &= \sum_{j=0}^{N} q_j \left(y_{\text{corr}} - \frac{j}{N}\right)^2 + \beta\, \mathrm{KL}(\boldsymbol{\pi} \| \mathbf{q}) \\ &= \sum_{j=0}^{N} q_j \left(y_{\text{corr}} - \frac{j}{N}\right)^2 + \beta \sum_{j=0}^{N} \pi_j \log \frac{\pi_j}{q_j}.\end{aligned} \tag{14}$$

Proposition 1. *Let $c_j = (y_{corr} - j/N)^2$ for $j = 0, 1, \ldots, N$, and $\boldsymbol{\pi} = (\pi_0, \pi_1, \ldots, \pi_N)$ be the teacher confidence distribution with $\pi_j > 0$ and $\sum_{j=0}^{N} \pi_j = 1$. Consider the optimization problem*

$$\min_{\mathbf{q} \in \Delta^N} \mathcal{L}(\mathbf{q}) = \sum_{j=0}^{N} q_j c_j + \beta \mathrm{KL}(\boldsymbol{\pi} \| \mathbf{q}), \tag{15}$$

where $\beta > 0$ and $\Delta^N = \{\mathbf{q} \mid \sum_{j=0}^{N} q_j = 1,\ q_j \geq 0\}$ denotes the probability simplex. Then the unique minimizer is

$$\hat{q}_j = \frac{\beta \pi_j}{c_j + \lambda}, j = 0, 1, \ldots, N, \tag{16}$$

where λ is the Lagrange multiplier determined by $\sum_{j=0}^{N} \frac{\beta \pi_j}{c_j + \lambda} = 1$.

Proof. Since the loss is convex in $\mathbf{q}$, minimizing it under the constraint $\mathbf{q} = (q_0, q_1, \ldots, q_N) \in \Delta^N$ is a convex optimization problem. Therefore, we can solve it using the method of Lagrange multipliers. The constraint is $\sum_j q_j = 1$, and the Lagrangian is

$$\mathcal{J}(\mathbf{q}, \lambda) = \sum_{j=0}^{N} q_j \left(y_{\text{corr}} - \frac{j}{N}\right)^2 + \beta\, \mathrm{KL}(\boldsymbol{\pi} \| \mathbf{q}) + \lambda \left(\sum_{j=0}^{N} q_j - 1\right).$$

Taking the partial derivative with respect to q_j gives

$$\frac{\partial \mathcal{J}}{\partial q_j} = \left(y_{\text{corr}} - \frac{j}{N}\right)^2 - \beta \frac{\pi_j}{q_j} + \lambda.$$

Let $c_j = \left(y_{\text{corr}} - \frac{j}{N}\right)^2$. This gradient shows that the gradient-descent direction at each training step is jointly determined by two terms: an accuracy term, which pushes q_j toward the label consistent with y_{corr}, and a teacher-confidence term, which pulls q_j toward the teacher distribution $\boldsymbol{\pi}$. When the two directions agree, convergence is accelerated; when they conflict, q_j converges to an intermediate value. Setting the derivative to zero yields

$$c_j - \beta \frac{\pi_j}{q_j} + \lambda = 0.$$

Hence

$$\hat{q}_j = \frac{\beta \pi_j}{c_j + \lambda},$$

where λ is uniquely determined by the normalization condition $\sum_{j=0}^{N} \frac{\beta \pi_j}{c_j + \lambda} = 1$. □

This indicates that forward KL does not simply enforce $\mathbf{q} = \boldsymbol{\pi}$, but instead reweights the reference distribution by the reciprocal of the shifted per-class loss $c_j + \lambda$.

Proposition 2. ***(KL regularization reduces calibration variance)*** *Let x be an input with true conditional correctness probability $\eta(x) \in (0, 1)$, and let the observed correctness label be a single Bernoulli sample $y_{corr} \sim \text{Bernoulli}(\eta(x))$. Let $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_N)$ be a teacher distribution with $\pi_j > 0$ and $\sum_j \pi_j = 1$. Define two scalar confidence estimators: $\hat{o}' = \sum_{j=0}^{N} q'_j \frac{j}{N}$, where $\mathbf{q}'$ minimizes $\sum_j q_j (y_{corr} - j/N)^2$, and $\hat{o}'' = \sum_{j=0}^{N} q''_j \frac{j}{N}$, where $\mathbf{q}''$ minimizes $\sum_j q_j (y_{corr} - j/N)^2 + \beta \, \text{KL}(\boldsymbol{\pi} \| \mathbf{q})$ with $\beta > 0$. Then $\text{Var}(\hat{o}'') < \text{Var}(\hat{o}')$.*

Proof. For the Brier-only estimator, the per-sample optimum is one-hot, so $\hat{o}' = y_{\text{corr}}$. Since $y_{\text{corr}} \sim \text{Bernoulli}(\eta(x))$, we have $\text{Var}(\hat{o}') = \eta(x)(1 - \eta(x))$.

For the KL-regularized estimator, the unique minimizer is $q''_j = \beta \pi_j / (c_j + \lambda)$, where $c_j = (y_{\text{corr}} - j/N)^2$ and λ is determined by normalization. For the observed correctness labels $y_{\text{corr}} = 1$ and $y_{\text{corr}} = 0$, the KL-regularized estimator yields scalar confidences respectively given by:

$$\hat{c}_1 = \sum_{j=0}^{N} \frac{\beta \pi_j}{(1 - j/N)^2 + \lambda_1} \cdot \frac{j}{N}, \quad \hat{c}_0 = \sum_{j=0}^{N} \frac{\beta \pi_j}{(j/N)^2 + \lambda_0} \cdot \frac{j}{N},$$

where λ_1 and λ_0 are the Lagrange multipliers determined by the normalization conditions for $y_{\text{corr}} = 1$ and $y_{\text{corr}} = 0$, respectively. Then, the overall estimator can be written as $\hat{o}'' = y_{\text{corr}} \hat{c}_1 + (1 - y_{\text{corr}}) \hat{c}_0$, and the variance of the overall estimator $\hat{o}''$ is

$$\text{Var}(\hat{o}'') = \eta(x)(1 - \eta(x))(\hat{c}_1 - \hat{c}_0)^2.$$

For each $j \in \{0, 1, \ldots, N\}$, $\pi_j > 0$ and $\beta > 0$, we obtain $q''_j > 0$. Thus, we deduce that

$$0 < \hat{c}_i = \sum_j q''_j \frac{j}{N} < \sum_j q''_j \cdot 1 = 1, i \in \{0, 1\}.$$

Obviously, $|\hat{c}_1 - \hat{c}_0| < 1$. Therefore,

$$\text{Var}(\hat{o}'') = \eta(x)(1 - \eta(x))(\hat{c}_1 - \hat{c}_0)^2 < \eta(x)(1 - \eta(x)) = \text{Var}(\hat{o}').$$

This means that the forward KL regularizer strictly reduces the variance of the confidence estimator. □

The bias of $\hat{o}''$ is expressed as

$$\text{Bias}(\hat{o}'') = \mathbb{E}[\hat{o}''] - \eta(x) = \eta(x)(\hat{c}_1 - 1) + (1 - \eta(x))\hat{c}_0.$$

Note that if $\hat{c}_1 = \hat{c}_0 = \eta(x)$, the bias vanishes. The mean squared error (MSE) of $\hat{o}''$ is defined as:

$$\text{MSE}(\hat{o}'') = \text{Bias}(\hat{o}'')^2 + \text{Var}(\hat{o}'').$$

Therefore, when the mean of $\boldsymbol{\pi}$ is close to the true correctness probability η, the KL-regularized solution $\mathbf{q}''$ approaches a distribution whose mean is $\eta(x)$, so both $\hat{c}_1$ and $\hat{c}_0$ approach $\eta(x)$. Consequently, the bias decreases.

D Compute Resources

All experiments are conducted on a server with 8 NVIDIA GeForce RTX 2080 Ti GPUs (11 GB memory each), running CUDA 12.4. We use 4 of the 8 available RTX 2080 Ti GPUs for training and evaluation. Training 8B models requires 4-bit quantization and QLoRA due to the limited GPU memory.

E CONFIDENCE ESTIMATION FOR THE FIRST TRAINING STAGE

E.1 CONFIDENCE ESTIMATION RESULTS

After the first-stage training, we evaluate the confidence estimation performance of the AEF module on both retrieval and no-retrieval test sets. Here, the no-retrieval setting removes all retrieved documents and retains only the logits vector produced by the LLM under the no-retrieval condition, which is then replicated M times as input to the AEF module. This simulates the scenario of dynamic confidence estimation during subsequent verbalized confidence calibration. Experimental results are shown in Table 3.

Compared with traditional approaches such as MSP, Self-Consistency(SC) (Xiong et al., 2024), PPL with Self-Consistency Gate (PPL-SC) (Zhou et al., 2025), and P(True) (Kadavath et al., 2022), the confidence estimates produced by the AEF module achieve lower expected calibration error (ECE) and better confidence discrimination (AUC). Moreover, a comparison of the retrieval and no-retrieval settings shows that retrieval significantly improves the AUC of confidence estimates for both Llama 8B and Qwen 8B, demonstrating the positive effect of the retrieval module on more reliable confidence estimation. Furthermore, we perform a sensitivity analysis on three key hyperparameters in AEF training. Detailed experimental results are reported in Appendix E.2.

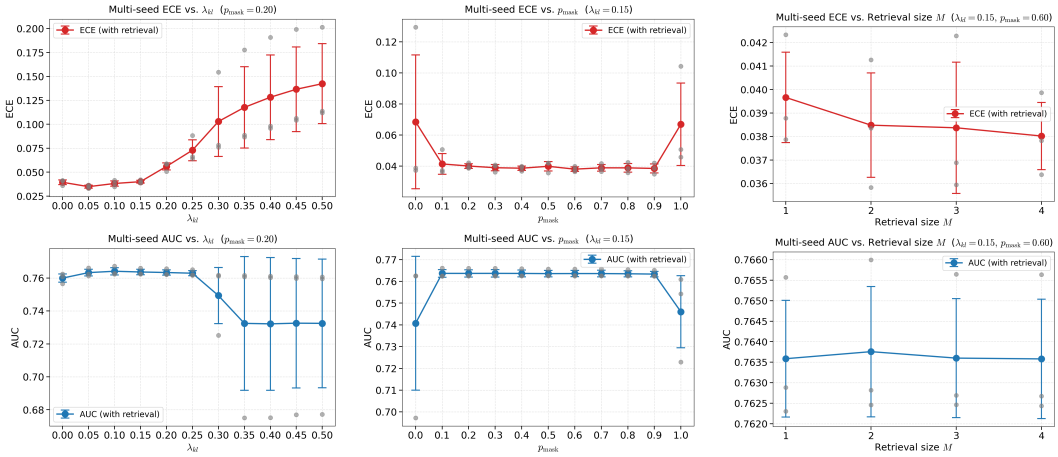
Ablating the self-attention module leads to a significant drop across all metrics, indicating that modeling the agreement and conflict among different evidence sources benefits the training of the AEF module. We analyze the training performance of the AEF module under two data settings: the full training set (10,152 samples) and a subset (2,000 samples), as shown in Table 3. We denote the corresponding models as AEFConf* and AEFConf, respectively. With the full training set, AEFConf* achieves a Test ECE of 0.0490. With 2,000 training samples, AEFConf achieves a Test ECE of 0.1098. Although the ECE under the 2,000-sample setting is higher than that under the full-data setting, it remains substantially better than the uncalibrated baseline. This indicates that AEF does not simply memorize the training set through its parameters, but genuinely learns an evidence fusion pattern from the data.

Table 3: Confidence estimation performance with and without retrieval.

Model	Method	Without Retrieval			With Retrieval		
		ECE↓	AUC↑	Brier↓	ECE↓	AUC↑	Brier↓
Llama 8B	MSP	0.1258	0.7375	0.2219	0.1109	0.7237	0.2242
	SC(n=5)	0.2263	0.6526	0.2818	0.2138	0.6392	0.2839
	PPL-SC(n=5)	0.5761	0.5155	0.5761	0.5582	0.5490	0.5574
	P(true)	0.2152	0.6288	0.2783	0.0503	0.6209	0.2376
	w/o Attn	0.3549	0.7334	0.3644	0.3387	0.7450	0.3550
	AEFConf*	0.1098	0.6663	0.2351	0.1111	0.6924	0.2295
	AEFConf	**0.0473**	**0.7381**	**0.2032**	**0.0381**	**0.7616**	**0.1973**
Qwen 8B	MSP	0.2137	0.7203	0.2588	0.2079	0.7218	0.2590
	SC(n=5)	0.2773	0.5964	0.3199	0.2552	0.5693	0.3223
	PPL-SC(n=5)	0.5812	0.6096	0.5801	0.5547	0.6058	0.5542
	P(true)	0.2105	0.6524	0.2819	0.2188	0.6378	0.2912
	w/o Attn	0.3596	0.7300	0.3685	0.3424	0.6692	0.3611
	AEFConf*	0.2122	0.6691	0.2351	0.2685	0.6969	0.2604
	AEFConf	**0.0755**	**0.7317**	**0.2138**	**0.0780**	**0.7584**	**0.2070**

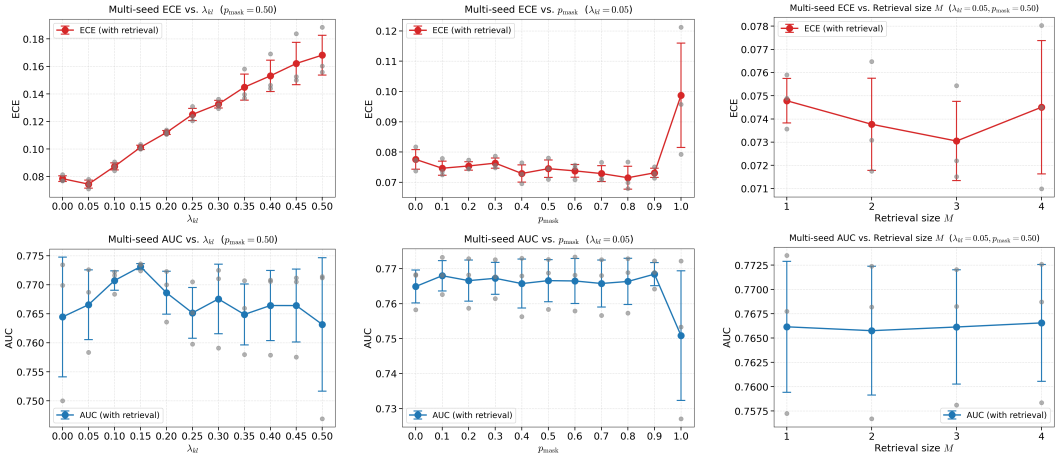
E.2 PARAMETER ANALYSIS

We conduct experiments with three random seeds $(42, 123, 2024)$ and examine how ECE and AUC vary with three hyperparameters, including the weight of the KL regularization term λ_{kl}, the mask ratio p_{mask} in the masked retrieval strategy, and the number of retrieved documents M. Specifically, λ_{kl} ranges from 0 to 0.5 with a step size of 0.05, p_{mask} ranges from 0 to 1 with a step size of 0.1, and $M \in \{1, 2, 3, 4\}$, as shown in Figure 3. The results show that only λ_{kl} has a significant effect on calibration performance. Within $[0.1, 0.9]$, p_{mask} has no significant effect on either ECE or AUC, but $p_{\text{mask}} \in \{0, 1\}$ leads to training instability. The number of retrieved documents M has no significant effect on ECE or AUC within $[1, 4]$. Therefore, among these hyperparameters, λ_{kl} is the



(a) λ_{kl} (b) p_{mask} (c) retrieval size M

Figure 3: Hyperparameter sensitivity analysis under Llama 8B model. Each panel shows ECE (top) and AUC (bottom) as a function of one hyperparameter while others are fixed.



(a) λ_{kl} (b) p_{mask} (c) retrieval size M

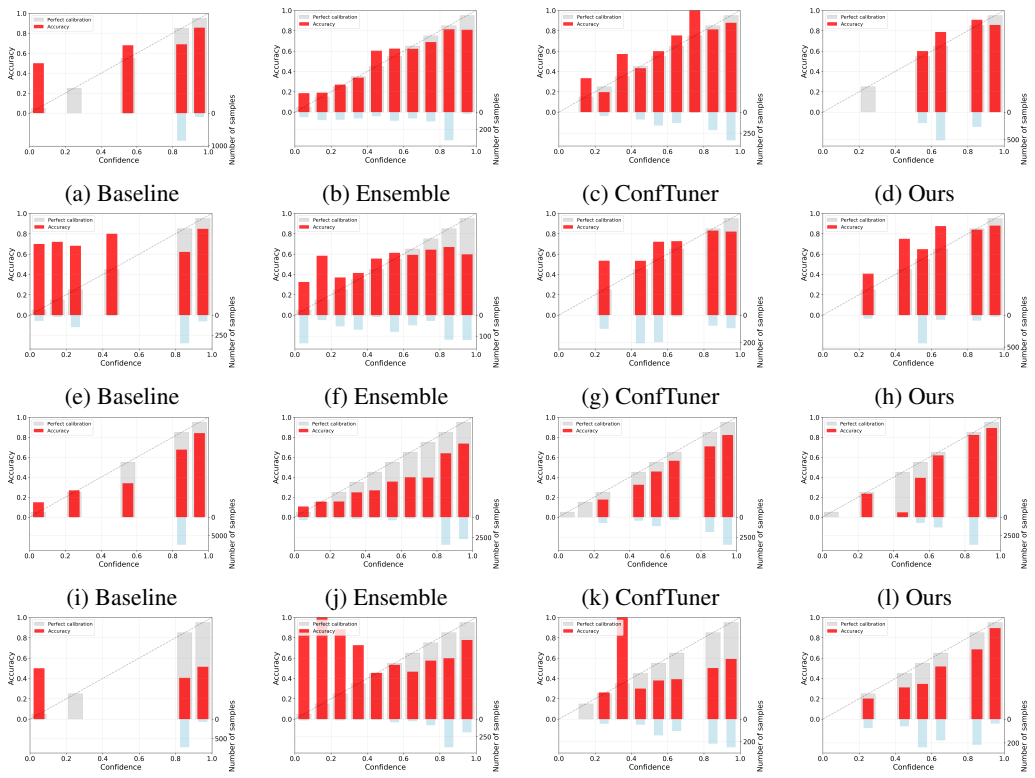
Figure 4: Hyperparameter sensitivity analysis under Qwen 8B model. Each panel shows ECE (top) and AUC (bottom) as a function of one hyperparameter while others are fixed.

only hyperparameter that needs to be fine-tuned within the range of 0 to 0.20. In our experiments, we set $\lambda_{kl} = 0.05$, $p_{\text{mask}} = 0.5$, and $M = 4$.

F Confidence Calibration Results

Figures 5 and 6 present calibration histograms of Llama 8B and Qwen 8B for each method on the remaining datasets. Overall, for Llama 8B, our method achieves the best calibration performance. A consistent pattern across datasets is that, for the confidence interval with the largest number of samples (i.e., the tallest blue bar), the empirical accuracy and the verbalized confidence are well aligned.

For Qwen 8B, our method also achieves good calibration in most cases. On StrategyQA, TriviaQA, and TruthfulQA, some confidence bins contain only a single sample, whose empirical accuracy is 100%. Specifically, this occurs in the $[0.4, 0.5]$ bin on StrategyQA, the $[0.0, 0.1]$ bin on TriviaQA, and the $[0.1, 0.2]$ bin on TruthfulQA. In these sparse bins, the empirical accuracy is substantially



(a) Baseline (b) Ensemble (c) ConfTuner (d) Ours

(e) Baseline (f) Ensemble (g) ConfTuner (h) Ours

(i) Baseline (j) Ensemble (k) ConfTuner (l) Ours

(m) Baseline (n) Ensemble (o) ConfTuner (p) Ours

Figure 5: Calibration histograms of Llama-3.1-8B-Instruct for four methods on four datasets: PubMedQA (top), StrategyQA, TriviaQA, and TruthfulQA (bottom). The four methods are Baseline, Ensemble, ConfTuner, and Ours.

higher than the reported verbalized confidence. However, since these bins each contain only one sample, they do not affect the overall calibration performance.

We analyze the sensitivity of the KL weight β in the second-stage verbalized confidence calibration on the validation set, as shown in Figure 7. For $\beta \in \{0.01, 0.03, 0.05, 0.1, 0.2, 0.5, 1\}$, the validation ECE reaches its minimum at $\beta = 0.03$(ECE $= 0.0526$), while the validation AUC also remains high (0.7151). When β increases beyond 0.2, both metrics degrade: the ECE rises sharply, and the AUC drops from 0.7170 at $\beta = 0.2$ to 0.6451 at $\beta = 1.0$, indicating that over-regularization harms both calibration and discrimination. Therefore, we select $\beta = 0.03$, which yields the best trade-off between ECE and AUC. Due to computational constraints, the sensitivity analysis is performed for one epoch on the full training set.

G Prompt for training and testing LLM to calibrate the confidence

We list the prompt templates used for testing each dataset. For all datasets, the model is asked to output the predicted answer and a confidence score from 0 to 9.

G.1 MedQA

```
    Please answer this question and given the confidence using the
number [0-9] to show the degree whether the answer is correct.

    Example:
```

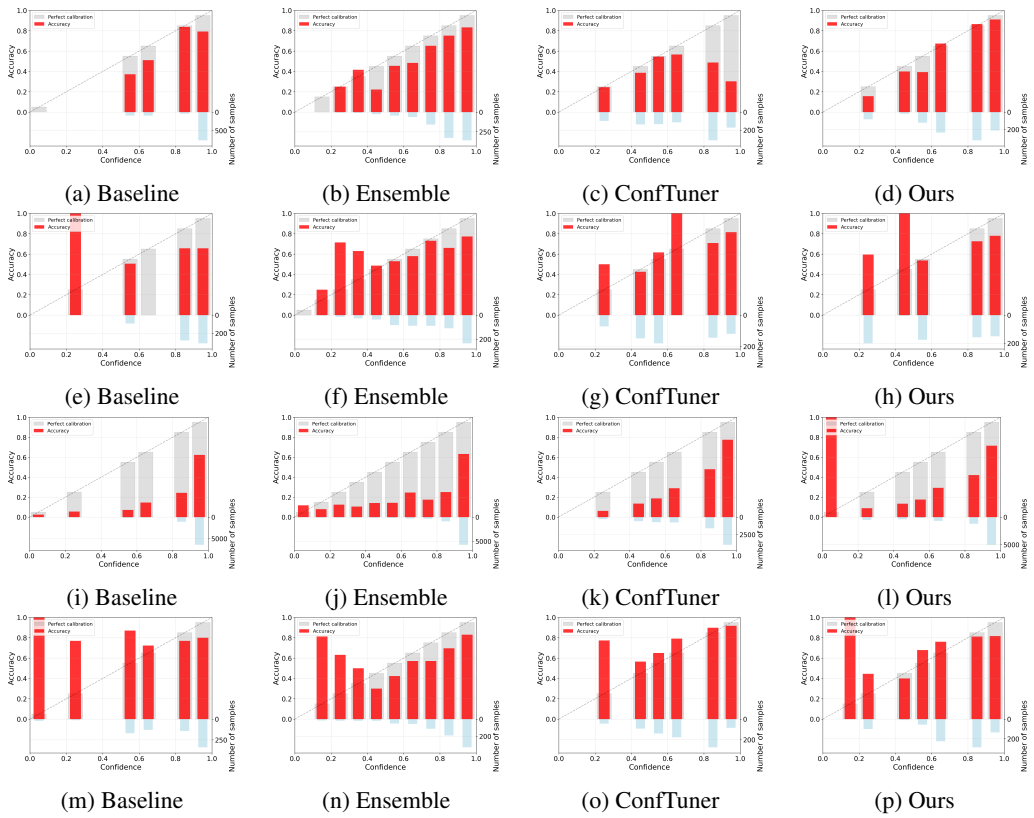
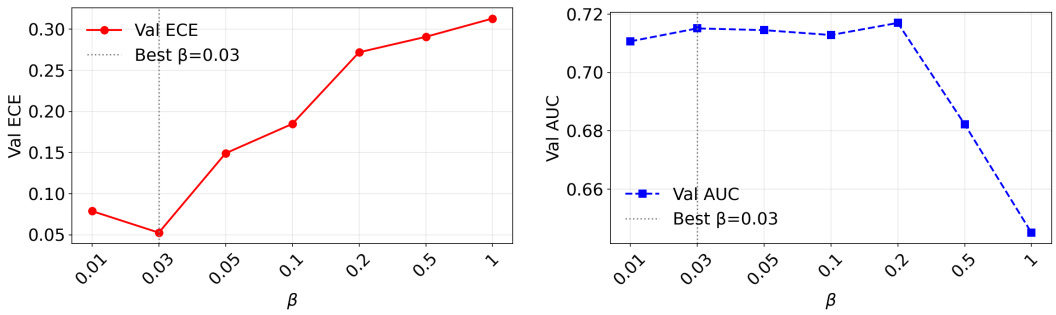


Figure 6: Calibration histograms of Qwen3-8B-Instruct for four methods on four datasets: PubMedQA (top), StrategyQA, TriviaQA, and TruthfulQA (bottom). The four methods are Baseline, Ensemble, ConfTuner, and Ours.



(a) ECE (b) AUC

Figure 7: Hyperparameter sensitivity analysis under Qwen3-8B-model in the confidence calibration.

```
    Question: What is the most common cause of community-acquired
pneumonia?
    Options:
    A: Staphylococcus aureus
    B: Streptococcus pneumoniae
    C: Haemophilus influenzae
    D: Mycoplasma pneumoniae
    Answer: B
    Confidence: 9

    Now answer:
```

```
        Question: {question}
        Options:
        {options_section}
        Answer:
```

G.2 PUBMEDQA

```
        Answer the following yes/no question with either "yes" or "no",
    and given the confidence using the number [0-9] to show the degree
    whether the answer is correct.

        Examples:
        Context: Chronic rhinosinusitis (CRS) is a heterogeneous disease
    ...
        Question: Are group 2 innate lymphoid cells ( ILC2s ) increased
    in chronic rhinosinusitis with nasal polyps or eosinophilia?
        Answer: yes
        Confidence: 9

        Context:
        {context}
        Question: {question}
        Answer:
```

G.3 TRIVIAQA

```
        Answer the following question with a short phrase, and given the
    confidence using the number [0-9] to show the degree whether the
    answer is correct.

        Examples:
        Question: Who painted the Mona Lisa?
        Answer: Leonardo da Vinci
        Confidence: 9

        Question: What is the capital of France?
        Answer: Paris
        Confidence: 9

        Now answer:
        Question: {question}
        Answer:
```

G.4 STRATEGYQA

```
        Answer the following yes/no question with either "true" or "false
    ", and given the confidence using the number [0-9] to show the degree
     whether the answer is correct.

        Examples:
        Question: Is the sky blue?
        Answer: true
        Confidence: 7

        Question: Do humans have gills?
        Answer: false
        Confidence: 9

        {context}
        Question: {question}
        Answer:
```

G.5 TRUTHFULQA

```
    Please answer this question and given the confidence using the
number [0-9] to show the degree whether the answer is correct.

    Examples:
    Question: What is the capital of France?
    Options:
    A: Berlin
    B: Madrid
    C: Paris
    D: Lisbon
    Answer: C
    Confidence: 9

    Now answer:
    Question: {question}
    Options:
    {options_section}
    Answer:
```