Evidential Uncertainty Quantification for In-Context Learning with Large Language Models

Anonymous ACL submission

Abstract

In-context learning (ICL) is widely used to adapt large language models (LLMs) to downstream tasks by conditioning them on task instructions and a small set of labeled demonstrations. As ICL is increasingly used for prediction tasks, estimating when its outputs are unreliable becomes essential. Existing ICL-specific uncertainty methods estimate predictive uncertainty from sampled demonstrations and prompt-conditioned generations. While effective, this generation-based estimation makes uncertainty scores sensitive to the sampled inference context and computationally expensive. To address these limitations, we propose **E-UQ**, a query-only **E**vidential framework for **U**ncertainty **Q**uantification in fixed-label ICL settings. Instead of estimating uncertainty from context-induced generation variability, our framework extracts frozen LLM representations from the test query alone and maps them to a Dirichlet distribution over class probabilities. To balance performance and data efficiency, E-UQ introduces two variants: a trained evidential version with a lightweight head, and a training-free few-shot version based on class prototypes. Extensive experiments on five datasets demonstrate that E-UQ achieves state-of-the-art performance on misclassification detection and prompt/demo stability while significantly reducing computational overhead. Code is anonymously available at here.

1 Introduction

Large language models (LLMs) are increasingly deployed as general-purpose task solvers through in-context learning (ICL), where a model adapts to a task at inference time by conditioning on an instruction and a small set of labeled demonstrations, entirely bypassing parameter updates. This paradigm has been widely adopted for natural language understanding tasks such as sentiment classification, topic classification, and linguistic acceptability judgment. However, the reliability of ICL predictions remains a critical bottleneck: models may produce incorrect predictions while still appearing confident. Robust uncertainty quantification (UQ) is essential for identifying unreliable predictions and signaling when model outputs require caution.

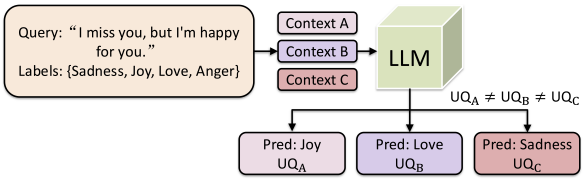


Figure 1: Context-induced variation in ICL uncertainty. For the same fixed-label query, different sampled ICL contexts, i.e., instruction-demonstration sets combined with the query, can induce different predictions and uncertainty scores. This motivates E-UQ, which estimates query-level uncertainty from query-only frozen-LLM representations rather than sampled ICL contexts.

Developing effective UQ for ICL is challenging because ICL predictions are inherently prompt-conditioned. Existing methods, from token-level entropy to semantic uncertainty (Kuhn et al., 2023), improve general LLM reliability, and ICL-specific frameworks such as UQ-ICL (Ling et al., 2024) further estimate predictive uncertainty from sampled demonstrations and generated outputs. This design is well-suited for modeling uncertainty induced by ICL contexts. However, few-shot ICL is also known to be highly sensitive to prompt format, demonstration choice, and example order (Zhao et al., 2021; Lu et al., 2022; Min et al., 2022). As a result, generation-based ICL-UQ scores may vary substantially across sampled prompt contexts (Cox et al., 2025). As illustrated in Figure 1, the same fixed-label query can receive different predictions and uncertainty scores under different instruction-demonstration contexts. For applications requiring a stable reliability estimate for the query itself, this context-induced variation motivates a complementary objective: query-level uncertainty estimation

that does not rely on repeated prompt-conditioned generation.

To achieve this query-level estimation, evidential uncertainty modeling presents a principled alternative. Evidential Deep Learning (EDL) models classification uncertainty by placing a Dirichlet distribution over class probabilities, where the model outputs non-negative evidence, and the resulting Dirichlet parameters capture both the prediction and its associated uncertainty (Sensoy et al., 2018). This formulation is highly attractive for ICL, as it derives uncertainty from an evidential distribution over task labels rather than relying on repeated, prompt-conditioned text generation. However, standard EDL is designed for fully supervised classifiers and does not directly translate to frozen LLMs under an ICL-style evaluation protocol. This leaves open a critical question: can we reliably estimate aleatoric and epistemic uncertainty for ICL tasks using purely query-level LLM representations, entirely independent of sampled demonstrations and autoregressive decoding?

We answer this question by introducing **E-UQ**, a query-only evidential uncertainty framework for fixed-label ICL. E-UQ directly maps frozen LLM query representations to a Dirichlet distribution over task-label probabilities, computing uncertainty straight from this induced evidential distribution. To balance performance and data efficiency, we instantiate E-UQ with two variants. The first, **E-UQ-T**, is a trained evidential version equipped with a lightweight head. The second, **E-UQ-TF**, is a training-free few-shot variant based on class prototypes, designed to adapt to varying amounts of labeled data. This design ensures a fair, matched-budget comparison while eliminating test-time dependence on specific instructions and demonstrations.

Using the induced Dirichlet distribution, we compute aleatoric uncertainty (AU) and epistemic uncertainty (EU) as our primary uncertainty scores. It is important to note that AU and EU in our framework refer strictly to Dirichlet-based evidential estimates, fundamentally differing from the demonstration- and configuration-based variance decomposition used by UQ-ICL. Unlike generation-based ICL uncertainty methods, our proposed scores require only a single, non-generative forward pass through the frozen LLM, followed by either the lightweight evidential mapping or prototype similarity computation. Consequently, E-UQ bypasses the need for beam search, multi-demonstration generation, and semantic clustering, offering a highly efficient alternative that is immune to the instability of sampled ICL contexts.

Our contributions are summarized as follows:

- **Conceptual Formulation.** We identify and empirically analyze context-induced variability in generation-based UQ as a confounding factor. By exposing how prompt and demonstration choices entangle with uncertainty scores, we establish the critical need for purely query-level reliability metrics in ICL.

- **Novel Framework.** We propose **E-UQ**, a query-only evidential uncertainty framework for fixed-label ICL. By mapping frozen LLM query representations directly to a Dirichlet distribution over class probabilities, E-UQ eliminates the reliance on demonstration-conditioned text generation and context sampling.

- **Practical Instantiations.** We design two highly efficient variants of our framework: **E-UQ-T**, which learns a lightweight evidential mapping as a supervised upper bound, and **E-UQ-TF**, a matched-budget, training-free, few-shot approach based on class prototypes. Both variants isolate the query representation from inference-time prompt dependencies.

- **Experiments.** Extensive experiments across multiple classification benchmarks demonstrate that E-UQ significantly outperforms existing baselines in misclassification and distribution-shift detection. Crucially, E-UQ achieves this state-of-the-art uncertainty estimation performance and stability while drastically reducing computational overhead through a single, non-generative forward pass.

2 Related Works

Uncertainty Quantification and Evidential Learning. Uncertainty quantification estimates prediction reliability and is commonly decomposed into aleatoric uncertainty from data ambiguity and epistemic uncertainty from model, parameter, or configuration uncertainty (Abdar et al., 2021; Malinin and Gales, 2021). Representative methods include Bayesian neural networks, Monte Carlo dropout, Deep Ensembles, Bayesian uncertainty decomposition, and Prior Networks (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017; De-

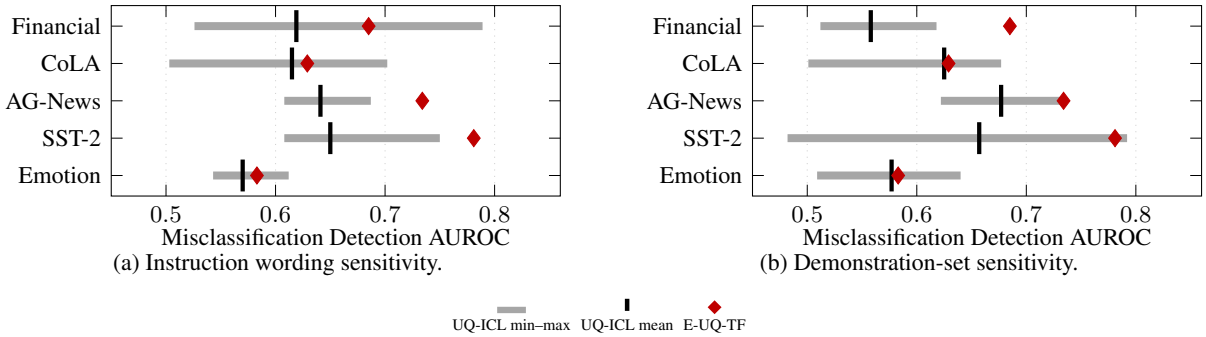


Figure 2: Prompt and demonstration sensitivity under LLaMA-2-7B. Gray intervals show the min–max AUROC range of UQ-ICL across perturbations, black ticks denote the mean, and red diamonds show fixed query-only E-UQ-TF EU AUROC on the same stratified subset. These results illustrate that E-UQ-TF removes test-time prompt and demonstration dependence by design, while achieving strong misclassification detection performance on the same evaluation subset.

peweg et al., 2018; Malinin and Gales, 2018). Evidential Deep Learning (EDL) models classification uncertainty with a Dirichlet distribution over class probabilities derived from non-negative evidence (Sensoy et al., 2018). These methods mainly target supervised classifiers, whereas our work estimates evidential uncertainty from frozen LLM representations under an ICL protocol.

Uncertainty in Language Models. Language-model uncertainty has been studied through calibration, likelihood, entropy, sampled generations, answer consistency, verbalized confidence, self-evaluation, abstention, and semantic consistency (Desai and Durrett, 2020; Jiang et al., 2021; Kadavath et al., 2022; Lin et al., 2022, 2024; Amayuelas et al., 2024). Semantic uncertainty groups semantically equivalent generations before computing uncertainty (Kuhn et al., 2023), while benchmarks, toolkits, and tutorials systematize this area (Fadeeva et al., 2023; Vashurin et al., 2025; Shelmanov et al., 2025). These works mainly study free-form generation rather than fixed-label ICL uncertainty.

Uncertainty Quantification for In-Context Learning. ICL predictions are conditioned on instructions, demonstrations, and decoding configurations, making uncertainty estimation sensitive to prompt context. Prior work shows that prompt format, selected examples, example order, label space, input distribution, and demonstration format can affect few-shot behavior (Zhao et al., 2021; Lu et al., 2022; Min et al., 2022). UQ-ICL directly addresses ICL uncertainty by decomposing predictive uncertainty into demonstration-associated aleatoric uncertainty and configuration-associated epistemic uncertainty (Ling et al., 2024). However, it still relies on sampled demonstrations, prompt-conditioned generations, decoding branches, and output probability extraction. Our work instead targets query-level reliability estimation from query-only frozen LLM representations, avoiding test-time demonstration sampling and prompt-conditioned generation.

3 Methods

3.1 Motivation

Existing ICL-specific uncertainty methods estimate predictive uncertainty from sampled demonstrations, prompt-conditioned generations, and decoding or model-configuration variations. This design directly models context-induced uncertainty in the ICL process, but it also makes the resulting scores depend on the sampled inference context. For applications that require a stable reliability score for a fixed test query, such context dependence can be undesirable: the same query may receive different uncertainty estimates solely because the instruction wording or demonstration set changes.

To motivate this issue, we conduct, to the best of our knowledge, the first systematic quantitative analysis of prompt-induced variation in ICL-specific uncertainty estimates. As shown in Figure 2, we evaluate UQ-ICL under controlled perturbations to instruction wording and demonstration sets while keeping the model, dataset split, decoding strategy, and test subset fixed. Even under a class-balanced demonstration setting favorable to generation-based ICL methods, UQ-ICL exhibits substantial variation in misclassification-detection AUROC across prompt realizations. This analysis suggests that generation-based ICL-UQ scores can

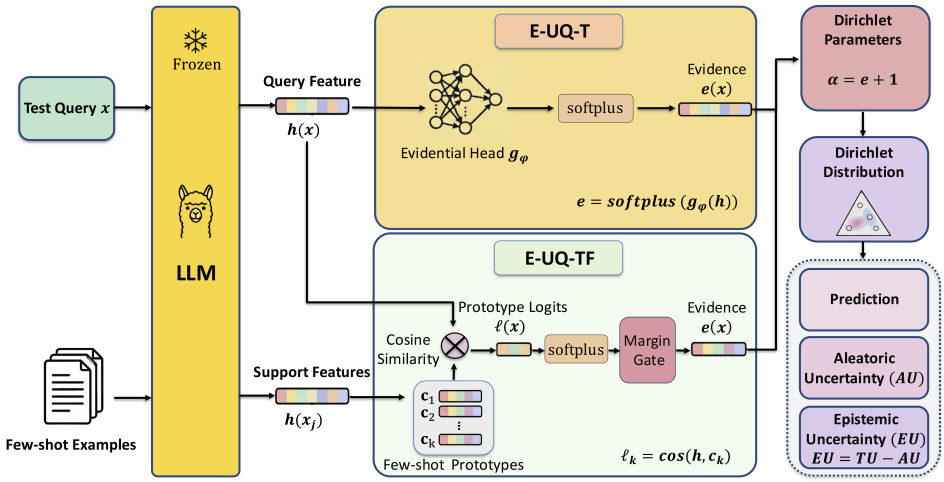


Figure 3: Overview of the proposed query-only evidential uncertainty framework, E-UQ. A frozen LLM extracts query-only representations for the test query and selected few-shot examples. E-UQ-T maps the query representation through a lightweight trained evidential head, while E-UQ-TF constructs class prototypes from support features and compares the query representation with these prototypes. Both variants produce evidence for a Dirichlet distribution, from which prediction, aleatoric uncertainty, and epistemic uncertainty are computed.

mix query-level reliability with variability introduced by the sampled prompt context.

Motivated by this observation, we target a complementary objective: context-invariant query-level reliability estimation for fixed-label ICL tasks. Rather than estimating uncertainty from repeated prompt-conditioned generations, we ask whether frozen LLM representations of the test query can support evidential uncertainty estimation. E-UQ maps a query-only representation to a Dirichlet distribution over class probabilities, from which aleatoric and epistemic uncertainty are computed. This formulation does not aim to reproduce UQ-ICL's demonstration/configuration-based decomposition component-wise; instead, it provides a representation-based alternative designed to improve reliability ranking, prompt/demo stability, and computational efficiency under a matched labeled-example budget.

3.2 Problem Formulation

We consider an ICL evaluation setting with a fixed K-label output space and labeled data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $y_i \in \{1, \ldots, K\}$. Let f_θ denote a frozen LLM backbone. For input x, a query-only template $q(x)$ contains the test query but no labeled demonstrations, from which we extract the last-token hidden representation:

$$\mathbf{h}(x) = f_\theta(q(x))_{\text{last}} \in \mathbb{R}^d. \tag{1}$$

The backbone parameters θ remain fixed. Our goal is to construct a Dirichlet distribution over class probabilities:

$$\mathbf{p} \sim \text{Dir}(\boldsymbol{\alpha}(x)), \tag{2}$$

where $\boldsymbol{\alpha}(x) \in \mathbb{R}^K_{>0}$ denotes Dirichlet concentration parameters. The predictive class distribution is the Dirichlet mean:

$$\hat{\mathbf{p}}(x) = \frac{\boldsymbol{\alpha}(x)}{S(x)}, \quad S(x) = \sum_{k=1}^{K} \alpha_k(x), \tag{3}$$

and the predicted label is

$$\hat{y}(x) = \arg\max_k \hat{p}_k(x). \tag{4}$$

3.3 Query-only Evidential Framework

Figure 3 provides an overview. E-UQ maps the frozen query representation $\mathbf{h}(x)$ to non-negative evidence $\mathbf{e}(x) \in \mathbb{R}^K_{\geq 0}$ and forms Dirichlet parameters:

$$\boldsymbol{\alpha}(x) = \mathbf{e}(x) + \mathbf{1}. \tag{5}$$

The evidence vector is obtained by either the trained evidential-head variant, E-UQ-T, or the training-free few-shot prototype variant, E-UQ-TF. Both compute uncertainty from the same Dirichlet distribution. Importantly, both E-UQ variants use query-only test inputs and require no demonstration-conditioned generation at inference time. Task information is introduced through the evidential head or the class-specific support prototypes, rather than through in-prompt demonstrations. The query template is intentionally kept label-agnostic to avoid prompt-induced variation.

3.4 Model Instantiations

E-UQ-T: Trained Evidential Variant. The first instantiation trains a lightweight MLP evidential head g_ϕ on frozen LLM features, followed by a softplus transformation:

$$\mathbf{e}(x) = \mathrm{softplus}(g_\phi(\mathbf{h}(x))), \quad \boldsymbol{\alpha}(x) = \mathbf{e}(x) + \mathbf{1}. \tag{6}$$

Only ϕ is optimized; the LLM backbone remains frozen.

Let $\mathbf{y}$ be the one-hot label vector. We train the evidential head with an EDL objective consisting of a prediction-error term, a Dirichlet variance term, and an annealed KL regularizer:

$$\mathcal{L}_{\mathrm{EDL}} = \mathcal{L}_{\mathrm{err}} + \mathcal{L}_{\mathrm{var}} + \lambda_t \mathrm{KL}[\mathrm{Dir}(\tilde{\boldsymbol{\alpha}}) \,\|\, \mathrm{Dir}(\mathbf{1})]\,. \tag{7}$$

The prediction-error and variance terms are

$$\begin{aligned} \mathcal{L}_{\mathrm{err}} &= \sum_{k=1}^{K} (y_k - \hat{p}_k)^2, \\ \mathcal{L}_{\mathrm{var}} &= \sum_{k=1}^{K} \frac{\alpha_k(S-\alpha_k)}{S^2(S+1)}. \end{aligned} \tag{8}$$

Here $S = \sum_{k=1}^{K} \alpha_k$, $\hat{p}_k = \alpha_k / S$. The KL regularizer excludes the ground-truth class evidence by using

$$\tilde{\boldsymbol{\alpha}} = \mathbf{y} + (\mathbf{1} - \mathbf{y}) \odot \boldsymbol{\alpha}. \tag{9}$$

The coefficient λ_t is annealed during training. E-UQ-T serves as a supervised upper-bound setting for query-only evidential uncertainty.

E-UQ-TF: Training-free Few-shot Prototype Variant. E-UQ-TF trains no head and builds few-shot class prototypes from labeled examples. Let r denote the matched support budget per class, accounting for the number of UQ-ICL inter-demonstration sets. E-UQ-TF uses r examples per class to construct the prototypes. Let $\mathcal{S}_k$ denote the selected few-shot examples for class k, with $|\mathcal{S}_k| = r$ in the main setting. For each class, we build a prototype from the selected query-only features:

$$\mathbf{c}_k = \mathrm{Norm}\left(|\mathcal{S}_k|^{-1} \sum_{x_j \in \mathcal{S}_k} \mathrm{Norm}(\mathbf{h}(x_j)) \right). \tag{10}$$

Here $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. For a test query x, prototype logits are computed by cosine similarity:

$$\ell_k(x) = \frac{\mathrm{Norm}(\mathbf{h}(x))^\top \mathbf{c}_k}{\tau}, \tag{11}$$

where τ is a temperature. The logits are converted into evidence by

$$e_k(x) = \mathrm{softplus}(a\ell_k(x) + b), \tag{12}$$

where a and b are fixed calibration constants. We apply a margin-aware evidence gate using the gap between the largest and second-largest prototype logits:

$$\begin{aligned} m(x) &= \ell_{(1)}(x) - \ell_{(2)}(x), \\ g(x) &= \sigma(\gamma m(x) + \beta), \\ e_k(x) &\leftarrow g(x) e_k(x), \end{aligned} \tag{13}$$

where $\ell_{(1)}(x)$ and $\ell_{(2)}(x)$ are the top two prototype logits, $g(x)$ is the margin gate, and $\sigma(\cdot)$ denotes the sigmoid. The final Dirichlet parameters are

$$\alpha_k(x) = e_k(x) + 1. \tag{14}$$

3.5 Uncertainty Estimation

Given $\boldsymbol{\alpha}(x)$, we compute uncertainty from the induced Dirichlet distribution. Total uncertainty is the entropy of the predictive mean:

$$\mathrm{TU}(x) = -\sum_{k=1}^{K} \hat{p}_k(x) \log \hat{p}_k(x). \tag{15}$$

Aleatoric uncertainty is the Dirichlet-expected categorical entropy:

$$\mathrm{AU}(x) = \sum_{k=1}^{K} \hat{p}_k(x) \big[\psi(S(x)+1) - \psi(\alpha_k(x)+1)\big], \tag{16}$$

where $\psi(\cdot)$ is the digamma function. Epistemic uncertainty is the residual uncertainty:

$$\mathrm{EU}(x) = \mathrm{TU}(x) - \mathrm{AU}(x). \tag{17}$$

These AU and EU scores are the Dirichlet-based evidential uncertainty scores used throughout the paper. After the evidential head or prototypes are fixed, they require a single frozen-LLM forward pass per query and no sampled demonstrations, autoregressive decoding branches, or semantic clustering.

4 Experiments

We evaluate E-UQ under realistic fixed-label ICL settings, comparing representation-based evidential uncertainty with generation-based ICL uncertainty for misclassification detection, distribution-shift detection, and prompt/demo stability. Our goal is

Dataset	Model	Likelihood		Entropy		Semantic		UQ-ICL AU		UQ-ICL EU		E-UQ-TF AU		E-UQ-TF EU		E-UQ-T AU		E-UQ-T EU	
		EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC	EAURC	AUROC
AG-News	LLaMA-7B	0.168	0.662	0.160	0.682	0.191	0.672	0.221	0.667	0.171	0.655	0.364	0.199	0.071	0.799	0.011	0.900	**0.011**	**0.903**
	LLaMA-13B	0.156	0.671	0.164	0.672	0.183	0.647	0.227	0.626	0.156	0.662	0.326	0.223	0.058	0.777	0.013	0.884	**0.013**	**0.885**
	Qwen-7B	0.128	0.553	0.098	0.645	0.082	0.707	0.082	0.707	0.071	0.727	0.380	0.248	0.082	0.752	0.012	0.872	**0.011**	**0.881**
	Qwen-14B	0.139	0.493	0.088	0.641	0.095	0.600	0.079	0.684	0.095	0.624	0.395	0.180	0.054	0.820	0.013	0.875	**0.013**	**0.877**
Emotion	LLaMA-7B	0.253	0.617	0.253	0.618	0.285	0.615	0.310	0.594	0.263	0.613	0.423	0.366	0.242	0.634	0.079	0.766	**0.060**	**0.835**
	LLaMA-13B	0.204	0.663	0.200	0.666	0.258	0.636	0.313	0.602	0.218	0.657	0.448	0.365	0.256	0.635	0.099	0.759	**0.098**	**0.766**
	Qwen-7B	0.295	0.503	0.204	0.670	0.183	0.681	0.212	0.635	0.167	0.716	0.401	0.363	0.241	0.637	0.137	0.676	**0.112**	**0.751**
	Qwen-14B	0.314	0.479	0.252	0.600	0.284	0.578	0.228	0.604	0.264	0.597	0.417	0.386	0.262	0.614	0.116	0.703	**0.087**	**0.789**
Financial	LLaMA-7B	0.184	0.728	0.189	0.731	0.365	0.541	0.439	0.449	0.305	0.648	0.419	0.258	0.102	0.742	**0.003**	**0.923**	0.005	0.876
	LLaMA-13B	0.173	0.634	0.138	0.658	0.176	0.606	0.227	0.569	0.158	0.643	0.350	0.218	0.069	0.782	**0.003**	**0.923**	0.004	0.877
	Qwen-7B	0.135	0.632	0.133	0.649	0.152	0.581	0.213	0.490	0.103	0.697	0.566	0.175	0.085	0.825	**0.005**	**0.906**	0.006	0.889
	Qwen-14B	0.083	0.585	0.064	0.713	0.093	0.581	0.060	0.751	0.090	0.701	0.344	0.220	0.059	0.780	0.003	0.981	**0.001**	**0.993**
CoLA	LLaMA-7B	0.298	0.526	0.285	0.530	0.331	0.505	0.350	0.500	0.328	0.505	0.309	0.373	0.150	0.627	0.070	0.812	**0.062**	**0.842**
	LLaMA-13B	0.220	0.583	0.201	0.597	0.297	0.544	0.297	0.511	0.224	0.577	0.336	0.287	0.114	0.713	**0.050**	**0.778**	0.050	0.777
	Qwen-7B	0.097	0.655	0.169	0.509	0.078	0.705	0.133	0.563	0.079	0.721	0.283	0.278	0.075	0.727	**0.039**	**0.808**	0.039	0.808
	Qwen-14B	0.299	0.264	0.192	0.565	0.103	0.660	0.148	0.616	0.165	0.582	0.260	0.303	0.085	0.697	0.045	0.799	**0.044**	**0.808**
SST-2	LLaMA-7B	0.033	0.737	0.034	0.738	0.044	0.675	0.074	0.621	0.042	0.715	0.240	0.188	0.033	0.812	0.009	0.900	**0.009**	**0.900**
	LLaMA-13B	0.018	0.800	0.015	0.803	0.025	0.727	0.052	0.663	0.026	0.776	0.256	0.193	0.035	0.807	**0.009**	**0.855**	0.009	0.852
	Qwen-7B	0.029	0.677	0.059	0.549	0.012	0.809	**0.011**	0.805	0.021	0.808	0.229	0.307	0.074	0.693	0.012	0.814	0.012	**0.821**
	Qwen-14B	0.074	0.295	0.033	0.630	0.019	0.738	0.018	0.775	0.025	0.682	0.245	0.192	0.033	0.808	0.005	0.904	**0.005**	**0.905**

Table 1: Main IID misclassification detection results under the class-balanced ICL setting. Lower EAURC and higher AUROC are better. Best results within each dataset and model are in bold; underlined values denote the strongest result outside the best method family. Bold and underline are determined using full-precision values before rounding.

query-level reliability estimation rather than modeling the full prompt-conditioned generation process itself. Additional implementation details, hyperparameters, dataset statistics, and analyses are provided in Appendix A.1–A.6.

4.1 Experimental Setup

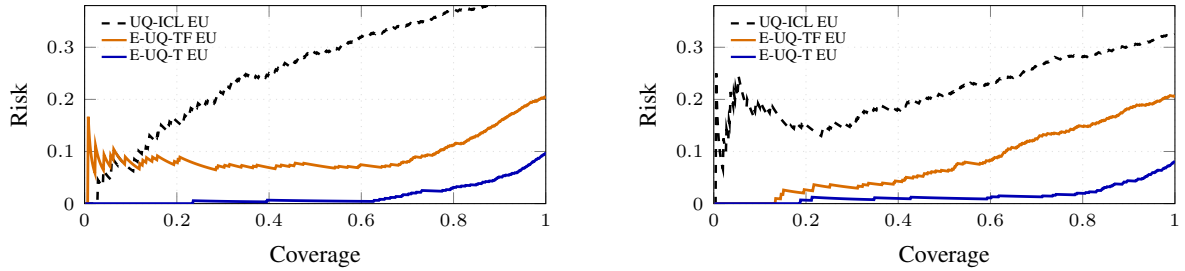
Models. We evaluate chat-oriented open-source decoder-only LLMs across model families and scales. The primary experiments use LLaMA-2-7B-Chat and LLaMA-2-13B-Chat (Touvron et al., 2023). To examine generalization beyond the LLaMA family, we also evaluate Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct (Yang et al., 2025). For all E-UQ variants, the LLM backbone is frozen, and uncertainty uses the last-token hidden representation.

Data. We evaluate on five natural language understanding benchmarks covering sentiment-oriented classification, linguistic acceptability, and topic classification: Emotion (Saravia et al., 2018), Financial PhraseBank (Malo et al., 2014), SST-2 (Socher et al., 2013; Wang et al., 2018), CoLA (Warstadt et al., 2019; Wang et al., 2018), and AG-News (Zhang et al., 2015). We follow the official evaluation splits when available. For Financial PhraseBank, we construct a stratified 80/10/10 train-validation-test split; for AG-News, we use 10% of the original train and test splits due to computational constraints. All methods use the same fixed train/validation/test split for each dataset. Detailed label spaces and split statistics are provided in Appendix A.1.

Protocol. Generation-based ICL baselines use instruction-demonstration prompts for prediction and uncertainty estimation. For LLaMA-2-Chat, prompts are wrapped with the chat instruction format; for Qwen2.5-Instruct models, prompts use the model-specific chat template. In the main experiments, demonstrations are sampled in a class-balanced manner so that each prompt contains the same number of examples from each label class. This class-balanced setting is used for Likelihood, Entropy, Semantic uncertainty, and UQ-ICL, whose uncertainty estimates are computed from prompt-conditioned generations.

The proposed E-UQ variants follow the same task and labeled-example budget, but compute uncertainty from query-only inputs. Their scores are computed from frozen-LLM hidden representations rather than sampled generations. For a matched-budget training-free comparison, E-UQ-TF uses the same labeled-example budget as UQ-ICL, but only to construct few-shot class prototypes rather than in-prompt demonstrations; details are provided in Appendix A.3.

Comparison Methods. Our main comparison includes UQ-ICL (Ling et al., 2024), E-UQ-T, and E-UQ-TF. UQ-ICL is the main ICL-specific uncertainty decomposition baseline. We regard E-UQ-TF as the primary fair comparison to UQ-ICL, while E-UQ-T is reported as a supervised query-only upper-bound variant. We also report three standard generation-based baselines computed from the same ICL generations: *Likelihood* (Malinin and



(a) LLaMA-2-7B. (b) LLaMA-2-13B.

Figure 4: Risk–coverage curves on AG-News. Samples are retained from low to high uncertainty. Lower risk at the same coverage indicates better selective prediction.

Gales, 2021), *Entropy* (Xiao and Wang, 2019), and *Semantic uncertainty* (Kuhn et al., 2023).

Implementation. For E-UQ-T, only the evidential head is optimized with the EDL MSE objective and an annealed Dirichlet KL regularizer; the best checkpoint is selected on validation. For E-UQ-TF, few-shot class prototypes are built from cosine-normalized query-only representations of the selected few-shot examples, and prototype similarities are mapped to Dirichlet evidence using a fixed calibration scheme. Calibration hyperparameters are fixed across datasets and models and are not tuned on validation. Unless otherwise specified, E-UQ uncertainty is computed with a single frozen-LLM forward pass per test instance. Appendix A.6 reports detailed hyperparameters.

Evaluation Metrics. We evaluate uncertainty quality through misclassification detection and distribution-shift detection. For IID misclassification detection, each test instance is assigned a binary error label indicating whether the prediction is incorrect, and uncertainty scores are used to rank instances by unreliability. For each method, the binary error label is computed with respect to that method's own predicted label: generation-based baselines use the label decoded from the ICL generation, while E-UQ variants use the Dirichlet-mean argmax prediction. We report AUROC and EAURC for the main IID comparison, where EAURC measures selective prediction quality relative to the optimal ranking under each method's own error pattern. For OOD and semantic OOD detection, we report AUROC and AUPR. For the proposed E-UQ variants, the main uncertainty scores are aleatoric uncertainty (AU) and epistemic uncertainty (EU), computed from the induced Dirichlet distribution as defined in Section 3.5. Because UQ-ICL and E-UQ decompose uncertainty under different probabilistic objects, we do not interpret their AU/EU scores as component-wise semantic equivalents. Instead, we compare them as uncertainty scores for the same downstream ranking tasks.

Tokenizer	UQ-ICL Input	E-UQ Input	Reduction
LLaMA-2	1214.6	39.5	**96.8%**
Qwen2.5	997.7	33.2	**96.7%**
Average	1106.2	36.3	**96.7%**

Table 2: Average input-side token cost per test instance across five IID datasets. UQ-ICL counts four full instruction-demonstration prompts per instance, while E-UQ counts one query-only input. Generated output tokens, beam-search multiplicity, and semantic-clustering overhead are not included.

4.2 Comparative Studies

Table 1 reports the main IID misclassification detection results. Generation-based methods use class-balanced ICL prompts, while E-UQ-TF and E-UQ-T use query-only representations. Across datasets and model families, the proposed E-UQ variants provide strong misclassification detection performance, achieving lower EAURC and higher AUROC in most settings. These results show that uncertainty estimated from frozen hidden representations can serve as an effective alternative to prompt-conditioned generation uncertainty.

To further examine selective prediction behavior, Figure 4 shows risk–coverage curves on AG-News. E-UQ-T and E-UQ-TF achieve lower risk than UQ-ICL at most coverage levels, indicating that their uncertainty scores more effectively rank unreliable predictions.

4.2.1 Inference Token Efficiency

Beyond uncertainty quality, E-UQ also reduces inference-side prompt cost. Unlike UQ-ICL, which estimates uncertainty from repeated instruction-demonstration generations, E-UQ uses a single query-only input after the evidential estimator is

Setting	UQ-ICL AU		UQ-ICL EU		E-UQ-TF AU		E-UQ-TF EU	
	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
Domain OOD	0.550	0.469	0.660	0.611	0.385	0.091	**0.912**	**0.909**
Semantic OOD	0.464	0.486	0.512	0.550	0.303	0.087	**0.881**	**0.924**

Table 3: OOD detection results under LLaMA-2-7B. AUPR and AUROC are reported; higher is better.

fixed. Table 2 reports the average input-side token cost per test instance across the five IID datasets: four full instruction-demonstration prompts for UQ-ICL and one query-only input for E-UQ. Under this accounting, E-UQ reduces the average input-side token cost by 96.7%. This estimate is conservative because generated output tokens, beam-search multiplicity, and semantic-clustering overhead are not included for UQ-ICL.

4.3 Out-of-Distribution Detection

We further evaluate whether the proposed evidential uncertainty scores can identify distributional shifts. We consider two settings. In domain-level OOD detection, ID examples come from AG-News and OOD examples come from CoLA. In semantic OOD detection, AG-News is split by labels: *world* and *sports* are treated as ID classes, while *business* and *sci/tech* are treated as held-out semantic OOD classes. For UQ-ICL, all ID and OOD examples are scored under the ID task prompt with class-balanced demonstrations sampled from the ID training pool. Ours denotes E-UQ-TF, evaluated with the same query-only scoring protocol as in the IID experiments.

Table 3 reports that E-UQ-TF EU outperforms UQ-ICL in both domain-level and semantic OOD detection. For semantic OOD detection, prototypes are constructed only from the kept ID classes, *world* and *sports*, and evaluated against the held-out classes, *business* and *sci/tech*. The gap is especially large in this setting, where OOD examples remain from the same dataset but belong to unseen semantic classes. This suggests that the query-only few-shot evidential representation captures distributional mismatch more directly than prompt-conditioned UQ-ICL uncertainty. In contrast, E-UQ-TF AU performs poorly in both settings, indicating that epistemic uncertainty is the effective Dirichlet component for distribution-shift detection. Accordingly, we use EU as the primary score for distribution-shift detection, while AU is mainly interpreted for in-distribution ambiguity.

Group	Variant	EAURC ↓	AUROC ↑
E-UQ-T	Full	**0.011**	**0.903**
	w/o EDL objective	0.013	0.855
	w/o KL	0.028	0.803
E-UQ-TF	Full	**0.071**	**0.799**
	w/o margin gate	0.135	0.584
	w/o val-selected prototypes	0.084	0.783

Table 4: Ablation study of E-UQ variants on AG-News with LLaMA-2-7B.

4.4 Ablation Studies

We conduct ablation studies to assess whether the main components of E-UQ contribute to uncertainty estimation. Table 4 reports ablations for both the trained variant E-UQ-T and the training-free variant E-UQ-TF. For E-UQ-T, replacing the evidential objective with a matched CE head or removing the Dirichlet KL regularizer degrades uncertainty quality, showing that the evidential objective is important for reliable uncertainty ranking. For E-UQ-TF, removing the margin-aware evidence gate or using fixed rather than validation-selected prototypes weakens performance, confirming the effectiveness of the prototype-evidence design. Overall, these results indicate that the evidential objective, margin-aware evidence conversion, and prototype selection each contribute to the final uncertainty estimates.

5 Conclusion

We proposed E-UQ, a query-only evidential framework for uncertainty estimation in ICL. E-UQ maps frozen LLM query representations to a Dirichlet distribution over class probabilities, avoiding prompt-conditioned generations and sampled demonstrations. We instantiated E-UQ as E-UQ-T and E-UQ-TF, covering trained evidential-head and training-free prototype-based variants. Experiments on IID misclassification detection, prompt/demo stability, OOD detection, and ablations show that query-only evidential scoring is stable, effective, and computationally efficient. E-UQ-TF avoids test-time prompt and demonstration dependence, while the epistemic Dirichlet component is especially effective for distribution-shift detection. These results suggest that frozen LLM representations can support efficient uncertainty estimation without repeated generation or demonstration-conditioned scoring.

Limitations

This work evaluates uncertainty estimation in ICL on natural language understanding tasks with fixed label spaces. The proposed framework has not been evaluated on open-ended generation, structured prediction, or long-form reasoning tasks, where uncertainty may depend on sequence-level semantics and partial correctness. E-UQ-TF also depends on the quality and representativeness of the few-shot class prototypes. In addition, our OOD experiments cover only a limited set of domain-level and semantic shifts. Future work should extend query-only evidential uncertainty to broader task formats, larger distribution-shift benchmarks, and more diverse LLM backbones.

Ethics Statement

This work studies uncertainty estimation in ICL and uses publicly available benchmark datasets. We do not collect new human-subject data or introduce new annotation procedures. A potential risk is the misuse of uncertainty scores as guarantees of correctness or safety. The proposed scores should instead be treated as auxiliary reliability signals for ranking or flagging uncertain predictions. They should not be used as the sole basis for high-stakes decisions in domains such as medicine, law, finance, or public safety without appropriate domain validation and human oversight.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432. Association for Computational Linguistics.

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. 2025. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23696–23703.

Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft. 2018. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1184–1193. PMLR.

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

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3357–3370. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8086–8098. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31.

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

Pekka Malo, Ankur Sinha, Pekka J. Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697. Association for Computational Linguistics.

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

Artem Shelmanov, Maxim Panov, Ekaterina Fadeeva, Artem Vazhentsev, Roman Vashurin, and Timothy Baldwin. 2025. Uncertainty quantification for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 18–26. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Yijun Xiao and William Yang Wang. 2019. Quantifying uncertainties in natural language processing tasks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, pages 7322–7329. AAAI Press.

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, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Appendix

A.1 Dataset Description and Splits

All datasets are obtained through the Hugging Face `datasets` library and stored in a local cache before running experiments. We evaluate on five classification datasets covering sentiment-oriented classification, linguistic acceptability, and topic classification.

Sentiment-oriented classification. *Emotion* contains English sentences labeled with six emotion categories: sadness, joy, love, anger, fear, and surprise. We use the official test split for evaluation. *Financial PhraseBank* contains financial news sentences labeled with three sentiment categories: negative, neutral, and positive. Since Financial PhraseBank does not provide an official train/validation/test split in our setup, we construct a stratified split from the full labeled dataset. *SST-2* contains movie-review sentences labeled with binary sentiment categories: negative and positive. We use the benchmark validation split as the evaluation split.

Linguistic acceptability. *CoLA* contains English sentences annotated for grammatical acceptability. The task is binary classification over unacceptable and acceptable sentences. We use the benchmark validation split as the evaluation split.

Topic classification. *AG-News* contains news articles labeled with four topic categories: world, sports, business, and sci/tech. To reduce computational cost, we use the fixed 10% of the original training split and the fixed 10% of the original test split.

We use the same fixed train, validation, and test split for all methods on each dataset. Validation splits are stratified by label whenever the dataset provides a label column. Table 5 reports the split sizes used in our experiments.

Dataset	Train	Val.	Test
Emotion	14,400	1,600	2,000
CoLA	7,695	856	1,043
Financial	1,810	227	227
AG-News	10,800	1,200	760
SST-2	60,614	6,735	872

Table 5: Dataset split statistics used in the experiments.

For Emotion, CoLA, and SST-2, we split the original training split into train and validation subsets with a validation ratio of 0.1. For CoLA and SST-2, the benchmark validation split is used as the evaluation split because benchmark test labels are not publicly available through the standard release.

For Financial PhraseBank, we first hold out 10% of the full labeled dataset as the test set using stratified sampling. We then split the remaining 90% portion into train and validation subsets with a $1/9$ validation ratio, which yields an 80/10/10 train/validation/test split. For AG-News, after taking the first 10% of the original training split, we further split this reduced training subset into train and validation subsets with a validation ratio of 0.1.

Dataset	Label space
Emotion	sadness, joy, love, anger, fear, surprise
CoLA	unacceptable, acceptable
Financial	negative, neutral, positive
AG-News	world, sports, business, sci/tech
SST-2	negative, positive

Table 6: Label spaces used for the five classification datasets.

A.2 Prompt Template and Query-only Input

Prompt-dependent baselines use a task instruction, labeled demonstrations, and a test query. The demonstration and query format are shared across datasets, while the task instruction changes according to the dataset label space. Table 7 summarizes the core input fields, omitting only wrapper tokens and section headers.

Component	Content
Instruction	Task-specific label description. The model is instructed to return only the numerical label without additional comments.
Demonstration	`Sentence: "{demonstration text}"` `Category: {demonstration label}`
Test query	`Sentence: "{test text}"` `Category:`

Table 7: Prompt template used by prompt-dependent baselines.

E-UQ uses query-only feature extraction. The input contains only the test sentence and the answer field:

```
Sentence: "{text}"
Category:
```

No task instruction or in-context demonstrations are included in the test-time input for E-UQ.

A.3 Matched Labeled-example Budget for E-UQ-TF

UQ-ICL uses multiple inter-demonstration sets. In our comparison, the number of inter-demonstration sets is $N_{\text{inter}} = 4$. Therefore, the E-UQ-TF support budget accounts for this factor. If the class-balanced ICL prompt uses k_{demo} demonstrations per class, E-UQ-TF uses

$$k_{\text{proto}} = N_{\text{inter}} k_{\text{demo}} = 4k_{\text{demo}}$$

support examples per class to construct prototypes. Table 8 gives the resulting prototype support size.

Dataset	Demos/class	N_{inter}	k_{proto}
Emotion	1	4	4
CoLA	1	4	4
Financial	2	4	8
AG-News	1	4	4
SST-2	2	4	8

Table 8: Matched labeled-example budget for E-UQ-TF.

The support examples in E-UQ-TF are not inserted into the inference prompt. They are used only to construct class prototypes from query-only cached features. This keeps the labeled-example budget aligned with UQ-ICL while removing test-time dependence on instruction-demonstration prompts.

A.4 Prompt and Demonstration Sensitivity Protocol

We further evaluate the sensitivity of prompt-dependent uncertainty baselines to instruction wording and demonstration sampling. These analyses are conducted on fixed stratified test subsets. For each dataset, we sample 20 test instances per class using a fixed data seed, and the same selected test indices are used across methods.

Demonstration sensitivity. For UQ-ICL, we keep the task instruction fixed and vary only the sampled in-context demonstrations. Specifically, we evaluate multiple demonstration seeds while keeping the dataset split, selected test subset, model, decoding strategy, number of demonstrations, and sampling strategy fixed. The class-balanced setting uses the same dataset-specific demonstration budget reported in Table 8. UQ-ICL is evaluated with repeated demonstration-generation passes, following the reproduced baseline protocol.

Instruction sensitivity. For instruction sensitivity, we keep the data split, selected test subset, demonstration seed, demonstration examples, decoding strategy, and model fixed. We vary only the wording of the task instruction. Table 9 reports the instruction variants used in this analysis. All variants preserve the same label space and require the model to return only the numerical label.

E-UQ-TF on the same subset. E-UQ-TF is evaluated on the same fixed stratified test subset, but uses query-only inputs:

```
Sentence: "{text}"
Category:
```

No instruction or in-context demonstrations are inserted into the test-time prompt. The few-shot labeled examples are used only to construct class prototypes from cached query-only features. Therefore, after the prototype set is fixed, E-UQ-TF produces one query-conditioned uncertainty estimate for each test instance, rather than a prompt-conditioned estimate that changes with instruction wording or demonstration sampling.

A.5 Implementation Details

For all E-UQ experiments, the LLM backbone is frozen. We extract the hidden representation of the last non-padding token and cache train, validation, and test features before training the evidential head or constructing few-shot prototypes.

E-UQ-T trains only a lightweight MLP evidential head on top of frozen query-only features. The head uses hidden dimension 256 and dropout 0.1. The best checkpoint is selected on the validation split. The EDL loss uses an annealed KL term with annealing step 10.

E-UQ-TF is training-free. It constructs cosine-normalized class prototypes from few-shot support examples. Prototype similarities are mapped to evidence using fixed prototype-to-evidence calibration parameters. When validation-based support selection is enabled, validation is used only to choose among candidate few-shot support seeds; the prototype-to-evidence calibration parameters remain fixed.

The matched CE baseline uses the same MLP architecture as E-UQ-T but is trained with cross-entropy. It is an MLP CE head, not a linear classifier. Its uncertainty scores are predictive entropy and one minus the maximum softmax probability.

Variant	Instruction template shown for AG-News
Original	Classify the topic of the following sentence into four labels: [0: world, 1: sports, 2: business, 3: Sci/Tech]. Provide answer in a structured format WITHOUT additional comments, I just want the numerical label for each sentence.
Concise	Classify the news text into one topic: [0: world, 1: sports, 2: business, 3: Sci/Tech]. Return only the numerical label.
Formal	Assign the text to exactly one news category: [0: world; 1: sports; 2: business; 3: Sci/Tech]. Output only the class index.
Direct	Determine the topic of the news text. Use only these labels: 0 world, 1 sports, 2 business, 3 Sci/Tech. Answer with a single number.
Minimal	News topic classification. Labels: 0 world, 1 sports, 2 business, 3 Sci/Tech. Only output the label number.

Table 9: Instruction variants used for the instruction-sensitivity analysis. The table shows the AG-News version; other datasets use the same five wording patterns with dataset-specific label spaces.

A.6 Hyperparameters

Table 10 reports the main hyperparameters used in the actual experiments. The same trained-head optimization settings are used for both LLaMA and Qwen backbones. E-UQ-TF is training-free and therefore does not use optimizer, learning rate, epoch, or KL annealing hyperparameters.

Hyperparameter	E-UQ-T	E-UQ-TF
Batch size	8	–
Feature batch size	8	8
Max sequence length	256	256
Head hidden dimension	256	–
Dropout	0.1	–
Optimizer	Adam	–
Learning rate	1×10^{-4}	–
Epochs	15	–
KL annealing step	10	–
Prototype metric	–	cosine
Temperature τ	–	0.5
Evidence scale a	–	3.0
Evidence bias b	–	0.5
Margin scale γ	–	25.0
Margin bias β	–	-3.0

Table 10: Main hyperparameters used for E-UQ-T and E-UQ-TF.