Explainable Text Classification with LLMs: Enhancing Performance through Dialectical Prompting and Explanation-Guided Training

Huaming Du[1], Lei Yuan[2], Cancan Feng[3], Guisong Liu[1], Gang Kou[2]*, Carl Yang[4]*
[1]School of Computing and Artificial Intelligence,
Southwestern University of Finance and Economics
[2]School of Business Administration,
Southwestern University of Finance and Economics
[3]Overseas Education College, Chengdu University
[4]Department of Computer Science, Emory University
{dhmfcc, kougang}@swufe.edu.cn, j.carlyang@emory.edu

Abstract

Large Language Models (LLMs) have achieved impressive success across a range of natural language processing tasks. However, they still underperform in text classification tasks compared to fine-tuned small models. This can be linked to complexities in addressing context-dependent expressions and complex linguistic phenomena. In contrast, fine-tuned small models typically achieve high prediction accuracy but often lack explanations for predictions. Existing explanation methods that generate keywords may be less effective due to missing critical contextual information. To mitigate these challenges, we propose a novel method termed Dialectical Explanation Training (DET). This method introduces a new prompting strategy, Dialectical Prompting, and integrates it with Explanation-Guided Training. Dialectical Prompting uses LLMs with our designed dialectical prompt to generate explanations for possible labels. These explanations handle context-dependent expressions and complex linguistic phenomena by considering multiple perspectives and providing rich, contextually relevant information. Explanation-Guided Training employs these explanations as features for training a small model, which combines the advantages of dialectical explanations and the predictive power of fine-tuned models to improve overall accuracy and interpretability. In addition, we incorporate the theory of Evidential Deep Learning, which further enhances the model's classification performance and quantify the uncertainty of its predictions. Extensive experiments on multiple datasets from diverse domains have demonstrated that our proposed model significantly improves accuracy and explanation quality over state-of the-art methods in text classification.

1 Introduction

Text classification is a foundational task underpinning many applications, where both interpretability and accuracy are crucial, especially in high-stakes domains like finance, law, and value alignment (Bhattacharjee et al., 2024). Large Language Models (LLMs), with their extensive pretraining on diverse datasets, have demonstrated remarkable success (Yao et al., 2024). However, their performance in text classification significantly lags behind that of small models such as BERT and DeBERTa (He et al., 2020; Bucher and Martini, 2024), which, despite their high prediction accuracy, often fail to provide explanations for their predictions. The dual challenge of optimizing both performance and interpretability in text classification drives our research. To tackle this, we propose a method that integrates Dialectical Prompting with Explanation-Guided Training, aiming to enhance both performance and interpretability.

Previous studies, such as those by (Sun et al., 2023a; Ziems et al., 2024), have shown that LLMs struggle with context-dependent expressions and complex linguistic phenomena due to their reliance on statistical patterns rather than deep semantic understanding (Chang and Bergen, 2024). Additionally, LLMs are often overly confident and prone to hallucinations. We aim to resolve these challenges via using dialectical explanations to strengthen the model's semantic understanding, thereby improving its ability to deal with linguistic subtleties. Figure 1 highlights a case of complex linguistic phenomena, where the sentence contains multiple layers of meaning. The conventional LLM misclassifies by favoring one interpretation, while the dialectical approach considers both, achieving correct classification with lower uncertainty[1]. An example of context-dependent expressions is the sentence "You have a strong personality." In American cultural context, this phrase is generally viewed positively, implying confidence and assertiveness.

*corresponding author

[1]Uncertainty refers to the level of confidence a model has in its predictions or outputs.

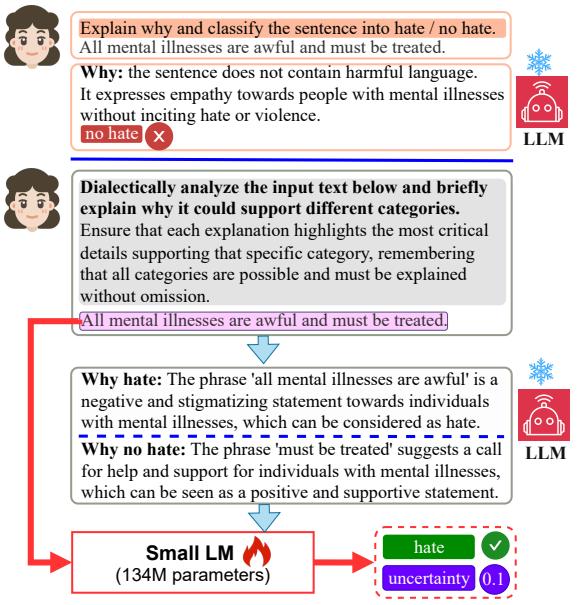


Figure 1: This figure compares two methods for classifying a sentence into "hate" or "no hate." The conventional method (top) uses a LLM with a direct explanation, often leading to incorrect classification, while the dialectical method (bottom) provides comprehensive explanations for both sides, allowing a smaller language model to fine-tune and achieve the correct classification.

However, in Chinese cultural context, the same phrase might be interpreted negatively, suggesting aggressiveness or difficulty in social interactions. This variability highlights the essential need for models to accurately understand and incorporate linguistic and context nuances.

Furthermore, small models like BERT and DeBERTa (Kenton and Toutanova, 2019; He et al., 2020) achieve high accuracy but offer limited explainability and often overlook uncertainty estimation, which is crucial for understanding model decisions, especially in sensitive applications like finance, legal, and security. Existing explanation methods (Ribeiro et al., 2016; Lundberg and Lee, 2017; Sekhon et al., 2023) typically generate a set of important keywords associated with their predictions. While these keywords can highlight significant aspects of the input text, they often fail to capture the broader context. For example, a model might identify the word "Great" as a positive keyword, but without context, it might miss that the phrase "Great, another meeting that could've been an email" actually conveys a negative sentiment. This omission of critical context information can result in incorrect explanations that undermine the trustworthiness of the model's predictions.

To tackle these limitations, we propose a novel method Dialectical Explanation Training (DET). This method introduces a new prompting strategy, Dialectical Prompting, and integrates it with Explanation-Guided Training. Dialectical Prompting utilizes LLMs with our specially designed dialectical prompt to generate diverse explanations, providing multiple perspectives to improve model understanding. By ensuring that all potential labels are thoroughly explained, this approach strengthens the model's capability to handle context-dependent expressions and complex linguistic phenomena, providing rich, context-aware information from various angles. These explanations are used as features in Explanation-Guided Training to train a small model. This combines the benefits of dialectical explanations with the predictive power of fine-tuned models to make the overall accuracy and explainability better. Finally, we introduce the theory of Evidential Deep Learning (Sensoy et al., 2018), which directly infers the model's classification uncertainty through evidence prediction.

Our main contributions are summarized as follows:

- We introduce a novel approach DET that integrates Dialectical Prompting with Explanation-Guided Training, specifically designed to tackle the dual challenges of enhancing both predictive performance and interpretability in text classification tasks.
- Our method effectively addresses the complexity of context-dependent expressions and complex linguistic phenomena by generating dialectical and high-quality explanations. In turn, these explanations, combined with evidential deep learning theory, are utilized to enhance the predictive capabilities of smaller models, thereby improving both interpretability and accuracy, while also enabling accurate measurement of the model's uncertainty.
- Through comprehensive evaluation across multiple datasets, DET demonstrates significant improvements in both classification performance and explanation quality , underscoring the method's applicability and effectiveness in high-stakes domains such as finance, law, and value alignment.

2 Related Works

Our work is related to Text Classification, Techniques for Explainability, LLM-Generated Expla-

nations, and Evidential Deep Learning. More details can be found in Appendix A.

Text Classification Approaches. Traditional text classification approaches, including fine-tuned models like BERT and DeBERTa, have set benchmarks in terms of accuracy (Kenton and Toutanova, 2019; He et al., 2020). Current methods typically focus on achieving high accuracy through LLMs because of their outstanding performance and ease of integration with various applications (Du et al., 2025; Chen et al., 2024; Zhang et al., 2025).

Techniques for Explainability. Existing explainability techniques achieve this by identifying important keywords or phrases (Lundberg and Lee, 2017; Sekhon et al., 2023). However, they often fail to capture the broader context, which can lead to misleading interpretations (Zhao et al., 2024). Additionally, generating natural language explanations alongside predictions can enhance user trust and model transparency, but may fall short in complex scenarios (Rudin, 2019).

LLM-Generated Explanations. Recent advancements have focused on using LLM-generated rationales to enhance model performance and interpretability (Kwon et al., 2024; Bhattacharjee et al., 2024). LLMs can explain their predictions by generating high-quality explanations, which improve few-shot or zero-shot performance when used to augment input prompts (Wei et al., 2022). These explanations have also been used as additional information to "self-improve" LLMs (Krishna et al., 2023; He et al., 2024a). However, the large size of LLMs limits their utility in many applications.

Evidential Deep Learning. Unlike traditional neural network classifiers that directly output the probability distribution for each sample, EDL parameterizes the Dirichlet distribution to obtain the density of classification probability assignments. Consequently, EDL leverages the properties of the Dirichlet distribution to distinguish between different types of uncertainties, demonstrating exceptional performance in uncertainty quantification and finding broad applications across various scenarios. For instance, GKDE (Zhao et al., 2020) proposed a multi-source uncertainty framework combining the Dempster-Shafer Theory (DST) for semi-supervised node classification based on Graph Neural Networks (Lu et al., 2024). (Soleimany et al., 2021) introduced evidential priors into the original Gaussian likelihood function to model uncertainty in regression networks. However, EDL has not yet been applied in the field of text classification.

3 Methodology

The framework DET we proposed, as shown in Figure 2, mainly includes Dialectical Prompting and Explanation-Guided Training to enhance both text classification and explanation performance.

3.1 Problem Formulation

Let X denote the input text space and Y the set of possible labels. Our objective is to learn a function $f : X \to Y$ that maps input text to labels while simultaneously producing explanations E for each label.

$$f : X \to Y \quad \text{with} \quad E = \{E_x(y) \,|\, x \in X, y \in Y\} \quad (1)$$

3.2 Dialectical Explanation Training

3.2.1 Dialectical Prompting

The Dialectical Prompting Strategy is an innovative method designed to enhance the explanation generation capabilities of LLMs in text classification tasks. The core idea is to encourage LLMs to generate arguments for each potential label, leading to a comprehensive and nuanced understanding of the input text. Formally, given an input sequence X and a set of possible labels Y, the task is to generate a collection of textual explanations E conditioned on a carefully crafted prompt X_{prompt}.

The construction of the prompt X_{prompt} involves several key components. Firstly, the prompting strategy X_{prompt} is designed to enhance the model's dialectical explaining ability, particularly in addressing complex linguistic phenomena. For example, the prompt might instruct the model to:

Dialectically analyze the input text below and briefly explain why it could support different categories. Ensure that each explanation highlights the most critical details supporting that specific category, remembering that all categories are possible and must be explained without omission.

Secondly, the input description X_{desc} generally outlines the domain of the input, which varies across different datasets. For example, for the StockEmotions dataset, the input description is, "comments regarding stock market activities from a financial social media platform."

Additionally, the prompt includes a set of possible labels Y and their description Y_{desc}, which are tailored to the specific classification task. For example, for the StockEmotions dataset, Y categories are "bullish" and "bearish," while the Y_{desc} is "investor sentiments." In order to get a standard

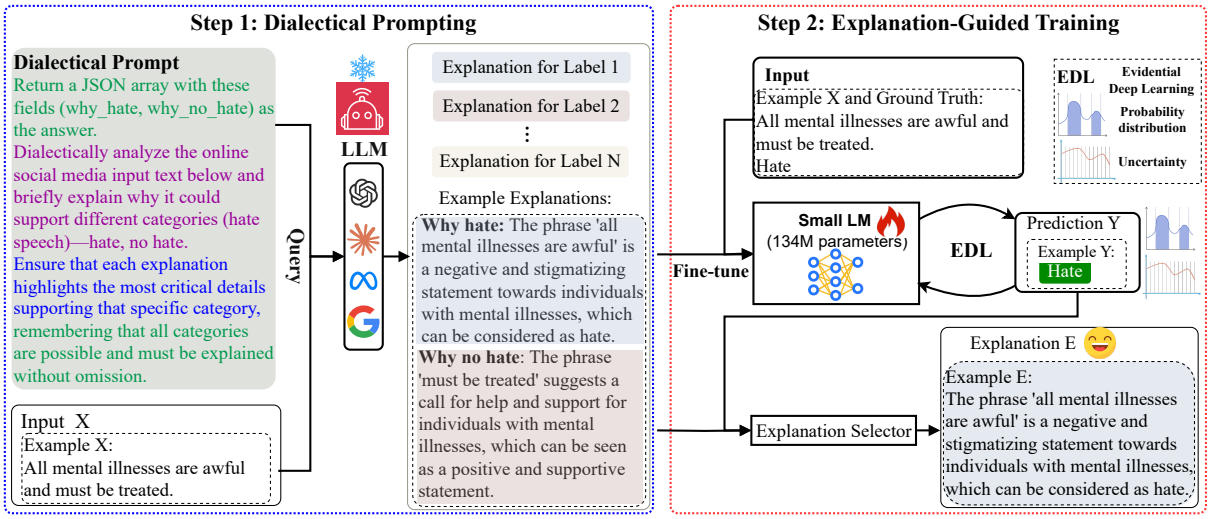


Figure 2: This is an overview of our DET, which consists of two primary steps. Step 1, the Dialectical Prompting uses dialectical prompts to generate explanations for different labels, highlighting critical details for each category. Step 2, the Explanation-Guided Training utilizes these dialectical explanations and EDL to fine-tune a smaller language model, enhancing classification accuracy and explainability.

output of the explanations, we prompt the model to return the JSON array with the "explain fields," constructed by prefixing the category types with "why" (e.g., "why_bullish" and "why_bearish" for the StockEmotions). The final prompt X_{prompt} combines all these components in the order shown in Figure 3, ensuring that the LLM provides a detailed analysis for each potential category. For each input x and label set Y, the LLM is capable of producing explanations for multiple labels simultaneously,

$$E_x = \text{LLM}(X_{\text{prompt}}(x, Y)) \,, \tag{2}$$

where $X_{\text{prompt}}(x, Y)$ is our crafted prompt based on the input x and the label set Y, and E_x contains the textual explanations generated for input x with respect to all labels in the set Y.

Dialectical Prompt:
Return a JSON array with these fields ([explain fields]) as the answer.
Dialectically analyze the $[X_{\text{desc}}]$ input text below and briefly explain why it could support different categories $[Y_{\text{desc}}]$—$[Y]$.
Ensure that each explanation highlights the most critical details supporting that specific category, remembering that all categories are possible and must be explained without omission.
Input: $[X]$.

Figure 3: Dialectical prompt template. The texts in blue are variables for different datasets.

3.2.2 Explanation-Guided Training

To fully leverage the rich contextual information contained in the explanations, the generated explanations E_x are used together with the original input text x to train a small classification model, aiming to improve both classification performance and explanation quality. Moreover, existing studies (Sun et al., 2023b; Liu et al., 2024b) mainly focus on enhancing text classification accuracy while overlooking model uncertainty. Therefore, we adopt the theory of Evidential Deep Learning (EDL) to assess the model's uncertainty.

The training process involves several key steps:
1. Data Preparation: The dataset $D = \{(x_i, E_i, y_i)\}_{i=1}^{m}$ is prepared, where y_i is the label corresponding to i-th original input x_i, and m denotes the number of samples.
2. Training Iterations: EDL provides a principled way to jointly formulate the classification problems and uncertainty modeling. Therefore, we innovatively use the EDL theory to guide the training of the model for text classification. The overall loss consists of *two parts*: Negative log Likelihood loss (improving classification accuracy) and KL divergence loss (regularized evidence distribution).

♠ **Negative log-likelihood loss:** Given a sample x_i for K-class classification, assuming that class probability follows a prior Dirichlet distribution. To improve the model's prediction accuracy, consistent with existing research (Sensoy et al., 2018; Bao et al., 2021) , we have only selected the negative

log-likelihood loss function to learn the evidence [2] $\mathbf{e}_{(i)} \in \mathbb{R}_{+}^{K}$, with the detailed derivation provided in Appendix B. The expression is as follows:

$$\begin{aligned}\mathcal{L}_{nll-edl,i}\left(y, e; \theta\right) &= -\log\left(\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B\left(\alpha_i\right)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} d\mathbf{P}_i\right) \\ &= \sum_{k=1}^{K} y_{ik}\left(\log\left(\mathcal{S}_i\right) - \log\left(e_{ik}+1\right)\right) ,\end{aligned} \tag{3}$$

where $\mathbf{e}_{(i)}$ can be expressed as $\mathbf{e}_{(i)} = g\left(\mathcal{F}\left(x_i, E_i; \theta\right)\right)$. Here, g is the evidence function to keep evidence $\mathbf{e}_k$ non-negative. $\mathcal{S}$ is the total strength of a Dirichlet distribution $\mathrm{Dir}\left(\mathbf{p} \mid \alpha\right)$, which is parameterized by $\alpha \in \mathbb{R}^K$, and $\mathcal{S}$ is defined as $\mathcal{S} = \sum_{k=1}^{K} \alpha_k$. $B\left(\alpha\right)$ represents the K-dimensional multinomial beta function. $\mathbf{P}_i$ is a simplex representing class assignment probabilities. Based on DST (Sentz and Ferson, 2002) and SL (Josang, 2016) theory, the α_k is linked to the learned evidence $\mathbf{e}_k$ by the equality $\alpha_k = \mathbf{e}_k + 1$. In the inference, the predicted probability of the k-th class is $\hat{\mathbf{p}}_k = \alpha_k / \mathcal{S}$ and the predictive uncertainty u can be deterministically given as $u = K/\mathcal{S}$.

♠ **KL divergence loss:** Inspired by existing research (Sensoy et al., 2018; Bao et al., 2021), we regularize our predictive distribution by incorporating a Kullback-Leibler (KL) divergence term into the loss function, which penalizes divergences that do not contribute to data fitting and deviate from the "*I don't know*" state, resulting in higher uncertainty for misclassified samples. Therefore, the specific loss function is as follows:

$$\mathcal{L}_{kl} = KL\left(\mathrm{Dir}\left(\mathbf{p}, \tilde{\alpha}\right), \mathrm{Dir}\left(\mathbf{p}, \mathbf{1}\right)\right) , \tag{4}$$

where $\tilde{\alpha}$ is the Dirichlet parameters after removal of the non-misleading evidence from predicted parameters α, $\mathbf{1}$ represents the parameter vector of K ones. Therefore, we can obtain the overall loss:

$$\begin{aligned}\mathcal{L}\left(\theta\right) = &\sum_{i=1}^{m} \mathcal{L}_{nll-edl,i}\left(y, e; \theta\right) \\ &+ \lambda_t \sum_{i=1}^{m} KL\left[\mathrm{Dir}\left(\mathbf{p_i} \mid \tilde{\alpha}_i\right) \| \mathrm{Dir}\left(\mathbf{p_i} \mid \langle 1, \cdots, 1\rangle\right)\right] ,\end{aligned} \tag{5}$$

where $\lambda_t = \min\left(1.0, t/10\right) \in [0, 1]$ is the annealing coefficient, t is the current training epoch.

3.2.3 Inference Stage

For the inference stage, new input x is processed to generate explanations E_x using the LLMs. The trained model then evaluates these explanations in conjunction with the original text to make the final prediction and estimate uncertainty:

$$\hat{y} = h(x, E_x), \qquad u = \frac{K}{\mathcal{S}} , \tag{6}$$

where h represents the classification model. Please note that when the value of u exceeds a certain threshold σ, DET refuses to make a prediction. We select the explanation corresponding to the predicted label of the trained classification model as the final explanation. Formally, let $\hat{y}$ be the predicted label for input x generated by the model. The final explanation $\hat{E}$ is defined as:

$$\hat{E} = E_x(\hat{y}) , \tag{7}$$

where $E_x(\hat{y})$ denotes the explanation generated for the predicted label $\hat{y}$ in Section 3.2.1.

4 Experiments

4.1 Experimental Setup

Datasets. In line with existing research (Liu et al., 2024b), we selected seven representative datasets to validate the effectiveness of our method: StockEmotions (Lee et al., 2023), Overruling (Zheng et al., 2021), Stance4Trump (Kawintiranon and Singh, 2021), Ethos (Mollas et al., 2022), RTE (Dagan et al., 2006), ValueNet (Qiu et al., 2022), AGNews (Zhang et al., 2015), and MASSIVE (Fitzgerald et al., 2023) as shown in Table 1. Detailed data descriptions can be found in Appendix C.1.

Setup. To balance performance and efficiency, we use GPT-3.5-turbo (OpenAI, 2024) as the explanation-generating LLM and DeBERTa-base (He et al., 2020) as the Small Language Model (SLM). DeBERTa-base comprises 134 million parameters, approximately three orders of magnitude fewer than GPT-3.5-turbo's 175 billion parameters. All results are averaged across five runs for robustness, with the LLM output generated at a temperature setting of 0.01. Model training was conducted on an A5000-24G GPU with a batch size of 9.

Baselines. To highlight the advantages of our DET method, we compared it against the following three categories of text classification approaches: (1) **Models without explanation**, including GPT-3.5-turbo with Zero-shot Learning (ZSL) and Few-shot Learning (FSL, k = 4), GPT-4o-mini (OpenAI, 2024) with ZSL and FSL (k = 4) , DeBERTa-base with the standard label supervision, CARP (Sun et al., 2023b) with clue prompting,

[2] The neural network estimates the strength of support for each class, which serves as the parameters of a Dirichlet distribution used to model uncertainty.

Dataset	Train	Validate	Test
StockEmotions	8,000	1,000	1,000
Overruling	1,920	168	312
Stance4Trump	875	112	263
Ethos	601	67	300
RTE	2,241	249	277
ValueNet	16,030	3,206	2,138
AGNews	120,000	-	7,600
MASSIVE en-US	11,514	2,033	2,974

Table 1: Statistics of the experimental datasets.

Metrics	Human-GPT4o	Human-Claude 3.5
Pearson coef.	0.673	0.572
Pearson P-value	2.291e-80	8.001e-45
K-alpha	0.668	0.572

Table 2: Comparison of evaluation metrics between human assessments and those of GPT-4o and Claude 3.5-sonnet. The Pearson correlation coefficient for GPT-4o (0.673) indicates a moderately strong positive correlation with human ratings, while Claude 3.5 (0.572) shows a moderate positive correlation. Both P-values are very low, indicating significant correlations. The K-alpha values suggest relatively high consistency in evaluations, supporting the validity and reliability of the designed EQ metric. Note: Claude 3.5 refers to Claude 3.5-sonnet, and K-alpha refers to Krippendorff's alpha.

and LLMEmbed (Liu et al., 2024b) with LLM for extracting text embeddings; (2) **Models with explanations**, specifically GPT-3.5-turbo with explanations, where the LLM outputs explanations along with its classification predictions; LLAMA 3.1-8B, GPT-4, GPT-4o-mini, o1-mini, and DeepSeek-R1 with explanations; DeBERTa with Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al., 2016) to explain DeBERTa's predictions, and DeBERTa with SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to explain DeBERTa's predictions; and (3) **Supervised methods with dialectical explanations**, specifically the variant of our method, DET W/O EDL. Please note that we do not use GPT-4o (OpenAI, 2024) as a baseline because it served as our metric EQ evaluation model. In addition, we did not use the more powerful o1 and DeepSeek for evaluation, primarily due to cost considerations.

Evaluation Metrics. Similar to previous methods (Liu et al., 2024b), we use ACC and F1 to evaluate classification performance. We have designed a novel metrics: **Explanation Quality (EQ)**, to assess the quality of the model-generated explanations for predictions, inspired by related studies (Hoffman et al., 2018; Chen et al., 2023). The evaluation method employs two advanced LLMs: GPT-4o and Claude 3.5-sonnet, and uses a 5-point scale to evaluate explanations based on five dimensions: *clarity*, *relevance*, *completeness*, *consistency*, and *credibility*. Additionally, we compare the LLM evaluations with those of five human experts on 120 samples, analyzing consistency using Pearson correlation coefficient and Krippendorff's alpha. As shown in Table 2, the LLM evaluations are highly consistent with those of human experts, validating the effectiveness and reliability of our designed EQ metric. More details can be found in Appendix C.3. Our code is available at https://github.com/trytodoit227/DET.

4.2 Main Results

We conducted extensive experiments on several benchmark datasets to evaluate our proposed model, DET, against various baselines. The results are summarized in Table 3. Overall, DET demonstrates strong performance across all datasets, consistently outperforming baselines in both classification performance and explanation quality. DET achieves the highest EQ scores on all seven datasets, clearly showcasing its ability to generate superior explanations. Additionally, DET leads in both ACC and F1 across most datasets. For example, on the RTE dataset, compared to the second-best baseline DeBERTa, DET achieves an 11.80% improvement in ACC, a 13.25% improvement in F1, and a remarkable 133% increase in EQ. Similarly, on the AGNews dataset, DET also outperforms the second-best baseline, LLMEmbed, with a 5.77% gain in ACC and a 5.57% gain in F1. We also conducted performance comparisons in the multi-label setting, as shown in Table 4. Our method achieved relative improvements of 4.57%, 5.27%, and 17.42% in ACC, F1, and EQ, respectively. These results indicate that DET excels not only in generating high-quality explanations but also in delivering accurate predictions, which makes it a robust solution across diverse text classification tasks.

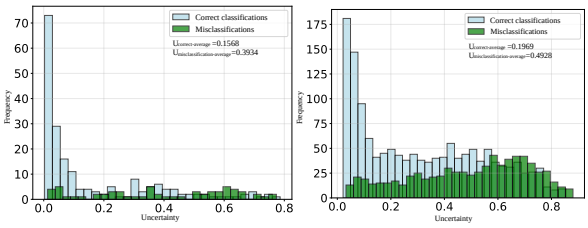
However, DET shows limited performance improvement on the StockEmotions and Ethos datasets, potentially because these datasets are closely related to sentiment analysis tasks and may have been used in the training of LLMs. *It is important to note that we use a smaller model, DeBERTa, for classification, which avoids data leakage issues while still achieving better performance than more powerful inference LLMs*. In contrast, DET demonstrates more significant improvements on domain-specific datasets such as TRE and Val-

Methods	Finance			Law			Politics			Hate			NLI			Values			News		
	StockEmotions			Overruling			Stance4Trump			Ethos			RTE			ValueNet			AGNews		
	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ
Models With No Explanation																					
GPT-3.5 (ZSL)	71.81	70.14	-	90.85	86.94	-	70.61	69.89	-	83.61	82.67	-	49.46	28.16	-	61.20	55.62	-	86.23	87.23	-
GPT-3.5 (FSL, k=4)	83.27	80.69	-	92.81	90.67	-	75.77	73.45	-	84.93	82.21	-	73.01	58.67	-	55.29	48.29	-	88.70	88.68	-
GPT-4o-m (ZSL)	79.00	76.31	-	95.51	94.55	-	76.05	76.13	-	85.33	85.00	-	49.10	29.61	-	57.34	41.80	-	83.43	83.00	-
GPT-4o-m (FSL, k=4)	79.70	78.50	-	94.55	93.26	-	76.81	75.68	-	86.50	86.66	-	71.84	40.64	-	60.57	49.67	-	87.70	87.44	-
o1-mini (ZSL)	83.10	80.00	-	91.99	94.40	-	73.59	74.81	-	86.00	85.92	-	49.82	31.51	-	61.27	59.85	-	85.41	84.69	-
o1-mini (FSL, k=4)	82.90	78.92	-	92.31	92.29	-	76.81	75.79	-	85.77	85.40	-	58.12	31.22	-	62.59	60.13	-	86.97	86.32	-
DeepSeek-R1 (ZSL)	78.60	80.79	-	93.91	93.57	-	77.51	76.31	-	85.00	84.98	-	48.01	30.47	-	59.96	41.92	-	86.60	86.56	-
DeepSeek-R1 (FSL, k=4)	81.00	79.26		90.38	90.35	-	77.75	**77.64**	-	85.33	85.25	-	49.10	30.94	-	60.24	57.67	-	88.16	86.38	-
DeBERTa	77.93	75.47	-	96.69	96.11	-	73.76	73.19	-	81.22	78.56	-	76.41	72.29	-	66.03	62.21	-	85.13	83.61	-
CARP	74.25	72.18	-	94.36	91.04	-	70.46	69.81	-	79.61	78.23	-	60.54	60.25	-	65.43	63.62	-	83.29	82.76	-
LLMEmbed	77.93	76.34	-	96.74	94.85	-	71.11	70.28	-	81.22	79.65	-	62.71	60.47	-	66.12	63.79	-	90.25	88.24	-
Models With Explanations																					
GPT-3.5	68.20	69.50	4.17	83.33	78.90	3.67	63.50	59.60	3.69	73.00	69.78	3.96	47.29	28.02	4.05	55.14	49.05	3.95	87.00	88.52	3.87
LLAMA 3.1-8B	70.46	69.84	4.13	89.83	89.74	4.31	56.27	55.06	3.86	74.41	53.62	2.08	50.90	26.18	4.17	55.43	45.60	3.14	12.67	45.25	3.33
GPT-4	68.20	68.05	4.17	93.35	93.44	4.34	74.43	72.66	3.90	84.67	85.81	4.24	48.01	30.26	4.31	55.14	52.36	3.95	88.57	87.94	4.65
GPT-4o-m	75.10	74.31	4.20	90.70	90.31	4.12	68.44	67.42	3.75	85.00	83.31	4.05	46.93	30.07	4.18	58.00	46.00	4.06	87.60	87.65	4.37
o1-mini	80.37	78.11	4.20	94.53	94.53	4.12	75.20	74.23	3.75	85.58	84.51	4.05	49.82	31.29	4.18	58.00	57.64	4.06	87.41	86.38	4.50
DeepSeek-R1	78.60	79.03	4.25	95.19	95.19	4.21	77.57	76.38	4.06	86.33	86.33	4.72	49.82	29.68	4.35	58.61	58.36	4.66	83.77	84.09	4.32
DeBERTa + LIME	77.93	75.47	1.95	96.69	96.11	2.00	73.76	73.19	1.85	81.22	78.56	1.90	76.41	72.29	1.80	66.03	62.21	1.87	85.13	83.61	1.84
DeBERTa + SHAP	77.93	75.47	2.10	96.69	96.11	2.15	73.76	73.19	2.05	81.22	78.56	2.10	76.41	72.29	2.00	66.03	62.21	2.08	85.13	83.61	2.37
DET (our)	**84.83**	**82.36**	**4.61**	**97.31**	**97.25**	**4.55**	**79.95**	76.17	**4.26**	**87.63**	**87.09**	**4.68**	**85.43**	**81.87**	**4.66**	**68.32**	**66.23**	**4.23**	**95.46**	**93.16**	**4.85**

Table 3: Performance comparison across datasets. Note: GPT-3.5 = GPT-3.5-turbo; GPT-4o-m = GPT-4o-mini. Bold numbers indicate the best results. See Table 10 in Appendix C.4 for the complete results.

Methods	ACC	F1	EQ
Llama3.1-8B	81.74	79.62	3.38
Qwen3-8B	74.48	68.53	3.32
GPT-3.5	70.12	66.35	3.15
GPT-4	80.25	80.49	3.37
o1-mini	86.95	82.37	3.56
DeepSeek-R1	86.47	80.29	3.51
DET	**90.93**	**86.71**	**4.18**

Table 4: Performance comparison on MASSIVE en-US.



(a) Stance4Trump (b) ValuesNet

Figure 4: Uncertainty distribution for different datasets.

Methods	Values		Politics		Finance		Hate	
	ValueNet		Stance4Trump		StockEmotions		Ethos	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1
DeBERTa	43.35	39.29	33.26	21.02	42.50	32.00	47.00	40.52
LLMEmbed	45.16	39.42	35.27	30.14	43.05	46.28	47.35	52.81
CARP	44.73	36.42	33.87	31.05	42.93	43.64	46.38	50.49
DET (our)	**59.81**	**59.13**	**51.16**	**45.51**	**54.60**	**71.63**	**60.33**	**71.86**

Table 5: Cross-domain experimental results.

ueNet. Since these datasets are less common and involve more complex content, LLMs may not perform as strongly, offering DET an opportunity to significantly improve prediction accuracy through richer explanations. This contrast highlights DET's varying performance across datasets, with greater potential in specialized domains, suggesting its value in less common or specialized datasets.

4.3 Uncertainty Analysis

Through the theory of EDL, we can obtain not only the classification probability for each sample but also the corresponding uncertainty. Figure 4 shows the uncertainty distribution across different datasets. Overall, the model assigns lower uncertainty to correctly classified samples, while higher uncertainty is observed for misclassified samples. In other words, when the uncertainty value of a classified sample is high, we should reject the model's prediction. For more experimental results, please refer to Appendix C.4.

4.4 Domain Adaptation

In this subsection, we conduct a robustness analysis using out-of-distribution datasets, as shown in Table 5. We observe a significant performance drop when the test set differs from the training set in the baseline model (e.g., when Stance is used as the train set and Values as the test set, the ACC of LLMEmbed drops from 66.12% to 45.16%, and F1 drops from 63.79% to 39.42%), indicating that the baseline model is highly sensitive to out-of-distribution data. In contrast, DET's ACC drops from 68.32% to 59.81%, and F1 from 66.23% to 59.13%, further demonstrating that our dialectical prompting can effectively alleviate this issue and exhibit stronger domain adaptation capabilities.

4.5 Ablation Studies

To evaluate the contribution of each component, We conduct the ablation study on the Ethos dataset

Variant	ACC	F1	EQ
DET	**87.63**	**87.09**	**4.68**
w/ Dial.explanations only	86.44	84.61	4.63
w/ Orig.Text + Dial.explanations (i.e., DET)	**87.63**	**87.09**	**4.68**
w/ Dial.explanations + Orig.Text	84.44	82.16	4.66
w/ Uni.explanation only	79.00	76.25	3.78
w/ Orig.Text + Uni.explanation	83.00	80.68	3.76
w/ Uni.explanation + Orig.Text	80.56	78.39	3.75
w/ Orig.Text only	81.22	79.17	3.82
w/ Cross-Entropy Loss	86.00	85.49	4.48

Table 6: Ablation study results on zero-shot Ethos.

Our DET with LLM	Law			Politics			NLI		
	Overruling			Stance4Trump			RTE		
	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ
LLAMA 3.1-8B	89.83	89.74	4.31	56.27	55.06	3.86	50.90	26.18	4.47
LLAMA 3.1-70B	96.90	95.84	**4.70**	**79.85**	**77.63**	**4.50**	84.84	60.42	4.53
LLAMA 3.1-405B	**97.01**	**96.21**	4.66	77.82	76.57	4.31	**90.25**	**70.36**	**4.82**

Table 7: Performance under LLMs of different sizes.

Methods	Overruling			Stance4Trump			RTE		
	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ
GPT4(+COT)	94.03	93.86	4.26	75.26	75.11	3.99	50.49	31.37	4.39
GPT-4O-m(+COT)	92.15	92.08	4.20	70.19	70.94	3.81	48.05	31.26	4.21
Llama 3.1-8B(+COT)	90.47	90.14	4.34	61.53	60.36	3.87	49.64	27.04	4.19
DET	**97.31**	**97.25**	**4.55**	**79.75**	**76.17**	**4.26**	**85.43**	**81.87**	**4.66**

Table 8: Performance Comparison between dialectical explanations and CoT .

Methods	MASSIVE fr-FR	MASSIVE zh-CN
GPT4	0.24	0.28
o1-mini	0.38	0.41
DeepSeek-R1	0.42	0.53
LLMEmbed	1.34	2.07
DeBERTa (W/O generate explanations)	0.76	0.95
DeBERTa + LIME	1.45	1.64
DeBERTa + SHAP	1.63	1.79
DET	1.07	1.36

Table 9: Average time (s) on the MASSIVE fr-FR and MASSIVE zh-CN.

with several variants, which are introduced as follows: 1) ***w/ Dial.explanations only*** variant, only dialectical explanations generated by GPT-3.5-turbo are used, excluding the original text. 2) ***w/ Orig.Text + Dial.explanations*** **(i.e., DET)** and ***w/ Dial.explanations + Orig.Text*** variants combine both elements, differing in input order. 3) ***w/ Uni.explanation only*** variant uses only unidirectional explanations. 4) ***w/ Orig.Text + Uni.explanation*** and ***w/ Uni.explanation + Orig.Text*** variants combine unidirectional explanations with the original text in different sequences. 5) ***w/ Orig.Text only*** variant serves as a baseline, using just the original text. 6) ***w/ Cross-Entropy Loss*** variant uses cross-entropy for training.

As shown in Table 6, using only dialectical explanations achieves the second-highest EQ score of 4.63, along with an accuracy of 86.44% and an F1 score of 84.61%, highlighting the significant impact of dialectical explanations. When the original text is combined with dialectical explanations, the model achieves high performance, while the baseline using only unidirectional explanations performs the worst. Using only the text also results in a significant drop in all metrics. Furthermore, replacing the loss function in DET with cross-entropy loss leads to performance degradation. These comprehensive ablation studies further validate the critical role of our method in enhancing both model performance and explanation quality.

Furthermore, to explore DET's performance with LLMs of varying sizes, we conducted additional experiments on the RTE, Stance4Trump, and Overruling to assess this impact. As shown in Table 7, with the increase in the size of LLMs, ACC, F1, and EQ values all improve, especially on the RTE, where a clear positive correlation is observed. However, we found some exceptions on the Overruling and Stance4Trump, and these variations may reflect the influence of specific task characteristics and dataset properties on the performance of DET.

To compare the proposed dialectical explanations with the traditional Chain-of-Thought (COT), we conducted experiments using GPT-4, GPT-4o-m, and Llama 3.1 on three datasets. As shown in the Table 8, COT can help enhance the model's classification performance, but our method is still the most optimal. Additionally, for the parameter and cost analysis of DET, please refer to Figures 7 and 8 in Appendix C.4, respectively.

4.6 Efficiency Comparison

We measured the average time required to generate explanations and perform training, as shown in Table 9. It can be seen that although the average runtime of our method is relatively longer, its performance surpasses that of methods that directly use LLMs, and the average runtime of our method is still shorter than that of methods that do not directly use LLMs. To further reduce the time and cost required for generating explanations, we can apply dialectical prompting only to ambiguous or unclear cases.

4.7 Case Studies

To further illustrate the effectiveness of DET, consider the input text "All mental illnesses are awful and must be treated" from the Ethos dataset. Using our Dialectical Prompting Strategy, we generated explanations for multiple potential labels. DET predicted it as "hate speech" and provided the explanation: "The phrase 'all mental illnesses are

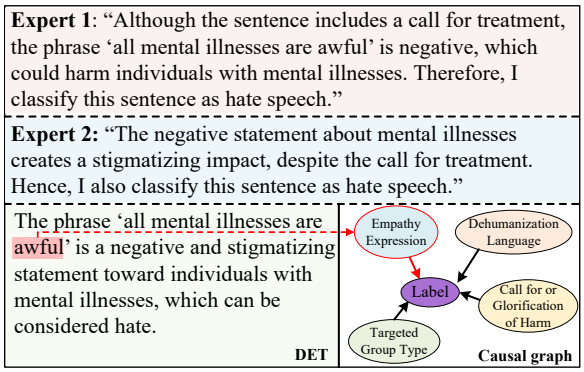


Figure 5: Classification explanation verification.

awful' is a negative and stigmatizing statement toward individuals with mental illnesses, which can be considered hate."

To validate the effectiveness and credibility of DET-generated explanations, we invited two human experts to interpret and classify the same example. As shown in Figure 5, the human experts agreed that the phrase "all mental illnesses are awful" is stigmatizing and potentially harmful, despite the call for treatment. They classified the sentence as hate speech, providing detailed reasoning similar to that of the DET method. Additionally, we invited other experts to evaluate the reasoning processes of both human and machine explanations. Finally, we employed a causal discovery algorithm COAT (Liu et al., 2024a) to derive a directed acyclic graph, which revealed a causal relationship between the word awful and the text label, further validating the effectiveness of our method. Detailed information can be found in the Appendix C.5.

5 Conclusion

In this paper, we propose a novel Dialectical Explanation Training method that combines Dialectical Prompting, Explanation-Guided Training, and EDL theory to enhance both the performance and explainability of text classification. By generating dialectical explanations and incorporating them directly into the training process, our method effectively tackles challenges such as context-dependent expressions, complex linguistic phenomena, and poor interpretability. Extensive experiments demonstrate that our method achieves significant improvements across multiple metrics. These results underscore the potential of our approach to make text classification more reliable and interpretable, particularly in high-stakes domains such as finance, law, and value alignment, where transparency and robustness are essential.

6 Acknowledgements

This research is partially supported by funding from Xiangjiang Laboratory (25XJ02002), grant from the National Natural Science Foundation of China (72495125, 62376227, 62376228), the science and technology innovation Program of Hunan Province (2024RC4008, AC2024040911247631ff26), and China Postdoctoral Science Foundation (2025M770766). Carl Yang is not supported by any funds from China.

Limitations

In this study, we propose DET to improve the performance and explainability of text classification. However, we also identify several limitations of DET: (1) DET still requires access to labeled data for text classification tasks in order to learn model parameters. In some real-world scenarios, label information may be restricted or unavailable, and this dependency limits its practicality. (2) Whether the fine-tuning method based on dialectical explanations is effective in LLMs training or fine-tuning remains to be further explored. (3) The generalization ability of DET still needs improvement. In future work, we will incorporate causal inference or stable learning to alleviate this issue.

Ethics Consideration

A potential issue is data leakage, where some test samples may overlap with the training data. This concern is particularly relevant when using and evaluating LLMs, as they are typically pre-trained on large-scale corpora spanning multiple domains. However, in our experiments, we use a smaller model, DeBERTa, instead of directly employing LLMs for text classification, thereby avoiding the risk of data leakage.

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. 2020. Deep evidential regression. In *Advances in neural information processing systems*, volume 33, pages 14927–14937.

Anthropic. 2024. Claude API models. https://docs.anthropic.com/en/docs/about-claude/models. Accessed: 2024-08-07.

Wentao Bao, Qi Yu, and Yu Kong. 2021. Evidential deep learning for open set action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13349–13358.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson Correlation Coefficient. In Israel Cohen, Yiteng Huang, Jingdong Chen, and Jacob Benesty, editors, *Noise Reduction in Speech Processing*, pages 1–4. Springer, Berlin, Heidelberg.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Towards Llm-Guided Causal Explainability for Black-Box Text Classifiers. In *AAAI 2024 Workshop on Responsible Language Models, Vancouver, BC, Canada*.

Martin Juan José Bucher and Marco Martini. 2024. Fine-Tuned 'Small' Llms (still) Significantly Outperform Zero-Shot Generative Ai Models in Text Classification. *Preprint*, arXiv:2406.08660.

Tyler A. Chang and Benjamin K. Bergen. 2024. Language Model Behavior: A Comprehensive Survey. *Computational Linguistics*, 50(1):293–350.

Ding Chen, Shichao Song, Qingchen Yu, Zhiyu Li, Wenjin Wang, Feiyu Xiong, and Bo Tang. 2024. Grimoire Is All You Need for Enhancing Large Language Models. *Preprint*, arXiv:2401.03385.

Zichen Chen, Jianda Chen, Mitali Gaidhani, Ambuj Singh, and Misha Sra. 2023. Xplainllm: A Qa Explanation Dataset for Understanding Llm Decision-Making. *Preprint*, arXiv:2311.08614.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The Pascal Recognising Textual Entailment Challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Huaming Du, Yujia Zheng, Baoyu Jing, Yu Zhao, Gang Kou, Guisong Liu, Tao Gu, Weimin Li, and Carl Yang. 2025. Causal discovery through synergizing large language model and data-driven reasoning. In *Proceedings of the KDD*, pages 543–554.

Jack Fitzgerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, and 1 others. 2023. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the ACL*, pages 4277–4302.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024a. Socreval: Large Language Models with the Socratic Method for Reference-Free Reasoning Evaluation. In *Findings of the NAACL*, pages 2736–2764. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-Enhanced Bert with Disentangled Attention. In *ICLR*.

Xiaoxin He, Xavier Bresson, Thomas Laurent, Adam Perold, Yann LeCun, and Bryan Hooi. 2024b. Harnessing Explanations: Llm-to-Lm Interpreter for Enhanced Text-Attributed Graph Representation Learning. In *ICLR*.

Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable Ai: Challenges and Prospects.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of ACL*, pages 8003–8017.

Linmei Hu, Hongyu He, Duokang Wang, Ziwang Zhao, Yingxia Shao, and Liqiang Nie. 2024. Llm Vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model. *Proceedings of AAAI*, 38(16):18234–18242.

Audun Josang. 2016. *Subjective logic*, volume 3. Springer.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of NAACL*, pages 4725–4735.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. 2024. Learning to Generate Explainable Stock Predictions Using Self-Reflective Large Language Models. In *Proceedings of WWW*, pages 4304–4315, Singapore Singapore. ACM.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. 2023. Post Hoc Explanations of Language Models Can Improve Language Models. In *Advances in Neural Information Processing Systems*, pages 65468–65483.

Taeyoon Kwon, Kai Tzu-iunn Ong, Dongjin Kang, Seungjun Moon, Jeong Ryong Lee, Dosik Hwang, Beomseok Sohn, Yongsik Sim, Dongha Lee, and Jinyoung Yeo. 2024. Large Language Models Are Clinical Reasoners: Reasoning-Aware Diagnosis Framework with Prompt-Generated Rationales. *Proceedings of AAAI*, 38(16):18417–18425.

Jean Lee, Hoyoul Luis Youn, Josiah Poon, and Soyeon Caren Han. 2023. Stockemotions: Discover Investor Emotions for Financial Sentiment Analysis and Multivariate Time Series. *Preprint*, arXiv:2301.09279.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. Bertgcn: Transductive text classification by combining gnn

and bert. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462.

Chenxi Liu, Yongqiang Chen, Tongliang Liu, Mingming Gong, James Cheng, Bo Han, and Kun Zhang. 2024a. Discovery of the hidden world with large language models. In *Proceedings of NeurIPS*.

Chun Liu, Hongguang Zhang, Kainan Zhao, Xinghai Ju, and Lin Yang. 2024b. Llmembed: Rethinking lightweight llm's genuine function in text classification. In *ACL*.

Weigang Lu, Yibing Zhan, Binbin Lin, Ziyu Guan, Liu Liu, Baosheng Yu, Wei Zhao, Yaming Yang, and Dacheng Tao. 2024. Skipnode: On alleviating performance degradation for deep graph convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):7030–7043.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: An Online Hate Speech Detection Dataset. *Complex & Intelligent Systems*, 8(6):4663–4678.

OpenAI. 2024. OpenAI API models. `https://platform.openai.com/docs/models`. Accessed: 2024-08-07.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. ValueNet: A New Dataset for Human Value Driven Dialogue System. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of KDD*, pages 1135–1144.

Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence*, 1(5):206–215.

Arshdeep Sekhon, Hanjie Chen, Aman Shrivastava, Zhe Wang, Yangfeng Ji, and Yanjun Qi. 2023. Improving Interpretability Via Explicit Word Interaction Graph Layer. *Proceedings of AAAI*, 37(11):13528–13537.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Proceedings of NeurIPS*, 31.

Kari Sentz and Scott Ferson. 2002. *Combination of evidence in Dempster-Shafer theory*. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); Sandia.

Ava P Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N Bhatia, and Connor W Coley. 2021. Evidential deep learning for guided molecular property prediction and discovery. *ACS central science*, 7(8):1356–1367.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023a. Text Classification Via Large Language Models. In *Findings of EMNLP*, pages 8990–9005, Singapore.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023b. Text classification via large language models. In *Findings of the EMNLP*, pages 8990–9005.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proceedings of the NeurIPS*, volume 35, pages 24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. In *Proceedings of the NeurIPS*, volume 36.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Proceedings of NeurIPS*, 28.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of WWW*, pages 2032–2042.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Xujiang Zhao, Feng Chen, Shu Hu, and Jin-Hee Cho. 2020. Uncertainty aware semi-supervised learning on graph data. *Proceedings of NeurIPS*, 33:12827–12836.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1):237–291.

A Related Work

A.0.1 LLM-Generated Explanations

Recent advancements have focused on using LLM-generated rationales to enhance model performance and interpretability (Kwon et al., 2024; Bhattacharjee et al., 2024). LLMs can explain their predictions by generating high-quality explanations, which improve few-shot or zero-shot performance when used to augment input prompts (Wei et al., 2022). These explanations have also been used as additional information to "self-improve" LLMs (Krishna et al., 2023; He et al., 2024a). However, the large size of LLMs limits their utility in many applications. To address this, generated rationales can be leveraged as informative supervision to train smaller, task-specific models that can be deployed with lower computation and memory costs (Hsieh et al., 2023; Koa et al., 2024; Hu et al., 2024; He et al., 2024b). Our method integrates explanation generation with prediction using dialectical prompting and explanation-guided training. This improves accuracy, provides multi-angle explanations, and addresses significant gaps in current literature.

A.0.2 Evidential Deep Learning

To accurately quantify the uncertainty of the model, recent EDL (Sensoy et al., 2018; Amini et al., 2020) is developed from the evidence framework of Dempster-Shafer Theory (DST) (Sentz and Ferson, 2002) and the subjective logic (SL) (Josang, 2016). For a K-class classification problem, the EDL treats the input x as a proposition and regards the classification task as to give a multinomial subjective opinion in a K-dimensional domain $\{1, ..., K\}$. The subjective opinion is expressed as a triplet $\omega = (b, u, a)$, where $b = \{b_1, ..., b_K\}$ is the belief mass, u represents the uncertainty, and $\mathbf{a} = \{a_1, ..., a_K\}$ is the base rate distribution. For any $k \in [1, 2, ..., K]$, the probability mass of a multinomial opinion is defined as:

$$p_k = b_k + a_k u \ . \tag{8}$$

To enable the probability meaning of p_k, i.e., $\sum_k p_k = 1$, the base rate a_k is typically set to $1/K$ and the subjective opinion is constrained by

$$u + \sum_{k=1}^{K} b_k = 1 \ . \tag{9}$$

Besides, for a K-class setting, the probability mass $\mathbf{p} = \{p_1, p_2, \cdots, p_K\}$ is assumed to follow a Dirichlet distribution parameterised by a K-dimensional Dirichlet strength vector $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_K\}$:

$$\mathrm{Dir}\left(\mathbf{P}|\alpha\right) = \begin{cases} \frac{1}{B(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, & \text{for } \mathbf{p} \in S_k \\ 0, & \text{otherwise} \ , \end{cases} \tag{10}$$

where $B\left(\alpha\right)$ is a K-dimensional Beta function, S_k is a K-dimensional unit simplex. The total strength of the Dirichlet is defined as $\mathcal{S} = \sum_{k=1}^{K} \alpha_k$. Note that for the special case when $K = 2$, the Dirichlet distribution reduces to a Beta distribution and a binomial subjective opinion will be formulated in this case.

According to the evidence theory, the term evidence is introduced to describe the amount of supporting observations for classifying the data x into a class. Let $\mathbf{e} = \{e_1, \cdots, e_K\}$ be the evidence for K classes. Each entry $e_k \geqslant 0$ and the Dirichlet strength α are linked according to the evidence theory by the following identity:

$$\alpha = \mathbf{e} + \mathbf{a}W \ , \tag{11}$$

where W is the weight of uncertain evidence. With the Dirichlet assumption, the expectation of the multinomial probability $\mathbf{P}$ is given by

$$\mathbb{E}\left(p_k\right) = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k} = \frac{e_k + \alpha_k W}{w + \sum_{k=1}^{K} e_k} \ , \tag{12}$$

With loss of generality, the weight W is set to K and considering the assumption of the subjective opinion constraint in Eq. 9 that $\alpha_k = 1/K$, we have the Dirichlet strength $\alpha_k = e_k + 1$ according to Eq. 11. In this way, the Dirichlet evidence can be mapped to the subjective opinion by setting the following equality's:

$$b_k = \frac{e_k}{\mathcal{S}} \quad \text{and} \quad u = \frac{K}{\mathcal{S}} \ . \tag{13}$$

Therefore, we can see that if the evidence e_k for the k-th class is predicted, the corresponding expected class probability in Eq. 8 (or Eq. 12) can be rewritten as $p_k = \alpha_k / \mathcal{S}$. From Eq. 13, it is clear that the predictive uncertainty u can be determined after α_k is obtained.

B Derivation of the Loss Function

Inspired by the idea of DEL, we leverage the DeBERTa-base model in DET to directly predict

the evidence e_i from the given input x_i for a K-class classification problem. In particular, the output of the DeBERTa-base model is activated by a non-negative evidence function. Considering the Dirichlet prior, the DeBERTa-base model is trained by minimizing the negative log-likelihood (NLL) loss [3]:

$$\begin{aligned}\mathcal{L}_{nll-edl,i}\left(\mathrm{y}, e; \theta\right) &= -\log\left(\int \prod_{k=1}^{K} p_{ik}^{y_{ik}} \frac{1}{B\left(\alpha_i\right)} \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}-1} d\mathbf{P}_i\right) \\ &= -\log\left[\frac{1}{B\left(\alpha_i\right)} \int \prod_{k=1}^{K} p_{ik}^{\alpha_{ik}+y_{ik}-1} d\mathbf{P}_i\right] \\ &= -\log\left[\frac{1}{B\left(\alpha_i\right)} B\left(\alpha_i + \mathbf{y}_i\right)\right] \\ &= -\log\left[\frac{B\left(\alpha_i\right)}{B\left(\alpha_i + \mathbf{y}_i\right)}\right] .\end{aligned} \tag{14}$$

Also note that:

$$B\left(\alpha_i\right) = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_{ik}\right)} = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\mathcal{S}_i\right)} , \tag{15}$$

where $\Gamma\left(\cdot\right)$ is the gamma function.

$$B\left(\alpha_i + \mathbf{y}_i\right) = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik} + y_{ik}\right)}{\Gamma\left(\sum_{k=1}^{K} \alpha_{ik} + y_{ik}\right)} = \frac{\prod_{k=1}^{K} \Gamma\left(\alpha_{ik} + y_{ik}\right)}{\Gamma\left(\mathcal{S}_i + 1\right)} . \tag{16}$$

Combining Eq. 15 and 16, we obtain:

$$\frac{B\left(\alpha_i\right)}{B\left(\alpha_i + \mathbf{y}_i\right)} = \frac{\Gamma\left(\mathcal{S}_i + 1\right)}{\Gamma\left(\mathcal{S}_i\right)} \prod_{k=1}^{K} \frac{\Gamma\left(\alpha_{ik}\right)}{\Gamma\left(\alpha_{ik} + 1\right)} = \prod_{k=1}^{K} \frac{\mathcal{S}_i}{\alpha_{ik}} = \prod_{k=1}^{K} \left(\frac{\mathcal{S}_i}{\alpha_{ik}}\right)^{y_{ik}} . \tag{17}$$

By combining Eq. 14 and 17, we get:

$$\begin{aligned}\mathcal{L}_{nll-edl,i}\left(\mathrm{y}, e; \theta\right) &= -\log\left[\frac{B\left(\alpha_i\right)}{B\left(\alpha_i + \mathbf{y}_i\right)}\right] \\ &= \textstyle\sum_{k=1}^{K} y_{ik} \left(\log\left(\mathcal{S}_i\right) - \log\left(e_{ik} + 1\right)\right) ,\end{aligned} \tag{18}$$

where $\mathrm{y}_i = \{y_{i1}, \cdots, y_{iK}\}$ is an one-hot K-dimensional label for sample x_i and e_i can be expressed as $e_i = g\left(\mathcal{F}\left(x_i; \theta\right)\right)$. Here, $\mathcal{F}$ is the DeBERTa-base model parameterized by θ and g is the evidence function such as exp, softplus, or ReLU.

Finally, the evidence of non-target classes is suppressed by minimizing the KL divergence between the modified Dirichlet distribution and the uniform distribution. Specifically, the regularization term has the following form:

$$\mathcal{L}_{kl} = \mathrm{KL}\left(\mathrm{Dir}\left(\mathbf{p}, \tilde{\alpha}_{x_i}\right), \mathrm{Dir}\left(\mathbf{p}, \mathbf{1}\right)\right) , \tag{19}$$

where $\mathrm{Dir}\left(\mathbf{p}, \mathbf{1}\right)$ is the uniform Dirichlet distribution, $\tilde{\alpha}_{x_i} = y + (1 - y) \odot \alpha_{x_i}$ is the Dirichlet parameter for sample x_i after removing non-misleading evidence from the predicted parameters, and $\odot$ represents the Hadamard product. Therefore, the overall loss function is as follows:

$$\begin{aligned}\mathcal{L}\left(\theta\right) = &\sum_{i=1}^{m} \mathcal{L}_{nll-edl,i}\left(y, e; \theta\right) \\ &+ \lambda_t \sum_{i=1}^{m} KL\left[\mathrm{Dir}\left(\mathbf{p_i} \mid \tilde{\alpha}_{x_i}\right) \| \mathrm{Dir}\left(\mathbf{p_i} \mid \langle 1, \cdots, 1 \rangle\right)\right] ,\end{aligned} \tag{20}$$

where $\lambda_t = \min\left(1.0, t/10\right) \in [0, 1]$ is the annealing coefficient, and t is the index of the current training epoch.

C More Details about Experiments

In this section, we provide more details of experiment setup for the reproducibility of the experiment results.

C.1 Datasets

To ensure a thorough evaluation of our method's ability to handle text classification challenges across different high-stakes application domains, such as finance, law, and value alignment, we selected eight representative datasets from high-stakes domains: StockEmotions (Lee et al., 2023) for finance, which contains text data related to stock market emotions; Overruling (Zheng et al., 2021) for law, focusing on the classification of judicial opinions to determine whether they have been overruled; Stance4Trump (Kawintiranon and Singh, 2021) for politics, which assesses the stance classification regarding support for or against former President Trump; Ethos (Mollas et al., 2022) for hate speech detection; RTE (Dagan et al., 2006) for natural language inference tasks to determine whether a given hypothesis logically follows from a given premise; ValueNet (Qiu et al., 2022) for value alignment, where sentences are classified based on their alignment with a stated value; AGNews (Zhang et al., 2015) consists of 4 types of news articles from the AG's corpus. The dataset contains 120000 training and 7600 testing examples, and the max length of words is 177. These datasets were chosen to provide a comprehensive evaluation

[3] Please note that, similar to existing research (Bao et al., 2021), we choose the NLL loss as the loss function. However, in other scenarios, the determination of $\mathcal{L}_{edl}$ should be flexibly chosen based on the specific task to achieve optimal model performance.

of our method's performance across high-stakes application domains; and MASSIVE (Fitzgerald et al., 2023) contains 1 million realistic, multilingual parallel-labeled virtual assistant utterances, covering 51 languages, 18 domains, 60 intents, and 55 slots.

C.2 Evaluation Materials

Return a JSON array with these fields ({fields}) as the answer. You are an expert in NLP and linguistic analysis. Please evaluate the explanation of every sample for a text classification task in the JSON file on a scale of 1-5 (1 being the lowest, 5 being the highest) based on the following criteria. Additionally, provide detailed justifications referencing specific aspects of the explanation and the text for each score.

C.2.1 Clarity.

- **Score 1:** The explanation is very unclear or difficult to understand.
- **Score 2:** The explanation is unclear in many parts and needs significant improvement.
- **Score 3:** The explanation is mostly clear but has some areas that need improvement.
- **Score 4:** The explanation is clear but could benefit from minor enhancements.
- **Score 5:** The explanation is very clear, with no ambiguity.

Guiding Questions:

- Does the explanation use language that is easy to understand?
- Does the explanation avoid complex terms or convoluted sentence structures?
- Are there specific examples or wording that make the explanation clearer?

C.2.2 Relevance.

- **Score 1:** The explanation is completely unrelated to the given text and classification task.
- **Score 2:** The explanation has limited relevance to the text and task, with many unrelated parts.
- **Score 3:** The explanation is mostly relevant, but some parts are not pertinent.
- **Score 4:** The explanation is relevant, with only minor deviations.
- **Score 5:** The explanation is entirely relevant, closely aligned with the text and classification task.

Guiding Questions:

- Does the explanation directly address the core issue of the classification task?
- Is the explanation closely related to the content of the text?
- Are there parts of the explanation that deviate from the text or task?

C.2.3 Completeness.

- **Score 1:** The explanation is very incomplete, missing critical factors.
- **Score 2:** The explanation is incomplete, missing several important details.
- **Score 3:** The explanation is mostly complete but lacks some details.
- **Score 4:** The explanation is complete, with only minor omissions.
- **Score 5:** The explanation is very complete, covering all critical factors with no omissions.

Guiding Questions:

- Does the explanation cover all the critical factors influencing the classification decision?
- Are any important details or factors missing from the explanation?
- Does the explanation provide sufficient background information to support the classification decision?

C.2.4 Consistency.

- **Score 1:** The explanation is completely inconsistent with the classification result.
- **Score 2:** The explanation has several inconsistencies with the classification result.
- **Score 3:** The explanation has some consistency with the classification result, but not fully.

- **Score 4:** The explanation is consistent with minor inconsistencies.
- **Score 5:** The explanation is entirely consistent with the classification result.

Guiding Questions:

- Is the explanation consistent with the classification result?
- Are there any parts of the explanation that contradict the classification result?
- Does the explanation provide enough evidence to support the classification result?

C.2.5 Credibility.

- **Score 1:** The explanation is not credible or contains obvious errors.
- **Score 2:** The explanation has low credibility with several doubtful elements.
- **Score 3:** The explanation is somewhat credible but contains elements that are doubtful.
- **Score 4:** The explanation is credible with minor issues.
- **Score 5:** The explanation is credible and aligns with known information and logic.

Guiding Questions:

- Is the explanation logical and credible?
- Does the explanation contain any obvious errors or false information?
- Are there any parts of the explanation that seem suspicious or unreliable?

Text: {original_text}
Model's Classification: {pred_label}
Model's Explanation: {explanation}

C.3 Evaluation Metrics

Similar to previous methods (Lin et al., 2021; Liu et al., 2024b), we use ACC to evaluate classification performance. To assess the quality of the models' generated explanations for predicted labels, we have designed a custom Explanation Quality (EQ) metric, inspired by (Hoffman et al., 2018; Chen et al., 2023), which demonstrated the results of using LLMs to evaluate explanation quality are highly correlated with human evaluations, proving the effectiveness of this method.

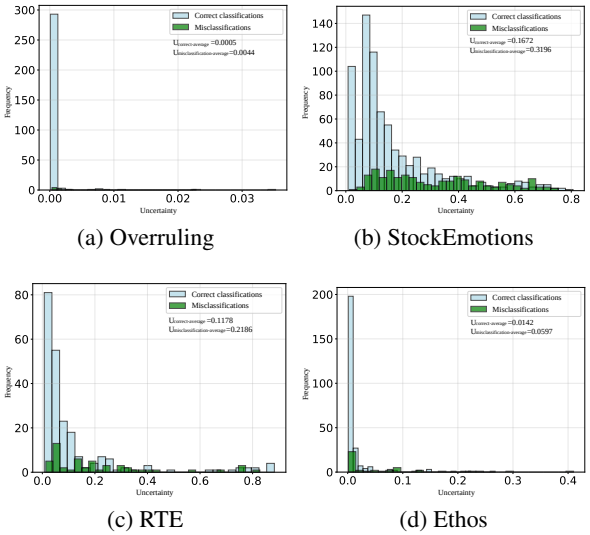
Our evaluation method employs two state-of-the-art LLMs: GPT-4o (OpenAI, 2024) and Claude 3.5-sonnet (Anthropic, 2024). And we assess explanations based on five key dimensions: clarity, relevance, completeness, consistency, and credibility, using a 5-point scale. These models were chosen for their capabilities in understanding complex linguistic phenomena and human-like reasoning. Appendix A contains the questions and instructions for instructing the LLMs on how to evaluate explanations.

The evaluation process involves assessing all samples in the test datasets. Each LLM independently evaluates all samples. Each sample's final EQ score is calculated as the average of its five-dimensional scores. Due to the presence of hate speech in the Ethos dataset and model safety usage restrictions, Claude 3.5-sonnet was unable to produce evaluation results for more than half of the samples in the Ethos dataset. Therefore, for Ethos dataset, we used EQ score from GPT-4o as the final score. For other datasets, we used the average of the GPT-4o and Claude 3.5-sonnet evaluations.

To validate this approach, we compared LLM evaluations with those of five human experts on a subset of 120 explanations, selecting 20 samples from each of the six datasets. The evaluation results were compared with those from GPT-4o (120 complete samples) and Claude 3.5-sonnet (100 samples excluding the Ethos dataset) using Pearson correlation coefficient (Benesty et al., 2009) and Krippendorff's alpha (Hayes and Krippendorff, 2007). Instead of comparing with multiple experts, which can introduce variability, we used the evaluation from the expert with the highest average consistency with the others. This method involved calculating Krippendorff's alpha to determine each expert's average consistency and selecting the most consistent one. This ensured a robust comparison with the most representative expert evaluation. The final results are shown in Table 2. The results show that the LLM evaluations and human expert evaluations are very relevant and consistent. This proves that our designed EQ metric is valid and reliable.

C.4 More Experimental Results

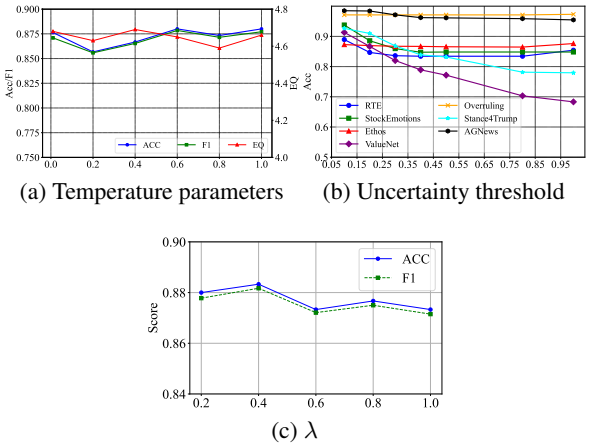
The complete experimental results with mean and variance are shown in Table 10.



(a) Overruling (b) StockEmotions

(c) RTE (d) Ethos

Figure 6: The uncertainty distribution across different datasets.



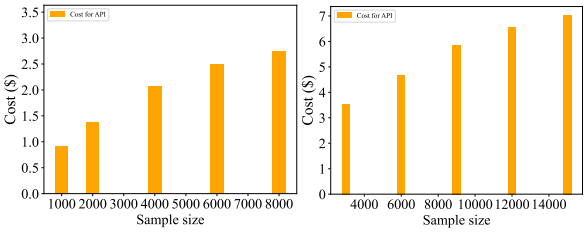
(a) Temperature parameters (b) Uncertainty threshold

(c) λ

Figure 7: Model performance under different parameter settings on the Ethos dataset.

Uncertainty Analysis. The uncertainty distributions of DET on the Overruling, StockEmotions, Stance4Trump, and ValueNet datasets are shown in Figure 6. Overall, correctly classified samples exhibit lower uncertainty, while DET assigns higher uncertainty values to misclassified samples. And uncertainty is higher on StockEmotions and values. In this scenario, introducing the EDL theory not only allows for measuring the classification uncertainty of each sample but also significantly enhances the model's classification performance.

Parameter Analysis . We mainly analyzed the impact of the temperature coefficient, uncertainty threshold, and λ in Equation 5 on model performance. As shown in Figure 7(a) and (c), the temperature parameter of large language models and λ have a relatively minor impact on model performance. In contrast, Figure 7(b) illustrates that accuracy increases when the uncertainty of predictions falls below a certain threshold.



(a) StockEmotions (b) ValuesNet

Figure 8: The cost statistics using our model with different sample size on StockEmotions and ValueNet datasets.

Cost Analysis. We reported the costs of our method with varying sample sizes, as shown in Figure 8. We can conclude that the costs produced by DET do not exhibit a linear relationship with increasing sample sizes, indicating good economic efficiency.

Expert 3: "The explanations generated by the DET method are very detailed and help me understand why the sentence is classified as hate speech. The multi-perspective analysis is particularly valuable."

Expert 4: "The machine-generated explanations are highly consistent with human reasoning. The DET method effectively understands complex contexts."

Expert 5: "The DET method has the ability to summarize and extract important evidence."

Figure 9: Expert evaluations on reasoning processes of both human and machine explanations.

C.5 More Evaluations in Case Study

As shown in Figure 9, the evaluating experts found that the explanations generated by the DET method exhibit two important properties: consistency with human reasoning and selective summarization of important evidence. The explanations align with how human experts comprehensively consider various factors in their judgments. For example, while the traditional classifier failed to recognize the neg-

Methods	Finance			Law			Politics			Hate			NLI			Values			News		
	StockEmotions			Overruling			Stance4Trump			Ethos			RTE			ValueNet			AGNews		
	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ	ACC	F1	EQ
Models With No Explanation																					
GPT-3.5 (ZSL)	71.81±0.05	70.14±0.05	-	90.85±0.04	86.94±0.09	-	70.61±0.02	69.89±0.11	-	83.61±0.07	82.67±0.09	-	49.46±0.02	28.16±0.06	-	61.20±0.08	55.62±0.05	-	86.23±0.10	87.23±0.09	-
GPT-3.5 (FSL, k=4)	83.27±0.09	80.69±0.06	-	92.81±0.11	90.67±0.13	-	75.77±0.06	73.45±0.08	-	84.93±0.08	82.21±0.11	-	73.01±0.07	58.67±0.12	-	55.29±0.08	48.29±0.14	-	88.70±0.15	88.68±0.10	-
GPT-4o-m (ZSL)	79.00±0.05	76.31±0.10	-	95.51±0.09	94.55±0.15	-	76.05±0.07	76.13±0.16	-	85.33±0.10	85.00±0.07	-	49.10±0.08	29.61±0.17	-	57.34±0.07	41.80±0.11	-	83.43±0.14	83.00±0.18	-
GPT-4o-m (FSL, k=4)	79.70±0.11	78.50±0.14	-	94.55±0.18	93.26±0.14	-	76.81±0.11	75.68±0.18	-	86.50±0.12	86.66±0.13	-	71.84±0.08	40.64±0.13	-	60.57±0.11	49.67±0.14	-	87.70±0.16	87.44±0.19	-
o1-mini (ZSL)	83.10±0.09	80.00±0.12	-	91.99±0.13	94.40±0.08	-	73.59±0.14	74.81±0.09	-	86.00±0.15	85.92±0.10	-	49.82±0.17	31.51±0.13	-	61.27±0.16	59.85±0.10	-	85.41±0.23	84.69±0.25	-
o1-mini (FSL, k=4)	82.90±0.12	78.92±0.17	-	92.31±0.14	92.29±0.09	-	76.81±0.08	75.79±0.12	-	85.77±0.13	85.40±0.18	-	58.12±0.12	31.22±0.19	-	62.59±0.25	60.13±0.19	-	86.97±0.16	86.32±0.20	-
DeepSeek-R1 (ZSL)	78.60±0.09	80.79±0.13	-	93.91±0.16	93.57±0.19	-	77.51±0.24	76.31±0.26	-	85.00±0.13	84.98±0.15	-	48.01±0.16	30.47±0.19	-	59.96±0.12	41.92±0.18	-	86.60±0.19	86.56±0.24	-
DeepSeek-R1 (FSL, k=4)	81.00±0.17	79.26±0.11	-	90.38±0.19	90.35±0.07	-	77.75±0.09	**77.64±0.13**	-	85.33±0.15	85.25±0.11	-	49.10±0.14	30.94±0.08	-	60.24±0.16	57.67±0.10	-	88.16±0.19	86.38±0.23	-
DeBERTa	77.93±0.06	75.47±0.13	-	96.69±0.11	96.11±0.15	-	73.76±0.13	73.19±0.21	-	81.22±0.17	78.56±0.20	-	76.41±0.14	72.29±0.09	-	66.03±0.21	62.21±0.16	-	85.13±0.20	83.61±0.27	-
CARP	74.25±0.20	72.18±0.23	-	94.36±0.17	91.04±0.12	-	70.46±0.11	69.81±0.09	-	79.61±0.18	78.23±0.14	-	60.54±0.17	60.25±0.13	-	65.43±0.19	63.62±0.22	-	83.29±0.19	82.76±0.24	-
LLMEmbed	77.93±0.09	76.34±0.08	-	96.74±0.13	94.85±0.15	-	71.11±0.14	70.28±0.12	-	81.22±0.09	79.65±0.11	-	62.71±0.12	60.47±0.14	-	66.03±0.09	63.79±0.14	-	90.25±0.19	88.24±0.25	-
Models With Explanations																					
GPT-3.5	68.20±0.09	69.50±0.07	4.17±0.05	83.33±0.11	78.90±0.14	3.67±0.09	63.50±0.17	59.60±0.14	3.69±0.08	73.00±0.17	69.78±0.19	3.96±0.11	47.29±0.19	28.02±0.22	4.05±0.13	55.14±0.19	49.05±0.27	3.95±0.16	87.00±0.15	88.52±0.19	3.87±0.12
LLAMA 3.1-8B	70.46±0.10	69.84±0.13	4.13±0.09	89.83±0.16	89.74±0.18	4.31±0.12	56.27±0.09	55.06±0.13	3.86±0.11	74.41±0.10	53.62±0.13	2.08±0.06	50.90±0.14	26.18±0.09	4.17±0.05	55.43±0.16	45.60±0.20	3.14±0.14	12.67±0.11	45.25±0.16	3.33±0.14
GPT-4	68.20±0.10	68.05±0.13	4.17±0.05	93.35±0.16	93.44±0.12	4.34±0.08	74.43±0.16	72.66±0.18	3.90±0.10	84.67±0.12	85.81±0.07	4.24±0.06	48.01±0.15	30.26±0.10	4.31±0.09	55.14±0.17	52.36±0.10	3.95±0.09	88.57±0.20	87.94±0.13	4.65±0.19
GPT-4o-m	75.10±0.11	74.31±0.13	4.20±0.05	90.70±0.13	90.31±0.09	4.12±0.12	68.44±0.10	67.42±0.16	3.75±0.11	85.00±0.13	83.31±0.18	4.05±0.09	46.93±0.07	30.07±0.12	4.18±0.10	58.00±0.13	46.00±0.11	4.06±0.17	87.60±0.17	87.65±0.21	4.37±0.09
o1-mini	80.37±0.14	78.11±0.12	4.20±0.07	94.53±0.16	94.53±0.10	4.12±0.11	75.20±0.17	74.23±0.13	3.75±0.12	85.58±0.18	84.51±0.15	4.05±0.12	49.82±0.11	31.29±0.12	4.18±0.08	58.00±0.07	57.64±0.11	4.06±0.13	87.41±0.22	86.38±0.18	4.50±0.11
DeepSeek-R1	78.60±0.13	79.03±0.17	4.25±0.10	95.19±0.12	95.19±0.09	4.21±0.08	77.57±0.16	76.38±0.19	4.06±0.13	86.33±0.16	86.33±0.11	4.72±0.12	49.82±0.14	29.68±0.12	4.35±0.08	58.61±0.09	58.36±0.12	4.66±0.07	83.77±0.16	84.09±0.19	4.32±0.10
DeBERTa + LIME	77.93±0.19	75.47±0.12	1.95±0.13	96.69±0.11	96.11±0.16	2.00±0.10	73.76±0.13	73.19±0.14	1.85±0.08	81.22±0.16	78.56±0.19	1.90±0.06	76.41±0.13	72.29±0.09	1.80±0.08	66.03±0.10	62.21±0.12	1.87±0.13	85.13±0.18	83.61±0.24	1.84±0.17
DeBERTa + SHAP	77.93±0.19	75.47±0.23	2.10±0.14	96.69±0.17	96.11±0.12	2.15±0.14	73.76±0.16	73.19±0.20	2.05±0.11	81.22±0.13±0.11	78.56±0.14	2.10±0.12	76.41±0.10	72.29±0.08	2.00±0.07	66.03±0.13	62.21±0.10	2.08±0.12	85.13±0.18	83.61±0.11	2.37±0.16
DET (our)	**84.83**±0.03	**82.36**±0.07	**4.61**±0.06	**97.31**±0.11	**97.25**±0.09	**4.55**±0.10	**79.95**±0.09	76.17±0.15	**4.26**±0.08	**87.63**±0.13	**87.09**±0.17	**4.68**±0.10	**85.43**±0.11	**81.87**±0.16	**4.66**±0.10	**68.32**±0.13	**66.23**±0.12	**4.23**±0.09	**95.46**±0.17	**93.16**±0.19	**4.85**±0.14

Table 10: Main results: performance comparison across datasets. Note: GPT-3.5 = GPT-3.5-turbo; GPT-4o-m = GPT-4o-mini. Bold numbers indicate the best results.

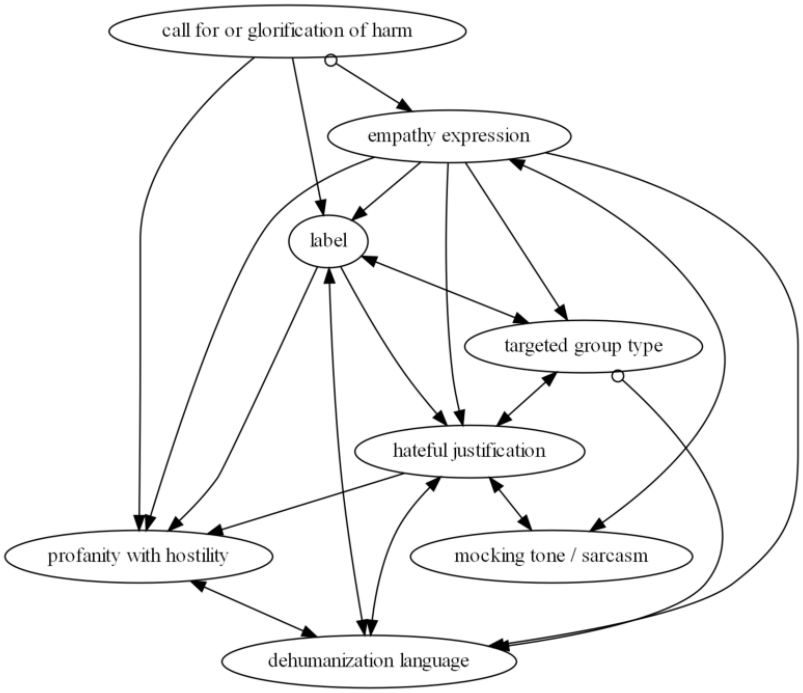


Figure 10: The complete causal graph for the Ethos dataset.

ative impact of the phrase "all mental illnesses are awful," the DET method identified and balanced these harmful implications, mirroring human reasoning. Furthermore, the DET method effectively summarizes and highlights critical information, which enhances the transparency and reliability of classification decisions. These properties make the DET method particularly suitable for high-stakes applications, demonstrating significant potential for practical use. Expert reviews confirmed that the method works and is useful, showing that DET-generated explanations are a strong and trustworthy foundation for difficult text classification tasks.

As shown in Figure 10, we obtain a complete causal graph using the causal discovery algorithm COAT. We find that *Empathy Expression*, *Dehumanization Language*, *Targeted Group Type*, and *Call for or Glorification of Harm* jointly determine the classification label of the text. In the case analysis, the term "awful" in the text corresponds to *Empathy Expression*, which further demonstrates the effectiveness of DET and the explanation metrics we proposed.