Towards Trustworthy Censorship Detection: Evidential Deep Learning and LLM Context in High-Uncertainty NLP Decisions

Joeun Yook
University of Toronto, Engineering Science
joeun.yook@mail.utoronto.ca

Abstract

Rising demand for better NLP models has enabled deeper understanding of human-generated content on social media, particularly for content moderation. As a result, keyword-based censorship is shifting toward capturing subtle semantic context. However, making binary censorship decisions is often oversimplified and less transparent compared to the written content policies users see. In this paper, we apply the Evidential Deep Learning (EDL) framework to replace the classical softmax head with an EDL head to estimate uncertainty and belief about censorship decisions on Twitter content. We focus not only on prediction accuracy but also on model trustworthiness through EDL outputs. We further analyze borderline cases with high uncertainty and investigate how adding LLM-generated context affects both model confidence and correctness.

1 Introduction

Over the past decade, a handful of social media platforms became the cornerstone of online communication, where effective content moderation became inevitable. Challenges in content moderation—such as the opacity of moderation decisions (Loveluck et al., 2022) and the methodological limitations of keyword-based detection that fail to distinguish between mention and usage (Gligorić et al., 2024)—are continuously discussed. These challenges often stem from major platforms like Twitter, Instagram, and Facebook using proprietary algorithms, securing their content moderation strategies and evaluation metrics—making moderation decisions a black box for users.

When it comes to defining the degree of free speech, content moderation becomes very subtle. Yet classical censorship primarily makes binary decisions—whether the content should be censored or not—and automatically deletes or retains the content. This binary result is oversimplified and lacks interpretability and trustworthiness. Although a platform's community guidelines are available as reference, slight differences in tone can change criticism into hate speech and vice versa—especially in borderline cases just between being censored or not. This paper emphasizes the importance of trustworthy censorship decisions made by seemingly black box models, starting from the core skepticism: "Did the model have enough evidence to claim such dichotomous decision?"

Thus, the focus is directed toward integrating the Evidential Deep Learning (EDL) framework into binary content moderation to assess the performance of EDL-head-enabled language models, gain insights into model confidence and decision behavior especially on borderline cases, and observe how adding LLM-generated context at inference affects decisions on high-uncertainty content. Our main contributions include:

- Replacing the standard softmax classification head with a custom EDL head, allowing belief and uncertainty estimation.
- Evaluating encoder-based language model performance using trust–performance metrics and analyzing decision tendencies quantitatively.
- Incorporating LLM-generated explanations at inference on high-uncertainty samples and statistically evaluating their impact, raising discussion on future LLM-guided context integration.

2 Related Work

2.1 Evidential Deep Learning

In recent years, the focus on uncertainty-aware models has grown across high-stakes NLP applications such as content moderation, misinformation detection, and toxic language classification.

While traditional deep neural networks typically use a softmax layer to output point estimates for class probabilities, this approach lacks a mechanism to express uncertainty—an especially critical gap when models make binary decisions that affect freedom of expression. *Evidential Deep Learning (EDL)* (Zhu et al., 2023) is a probabilistic deep learning approach that interprets the categorical predictions of a neural network as a *distribution* over class probabilities by placing a Dirichlet prior upon the class probabilities. This formulation enables the network to quantify both belief and uncertainty for each prediction.

Formally, for a given input $\mathbf{x}$, a standard deep neural network (DNN) would produce logits passed through a softmax function to yield a categorical probability vector $\mathbf{p}$. In contrast, EDL replaces the softmax with an evidential head that outputs a set of non-negative evidence scores $\{e_c\}_{c=1}^{C}$, which are used to parameterize a Dirichlet distribution as $\alpha_c = e_c + 1$. The resulting distribution over class probabilities is given by:

$$\text{Dir}(p \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{c=1}^{C} p_c^{\alpha_c - 1}$$

where $B(\boldsymbol{\alpha})$ is the multivariate beta function that normalizes the distribution, and C is the number of classes. In our work, we focus on a binary classification task ($C = 2$), where the model must decide whether a piece of user-generated content should be censored ($y = 1$) or not ($y = 0$).

The strength of the Dirichlet formulation lies in its ability to produce not only class probabilities but also epistemic uncertainty: belief mass for each class is computed as $b_c = \frac{\alpha_c - 1}{S}$, and overall uncertainty is captured as $u = \frac{C}{S}$, where $S = \sum_{c=1}^{C} \alpha_c$. High uncertainty values signal low model confidence and are especially useful for detecting borderline cases near the decision boundary. By integrating an EDL head into the encoder architecture, our model outputs both a classification and a calibrated uncertainty estimate for every input.

This capability is particularly valuable in content moderation, where edge cases often involve nuanced tone, sarcasm, or implicit intent—contexts where softmax-based models may overcommit to incorrect predictions. Our work builds on the growing body of literature that incorporates uncertainty quantification into NLP systems, but adapts it for binary content decisions where legal and ethical interpretations matter. This application of evidential uncertainty estimation specifically tailored to real-world censorship detection opens a promising path for improving model interpretability and trust in moderation pipelines.

2.2 Content Moderation and Binary Classification

Prior work on online moderation has focused on tasks like hate speech detection (Waseem and Hovy, 2016), political censorship (Elmas et al., 2021), and keyword-based filtering (MacKinnon, 2009). However, such systems often oversimplify decisions and lack trust calibration. Recent work has explored model calibration and uncertainty in moderation settings (Gligorić et al., 2024).

2.3 Uncertainty-Aware Models and Calibration

Trust-aware models in NLP have been explored using Bayesian methods, dropout uncertainty, and more recently Evidential Deep Learning (EDL) (Zhu et al., 2023). While EDL has been applied in computer vision and tabular tasks, its use in binary NLP classification and moderation tasks remains limited. Model calibration metrics such as Expected Calibration Error (ECE) (Guo et al., 2017) have been increasingly adopted in safety-critical NLP domains.

2.4 LLM Explanations at Inference Time

Large language models (LLMs) are increasingly used for post-hoc explanation or context augmentation (Bai et al., 2023). While LLMs have shown promise in rationalizing decisions, their impact on predictive trustworthiness in high-uncertainty moderation tasks remains underexplored. Our work complements this by injecting LLM-generated context at inference time and analyzing its effect on EDL-based belief and uncertainty estimates.

3 Methodology

This section provides a preliminary explanation of the model architecture we adopted, the embedding and classifier pipeline, and the method for injecting LLM-guided tweet context. It is intended to clarify our task setup: (i) integrating EDL and (ii) adding LLM-generated context for censorship classification.

3.1 Model Architecture

We adopt `RoBERTa-base` as the base encoder model for our trustworthy censorship classification

task, motivated by both experimental performance from related research (Yadav et al., 2024) and architectural compatibility with the evidential reasoning framework. In terms of its architecture, as opposed to decoder-style architectures such as GPT or Pythia, encoder-based models like RoBERTa—built upon the BERT architecture—facilitate *multi-input conditioning through bidirectional attention*, which is a key feature that enables injecting both LLM-generated context at inference time as well as the raw content.

RoBERTa is based upon the BERT model, with its key structural features summarized as: *embedding layer*, *repeated blocks of multi-head self-attention*, *layer normalization*, and *position-wise feed-forward networks*, each designed to iteratively refine token representations using bidirectional context. The performance of RoBERTa over BERT, discussed in previous research on the same data used in this paper, demonstrated the improved ability of RoBERTa to identify subtle moderation cues.

The input sequence is tokenized using a byte-level BPE tokenizer and encoded into embeddings, which are then passed through $L = 12$ transformer layers. For comparison, the baseline performance with a softmax classification head was first evaluated. Each layer performs:

- **Multi-Head Self-Attention**:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

  enabling token-level alignment and bidirectional dependency modeling.

- **Feed-Forward Network (FFN)**:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

3.2 Embedding and Classifier Pipeline

We used Hugging Face's `AutoModel` to load the pretrained `roberta-base` encoder, freezing the transformer weights and appending a custom `EDLHead` to replace the standard softmax classification head. The final hidden layer of RoBERTa produces contextual embeddings of shape $(\text{batch\_size}, \text{seq\_len}, 768)$, from which we extract the `[CLS]` token for classification as a 768-dimensional feature vector.

The key architectural deviation is replacing the typical pipeline:

`AutoModelForSequenceClassification`

with:

`AutoModel + EDLHead`

The `EDLHead` (defined in `models.py`) receives the `[CLS]` vector, after which it computes *evidence scores* $e \in \mathbb{R}_+^C$, where $C = 2$ for our binary task (Censored vs. Not Censored). The evidence is converted into Dirichlet concentration parameters via $\alpha_c = e_c + 1$, enabling the model to express both class belief and uncertainty.

This change to a Dirichlet-based approach allows better interpretability of model outputs in high-stakes moderation contexts. The classifier no longer outputs raw logits, but a distribution that retains both *confidence* and *ambiguity*. The final output shape is $[\text{batch\_size}, 2]$, aligning with binary labels and used for trust estimation during the inference phase.

Our architectural decision aligns with trends in uncertainty-aware NLP research (e.g., Sensoy et al., 2018), emphasizing *calibrated decision-making* and *transparent error bounds*. By leveraging RoBERTa's encoder structure and the probabilistic expressiveness of EDL, we design a system that supports both high-performance moderation and interpretability—a key requirement when moderating borderline content where unjustified censorship is especially harmful.

3.3 LLM-Guided Context Injection

Past research discussed the potential of using LLMs within the content moderation framework by evaluating both the accuracy and the qualitative explainability of their outputs—including fluency and helpfulness as assessed by human explanations—where findings suggest raw accuracy is around 50% and 72% of explanations were judged helpful in explaining the result (Yadav et al., 2024).

Building on such earlier research, this paper explores the effectiveness of LLM-generated context in reducing decision uncertainty, particularly in borderline cases. We examine whether injecting LLM-guided context—extra information about the content—can help the model make more assertive decisions on censorship. Three design options for LLM content injection were considered, though only one was evaluated in this work:

1. **LLM-based context augmentation at inference time**, where explanatory context is concatenated with the original input:

   Input: [Original post] + [Gemini explanation]

2. **LLM-generated context used during training**, so that the model learns to incorporate and weigh contextual guidance

3. **Dual-embedding structure with separate input and context representations**, using two `[CLS]` tokens and fusing only at the EDL head.

The first option was adopted mainly due to structural constraints of the EDL architecture, which expects a single contextualized input representation in the form of a single `[CLS]` token. Splitting embeddings or tagging parts of the input (e.g., distinguishing context tokens) could confuse the evidence function or disrupt calibration. Training with LLM-generated context is also computationally expensive and may result in the model simply learning the LLM's guideline. In contrast, direct concatenation of the context before tokenization is simple to implement and requires no architectural changes. However, this method has clear limitations as follows, which is also discussed in the future work section:

- EDL treats the entire sequence uniformly—the model cannot explicitly distinguish between user content and LLM-generated context.
- The `[CLS]` token aggregates the full sequence, potentially mixing noisy context with the original signal.

Gemini 1.5 Flash was used to generate the explanatory context, using the following prompt:

```
In under 50 words, objectively
explain the social, cultural, or
political context relevant to
the content of this specific
tweet.  Focus solely on the
tweet's context.
```

Due to computational constraints, we selected 483 borderline samples with extremely high uncertainty for evaluation. These were used to compare model performance with and without the Gemini-generated context.

4 Dataset and Experimental Setup

4.1 Dataset Source and Focus

We use the same censorship dataset as introduced by Yadav et al. (2024), which comprises tweets originally posted in five countries—Germany, France, India, Turkey, and Russia—based on filtering criteria defined in Elmas et al.(2021). While the original work emphasized cross-country censorship patterns, multi-national focus was not considered. In our case, accessible dataset does not contain ground truth annotations of which specific country's regime censored a given tweet. Therefore, our task reduces to a binary classification problem: whether a tweet was censored or not, without distinguishing the country.

4.2 Preprocessed Inputs and Format

The dataset used in this paper was made available by the authors of the previous work, with tweets already matched via the Twitter API, cleaned, and annotated with censorship labels. Below are the key columns in our dataset.

- `text_proc`: Preprocessed tweet text in English.
- `censored`: Binary label (1 = censored, 0 = not censored) from Twitter.
- `withheld`: Optional metadata indicating country-level takedown, not used in this paper due to missing labels in test-time.

For EDL inference uncertainty experiments, we tested Gemini-guided inference on a subset of 483 samples with extreme uncertainty (uncertainty $>$ 0.975). Gemini's context was not provided during the training phase, but only at during the testing phase.

4.3 Data Splits and Preprocessing

Data was splitted into 80/12/8 split for training, validation, and test sets. Each tweet is tokenized using RoBERTa's byte-level BPE tokenizer. For Gemini-context experiments, the context is appended after the original tweet prior to tokenization, resulting in longer sequences but still constrained to the maximum length.

4.4 Training Setup

All models are trained using the AdamW optimizer. We fine-tuned a pretrained `RoBERTa-base` encoder on our censorship classification dataset, appending a custom `EDLHead` for uncertainty-aware prediction. For the EDL model, the custom loss function combines mean squared error between predicted belief and true labels, with a regularization term on the

Dirichlet evidence. The final classification uses the belief values inferred from Dirichlet parameters, rather than softmax probabilities.

5 Results and Analysis

5.1 Results for Reproducing Content Moderation Decision

Below in Table 1, the summary of model performance is presented. Our model, RoBERTa with EDL head fine-tuned, as discussed in the methodology section, achieved a high F1 score of 0.9124. Compared to the base RoBERTa model with a regular softmax head (also fine-tuned), which had an F1 score of 0.9036, it performed slightly better.

Out of all samples, the RoBERTa + EDL model correctly predicted 5,326 positives and 9,931 negatives, with 338 Type I errors (false positives) and 685 Type II errors (false negatives). This result is computed for performance metrics in Table 1.

Performance Metrics:

Metric	Value
Precision	0.9403
Recall	0.8860
F1 Score	0.9124
Type I Error Rate	0.0329
Type II Error Rate	0.1140
Expected Calibration Error(ECE)	0.4008

Table 1: Performance metrics of RoBERTa + EDL on the test set.

The high ECE value of 0.4008 indicates that the model's predicted belief scores are poorly calibrated—i.e., it is often overconfident in incorrect predictions. This highlights the need to analyze the relationship between belief and actual performance.

5.2 Belief Score Analysis

The belief score for the binary classification was plotted as a box plot in Figure 1. In this figure, the belief score for the `censored` class ($C =$ `belief_1`) is plotted. Note that belief scores for the two classes add up to 1. So, for example, a y-axis value of 0.525 means that for that particular content, the belief score for it to be censored is 0.525, and automatically, the belief score for it to *not* be censored (which is not plotted) is 0.475.

Overall, belief scores are very close to a balanced value of 0.5, meaning the model was not particularly assertive in leaning toward one class—this is expected given the inherently ambiguous nature of content moderation for speech.

Figure 1 offers insight into the model's assertiveness and consistency across different output types, differentiated by color as follows:

- **FP (censored, wrong)** cases show relatively moderate to high belief values with a wide interquartile range (IQR), suggesting that the model was inconsistent in its belief level (often overconfident) when it incorrectly predicted censorship.
- **TP (censored, correct)** cases show relatively high belief values, all above 0.5375, with a tight IQR, indicating relatively high confidence and consistency when the model correctly predicted censorship.
- **FN (not censored, wrong)** cases show relatively low to moderate belief values with a relatively wide spread, revealing underconfidence on samples that were actually censored.
- **TN (not censored, correct)** cases show the lowest belief scores with a narrow IQR, reflecting confident and reliable predictions for non-censorship.

Interestingly, a common trend is observed: the model tends to be more assertive and consistent when making correct decisions (True Positive and True Negative), whereas it is less assertive and more inconsistent when making wrong decisions (False Positive and False Negative), indicating that these misclassifications are more weakly believed and less stable.

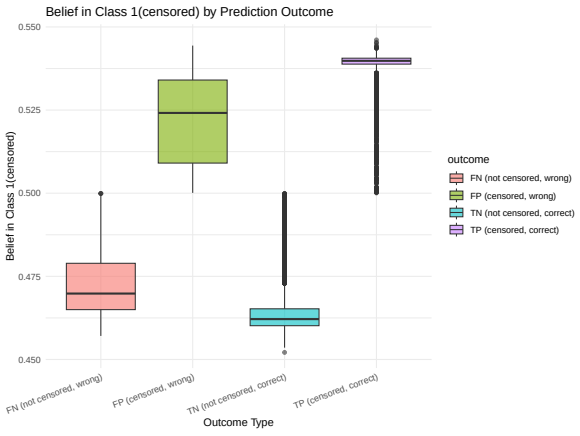

Figure 1: Box plot of `belief_1` by prediction outcome.

5.3 Model Uncertainty distribution

Aligning with the insight from the previous section—that the model does not make assertive decisions—the uncertainty value for the decision is found to be very high. Mathematically, this result makes sense, as the Dirichlet parameter α is used in both belief score and uncertainty calculation. Figure 2 shows the visual distribution of model uncertainty, and Table 2 summarizes the statistics. The mean uncertainty was 0.9274, with some extremely uncertain samples reaching a maximum of 0.9956. This high level of uncertainty motivates further analysis on high-uncertainty cases, and the following section extracts such borderline samples for additional testing.

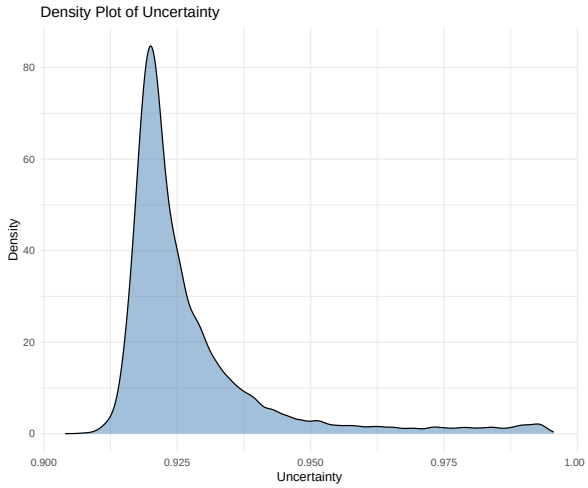

Figure 2: Density plot of model uncertainty on test samples.

Statistic	Uncertainty Value
Mean	0.9274
Median	0.9225
Standard Deviation	0.0144
Minimum	0.9040
25th Percentile (Q1)	0.9194
75th Percentile (Q3)	0.9294
Maximum	0.9956
Interquartile Range	0.0101

Table 2: Summary statistics of `uncertainty` values from the RoBERTa-EDL model on the test set.

5.4 Effect of LLM-Generated Context on High-Uncertainty Samples

Among the entire dataset, high-uncertainty contents with uncertainty over 0.975—a total of 483 samples—were selected to test the effect of embedding Gemini-generated context. Two questions were tested using these high-uncertainty contents: (i) whether LLM-generated context input increases model performance, and (ii) whether LLM-generated context reduces decision uncertainty.

Results show that embedding Gemini-generated context decreases model performance across all metrics (F1, precision, recall, type I error, type II error, accuracy), as summarized in Table 3.

Metric	Before Context	After Context
F1 Score	0.4976	0.3866
Precision	0.4976	0.3333
Recall	0.4976	0.4600
Type I Error	0.3768	0.4144
Type II Error	0.5024	0.5400
Accuracy	0.5694	0.5466

Table 3: Performance comparison on high-uncertainty test samples (uncertainty > 0.975) with and without Gemini-generated context.

For the change in uncertainty, each of the 483 contents' uncertainty was compared before and after adding the Gemini context. Based on the above result on poorer performance from Gemini context input, the null hypothesis (H_0) was set as: uncertainty before adding Gemini context is less than or equal to uncertainty after—i.e., Gemini does not reduce uncertainty. The alternative hypothesis (H_1) was that Gemini-generated context reduces uncertainty.

Since the data are paired and approximately normally distributed (see Appendix), a paired t-test was conducted. Table 4 summarizes the hypothesis setup and test results. Interestingly, the test rejected the null hypothesis with a very low p-value, indicating that the alternative hypothesis—Gemini-generated context reduces uncertainty—was accepted.

Overall, Gemini-generated context embedding resulted in worse performance, but the model's decisions were made with less uncertainty. This means the LLM helped the model become more confident in wrong choices, which can be dangerous in borderline cases.

Test Type	Paired t-test
Sample Size	483, paired
Null Hypothesis (H_0)	$\mu_{\text{before}} \leq \mu_{\text{after}}$
Alternative Hypothesis (H_1)	$\mu_{\text{before}} > \mu_{\text{after}}$
t-statistic	14.446
Degrees of Freedom	482
p-value	$< 2.2 \times 10^{-16}$
Mean Difference	0.0114
90% Confidence Interval	$[0.0104, \infty)$

Table 4: Summary of paired t-test evaluating whether Gemini-generated context reduces model uncertainty. The test strongly supports the alternative hypothesis with statistical significance at the 90% confidence level.

6 Discussion

6.1 How Effective is EDL Integration in Censorship?

Effectiveness of combining the Evidential Deep Learning (EDL) framework into censorship classification can be discussed from two perspectives: (i) whether it improves performance, and (ii) whether it helps ensure the model's trustworthiness.

In terms of performance, although the F1 score increased from 0.9036 with a softmax head to 0.9124 with an EDL head, the improvement is relatively small. Therefore, no strong claim can be made regarding significant performance gains.

However, from a trustworthiness standpoint, the EDL-based approach offers a meaningful advantage by providing additional output signals such as belief scores, uncertainty, and evidence. These outputs allow for deeper analysis of model behavior, especially in borderline cases as it was used to test the effectiveness of LLM generated context input in this paper, and open up pathways for building trust-aware evaluation metrics.

For instance, it can be proposed to use an uncertainty-aware learning framework that uses EDL outputs as feedback for secondary fine-tuning. High ECE value motivates incorporating calibration-aware training or penalty mechanisms for overconfidence to improve trustworthiness. Retraining with "confidence-weighted" labels allows the model to better align predictions with trustworthiness.

Overall, our integration of EDL into the censorship task contributes to the development of a new evaluation paradigm—one that incorporates uncertainty to assess the reliability of predictions beyond standard performance metrics like accuracy.

6.2 Takeaways from LLM guided context input result

Our result shows that the additional input of LLM-generated context worsens performance in all metrics for borderline cases. More concerning, such wrong decisions are made with lower uncertainty (i.e., higher confidence), suggesting the LLM's context causes the model to become overconfident. While overconfidence and hallucination are common issues with LLMs themselves (Nguyen et al., 2025; Ji et al., 2025), the fact that embedding LLM outputs into a secondary model like RoBERTa still leads to overconfident errors is concerning and cannot be logically explained without deeper analysis.

Figure 3 compares confidence distribution before and after adding Gemini-generated context:

- **Before Gemini context (blue)**: Confidence values are sharply concentrated near 0 (high uncertainty), forming a narrow, left-skewed distribution.
- **After Gemini context (red)**: Confidence values are spread more broadly, with a smoother, near-normal shape and a slightly higher peak.

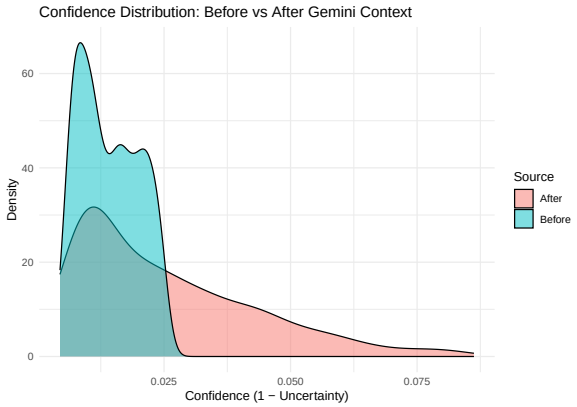

Figure 3: Confidence distribution: before vs. after Gemini-generated context.

One suggestion with this result is that LLM-generated context may be embedding structural signals that normalize the latent feature space, soften sharp uncertainty regions, and encourage smoother predictions across borderline inputs. As a result, the model assigns higher confidence—even when incorrect.

What may be happening is that Gemini context anchors the `[CLS]` embedding or adjacent tokens with extra semantic cues. These cues might flatten

or reshape the uncertainty surface in EDL, widening belief margins. This can mimic calibration behavior—but it's risky if the confidence boost is not backed by grounded evidence.

Still, it is premature to conclude that LLM-guided censorship is ineffective due to some key limitation in the design choice adopted in this paper: Gemini context was appended only during inference, with no exposure during training. Additionally, a single embedding of `[Content + Context]` was used for simplicity and compatibility with the EDL structure. As a result, the model have treated both user content and context as the same source, without structural distinction. Future work should explore segment tagging, dual-encoder designs, or EDL adaptations that allow structured awareness of mixed inputs.

7 Conclusion

In this paper, we implemented and evaluated a censorship classification model integrated with Evidential Deep Learning (EDL), which not only improves classification performance but also enables deeper analysis of the model's decision-making tendencies to ensure trustworthiness. We further investigated the use of LLM-generated context as an additional input and evaluated its impact on borderline censorship cases. Our results show that while LLM-guided context can reduce uncertainty, it also leads to overconfident errors, raising concerns about reliability. Future work should explore leveraging EDL outputs during training to develop trustworthy learning metrics and further investigate the implications of integrating LLMs into censorship pipelines.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. 2021. A dataset of state-censored tweets. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):1009–1015.

Kristina Gligorić, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. Nlp systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 1321–1330. PMLR.

Xiaozhong Ji, Laziz Abdullaev, and Tan M. Nguyen. 2025. Twicing attention: Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Benjamin Loveluck, Ksenia Ermoshina, and Francesca Musiani. 2022. A market of black boxes: The political economy of internet surveillance and censorship in russia. *Journal of Information Technology & Politics*, 19(1):18–33.

Rebecca MacKinnon. 2009. China's censorship 2.0: How companies censor bloggers. *First Monday*, 14(2).

Hieu Nguyen, Zihao He, Shoumik Atul Gandre, Ujjwal Pasupulety, Sharanya Kumari Shivakumar, and Kristina Lerman. 2025. Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation. *arXiv preprint arXiv:2502.11306*.

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

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Vikas Yadav, Qinyuan Qian, Vivek Kulkarni, and Yunyao He. 2024. Revealing hidden mechanisms of cross-country content moderation with natural language processing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

Jiaqi Zhu, Shaofeng Cai, Fang Deng, Beng Chin Ooi, and Wenqiao Zhang. 2023. METER: A dynamic concept adaptation framework for online anomaly detection. *Proceedings of the VLDB Endowment*, 17(4):794–807.

Appendix

The Q-Q plot of uncertainty differences appears approximately linear along the 45-degree reference line, indicating that the differences are roughly normally distributed. This supports the validity of using a paired t-test for statistical comparison of

uncertainty before and after LLM-generated context. Deviations at the tails are minor and expected for moderate sample sizes.

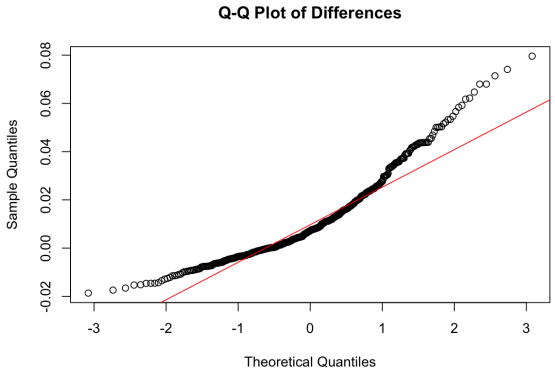

Figure 4: Q-Q plot of uncertainty differences between model predictions before and after LLM context injection.