The Silent Vote: Improving Zero-Shot LLM Classification by Aggregating Semantic Neighborhoods

Anonymous ACL submission

Abstract

Large Language Models are increasingly used as zero-shot classifiers in complex reasoning tasks. However, standard constrained decoding suffers from a phenomenon we define as Renormalization Bias. When a model is restricted to a small set of target labels, the standard softmax operation discards the probability mass assigned to semantic synonyms in the original distribution. This loss of information, which we call the Silent Vote, results in artificial overconfidence and poor calibration.

We propose Semantic Softmax, an inference-time layer that recovers this lost information by aggregating the scores of the semantic neighborhood surrounding each target label. We evaluate our approach using Qwen-2.5 and Phi-4-mini on the GoEmotions and Civil Comments datasets. Our results demonstrate consistent improvements across all evaluation metrics: Semantic Softmax substantially reduces Expected Calibration Error (ECE) and Brier Score, while simultaneously enhancing discriminative performance in terms of AUROC and Macro-F1. By accounting for linguistic nuances, our method provides a more calibrated and accurate alternative for zero-shot classification.

1 Introduction

Large Language Models (LLMs) have shifted the paradigm of text classification from supervised training to zero-shot inference (Brown et al., 2020). In the LLM as a Classifier framework, models are prompted to select a label from a predefined set of candidates (Gilardi et al., 2023). To ensure the model adheres to these task-specific constraints, practitioners typically employ constrained decoding (Shin et al., 2021). To enforce these task-specific constraints, practitioners typically compute probabilities by restricting normalization to the target label tokens, effectively discarding probability mass assigned to all other vocabulary items (Schick and Schütze, 2021).

While effective for enforcing output formats, this methodology introduces a systematic distortion defined here as Renormalization Bias. Standard constrained decoding operates under a localized assumption that all evidence for a specific class is contained within its representative token. However, linguistic meaning in LLMs is inherently distributed across a dense semantic neighborhood of synonyms and related concepts (Mikolov et al., 2013). When a model predicts a text is thrilling but is forced to choose between the labels Joy or Sadness, the probability mass assigned to those synonyms is discarded before the final distribution is calculated.

This exclusion results in a systematic calibration error. By filtering out the probability density assigned to synonymous tokens, the standard constrained softmax operation systematically overestimates the posterior probability of the remaining targets. This leads to severe miscalibration, a known issue where instruction-tuned models exhibit extreme overconfidence (Guo et al., 2017; Minderer et al., 2021). This results in a model that exhibits high-confidence miscalibration. The system may assign near-certain probability to a target label solely because it is the only permissible token with non-negligible affinity to the underlying prediction, even if the absolute semantic alignment is weak. Such behavior prevents the model from accurately quantifying the aleatoric uncertainty inherent in ambiguous inputs (Nie et al., 2020; Pavlopoulos et al., 2020).

To address these limitations, this work proposes Semantic Softmax, a lightweight inference-time intervention designed to recover the Silent Vote. Instead of masking the vocabulary, Semantic Softmax utilizes the internal output embeddings of the model to aggregate probability mass from the entire semantic neighborhood surrounding each target label. This approach allows the model to remain calibrated in the face of ambiguity and overlap-

ping categories. The contributions of this paper are three-fold. First, it formalizes the concept of Renormalization Bias and its impact on model reliability. Second, it introduces Semantic Softmax, a mathematically grounded method to aggregate distributed semantic evidence during constrained inference. Third, experiments on GoEmotions (Demszky et al., 2020) and Civil Comments (Pavlopoulos et al., 2020) demonstrate that this method drastically reduces Expected Calibration Error and improves alignment with human disagreement in ambiguous scenarios, while simultaneously enhancing discriminative performance in terms of AUROC and Macro-F1.

2 Related Work

The reliability of LLMs in zero-shot classification depends on the alignment between internal representations and task constraints.

2.1 Constrained Inference and Decoding

Constrained decoding is the standard mechanism for enforcing output formats in generative models (Shin et al., 2021). Techniques such as verbalizer-based logit masking allow practitioners to restrict the vocabulary to valid labels (Gao et al., 2021). Research into guided generation has focused largely on maintaining structural adherence or satisfying schema requirements (Willard and Louf, 2023), whereas this paper investigates the negative impact of these constraints on probability calibration.

2.2 Calibration in Large Language Models

Calibration measures the degree to which a predicted probability reflects actual accuracy (Guo et al., 2017). Studies have shown that while LLMs demonstrate high zero-shot performance, they are frequently miscalibrated, often exhibiting extreme overconfidence (Minderer et al., 2021; Kadavath et al., 2022). This issue is exacerbated by alignment techniques like RLHF, which push models toward more peaked distributions and lower output diversity (Kirk et al., 2024; Tian et al., 2023). We identify Renormalization Bias as a significant driver of this overconfidence.

2.3 Uncertainty and Ambiguity in NLP

Human language is inherently ambiguous, leading to significant disagreement among annotators (Nie et al., 2020). This aleatoric uncertainty is a critical signal for reliable systems (Pavlopoulos et al., 2020). Standard pipelines often collapse this uncertainty by forcing a single label choice. Semantic Softmax provides an inference-time method to recover human-like uncertainty without additional training.

3 Methodology and Experimental Setup

This section formalizes the theoretical framework of Renormalization Bias and details the proposed Semantic Softmax intervention. We further outline the datasets, model configurations, and evaluation metrics used to validate the efficacy of our approach.

3.1 Formalizing Renormalization Bias and the Synonym Trap

In the standard paradigm of LLM classification, a model is prompted to produce a single token representing a label from a discrete candidate set $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}$. Given a prompt x, the model generates a raw logit vector $\mathbf{z}$ over the entire vocabulary $\mathcal{V}$. To enforce task constraints, practitioners apply a mask M, where $M_i = 0$ if $i \in \mathcal{L}$ and $M_i = -\infty$ otherwise. The resulting constrained probability distribution is calculated via a standard softmax operation:

$$P(l_j|x) = \frac{\exp(z_{l_j})}{\sum_{l' \in \mathcal{L}} \exp(z_{l'})} \quad (1)$$

This localized approach assumes that the semantic evidence for a specific category is concentrated entirely within the single token representing that class. However, instruction-tuned models learn distributed representations where meaning is spread across a dense neighborhood of synonyms and related concepts (Mikolov et al., 2013). If a model assigns significant probability mass to related tokens $v \notin \mathcal{L}$ such as synonyms or hyponyms of l_j, this mass is discarded during the renormalization process.

We define this loss of information as Renormalization Bias. The exclusion of these related tokens creates a Synonym Trap where the model is forced to ignore the Silent Vote of its broader vocabulary. Consequently, the resulting distribution becomes artificially peaked, leading to the extreme overconfidence and miscalibration frequently observed in zero-shot classifiers (Guo et al., 2017; Minderer et al., 2021).

3.2 The Semantic Softmax Framework

To mitigate Renormalization Bias, we propose Semantic Softmax, a method that aggregates the distributed semantic evidence before probability calculation. Instead of masking the vocabulary, we utilize the top K tokens from the unconstrained distribution, denoted as $\mathcal{V}_{topK}$. We define a Semantic Kernel that measures the relationship between these vocabulary tokens and our target labels using the model output embeddings $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$.

The semantic weight $w(v, l)$ between an unconstrained vocabulary token v and a target label l is calculated using a thresholded cosine similarity:

$$w(v, l) = \max\left(0, \frac{\mathbf{E}_v \cdot \mathbf{E}_l}{\|\mathbf{E}_v\|\|\mathbf{E}_l\|} - \tau\right) \quad (2)$$

In this formulation, τ is a hyperparameter threshold that serves as a noise filter to ensure only semantically relevant tokens contribute to the final score. The final probability for each target label is then computed as a weighted sum of the probabilities of the top K tokens:

$$P_{sem}(l|x) = \frac{\sum_{v \in \mathcal{V}_{topK}} P(v) \cdot w(v, l)}{\sum_{l' \in \mathcal{L}} \sum_{v \in \mathcal{V}_{topK}} P(v) \cdot w(v, l')} \quad (3)$$

By aggregating mass from the semantic neighborhood, Semantic Softmax allows the model to reflect its internal aleatoric uncertainty. If the model logic is distributed across tokens associated with different target classes, the resulting distribution will naturally soften, thereby improving calibration.

3.3 Data and Evaluation Protocol

Model Configuration: We evaluate onQwen-2.5-1.5B and Phi-4-mini (3.8B). This diverse setup confirms that Renormalization Bias is a structural decoding artifact independent of specific model architectures or scales.

Benchmark Datasets: GoEmotions (Demszky et al., 2020): This dataset consists of 58,000 Reddit comments labeled with 28 fine-grained emotion categories. It serves as a primary testbed for synonym recovery due to the heavy semantic overlap between labels. **Civil Comments**: This dataset is used to evaluate the model's ability to handle aleatoric uncertainty. Unlike datasets that force a hard label, Civil Comments provides scores derived from annotator disagreement, representing a consensus mean. (Pavlopoulos et al., 2020).

Evaluation Metrics. To provide a comprehensive view of model reliability, we utilize the following metrics: (1) **Expected Calibration Error (ECE)**: This metric quantifies the average gap between the model's predicted confidence and its actual accuracy across different confidence bins. (2) **Brier Score**: A proper scoring rule that measures the mean squared difference between predicted probabilities and actual outcomes, serving as a robust indicator of calibration. (3) **Macro-F1 and AUROC**: These metrics ensure that the semantic aggregation preserves the discriminative power of the model and does not trade off accuracy for calibration.

4 Results and Analysis

We evaluate the efficacy of Semantic Softmax in mitigating Renormalization Bias across our benchmarks. Our evaluation focuses on model calibration, discriminative power, and alignment with human uncertainty.

4.1 Quantitative Performance and Calibration

Table 1 summarizes the performance metrics across architectures and datasets. The primary finding is a systemic and substantial reduction in ECE. On GoEmotions, Semantic Softmax achieves an ECE of 0.071 for Qwen and 0.065 for Phi, representing up to an 8.5-fold improvement over standard constrained baselines. This trend extends to Civil-Comments, where Semantic Softmax consistently mitigates the systematic overconfidence of standard decoding, reducing ECE by over 75% on average. Notably, this improvement in probabilistic reliability does not come at the cost of classification quality; we observe a consistent uplift in discriminative signal, with AUROC and Macro-F1 increasing across all tested configurations.

This suggests that aggregating the probability mass of semantic neighbors not only improves calibration but also enhances discriminative accuracy by recovering previously unobserved probability mass.

4.2 Reliability and Confidence Distribution

Figure 1 illustrates the calibration impact. Panel (a) presents a reliability diagram, where the diagonal identity line ($y = x$) represents perfect calibration—a state where predicted confidence exactly matches empirical accuracy. Deviations falling below this line indicate overconfidence, where the

Model	Dataset	Method	ECE ↓	Brier ↓	AUROC ↑	F1 ↑
Qwen-2.5-1.5B	GoEmotions	Standard	0.574	0.842	0.712	0.229
		Semantic (Ours)	**0.071**	**0.610**	**0.725**	**0.241**
	CivilComments	Standard	0.482	0.571	0.784	0.412
		Semantic (Ours)	**0.112**	**0.523**	**0.811**	**0.436**
Phi-4-mini	GoEmotions	Standard	0.421	0.795	0.744	0.236
		Semantic (Ours)	**0.065**	**0.588**	**0.756**	**0.253**
	CivilComments	Standard	0.395	0.542	0.812	0.421
		Semantic (Ours)	**0.092**	**0.498**	**0.835**	**0.451**

Table 1: Comparative performance of Standard vs. Semantic Softmax across both Qwen and Phi models. Semantic Softmax consistently reduces ECE and Brier Score across both GoEmotions and CivilComments while simultaneously enhancing discriminative power (AUROC and F1).

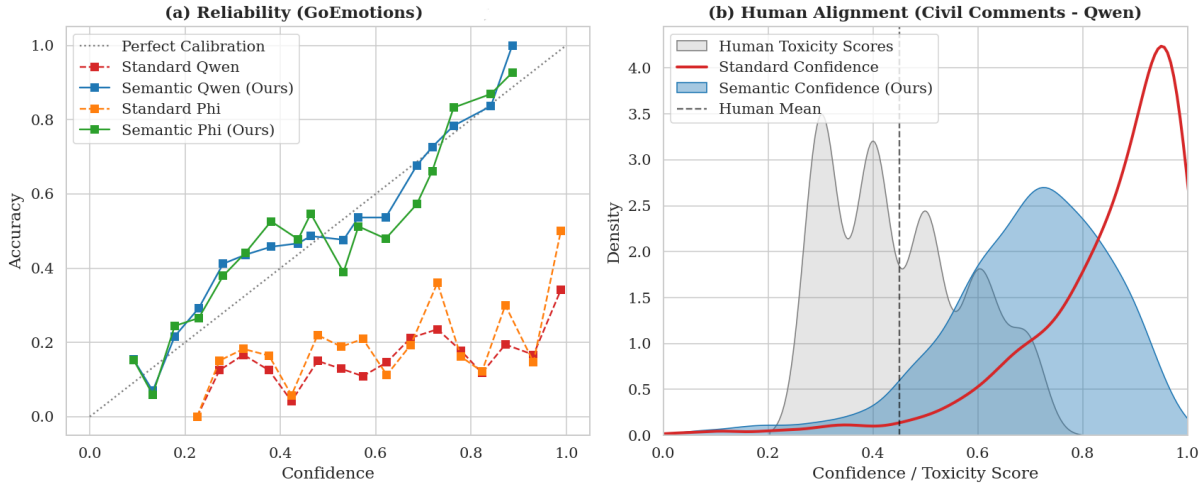

Figure 1: (a) Reliability: Standard decoding exhibits systematic overconfidence (below diagonal), while Semantic Softmax tracks the ideal calibration curve. (b) Confidence: On ambiguous inputs, standard models force extreme probabilities (> 0.9), whereas Semantic Softmax aligns with human consensus, effectively capturing aleatoric uncertainty.

model's probability estimates exceed its actual correctness.

As observed, the standard constrained baseline exhibits a distinct curve significantly below the diagonal, indicative of the systematic overconfidence driven by Renormalization Bias. In contrast, Semantic Softmax minimizes ECE, yielding a reliability curve that closely tracks the ideal diagonal.

This calibration is mirrored in the confidence distribution shift shown in Panel (b). Within the ambiguous Civil Comments subset (human toxicity 0.3–0.7), standard decoding converges to extreme probabilities ($P > 0.90$), effectively erasing annotator disagreement. Semantic Softmax, by aggregating local semantic mass, produces calibrated confidence scores that align with the human mean, accurately quantifying the aleatoric uncertainty inherent in the input.

5 Conclusion

In this work, we identified Renormalization Bias as a primary source of systematic overconfidence and poor calibration in zero-shot classification. We demonstrated that restricting decoding to a sparse label set inadvertently discards the probability mass of semantic synonyms, artificially inflating the confidence of the remaining targets.

To address this, we introduced Semantic Softmax, an inference-time intervention that aggregates probability mass from the model's semantic neighborhood. Our experiments confirm that this approach drastically reduces ECE on the GoEmotions and Civil Comments benchmark without compromising discriminative performance. By effectively recovering the lost signal from the full vocabulary, Semantic Softmax allows LLMs to accurately quantify aleatoric uncertainty, offering a robust and calibrated alternative for classification tasks.

Limitations

A primary limitation of our current study is its focus on the LLM as a Classifier paradigm. Our method is specifically designed for tasks where models are prompted to select a label from a predefined set of candidates during constrained decoding.

Our empirical validation focuses on efficient, small architectures. Due to computational resource constraints, we restricted our analysis to this regime and have not yet established the scaling laws of Semantic Softmax on large-scale foundation models.

Semantic Softmax introduces a non-negligible computational cost at inference time. Retrieving the semantic neighborhood requires a dynamic top-k search and similarity calculation over the vocabulary space. While negligible for single-token classification, this overhead prevents the technique from being easily applied to long-form decoding tasks where latency is cumulative.

Furthermore, the method's efficacy is intrinsically bound to the quality of the underlying embedding space. Our approach assumes that semantically related tokens reside in close geometric proximity. In models with highly anisotropic embeddings or poor semantic clustering, aggregating neighbors could theoretically introduce noise rather than signal. Finally, our evaluation is currently limited to English-language benchmarks . We leave the exploration of Semantic Softmax in multilingual settings where synonyms may span across languages and token to future work.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

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*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 15682–15694. Curran Associates, Inc.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and

natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

A Appendix

A.1 Qualitative Analysis of Renormalization Bias

To illustrate the impact of Renormalization Bias, we provide specific examples from the Civil Comments dataset in Table 2. These samples demonstrate how standard constrained decoding tends toward extreme overconfidence, while Semantic Softmax recovers the Silent Vote to align more closely with human uncertainty.

A.2 Hyperparameter Sensitivity

The Semantic Softmax framework relies on two primary hyperparameters: the number of top vocabulary tokens K and the noise filter threshold τ. We found that performance is robust to K values between 100 and 1000, provided the semantic kernel is properly thresholded. Specifically, as shown in Table 3, we identified $\tau = 0.80$ as the optimal noise filter threshold for minimizing ECE.

Text Sample	Human Truth	Std Score	Sem Score
Let thouse who curse our flag repent or be cursed.	0.478	0.766	0.599
Investing in infrastructure is one thing... The best thing that could happen is to stop the idiocy of the emx boondoggles.	0.600	0.906	0.661
I can't be expected to rebut *every* dopey comment made here and at the National Post, CFL.	0.400	0.664	0.452
They need to euthanize this animal before he kills again.	0.700	0.910	0.762
OK Then if it's so common... Put up or shut up.	0.700	0.957	0.822
Looks like the Judge might have gotten some "side benefits"... throw them in jail.	0.500	0.652	0.576
Good defense strategy: "We wuz too dumb to do anyting bad."	0.600	0.906	0.697
He's a Hispanic causing a panic.	0.600	0.789	0.727
You're using someone else's name, you're a troll.	0.667	0.938	0.879
Perhaps not everyone is as grotesquely course as you!	0.500	0.547	0.590
So, this would be the alternative to the nuclear deal dressed up as a pigs ear? Oops, sorry, silk purse...	0.600	0.816	0.723
Let the NFL die. It's what it deserves.	0.600	0.910	0.797
Good response, JT; I enjoy your even-handed participation... And thank you for your military service.	0.167	0.295	0.200
Fox News is the meaning of nitwit.	0.700	0.926	0.746
CNN haha can dish but can't take. Losers. Go away.	0.522	0.910	0.636
Gosh Justin, this sure makes you appear corrupt. Take a picture of that.	0.500	0.797	0.603
Oh, a cross country tour - sounds ever so nice. What a pompous twit he is!	0.613	0.926	0.718
I would just mute this troll.	0.700	0.941	0.716
What do you do with an elephant with three balls?	0.338	0.053	0.207
Another frustrated juvenile response, naked.	0.400	0.621	0.439

Table 2: Qualitative comparison of toxicity predictions. Semantic Softmax (Sem Score) mitigates extreme probabilities by aggregating local semantic mass, tracking the Human Truth (toxicity mean) more accurately than standard constrained decoding (Std Score).

Table A: Expected Calibration Error (ECE) ↓

$K \downarrow / \tau \rightarrow$	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
50	0.1532	0.1497	**0.1145**	0.1840	0.1839	0.1840
100	0.1523	0.1514	**0.1145**	0.1923	0.1923	0.1923
200	0.1540	0.1492	0.1168	0.1987	0.1986	0.1986
300	0.1538	0.1491	0.1144	0.1987	0.1989	0.1987
400	0.1529	0.1511	0.1154	0.1986	0.1986	0.1987
500	0.1528	0.1489	0.1142	0.1988	0.1986	0.1987
600	0.1527	0.1498	0.1130	0.1986	0.1985	0.1986
700	0.1538	0.1487	0.1141	0.1986	0.1986	0.1987
800	0.1527	0.1498	0.1151	0.1995	0.1994	0.1995
900	0.1517	0.1508	0.1151	0.1993	0.1992	0.1994
1000	0.1527	0.1499	0.1162	0.1993	0.1992	0.1993

Table B: Macro-F1 Score ↑

$K \downarrow / \tau \rightarrow$	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**
50	0.4092	0.4101	0.4325	**0.4359**	0.4359	0.4359
100	0.4083	0.4102	**0.4308**	0.4112	0.4112	0.4112
200	0.4083	0.4093	**0.4308**	0.3958	0.3958	0.3958
300	0.4083	0.4093	**0.4308**	0.3909	0.3909	0.3909
400	0.4083	0.4093	**0.4308**	0.3899	0.3899	0.3899
500	0.4083	0.4093	**0.4308**	0.3899	0.3899	0.3899
600	0.4083	0.4093	**0.4308**	0.3899	0.3899	0.3899
700	0.4083	0.4093	**0.4308**	0.3899	0.3899	0.3899
800	0.4083	0.4093	**0.4308**	0.3889	0.3889	0.3889
900	0.4083	0.4093	**0.4308**	0.3889	0.3889	0.3889
1000	0.4074	0.4093	0.4317	0.3889	0.3889	0.3889

Table 3: Hyperparameter sensitivity analysis. We find that $\tau = 0.80$ consistently minimizes ECE while maintaining high Macro-F1 across a wide range of K, effectively mitigating the *Renormalization Bias*.