Confidence-Gated LLM Synthesis for Enhanced Multi-Class Sentiment Analysis in Financial Texts

YongKyung Oh
Medical & Imaging Informatics (MII) Group
University of California, Los Angeles (UCLA)
Los Angeles, CA, USA
yongkyungoh@mednet.ucla.edu

Jaesun Yeom
Department of Industrial Management Engineering
Hanbat National University
Daejeon, Republic of Korea
jsyeom@hanbat.ac.kr

Abstract

Accurate sentiment analysis of news headlines, particularly in specialized domains like finance, is challenging due to textual nuances and inherent ambiguities. While Large Language Models (LLMs) excel at text understanding, their zero-shot performance can be inconsistent. We propose a novel framework where an LLM acts as an intelligent synthesizer, combining raw text semantics with confidence-gated probabilistic outputs from a specialized intermediate multi-class sentiment classifier. This confidence-gating mechanism dynamically adjusts the information provided to the LLM, guiding its reasoning process. Experiments on established financial sentiment datasets, SentFiN, demonstrate that our approach significantly outperforms zero-shot LLM baselines and traditional machine learning combiners, when leveraging insights from a confident intermediate model.

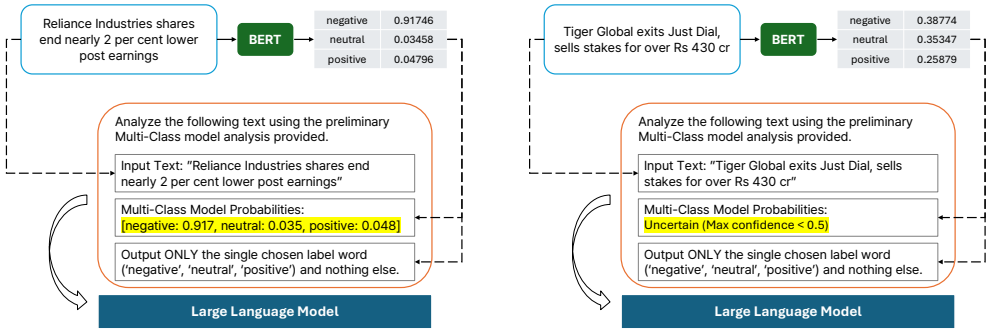
1 Introduction

Interpreting sentiment in textual data, such as news headlines, is crucial for applications ranging from financial market analysis [17, 6] to understanding public opinion [14]. Financial texts are characterized by domain-specific terminology and subtle contextual cues that can lead to high ambiguity, making accurate multi-class sentiment classification a difficult task [11, 19, 2]. While Large Language Models (LLMs) have shown impressive zero-shot capabilities in various NLP tasks [10], their out-of-the-box performance on specialized classification tasks can be further improved.

Existing approaches often involve fine-tuning LLMs, which can be resource-intensive, or employing traditional ensemble methods [15, 8] that may not fully leverage the nuanced reasoning capabilities of LLMs. This paper explores a novel direction: using an LLM as an intelligent synthesizer that integrates raw text with conditionally provided insights from a fine-tuned, task-specific intermediate classifier. This approach is particularly relevant as recent studies highlight both the potential and the challenges of applying LLMs directly to complex financial sentiment tasks [12, 3].

The core of our contribution is a confidence-gating mechanism. This mechanism dynamically determines whether to provide the LLM with the probabilistic output of the intermediate MC classifier based on its prediction confidence. If the MC classifier is highly confident, its detailed analysis is passed to the LLM to guide its final decision. If the MC classifier is uncertain, this uncertainty is

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .



(a) BERT model predicts 'negative' with high confidence (0.917), so the full probability distribution is included in the prompt to the LLM.

(b) BERT model shows low confidence (max probability 0.388, which is lower than 0.5), thus an 'Uncertain' message is passed to the LLM.

Figure 1: Illustration of the confidence-gating mechanism

signaled, prompting the LLM to rely more on its inherent understanding of the raw text. This process is visualized in Figure 1. We hypothesize that this selective guidance allows the LLM to make more robust and accurate multi-class sentiment predictions by leveraging specialized knowledge when it is reliable, while avoiding noise from uncertain intermediate predictions. This aligns with broader efforts to enhance LLM reasoning by integrating external, structured information or tools [4, 23]. We evaluate our method on established financial sentiment datasets, SentFiN [18], demonstrating its advantages over zero-shot LLMs and traditional machine learning combiners.

2 Related Work

The advent of LLMs has significantly impacted text classification [7], with many studies exploring their zero-shot and few-shot learning capabilities [10]. Beyond direct classification, LLMs have demonstrated sophisticated reasoning abilities, often elicited through prompting techniques like Chain-of-Thought (CoT) [20] or more complex frameworks like Tree of Thoughts (ToT) [23]. Furthermore, LLMs are increasingly being integrated with external tools or information sources [22, 4, 13], a paradigm conceptually related to how our LLM utilizes outputs from the intermediate classifier. Our method distinguishes itself by focusing on the confidence-gated integration of insights from a task-specific, fine-tuned classifier.

The principle of combining multiple models to achieve superior performance is well-established in ensemble learning [15, 8]. Traditional stacking approaches utilize a meta-learner to aggregate predictions from base models. While our framework's LLM acts as a high-level decision-maker, it differs by processing a natural language summary of the intermediate model's probabilistic output and confidence, facilitating a more nuanced synthesis than typical feature-based meta-learners. Recent studies also explore ensembling LLMs directly [24, 1].

3 Methodology

3.1 Intermediate Multi-Class (MC) Classifier

First, we fine-tune a standard transformer-based model hereafter referred to as the MC classifier (M_{MC}), on the target multi-class sentiment task. For any input headline x, M_{MC} produces a probability distribution $\mathbf{p}_{mc}(x)$ across the K sentiment classes. This provides an initial, task-specific sentiment assessment.

3.2 Confidence-Gated Prompt Construction

A key novel component is how we utilize $\mathbf{p}_{mc}(x)$. Instead of always passing the full probability vector to the LLM, we employ a confidence-gating mechanism based on a predefined threshold τ_{mc}.

Let $p_{mc}^{\max}(x)$ be the maximum probability in $\mathbf{p}_{mc}(x)$, which is denoted as confidence of M_{MC}:

- Certain Case: If $p_{mc}^{\max}(x) \geq \tau_{mc}$, the LLM prompt includes a textual representation of the full probability vector from M_{MC}. Described in Figure 1(a).
- Uncertain Case: if $p_{mc}^{\max}(x) < \tau_{mc}$ (Uncertain Case, Figure 1b), the LLM prompt includes a message indicating M_{MC}'s uncertainty. Described in Figure 1(b).

This conditional summary is denoted as $s_{mc}(x)$.

3.3 LLM as a Synthesizer

The final sentiment classification is performed by an pretrained LLM (M_{LLM}). The LLM is provided with a system prompt defining its role as an expert sentiment analysis assistant and constraining its output to one of the predefined labels, as represented in Figure 2. The user prompt then contains the original headline x and the confidence-gated summary $s_{mc}(x)$ from Section 3.2.

> You are an expert sentiment analysis assistant. Your task is to determine the final sentiment label for the text provided in the user message. You MUST choose exactly one label from the following list: ('negative', 'neutral', 'positive'). Output ONLY the single chosen label word ('negative', 'neutral', 'positive') and nothing else. Do not include explanations, introductions, or any other text.

Figure 2: Zero-shot instruction prompt for LLM

The LLM's task is to synthesize these pieces of information—the raw semantics of x and the (potentially uncertain) preliminary analysis $s_{mc}(x)$—to produce the final sentiment label $\hat{y}_{LLM}$. This approach allows the LLM to incorporate contextual understanding and external guidance from M_{MC}, leading to more informed predictions than relying solely on x or rigidly copying M_{MC}'s output.

4 Experiments

For the intermediate MC classifier (M_{MC}), We fine-tune a BERT-based model, specifically DistilBERT [16], as our M_{MC} on the combined training splits to predict one of classes. In case of LLM Combiners (M_{LLM}), we test two pretrained LLMs as synthesizers, including Mistral-7B-Instruct-v0.3 [9][1] and Llama-3.1-8B-Instruct [5][2], using Huggingface [21]. The system prompt (detailed in Section 3) instructs the LLM to act as an expert sentiment analyzer and output a single label. We compare our proposed framework against:

1. M_{MC} only: The fine-tuned intermediate MC classifier's direct predictions after fine-tuning.
2. Traditional ML classifiers: Logistic Regression, SVM, Random Forest, XGBoost, and LightGBM, trained on text embeddings from M_{MC}'s base model, with supervised manner.
3. Zero-shot LLMs: The LLMs classifying headlines directly without M_{MC} guidance. We evaluated both OpenAI's ChatGPT models (GPT-3.5-turbo and GPT-4o-mini) and our proposed backbone models (Mistral-7B and Llama-3.1-8B).

Each method is run three times, and we report the average performance, including the macro F1 score and class-wise F1 scores.

4.1 SentFiN

We use the SentFiN dataset [18], which contains more than 10000 financial news sentences annotated by domain experts with sentiment labels: 'negative', 'neutral', or 'positive'. While the original dataset includes multi-target sentiment annotations, we use only instances with a single target. We split the data into training and test sets using an 80/20 stratified split, as summarized in Table 1.

[1] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`

[2] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

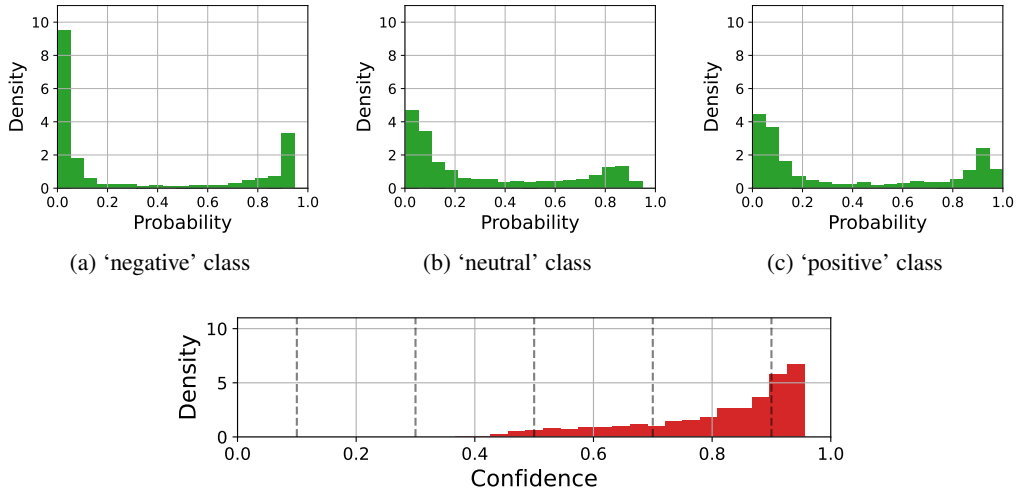
Table 1: Class-wise sample counts in the SentFiN dataset

Label \ Split	train	test
negative	1901	475
neutral	2151	538
positive	2270	568
Total	6322	1581

Table 2 showcases the performance comparison of our proposed framework against several baseline methods on the SentFiN dataset. The best-performing and second-best models are indicated with bold and underline, respectively. Our framework significantly improves the classification performance when using both tested backbone LLMs without training or fine-tuning.

Table 2: Performance comparison using SentFiN dataset

Method	F1 Macro	F1_negative	F1_neutral	F1_positive
BERT only	0.7980	0.8171	0.7518	0.8249
BERT + LogisticRegression	0.7747	0.7634	0.7744	0.7861
BERT + SVM	0.7748	0.7677	0.7737	0.7831
BERT + DecisionTree	0.5434	0.5056	0.5609	0.5639
BERT + RandomForest	0.7026	0.6557	0.7376	0.7143
BERT + XGBoost	0.7312	0.7087	0.7458	0.7390
BERT + LightGBM	0.7270	0.7078	0.7407	0.7324
Zero-shot GPT-3.5-turbo	0.7653	<u>0.8671</u>	0.6217	0.8070
Zero-shot GPT-4o-mini	0.7757	0.8325	0.6770	0.8174
Proposed method (Mistral-7B)	<u>0.8456</u>	**0.8682**	<u>0.7915</u>	**0.8773**
–Zero-shot Mistral-7B	0.7482	0.8530	0.6092	0.7824
Proposed method (Llama-3.1-8B)	**0.8551**	0.8655	**0.8281**	<u>0.8718</u>
–Zero-shot Llama-3.1-8B	0.7293	0.8296	0.5751	0.7831



(a) 'negative' class (b) 'neutral' class (c) 'positive' class

(d) Distribution of model confidence scores

Figure 3: MC classifier's probability distributions of each class and model confidence scores

Figure 3 (a), (b), and (c) show the probability distribution of each class ('negative', 'neutral', or 'positive') as predicted by our intermediate MC classifier (M_{MC}) on the SentFiN dataset. We utilize the maximum value from these probability distributions as the confidence score for a given prediction.

The overall distribution of these maximum confidence scores from M_{MC} is presented in Figure 3 (d). These distributions inform how our confidence-gating mechanism, which uses a threshold τ_{mc} against these maximum probabilities, affects the information flow to the LLM.

5 Conclusion

We presented a framework for multi-class sentiment analysis that combines a confidence-aware intermediate classifier with a large language model. The LLM acts as a synthesizer, integrating raw text with classifier outputs to produce label-consistent predictions without requiring fine-tuning. Experiments on financial sentiment datasets demonstrate that this approach outperforms both zero-shot LLMs and traditional model combiners, particularly under ambiguous or uncertain inputs.

Future work includes improving uncertainty communication, exploring adaptive thresholding, refining prompts, and distilling the framework into more efficient models. Extending this approach to other NLP tasks may further validate its general applicability.

References

[1] Ahmed Alsayat. Improving Sentiment Analysis for Social Media Applications Using an Ensemble Deep Learning Language Model. *Arabian Journal for Science and Engineering*, 47(2):2499–2511, February 2022. doi: 10.1007/s13369-021-06227-w.

[2] Samuel W.K. Chan and Mickey W.C. Chong. Sentiment analysis in financial texts. *Decision Support Systems*, 94:53–64, February 2017. doi: 10.1016/j.dss.2016.10.006.

[3] Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. Financial Sentiment Analysis: Techniques and Applications. *ACM Comput. Surv.*, 56(9), April 2024. doi: 10.1145/3649451.

[4] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR, July 2023.

[5] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 Herd of Models, 2024.

[6] Michael Hagenau, Michael Liebmann, and Dirk Neumann. Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems*, 55(3):685–697, June 2013. doi: 10.1016/j.dss.2013.02.006.

[7] J. Fields, K. Chovanec, and P. Madiraju. A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe? *IEEE Access*, 12:6518–6531, 2024. doi: 10.1109/ACCESS.2024.3349952.

[8] Jianguo Jia, Wen Liang, and Youzhi Liang. A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing, 2023.

[9] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B, 2023.

[10] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9):195:1–195:35, 2023. doi: 10.1145/3560815.

[11] Tim Loughran and Bill Mcdonald. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65, February 2011. doi: 10.1111/j.1540-6261.2010.01625.x.

[12] Bo Peng, Emmanuele Chersoni, Yu-yin Hsu, Le Qiu, and Chu-ren Huang. Supervised Cross-Momentum Contrast: Aligning representations with prototypical examples to enhance financial sentiment analysis. *Knowledge-Based Systems*, 295:111683, July 2024. doi: 10.1016/j.knosys.2024.111683.

[13] Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. Empowering Multi-step Reasoning across Languages via Program-Aided Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 12171–12187, 2024.

[14] David Rozado, Ruth Hughes, and Jamin Halberstadt. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE*, 17(10): e0276367, October 2022. doi: 10.1371/journal.pone.0276367.

[15] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1249, July 2018. doi: 10.1002/widm.1249.

[16] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, March 2020.

[17] Robert P. Schumaker, Yulei Zhang, Chun-Neng Huang, and Hsinchun Chen. Evaluating sentiment in financial news articles. *Decision Support Systems*, 53(3):458–464, June 2012. doi: 10.1016/j.dss.2012.03.001.

[18] Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. SEntFiN 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9): 1314–1335, 2022. doi: 10.1002/asi.24634.

[19] Yongyong Sun, Haiping Yuan, and Fei Xu. Financial sentiment analysis for pre-trained language models incorporating dictionary knowledge and neutral features. *Natural Language Processing Journal*, 11: 100148, June 2025. doi: 10.1016/j.nlp.2025.100148.

[20] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022.

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6.

[22] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J. Cai. PromptChainer: Chaining Large Language Model Prompts through Visual Programming. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022, Extended Abstracts*, pages 359:1–359:10, 2022. doi: 10.1145/3491101.3519729.

[23] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.*, 2023.

[24] Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. Pushing The Limit of LLM Capacity for Text Classification. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, pages 1524–1528, New York, NY, USA, 2025. Association for Computing Machinery. doi: 10.1145/3701716.3715528.