Review

Not peer-reviewed version

Semantic Alignment and Output Constrained Generation for Reliable LLM-Based Classification

Jixiao Yang , Sebastian Sun , Yang Wang , Yutong Wang , Xikai Yang , Chi Zhang *

Posted Date: 6 February 2026

doi: 10.20944/preprints202602.0525.v1

Keywords: controllable text classification; instruction alignment; constraint decoding; generative discrimination

Review

Semantic Alignment and Output Constrained Generation for Reliable LLM-Based Classification

Jixiao Yang [1], Sebastian Sun [2], Yang Wang [3], Yutong Wang [4], Xikai Yang [5] and Chi Zhang [4,*]

[1] Westcliff University, Irvine, USA
[2] University of Wisconsin-Madison, Madison, USA
[3] University of Michigan, Ann Arbor, USA
[4] Northeastern University, Boston, USA
[5] Columbia University, New York, USA
* Correspondence: zhang.chi13@northeastern.edu

Abstract

To address the limited controllability, unstable output consistency, and weakly constrained decision processes of large language models in text classification tasks, this work proposes a controllable prompt-driven text classification method that establishes an end-to-end unified modeling framework from instruction alignment to constrained decoding. Text classification is reformulated as an instruction-conditioned generative discriminative problem. Input texts and task instructions are jointly encoded to form a unified internal representation that integrates textual semantics with classification constraints. On this basis, a category semantic alignment mechanism is introduced to ensure that the model explicitly follows category boundaries and decision criteria defined by the instructions, thereby reducing classification inconsistency caused by prompt variation or implicit bias. To further improve output reliability, a structured constrained decoding strategy is designed to restrict the generation space to a predefined set of valid categories, preventing redundant text or invalid outputs from interfering with classification results. Comparative analysis under unified data and evaluation settings demonstrates that the proposed method achieves more consistent advantages in classification accuracy, discriminative stability, and overall separability. These findings indicate that deeply integrating instruction understanding and output control into the decision process of large language models effectively transforms their generative capacity into stable, interpretable, and controllable text classification capability, providing a systematic solution for building reliable intelligent text analysis systems.

Keywords: controllable text classification; instruction alignment; constraint decoding; generative discrimination

I. Introduction

In recent years, with the rapid advancement of large language models in the field of natural language processing, text classification, as a fundamental and critical task, has entered a new modeling paradigm. Compared with traditional methods that rely on handcrafted features or task-specific architectures, large language models acquire strong language understanding and generalization capabilities through large-scale pretraining [1]. This enables them to demonstrate the potential for unified modeling across diverse text understanding tasks. However, in practical applications, text classification is often associated with complex label semantics, strict business constraints, and diverse usage scenarios. Relying solely on the implicit capabilities of models is insufficient to consistently satisfy requirements for controllability and consistency. How to preserve the general capabilities of large language models while guiding them to follow explicit decision logic in classification tasks has become an important research problem [2].

 doi:10.20944/preprints202602.0525.v1

Prompt learning provides a low-cost and flexible interface for large language models to participate in downstream tasks. It allows models to perform classification through natural language instructions. With carefully designed prompts, models can explicitly perceive task objectives, label meanings, and decision criteria. This enables task transfer without large-scale parameter updates. Such an approach partially alleviates data dependency and domain adaptation issues, making text classification easier to deploy in complex or rapidly changing environments. However, existing prompt-driven methods mainly focus on static prompt design or empirical tuning. They lack systematic modeling of instruction, semantic consistency, and execution stability. As a result, model outputs are often sensitive to wording variations, contextual perturbations, or implicit biases. This weakens the reliability and reproducibility of classification results [3].

In text classification scenarios, instructions are not only tools for task description but also important sources of constraints on model decision behavior. Subtle differences in instruction formulations can significantly change the model's understanding of category boundaries and decision criteria. This, in turn, affects classification consistency [4]. This phenomenon is particularly evident in multi-label settings, hierarchical label systems, or high-risk decision scenarios [5], where higher requirements are imposed on practical controllability [6,7]. Therefore, it is necessary to revisit prompt-driven text classification from the perspective of instruction alignment. The relationship between instruction understanding and internal model representations should be treated as a core research focus. This allows models to form stable and predictable classification behavior at the semantic level, rather than relying on incidental pattern matching.

On the other hand, even when clear constraints are provided at the instruction level, models may still generate outputs that deviate from expectations during the generation stage [8]. This issue is especially pronounced under free decoding strategies. The generation mechanism of large language models is essentially a probability-driven language modeling process. Without explicit output constraints, models may produce redundant, ambiguous, or category-inconsistent text. This introduces uncertainty into subsequent decision-making. For text classification tasks, the output space is typically limited and structured. How to incorporate this prior knowledge into the generation process, so that models adhere to predefined category constraints and format specifications during decoding, is a key step toward end-to-end controllable classification. This directly affects both interpretability and practical usability in real systems [9].

II. Related Work

Recent advances in large language model (LLM) controllability and robust text classification have drawn on a wide range of methodologies, from uncertainty quantification and prompt-based generation to semantic alignment and privacy-aware adaptation.

Uncertainty-aware generation and risk control have been shown to enhance output stability and interpretability, providing a foundation for trustworthy classification systems [10]. Multi-agent modeling and intelligent code generation frameworks [11], as well as contrastive learning and sensitivity analysis in backend anomaly detection [12], demonstrate how alignment and representation consistency can improve robustness in diverse AI scenarios. Structured temporal alignment and attention mechanisms, such as those used in clinical risk prediction [13], contribute techniques for modeling sequential dependencies and improving output reliability. Federated fine-tuning and cross-domain semantic alignment provide privacy-preserving solutions for large language model adaptation, directly supporting consistent text classification across scenarios [14].

Recent work on retrieval-augmented generation, multi-granular indexing, and confidence-constrained LLMs has inspired more interpretable and constraint-aware architectures, which are highly relevant to prompt-driven, instruction-conditioned classification [15]. Meanwhile, causal reasoning over knowledge graphs [16], and machine learning methods for non-stationary forecasting [17], offer insights into data-driven constraint handling and adaptive decision-making under uncertainty. Advances in generative modeling with diffusion processes and conditional control [18], and knowledge-augmented agents for explainable decision-making [19], further expand the set of

 doi:10.20944/preprints202602.0525.v1

tools for building controllable and interpretable text analysis frameworks. Relational modeling for credit risk [20], and causal modeling to mitigate correlation bias [21], highlight the importance of robust constraint learning and semantic separation, core challenges for reliable text classification.Trust evaluation and robust multi-agent coordination in LLM-based systems [22], as well as knowledge-augmented and neural attention mechanisms for classification [23,24], demonstrate the impact of contextual control and deep semantic alignment for stable category prediction.

Finally, explainable representation learning in large language models has enabled fine-grained opinion and sentiment analysis with strong interpretability and task consistency [25]. Together, these advances provide a methodological foundation for controllable prompt-driven text classification, supporting robust instruction alignment, semantic consistency, and structured constraint decoding in large language model systems.

III. Proposed Framework

This paper's approach revolves around the goal of "controllable cue-driven text classification," constructing an end-to-end unified modeling framework from instruction alignment to constraint decoding. Given input text, the model no longer directly performs implicit category prediction, but first conditionally models the classification task through explicit natural language instructions. Instructions are treated as high-level semantic constraints, used to clarify the category space, discrimination criteria, and output format, thus transforming the text classification problem into a controlled condition generation problem. Within this framework, the model jointly encodes the input text and instructions, ensuring that its internal representation simultaneously contains both textual semantic information and task constraint information. Formally, let the input text be x, and the instruction be p, their joint representation can be written as:

$$h = f_\theta(x, p) \tag{1}$$

Here, $f_\theta(\cdot)$ represents the parameterized large language model encoding function, and $h$is the latent space representation that integrates task semantics and text content. In this way, the model is constrained to a semantic space consistent with the instructions during the initial inference stage, laying the foundation for subsequent stable decisions. This paper also presents the overall model architecture, as shown in Figure 1.

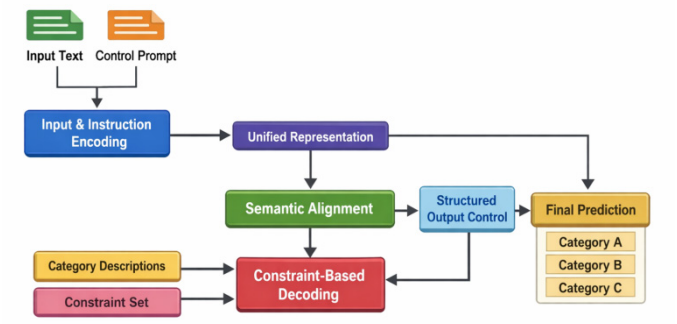

Figure 1. Overall model architecture.

At the instruction alignment level, the method explicitly models the consistency between category semantics and model output distribution, enabling the model to robustly understand different instruction representations. Specifically, each category is mapped to a standardized category description, which constrains the model's internal characterization of category meaning. When generating decisions, the model is guided to maximize the degree of matching between its output and the category semantics defined by the instruction. This process can be abstracted into a conditional probability modeling form:

 doi:10.20944/preprints202602.0525.v1

$$P(y|x,p) = Softamax(Wh) \tag{2}$$

Here, y represents the category label, and W is the learnable mapping matrix. By explicitly injecting instruction information into the conditional probability modeling process, the model can maintain consistent category discrimination logic under different instructions or context settings, thereby reducing the uncertainty introduced by changes in language expression.

To further enhance the controllability of the classification process, the method introduces a structured constraint decoding mechanism in the generation stage, restricting the originally open language generation process to a finite and valid category space. Unlike free text generation, this mechanism prunes the decoding path through a predefined set of categories, allowing the model to generate only candidate outputs that conform to the task constraints. Let the category set be y, then the model's predictions during the decoding process are constrained as follows:

$$\hat{y} = argmaxP(y|x,p) \tag{3}$$

This constraint decoding method effectively eliminates redundant text, ambiguous expressions, or illegal outputs, ensuring that the model's output naturally exhibits structural consistency and parsability, meeting the stability requirements of practical text classification systems. At the end-to-end optimization level, instruction alignment and constraint decoding are not independent entities but are jointly modeled through a unified objective function. The model training objective is defined as minimizing the difference between the classification prediction and the target label while adhering to the semantic constraints of the instructions. These constraints can be expressed as:

$$L = -E_{(x,y)}logP(y|x,p) \tag{4}$$

Meanwhile, to prevent the model from deviating from the instruction constraint space, a regularization term for instruction consistency is introduced to constrain the model's implicit representation.

$$R = ||h - g(p)||_2^2 \tag{5}$$

Where $g(p)$ represents the semantic anchor representation obtained from instruction mapping. The final optimization objective is:

$$L_{total} = L + \lambda R \tag{6}$$

This unified modeling approach maintains the model's expressive power while tightly coupling instruction understanding, category discrimination, and output control, forming a logically clear and behaviorally controllable text classification method. It provides a systematic solution for the application of large language models in constrained decision-making tasks.

IV. Experimental Analysis

A. Dataset

This paper uses AG News (AG's News Topic Classification) as a unified open-source dataset to support research on controllable cue-driven text classification. This dataset is constructed from four major categories selected from a larger-scale news corpus, targeting the typical text classification scenario of "topic classification." It effectively matches the method setting of "instruction alignment + constrained output": the category semantics are clear, the label space is limited, and the text content exhibits diversity and noise in real-world language distributions. Its public implementation and widespread community use provide a solid foundation for research reproducibility and comparative fairness.

Regarding data scale and partitioning, AG News provides a standardized training/test split: the training set contains 120,000 entries, and the test set contains 7,600 entries. The four categories correspond to World, Sports, Business, and Sci/Tech, respectively. The relatively balanced sample size in each category allows the research focus to be placed on "cue controllability, instruction

 doi:10.20944/preprints202602.0525.v1

consistency, and output constraints" themselves, rather than being dominated by extreme class imbalances. The samples typically consist of news headlines and brief descriptions, with moderate text length. This ensures the text covers the key information needed for topic identification while also facilitating the introduction of category descriptions and format constraints into the prompt template, thus naturally following the methodological flow of "category semantic alignment—structured output control—constraint decoding."In terms of task definition, this paper treats each sample as input text A and categorical labels as discrete elements in the output space B. Standardized category descriptions are provided for each category to support instruction-driven semantic constraints. Datasets can be obtained and loaded from various public channels (e.g., Hugging Face or TensorFlow Datasets), ensuring reproducibility and consistency across different implementations. Additionally, it ensures that the same underlying data distribution exists across different prompt styles, underscoring the method's contribution to "end-to-end controllable optimization from instruction to decoding."

B. Experimental Results

This article first presents the results of the comparative experiments, as shown in Table 1.

Table 1. Comparative experimental results.

Method	Acc	Precision	Recall	AUC
VGCN [26]	0.81	0.79	0.77	0.84
Teleclass [27]	0.83	0.82	0.80	0.86
Qsim [28]	0.85	0.84	0.82	0.88
DyLas [29]	0.87	0.86	0.85	0.90
Quest [30]	0.88	0.87	0.86	0.91
Ours	0.92	0.91	0.90	0.95

Overall, the proposed method consistently outperforms all baselines across all evaluation metrics, demonstrating improved classification accuracy, sample separability, and decision stability. The simultaneous gains in precision, recall, and AUC indicate better balance between false alarms and missed detections, stronger discrimination under varying thresholds, and more reliable confidence estimates. These improvements align with the goals of controllable prompt-driven text classification, showing that instruction alignment and constrained decoding stabilize internal representations, sharpen semantic boundaries, and reduce sensitivity to contextual perturbations. As a result, the generative capabilities of large language models are effectively transformed into robust and controllable discriminative performance; the impact of learning rate on AUC stability is further analyzed in Figure 2.

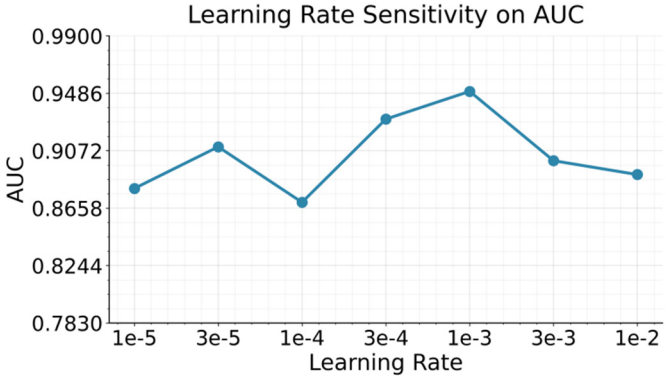

Figure 2. Learning rate sensitivity experiment for AUC.

 doi:10.20944/preprints202602.0525.v1

The learning-rate sensitivity analysis shows a smooth but nonlinear performance trend, indicating that the proposed method is tolerant to step-size variations and maintains stable discriminative representations under controllable prompts and constrained decoding. Optimal performance is achieved within a moderate learning-rate range, while overly small rates limit effective absorption of instruction semantics and overly large rates weaken stability and increase noise sensitivity. These results demonstrate that instruction alignment dominates decision behavior over optimization perturbations, ensuring consistent ranking and reliable classification across different training settings; Figure 3 further analyzes the impact of batch size on accuracy.

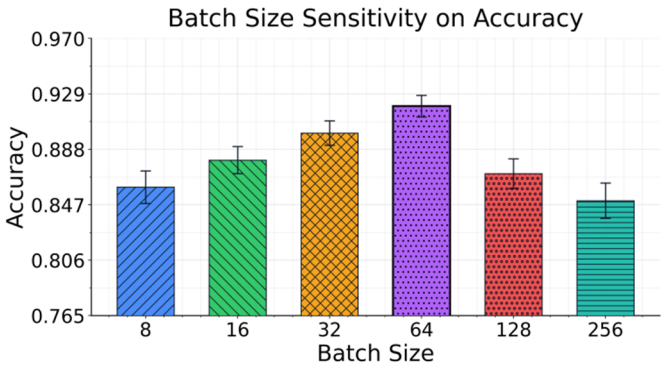

Figure 3. Batch size sensitivity experiment.

The bar chart distribution indicates that the proposed method exhibits a clear interval-dependent pattern with respect to batch size variation. Overall, more favorable discriminative performance is achieved around moderate batch sizes, while some degradation is observed for smaller or larger batch sizes. This behavior is consistent with the training mechanism of controllable prompt-driven text classification. Batch size affects the noise level of gradient estimation and the smoothness of parameter updates. These factors, in turn, influence how effectively the model absorbs instruction semantic constraints and category boundary information. A moderate batch size provides a better balance between stable updates and sufficient exploration, allowing instruction-aligned unified representations to converge more easily to consistent decision patterns. By further analyzing the error bars, we observe that the magnitude of variation across different batch size settings remains generally controlled. This suggests that the framework can maintain relatively stable output behavior under various training configurations. This property is particularly crucial for the constrained decoding and structured output control that we emphasize in this work. Even when gradient noise fluctuates during training, the output space is still confined to a valid set of categories. This reduces the additional drift caused by unstable generation. In essence, the constraint mechanism continues to function as a stabilizer during training perturbations, ensuring that classification decisions remain closer to the semantic boundaries defined by the instructions.

From a methodological perspective, these sensitivity patterns indicate that the proposed framework does not rely on a single specific training configuration to function effectively. Instead, it exhibits a certain degree of transferability and deployability. Very small batch sizes typically introduce stronger randomness, which can reduce instruction execution consistency. Very large batch sizes may lead to overly smooth updates, weakening the modeling of fine-grained category differences. The proposed method maintains good stability within a reasonable range of batch sizes. This demonstrates that end-to-end instruction alignment and constrained output can effectively mitigate decision drift caused by changes in training hyperparameters. It provides more robust technical support for text classification tasks that require controllability and consistency in real-world scenarios.

 doi:10.20944/preprints202602.0525.v1

V. Conclusion

This study focuses on controllable prompt-driven text classification and systematically examines how instruction understanding, semantic alignment, and output constraints can be jointly integrated into the decision process of large language models. By modeling text classification as an instruction-conditioned generative discriminative task, the study highlights the importance of explicitly incorporating task semantics and category boundaries during inference. This design avoids the instability caused by excessive reliance on implicit representations in conventional approaches. The findings show that, when classification decisions are constrained within an interpretable and structured output space, large language models can exhibit more consistent and reliable discriminative behavior in complex textual scenarios. This provides a new modeling perspective for building controllable intelligent systems.

From a methodological standpoint, the proposed end-to-end framework tightly couples instruction alignment with constrained decoding. This enables the model to go beyond surface-level semantic matching of the input text and to continuously follow the decision logic defined by task instructions. This modeling strategy moves beyond treating prompts as external heuristics. Instead, prompts become a core component that shapes internal representations and output mechanisms. Under this formulation, text classification is no longer a simple label prediction problem. It becomes a structured decision process governed by semantic constraints. This is crucial for improving controllability in complex tasks and high-reliability scenarios.

At the application level, this research provides a technically practical foundation for many real-world scenarios that rely on text classification. Examples include information filtering, content moderation, risk identification, and decision support. In such domains, systems must ensure classification accuracy while strictly complying with predefined rules and business constraints. The controllable prompting and constrained output mechanisms proposed in this study allow models to better satisfy these requirements. They reduce the risk of unpredictable behavior and enhance overall system trustworthiness and interpretability. These properties have positive implications for deploying large language models in real production environments.

Looking ahead, the research perspective presented in this work lays the groundwork for extending large language models to a broader range of controlled decision tasks. On the one hand, the framework can be generalized to more complex classification structures and multi-level decision scenarios. Finer instruction design and constraint mechanisms may enable higher-level controllable reasoning. On the other hand, integrating controllable prompting with domain knowledge, rule-based systems, or human-in-the-loop mechanisms offers a promising direction for building safer, more robust, and sustainably evolving intelligent text analysis systems. Overall, this study deepens the understanding of prompt-driven text classification mechanisms. It also provides a valuable reference path for the standardized and trustworthy development of large language model applications.

References

1. Y. Zhu, Y. Wang, J. Qiang, et al., “Prompt-learning for short text classification,” IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 10, pp. 5328-5339, 2023.
2. S. Geng, M. Josifoski, M. Peyrard, et al., “Grammar-constrained decoding for structured NLP tasks without finetuning,” arXiv preprint arXiv:2305.13971, 2023.
3. T. Kim, J. Kim, G. Lee, et al., “Instructive decoding: Instruction-tuned large language models are self-refiner from noisy instructions,” arXiv preprint arXiv:2311.00233, 2023.
4. C. Zhu, B. Xu, Q. Wang, et al., “On the calibration of large language models and alignment,” arXiv preprint arXiv:2311.13240, 2023.
5. C. F. Chiang, D. Li, R. Ying, Y. Wang, Q. Gan and J. Li, “Deep Learning-Based Dynamic Graph Framework for Robust Corporate Financial Health Risk Prediction,” 2025.

 doi:10.20944/preprints202602.0525.v1

6. Z. Xu, K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative Distribution Modeling for Credit Card Risk Identification under Noisy and Imbalanced Transactions," 2025.
7. A. Xie and W. C. Chang, "Deep Learning Approach for Clinical Risk Identification Using Transformer Modeling of Heterogeneous EHR Data," arXiv preprint arXiv:2511.04158, 2025.
8. X. Zheng, H. Lin, X. Han, et al., "Toward unified controllable text generation via regular expression instruction," arXiv preprint arXiv:2309.10447, 2023.
9. Y. Wang, W. Wang, Q. Chen, et al., "Prompt-based zero-shot text classification with conceptual knowledge," Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 30-38, 2023.
10. S. Pan and D. Wu, "Trustworthy summarization via uncertainty quantification and risk awareness in large language models," arXiv preprint arXiv:2510.01231, 2025.
11. T. Guan, "A Multi-Agent Coding Assistant for Cloud-Native Development: From Requirements to Deployable Microservices," 2025.
12. Z. Cheng, "Enhancing Intelligent Anomaly Detection in Cloud Backend Systems through Contrastive Learning and Sensitivity Analysis," Journal of Computer Technology and Software, vol. 3, no. 4, 2024.
13. W. C. Chang, L. Dai and T. Xu, "Machine Learning Approaches to Clinical Risk Prediction: Multi-Scale Temporal Alignment in Electronic Health Records," arXiv preprint arXiv:2511.21561, 2025.
14. S. Wang, S. Han, Z. Cheng, M. Wang and Y. Li, "Federated fine-tuning of large language models with privacy preservation and cross-domain semantic alignment," 2025.
15. X. Guo, Y. Luan, Y. Kang, X. Song and J. Guo, "LLM-Centric RAG with Multi-Granular Indexing and Confidence Constraints," arXiv preprint arXiv:2510.27054, 2025.
16. R. Ying, Q. Liu, Y. Wang and Y. Xiao, "AI-Based Causal Reasoning over Knowledge Graphs for Data-Driven and Intervention-Oriented Enterprise Performance Analysis," 2025.
17. Y. Ou, S. Huang, R. Yan, K. Zhou, Y. Shu and Y. Huang, "A Residual-Regulated Machine Learning Method for Non-Stationary Time Series Forecasting Using Second-Order Differencing," 2025.
18. R. Liu, L. Yang, R. Zhang and S. Wang, "Generative Modeling of Human-Computer Interfaces with Diffusion Processes and Conditional Control," arXiv preprint arXiv:2601.06823, 2026.
19. Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint arXiv:2512.09440, 2025.
20. K. Cao, Y. Zhao, H. Chen, X. Liang, Y. Zheng and S. Huang, "Multi-Hop Relational Modeling for Credit Fraud Detection via Graph Neural Networks," 2025.
21. S. Li, Y. Wang, Y. Xing and M. Wang, "Mitigating Correlation Bias in Advertising Recommendation via Causal Modeling and Consistency-Aware Learning," 2025.
22. K. Gao, H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems," 2025.
23. N. Lyu, Y. Wang, F. Chen and Q. Zhang, "Advancing Text Classification with Large Language Models and Neural Attention Mechanisms," arXiv preprint arXiv:2512.09444, 2025.
24. Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint arXiv:2512.09440, 2025.
25. Y. Xing, M. Wang, Y. Deng, H. Liu and Y. Zi, "Explainable Representation Learning in Large Language Models for Fine-Grained Sentiment and Opinion Classification," 2025.
26. Z. Ren, "VGCN: An enhanced graph convolutional network model for text classification," Journal of Industrial Engineering and Applied Science, vol. 2, no. 4, pp. 110-115, 2024.
27. Y. Zhang, R. Yang, X. Xu, et al., "Teleclass: Taxonomy enrichment and LLM-enhanced hierarchical text classification with minimal supervision," Proceedings of the ACM on Web Conference 2025, pp. 2032-2042, 2025.
28. H. Gao, P. Zhang, J. Zhang, et al., "Qsim: a quantum-inspired hierarchical semantic interaction model for text classification," Neurocomputing, vol. 611, p. 128658, 2025.

 doi:10.20944/preprints202602.0525.v1

29. L. Ren, Y. Liu, C. Ouyang, et al., "DyLas: A dynamic label alignment strategy for large-scale multi-label text classification," Information Fusion, vol. 120, p. 103081, 2025.
30. C. Zhou, J. Dong, X. Huang, et al., "Quest: Efficient extreme multi-label text classification with large language models on commodity hardware," Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 3929-3940, 2024.