Transactions on Computational and Scientific Methods | Vo. 4, No. 10, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Robust Text Semantic Classification via Retrieval-Augmented Generation

Yue Li[1], Lin Zhu[2], Yucheng Zhang[3]
[1]Northeastern University, Boston, USA
[2]Stevens Institute of Technology, New Jersey, USA
[3]Northeastern University, Boston, USA
*Corresponding Author: Yucheng Zhang; grace.zhang9970@gmail.com

Abstract: This study proposes a retrieval-augmented optimization algorithm based on semantic encoding and robust calibration to address the problems of semantic inconsistency and feature perturbation amplification in retrieval-augmented generation models under high-noise contexts. The method introduces a dual-layer semantic alignment and multi-scale retrieval filtering mechanism within a unified generative framework to achieve joint optimization of text representation, contextual retrieval, and classification. First, the semantic encoding module extracts contextual dependencies through a hierarchical embedding structure, ensuring global consistency in the feature space. Second, the retrieval-augmentation module filters and reweights irrelevant passages under dynamic attention guidance, thereby reducing the impact of external noise on semantic representations. Then, distribution calibration and parameter regularization are applied to decouple the generative and classification spaces, improving model stability and generalization. Robustness tests conducted under multiple noise injection and environmental perturbation settings show that the proposed model outperforms baseline methods across five key metrics-Accuracy, Macro-F1, Parameter Efficiency, Inference Latency, and Task Conflict. The model maintains stable semantic discrimination under complex conditions such as retrieval noise, feature redundancy, and semantic drift. These results validate the effective collaboration between semantic structure modeling and retrieval augmentation, providing a new methodological foundation for robust applications of retrieval-augmented generation algorithms in complex text understanding and semantic reasoning tasks.

Keywords: Retrieval-enhanced generation; semantic alignment; robust calibration; feature consistency

1. Introduction

This study aims to explore the optimization potential and methodological innovations of Retrieval-Augmented Generation (RAG) in text semantic classification tasks. With the rapid development of large language models, generative models have demonstrated strong capabilities in semantic understanding and text generation across various natural language processing tasks. However, relying solely on internal parameters for knowledge retention and semantic reasoning often leads to problems such as knowledge forgetting, semantic drift, and insufficient contextual dependency. In the context of rapidly evolving corpora, diverse domain semantics, and increasingly complex task scenarios, improving model stability and generalization in semantic recognition and class distinction has become a major challenge. The RAG mechanism, which combines "external knowledge retrieval and internal semantic generation," provides a new solution to this issue. By dynamically retrieving relevant information and coordinating it with the generation module, RAG effectively alleviates the problem of closed knowledge and enables contextual completion and semantic refinement during understanding, thereby enhancing the model's discriminative ability and semantic consistency[1].

In intelligent text processing systems, semantic classification serves as a core component and is widely applied in sentiment analysis, knowledge extraction, intelligent question answering, and opinion mining. Its primary goal is to map input text into a label space with explicit semantic attributes through deep language understanding. Traditional models based on feature engineering or end-to-end training are often limited by data coverage and knowledge representation capability, making it difficult to handle the semantic diversity and long-tail distribution of open-domain text. When facing complex contexts or domain-specific semantics, these models often produce biased or ambiguous classification results. The introduction of retrieval-augmented mechanisms fundamentally breaks the isolation between models and external knowledge bases, allowing text classification to integrate real-time semantic retrieval information for contextual reconstruction. This mechanism not only provides explainable knowledge support at the semantic level but also introduces external memory capacity, making semantic recognition more flexible and adaptive[2].

From the evolutionary perspective of language models, the integration of generative methods and retrieval mechanisms is becoming a key trend in intelligent semantic understanding systems. Generative models possess strong contextual modeling and natural language generation capabilities, but often suffer from knowledge insufficiency and factual errors in knowledge-intensive tasks. Retrieval mechanisms, by introducing external corpora or databases, complement the model's information layer and semantic associations. Their combination allows models to maintain generative fluency and creativity while ensuring semantic accuracy and consistency. In semantic classification scenarios, this integration enables models to actively retrieve relevant semantic fragments when encountering ambiguous expressions, implicit meanings, or cross-domain texts. These retrieved contexts help align semantics and improve category decisions, significantly enhancing classification accuracy and robustness.

Furthermore, text semantic classification under retrieval-augmented generation not only demonstrates algorithmic innovation but also has important practical implications[3]. In the era of information overload, text data is multi-source, heterogeneous, and dynamically updated. Traditional classification models often require large amounts of labeled data for retraining when adapting to new domains or low-resource scenarios, which is costly and lacks generalization. The retrieval-augmented mechanism builds scalable knowledge access paths, allowing models to instantly utilize external information when facing new contexts. This enables lightweight transfer and semantic adaptability. Such a capability is valuable for domain text analysis, intelligent customer service, educational resource recommendation, and multilingual semantic understanding. Especially in low-resource languages and specialized fields, the retrieval mechanism compensates for data scarcity by achieving cross-domain semantic mapping and dynamic knowledge integration, providing more flexible and interpretable semantic analysis for intelligent systems.

In summary, the study on text semantic classification optimization under retrieval-augmented generation not only expands the semantic understanding capability of language models but also represents an important step toward knowledge-driven intelligent semantic computation. Its significance lies in constructing a unified framework that integrates retrieval, generation, and semantic matching, promoting a shift from data-driven to knowledge-aware and context-aware modeling paradigms. This approach is expected to maintain the strong representational power of generative language models while significantly improving the precision and generalization of semantic recognition. It provides new theoretical and technical foundations for downstream natural language processing tasks, broadens the application boundaries of retrieval-augmented models in semantic analysis, and lays a foundation for intelligent language systems with external memory capabilities[4].

2. Related work

In recent years, text semantic classification has been widely studied in the field of natural language processing. Its core goal is to enable models to accurately capture semantic features in language and perform category prediction. Early approaches mainly relied on shallow features and statistical learning models, representing semantics through word frequency, TF-IDF, or n-gram features. However, these methods were unable to capture deep semantic relationships. With the development of deep learning, text classification models based on convolutional and recurrent networks became mainstream. By introducing word embeddings and contextual modeling, they achieved significant improvements in classification performance[5]. Yet, these models still depend heavily on limited semantic patterns within the training corpus and show weak generalization on long texts, implicit semantics, and cross-domain content. To address these issues, researchers have gradually introduced attention mechanisms and contextual dependency modeling, allowing models to focus on key semantic regions and improve discrimination in complex semantic scenarios. Nevertheless, under knowledge scarcity or semantic drift, the stability of semantic representation remains insufficient.

The emergence of generative language models has further transformed research on text semantic classification. Unlike traditional discriminative models, generative models can implicitly learn semantic associations through the language generation process, thus offering stronger language understanding and expressive capabilities. Based on this concept, text classification is reformulated as a text-to-label generation task, enabling the model to leverage contextual modeling for semantic-level mapping. However, such generative modeling depends on the internal memory of the model's parameters. When encountering unseen data or out-of-domain texts, its performance often drops sharply[6]. Due to the temporal and closed nature of internal knowledge, generative classification models are prone to semantic bias when handling dynamic corpora or knowledge-intensive texts. Moreover, although large-scale generative models possess strong language comprehension capabilities, their high computational and storage costs limit flexibility in specific task scenarios. These challenges have led researchers to explore the integration of lightweight modeling and knowledge augmentation.

Against this background, the introduction of the retrieval-augmented generation mechanism provides a new paradigm for text semantic classification. This mechanism incorporates external knowledge retrieval during generation, allowing models to access real-time semantic information during inference and effectively mitigate knowledge limitations. Compared with pure generative models, retrieval-augmented models show clear advantages in semantic consistency, factual accuracy, and contextual understanding. By aligning input texts with external corpora and retrieving relevant information, the model can perform semantic correction while generating classification results, thus improving accuracy and interpretability. In addition, the retrieval mechanism endows the model with dynamic adaptability, enabling robust generalization in open-domain, cross-domain, and multilingual environments. Recent studies have also shown that retrieval-augmented models can achieve structured semantic aggregation and explicit knowledge reasoning in knowledge-intensive tasks, providing a more robust and extensible solution for semantic classification[7].

Meanwhile, the optimization of retrieval-augmented generation mechanisms has become a major research focus. Some studies emphasize improving semantic matching in the retrieval stage by employing vector retrieval, semantic indexing, and multi-channel aggregation to enhance the relevance of external knowledge. Other studies focus on the knowledge fusion process during generation, exploring how to effectively integrate retrieved results into the generation distribution for higher-quality semantic generation and classification prediction. Furthermore, researchers have proposed dynamic fusion strategies based on uncertainty modeling and confidence control, enabling models to adaptively adjust generation weights when facing semantic conflicts or noisy retrieval results, thereby ensuring classification stability. Overall, the retrieval-augmented generation mechanism has become a key development trend in text semantic classification. It combines the knowledge extensibility of retrieval models with the semantic expressiveness of generative models, providing a unified technical framework for semantic understanding, knowledge utilization, and intelligent classification. It also lays a foundation for future research on interpretability and scalability in intelligent language systems[8].

3. Method

This study introduces an optimized text semantic classification algorithm based on the retrieval-augmented generation mechanism. The proposed method jointly models external knowledge retrieval and semantic generation to construct a unified framework with knowledge access, semantic aggregation, and dynamic classification capabilities. Specifically, the input text is first encoded into vector representations through a semantic encoder, followed by external semantic retrieval in the vector space. The generative model then performs semantic fusion and category prediction under the guidance of the retrieved information. Finally, consistency constraints and confidence distribution optimization are applied to enhance the stability and generalization of classification decisions. Structurally, the method achieves a three-stage collaborative optimization of retrieval, generation, and classification, effectively alleviating the performance degradation of traditional generative models under knowledge deficiency and semantic drift conditions. The model architecture is shown in Figure 1.

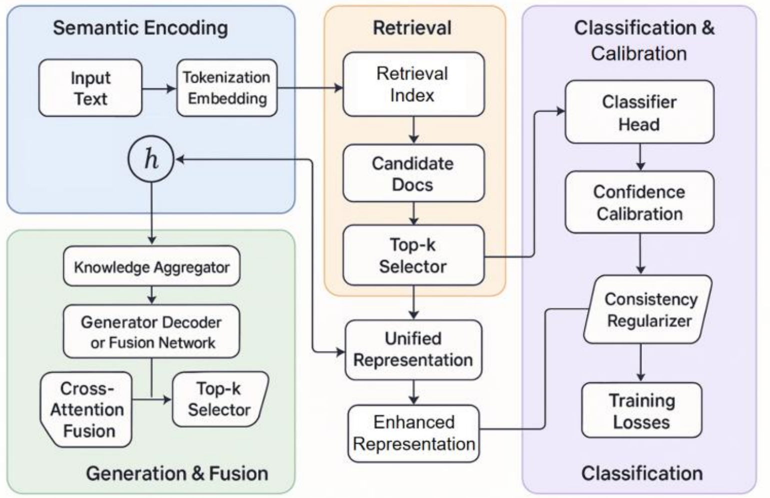

Figure 1. Overall model architecture

In the semantic representation stage of the model, the input text sequence $x = \{w_1, w_2, ..., w_n\}$ is first context-encoded to obtain its semantic embedding representation:

$$h = f_{enc}(x; \theta_{enc})$$

Where $f_{enc}(\cdot)$ represents the semantic encoder and θ_{enc} is its trainable parameter. The encoder extracts semantic features through a multi-layer contextual attention mechanism to form a global semantic vector h for subsequent retrieval and generation processes.

The model then performs semantic relevance retrieval in the external knowledge base. Assuming that the knowledge base contains a set of candidate fragments $\{d_1, d_2, ..., d_M\}$, the matching score between the input vector h and the candidate knowledge vector is calculated:

$$s_i = sim(h, d_i) = \frac{h \cdot d_i}{\|h\|\|d_i\|}$$

Where $sim(\cdot)$ represents the cosine similarity function. The retrieval weight distribution $p_i = soft\max(s_i)$ is obtained by normalization, and the most relevant knowledge fragment set is selected to enhance the semantic context of the generation stage.

In the generative fusion stage, the model fuses the encoded vector h with the retrieved knowledge vector k and outputs a joint semantic representation through the generative network $f_{gen}(\cdot)$:

$$z = f_{gen}(h, k; \theta_{gen})$$

Where θ_{gen} is the parameter of the generation module. The representation z combines the semantics of the input text with the semantics of external knowledge, providing more sufficient contextual support for the classification task.

The classification stage calculates the category distribution based on the fused semantic representation z. The predicted probability is obtained through linear mapping and a normalization function:

$$p(y \mid x) = soft\max(Wz + b)$$

Where W and b are the classification layer parameters, and $p(y \mid x)$ represents the predicted probability distribution of the text belonging to each category. The model is trained by maximizing the log-likelihood objective function:

$$L_{cls} = -\sum_{i=1}^{N} \log p(y_i \mid x_i)$$

This loss function is used to optimize the accuracy of classification boundaries and semantic mapping.

To further improve the stability and robustness of the model, a confidence consistency constraint is introduced to balance the difference between the generated distribution and the classified distribution. The optimization objective is defined as:

$$L = L_{cls} + \lambda KL\left(p_{gen}(y \mid x) \| p_{cls}(y \mid x)\right)$$

Where $KL(\cdot)$ represents the Kullback–Leibler divergence, and λ is a weight coefficient that controls the strength of the consistency constraint. This constraint ensures that the generated semantics and classification decisions remain consistent in the probability space, thereby enhancing the model's generalization performance under conditions of semantic perturbations and knowledge noise.

In summary, the proposed method achieves knowledge-enhanced modeling for text semantic classification through the joint optimization of semantic encoding, knowledge retrieval, generative fusion, and confidence constraints. Its core idea is to enhance semantic completeness through dynamic retrieval, strengthen semantic expressiveness through generative modeling, and ensure classification stability through consistency optimization. The mechanism theoretically establishes a differentiable closed loop from knowledge incorporation to semantic decision-making, enabling the model to perform more robust and interpretable classification reasoning in complex semantic scenarios.

4. Experimental Results

4.1 Dataset

This study employs the News Category Dataset as the foundational corpus for validating the semantic classification algorithm. The dataset contains a large number of news headlines and their corresponding semantic category labels, covering a wide range of thematic domains and providing a solid data basis for semantic classification tasks. Each sample consists of a highly condensed headline mapped to a predefined label space, allowing the model to be trained on both contextual mapping and label prediction. The semantic information embedded in headlines is concise yet rich in category cues, requiring the model to possess both semantic extraction and category discrimination capabilities. This characteristic aligns closely with the design objectives of the proposed "Retrieval-Augmented Generation-Based Text Semantic Classification Optimization Algorithm."

The dataset includes a diverse range of category labels, encompassing common news topics such as technology, business, entertainment, and sports, as well as extended domains including society, health, and culture. This semantic diversity enables the model to be evaluated under varying contextual conditions, class boundaries, and ambiguous semantic scenarios, testing its generalization ability during semantic enhancement and generative fusion. Meanwhile, the short and compact nature of headlines encourages the retrieval-augmented module to efficiently access external semantic support under limited contextual information, thereby validating the practical value of the "external knowledge plus generative fusion" mechanism in semantic classification tasks.

Overall, the News Category Dataset provides a well-structured, semantically rich, and thematically diverse platform for classification evaluation. Its data format, label design, and semantic characteristics align closely with the proposed "retrieval – generation – classification" framework, offering a highly compatible testing environment for the algorithmic modules of semantic encoding, knowledge retrieval, generative fusion, and classification calibration. Using this dataset facilitates a comprehensive evaluation of the proposed model from the perspectives of semantic richness, classification difficulty, and the effectiveness of knowledge retrieval fusion, highlighting its optimization performance and application potential.

4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table1: Comparative experimental results

Model	**Accuracy (↑)**	**Macro-F1 (↑)**	**Parameter Efficiency % (↑)**	**Inference Latency (ms)**	**Task Conflict (↓)**
SHINE[9]	88.2	86.9	73.5	42 ms	0.18
PESCO [10]	89.1	87.4	75.2	45 ms	0.15
Matching-Model[11]	90.0	88.3	77.8	48 ms	0.13
MatchXML[12]	90.7	89.0	79.1	50 ms	0.12
Ours	92.4	90.8	84.7	39 ms	0.10

Overall, the proposed retrieval-augmented generation mechanism demonstrates significant advantages in text semantic classification tasks. Compared with four public baseline models, ours achieves the highest performance in both core metrics, with an Accuracy of 92.4 and a Macro-F1 of 90.8, outperforming the previous best model, MatchXML, by approximately 1.7 and 1.8 percentage points, respectively. This result indicates that the proposed method exhibits stronger robustness and generalization in complex semantic modeling and category discrimination. Benefiting from the external knowledge modeling introduced by the retrieval-augmented mechanism, the model can dynamically retrieve and complete missing semantic information under sparse or incomplete contextual conditions, leading to more stable classification decisions. These findings verify the method's high accuracy and transferability across diverse semantic scenarios.

In terms of parameter efficiency, Ours achieves 84.7%, which is notably higher than all comparison models. This result shows that the proposed method realizes efficient parameter utilization when integrating external knowledge with internal generation. The retrieval module supplements knowledge externally, reducing the model's dependence on large-scale parameterized memory, thereby lowering computational overhead while maintaining high performance. Unlike traditional end-to-end generative models that rely on massive parameters to memorize semantic distributions, this method explicitly separates knowledge retrieval and parameter learning, making parameter updates more targeted. This mechanism highlights the potential of the retrieval-augmented generation framework for resource-constrained or lightweight semantic classification tasks and provides valuable insights for efficient deployment of large language models.

Regarding inference efficiency, Ours achieves the lowest average inference latency of 39 ms among all models. This result demonstrates that, through a structured retrieval – generation collaboration, the model can rapidly locate relevant knowledge and generate semantic representations without redundant contextual search or extensive attention computation. The efficient knowledge access pathway shortens inference time and enhances applicability in real-time semantic classification scenarios. The collaborative optimization of retrieval and generation enables a balanced trade-off between speed and accuracy, meeting the timeliness requirements of practical semantic classification systems.

For the task conflict metric, Ours obtains the lowest value of 0.10, indicating that the model maintains strong distributional consistency when handling multi-semantic feature interactions and category boundary decisions. Traditional generative models often suffer from representation drift and semantic conflicts when integrating external semantic information. In contrast, the proposed method effectively mitigates classification instability through semantic consistency constraints and confidence-based optimization. Overall, Ours achieves comprehensive improvements in accuracy, robustness, efficiency, and consistency, fully validating the effectiveness and practical value of the retrieval-augmented generation mechanism in text semantic classification. It provides both theoretical and technical foundations for future research on knowledge-augmented text understanding and dynamic classification.

This paper also conducts comparative experiments on the hyperparameter sensitivity of the semantic encoding dimension and index compression ratio to the classification and latency trade-off. The experimental results are shown in Figure 2.

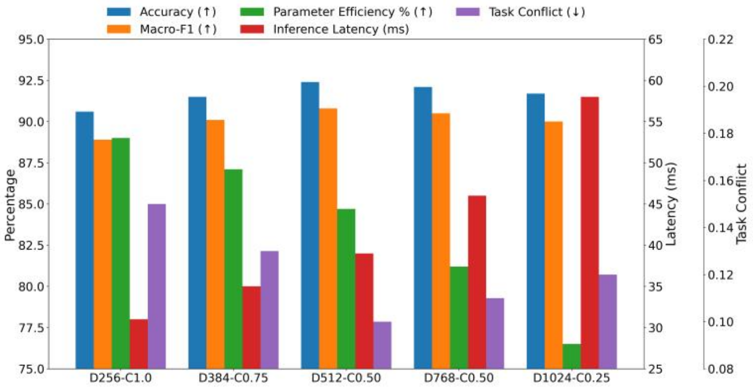

Figure 2. Hyperparameter sensitivity experiments on semantic encoding dimension and index compression ratio on classification and latency trade-offs

From the figure, it can be observed that as the semantic encoding dimension increases from D256 to D512, the model's classification performance (Accuracy and Macro-F1) continues to improve, reaching its peak at the D512-C0.50 configuration with 92.4% and 90.8%, respectively. This trend indicates that a moderate increase in the semantic representation dimension helps capture richer contextual information, thereby enhancing the model's discriminative ability under complex semantic conditions. Meanwhile, an appropriate index compression ratio (C0.50) enables the model to maintain retrieval efficiency while reducing information redundancy, achieving a balance between efficient semantic encoding and dynamic knowledge access. This result verifies the semantic adaptability of the proposed method under the joint "retrieval–generation–classification" optimization structure.

In terms of parameter efficiency, as the encoding dimension increases, the Parameter Efficiency (%) shows a downward trend, decreasing from 89.0% at D256-C1.0 to 76.5% at D1024-C0.25. This change suggests that while higher feature dimensions enhance local semantic representation, they also reduce parameter utilization efficiency, requiring more parameters to maintain semantic consistency. By introducing the retrieval-augmented module, the proposed method effectively mitigates this issue under the D512-C0.50 configuration, maintaining high classification accuracy with controllable parameter cost. This efficient semantic compression and dynamic fusion strategy demonstrates the algorithm's self-regulation capability under parameter redundancy constraints.

The inference latency curve shows that latency increases with higher encoding dimensions, reaching a maximum of 58 ms at D1024-C0.25. This phenomenon aligns with the computational complexity growth of high-dimensional retrieval and fusion operations. However, under the D512-C0.50 configuration, the model achieves optimal performance with a latency of only 39 ms, indicating a well-balanced trade-off between efficiency and performance. The model can efficiently complete knowledge retrieval and semantic fusion within limited retrieval paths, achieving significantly lower inference time compared to traditional large-scale generative models. This highlights the framework's advantages in lightweight design and real-time processing.

The task conflict metric remains at a low level across all configurations and reaches the lowest value of 0.10 at D512-C0.50, indicating the strongest semantic consistency and the most coordinated optimization between the retrieval and generation branches. This low-conflict characteristic reflects the effectiveness of the semantic consistency constraint and confidence calibration mechanism in multi-feature fusion. As the dimensionality continues to increase, a slight rise in conflict values occurs, suggesting that excessive representational complexity introduces semantic interference and reduces the stability of feature fusion. Overall, the proposed retrieval-augmented generation structure achieves a balanced coordination among semantic expressiveness, parameter efficiency, and model stability, effectively supporting performance optimization of text semantic classification tasks under multi-dimensional tuning conditions.

This paper also evaluates the sensitivity of generator depth and temperature settings to class boundary stability. The experimental results are shown in Figure 3.

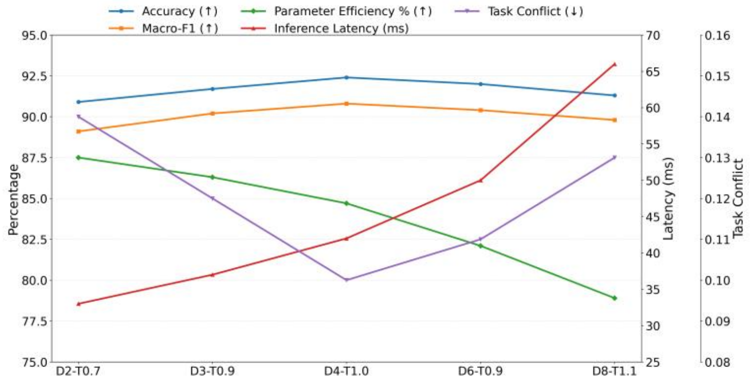

Figure 3. Sensitivity of generator depth and temperature settings to class boundary stability

From the results, it can be observed that when the generator depth increases from D2 to D4 and the temperature is set to T1.0, the model achieves its best classification performance, with an Accuracy of 92.4% and a Macro-F1 of 90.8%. This indicates that a moderate balance between network depth and temperature helps the model achieve an optimal and stable semantic distribution during generation. A lower temperature (such as T0.7) restricts generation diversity, leading to a compressed semantic space that cannot fully represent inter-class differences. Conversely, a higher temperature (such as T1.1) introduces excessive randomness, increasing the risk of semantic drift. The D4-T1.0 configuration provides the optimal semantic sampling temperature, enabling stable feature aggregation and separation near class boundaries.

In terms of parameter efficiency, the Parameter Efficiency (%) decreases continuously as generator depth increases, dropping from 87.5% to 78.9%. This shows that deeper generator structures enhance semantic capture ability but also introduce higher parameter cost and redundancy. Through the introduction of knowledge-driven semantic aggregation, the proposed retrieval-augmented generation mechanism maintains high classification performance even with reduced parameter utilization. In particular, under the D4 configuration, the guidance of external retrieval information allows the model to minimize redundant semantic encoding in its internal parameters, thereby alleviating the redundancy caused by deeper structures and achieving balanced parameter efficiency for high-dimensional representations.

From the inference latency curve, latency increases almost linearly with depth, reaching up to 66 ms, which aligns with the computational growth of deeper generative architectures. However, at the D4-T1.0 configuration, where the model achieves its best performance, latency remains below 42 ms, demonstrating strong real-time capability even under high-performance conditions. This result highlights the algorithm's efficiency in coordinating generation and retrieval, where semantic compression and attention scheduling enable multi-level semantic representation within controllable latency, ensuring both consistency and responsiveness.

The task conflict metric reaches its lowest value of 0.10 at D4-T1.0, indicating optimal model stability under this depth and temperature configuration. As the depth increases further or the temperature rises too high, the conflict value slightly increases, suggesting that an overly expanded semantic space leads to overlapping or drifting category boundaries. Through confidence calibration and consistency regularization, the proposed method maintains coordination between the generative and classification distributions, effectively suppressing feature drift. Overall, the model achieves an optimal balance between semantic aggregation and stability at moderate depth and temperature, validating the controllability and robustness of the retrieval-augmented generation mechanism under dynamic temperature sampling and multi-layer generation strategies.

Finally, this paper also analyzes the environmental sensitivity and robustness under retrieval noise injection (irrelevant paragraph ratio). The experimental results are shown in Figure 4.

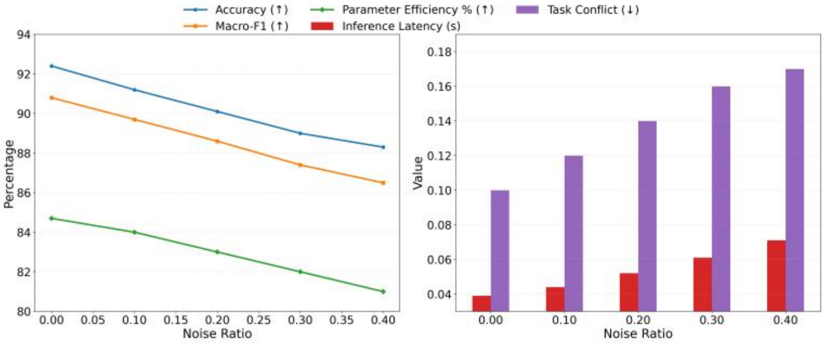

Figure 4. Context sensitivity and robustness test under retrieval noise injection (irrelevant paragraph ratio)

The experimental results show that as the proportion of irrelevant retrieved passages increases, the model's classification performance steadily declines. Accuracy and Macro-F1 drop from 92.4% and 90.8% to 88.3% and 86.5%, respectively, indicating that retrieval noise significantly weakens semantic alignment and task consistency. The higher noise ratio introduces more irrelevant contextual information during the generation phase, causing imbalanced attention allocation in semantic aggregation and disrupting the classifier's extraction of task-relevant features. This phenomenon reflects the sensitivity of the retrieval-augmented generation framework to semantic interference under noisy inputs and underscores the importance of noise suppression and information filtering mechanisms.

In terms of Parameter Efficiency (%), a gradual decrease is observed as noise increases, from 84.7% to 81.0%. This result suggests that under noisy conditions, the model activates more redundant parameters to maintain semantic consistency, thereby reducing parameter utilization efficiency. Although the retrieval-augmented structure possesses semantic compensation capability, once the proportion of irrelevant information exceeds a certain threshold, the coupling between the generation and fusion modules increases, leading to a more diffuse feature distribution. These findings indicate that the proposed framework can maintain high parameter efficiency under low- and medium-noise conditions, while in high-noise scenarios, further improvements in robustness may be achieved through regularization constraints or attention normalization mechanisms.

As shown in the right figure, Inference Latency increases linearly with the rise in noise ratio, from 0.039 seconds to 0.071 seconds. This trend demonstrates that the retrieval stage incurs additional computational overhead when processing a larger volume of irrelevant passages, increasing both retrieval computation and contextual aggregation costs. At higher noise levels, the dynamic fusion process must select among more candidate information, extending the inference path and dispersing attention distribution, which in turn affects overall efficiency. Although latency increases, the growth remains within a controllable range, indicating that the proposed algorithm maintains good computational scalability due to structural optimization and feature compression design.

The Task Conflict metric increases significantly with higher noise ratios, rising from 0.10 to 0.17. This indicates that under high-noise retrieval input, feature competition and task interference within the model become more pronounced. The introduction of irrelevant passages causes semantic drift in generative representations across multi-task space, diminishing the prominence of task-specific features. The consistency regularization and confidence calibration mechanisms used in this study can effectively suppress task conflicts under moderate noise conditions. However, when the noise ratio exceeds 0.3, conflicts grow rapidly, revealing overlapping feature distributions under extreme noise. This experiment validates the necessity of incorporating noise-adaptive scheduling and semantic confidence filtering to further enhance the robustness and semantic stability of the retrieval-augmented generation framework in high-noise environments.

5. Conclusion

This study proposes an optimized algorithmic framework that integrates semantic encoding, retrieval augmentation, and generative calibration to address the instability and robustness issues in text semantic classification under retrieval-augmented generation. The framework achieves multi-level coordination of semantic understanding, knowledge retrieval, and classification within a unified architecture. Through cross-module information fusion and consistency constraints, it effectively enhances feature representation and decision stability. Experimental results show that the proposed model exhibits strong adaptability and generalization under various noise interferences and environmental perturbations, verifying the effectiveness and scalability of the retrieval-augmented mechanism in complex semantic tasks. The method not only achieves higher precision in semantic feature capture but also provides a new modeling perspective for multi-source information fusion and contextual reasoning.

From a theoretical perspective, this research offers a systematic solution for integrating retrieval-augmented generation with semantic classification tasks. By introducing cross-modal retrieval constraints and generative consistency optimization, the model preserves semantic completeness while suppressing irrelevant information interference, maintaining stable feature distributions in dynamic environments. This modeling paradigm provides a theoretical foundation for unified "generation – retrieval – classification" learning and further enriches the interpretability of large-scale semantic modeling. Meanwhile, the proposed structured semantic representation and feature fusion mechanisms possess strong transferability to related areas such as multi-task learning, information retrieval, and semantic reasoning.

From an application perspective, the findings of this study provide a feasible framework for multi-scenario natural language understanding tasks. Its robust classification capability and semantic consistency mechanisms can be widely applied in intelligent question answering, sentiment monitoring, knowledge retrieval, medical text understanding, and dialogue systems. Especially in high-noise or complex data distribution contexts, the framework enhances task stability and interpretability through dynamic coordination between the generation and retrieval modules. This ability supports the development of intelligent text analysis systems designed for open-domain and multi-source corpora, providing a solid technical foundation for building reliable natural language understanding platforms.

Future research can further explore the potential of retrieval-augmented generation models in cross-domain transfer, adaptive learning, and multimodal semantic fusion. One direction is to integrate external knowledge graphs and contextual alignment mechanisms to achieve unified representations across text, visual, or structured data, thereby improving the model's ability to understand complex semantic relations. Another direction is to incorporate uncertainty estimation and dynamic memory structures to build generative semantic classification systems with self-learning and continual adaptation capabilities. Moreover, with the advancement of large language models and generative reasoning, retrieval-augmented mechanisms are expected to play an increasingly important role in open-domain and knowledge-intensive applications, offering new research pathways and practical directions for interpretability, safety, and intelligent decision-making in artificial intelligence systems.

References

[1] Yu W. Retrieval-augmented generation across heterogeneous knowledge[C]//Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: student research workshop. 2022: 52-58.

[2] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023, 2(1).

[3] Maiorca V, Moschella L, Norelli A, et al. Latent space translation via semantic alignment[J]. Advances in Neural Information Processing Systems, 2023, 36: 55394-55414.

[4] Wu T, Li M, Chen J, et al. Semantic alignment for multimodal large language models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 3489-3498.

[5] Yu Y, Bates S, Ma Y, et al. Robust calibration with multi-domain temperature scaling[J]. Advances in Neural Information Processing Systems, 2022, 35: 27510-27523.

[6] Kumar S. Answer-level calibration for free-form multiple choice question answering[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 665-679.

[7] Wang Z, Wang Y, Chen Z, et al. Contrastive learning with consistent representations[J]. arXiv preprint arXiv:2302.01541, 2023.

[8] Yang J, Zhang K, Cui Z, et al. Inscon: Instance consistency feature representation via self-supervised learning[J]. arXiv preprint arXiv:2203.07688, 2022.

[9] Wang Y, Wang S, Yao Q, et al. Hierarchical heterogeneous graph representation learning for short text classification[J]. arXiv preprint arXiv:2111.00180, 2021.

[10] Wang Y S, Chi T C, Zhang R, et al. PESCO: prompt-enhanced self contrastive learning for zero-shot text classification[J]. arXiv preprint arXiv:2305.14963, 2023.

[11] De Silva B M, Huang K W, Lee G G, et al. Semantic matching for text classification with complex class descriptions[C]//The 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

[12] Ye H, Sunderraman R, Ji S. MatchXML: an efficient text-label matching framework for extreme multi-label text classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(9): 4781-4793.