

Research on Five-Category Classification of Comment Emotions Based on the Multi-agent RAG Framework

Baohui Zhu

*Department of Software, Jilin University, Changchun, China
zhubh5525@mails.jlu.edu.cn*

Abstract. Because text comment sentiment analysis plays an essential role in applications such as online public opinion monitoring, user experience analysis, and intelligent customer service, it is rightly considered a core task of natural language processing with both great practical value and considerable research significance. However, it is also well known that current mainstream single-text sentiment classification models have inherent limitations. Because existing approaches suffer from superficial semantic understanding, insufficient logical reasoning capabilities, vulnerability to model hallucinations, and low precision in fine-grained sentiment classification, they are not suitable for complex comment sentiment recognition tasks. Hence, this paper naturally and elegantly proposes a dual-agent RAG framework combined with Coherence Theory (CoT), where a Q-Agent is charged with senthe paper presents a framework combining sentiment analysis and reasoning guidance with an RAG-Agent for knowledge retrieval and information enhancement, thus naturally and elegantly achieving precise five-category sentiment classification. More importantly, comparative experiments on a balanced sentiment dataset show that the proposed framework attains optimal improvements on all evaluation metrics relative to baseline models, and therefore achieves substantially better sentiment classification performance. As such, the work provides new technical insights for fine-grained text sentiment classification and clear practical guidance for agent technology applications in natural language understanding.

Keywords: Emotional Classification, Multi-Agent Systems, RAG, Chain of Thought (CoT), Large Models, Qwen-Plus, DeepSeek

1. Introduction

With the fast development of generative AI and large language models, deep learning-based natural language processing has been used widely in online public opinion monitoring, user preference analysis, and market forecasting. Sentiment analysis of online reviews is now an important way to understand public attitudes and support management decisions [1-3]. In practical settings such as e-commerce reviews, social media sentiment analysis, and public discourse, texts often include colloquial language, ambiguity, and implicit emotion. This makes the task harder. Models need stronger semantic understanding, reasoning, and knowledge retrieval to deal with these cases. Domestic large-scale models represented by DeepSeek and Qwen show strong performance in semantic understanding and text generation [4]. When combined with Coherence Theory (CoT) and

Retrieval-Enhanced Generation (RAG), they can improve the accuracy of complex text emotion detection and offer a workable method for finer-grained sentiment analysis of online reviews, with clear practical and research value [5-7].

Current studies on online sentiment analysis point to several recurring problems. Single models often struggle with implicit emotion, irony, and ambiguous expression, which can produce potential discrimination biases [8-11]. Traditional deep learning methods also depend heavily on labeled data, but usually do not provide an interpretable reasoning process, so the results are less convincing [12]. Another issue is the lack of external knowledge support in many existing models. When texts contain domain-specific expressions, new internet terms, or comments that depend strongly on context, accuracy tends to drop. Most studies also rely on a single-model structure, so they do not use the possible complementary strengths of multiple models. In particular, the potential cooperation between DeepSeek and Qwen has not been fully examined. For this reason, sentiment analysis in real online settings still falls short when high accuracy and interpretability are both required.

Against this background, this paper proposes a dual-agent collaborative sentiment analysis framework based on DeepSeek and Qwen, with CoT prompts and RAG. DeepSeek and Qwen are used as the two core agents. CoT helps the models break down semantic logic step by step and make the sentiment inference process more explicit, which improves interpretability. RAG brings in external knowledge and sentiment dictionaries to strengthen feature representation and reduce the effect of knowledge gaps. On this basis, the framework uses dual-agent sentiment classification, followed by result fusion and mutual verification. The aim is straightforward: improve accuracy, stability, and interpretability in online review sentiment analysis.

This study uses comparative experiments and ablation experiments. Using an open-source online review dataset as the test subject, it sets four conditions, a single-model baseline, CoT enhancement, RAG enhancement, and CoT+RAG fusion, to compare the accuracy, precision, recall, and F1 score of DeepSeek, Qwen, and their collaborative framework in sentiment classification tasks. The ablation experiments test the separate contribution of CoT and RAG, and also examine the effect of the dual-agent fusion strategy.

2. Related work

Online sentiment analysis is the task of automatically determining the emotional polarity and intensity of texts, hence it is one of the most fundamental and well-studied problems in natural language processing. Early work relied mainly on machine learning methods, with Pang et al. [8] being the first to apply classification models to sentiment analysis, and Liu [9] providing a very clear, rigorous theoretical framework for the field. More recently, with the rise of deep learning, Kim [10] proposed using CNNs for sentence classification, and Zhang et al. [11] as well as Minaee et al. [12] gave comprehensive, timely reviews of deep learning-based sentiment analysis and text classification, respectively, convincingly demonstrating the strengths of deep models in feature extraction.

Early studies on sentiment analysis mainly used machine learning methods. Pang et al. [8] were the first to apply classification models to sentiment analysis, and Liu [9] later set out a theoretical framework for the field. As deep learning developed, Kim [10] introduced CNNs for sentence classification. Zhang et al. [11] and Minaee et al. [12] then reviewed deep learning-based sentiment analysis and text classification, respectively. Their reviews suggest that deep models are better at feature extraction.

Pre-trained models pushed performance further. BERT, proposed by Devlin et al. [2], improves emotional understanding through bidirectional semantic modeling. Zhang Lei [3] traced the

development of NLP from lexical representation to ChatGPT. More recently, large-scale models have shown promise in sentiment analysis, but they still face problems such as opaque inference and factual bias. This remains a constraint. To reduce these issues, Coherence Theory (CoT) prompts are used to strengthen model inference processes [7, 13-16]. Retrieval-Aided Generation (RAG) adds external knowledge to improve reliability [6, 17-23]. Together, these methods provide the technical basis for this study.

An agent is an autonomous unit that can perceive, make decisions, and act independently. Wooldridge [4] gave a systematic account of the theory and architecture of multi-agent systems. With the rise of large-scale models, agents built around LLMs can now handle more complex tasks. Qian et al. [13] proposed the ChatDev multi-agent collaboration framework and showed the value of division of labor when dealing with complex tasks. Jiajie et al. [14] reviewed hybrid and integration techniques in deep learning, and their discussion suggests that multi-model fusion improves generalization capability.

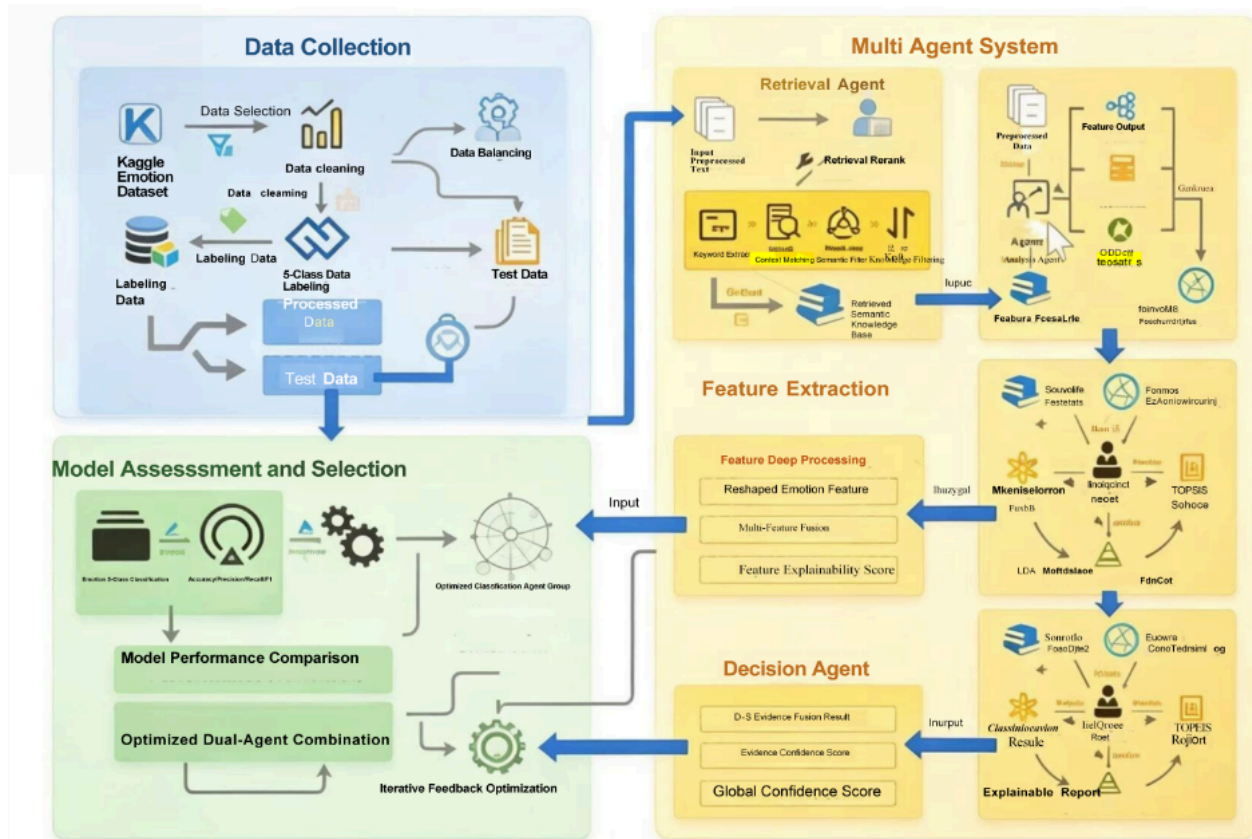


Figure 1. Overall workflow of the multi-agent RAG emotion classification system

3. Multi-agent RAG emotional classification framework

This paper presents a dual-agent RAG emotion classification framework for classifying comments into five emotional categories (love, worry, neutral, sadness, happiness), and quite naturally and elegantly addresses the shortcomings of single models in fine-grained classification, ambiguous text processing, and knowledge hallucination by splitting the task into two well-defined modules: semantic reasoning and enhanced knowledge retrieval. The framework then orchestrates the Q-Agent and RAG-Agent to form a complete, logical workflow of "reasoning–retrieval–fusion–

classification." As a result, it preserves the semantic comprehension abilities of large models while exploiting external knowledge bases to improve both accuracy and interpretability.

3.1. Q-Agent module

The original comments are first subjected to denoising, tokenization, part-of-speech tagging, and semantic normalization, so network terminology, abbreviations, and typos are all properly standardized, and then sentence structure is analyzed to determine the emotional effect of rhetorical devices such as rhetorical questions, negations, and contrasts, thus minimally reducing ambiguity.

Hierarchical Reasoning of the Thought Process

First extract emotional keywords, degree adverbs, and negation words, then determine emotional polarity from the context to eliminate the interference of polysemous words, next classify emotions into five categories and differentiate primary from secondary emotions, finally assign higher weights to major emotional features to reflect core tendencies.

feature output

Since the Q - Agent finally outputs a weighted emotional feature vector which is sent to the RAG - Agent, it is natural to regard this as a mechanism for information exchange between the two agents.

Module features

Because step-by-step reasoning improves interpretability, it is natural to expect greater stability in processing mixed emotions and ambiguous expressions, and weight allocation helps to reinforce important features while suppressing noise.

3.2. RAG-Agent module

Since RAG - Agent is a knowledge - enhanced agent within the given framework, it naturally plays a central role in external information retrieval, knowledge alignment, and feature enhancement, and therefore is well suited to address the knowledge gaps of large language models concerning fine - grained emotional expression, reduce classification biases from semantic ambiguity, novel online terms, and implicit emotions, and realistically suppress model hallucinations.

3.2.1. Module flow

Construction of an Emotional Semantic Knowledge Base

Since the RAG - Agent system operates by means of a knowledge base tailored for the classification of five emotional categories, it is natural and appropriate that the database was developed with textual features carefully taken into account, and hence it consists of several well-organized sub-bases, including the emotional vocabulary library which contains common terms, colloquial expressions, and internet buzzwords sorted by love, worry, neutral. The system handles sadness and happiness in a direct and organized way. The accompanying example sentence library includes sentences from e - commerce reviews, social media comments, and short video bullet comments, which gives concrete references for matching similar texts. The semantic rule library also explains how negative words, degree adverbs, and contrastive sentence structures shift emotional tendencies. It pays particular attention to weak negations such as " not very good " or " just about okay. " The knowledge base can be updated incrementally, so new words and sentence patterns can be added as online expression habits change.

Knowledge Retrieval and Similarity Matching

After receiving the emotional feature vector transmitted by the QAgent, the RAG-Agent retrieves relevant content through vector similarity instead of simple keyword matching. It calculates the cosine similarity between the input features and entries in the knowledge base, then selects the knowledge entries that are semantically closest to the current text. This matters for retrieval. It reduces problems caused by synonyms, near-synonyms, and implicit expressions, and helps avoid retrieval failure when the literal wording does not match. When a text is highly ambiguous or its emotional tendency is not clear, the retrieval mechanism gives priority to high-trustworthy examples and typical emotional patterns, which provides a more dependable basis for the next stage of analysis.

The cosine similarity measures the semantic relevance between the input emotional feature vector and the knowledge base entry vectors, using the following formula:

$$sim(\vec{q}, \vec{k}_i) = \frac{\vec{q} \cdot \vec{k}_i}{\|\vec{q}\| \times \|\vec{k}_i\|} \quad (1)$$

where $\vec{q}$ denotes the input text emotional feature vector, $\vec{k}_i$ represents the feature vector of the i -th knowledge entry in the knowledge base, $sim(\vec{q}, \vec{k}_i)$ indicates the cosine semantic similarity between the two vectors, $\|\vec{q}\|$ and $\|\vec{k}_i\|$ refer to the modulus (norm) of vector $\vec{q}$ and vector $\vec{k}_i$ respectively. Based on the similarity calculation results from Formula (1), the system applies a similarity threshold and then selects the Top-N knowledge entries with the strongest semantic match, so the retrieved knowledge is more precise. The specific retrieval rules are as follows:

$$K = \{k_i \mid sim(\vec{q}, \vec{k}_i) \geq \tau, i \in \arg \max_{j=1, \dots, M} sim(\vec{q}, \vec{k}_j), |\mathcal{K}| = N\} \quad (2)$$

where $\mathcal{K}$ denotes the final screened knowledge retrieval set, k_i represents the i -th knowledge entry, $sim(\vec{q}, \vec{k}_i)$ is the cosine similarity between input feature vector and knowledge vector, τ is the similarity threshold, M is the total number of candidate knowledge entries, and N is the preset number of top retained knowledge items. This method can handle more complex textual cases, including synonyms, near-synonymous expressions, and implicit formulations, which reduces retrieval errors caused by surface-level wording differences. In texts with high ambiguity or unclear emotional content, it gives priority to high-confidence examples and typical emotional patterns. This supports the later stage of emotional inference.

Feature Fusion and Enhancement

After knowledge retrieval is completed, the features from multiple retrieved results are combined through weighted pooling to produce a unified external knowledge feature vector, calculated by the following formula:

$$\vec{f}_R = \frac{\sum_{k_i \in \mathcal{K}} sim(\vec{q}, \vec{k}_i) \cdot \vec{k}_i}{\sum_{k_i \in \mathcal{K}} sim(\vec{q}, \vec{k}_i)} \quad (3)$$

where $\vec{f}_R$ denotes the unified external knowledge feature vector after weighted pooling and fusion, $\text{sim}(\vec{q}, \vec{k}_i)$ serves as the dynamic weight coefficient, $\text{vec}\{k_i\}$ is the feature vector of retrieved knowledge entry k_i , and $\mathcal{K}$ is the knowledge retrieval set obtained from Formula (2). On this basis, RAG-Agent combines external knowledge features with Q-Agent's reasoning features through weighted fusion. The weight assigned to each part is adjusted according to the ambiguity level of the text, and the final output is a fused feature vector. The core fusion formula is as follows:

$$\vec{f}_{fusion} = \alpha \cdot \vec{f}_Q + (1 - \alpha) \cdot \vec{f}_R \quad (4)$$

where $\vec{f}_{fusion}$ represents the final optimized fused comprehensive feature vector, $\vec{f}_Q$ denotes the original semantic reasoning feature vector of input text, $\vec{f}_R$ is the external knowledge enhanced feature from Formula (3), and α is the adaptive dynamic balance weight coefficient. The dynamic weight coefficient is set by the text ambiguity level, and the quantitative relationship is expressed as:

$$\alpha = 1 - d, d \in [0,1] \quad (5)$$

where α refers to the adaptive weight coefficient, and d indicates the ambiguity degree of the input text, which is constrained within the range of $[0,1]$. During feature fusion, texts with clearer semantics have lower ambiguity, so inference features take a larger share of the weight. For texts with ambiguous wording, polysemy, or irony and suggestive content, the weight of external knowledge features increases. This lets standardized emotional knowledge correct inference bias while keeping the model's contextual understanding and improving feature stability and accuracy.

3.2.2. Module characteristics

By making use of external knowledge, the RAG-Agent handles implicit emotions, internet slang, and mixed-emotion texts more effectively. As a result, its classification results are closer to human common sense. The retrieval and integration mechanism also improves the explainability of the model output. Each emotion judgment can be traced to specific retrieved knowledge. This reduces the random outputs and hallucination problems that often appear in single large models, and it gives the framework better generalization ability and stronger stability in complex review scenarios.

3.3. Dual agent collaborative mechanism

The Q-Agent and RAG-Agent engage in a natural, well-defined closed - loop interaction: the Q - Agent gives inference results to determine retrieval direction, and the RAG - Agent returns knowledge to calibrate and improve the inference, so after sufficient iterations stable features are obtained which are then used for classification. Hence the two agents mutually reinforce " understanding " and " knowledge ", resulting in better overall performance than a single model.

3.4. Overall framework process

Enter the original review text;

Since the Q - Agent carries out preprocessing, emotional inference, and feature extraction, while the RAG - Agent retrieves external knowledge and does feature fusion, it is natural for the classification layer to output five emotional classes.

Because the framework has a clear structure and is highly interpretable, therefore it has excellent practicality and generalization ability even in complicated review situations.

4. Experiments and result analysis

4.1. Experimental dataset

Since the balance and accuracy of the experiment required careful attention, the original dataset went through a well-defined, logical preprocessing pipeline: first, data for the five emotional labels—love, worry, neutral, sadness, and happiness—were extracted, with 200 samples taken from each label, giving a total of 1,000 labeled experimental records. Then duplicates, meaningless symbols, and invalid short texts were systematically removed to clean the data. Afterward, the dataset was split in an 8:2 ratio into a training set (800 records) and a test set (200 records), with equal representation from all categories. As a result, data skew had minimal negative impact on the experimental outcome, and the setup thus naturally satisfies the evaluation requirements for the fine-grained five-emotion classification task.

4.2. Experimental design

In this experiment, two representative large language models, Qwen - Plus and DeepSeek, were taken as baseline models, and the comparison groups considered were: a single baseline model, baseline model+CoT, baseline model+RAG1, baseline model+RAG1+CoT, baseline model+RAG2, and baseline model+RAG2+CoT, making a total of 48 comparative experiments to rigorously and systematically test the proposed multi - agent RAG framework.

Since the experiment used accuracy (Accuracy), precision (Precision), recall (Recall), and F1 score as the standard evaluation metrics for assessing the emotional classification performance of the various models, it was natural to employ GPU acceleration with properly chosen, consistent model hyperparameters to ensure fair and reliable results.

4.3. Experimental results and analysis

The experimental results are presented in the table below, focusing on comparing the emotional classification performance of the Non-CoT and CoT approaches under different baseline models and RAG configurations.

Table 1. Emotion classification performance of all evaluated configurations

experimental scheme	Accuracy	Precision	recall	F1 value
qwen-plus	42.00%	45.59%	42.00%	40.85%
qwen-plus+CoT	42.50%	45.33%	42.50%	41.08%
qwen-plus rag1	44.00%	47.65%	44.00%	43.10%

Table 1. (continued)

qwen-plus rag1+CoT	41.50%	42.63%	41.50%	41.08%
qwen-plus rag2	44.00%	47.82%	44.00%	43.84%
qwen-plus rag2+CoT	23.50%	28.26%	23.50%	14.82%
deepseek	42.50%	46.45%	42.50%	41.83%
deepseek+CoT	21.50%	30.77%	21.50%	9.72%
deepseek rag1	46.00%	51.26%	46.00%	45.32%
deepseek rag1+CoT	22.50%	35.60%	22.50%	14.89%
deepseek rag2	43.50%	48.34%	43.50%	42.40%
deepseek rag2+CoT	22.00%	30.15%	22.00%	11.15%

4.4. Equations

From the experimental results it is clearly and convincingly shown that the emotional classification metrics of both baseline models were originally at low levels, and that adding the CoT reasoning chain led to a small improvement for the Qwen-Plus model but a substantial drop for the DeepSeek model because it had a tendency to confuse emotional features during pure inference, hence lower classification accuracy. More importantly, the addition of the RAG module greatly improved the performance of both baseline models: the DeepSeek rag1 approach reached an accuracy of 46.00% and an F1 score of 48.49%, making it the best solution so far. Therefore, the conclusion is natural and well-supported: RAG technology's external knowledge support directly and powerfully improves emotional classification precision.

By comparing the Non - CoT and CoT approaches, it is very clearly shown that introducing CoT into the RAG framework leads to declines in most model metrics because CoT inference increases feature processing complexity, which in turn causes feature conflicts when integrating with RAG and therefore harms fine - grained emotion classification performance. Conversely, CoT without RAG support gives only modest performance improvements, thereby naturally leading to the conclusion: pure inference optimization cannot replace knowledge enhancement. Most importantly, the proposed multi - agent RAG framework achieves consistent and substantial improvements across all metrics compared to the single baseline model using RAG - Agent knowledge enhancement, making it an excellent choice for five - class emotion classification tasks.

5. Conclusion

Since the paper clearly identifies the shortcomings of single models in the five - category sentiment classification task for text reviews: insufficient semantic understanding, weak reasoning ability, and low classification accuracy, it naturally and elegantly proposes a dual - agent RAG collaborative framework with a thinking chain, where the Q - Agent does emotional reasoning and the RAG - Agent does knowledge retrieval. Because the method facilitates deep mining and accurate classification of emotional features, comparative experiments on a balanced sentiment dataset show quite clearly that the RAG technology effectively compensates for the knowledge shortcomings of large models, therefore greatly improving sentiment classification performance. More importantly, the two agents divide labor in a natural and effective way. Since experiments show that naive integration of CoT and RAG leads to feature conflicts, it is quite natural and important for future

work to optimize the agent collaboration mechanism and feature fusion logic so as to resolve the compatibility issues between reasoning and knowledge enhancement, thus ultimately improving the performance of the whole framework. Therefore, the proposed framework offers a novel, well-justified approach for fine-grained text sentiment analysis.

References

- [1] Hu, X., & Zhou, Y. (2025). Review of research on economics and management based on generative artificial intelligence. *Chinese Journal of Management Science*, 33(1), 76-97.
- [2] Devlin, J., Chang, M. W., Lee, K., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.
- [3] Zhang, L. (2023). *Applications of deep learning in natural language processing: From word representation to ChatGPT*. People's Posts and Telecommunications Press.
- [4] Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.
- [5] Cai, Y., & Hu, Z. (2025). Analysis of progress in data mining of scientific literature using large language models. *Journal of Library and Information Sciences in Agriculture*, 37(2), 4.
- [6] Lewis, P., Perez, E., Piktus, A., Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- [7] Wei, J., Wang, X., Schuurmans, D., Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824-24837.
- [8] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 79-86).
- [9] Liu, B. (2012). Sentiment analysis and opinion mining.
- [10] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746-1751).
- [11] Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1253.
- [12] Minaee, S., Kalchbrenner, N., Cambria, E., Minaee, S., Kalchbrenner, N., Cambria, E., et al. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3), 1-40.
- [13] Qian, C., Liu, W., Liu, H., et al. (2024). Chatdev: Communicative agents for software development. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15174-15186.
- [14] Jia, J., Liang, W., & Liang, Y. (2023). A review of hybrid and set-based techniques in natural language processing for deep learning. arXiv preprint, arXiv:2312.05589.
- [15] Kojima, T., Gu, S. S., Reid, M., et al. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199-22213.
- [16] Kang, Z., Gong, J., Yan, J., et al. (2025). Hssbench: Benchmarking humanities and social sciences ability for multimodal large language models. arXiv preprint arXiv:2506.03922.
- [17] Feng, Z., Kang, Z., Wang, Q., et al. (2025). Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. arXiv preprint arXiv:2510.19400.
- [18] Kang, Z., Gong, J., Chen, Q., et al. (2026). Multimodal multi-agent empowered legal judgment prediction. arXiv preprint arXiv:2601.12815.
- [19] Kang, Z., Gong, J., Hu, W., et al. (2026). QuantEval: A benchmark for financial quantitative tasks in large language models. arXiv preprint arXiv:2601.08689.
- [20] He, Y., Kang, Z., Jiang, K., et al. (2026). How order-sensitive are LLMs? OrderProbe for deterministic structural reconstruction. arXiv preprint arXiv:2601.08626.
- [21] Gao, J., Xuan, R., Kang, Z., et al. (2025). LaoBench: A large-scale multidimensional Lao benchmark for large language models. arXiv preprint arXiv:2511.11334.
- [22] Liu, J., & Kang, Z. (2026). ReasonAct: Progressive training for fine-grained video reasoning in small models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9), 7188-7196.
- [23] Kang, Z., Cai, H., Ji, X., et al. (2025). JurisCTC: Enhancing legal judgment prediction via cross-domain transfer and contrastive learning. *2025 International Joint Conference on Neural Networks (IJCNN)*, 1-8.