Proceedings of the 4th International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2026.33551

StressAgent: A Dual-Agent System for Human Stress Classification with Retrieval Augmentation

Ziyue Zhu

School of Mathematical Sciences, Anhui University, Hefei, China
a02314161@stu.ahu.edu.cn

Abstract. Identifying stress states from user text is important for mental health monitoring, but it remains difficult because stress expression is subjective and varied, and language cues across states overlap. In closed-set classification, the model must select a label from a fixed set of labels, as vague or weak evidence often leads to errors. While instruction follows LLMs to accomplish this task, the output can be word-sensitive and use unconstrained reasoning when the evidence is insufficient. We propose StressAgent, an evidence-enhanced dual-agent framework, to improve robustness. StressAgent breaks down closed-set predictions into two steps: a Reasoning and Decision (RD) agent and a Retrieval-Augmented Generation (RAG) agent. The RAG agent retrieves the 10 most similar training instances from the search index and returns their category labels and short text fragments associated with the tags as evidence. The RD agent then integrates the retrieved evidence into a restricted closed-set output format to generate a final label. Experiments were conducted on Qwen2.5-7B-Instruct and Deepseek-Chat to evaluate all combinations of enabled or disabled retrieval and enabled or disabled chains of thought. The results show that retrieval enhancement provides continuous accuracy improvements and explains most improvements. Search using cosonic similarity outperformed Euclidean distance, highlighting the impact of similarity measures on the usefulness of evidence. In the absence of chain of thought, cosine similarity retrieval improved the accuracy of Qwen2.5-7B-Instruct from 61.97% to 71.37%, and Deepseek-Chat from 67.52% to 74.89%. Once retrieval is enabled, the further benefits of chains of thought are limited. Overall, StressAgent provides an interpretable, controllable, evidence-based driven classification paradigm for closed set stresses.

Keywords: Closed-Set Classification, StressAgent, Retrieval-Augmented Generation, Cosine Similarity, Chain-of-Thought

1. Introduction

Stress detection from user posts is challenging due to diverse expressions: the same underlying stress state may be described differently, while different states can still share overlapping linguistic cues.

In a closed-set setting, the model must accurately select a label from the predefined set of categories, so that label ambiguity becomes the main source of errors. Although the instruction-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33551

following LLMs are effective for text classification, the prompt model may only predict from input, which may cause it to rely on incomplete clues and make the final decision sensitive to prompt wording.

To improve robustness, we proposed StressAgent, which is a dual-agent framework that decomposes closed set classification into two sequential stages. First, the Retrieval-Augmented Generation (RAG) agent retrieves the evidence of label alignment from the database of the annotation training instance. Secondly, the Reasoning and Decision (RD) agent generates the final label under the restricted closed set output format. Specifically, the RAG agent retrieves the most similar training examples of top-k and provides its ground truth category labels and short text fragments as supporting references. This guides the RD agent to reason with consistent evidence and reduce its dependence on unconstrained internal associations.

StressAgent has a configurable retrieval similarity function. We checked two key indicators: Euclidean distance and cosine similarity. Lightweight prompt style control is also integrated to reduce the confusing effect of prompt wording changes. All evaluation indicators are obtained by analyzing the final prediction label from the constraint output of the model.

We further explore the chain-of-thought (CoT) prompt as an auxiliary reasoning variant. However, once the retrieval enhancement is enabled, the impact is insignificant and sometimes slightly disadvantageous. This shows that providing retrieval-based evidence consistent with the label is the main driving force for the observed improvement.

The experiment of the six-way stress classification benchmark shows that the injection of evidence based on retrieval and label consistency has greatly improved closed stress identification. In addition, the choice of retrieval configuration will significantly affect the utility of the retrieved reference.

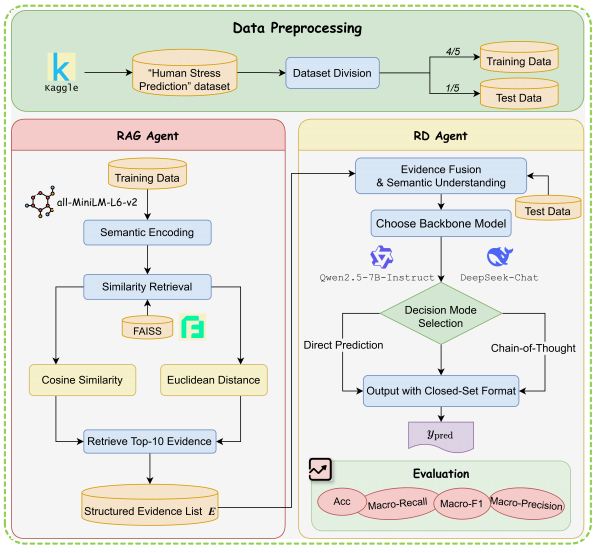


Figure 1. Overview of our dual-agent framework

The main contributions of this study are as follows:

1. We propose and implement StressAgent, an LLM-based dual-agent framework that decomposes closed-set classification into two sequential stages: RAG agent for evidence retrieval and RD agent for constraint label prediction.

DOI: 10.54254/2755-2721/2026.33551

2. We design a label-aligned evidence injection mechanism in which the RAG agent retrieves the top ten similar training instances and uses its category labels and text fragments as supporting references for decision agents, using a closed-set constrained output format.

3. We conduct extensive comparative and ablation studies to quantify the unique benefits of inverse finding evidence and to analyse the impact of the anti-seeking similarity function—especially comparing cosine similarity to Euclidean distance under fixed evidence and prediction schemes.

4. We introduce prompt style control and label-driven evaluation, including reasoning-oriented prompt variations, and final label calculation metrics extracted directly from constrained output to reduce confounding factors from free intermediate generation..

2. Related work

2.1. Evolution of methods for stress state detection and emotion analysis

Early stress detection methods have gone through three distinct stages of development. In the initial stage, researchers mainly relied on approaches such as feature engineering and support vector machines, using emotional vocabulary together with N-gram features [1]. However, this line of work was often constrained by the dependence on domain expertise, which made it difficult to capture complex and subjective stress states.

With the rise of deep learning, the field has undergone significant changes. Convolutional neural networks have enabled automatic feature extraction from text [2], while recurrent neural networks have shown stronger capability in modeling sequential context [3]. More recently, Transformer architectures and attention mechanisms have brought major improvements in feature fusion [4], leading to pre-trained models such as BERT [5] and setting new performance baselines. Even so, these models are still frequently viewed as "black boxes," and interpretability remains limited [6]. In addition, their reliability can drop when the input deviates from the training data distribution, or when labels are ambiguous [7,8].

In the current stage, large language models allow more effective classification for minority samples through natural language instructions [9]. Still, applying them to closed-set classification remains challenging, mainly because the predictions are highly sensitive to suggestive expressions [10], and the models may also produce "hallucinations" that lead to factual inconsistencies [11]. To improve robustness, retrieval augmentation methods that provide external evidence have therefore become important [12]. Based on this motivation, our StressAgent framework aims to build an evidence-based and controllable classification paradigm for LLM stress states.

2.2. Agentic multi-agent for reliable decisions

Research on multi-agent collaboration based on large language models started from early attempts to explore how agents could interact with each other. The proposed debate mechanism also supports the idea that multi-perspective collaboration can improve the reliability of the resulting solutions.

With technological advancements, the field has entered a stage of system framework development. Represented by ChatDev [13], role-based collaboration, the "reasoning-action" paradigm imitating ReAct [14], and the open dialogue advocated by CAMEL [15] have jointly promoted the formation of common interaction and cooperation models in multi-agent systems.

After that, the focus turned to enhancing the metacognitive ability of the intelligent body. Self-reflection and iterative optimization mechanisms from frameworks such as Reflexion [16] and Self-

DOI: 10.54254/2755-2721/2026.33551

Refine [17] are introduced, which significantly improves the robustness of intelligent body execution.

At present, research is developing in the direction of diversified applications, especially in closed and high-precision decision-making tasks, facing new challenges in explainability and evidence sensitivity. For this reason, this article puts forward the StressAgent framework. Through a controlled dual-agent architecture and label alignment evidence injection mechanism, it provides a reliable and traceable collaborative solution for fine-grained classification tasks. In StressAgent, the RAG agent retrieves the relevant evidence, and the RD agent uses these evidences and fragments to judge the final label. By reasoning based on traceable facts, it reduces dependence on internal a priori and suggestive wording. Overall, it has formed an interpretable, robust and controllable text classification system.

3. The StressAgent framework

3.1. Retrieval-Augmented Generation (RAG) agent

Retrieval-Augmented Generation (RAG) is the core of knowledge retrieval and evidence provision in the StressAgent framework. Its main responsibility is, for each user input to be classified, to efficiently and in real time retrieve the semantically most similar instances from the labeled training data. It then passes the text snippets of these instances along with their true labels as label-alignment evidence to the downstream inference and decision-making agent. This design aims to anchor the open-domain reasoning capability of large language models (LLMs) in factual data that is specific, relevant, and accompanied by explicit supervised signals, thereby fundamentally improving the robustness, traceability, and insensitivity of the final decisions to the wording of prompts.

3.1.1. Semantic encoding and vectorization

The first step of the RAG agent is to transform textual data into computable dense semantic vectors. Given a training set $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is the text and $y_i \in Y$ is the corresponding pressure-state label (where Y is a predefined closed label set), we adopt a pretrained Sentence Transformer encoder f_{enc} to map each text segment into a d-dimensional semantic vector representation:

$$v_i = f_{\text{enc}}(x_i) \in \mathbb{R}^d \tag{1}$$

In this work, f_{enc} is instantiated as the all-MiniLM-L6-v2 model. This model is optimized through contrastive learning to ensure that semantically similar sentences are close to each other in the vector space. To support different similarity-metric schemes, the encoding process includes a key configuration option: for cosine similarity retrieval, the generated vectors are L2-normalized, defined as $\hat{v}_i = \frac{v_i}{\|v_i\|_2}$, which projects the vectors onto the unit hypersphere, ensuring that similarity is determined only by direction; for Euclidean distance retrieval, the vectors are kept in their original, non-normalized form v_i, so as to preserve their absolute positional information in the semantic space, facilitating meaningful distance computation.

The collection of all training vectors forms a searchable vector database, denoted by the matrix V_{db}.

DOI: 10.54254/2755-2721/2026.33551

3.1.2. Similarity metrics and retrieval mechanism

For a query text q to be classified, the RAG agent first encodes it into v_q or $\hat{v}_q$. The core of retrieval is to compute the similarity between the query vector and all vectors in the database, in order to find K most relevant samples (in this study, $K = 10$). The framework regards the similarity index as a configurable core module; the system compares the two standard indicators realized through different FAISS index types:

• Cosine similarity retrieval: This method focuses on the consistency of semantic directions between vectors. For L2-normalized vectors, cosine similarity is equivalent to the inner product.

$$s_{\cos}\left(\hat{v}_q, \hat{v}_i\right) = \hat{v}_q \cdot \hat{v}_i = \frac{v_q \cdot v_i}{\|v_q\|_2 \|v_i\|_2} \tag{2}$$

Its range is $[-1, 1]$; the larger the value, the higher the semantic similarity. In implementation, this mode uses FAISS's IndexFlatIP (inner product) index and works with normalized vectors.

• Euclidean distance retrieval: This method measures the absolute geometric distance of vectors in a high-dimensional space. FAISS's IndexFlatL2 index computes the squared Euclidean distance.

$$d_{L2}^2\left(v_q, v_i\right) = \| v_q - v_i \|_2^2 \tag{3}$$

Its range is $[0, +\infty)$; the smaller the value, the more similar the vectors. In implementation, this mode uses the IndexFlatL2 index together with the original, non-normalized vectors. This setup serves as a key comparative experimental configuration to analyze whether direction-based cosine retrieval outperforms absolute distance-based Euclidean retrieval in subjective text classification tasks.

Based on the selected similarity metric s or distance d^2, the retrieval process is formalized as finding Top- K nearest neighbors:

$$I_{\text{top}K} = \underset{i \in [1,N]}{\text{argtopK}}\, s\left(v_q, v_i\right) \tag{4}$$

where $I_{\text{top}K}$ is the index set of the retrieved top- K training samples.

3.1.3. Evidence formatting and output

For each retrieved sample index $j \in I_{\text{top}K}$, the RAG agent extracts its true label y_j and the original text x_j from the training data. To manage the context length of the downstream model, the text x_j is truncated to a predefined maximum length of L characters (set to 200 in this study). Finally, the RAG agent outputs a structured evidence list E as its final artifact:

$$E = \left[\left(y_j, \text{truncate}\left(x_j, L\right), s_j\right)\right]_{j \in I_{\text{top}K}} \tag{5}$$

DOI: 10.54254/2755-2721/2026.33551

where s_j is the corresponding similarity or distance score. This evidence list E, together with the original query text q, forms the complete input for the inference and decision-making agent. Through this label-matching evidence injection mechanism, the RAG agent provides traceable and valuable data support for subsequent discriminant reasoning. The in-depth analysis of retrieval similarity indicators in this framework shows that the quality of retrieval evidence is a key factor affecting the performance of the entire system, which is also the core contribution of this study.

3.2. Reasoning and Decision (RD) agent

The Reasoning and Decision (RD) agent is the ultimate decision-making module of the StressAgent framework. It receives a tuple input from the RAG agent: the original user query text q and the structured evidence list E. Its core responsibility is to integrate the information of these two parts for discriminant reasoning, and output a single, definitive prediction label y_{pred}, which must belong to the predefined closed label set Y.

This study selects two lightweight and open-source instruction-tuned large-scale language models, Qwen2.5-7B-Instruct and Deepseek-Chat, as the backbone of the RD agent. These two models achieve an excellent balance between instruction tracking ability, context depth understanding and computational efficiency. Its open-source features support complete local deployment. By integrating the external evidence of label alignment provided by the RAG agent, the reasoning process of the RD agent is actually limited to relevant facts. This helps to reduce the "illusion" of the model and improve the practicality and maintainability of the framework by updating and retrieving the database instead of retraining the model.

In order to explore the effectiveness of different decision-making models supported by evidence and adapt to diversified application scenarios, the RD agent designed and implemented two switchable prompt strategies:

• Direct Prediction Prompting: This strategy instructs the model to directly output the final label after comprehensive analysis based on query q and evidence list E, without presenting any intermediate inference steps. This strategy pursues the efficiency of final decision-making and is suitable for lightweight application scenarios that require high-speed response or clear evidence quality and high confidence.

• Chain-of-Thought Prompting: The strategy guides the model to conduct internal step-by-step reasoning, such as evaluating the correlation between each evidence and the query, and comparing the support and weight of the evidence in different categories before giving the final answer. By clarifying the "thinking process" of the model, the strategy aims to improve the reliability of decision-making and improve the interpretability of the results when dealing with complex or vague cases. It is suitable for in-depth evaluation or diagnosis support scenarios that require a more rigorous analysis process.

No matter what strategy is adopted, the RD agent must strictly abide by the closed set output format. This key design strictly restricts the originally open output space used to generate large language models, limiting them from using limited predefined categories, so as to fundamentally ensure the certainty and controllability of decision-making. In the process of implementation, the prediction results can be automatically and steadily extracted through subsequent regular expression analysis, which greatly promotes the automatic evaluation of system performance.

DOI: 10.54254/2755-2721/2026.33551

4. Experiment

4.1. Dataset

This study evaluates the proposed StressAgent framework on the publicly available "Human Stress Prediction" dataset. The dataset is collected from online mental health support communities and contains texts where users express various psychological stressors and life difficulties in an unstructured, narrative manner. It provides a realistic scenario for researching the classification robustness of subjective texts.

The dataset contains a total of 1,168 annotated samples, covering 6 major stress state categories, which constitute the closed label set Y of this study: anxiety, assistance, domestic violence, homelessness, post-traumatic stress disorder (PTSD), and relationship problems. Its raw distribution shows significant natural imbalances. This reflects the real difference in the frequency of discussion of different stress topics within the community, and also poses additional challenges for the model to achieve unbiased classification in imbalanced data.

In order to carry out strict evaluation and prevent data leakage, we use random layered sampling to divide the original data set into mutually exclusive training sets and test sets in a ratio of 4:1. This division ensures the complete sample separation between the knowledge base (training set) retrieved by the RAG agent and the test set used for evaluation, and fundamentally eliminates the risk of evaluation deviation caused by the leakage of test data during the training process. While maintaining the original distribution ratio of each category, this split also ensures the statistical significance of the test set. The detailed division of the dataset is shown in Table 1.

Table 1. Dataset division and sample distribution

Category	Training Samples	Test Samples	Category Total
anxiety	152	48	200
assistance	168	32	200
domestic violence	170	30	200
homeless	130	38	168
ptsd	155	45	200
relationships	159	41	200
Total	934	234	200

The text in the data set is highly subjective, diverse in form and ambiguous in semantics. For example, similar stress states may be described by completely different personal experiences, while different states may share similar emotional words. This inherent semantic ambiguity, coupled with the above category imbalance, is the core challenge of closed stress classification: label ambiguity and weak discrimination clues. Therefore, the data set provides an ideal and rigorous test platform to verify whether the StressAgent framework can enhance the discrimination and decision-making robustness of the model on fuzzy and unbalanced data by retrieving and injecting high-quality external evidence.

4.2. Experimental variable and evaluation metrics

To systematically evaluate the effectiveness of the StressAgent framework and the impact of its components on the final performance, controlled and ablation experiments were designed. The

DOI: 10.54254/2755-2721/2026.33551

experimental design focused on the core variable of retrieval enhancement, and investigated the impact of retrieval similarity index, inference prompting strategy, and basic model configuration.

The core mechanism variables test the fundamental impact of introducing external evidence by comparing two settings: in the baseline setting, the RAG agent is disabled, and the RD agent relies only on its internal knowledge for zero-shot classification; the other setting supports a complete two-agent collaboration framework. Within the main framework, we further compare two retrieval similarity metrics: cosine similarity retrieval, inner product calculation based on normalized vectors, and Euclidean distance retrieval based on the square Euclidean distance of the original vector, to analyze how different similarity definitions affect the quality of evidence and subsequent task performance. The prompt strategy variable explores how the RD agent's internal reasoning mode affects its decision-making efficiency after obtaining evidence, and compares the efficient direct prediction strategy with the chain of thought strategy that enhances interpretability through explicit stepwise reasoning. To test the generalization of the framework, two lightweight open-source models, Qwen2.5-7B-Instruct and Deepseek-Chat, are used as the basis backbone models for validation.

We employed Accuracy (Acc) , Macro-averaged Recall (MR), Macro-averaged F1-score (F1) and Macro-averaged Precision (MP) as evaluation metrics for closed-set text classification. All metrics are calculated based on the final predicted labels and true labels parsed in the RD agent's strictly constrained output format. During the evaluation, we filter out invalid predictions and process unseen categories by zero division to ensure that the classification performance quantitative results obtained on complex, unbalanced stress text data are robust and reproducible.

4.3. Experimental result and analysis

Table 2 and Table 3 present the experimental results when using the direct prediction strategy and the Chain-of-Thought strategy, respectively.

Table 2. Experimental results using the direct prediction

Backbone Model	Experimental Group	Similarity Metric	Acc	MR	F1	MP
Qwen2.5	Baseline	\	0.6197	0.6156	0.6223	0.6970
	StressAgent	Cosine Similarity	0.7137	0.6998	0.7007	0.7327
	StressAgent	Euclidean Distance	0.6787	0.6792	0.6737	0.7054
DeepSeek	Baseline	\	0.6752	0.6648	0.6686	0.7122
	StressAgent	Cosine Similarity	0.7489	0.7496	0.7450	0.7643
	StressAgent	Euclidean Distance	0.6496	0.6497	0.6414	0.7444

Table 3. Experimental results using the chain-of-thought

Backbone Model	Experimental Group	Similarity Metric	Acc	MR	F1	MP
Qwen2.5	Baseline	\	0.6239	0.6185	0.6210	0.6948
	StressAgent	Cosine Similarity	0.6838	0.6705	0.6691	0.6976
	StressAgent	Euclidean Distance	0.6795	0.6747	0.6717	0.6952
DeepSeek	Baseline	\	0.7179	0.7082	0.7120	0.7487
	StressAgent	Cosine Similarity	0.7436	0.7448	0.7400	0.7597
	StressAgent	Euclidean Distance	0.6453	0.6432	0.6366	0.7343

DOI: 10.54254/2755-2721/2026.33551

In Table 2 and 3, Qwen stands for Qwen2.5-7B-Instruct, and DeepSeek refers to Deepseek-Chat. Experimental results show that the retrieval enhancement mechanism of the StressAgent framework can significantly and stably improve the performance of stress state recognition in closed sets. Compared with a single model baseline without RAG agents, enabling RAG agents enables the model to obtain external evidence that is more semantically relevant to the current input with clear category labels. This makes it easier for the RD module to form evidence citations when selecting categories in closed label collections, thereby reducing the risk of label ambiguity and misjudgment caused by relying only on surface lexical cues.

4.3.1. RAG agent brings stable performance improvement

In the configuration without CoT enabled, RAG retrieval enhancement provides an overall performance improvement compared to the baseline without retrieval. Taking Qwen2.5-7B-Ininstruction as an example, the accuracy under the cosine retrieval index increases from 61.97% to 71.37%, and the macroscopic average F1 score increases from 62.23% to 70.07%. This indicates that the retrieval enhancement provides a more stable clue of similarity to the model, thereby improving the overall quality of closed set discrimination. In addition, in the Deepseek-Chat configuration with CoT enabled and using the cosine retrieval metric, the accuracy improves from 71.79% to 74.36%. Therefore, it can be concluded that RAG retrieval enhancement still achieves a net performance improvement under the condition of requiring organized reasoning. However, in the Deepseek-Chat configuration with both CoT and Euclid retrieval metrics enabled, the accuracy drops to 64.53%. This phenomenon suggests that the intensity of RAG gain is affected by retrieval metrics and evidence matching, but this does not change the overall conclusion that RAG improves performance in representative configurations compared to no search baseline.

4.3.2. Cosine retrieval evidence is more usable

Comparing the cosine and Euclidean inversion methods shows that the evidence constraints provided by cosine inversion are more effective in closed classification. Therefore, the model finds that it is easier to gain an advantage in key discriminant dimensions. In the absence of CoT, the accuracy of the cosine gauge of Qwen2.5-7B-Ininstruction reaches 71.37%, which is higher than that of the Euclid gauge of 67.87%. Deepseek-Chat achieves 74.89% under the cosine gauge, and the use of the Euclidean gauge even reduces the accuracy to 64.96%, showing a larger performance gap. This model is still valid when using CoT.

4.3.3. Gains from cot become limited or even unstable after enabling RAG

Comparing scenarios with and without deep thinking, it is clear that CoT does not bring a monotonous improvement. For Qwen2.5-7B-Ininstruction, the accuracy increased from 61.97% to 62.39% in the non-RAG scenario. However, after enabling cosine repetition, the accuracy rate decreased from 71.37% to 68.38%. Under the Euclide index, the performance remained basically unchanged, and the accuracy rate increased from 67.87% to 67.95%. For Deepseek-Chat, CoT provides a more significant improvement without RAG, with an accuracy rate increased from 67.52% to 71.79%. After enabling the retrieval enhancement, the accuracy of the cosine only decreased slightly, from 74.89% to 74.36%. Euclide's configuration is still relatively low, with an accuracy rate reduced from 64.96% to 64.53%. Therefore, when the external evidence is relatively

DOI: 10.54254/2755-2721/2026.33551

clear, additional thought chain reasoning is more likely to trigger fluctuations in the redistribution of evidence weights and interpretation paths, limiting benefits and even leading to declines.

4.3.4. Deepseek-Chat shows more pronounced advantages with cosine evidence

From the perspective of overall performance, Deepseek-Chat gets a more stable return under the optimal retrieval configuration. In the absence of deep thinking, the total accuracy rate of Deepseek-Chat and cosine search indicators reached 74.89%. After enabling CoT, the combination is still in the highest range, with an accuracy rate of 74.36%. This shows that Deepseek-Chat can more effectively convert the category-related evidence retrieved into the discrimination advantage of closed collections. In contrast, the accuracy rate of Qwen2.5-7B-Ininstruction under the cosine index is 71.37% and 68.38% respectively. Its increase is slightly lower than that of Deepseek-Chat, reflecting that the absorption and utilization efficiency of Qwen under the same evidence input is relatively limited, so it fails to reach the same upper limit under the optimal retrieval configuration.

4.4. Chapter summary and experimental implications

The experiment in this chapter verifies the effectiveness of the StressAgent framework in classifying closed set rereading text. One core finding is that the retrieval enhancement generation mechanism is the fundamental driving force for performance improvement. By introducing external evidence of tag alignment, the mechanism effectively compensates for common semantic ambiguities in short texts, thus significantly and continuously improving the accuracy and robustness of classification.

Further comparative analysis reveals the key insights of system configuration. First of all, the selection of retrieval similarity indicators is very important. The quality consistency and semantic alignment of evidence using cosine similarity retrieval are significantly better than the method of using Euclidean distance retrieval to strengthen closed classification constraints. Secondly, the effectiveness of the thinking chain prompt has significant conditions. In the absence of external evidence, the strategy has certain advantages. However, once the system can provide high-quality retrieval evidence, its marginal benefits will be significantly reduced, which may lead to decision-making fluctuations. Therefore, this strategy is more suitable as an auxiliary tool for insufficient or uncertain situations.

In a word, the experimental results provide clear guidance for practical operation. In order to build a robust classification system, the cosine similarity inversion enhancement component should be given priority as the main basis for decision-making. The thinking chain strategy should be used cautiously, mainly for situations where the retrieval evidence is not enough to support high-reliability decision-making.

5. Conclusion

To address evidence sparsity and semantic ambiguity in closed-set stress state classification, this study proposes StressAgent, a dual-agent framework combining a 7B-scale open-source LLM with a RAG agent. RAG agent retrieves label-aligned evidence from a case database, providing explicit constraints for classification. Experiments confirm that RAG delivers stable gains, with cosine similarity outperforming Euclidean distance by retrieving higher-quality, semantically aligned evidence. The benefit of Chain-of-Thought (CoT) prompting is conditional. It offers diminishing returns when high-quality evidence is already provided. Therefore, the primary recommendation is to prioritize cosine-based retrieval augmentation and use CoT only when evidence is insufficient.

DOI: 10.54254/2755-2721/2026.33551

References

[1] Pang B, Lee L. Opinion mining and sentiment analysis [M]. Now Publishers Inc, 2008.
[2] Xu J, Chen D, Qiu X, et al. Cached long short-term memory neural networks for document-level sentiment classification [C]//Proceedings of the 2016 conference on empirical methods in natural language processing. 2016: 1660-1669.
[3] Kim Y. Convolutional neural networks for sentence classification [C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1746-1751.
[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.
[5] Lee J, Toutanova K. Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv: 1810.04805, 2018, 3(8): 4171-4186.
[6] Zhang Y, Tiňo P, Leonardis A, et al. A survey on neural network interpretability [J]. IEEE transactions on emerging topics in computational intelligence, 2021, 5(5): 726-742.
[7] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks [J]. arXiv preprint arXiv: 1610.02136, 2016.
[8] Gao B B, Xing C, Xie C W, et al. Deep label distribution learning with label ambiguity [J]. IEEE Transactions on Image Processing, 2017, 26(6): 2825-2838.
[9] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [J]. Advances in neural information processing systems, 2022, 35: 27730-27744.
[10] Zhuo J, Zhang S, Fang X, et al. ProSA: Assessing and understanding the prompt sensitivity of LLMs [C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 1950-1976.
[11] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation [J]. ACM computing surveys, 2023, 55(12): 1-38.
[12] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [J]. Advances in neural information processing systems, 2020, 33: 9459-9474.
[13] Qian C, Liu W, Liu H, et al. Chatdev: Communicative agents for software development [C]//Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). 2024: 15174-15186.
[14] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models [C]//The eleventh international conference on learning representations. 2022.
[15] Li G, Hammoud H, Itani H, et al. Camel: Communicative agents for" mind" exploration of large language model society [J]. Advances in neural information processing systems, 2023, 36: 51991-52008.
[16] Shinn N, Cassano F, Gopinath A, et al. Reflexion: Language agents with verbal reinforcement learning [J]. Advances in neural information processing systems, 2023, 36: 8634-8652.
[17] Madaan A, Tandon N, Gupta P, et al. Self-refine: Iterative refinement with self-feedback [J]. Advances in neural information processing systems, 2023, 36: 46534-46594.