Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28513

AgentPress: A Multi-Agent Framework for News Topic Classification with Retrieval, Reasoning, and Reflection

Qi Li

Lanzhou University, Lanzhou, Gansu, China
liq23@lzu.edu.cn

Abstract: Efficient news topic classification is a critical challenge in the era of information overload. While Large Language Models (LLMs) offer a promising alternative to traditional methods, they still face issues with reliability and interpretability. This paper introduces AgentPress, a collaborative multi-agent framework designed to enhance LLM performance and robustness for this task. AgentPress decomposes classification into a pipeline of three specialized agents: initial analysis, Retrieval-Augmented Generation (RAG), and reflective correction. We constructed an eight-category news benchmark and conducted comprehensive ablation studies on the Qwen3-4B and Llama-3.1-8B models. Experiments reveal that the RAG agent is the single most critical component, delivering a substantial performance gain by boosting the macro F1-score of Qwen3-4B from 56.23% to 82.83%. Our findings empirically demonstrate that for structured tasks like news classification, optimizing retrieval-augmentation is a more effective and economical strategy than implementing complex reasoning or reflection mechanisms.

Keywords: Large Language Models, Multi-Agent Systems, News Topic Classification, Retrieval-Augmented Generation, Chain-of-Thought.

1. Introduction

In the current era of information overload, the rapid proliferation of news data has made efficient and accurate topic classification a pivotal task for news media editors, information retrieval systems, and public opinion monitoring platforms. It is also a significant research area within Natural Language Processing (NLP). Precise news topic classification not only enables users to quickly access key information but also provides foundational support for downstream applications such as news recommendation, event tracking, and knowledge graph construction.

Traditional approaches to news classification have relied on feature engineering and machine learning, or on deep learning-based text representations. Although these methods perform well on standard benchmarks, they face persistent limitations, including poor domain adaptability, limited reasoning capabilities for ambiguous texts, and knowledge latency in the face of new events. More recently, Large Language Models (LLMs) have introduced a new paradigm through In-Context Learning (ICL). However, for tasks demanding high accuracy and stability, the reasoning process of LLMs

© 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: 10.54254/2755-2721/2025.LD28513

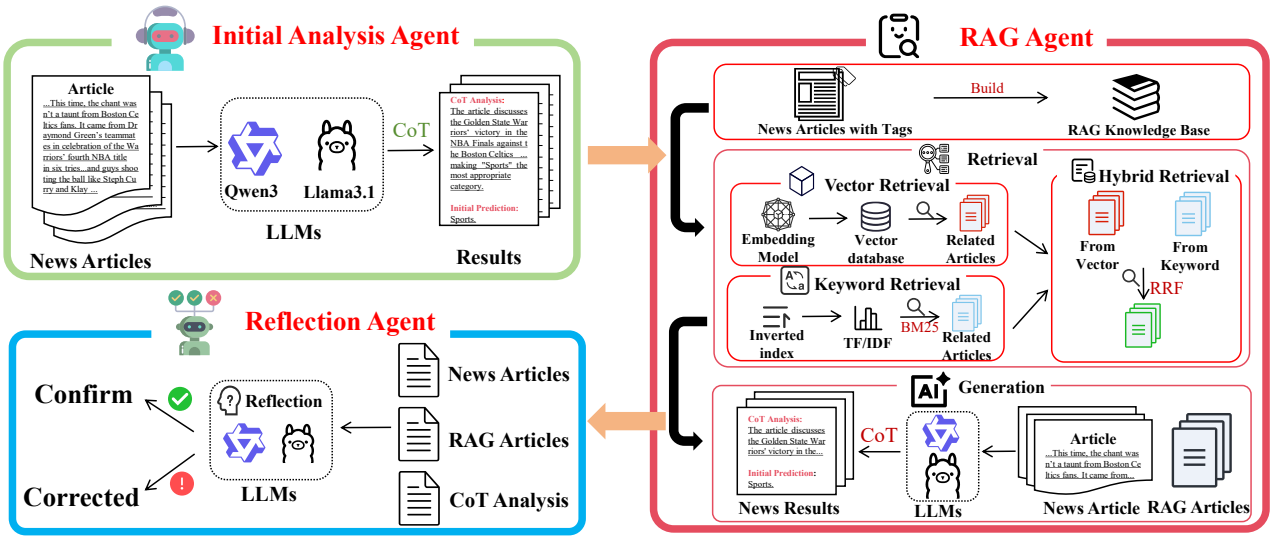

Figure 1: Overview of our multi-agent framework

is often perceived as a black box, and issues such as hallucinations and outdated knowledge [1] can lead to erroneous judgments.

To address these challenges, we introduce AgentPress, a collaborative multi-agent framework for news topic classification. The AgentPress framework consists of three specialized agents: an Initial Analysis Agent for preliminary predictions, a Retrieval-Augmented Generation (RAG) Agent that retrieves relevant examples from external knowledge sources, and a Reflection Agent that validates and corrects the preceding output. Through its modular design, AgentPress synergistically integrates the strengths of each agent to enhance overall performance and robustness.

To systematically evaluate AgentPress, we constructed a high-quality benchmark, which we term NewsTopic-8. This dataset was created by cleaning and restructuring a public dataset from Hugging Face, consolidating 120 fine-grained categories into eight distinct macro-categories. Using NewsTopic-8 with the Qwen3-4B [2] and Llama-3.1-8B [3] models, we assessed performance changes from a zero-shot baseline by progressively integrating RAG, CoT, and reflection mechanisms. Our experimental results indicate that the RAG agent is the core driver of performance, as introducing relevant contextual examples significantly improves both classification accuracy and the macro F1-score. While the subsequent addition of CoT and reflection did not yield significant further improvements, their fine-grained control over the reasoning process enhanced the framework's overall robustness, although their contribution to the main performance metric remained limited.

The main contributions of this paper are as follows:

1. We propose and implement AgentPress, an LLM-based multi-agent collaborative framework that combines initial reasoning, retrieval-augmentation, and reflection to offer an efficient and robust solution for news topic classification.
2. We construct NewsTopic-8, a high-quality dataset for 8-category English news classification comprising 7,195 articles, by restructuring labels to be better suited for evaluating the macro-topic classification capabilities of LLMs.
3. Through extensive comparative and ablation studies, we quantitatively analyze the distinct contributions of RAG, CoT, and reflection mechanisms in news classification, providing empirical evidence to guide the optimization of LLMs in vertical domains.

DOI: 10.54254/2755-2721/2025.LD28513

2. Related work

2.1. News topic classification

Early research in news topic classification relied on a combination of feature engineering and traditional machine learning. Methods such as Bag-of-Words (BoW) and TF-IDF were used to represent text, with classifiers like Support Vector Machines (SVM) demonstrating strong performance [4]. However, these approaches are heavily dependent on feature engineering and struggle to capture deeper semantic information, limiting their performance ceiling.

The advent of deep learning led to methods that could automatically learn text representations. Convolutional Neural Networks (CNNs) proved effective at capturing local n-gram features [5], while Long Short-Term Memory (LSTM) networks were better suited for handling sequential information [6]. The paradigm shifted with the introduction of Pre-trained Language Models (PLMs) like BERT [7], which established new performance benchmarks through a pre-train and fine-tune approach. Nevertheless, this paradigm's reliance on large, in-domain annotated datasets and significant computational resources limits its flexibility and timeliness.

To overcome the limitations of fine-tuning, research has shifted towards leveraging the In-Context Learning (ICL) capabilities of Large Language Models (LLMs) [8]. ICL enables models to perform classification tasks using only a few examples provided in the prompt, without updating model parameters. The performance of ICL is highly dependent on the quality and selection of these examples [9]. To further enhance the reasoning abilities of LLMs, techniques like CoT prompting were developed to guide models through step-by-step reasoning processes [10]. Concurrently, RAG has become a mainstream solution [11,12], with advanced methods like Self-RAG incorporating reflection mechanisms to make retrieval adaptive [13]. Our work is built upon the foundations of ICL and RAG, aiming to identify their optimal application in news classification.

2.2. LLM-based multi-agent collaboration

The paradigm of LLM-based agents has recently seen a surge in interest and development [14]. Multi-agent systems based on LLMs generally follow two primary collaboration patterns. The first is a role-playing debate model, where agents achieve goals through simulated social interaction, as seen in the CAMEL framework [15]. The second is a pipeline-based task decomposition model, which breaks down complex tasks into sequential sub-tasks, often following a Reason-Act paradigm [16], each handled by a specialized agent. This approach is often implemented using general-purpose frameworks like AutoGen [17].

To improve the decision-making quality of agents, researchers have introduced self-reflection mechanisms. For instance, the Reflexion framework enables an agent to learn from its past actions by generating reflective feedback [18]. Similarly, the Self-Refine framework proposes an iterative process where an LLM provides feedback on and refines its own outputs [19]. These approaches, which simulate human metacognitive abilities, offer a promising direction for building more robust systems. In contrast to prior multi-agent systems focused on open-ended tasks like code generation [20] and scientific discovery [21], our work applies a pipeline-based, reflective framework to the structured task of news topic classification to analyze its practical utility.

3. The AgentPress framework

DOI: 10.54254/2755-2721/2025.LD28513

3.1. Overall framework architecture

We designed and implemented AgentPress, a collaborative multi-agent framework for news topic classification. It decomposes the complex classification task into a three-stage pipeline, with each stage executed by a specialized agent. The overall workflow is illustrated in Figure 1.

The framework is composed of three core agents:

1. An Initial Analysis Agent that rapidly generates a candidate classification.
2. An RAG agent that refines the decision using external knowledge.
3. A Reflection Agent that performs final verification and correction from the preceding stage.

3.2. Initial analysis agent

As the starting point of the framework, this agent performs a preliminary analysis and prediction on the input news text. Its core function relies on interacting with an LLM through meticulously designed prompts. We implemented two operational modes to assess the impact of varying reasoning depths:

Zero-shot Direct Inference Mode: In this mode, the agent is given a concise prompt instructing it to act as a professional news editor and select the most appropriate category from the eight predefined options. The prompt enforces a strict output format to establish a cost-effective baseline for the model's classification capability.

Chain-of-Thought (CoT) Mode: Here, the prompt guides the model through a structured three-step process: (1) analyze the article, (2) evaluate the match with potential categories, and (3) derive a final conclusion. The model is required to output its detailed analysis before presenting the final classification. This mode aims to improve accuracy on complex or ambiguous texts and enhance the interpretability of the results.

3.3. RAG agent

The RAG Agent is a core component of the framework, functioning as a knowledge retrieval expert. Its primary task is to provide the LLM with high-quality, task-relevant contextual information in the form of few-shot examples. Upon receiving a news article, this agent retrieves semantically similar and correctly labeled news-category pairs from a pre-built corpus. By integrating these examples into the prompt, it transforms the task from zero-shot learning into few-shot In-Context Learning (ICL), providing the model with immediate, relevant paradigms during inference.

To comprehensively evaluate retrieval effectiveness, we implemented and compared three mainstream retrieval techniques:

1. **Vector Retrieval:** A semantic search method based on dense vector representations. We use the BGE-M3 embedding model [22] to map news articles into high-dimensional vectors. For a query text q and a document d in the corpus, the similarity score is calculated by the cosine similarity between their vector representations $E(q)$ and $E(d)$:

$$\text{Score}_{\text{vec}}(q, d) = \frac{E(q) \cdot E(d)}{\|E(q)\| \|E(d)\|} \tag{1}$$

2. **Keyword Retrieval (BM25):** A classic lexical matching algorithm based on sparse vector representations [23]. For a query q composed of terms t_i, the relevance score for a document d is calculated as:

$$\text{Score}_{\text{BM25}}(q, d) = \sum_{t_i \in q} \text{IDF}(t_i) \cdot \frac{f(t_i, d) \cdot (k_1 + 1)}{f(t_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \tag{2}$$

DOI: 10.54254/2755-2721/2025.LD28513

Where $\text{IDF}(t_i)$ is the inverse document frequency of term t_i, $f(t_i, d)$ is its term frequency in d, and $|d|$ and avgdl are the document length and average document length, respectively. The hyperparameters k_1 and b use the default values from the LangChain library in this study.

3. **Hybrid Retrieval:** To leverage the advantages of both semantic and lexical matching, we use the Reciprocal Rank Fusion (RRF) algorithm [24] to combine the ranked lists from the two methods above. For any given document d, its final RRF score is calculated as:

$$\text{Score}_{\text{RRF}}(d) = \frac{1}{k + \text{rank}_{\text{vec}}(d)} + \frac{1}{k + \text{rank}_{\text{BM25}}(d)} \tag{3}$$

where $\text{rank}_{\text{vec}}(d)$ and $\text{rank}_{\text{BM25}}(d)$ are the ranks of the document in the vector and BM25 retrieval lists, respectively. The smoothing constant k is a hyperparameter, for which we use the default value in LangChain. This approach effectively combines the strengths of different retrieval systems, typically yielding superior retrieval results.

3.4. Reflection agent

The Reflection Agent constitutes the final stage in the AgentPress pipeline, performing rapid verification and, if necessary, correction of the previous stage's output. This agent assumes the persona of an efficient reviewer. Its input includes the original news text x, the predicted category from the previous stage c_{pred}, and the corresponding CoT analysis process t_{trace}. Its decision logic can be formally expressed as a function F_{reflect}:

$$F_{\text{reflect}}(x, c_{\text{pred}}, t_{\text{trace}}) = \begin{cases} \text{CONFIRM}, & \text{if } V(x, c_{\text{pred}}, t_{\text{trace}}) = \text{True} \\ c_{\text{corrected}}, & \text{if } V(x, c_{\text{pred}}, t_{\text{trace}}) = \text{False} \end{cases} \tag{4}$$

where $V(\cdot)$ represents the agent's implicit validation function that assesses the correctness of the initial prediction. If the prediction is deemed correct, the agent outputs a simple CONFIRM signal. If it is deemed incorrect, the agent directly outputs the corrected category, $c_{\text{corrected}}$. This minimalist output design significantly reduces the computational overhead of the reflection stage, making it a lightweight yet effective error-correction module.

4. Experiments and results analysis

4.1. Dataset

Our study utilizes the NewsTopic-8 dataset, which we constructed from the public Topic_Classification corpus on Hugging Face. The original dataset, containing approximately 22,462 articles across 120 fine-grained topics, was refined through a two-step process: (1) Data Cleaning, where we removed samples with missing content or noisy labels and standardized text formats; and (2) Label Reconstruction, where we consolidated the 120 original labels into eight distinct and well-defined macro-categories: Business & Economics, Culture & Entertainment, Health & Medicine, Politics, Science & Environment, Society & Law, Sports, and Technology.

The resulting dataset comprises 7,195 high-quality articles. Following standard practice, we performed a random 80/20 split, yielding a RAG retrieval corpus of 5,756 articles and a test set of 1,439 articles for evaluation. The class distribution, detailed in Table 1, is relatively balanced.

DOI: 10.54254/2755-2721/2025.LD28513

Table 1: Class Distribution and Dataset Splits for NewsTopic-8

Category	Number of Articles
Business & Economics	898
Culture & Entertainment	900
Health & Medicine	901
Politics	901
Science & Environment	901
Society & Law	894
Sports	901
Technology	899
Total	**7,195**
RAG Corpus (80%)	**5,756**
Test Set (20%)	**1,439**

4.2. Experimental design

To systematically evaluate the efficacy of each component within the AgentPress framework, we conducted a series of comparative and ablation studies on the Qwen3-4B and Llama-3.1-8B models. The configurations were as follows:

1. **Baseline (LLM only):** A zero-shot direct inference mode to assess the LLM's foundational classification ability without any enhancements.
2. **Single Component Efficacy:**
   - **+CoT only:** The baseline enhanced solely with CoT reasoning to measure the impact of explicit reasoning.
   - **+RAG:** The baseline augmented with the RAG agent, testing three retrieval strategies (Vector, BM25, and Hybrid) to identify the optimal approach.
3. **Combined Module Efficacy:**
   - **+CoT+RAG:** Combines RAG with CoT to determine if CoT provides additional benefits on top of retrieval-augmented context.
   - **+CoT+RAG+Reflection:** The full framework, adding the Reflection agent to assess the final contribution of the verification and correction step.

4.3. Results and analysis

Evaluation Metrics: We employ four standard metrics: Accuracy (**Acc.**), Macro Precision (**MP**), Macro Recall (**MR**), and Macro F1-Score (**F1**). The **F1** serves as our primary evaluation metric, as it provides a balanced assessment of performance across all categories, which is particularly suitable for our well-distributed dataset.

The comprehensive results of our experiments are detailed in Table 2. Our analysis focuses on three key findings.

RAG is the decisive factor for performance enhancement. As clearly shown in the results table, the introduction of the RAG module brings about a massive and significant performance improvement on both the Qwen3-4B and Llama-3.1-8B models. For example, the macro F1-score for Qwen3-4B in a zero-shot setting was 0.5623. After integrating RAG with hybrid retrieval, this score surged to

DOI: 10.54254/2755-2721/2025.LD28513

Table 2: Main Experimental Results on the NewsTopic-8 Test Set.

	Qwen3-4B				Llama-3.1-8B			
Configuration	**Acc.**	**MP**	**MR**	**F1**	**Acc.**	**MP**	**MR**	**F1**
Baseline (LLM only)	0.5782	0.6435	0.5780	0.5623	0.6685	0.6960	0.6687	0.6636
+RAG(Vector)	0.8165	0.8215	0.8165	0.8166	0.7936	0.8067	0.7936	0.7942
+RAG(BM25)	0.8082	0.8131	0.8082	0.8086	0.7867	0.7973	0.7868	0.7859
+RAG(Hybrid)	**0.8284**	**0.8309**	**0.8283**	**0.8283**	**0.8096**	**0.8212**	**0.8097**	**0.8100**
+CoT only	0.6206	0.6612	0.6207	0.6272	0.6185	0.6765	0.6185	0.6333
+CoT+RAG(Vector)	0.8089	0.8135	0.8089	0.8083	0.7700	0.7880	0.7700	0.7731
+CoT+RAG(BM25)	0.7908	0.8081	0.7908	0.7962	0.7457	0.7710	0.7457	0.7508
+CoT+RAG(Hybrid)	0.8131	0.8284	0.8130	0.8187	0.7589	0.7969	0.7590	0.7720
+CoT only+Reflection	0.6303	0.6701	0.6305	0.6354	0.6595	0.6807	0.6590	0.6547
+CoT+RAG(Vector)+Reflection	0.7665	0.7813	0.7666	0.7664	0.7158	0.7382	0.7156	0.7171
+CoT+RAG(BM25)+Reflection	0.7867	0.8054	0.7867	0.7918	0.7220	0.7456	0.7220	0.7233
+CoT+RAG(Hybrid)+Reflection	0.8054	0.8239	0.8054	0.8115	0.7665	0.7824	0.7665	0.7689

0.8283, an increase of approximately 26.6 percentage points. This phenomenon powerfully demonstrates that for a news classification task, providing the LLM with high-quality, relevant in-context examples is the most direct and effective means of boosting its accuracy. Among the three retrieval strategies, hybrid retrieval generally performed best. On Qwen3-4B, for instance, the macro F1-score of hybrid retrieval (0.8283) was superior to both vector retrieval (0.8166) and BM25 (0.8086). This indicates that a combined strategy leveraging both semantic similarity (Vector) and lexical matching (BM25) can more consistently provide high-quality context, thereby achieving better classification performance.

The marginal utility of CoT and reflection is limited. Compared to the substantial gains brought by RAG, the additional contribution of CoT and the reflection mechanism is relatively limited when RAG is already in place. In some scenarios, their inclusion even led to a slight performance degradation.

- **Efficacy of CoT:** Using CoT alone (+CoT only) resulted in a modest performance increase compared to the baseline, suggesting that guiding the model's thought process is somewhat helpful. However, when CoT was added on top of a powerful RAG context, the macro F1-score for Qwen3-4B dropped from 0.8283 (RAG Hybrid) to 0.8187. This may imply that for a classification task with a well-defined output space, once RAG provides clear contextual examples, the LLM can complete the task efficiently via direct pattern matching. A complex, explicit reasoning chain might, in fact, introduce unnecessary interference or noise.
- **Efficacy of Reflection:** The Reflection Agent demonstrated some corrective capabilities in certain configurations but failed to systematically surpass the optimal RAG-only setup. For instance, the macro F1-score for the full +CoT+RAG(Hybrid)+Reflection configuration was 0.8115 on Qwen3-4B and 0.7689 on Llama-3.1-8B, both of which are lower than their respective +CoT+RAG(Hybrid) counterparts. This suggests that the effectiveness of the reflection mechanism is highly dependent on the quality of the CoT analysis. When the CoT process itself does not yield a performance benefit, the Reflection Agent struggles to fulfill its corrective role and may even amplify biases.

DOI: 10.54254/2755-2721/2025.LD28513

Comparison of model performance. A comparison between the base models reveals interesting differences in their capabilities. In the zero-shot scenario, Llama-3.1-8B (Macro F1-score: 0.6636) outperformed Qwen3-4B (0.5623). However, this gap narrowed significantly and even reversed after introducing RAG. In the optimal RAG (Hybrid) configuration, Qwen3-4B (0.8283) surpassed Llama-3.1-8B (0.8100). This suggests that Llama-3.1-8B may possess stronger inherent zero-shot reasoning, while Qwen3-4B excels at instruction-following and leveraging the contextual patterns provided by RAG.

5. Conclusion

In this study, we proposed and evaluated AgentPress, a multi-agent framework for news topic classification, comprising initial analysis, Retrieval-Augmented Generation (RAG), and reflection agents. Experiments on our eight-category benchmark with Qwen3-4B and Llama-3.1-8B show that RAG is the primary driver of performance, substantially improving accuracy and macro-F1, while explicit reasoning (CoT) and reflection offer limited or even negative gains in this structured task. Moreover, Qwen3-4B leveraged RAG context more effectively than Llama-3.1-8B, achieving the best results. These findings highlight that, in vertical domains, optimizing retrieval-augmentation is more effective than adding complex reasoning or reflection mechanisms, and provide a solid foundation for future research and practical applications in multi-agent LLM systems.

References

[1] Zhibin Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, H. T. Madabushi, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[2] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, et al. Qwen3 Technical Report. arXiv preprint arXiv:2505.09388, 2025.

[3] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, 2024.

[4] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, pages 137–142, 1998.

[5] Yoon Kim. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, 2014.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 4171–4186, 2019.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[9] Jiachang Liu, Ding-bu Shen, Yizhong Zhang, Bill Chen, Wang-Chiew Lee, and Dragomir Radev. What Makes Good In-Context Examples for GPT-3? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 314–321, 2022.

DOI: 10.54254/2755-2721/2025.LD28513

[10] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 601–613, 2022.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Myle Ott, Wen-tau Chen, Alexis Conneau, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[12] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinlong Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Han. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997, 2023.

[13] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. arXiv preprint arXiv:2310.11511, 2023.

[14] Zhen-guo Xi, Wen-hua Chen, Xin Guo, Wei He, Yi-hui Ding, Bei-da Hong, Cheng Zhang, Xu Wang, Hui Lin, Rui-nan Li, et al. The Rise and Potential of Large Language Model Based Agents: A Survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14087–14110, 2023.

[15] Guohao Li, Müge Mair, Matei Zaharia, and Joseph E. Gonzalez. CAMEL: Communicative Agents for "Mind" Exploration of Large Scale Language Model Society. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[16] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Ekin Durmus, Vered Liptchinsky, Ilya Sutskever, and Quoc Le. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

[17] Qingyun Wu, Gagan Bansal, Jie Zhang, Yiran Wu, Shaokun Li, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv preprint arXiv:2308.08155, 2023.

[18] Noah Shinn, Federico Cassano, Ashwin Gopinath, and Karthik Labash. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[19] Aman Madaan, Niket Tandon, Prakhar Gupta, Kevin Hall, Luyu Gao, Bodhisattwa Majumder, S. Nagpal, Chen-tsung Wu, Uri Alon, A. Srivatsan, et al. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[20] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yu Yang, Lain Li, Weinan Zhang, Yexiang Wang, Liyuan He, Han-lin Li, and Jun Wang. MetaGPT: Meta Programming for Multi-Agent Collaborative Framework. arXiv preprint arXiv:2308.00352, 2023.

[21] Andres M. Bran, Sam Cox, Oliver Schilter, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools. arXiv preprint arXiv:2304.05376, 2023.

[22] Jian-hui Chen, Shuo Xiao, Peng Zhang, Kun Luo, De-qiang Lian, and Zhi-zhong Liu. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv preprint arXiv:2402.03216, 2024.

[23] Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic information retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 232–241, 1994.

[24] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion for extreme scale information retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management (CIKM)*, pages 1839–1842, 2009.