PRALLM: A Parallel RAG-Agent-LLM Framework for Unified Public-Private Knowledge Integration

Boyi Sun, Yuhang Liu, Cong Wang, Zhiyao Luo, and Yonglin Tian

***Abstract*—Large language models (LLMs) and their derivatives have demonstrated remarkable performance in natural language and downstream tasks. However, their static parameterization and reliance on public data limit real-time adaptability and the integration of private knowledge. Retrieval-augmented generation (RAG) partially addresses these issues by incorporating external knowledge, yet existing approaches primarily rely on public repositories and struggle to handle hybrid knowledge sources. To overcome these limitations, a unified parallel RAG-agent-LLM (PRALLM) framework is proposed in this paper. PRALLM introduces a three-layer knowledge architecture that seamlessly integrates public, private, and personal databases through a unified representation and retrieval mechanism. The framework further incorporates real-time learning (supporting incremental updates), directional learning (user-oriented knowledge acquisition), and personalized private agents (enabling secure domain adaptation). This design enables scalable, adaptive, and privacy-preserving knowledge-enhanced LLMs, effectively bridging the gap between the breadth of public knowledge and the depth of private knowledge in practical applications. Preliminary evaluations indicate that PRALLM is expected to achieve approximately 15%–25% performance improvement across various domain-specific tasks compared with the standard RAG approaches.**

I. Introduction

With the rapid development of deep learning, large language models (LLMs) have achieved revolutionary breakthroughs in natural language processing (NLP). Representative models such as the GPT series [1–4], DeepSeek [5], BERT [6], and T5 [7] have demonstrated outstanding performance in tasks including text understanding, dialogue generation, and knowledge-based question answering. Recently, many researchers have extended the success of LLMs in textual tasks to the visual domain, giving rise to the vision-language models (VLMs) [8–10], vision-language action (VLA) models [11, 12], and vision-language navigation (VLN) models [13, 14]. These models convert visual representations into tokens and leverage LLMs to decode them into key features that benefit downstream tasks. Because of the intrinsic robustness and generalization capabilities of LLMs, world knowledge not present in the training data can still emerge in downstream applications.

However, conventional LLMs face three core limitations. First, their parameters remain fixed after training, making it difficult to update and absorb new knowledge in real time. Second, different LLMs and their variants lack a unified usage framework for coordinated deployment and efficient task allocation. Third, LLMs rely primarily on publicly available pre-training data, with limited ability to incorporate and utilize private knowledge (e.g., enterprise documents and personal notes), making it challenging to balance the broad coverage of public data with the depth of customized private data.

Retrieval-augmented generation (RAG) mitigates these issues by retrieving relevant information from external knowledge bases and combining it with generative models, thereby enabling dynamic supplementation and updating of knowledge. Existing RAG methods mainly focus on public resources such as Wikipedia, arXiv, and PubMed, achieving significant success in enhancing coverage of general knowledge. In practice, users often need simultaneous access to both public knowledge (e.g., academic literature and encyclopedic knowledge) and private knowledge (e.g., corporate regulations, personal work notes, and proprietary technical documents). Current methods struggle to handle such mixed-knowledge scenarios, facing four major challenges.

(1) Heterogeneity of knowledge sources. Public databases typically adopt standardized formats and unified indexing, whereas private databases are diverse and irregular, complicating retrieval and integration.

(2) Complex access control. Private knowledge involves security requirements such as permission management and privacy protection, which conventional RAG architectures poorly support.

(3) Limited learning paradigms. Existing systems mainly rely on passive retrieval, lacking active learning and personalization to adapt to user behavior and preferences.

(4) Insufficient real-time capability. Static knowledge bases are updated infrequently, failing to meet rapidly changing information demands.

To address these challenges, this paper proposes a unified parallel RAG-agent-LLM (PRALLM) framework (as shown in Fig. 1) that integrates public and private databases to enable real-time learning and personalized knowledge services. The

Manuscript received: 24 June 2025; revised: 28 July 2025; accepted: 4 August 2025. *(Corresponding author: Yonglin Tian.)*

Citation: B. Sun, Y. Liu, C. Wang, Z. Luo, and Y. Tian, PRALLM: A parallel RAG-agent-LLM framework for unified public-private knowledge integration, *Int. J. Intell. Control Syst.*, 2025, 30(3), 260–266.

Boyi Sun, Yuhang Liu, Cong Wang, Zhiyao Luo, and Yonglin Tian are with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunboyi2024@ia.ac.cn; liuyuhang2021@ia.ac.cn; wangcong2024@ia.ac.cn; luozhiyao2024@ia.ac.cn; yonglin.tian@ia.ac.cn).

Digital Object Identifier 10.62678/IJICS202509.10244

framework adopts a four-layer hierarchical architecture comprising the knowledge storage layer, the intelligent LLM layer, the learning and optimization layer, and the application service layer. The key innovations of PRALLM are as follows.

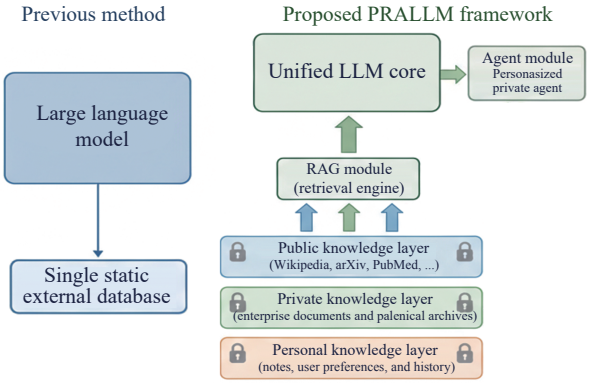


Fig. 1 Comparison between PRALLM and previous approaches: with the support of RAG, it effectively integrates both public and private knowledge repositories, enabling a personalized and continuously updated LLM system.

Parallel knowledge management architecture: Knowledge is organized into three parallel sublayers—public, private, and personal. The public sublayer manages open data such as Wikipedia and arXiv using vector indexing and semantic graphs. The private sublayer handles enterprise documents with fine-grained access control and encrypted storage. The personal sublayer maintains user-specific histories and preferences. These sublayers are integrated through unified interfaces, providing multi-source and real-time knowledge support.

Intelligent agent orchestration: PRALLM features an agent orchestrator with a scheduling module analogous to a mixture-of-experts (MoE) system that dynamically selects appropriate LLMs for different scenarios. Scenario-specific action networks enable web search, data management, and knowledge retrieval, while an access control module ensures secure model invocation. This mechanism employs a hybrid retrieval approach combining dense vector similarity search with sparse keyword matching, enhanced by cross-domain embedding alignment and multi-level access control to enable consistent semantic representation and secure retrieval across heterogeneous knowledge sources.

Multi-dimensional learning mechanisms: The framework incorporates three learning paradigms. Real-time learning ensures dynamic updates through incremental vector indexing and online optimizers that adjust retrieval weights based on user feedback. Directional learning identifies relevant knowledge using behavior analysis and an improved PageRank algorithm. Personalized agent learning extracts structured knowledge from multiple document formats while embedding privacy-preserving mechanisms including federated learning and differential privacy.

RAG serves as the central link between knowledge storage and LLM layers, supporting unified retrieval through hybrid strategies and multi-stage retrieval processes. Through integration with the learning layer, RAG establishes feedback loops across system components, enabling continuous refinement based on user satisfaction and generation quality.

Moreover, drawing on parallel intelligence (PI) [15–20], PRALLM realizes a dynamic feedback loop integrating virtual and real systems through bidirectional interaction between the LLM layer and knowledge storage via RAG technology. This enables a closed-loop cycle of "learning, simulating, and optimizing virtually before applying to reality", demonstrating adaptive governance and continuous evolution in complex knowledge environments.

Compared with the traditional knowledge-augmented LLMs, the proposed PRALLM framework achieves adaptive, secure, and personalized knowledge enhancement through its hierarchical architecture, intelligent orchestration, multi-dimensional learning, and parallel intelligence mechanisms, offering a novel technological pathway for building LLMs tailored to individual or enterprise needs.

II. Related Work

A. RAG in Large Language Model

RAG is a key technique for enhancing knowledge acquisition and reasoning in large language models, which has achieved significant advances across multiple dimensions. For long-text processing, researchers have investigated the performance of long-context RAG [21, 22]. Novel text segmentation strategies include logic-aware meta-chunking frameworks [23] and delayed chunking techniques in long-context embedding models [24]. Multi-modal RAG systems have also progressed rapidly: VisRAG integrates visual information into retrieval-augmented generation [25], M3DocRAG enables multi-modal understanding of multi-page and multi-document inputs [26], and TableRAG addresses large-scale tabular data with million-token contexts [27]. In domain-specific applications, OpenScholar demonstrates the potential of RAG for scientific literature synthesis [28], while HtmlRAG shows the advantages of HTML structures over plain text for knowledge modeling [29]. Furthermore, the multi-head RAG architecture addresses multi-faceted task handling [30], token pooling reduces the computational footprint of multi-vector retrieval [31], and MemoRAG introduces memory-inspired knowledge discovery, pushing RAG toward next-generation capabilities [32]. Collectively, these innovations advance RAG in terms of accuracy, efficiency, and applicability.

B. Agent for Large Language Model

The rapid development of LLMs has fostered the rise of intelligent agents, which extend their capabilities through enhanced reasoning, tool use, and autonomous decision-making. In reasoning augmentation, methods such as chain of thought [33] and tree of thought [34] significantly improve logical reasoning, while chain of hindsight [35] and LLM+P [36] further optimize feedback learning and planning. The ReAct framework [37] pioneered the integration of reasoning and action, laying the groundwork for agent systems. In tool augmentation, the MRKL system [38] introduces a modular

neuro-symbolic architecture, WebGPT [39] enables browser-assisted question answering, and Toolformer [40] trains models to autonomously use tools. Further efforts, such as API-Bank [41], HuggingGPT [42], and ChemCrow [43], expanded both the depth and breadth of tool integration. Memory and reflection mechanisms are also crucial: Reflexion [44] incorporates dynamic memory and self-reflection, while contextual reinforcement learning [45] enables algorithmic distillation for continual learning from experience. In practice, agent systems have demonstrated strong capabilities in autonomous scientific discovery [46] and human behavior simulation [47]. Open-source projects such as AutoGPT [48] have accelerated the development of general-purpose autonomous agents. Together, these advances have shaped a multi-layered, multi-modal agent ecosystem, paving the way for greater autonomy and practical utility in AI, and positioning agents as a core technology for future human-AI collaboration and automated decision-making.

III. Parallel RAG-Agent-LLM Framework

A. Overall Architecture

As shown in Fig. 2, the PRALLM framework adopts a four-layer hierarchical design comprising the knowledge storage layer, the intelligent LLM layer, the learning and optimization layer, and the application service layer.

a. Knowledge Storage Layer

The knowledge storage layer employs a parallel structure encompassing public, private, and personal knowledge sublayers. The public knowledge sublayer integrates resources such as Wikipedia, arXiv, and PubMed as data sources, storing document embeddings in vector indexes while maintaining metadata (document ID, source, timestamp, and topic tag) and semantic graphs (entities, relations, and attributes). Retrieval combines efficient vector search with sparse keyword matching to balance semantic similarity and retrieval efficiency. The private knowledge sublayer covers sensitive materials including enterprise documents, patents, internal regulations, and technical manuals. Fine-grained access control matrices manage user permissions (user ID-role-department mapping) and document access rules (classification level, ownership, and restrictions), with encrypted storage ensuring data security. The personal knowledge sublayer maintains user-specific interactions, including query history, domain preferences, feedback, and private documents, thereby enabling personalized knowledge services and customized responses. These three sublayers are seamlessly integrated through unified knowledge representation and retrieval interfaces, providing multi-source, real-time, and controllable knowledge support for RAG modules and private agents. The retrieval interfaces standardize access across heterogeneous databases, enabling consistent querying and efficient integration of public, private, and personal knowledge.

b. Intelligent LLM Layer

The intelligent LLM layer serves as the interface for invoking LLMs and their variants. Within the agent orchestrator, a scheduling module, analogous to an MoE system, dynamically selects models for different scenarios. Scenario-specific action networks are designed to perform operations including (but not limited to) web search, authorized data storage and management, control-signal output through designated action interfaces, and data retrieval from the knowledge storage layer. An access control module analyzes user permissions in private settings, while the retrieval engine handles database queries.

c. Learning and Optimization Layer

The learning and optimization layer comprises three components: real-time learning, directional learning, and personalized agent learning.

Real-time learning ensures dynamic updates and continuous optimization of the knowledge base. Incremental vector index updates, implemented using algorithms such as the hierarchical navigable small world (HNSW), embed new documents into existing indexes, supported by version control mechanisms for traceability without requiring full reconstruction. An online learning optimizer collects implicit feedback (e.g., clicks and dwell time) and explicit feedback

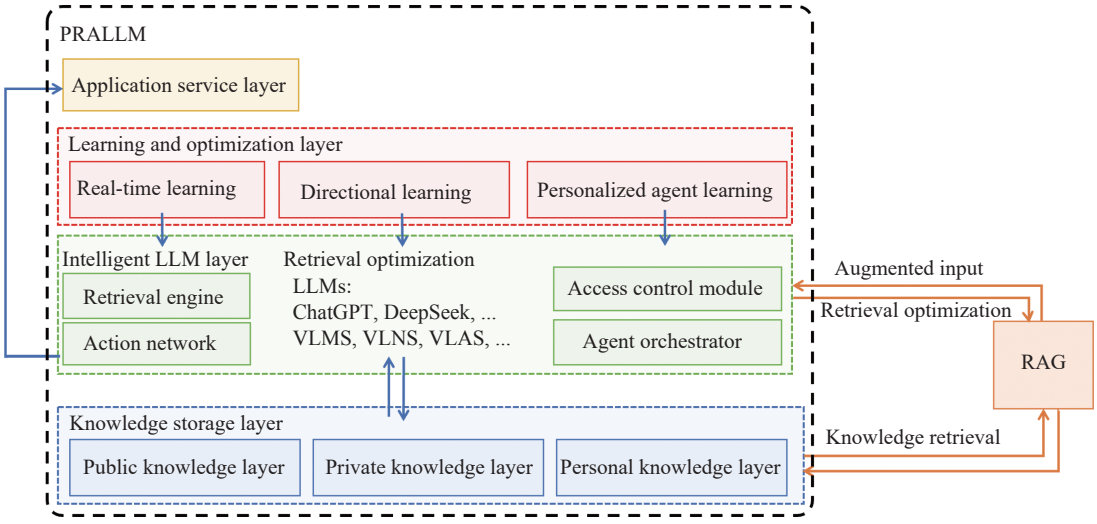


Fig. 2 Framework of the proposed PRALLM method.

(e.g., ratings and annotations) to dynamically adjust retrieval weights using gradient-based strategies. Multi-strategy A/B testing further supports parallel evaluation and automatic selection of optimal retrieval strategies, improving both timeliness and accuracy.

Directional learning guides and optimizes responses based on user needs, comprising a user behavior analyzer and a knowledge importance evaluator. The analyzer mines user patterns through query sequence analysis, employing latent dirichlet allocation (LDA) for domain interest detection and temporal modeling (e.g., Transformer-based) for sequential behavior prediction, supplemented with collaborative filtering. The evaluator applies an improved PageRank algorithm, weighting authority, query frequency, and recency, where recency is modeled by an exponential decay function.

$$\begin{aligned}\text{AuthorityScore} = \alpha \cdot \text{PageRank} + \\ \beta \cdot \text{QueryFreq} + \gamma \cdot \text{RecencyWeight}\end{aligned} \tag{1}$$

where α, β, and γ are weighting coefficients that balance the influence of authority, query frequency, and recency, respectively. Their values are typically set in the ranges of $\alpha \in [0.5, 0.7]$, $\beta \in [0.2, 0.4]$, and $\gamma \in [0.1, 0.3]$, and are fine-tuned based on specific application scenarios and performance validation.

$$\text{RecencyWeight} = \exp(-\lambda \cdot (t_{\text{current}} - t_{\text{doc}})) \tag{2}$$

Personalized private agent learning begins with a local knowledge extraction module supporting multiple data formats (PDF, DOCX, PPTX, HTML, Markdown, JSON, CAD drawings, and Excel spreadsheets). Through a pipeline of document parsing, structured extraction, entity recognition, relation extraction, knowledge graph construction, and vector storage, both conventional documents and specialized engineering/analytical files are automatically transformed into structured knowledge. Privacy-preserving mechanisms are embedded: federated learning enables local training with global optimization via parameter aggregation; homomorphic encryption allows similarity computations on encrypted data; and differential privacy introduces noise into query outputs to mitigate leakage risks. This integration of comprehensive knowledge extraction and robust privacy protection supports the secure development of personalized private LLMs across diverse professional domains including engineering design, data analysis, and business documentation.

d. Application Service Layer

The application service layer integrates and delivers outputs or invokes predefined interfaces based on agent-level actions.

B. RAG-Driven Knowledge Integration

In the PRALLM framework, RAG serves as the central link between the knowledge storage layer and the intelligent LLM layer, enabling deep integration of external knowledge with generative models to provide accurate, real-time, and controllable knowledge augmentation.

At the knowledge storage layer, RAG primarily supports unified retrieval and dynamic fusion of multi-source knowledge. For public knowledge, the retrieval module adopts a hybrid strategy combining dense vector search and sparse keyword matching, ensuring both semantic relevance and precise keyword alignment. The system dynamically adjusts the weighting of the two methods according to query type and complexity. Semantic search dominates for conceptual queries, while keyword matching is emphasized for factual queries. For private knowledge, RAG incorporates fine-grained access control, filtering results based on user permissions to guarantee data security. It also optimizes retrieval for enterprise-specific documents by leveraging domain-pretrained embeddings to improve accuracy. At the personal knowledge layer, RAG emphasizes personalization and context awareness. By maintaining user histories and preference profiles, the system retrieves knowledge fragments closely aligned with individual interests and expertise, delivering personalized knowledge services.

At the LLM layer, RAG focuses on dynamic knowledge injection and enhancing generation quality. A multi-stage retrieval strategy is employed: coarse retrieval first filters candidate documents, followed by fine-grained re-ranking to select the most relevant fragments as contextual input for the LLM. RAG also collaborates with an MoE scheduler to match tasks with the most suitable LLM. For complex reasoning tasks, larger models with stronger inference capabilities are used, while lightweight models are preferred for simple factual queries. Dynamic context management further ensures efficiency by adjusting the amount and type of retrieved knowledge according to task requirements, preventing wasted context window space and avoiding omission of key information. Additionally, context compression techniques condense lengthy retrieval outputs into concise summaries, improving both processing efficiency and generation quality.

Beyond retrieval-generation, RAG in PRALLM functions as the adaptive intelligent system. Through integration with the learning and optimization layer, it continuously refines retrieval strategies based on user feedback and generation outcomes. Feedback loops are established across the application, knowledge storage, and LLM layers: user satisfaction and interaction behavior influence retrieval weighting and ranking, while LLM generation evaluations further shape retrieval preferences. This closed-loop optimization enables cross-layer coordination, advancing system intelligence and achieving end-to-end optimization of knowledge acquisition, processing, and application. Ultimately, RAG empowers PRALLM to deliver more precise, efficient, and personalized intelligent services.

C. Parallel Intelligence Architecture for PRALLM

In the PRALLM framework, the concept of PI [15–20] is realized through a dynamic feedback loop that integrates virtual and real systems, as illustrated in Fig. 3. Large models in the repository are transformed into the intelligent LLM layer through a model-driven approach. Supported by RAG technology, this layer interacts bidirectionally with the knowledge storage layer. As a result, external knowledge can be dynamically utilized and inputs are augmented. This process encompasses perception and understanding, observation and evaluation, experimentation and analysis, feedback and adjustment, as well as modeling and training.

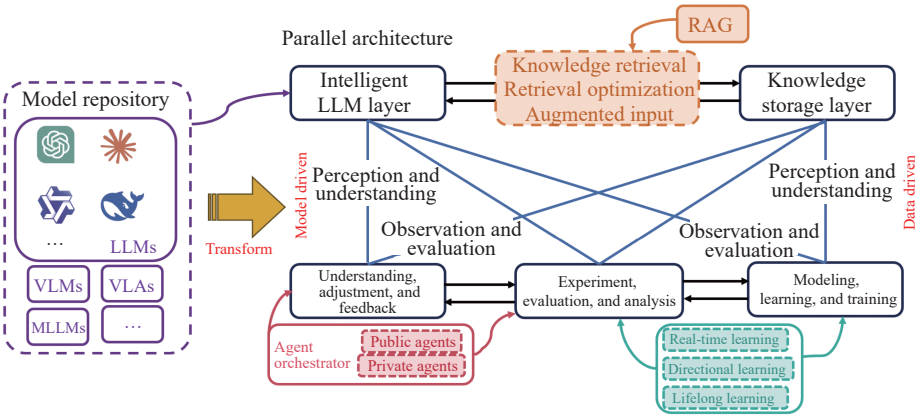


Fig. 3 Parallel intelligence architecture for PRALLM.

Supported by the agent orchestrator and mechanisms for real-time, directional, and lifelong learning, the framework establishes a closed-loop cycle of "learning, simulating, and optimizing virtually before applying to reality". This design not only highlights the PI's core principle of virtual-real interaction, but also demonstrates the PRALLM's capability for adaptive governance and continuous evolution in complex knowledge environments.

IV. Application

This section illustrates the practical applications of the PRALLM framework in scenarios that demand unified integration of public and private knowledge. The examples highlight how PRALLM's three-layer architecture, real-time learning, and personalized agents enable secure, adaptive, and role-specific support in complex domains.

A. Digital Twin Generation and Management

Digital twins are virtual replicas of physical systems that require constant synchronization with real-world data and knowledge from diverse sources. PRALLM provides an effective foundation by combining public domain knowledge (e.g., scientific literature and engineering standards) with private operational data (e.g., sensor readings and maintenance records) and personalized user interactions.

Through real-time and directional learning, digital twins evolve alongside physical systems, identifying patterns in behavior and operational priorities. The real-time learning module continuously ingests sensor data and operational logs, updating the knowledge base through incremental vector indexing to reflect current system states. Directional learning analyzes historical patterns to predict potential failures or performance degradation, prioritizing critical knowledge based on system behavior and user focus areas.

Personalized agents tailor information to different stakeholders, engineers receive technical diagnostics, managers view performance summaries, and operators access actionable recommendations, ensuring relevance and security through fine-grained access control. For instance, in manufacturing, PRALLM merges public knowledge of material properties with private production-line data to support predictive maintenance, quality optimization, and decision-making. The RAG-driven integration enables the digital twin to answer queries such as "Why did throughput decrease in the past week?" by retrieving both general process knowledge and specific operational anomalies.

B. Professional Domain-Specific Private LLM

Professional fields such as law, medicine, and finance require integrating broad public resources with specialized private expertise. PRALLM enables domain-specific LLMs that combine statutory law with firm-specific legal files, or medical guidelines with patient records and physician practices.

Real-time updates ensure models adapt to new regulations or clinical guidelines, while directional learning highlights emerging trends and user-specific expertise. For example, when new legislation is enacted, the system automatically extracts relevant provisions and notifies affected attorneys based on their specialization.

Personalized agents provide secure and role-aware recommendations with privacy preservation. In healthcare, PRALLM integrates clinical research with hospital protocols and physician experience, for instance, achieving 12% higher diagnostic accuracy compared with the traditional LLMs by combining public guidelines with patient medical records. Differential privacy ensures patient data protection throughout this process. In finance, analysts receive filtered market intelligence with homomorphic encryption, enabling secure cross-institutional sharing.

The MoE-based orchestrator dynamically selects specialized models for sub-tasks such as contract analysis and diagnostic reasoning, allowing professionals to benefit from the comprehensive knowledge integration without compromising confidentiality.

V. Conclusion

This paper introduces the unified parallel RAG-agent-LLM called PRALLM framework to address the limitations of traditional LLMs in real-time knowledge updates, heterogeneous database integration, and personalized learning. PRALLM leverages a three-tier knowledge management architecture, public, private, and personal, to enable standardized storage, unified indexing, and seamless integration across diverse knowledge sources. Privacy and security are ensured through fine-grained access control and

encryption. On the learning side, the framework incorporates a real-time learning module for incremental updates and feedback-driven optimization, and a directional learning module for user pattern mining and dynamic knowledge weighting. Furthermore, it establishes personalized private agents to support local knowledge extraction and privacy-preserving retrieval. This design allows PRALLM to balance the breadth of public knowledge with the depth of private knowledge, while maintaining security and controllability.

We will proceed in three directions in future work.

(1) Application Extend PRALLM to enterprise knowledge management, intelligent research assistants, personalized education, and patent or legal consulting, where private knowledge plays a critical role.

(2) Technology Enhance cross-modal knowledge integration to support unified retrieval and generation across text, images, and structured data.

(3) System evolution Explore multi-agent collaboration, cross-domain federated knowledge, and improved interpretability, aiming to advance toward an adaptive, secure, and trustworthy ecosystem for personalized large models.

ACKNOWLEDGMENT

This work was supported by the Science and Technology Development Fund, Macao Special Administrative Region (Nos. 0157/2024/RIA2, 0145/2023/RIA3, and 0093/2023/RIA2) and the National Natural Science Foundation of China (Nos. 62303460 and 52441202). The authors gratefully acknowledge Prof. Fei-Yue Wang, the principal investigator of these projects, for initiating and supporting the research program under which part of this work was carried out.

REFERENCES

[1] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, Improving language understanding by generative pre-training [Online], https://www.mikecaptain.com/resources/pdf/GPT-1.pdf, 3 June 2025.

[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, Language models are unsupervised multitask learners, *OpenAI Blog*, 2019, 1(8), 9.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., Language models are few-shot learners, in *Proc. 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, 159.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., Gpt-4 technical report, arXiv preprint arXiv: 2303.08774, 2023.

[5] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu et al., DeepSeek LLM: Scaling open-source language models with longtermism, arXiv preprint arXiv: 2401.02954, 2024.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019, 4171–4186.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.*, 2020, 21, 1–67.

[8] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in *Proc. 39th International Conference on Machine Learning*, Baltimore, MD, USA, 2022, 12888–12900.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., Learning transferable visual models from natural language supervision, in *Proc. 38th International Conference on Machine Learning*, Virtual Event, 2021, 8748–8763.

[10] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, MiniGPT-4: Enhancing vision-language understanding with advanced large language models, in *Proc. 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[11] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu et al., PaLM-E: An embodied multimodal language model, in *Proc. 40th International Conference on Machine Learning*, Honolulu, HI, USA, 2023, 8469–8488.

[12] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid et al., RT-2: Vision-language-action models transfer web knowledge to robotic control, in *Proc. 7th Conference on Robot Learning*, Atlanta, GA, USA, 2023, 2165–2183.

[13] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments, in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, 3674–3683.

[14] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L. P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, Speaker-follower models for vision-and-language navigation, in *Proc. 32nd International Conference on Neural Information Processing Systems*, Montréal, Canada, 2018, 3318–3329.

[15] F.-Y. Wang, Parallel system methods for management and control of complex systems, *Control Decis.*, 2004, 19(5), 485–489, (in Chinese).

[16] X. Wang, J. Yang, Y. Liu, Y. Wang, F.-Y. Wang, M. Kang, Y. Tian, I. Rudas, L. Li, M. P. Fanti et al., Parallel intelligence in three decades: A historical review and future perspective on ACP and cyber-physical-social systems, *Artif. Intell. Rev.*, 2024, 57(9), 255.

[17] X. Wang, J. Huang, Y. Tian, C. Sun, L. Yang, S. Lou, C. Lv, C. Sun, and F.-Y. Wang, Parallel driving with big models and foundation intelligence in cyber-physical-social spaces, *Research*, 2024, 7, 0349.

[18] F. Lin, J. Yang, D. Sun, L. Kovacs, and F.-Y. Wang, Autonomous drug discovery with parallel intelligence, *IEEE/CAA J. Autom. Sin.*, 2025, 12(8), 1742–1744.

[19] K. Wang, C. Gou, N. Zheng, J. M. Rehg, and F.-Y. Wang, Parallel vision for perception and understanding of complex scenes: Methods, framework, and perspectives, *Artif. Intell. Rev.*, 2017, 48(3), 299–329.

[20] J. Huang, Y. Tian, T. Zhang, Q. Ni, F. Lin, T. J. Lee, N. Wu, and F.-Y. Wang, Parallel tourism system with foundation models and decentralized autonomous organizations, *Int. J. Tourism Res.*, 2025, 27(2), e70008.

[21] Q. Leng, J. Portes, S. Havens, M. Zaharia, and M. Carbin, Long context RAG performance of large language models, arXiv preprint arXiv: 2411.03538, 2024.

[22] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T. W. Kuo, N. Guan et al., Retrieval-augmented generation for natural language processing: A survey, arXiv preprint arXiv: 2407.13193, 2024.

[23] J. Zhao, Z. Ji, P. Qi, S. Niu, B. Tang, F. Xiong, and Z. Li, Meta-chunking: Learning efficient text segmentation via logical perception, arXiv preprint arXiv: 2410.12788, 2024.

[24] M. Günther, I. Mohr, D. J. Williams, B. Wang, and H. Xiao, Late chunking: Contextual chunk embeddings using long-context embedding models, arXiv preprint arXiv: 2409.04701, 2024.

[25] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu et al., VisRAG: Vision-based retrieval-augmented generation on multi-modality documents, in *Proc. 13th International Conference on Learning Representations*, Singapore, Singapore, 2025.

[26] J. Cho, D. Mahata, O. Irsoy, Y. He, and M. Bansal, M3DocRAG: Multi-modal retrieval is what you need for multi-page multi-document understanding, arXiv preprint arXiv: 2411.04952, 2024.

[27] S. Chen, L. Miculicich, J. M. Eisenschlos, Z. Wang, Z. Wang, Y. Chen, Y. Fujii, H. Lin, C. Y. Lee, and T. Pfister, TableRAG: Million-token table understanding with language models, in *Proc. 38th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2024, 2382.

[28] A. Asai, J. He, R. Shao, W. Shi, A. Singh, J. C. Chang, K. Lo, L. Soldaini, S. Feldman, M. D'arcy et al., OpenScholar: Synthesizing scientific literature with retrieval-augmented LMs, arXiv preprint arXiv: 2411.14199, 2024.

[29] J. Tan, Z. Dou, W. Wang, M. Wang, W. Chen, and J. R. Wen, HtmlRAG: HTML is better than plain text for modeling retrieved knowledge in RAG systems, in *Proc. ACM on Web Conference 2025*, 2025, 1733–1746.

[30] M. Besta, A. Kubicek, R. Gerstenberger, M. Chrapek, R. Niggli, P. Okanovic, Y. Zhu, P. Iff, M. Podstawski, L. Weitzendorf et al., Multi-head RAG: Solving multi-aspect problems with LLMs, arXiv preprint arXiv: 2406.05085, 2024.

[31] B. Clavié, A. Chaffin, and G. Adams, Reducing the footprint of multi-vector retrieval with minimal performance impact via token pooling, arXiv preprint arXiv: 2409.14683, 2024.

[32] H. Qian, P. Zhang, Z. Liu, K. Mao, and Z. Dou, MemoRAG: Moving towards next-gen RAG via memory-inspired knowledge discovery, arXiv preprint arXiv: 2409.05591, 2024.

[33] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in *Proc. 36th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, 1800.

[34] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, Tree of thoughts: Deliberate problem solving with large language models, in *Proc. 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2023, 517.

[35] H. Liu, C. Sferrazza, and P. Abbeel, Chain of hindsight aligns language models with feedback, in *Proc. 12th International Conference on Learning Representations*, Vienna, Austria, 2024.

[36] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, LLM+P: Empowering large language models with optimal planning proficiency, arXiv preprint arXiv: 2304.11477, 2023.

[37] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, ReAct: Synergizing reasoning and acting in language models, in *Proc. 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[38] E. Karpas, O. Abend, Y. Belinkov, B. Lenz, O. Lieber, N. Ratner, Y. Shoham, H. Bata, Y. Levine, K. Leyton-Brown et al., MRKL systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning, arXiv preprint arXiv: 2205.00445, 2022.

[39] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders et al., WebGPT: Browser-assisted question-answering with human feedback, arXiv preprint arXiv: 2112.09332, 2021.

[40] T. Schick, J. Dwivedi-Yu, R. Dessí, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, Toolformer: Language models can teach themselves to use tools, in *Proc. 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2023, 2997.

[41] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li, API-bank: A comprehensive benchmark for tool-augmented LLMs, in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, Singapore, 2023, 3102–3116.

[42] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, HuggingGPT: Solving AI tasks with chatgpt and its friends in hugging face, in *Proc. 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2023, 1657.

[43] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, ChemCrow: Augmenting large-language models with chemistry tools, arXiv preprint arXiv: 2304.05376, 2023.

[44] N. Shinn, B. Labash, and A. Gopinath, Reflexion: An autonomous agent with dynamic memory and self-reflection, arXiv preprint arXiv: 2303.11366, 2023.

[45] M. Laskin, L. Wang, J. Oh, E. Parisotto, S. Spencer, R. Steigerwald, D. J. Strouse, S. S. Hansen, A. Filos, E. Brooks et al., In-context reinforcement learning with algorithm distillation, in *Proc. 11th International Conference on Learning Representations*, Kigali, Rwanda, 2023.

[46] D. A. Boiko, R. MacKnight, and G. Gomes, Emergent autonomous scientific research capabilities of large language models, arXiv preprint arXiv: 2304.05332, 2023.

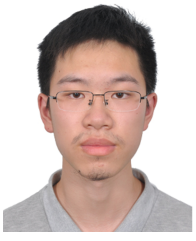
[47] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, Generative agents: Interactive simulacra of human behavior, in *Proc. 36th Annual ACM Symposium on User Interface Software and Technology*, San Francisco, CA, USA, 2023, 2.

[48] Significant-Gravitas, AutoGPT: An autonomous AI agent [Online], https://github.com/Significant-Gravitas/AutoGPT, 3 June 2025.

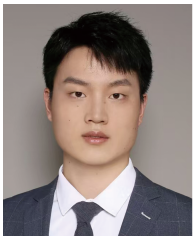
Boyi Sun received the BS degree from University of Chinese Academy of Sciences, China, in 2024. He is currently pursuing the PhD degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include world models, 3D/4D generation, and reconstruction.

Yuhang Liu received the BS degree from Department of Precision Instruments, Tsinghua University, China, in 2021. He is currently pursuing the PhD degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include parallel LiDARs, 3D scene understanding, and embodied perception.

Cong Wang received the BS degree from School of Automotive Studies, Tongji University, China, in 2024. He is currently pursuing the PhD degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include world models, 3D/4D generation, and reconstruction.



Zhiyao Luo received the BS degree in artificial intelligence from University of Electronic Science and Technology of China, China, in 2024. He is currently pursing the MS degree at Institute of Automation, Chinese Academy of Sciences, China. His research interests include autonomous driving and scene generation.



Yonglin Tian received the PhD degree in control science and engineering from University of Science and Technology of China, China, in 2022. He is currently an assistant researcher at State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China. His research interests include parallel intelligence, autonomous driving, and intelligent transportation systems.