

Foundation Models for Information Retrieval and Knowledge Processing

Shuo Shang¹, Renhe Jiang², Ryosuke Shibasaki², Rui Yan³

¹University of Electronic Science and Technology of China, North Jianshe Road, Chengdu 610054, China

²University of Tokyo, Japan

³Renmin University of China, Zhongguancun Street, Beijing 100872, China

Citation: Shang S., Jiang R.H., Shibasaki R., et al.: Foundation Models for Information Retrieval and Knowledge Processing. Data Intelligence 6(4), 891-892 (2024). doi: <https://doi.org/10.3724/2096-7004.di.2024.0067>

This special issue on *Foundation Models for Information Retrieval and Knowledge Processing* highlights the transformative potential of foundation models across a range of applications. The six selected papers explore innovative methodologies and practical implementations that enhance the efficiency, accuracy, and scalability of information retrieval systems and knowledge processing frameworks, which are summarized as follow.

Wei et al. [1] introduce a framework that leverages contrastive learning and attention mechanisms to enhance remote sensing image change detection. The proposed model, which employs dictionary learning to refine feature extraction, demonstrates significant improvements in cross-domain learning tasks.

Luo et al. [2] present Dr.ICL, a demonstration-retrieved framework designed to enhance the performance of large language models (LLMs) in few-shot learning tasks. By retrieving semantically similar demonstrations, Dr.ICL improves both general and instruction-finetuned LLMs. The research demonstrates improved reasoning capabilities through the use of Chain-of-Thought prompting and highlights the advantages of integrating retrieval-based techniques for complex reasoning tasks.

Zhang et al. [3] propose a multi-agent framework that addresses the challenges of exploratory data analysis (EDA) using large language models (LLMs). By decomposing the data analysis task into subtasks, each assigned to different agents, the framework leverages tools and external knowledge sources to optimize the accuracy and depth of analysis.

Oida-Onesa and Ballera [4] explore the fine-tuning of pre-trained transformer models on a newly developed parallel corpus for two low-resource languages, Bikol and Filipino. BFPaCo corpus is introduced, demonstrating fine-tuning models like T5 and mT5 are capable of exhibiting significant improvements in translation accuracy. The results highlight the importance of creating specialized corpora for underrepresented languages in the field of natural language processing.

Dai et al. [5] present a multi-modality fusion-based framework, DMDroid, to detect new malware classes. By integrating static features from graph based and image based modalities, enhanced with a multi-head attention mechanism, DMDroid addresses the challenges of detecting malware in dynamic environments.

Gao et al. [6] present a privacy preserving publish/subscribe system for processing spatial keyword queries over geo-textual data streams. The authors introduce a novel privacy-preserving publish/subscribe framework that efficiently processes large-scale spatial keyword queries while maintaining user privacy. By employing advanced indexing techniques and privacy-enhancing algorithms, the system is able to match queries in real-time without compromising user location or preference data (DI-2024-0043).

In conclusion, the six papers in this issue collectively advance the field of information retrieval and knowledge processing by introducing novel frameworks and methodologies to address critical challenges such as privacy preservation, domain-specific language model fine-tuning, and multi-modality data fusion. These contributions demonstrate the broad applicability and transformative potential of foundation models across diverse domains, offering valuable insights for future research and development.

REFERENCES

- [1] Wei T., Liu Y.F., Zhao R.: Dictionary Attention-Weighted Cross-Domain Contrastive Learning for Remote Sensing Image Change Detection. *Data Intelligence* 6(4), 893–908 (2024)
- [2] Luo M., Xu X., Dai Z.Y., Panupong P., Mehran K., Chitta, Vaiva I., Vincent Y Zhao.: Dr. ICL: Demonstration-Retrieved In-context Learning. *Data Intelligence* 6(4), 909–922 (2024)
- [3] Zhang X.L., Mao Z.X., Chen Z.W., et al.: Effective Tool Augmented Multi-Agent Framework for Data Analysis. *Data Intelligence* 6(4), 923–945 (2024)
- [4] Rosel O.O., Melvin A. B.: Fine Tuning Language Models: A Tale of Two Low-Resource Languages. *Data Intelligence* 6(4), 946–967 (2024)
- [5] Dai X., Yu Z., Liang C.L., Gao C.Y., He Q.D., Wu D., Xu Z.C.: Detecting Novel Malware Classes with a Foundational Multi-Modality Data Analysis Model. *Data Intelligence* 6(4), 968–993 (2024)
- [6] Gao Y., Ai Q.Y., Song J.: Privacy Preserving Publish/Subscribe for Geo-Textual Data Streams. *Data Intelligence* 6(4), 994–1013 (2024)