Industrial large model: A survey

Jiehan Zhou[1,2*], *Yang* Cao[3*], *Quanbo* Lu[4], *Yan* Zhang[5], *Cong* Liu[6], *Shouhua* Zhang[2] and *Junsuo* Qu[7]

[1]Shandong University of Science and Technology, Qingdao, China
[2]University of Oulu, Oulu, Finland
[3]Microsoft Advanced Technology Centre, Beijing, China
[4]China University of Geosciences, Beijing, China
[5]Shandong Electric Power Research Institute, Jinan, China
[6]Shandong University of Science and Technology, Qingdao, China
[7]Xi'an University of Posts and Telecommunications, Xi'an, China

Abstract. Industrial large models are attracting significant attention for their roles in improving industrial production efficiency and product quality. This paper categorises and reviews current research on industrial large models in three main areas: pre-training, fine-tuning, and Retrieval-Augmented Generation (RAG). It also introduces a generic platform for industrial large models, including a model for interaction between industrial large and small models. Furthermore, it specifies the application areas of large industrial models within product lifecycle management, and discusses the challenges encountered during their development.

1 Introduction

The emergence of the industrial large model (ILM) marks a new era in the application of artificial intelligence in industry. With the development of AI technology, especially the progress of large language models (LLMs), industrial intelligence has entered a new development phase. These large models can not only process and understand massive amounts of industrial data but also support various complex industrial tasks, thereby driving industrial automation and intelligence. For example, large models can optimise manufacturing processes through deep learning and machine learning algorithms, improve production efficiency, reduce costs, and also enhance product quality and safety.

We define ILM as a transformer-based model used in industrial applications, especially those with a vast number of parameters capable of processing and generating multimodal data (such as text, images, and sound). Through large-scale data pre-training, these models possess strong generalisation capabilities and can handle multiple tasks without explicit guidance. Industrial large models are supposed to be used in automated design, production process optimisation, product quality control, and more.

Three primary technical strategies are employed to integrate industrial Knowledge Graphs (KGs) with Large Language Models (LLMs) to develop industrial large models: Pre-training, fine-tuning, and Retrieval-Augmented Generation (RAG). Pre-training is primarily based on training with a large amount of unlabelled data, while fine-tuning refers to further

* Corresponding authors: jiehan.zhou@ieee.org; sycao5@gmail.com

training a pre-trained model using specific datasets to adapt the model to specific tasks or domains. Third, without changing the model parameters, (RAG) can provide additional data to the models, supporting the acquisition and generation of industrial knowledge. These three techniques are not used independently but are often applied together.

This paper aims to review existing work on industrial large models, providing insights for researchers and practitioners to further their work. The remainder of the paper is organised as follows: Section 2 reviews existing work. Section 3 introduces a generic platform for industrial large models. Section 4 identifies the applications within product lifecycle management. Section 5 presents the challenges facing the development. Section 6 concludes the paper.

2 Review

2.1 Pre-trained Industrial Models

The strategic partnership between SymphonyAI [1] and Microsoft aims to address challenges in the retail sector by leveraging AI technology. Utilising Microsoft Azure OpenAI Service, SymphonyAI offers retail-specific AI applications, enabling faster and more effective decision-making for retailers worldwide. These applications, such as "Category Manager Copilot" and "Demand Planner Copilot" utilise finely tuned retail-specific language models to provide insights and optimise inventory levels, ultimately improving business performance. Both SymphonyAI and Microsoft believe that AI technology can enhance retail operations and achieve true end-to-end connectivity.

Zhang et al. [2] proposed ERNIE to augment pre-trained models with knowledge graphs. They used TAGME to extract the entity mentions in the sentences and linked them to their corresponding entities in KGs; they encoded KGs with the TransE algorithm. For information fusion, the BERT model is the backbone model. ERNIE has a T-encoder and a K-encoder. ERNIE performs better for entity typing and relation classification tasks.

Zhang et al. [3] explored the limitations of traditional fault diagnosis methods, particularly in capturing temporal features using convolutional neural networks (CNNs). They proposed a novel model tailored explicitly for fault diagnosis, combining convolutional layers with Transformer encoders to capture both local features and long-term temporal information.

2.2 Fine-tuned Industrial Models

Zhang et al. [4] introduce a plug-and-play framework that enables LLMs to induce programs in knowledge bases (KBs) with scarce resources. The KB-Plugin framework consists of two pluggable modules: the Schema Plugin, which encodes detailed schema information of a given KB using self-supervised learning, and the PI Plugin, which leverages annotated data from resource-rich KBs to help LLMs extract the relevant schema from any KB's Schema Plugin. Experiments demonstrate that KB-Plugin achieves comparable or superior performance to state-of-the-art low-resource KBs, offering a viable solution for enhancing LLM capabilities in sparse data environments.

Li et al. [5] re-examined the knowledge distillation learning paradigm for few-shot object detection tasks from the perspective of causal theory for the first time. Based on the established structural causal model, they proposed a knowledge distillation method called "Separate and Reassemble" (D&R). It conducts conditional causal interventions on the corresponding structural causal model, significantly improving performance on few-shot object detection tasks.

Song et al. [6] introduced Bayesprompt, a method for enhancing the performance of large-scale pre-trained language models on few-shot inference tasks. Bayesprompt employs de-biased domain abstraction to generate prompts, which are specialised instructions provided to the model for a particular task. These prompts help the model generalise better from limited examples by capturing essential task information while minimising the influence of biases presented in the training data.

Yuan et al. [7] explored the capabilities and limitations of foundation LLMs. It utilises category theory to examine both the limitations and potentials of foundation models. They investigated how these models can perform on a wide range of tasks without task-specific fine-tuning. They also examined various factors affecting foundation models' performance, such as dataset size, model architecture, and training objectives and proposed strategies for improving their effectiveness in different domains.

Liu et al. [8] presented a model that integrates knowledge graphs (KG) with LLMs for fault diagnosis, specifically applied in aviation assembly. This model embeds aviation assembly knowledge graph (AAKG) into LLMs for prefix fine-tuning. The joint model generates knowledge sub-graphs in test scenarios and fine-tunes the LLM's parameters through retrieval augmentation.

Liu et al. [9] proposed a novel framework, "Data-centric FinGPT," addressing data scarcity in financial LLMs. It tackles the performance gap of existing LLMs in finance due to differences in text data and the lack of open financial text datasets.

2.3 Retrieval-Augmented Generation

Jiang et al. [10] presented the KG-Agent framework to enhance the reasoning capabilities of LLMs over KGs. KG-Agent integrates LLMs, a versatile toolbox, KG-based executors, and a knowledge store, using an iterative mechanism to select tools and update memory for reasoning over KGs autonomously. Using programmatic languages, they devised multi-hop reasoning processes on KGs and synthesised code-based instruction datasets to fine-tune the underlying LLM. The results demonstrated that fine-tuning a LLaMA-7B model with only 10K samples surpasses existing methods using larger LLMs or more data, both within and outside the domain.

Luo et al. [11] proposed a CHATKBQA framework for knowledge-based question-answering tasks. This framework employs a generate-then-retrieve approach, where LLM is fine-tuned to generate candidate answers, which are then retrieved from a knowledge base. It demonstrated the effectiveness of CHATKBQA in accurately answering questions from knowledge bases, showcasing its potential for improving KBQA performance using fine-tuned LLMs.

Aiming to integrate advanced AI technologies to lower entry barriers for modelling digital twins, Zhou et al. [12] proposed a Model, AI4CDT, for facilitating accurate reflection of physical entities in digital worlds. AI4CDT combines deep learning and LLMs to provide hybrid intelligence for modelling and simulating digital twins.

Guu et al. [13] introduced a method REALM that enhances model pre-training by incorporating a retrieval mechanism. They embedded a knowledge-based document retrieval component into the language model, where the input first retrieves the most relevant batch of documents and then jointly enters the encoder to predict the language model to improve accuracy.

2.4 Others

Pan et al. [14] explored the integration of LLMs such as ChatGPT and GPT-4 with KGs. It addresses the complementary strengths of LLMs, which often operate as black-box models

that struggle to capture factual knowledge. They proposed three frameworks to leverage the synergies between LLMs and KGs: enhancing LLMs with KGs during training and inference, enhancing KG tasks with LLM capabilities such as embedding and completion, and a collaborative model where LLMs and KGs interact bi-directionally to enrich data and knowledge.

Xu et al. [15] addressed the challenge of integrating outputs from diverse LLMs during generation processes effectively and introduced a novel approach called Explicit Vocabulary Alignment (EVA); it enables fine-grained integration by learning vocabulary mappings between LLMs at each generation step. EVA resolves vocabulary disparities among LLMs, facilitating precise integration and outperforming previous methods in various tasks like commonsense reasoning and machine translation, leveraging knowledge from different models consistently, thus maximising ensemble effectiveness.

Shnitzer et al. [16] introduced a method for routing queries to LLMs based on benchmark datasets. They proposed applying benchmark datasets to train a "router" model to select the best LLM, converting the selection process into a series of binary classification tasks corresponding to each benchmark dataset. The use of benchmark datasets as reliable metrics for evaluating LLMs' efficiency and accuracy in routine tasks is highlighted, demonstrating improved performance across all tasks by employing different benchmark datasets.

Duan et al. [17] presented a solution to the scarcity of defect images in industrial settings, crucial for defect detection performance, by proposing a method for generating diverse defect images under challenging few-shot conditions. The approach involves two training stages: utilising data-efficient StyleGAN2 as the backbone network under limited defect image availability and incorporating defect-aware residual blocks to generate plausible defect masks and manipulate features within these masks. The methodology leverages StyleGAN2's efficiency and introduces defect-aware residual blocks for generating realistic defect masks and manipulating features, demonstrating effectiveness in developing realistic and diverse defect images while benefiting downstream tasks. This innovative approach addresses the scarcity of defect images in industrial settings, enhancing defect detection performance.

Fatouros et al. [18] introduced MarketSenseAI, an innovative framework that leverages GPT-4's advanced reasoning to select stocks in financial markets. By integrating "thought chains" and "contextual learning," the framework analyses diverse data sources, including market trends, news, fundamentals, and macroeconomic factors, to simulate expert investment decisions and generate actionable and interpretable signals. Notably, GPT-4 serves as both a prediction mechanism and signal evaluator, highlighting AI's transformative potential in financial decision-making.

3 Interactive evolution and platform

3.1 Interactive evolution

Figure 1 illustrates the interactive evolution of industrial large models with small models as follows:

Independent Development: In this stage, the large model platform and MaaS (Model as a Service) services develop independently. The large model platform is primarily responsible for training and deploying large-scale models, while MaaS services provide various model services for client usage.

Integration and Support: With the growth in demand, the large model platform begins to integrate with MaaS services and provide support. This may involve optimising model deployment processes, offering more efficient training environments, and integrating model monitoring and management tools.

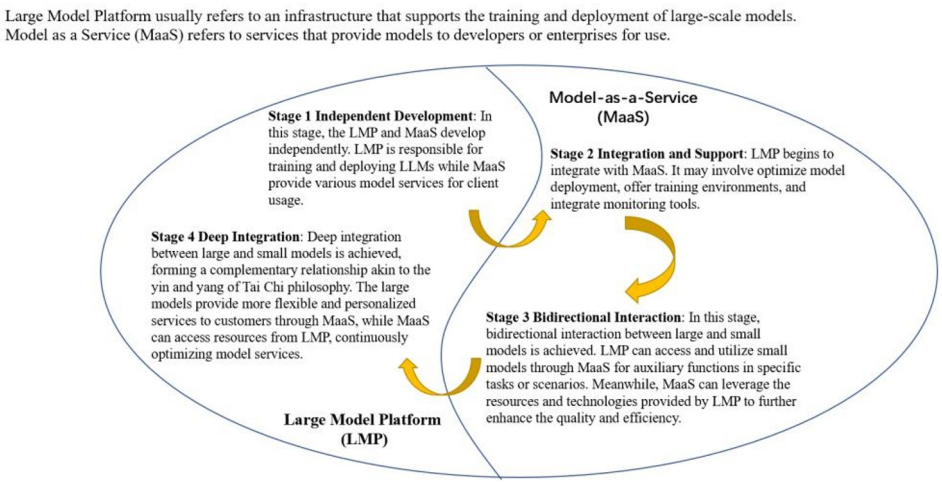

Fig. 1. The interactive evolution of industrial large models with small models.

Bi-directional Interaction: In this stage, bi-directional interaction between large and small models is achieved. The large model platform can access and utilise small models through MaaS services for auxiliary functions in specific tasks or scenarios. Meanwhile, MaaS services can also leverage the resources and technologies provided by the large model platform further to enhance the quality and efficiency of their services.

Deep Integration: In this stage, deep integration between large and small models is achieved, forming a complementary relationship. The large models provide more flexible and personalised services to customers through MaaS services, while MaaS services can access more substantial support and resources from the large model platform, continuously optimising and improving the model.

3.2 ILM platform

Figure 2 presents a platform for developing ILMs. The ILM combines general domain data and industrial data during the pre-training phase to ensure that the model possesses the "common sense" of large models and the "expertise" of industrial models.

The ILM is based on cloud-based high-performance computing clusters. The cloud platform can be public, private, or hybrid. Distributed training frameworks such as DeepSpeed and Mindspore are adopted. The base large model is based on the Transformer's Encoder-Decoder or Decoder-Only architecture model.

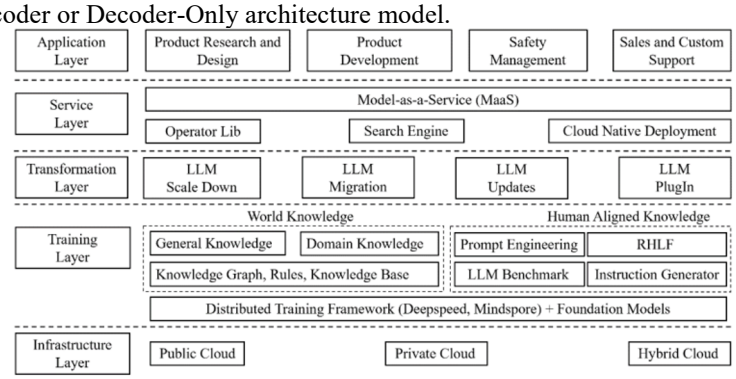

Fig. 2. A platform for industrial large models.

In acquiring domain-specific knowledge, the ILM can incorporate its industrial data and provide high-quality data through data governance and preprocessing. In addition to model training, the ILM also delves into self-developed algorithms at the model transformation and service layers, improving the inference speed and deployment performance of large models. This enables the capabilities of large models to fully integrate into the business end and create accurate models and services through integrated platform capabilities.

4 ILM-driven applications

Taking aviation engines as an example, ILMs can be employed in the product research and development, design, simulation, production, testing, operation, and after-sales phases as follows:

Research and Development phase. It utilises ILMs for data analysis and modelling to identify potential technical challenges and propose solutions. The knowledge base and information retrieval function provided by ILMs accelerate researchers' understanding and in-depth learning in relevant fields. The phase generates innovative ideas and research directions and conducts preliminary verification.

Design phase. It carries out detailed design work based on the suggestions and plans generated by ILMs, utilises ILMs for design optimisation and validation to enhance design efficiency and quality, and combines principles of human engineering with ILMs to generate designs that meet ergonomic requirements.

Simulation phase. It generates simulation models using ILMs for thermodynamics and structural strength analyses, guides design improvements and optimisation processes based on the simulation results produced by ILMs, and interprets simulation results and provides further suggestions.

Production phase. It enhances production efficiency and quality by utilising manufacturing process plans and fixture designs generated by ILMs, collaborates with suppliers to ensure that component production meets standard requirements using the technical documents and process flows generated by ILMs, and ensures that components produced meet design requirements by implementing quality inspection plans and guidelines generated by ILMs.

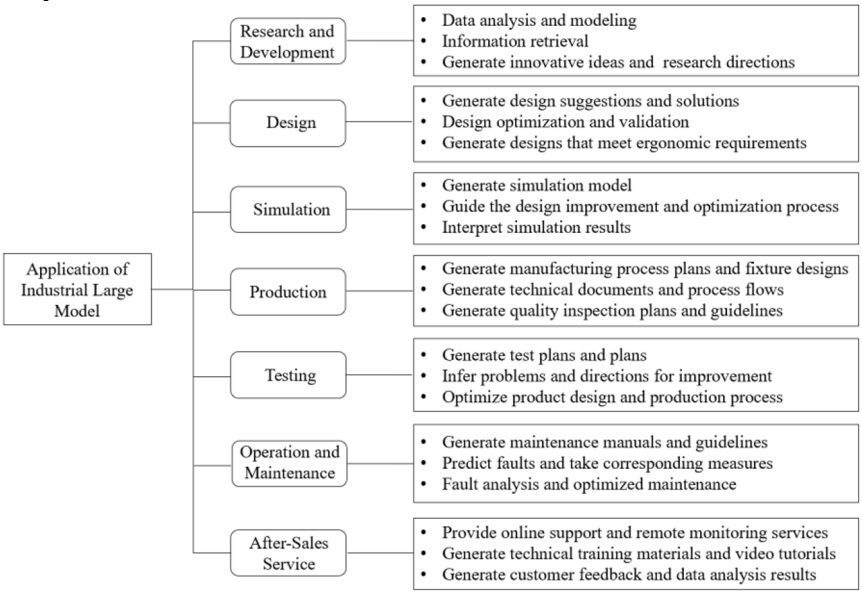

Fig. 3. Applications of industrial large models.

Testing phase. It conducts ground stand tests and flight tests based on the test plans and proposals generated by ILMs, analyses test data and feedback, infers issues and improvement directions using the natural language processing and data analysis functions of ILMs, and optimises product design and production processes by incorporating reports and summaries generated by ILMs.

Operation and Maintenance phase. It assists operations and maintenance personnel in correctly maintaining engines by utilising maintenance guides generated by ILMs, predicts faults in real-time by monitoring equipment status and using ILMs, and conducts fault analysis and optimises maintenance strategies.

After-Sales Service phase. It offers online support and remote monitoring services to address customer issues using ILMs, helps better utilise products by providing technical training materials and video tutorials generated by ILMs, and continuously improves products and services.

5 Challenges

The development of ILMs faces many practical challenges, summarised as follows:

• Cost: The financial and manpower investment required to implement ILMs is substantial, making it difficult for companies to afford.

• Interpretability: Current ILMs lack strong interpretability, making it hard for them to explain their inference results. This limits their direct application in decision-making.

• Real-time: ILMs do not meet the requirements for real-time computation and updates, which are in demand by high-speed manufacturing processes.

• Sample quantity and quality: The performance of ILMs heavily depends on the amount and quality of data, but industries generally face difficulties in acquiring high-quality samples.

• Generalisation: ILMs face highly diverse scenarios and needs, with significant task variability, requiring redeployment for different industries/products/processes.

• Intellectual property: The security and intellectual property issues related to the training data must be addressed.

6 Conclusion

Industrial large models have significant potential to enhance production efficiency and improve product quality across various industrial sectors. This paper overviews the existing work from three aspects: pre-training, fine-tuning, and Retrieval-Augmented Generation (RAG). Additionally, we introduce a generic platform for developing industrial large models, including a model for interaction between large and small models. Furthermore, we identify specific application areas and outline the predominant challenges faced during their development.

References

1. SymphonyAI. *"Industrial LLM - SymphonyAI with Microsoft".* https://www.symphonyai.com/ , accessed April 29, 2024 (2024)
2. Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun and Q. Liu, *"ERNIE: Enhanced language representation with informative entities"*, arXiv preprint arXiv:1905.07129, (2019)
3. S. Zhang, J. Zhou, X. Ma, C. Wen, S. Pirttikangas, C. Yu and C. Yang, *"TSViT: A Time Series Vision Transformer for Fault Diagnosis"*, arXiv preprint arXiv:2311.06916, (2023)

4. J. Zhang, S. Cao, L. Hu, L. Feng, L. Hou and J. Li, *"KB-Plugin: A Plug-and-play Framework for Large Language Models to Induce Programs over Low-resourced Knowledge Bases"*, arXiv preprint arXiv:2402.01619, (2024)
5. J. Li, Y. Zhang, W. Qiang, L. Si, C. Jiao, X. Hu and F. Sun, *"Disentangle and remerge: interventional knowledge distillation for few-shot object detection from a conditional causal perspective"*, in proceedings of the AAAI Conference on Artificial Intelligence, **37** (1), pp. 1323-1333, (2023)
6. J. Li, F. Song, Y. Jin, W. Qiang, C. Zheng, F. Sun, and H. Xiong, *"BayesPrompt: Prompting Large-Scale Pre-Trained Language Models on Few-shot Inference via Debiased Domain Abstraction"*, arXiv preprint arXiv:2401.14166, (2024)
7. Y. Yuan, *"On the power of foundation models"*, arXiv:2211.16327, (2023)
8. P. Liu, L. Qian, X. Zhao, B. Tao, IEEE T. Indust. Inform, **20** (6) pp. 1-10, (2024)
9. X. Liu, G. Wang, H. Yang, D. Zha, *"Data-centric FinGPT: Democratising internet-scale data for financial large language models"*, arXiv:2307.10485v2, (2023)
10. J. Jiang, K. Zhou, W. Zhao, Y. Song, C. Zhu, H. Zhu and J. Wen*,"KG-Agent: An Efficient Autonomous Agent Framework for Complex Reasoning over Knowledge Graph"*, arXiv preprint arXiv:2402.11163, (2024)
11. H. Luo, Z. Tang, S. Peng, Y. Guo, W. Zhang, C. Ma and W. Lin, *"Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models"*, arXiv preprint arXiv:2310.08975, (2023)
12. J. Zhou, Q. Lu, X. Chai, C. Liu, W. Shen, *"A Data-Driven and Knowledge Graph Enhanced Intelligent Framework for Modeling Cognitive Digital Twins"*, in Int. Conf. on SMC (2024)
13. K. Guu, K. Lee, Z. Tung, P. Pasupatand, M. Chang, *"Retrieval augmented language model pre-training",* in international conference on machine learning, pp. 3929-3938, PMLR, (2020)
14. S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang and X. Wu, IEEE Trans. Knowl. Data Eng, **36** (7), 2024.
15. Y. Xu, J. Lu and J. Zhang, *"Bridging the Gap between Different Vocabularies for LLM Ensemble"*, arXiv preprint arXiv:2404.09492 (2024)
16. T. Shnitzer, A. Ou, M. Silva, K. Soule, Y. Sun, J. Solomon and M. Yurochkin, *"Large language model routing with benchmark datasets"*, arXiv preprint arXiv:2309.15789, (2023)
17. Y. Duan, Y. Hong, L. Niu and L. Zhang, *"Few-shot defect image generation via defect-aware feature manipulation"*,in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 1, pp. 571-578, (2023)
18. G. Fatouros, K. Metaxas, J. Soldats and D. Kyriazis, *"Can large language models beat wall street? unveiling the potential of ai in stock selection"*, arXiv preprint arXiv:2401.03737 (2024)