Paper 96-3 has been designated as a Distinguished Paper at Display Week 2025. The full-length version of this paper appears in a Special Section of the *Journal of the Society for Information Display (JSID)* devoted to Display Week 2025 Distinguished Papers. This Special Section will be freely accessible until December 31, 2025 via:

https://sid.onlinelibrary.wiley.com/doi/full/10.1002/jsid.2064

Authors that wish to refer to this work are advised to cite the full-length version by referring to its DOI:

https://doi.org/10.1002/jsid.2064

Developing Large Language Models for Display Industrial Knowledge: Data Augmentation, Training Techniques, and Evaluation Strategies

Bingqian Wang, Lixin Wang, Qingqing Sun, Yulan Hu, Yuyu Liu, Xingqun Jiang
BOE Technology Group Co., Ltd., Beijing, China

Abstract
Large Language Models (LLMs) hold numerous potential applications within the display industry. However, mainstream LLMs generally lack the domain-specific knowledge. Especially, in display knowledge question-answering(Q&A) scenarios, the lack of understanding of specialized terminology leads to low accuracy in responses. Consequently, this paper proposes a development framework for a Display Industry Knowledge Large Language Model (DIK-LLM), aimed at enhancing the model's comprehension of semiconductor display industry field through specialized data governance, knowledge distillation, data augmentation strategies, and continual pre-training mechanisms. This approach not only significantly improves the model's performance in Q&A applications within the display industry but also prevents catastrophic forgetting. It is hoped that these contents will provide guidance and reference for researchers and practitioners in the customization of LLM for specialized domains.

Author Keywords
Continual Pre-training; Industrial Knowledge Large Language Model; Data Synthesis; Model Fine-tuning; Model Evaluation

1. Introduction
Large language models (LLMs) have achieved remarkable advancements in natural language processing (NLP), improving performance across a variety of tasks [1]. These models, trained on expansive open-web datasets, have evolved into versatile general-purpose LLMs. Consequently, when confronted with specialized scenarios such as industry-specific knowledge questioning and report generation, their performance often falls short of professional standards. This limitation primarily stems from the general-purpose LLMs' lack of in-depth understanding and specialized training in domain-specific knowledge.

Currently, numerous studies have focused on enhancing the expertise of general LLMs in domain-specific scenarios, with notable examples including the Med-PaLM for the medical field [2], ChatLaw for legal [3], and BloombergGPT for finance [4]. However, to date, no dedicated research has been conducted for the display industry. The implementation and development of LLMs in the display industry face a series of challenges and limitations, primarily manifested as follows: 1). Insufficient understanding of industry knowledge: The inability to accurately comprehend industry field and professional concepts leads to the occurrence of hallucination phenomena; 2). Data Scarcity: The display industry is highly specialized, with most data housed within companies and only a small portion available in patents, papers, and reports, posing challenges for assembling training datasets; 3). Lack of professional evaluation benchmarks: Currently, there are no evaluation datasets that focus on knowledge specific to the display industry; 4). Balancing professional and general knowledge: Enhancing domain-specific expertise while retaining the model's general capabilities in low-resource scenarios presents a significant challenge. This study proposes a construction method for a LLM tailored to the display industry. The method aims to enhance the model's comprehension of professional terminology and improve its application performance in industrial question-and-answer systems through specialized data governance, knowledge distillation, data augmentation strategies, and continual pre-training mechanisms.

Our main contributions are summarized as follows:

- We propose a data processing and synthesis scheme tailored for specialized domains, effectively addressing the issues of poor training data quality and insufficient data volume.
- We validate the effectiveness of continual pre-training in constructing domain-specific LLMs, providing a replicable case for the development of models within the same industry.
- We establish an evaluation benchmark for LLMs in the display domain, filling the gap in assessment standards for this field.

These contributions significantly advance NLP in the display industry and provide valuable insights for related research areas.

2. Display Industry Corpora: Data Collection and Cleaning
Data Collection: Data in the display industry is complex in terms of sources and formats, with primary data coming from journals, papers, books, training materials, and production line documents. To prevent catastrophic forgetting, we have also incorporated open-source general domain data.

- Papers: We have collected papers from conferences and journals related to the display industry, such as The Society for Information Display (SID).
- Patents: We have gathered publicly available patent information related to Liquid Crystal Display (LCD) and Organic Electroluminescence Display (OLED) yield analysis, process flow, and material development.
- Books: We have collected professional books related to LCD and OLED.
- Industrial production data: Training materials, work instruction manuals, and equipment operation and maintenance manuals.
- Open-source general data: WuDaoCorpora [5] and other open datasets.

Data Processing: These data must be converted into text data through document processing techniques such as Optical Character Recognition(OCR), layout analysis, and formula recognition. Subsequently, they undergo data cleaning and filtering steps to transform them into relatively clean text data. The general data processing pipeline is depicted in Figure 1.

ISSN 0097-996X/25/5602-1334-$1.00 © 2025

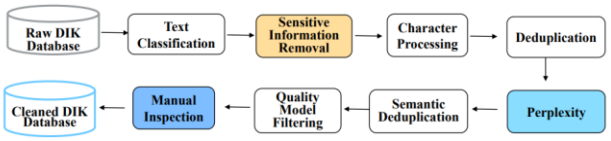

Figure 1. Data Processing Workflow

The main processing steps include text classification, sensitive information filtering and deduplication (e.g., document-based, perplexity-based and semantic deduplication). Finally, manual sampling verifies threshold settings, ensuring data quality for effective model training.

3. Methodology

Our method of building display industry-specific LLM consists of two stages: domain specific continual pre-training(CPT) and supervised instruction tuning.

Continual Pre-training: The display industry data is highly specialized, with certain knowledge points exhibiting a long-tail distribution characteristic. To further enrich the diversity of data and increase the overall data volume, this study proposes a knowledge and information synthesis and filtering method targeted at display industry data, aiming to effectively enhance the diversity and expansion of specialized knowledge and data scale. To ensure the authenticity and accuracy of the data, the paper synthesizes information based on display industry knowledge. By employing diverse, multi-level and multi-dimensional prompt designs, we guided LLMs to integrate both rich and scarce knowledge. This approach helps the model learn a broader range of information and mitigates the knowledge silo effect, with some prompt examples shown in Table 1.

Table 1. Prompts Examples for Synthetic Data Generation

Function	Prompt Examples
Glossary	Please extract all technical terms from the article excerpt, provide a brief explanation and practical application examples for each term, ensuring that readers have a clear understanding of the terminology.
Summary	Please compose an abstract based on the article excerpt that accurately reflects the core ideas and themes of the article.
Q&As	Please extract three sets of professional interactive Q&As from the article, ensuring that there is a certain logic between the questions. Make sure that both the questions and answers are derived from this article.

Leveraging multiple large language models as domain experts, we synthesize various types of data based on specific instructions. Concurrently, to ensure data quality, the synthesized data undergoes a selection process. Initially, the synthesized data is assessed to determine if it pertains to the display industry domain, with non-relevant data being eliminated. Subsequently, domain-specific keywords are incorporated for filtering, and the semantic relevance of the synthesized data to the original articles is evaluated, with weakly related content being discarded. Additionally, the stylistic consistency of the generated data with the original articles is assessed, along with the semantic similarity threshold; data failing to meet these criteria are removed. Ultimately, this process yields high-quality and diverse domain data, with the specific steps illustrated in Figure 2.

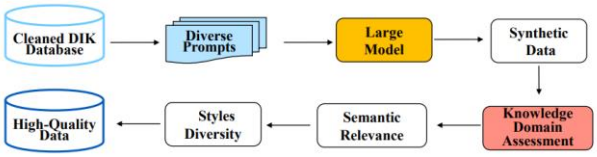

Figure 2. Data Synthesis Workflow

We trained our foundational model by fine-tuning open-source LLMs with domain-specific datasets, resulting in a model enriched with industry-specific knowledge. This method conserves resources by minimizing the requirement for extensive computational resources and generic datasets. To prevent catastrophic forgetting during incremental training, we blended general open data at a 5:1 ratio with domain-specific data. Throughout the CPT phase, the model concentrated on acquiring new professional knowledge while concurrently reviewing general knowledge.

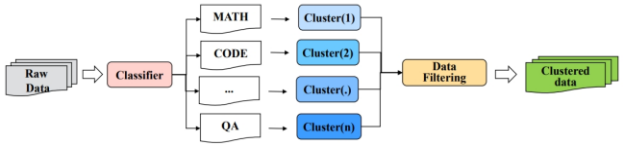

Figure 3. SFT Dataset Construction Process

Domain-Specific Instruct Tuning: During (domain-specific) pre-training, LLMs learn and acquire general knowledge but they might not be great at chatting with users. Therefore, pre-trained LLMs should be further instruct tuned to follow instructions to interact effectively with users and other LLMs through Supervised Fine-Tuning (SFT). During the phase, it is essential to incorporate instruction data from both the general domain and the specific display industry domain. Data from the general domain is intended to stimulate the large model's performance in general tasks, whereas the SFT for the display industry domain is designed to enhance the model's knowledge-based Q&A capabilities within that sector. The methodology for constructing the SFT dataset is illustrated in Figure 3.
The primary steps involved are as follows:

- Step 1: Categorize the open-source SFT data into types such as question-and-answer, code, generation, and logical reasoning.
- Step 2: Employ clustering methods, such as K-means or DBSCAN, to segment each category of data into distinct clusters, as illustrated in Figure 3.
- Step 3: For each cluster of instruction pairs, utilize the base model under incremental training to generate answers. Subsequently, have a larger LLMs evaluate the compliance with the instructions and assign a score. This results in a set of instruction pairs along with their corresponding compliance scores.
- Step 4: Dynamically configure the quantity of instruction data based on requirements. For instance, to enhance the model's Q&A capabilities, if a total of 10,000 SFT data points are needed across four categories (Q&A, code, generation, and logical reasoning), one could follow a ratio of 7:1:1:1.

For selecting of vertical domain SFT data, both model synthesis and manual screening methods can be applied. The synthetic approach involves generating instruction pairs from high-

knowledge-density texts or paragraphs using large models based on existing content. In contrast, the manual approach involves inviting domain experts to construct Q&A pairs to form the instruction set.

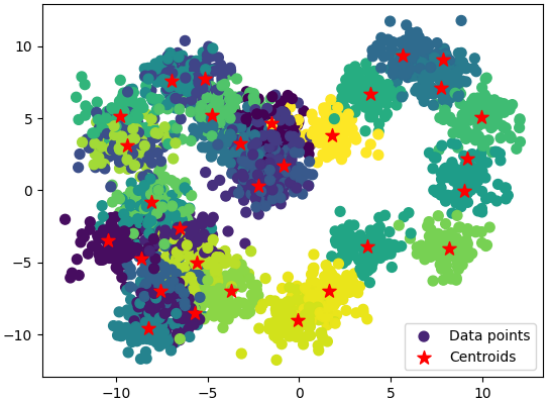

Figure 4. Example of SFT Data Clustering Results

Evaluation Benchmark Construction: The display industry lacks an authoritative evaluation benchmark. To address this gap, we have developed a domain-specific evaluation benchmark by extracting and validating data from publicly accessible specialized examinations, including undergraduate, postgraduate, and professional qualification tests in the field of semiconductor display.

Table 2. The statistics of Evaluation Dataset

Category	DIK-Eval	DA-Eval
Basic	13	20
Professional	43	20
Advanced	44	10
Total	100	50

Table 3. Examples of Evaluation Dataset

Question Type	Sample Question
Single Choice (Basic Knowledge)	In luminescent display devices, the one with the lowest power consumption is: A. Plasma Display Panel (PDP) B. Cathode Ray Tube (CRT) C. Field Emission Display (FED) D. Electron Luminescent Display (ELD) Answer: D
Multiple Choice (Yield Analysis)	Which of the following options belong to foreign particle defects? A. PT Gap B. Cell Particle C. Clumped Zara D. Striped Zara Answer: ABCD
True/False (Manufacturing Knowledge)	If alternating current power is used to drive LCD panels, it is likely to cause chemical reactions in the liquid crystal material, leading to a reduction in lifespan. A. True B. False Answer: B

This benchmark encompasses 150 questions, divided into two sets: the Display Industry Knowledge Evaluation (DIK-Eval) and the Defect Analysis Evaluation (DA-Eval). The DIK-Eval spans various scenarios, including foundational knowledge, research and development design, production management, manufacturing, and processes. The DA-Eval focuses on the understanding of defective terminology, phenomena, causes, and improvement measures, with a particular emphasis on professional knowledge assessment. This evaluation set has been reviewed and validated by industry experts. The questions are divided into three difficulty levels: basic, professional, and advanced, reflecting an increasing level of challenge. The question types primarily consist of single choice, multiple-choice and true/false formats. Table 2 summarizes the distribution of difficulty levels in the evaluation benchmark we proposed, while Table 3 presents some sample test data.

4. Experiments and Results

Experimental Setup: We selected Qwen2-7B [6] and Yi-1.5-34B [7] as our base models and performed CPT and instruction fine-tuning using the constructed dataset for display industry (6B tokens) and an instruction dataset (120,000 samples). Ultimately, we obtained display industry knowledge large language models (DIK-LLMs), including DIK-7B and DIK-34B. All experiments were conducted on two servers equipped with 8 NVIDIA A800 GPUs each. To accelerate model training and iteration, we utilized the Megatron-LM training framework. Detailed training parameter configurations are shown in Table 4.

Table 4. Training Hyper Parameter Settings

Hyper parameter	CPT	SFT
Precision	bf16	bf16
Epochs	5	3
Global Batch size	64	64
Learning rate	7e-6	5e-6
LR scheduler type	cosine	cosine

Performance on Industrial Evaluation Datasets: To validate the model's effectiveness, we evaluated its performance on a display industry-specific benchmark using a 5-shot setting within the OpenCompass [9] evaluation framework. As shown in Table 5, our model demonstrated a significant improvement over the current SOTA model, GPT-4o, in the display industry domain. Specifically, DIK-7B achieved an 11% improvement on the DIK-Eval benchmark (63% *vs.* 74%) and a 20% improvement on the DA-Eval benchmark (62% *vs.* 82%). Similarly, DIK-34B demonstrated a 15% improvement on the DIK-Eval benchmark (63% *vs.* 78%) and a 26% improvement on the DA-Eval benchmark (62% *vs.* 88%). These results confirm the effectiveness of the proposed model development framework.

Table 5. Results on Industrial Evaluation Datasets

Model	DIK-Eval	DA-Eval
GPT-4o	63	62
Qwen2-7B	52	48
Yi-1.5-34B	60	56
DIK-7B	74(+11)	82(+20)
DIK-34B	78(+15)	88(+26)

Performance on General Evaluation Datasets: To assess whether the model's original general capabilities were retained after incremental training, we evaluated its performance on benchmarks such as C-Eval, CMMLU, MMLU, and MATH before and after incremental training, shown in Table 6. The experimental results indicate that our model's general capabilities did not suffer any significant loss. Compared to the original

model Qwen2-7B, DIK-7B showed only a minor accuracy decrease of 0.37% on the MMLU benchmark (69.17% *vs.* 68.8%). For DIK-34B, there were slight decreases on the MMLU (76.05% *vs.* 76.02%), and MATH (53.42% *vs.* 47.68%) benchmarks. However, the model demonstrated varying degrees of improvement on other benchmarks, including C-Eval and CMMLU.

Table 6. Results on General Evaluation Datasets

Model	C-Eval	CMMLU	MMLU	MATH
Qwen2 -7B	79.05	81.57	69.17	49.34
Yi-1.5 -34B	82.1	82.39	76.05	53.42
DIK -7B	81.28 **+2.23**	81.64 **+0.07**	68.8 **-0.37**	51.06 **+1.72**
DIK -34B	83.51 **+1.14**	83.00 **+0.61**	76.02 **-0.03**	47.68 **-5.56**

5. Applications

LLMs offer numerous applications in the display industry, including knowledge-based Q&A, equipment maintenance, employee training, and yield analysis. We deployed a display industry-specific yield knowledge Q&A system for internal factory trias, showing improved accuracy in specialized queries and reduced hallucination. Additionally, we compared the performance of general LLMs and DIK-LLMs in the RAG system. Utilizing the RAGAS [8] framework, we assessed the overall performance with a focus on three key metrics: Faithfulness (derived from context), Correctness (accuracy against the actual answers), and Similarity (semantic similarity to human-annotated responses).

Table 7. Results on Q&A System

Model	Faithfulness	Correctness	Similarity	Average
Qwen2-7B	0.8902	0.5266	0.6240	0.678
DIK-7B	0.9219	0.7087	0.7113	0.781
Yi-1.5-34B	0.8767	0.6043	0.6324	0.704
DIK-34B	0.9623	0.7530	0.6872	0.8011

As shown in Table 7, our model demonstrated substantial improvements over the original model across these three metrics in defect-related Q&A scenarios. Specifically, DIK-7B achieved a 15.1% improvement (0.678 *vs.* 0.781), while DIK-34B showed a 9.7% improvement (0.704 *vs.* 0.801). These results indicate that using a display industry-specific large language model enhances the effectiveness of the entire Q&A system in real-world applications.

6. Conclusion

This research introduces a technical framework for developing LLMs specialized in the display industry, focusing on data processing, training strategies, and evaluation protocols. Experiments on a display industry knowledge dataset and Q&A applications confirm that our approach significantly enhances LLM performance on industry-specific tasks while maintaining general domain capabilities at a lower cost. This foundational work paves the way for the development of specialized LLMs in the display industry, with the potential to broaden their application across related sectors.

7. References

1. 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. Advances in neural information processing systems. 2020; 33:1877-1901.
2. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature. 2023; 620:172-80.
3. J. Cui, Z. Li, Y. Yan, B. Chen, and L. Yuan, "Chatlaw: Open-source legal large language model with integrated external knowledge bases," arXiv preprint. 2023; arXiv:2306.16092.
4. Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance. CoRR.2023; abs/2303.17564.
5. Yuan S, Zhao H, Du Z, et al. WuDaoCorpora: A Super Large-scale Chinese Corpora for Pre-training Language Models. AI Open, 2021;2;65-68.
6. An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang. et al. Qwen technical report. arXiv preprint 2024; arXiv:2407.10671.
7. Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang. et al. Yi: Open Foundation Models by 01.AI. arXiv preprint 2024;arXiv:2403.04652.
8. Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated Evaluation of Retrieval Augmented Generation. 2023; arXiv:2309.15217.
9. https://github.com/open-compass/OpenCompass/