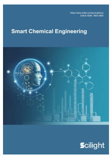
Smart Chemical Engineering
https://www.sciltp.com/journals/sce



Article

Domain-Specific Fine-tuning of Large Language Models and Intelligent Question-Answering System for Industrial Catalysis

Shican Wu [1,2,†], Xin Chang [1,2,†], Xiao Ma [1,2], Xiaoyun Lin [1,2], Ran Zhao [1,2] and Zhi-Jian Zhao [1,2,*]

[1] Key Laboratory of Green Synthesis and Transformations, Ministry of Education, School of Chemical Engineering and Technology, Tianjin University, Tianjin 300072, China
[2] International Joint Laboratory of Low-Carbon Chemical Engineering, Ministry of Education, Tianjin 300350, China
* Correspondence: zjzhao@tju.edu.cn
† These authors contributed equally to this work.

How To Cite: Wu, S.; Chang, X.; Ma, X.; et al. Domain-Specific Fine-tuning of Large Language Models and Intelligent Question-Answering System for Industrial Catalysis. *Smart Chemical Engineering* **2026**, *2*(1), 2. https://doi.org/10.53941/sce.2026.100002

Received: 30 December 2025
Revised: 23 February 2026
Accepted: 5 March 2026
Published: 16 March 2026

Abstract: Industrial catalysis, as a core field of chemical engineering, is characterized by intensive professional terminology and complex knowledge structures, making it challenging for general-purpose large language models to accurately understand and apply relevant professional knowledge. This research presents a domain-specific fine-tuning technique and retrieval-augmented generation system for the industrial catalysis field. Through a multi-model collaborative data processing pipeline, we construct high-quality training corpora, employ parameter-efficient fine-tuning techniques to train specialized domain models, and design a retrieval-augmented generation workflow based on consistency verification. The research first establishes a training corpus containing 2.3 billion tokens, including 1.1 billion domain-specific tokens and 1.2 billion general tokens with a balanced 1:1 ratio strategy. Subsequently, we apply rank-stabilized low-rank adaptation (rsLoRA) method to perform parameter-efficient fine-tuning on the Yi-1.5-6B model, resulting in the PeiYang Micro-Emergence model, which achieves a score of 76.81 in industrial catalysis field evaluation, significantly outperforming the general-purpose model Qwen2.5-72B-Instruct (65.45 points) with 12 times the parameters, while maintaining good general capabilities. We further construct a 3.37 million domain-specific retrieval pair dataset and optimize the embedding model using Matryoshka representation learning (MRL) techniques, achieving an average improvement of 2.87 percentage points in domain retrieval recall@3 while slightly enhancing general capabilities. Finally, we design a professional retrieval-augmented generation workflow integrating bilingual hypothetical document generation, dual-path retrieval, and consistency verification, achieving high-quality professional knowledge services. This system provides accurate and reliable professional knowledge services for the industrial catalysis field, demonstrates the application value of domain-specific large language models in resource-constrained environments, and offers a replicable technical pathway for artificial intelligence applications in other specialized domains.

Keywords: industrial catalysis; large language models; domain-specific fine-tuning; retrieval-augmented generation; intelligent question-answering

1. Introduction

Industrial catalysis, as the cornerstone of modern industry, significantly accelerates chemical reactions, reduces energy consumption, and improves selectivity. It not only serves as critical support for core industries such as petrochemicals, pharmaceuticals, and energy but also acts as an important driving force for promoting green chemistry and sustainable development [1,2]. Catalysts play a central role in over 90% of chemical processes and most industrial product manufacturing [3,4]. Facing major challenges such as global climate change, energy crises, and resource shortages, industrial catalysis provides fundamental solutions for constructing efficient and clean sustainable industrial systems through its unique advantages [5].

However, industrial catalysis, as a core research field in chemical engineering and materials science, is characterized by knowledge intensity, numerous professional terms, and strong interdisciplinary nature [6]. With the exponential growth of the amount of scientific literatures [7], researchers face the challenge of information overload, urgently requiring efficient knowledge acquisition and processing tools. In recent years, large language models (LLMs), as breakthrough progress in natural language processing [8], have provided new possibilities for knowledge retrieval and professional consultation in the industrial catalysis field [9–11].

Nevertheless, general-purpose large language models have significant limitations in professional domain applications. General models are primarily trained on general corpora and have insufficient mastery of highly specialized fields such as industrial catalysis, manifesting as limited understanding of professional terminology, insufficient depth of domain knowledge, and inadequate professional reasoning capabilities [12]. Secondly, general models are large in scale, facing challenges such as high computational resource requirements and slow inference speeds during actual deployment [13], making it difficult to meet the practical application needs of research institutions. Finally, knowledge in professional fields updates rapidly, and static pre-trained models struggle to keep pace with the latest research developments [14].

In the industrial catalysis field, existing research has explored applications of machine learning and artificial intelligence technologies, including catalyst design [15], reaction condition optimization [16], and mechanism prediction [17]. Although existing general scientific artificial intelligence assistants can handle some basic questions, they often perform poorly when facing complex catalysis professional questions [9,12,18]. Recent advances in retrieval-augmented generation (RAG) methodologies have shown promise in chemical and materials sciences [19], providing frameworks for integrating domain knowledge with language model capabilities.

Addressing these challenges, domain-specific model fine-tuning [20] and retrieval-augmented generation [21] technologies provide effective solutions for professional domain applications. This research constructs a complete large language model solution for the industrial catalysis field, designing a multi-model collaborative corpus construction process to achieve full-chain data processing from domain literature to training corpus synthesis. On this basis, we build a large-scale training corpus library containing 1.2 billion general tokens and 1.1 billion domain-specific tokens, and fine-tune the Yi-1.5-6B model [22] using rank-stabilized low-rank adaptation (rsLoRA) method [23] to construct the PeiYang Micro-Emergence model, which achieves 76.81 points in the propane dehydrogenation field evaluation, significantly exceeding the general model Qwen2.5-72B-Instruct [24] (65.45 points) with 12 times the parameters, while effectively maintaining general capabilities.

The system further designs a catalysis domain retrieval pair synthesis workflow, constructs a large-scale dataset containing 3.37 million domain retrieval pairs, optimizes the bge-m3 embedding model [25] using Matryoshka representation learning (MRL) method [26], achieving significant improvement in domain retrieval performance while maintaining general capabilities. The finally constructed professional retrieval-augmented generation workflow combines bilingual hypothetical document generation [27], dual-path retrieval, and consistency verification mechanisms, effectively solving core problems such as insufficient professional query expression and answer reliability. This system provides efficient, accurate, and reliable intelligent support for industrial catalysis research.

2. Materials and Methods

2.1. Domain Training Corpus Construction

This research systematically collected over 50,000 industrial catalysis-related publications from professional academic websites and databases based on Selenium [28], strictly adhering to academic usage guidelines for text mining purposes. Literature selection criteria mainly focused on core areas of computational catalysis and thermal catalysis, including catalytic process simulation, catalytic material design, and catalytic reaction mechanisms. The data cleaning stage included format matching based on regular expressions, and document-level precise deduplication based on literature DOI.

To adapt to bilingual training needs, we designed a complete large language model translation process. This process uses prompt engineering to guide large language models for translation, optimized for the Chinese conversion needs of large amounts of English content in academic literature. The translation process includes document structure recognition, paragraph segmentation, segmented translation, and post-processing steps, using a sliding window mechanism to ensure context coherence and rule truncation for repeated outputs. Both original literature and translated literature serve as domain-specific continuous pre-training corpora. This bilingual strategy is designed to serve the Chinese industrial catalysis research community, our primary target users, while leveraging cross-lingual knowledge transfer [29] through shared semantic representations.

Domain question-answer pair generation technology adopts a knowledge extraction method based on large language models, constructing question-answer pairs from industrial catalysis literature. The synthesis process first conducts in-depth analysis of literature content, identifying key knowledge points, core concepts, and important conclusions, then designs multi-level question types, including factual inquiries, concept explanations, principle elaborations, method comparisons, and application reasoning, covering multiple cognitive levels from basic to advanced. To enhance training data diversity, we implemented systematic question generation strategies, designing various variations from perspectives of question types, cognitive difficulty, and target audiences.

Multi-turn logical dialogue construction adopts a role-playing-based large language model synthesis method. We designed multiple role settings such as researchers, review experts, research supervisors, technical experts, senior researchers, and domain experts, with each role generating dialogues based on preset professional backgrounds and evaluation frameworks. Through large language model logical thread analysis of original literature, identifying core arguments, supporting evidence, research methods, and other key logical elements, then designing dialogue progressive development paths based on extracted logical threads, ensuring dialogues gradually deepen from surface questions to core issues, thereby constructing multi-turn logical dialogues. Additionally, logical threads are integrated into conference record-like text through large language models, which can serve as continuous pre-training corpora.

The domain evaluation set adopts an automated synthesis method based on large language models, extracting key knowledge points from collected propane dehydrogenation literature and generating corresponding multiple-choice questions. To ensure evaluation objectivity, all questions undergo multiple rounds of large language model verification [30] for quality control, forming a comprehensive evaluation system containing 95,206 multiple-choice questions.

General pre-training corpus screening selects from large-scale general datasets such as redpajama-refine [31] for cleaning processing. Through entity extraction from aforementioned literature via large language model, followed by large language model annotation and manual verification, we determined 3768 high-quality domain-specific terms, forming the core of the domain vocabulary (as shown in Supplementary Materials Figure S1). Based on the professional vocabulary, we established screening mechanisms, setting different professional term density requirements for different content types, effectively improving the relevance of general corpora to the domain. General supervised fine-tuning corpora use open-source datasets such as Infinity-Instruct [32] for simple cleaning and format alignment.

2.2. Parameter-Efficient Fine-tuning Technology

After systematic comparison and evaluation, we selected the Yi-1.5-6B pre-trained model [22], which demonstrates competitive performance and strong bilingual capabilities among models of comparable scale (see Supplementary Materials Figure S2 for detailed benchmark comparison), as the fine-tuning foundation. This model possesses good Chinese-English processing capabilities and can effectively handle scenarios containing large amounts of translated content. Considering sufficient data volume, we adopt the pre-trained base model as the starting point for domain fine-tuning.

This research employs parameter-efficient fine-tuning technology for domain adaptation, selecting the rsLoRA (rank-stabilized Low-Rank Adaptation) [23] method based on the need to maximize rank values. This method effectively solves high-rank gradient collapse problems by correcting adapter scaling factors, maintaining adapter stability across different rank values. In specific implementation, the adaptation matrix adopts rank value 64 and scaling coefficient 128, applying low-rank adaptation to all linear layers and embedding layers in the model, requiring updates to only 2.41% of the base model parameters, greatly reducing computational resource requirements and training costs.

For the adaptation matrix, we use principal singular values and singular vectors adaptation (PISSA) [33] initialization method, initializing adapter matrices through singular value decomposition of pre-trained model weight matrices using principal singular values and singular vectors. We select the AdamW [34] optimizer, with learning rate scheduling adopting a multi-stage strategy including linear warmup, containing warmup (1×10^{-8} to 1×10^{-4}), stable (1×10^{-4}), and cosine decay (1×10^{-4} to 1×10^{-8}) three stages [35].

To suppress catastrophic forgetting phenomena, we merge continuous pre-training and supervised fine-tuning into joint training [36]. Literature and general text from continuous pre-training inject domain knowledge into the model, while question-answer pairs from supervised fine-tuning endow the model with dialogue response capabilities. By mixing both parts of corpora for training, we reduce problems of excessive original parameter drift caused by multiple adapter applications.

The training process was conducted on a hundred-card cluster of 14 machines with 8 cards each, adopting distributed data parallel strategy [37], achieving efficient communication through InfiniBand network. Training uses MS-Swift [38] as the core framework for distributed training, ultimately fine-tuning to obtain the PeiYang Micro-Emergence model, with performance evaluation based on the OpenCompass [39] evaluation framework. The overall technical architecture diagram of the PeiYang Micro-Emergence model is shown in Figure 1.

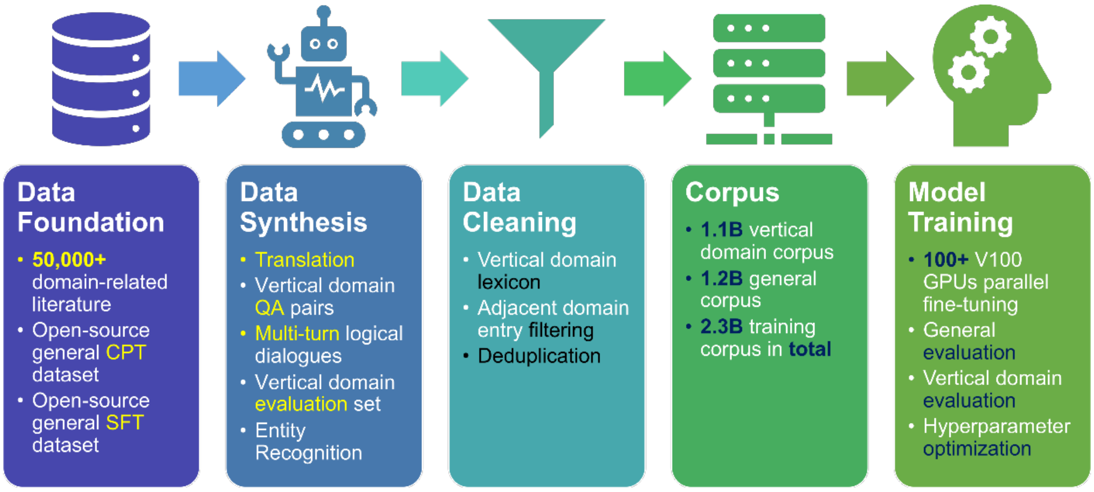

Data Foundation
• 50,000+ domain-related literature
• Open-source general CPT dataset
• Open-source general SFT dataset
Data Synthesis
• Translation
• Vertical domain QA pairs
• Multi-turn logical dialogues
• Vertical domain evaluation set
• Entity Recognition
Data Cleaning
• Vertical domain lexicon
• Adjacent domain entry filtering
• Deduplication
Corpus
• 1.1B vertical domain corpus
• 1.2B general corpus
• 2.3B training corpus in total
Model Training
• 100+ V100 GPUs parallel fine-tuning
• General evaluation
• Vertical domain evaluation
• Hyperparameter optimization

Figure 1. Overall technical architecture diagram of PeiYang Micro-Emergence model.

2.3. Intelligent Question-Answering System Design

The intelligent question-answering system is overall based on customized development and implementation of Langchain-ChatChat2.0, an open-source framework for building retrieval-augmented generation systems. We designed a professional workflow including hypothetical document generation, bilingual dual-path retrieval, and consistency verification (as shown in Figure 2a). Early query processing includes query rewriting, query splitting, and bilingual query translation, where each query is translated into both Chinese and English to produce parallel bilingual query versions, optimized for the characteristics of professional catalysis queries.

We used hypothetical document generation technology [27], using the PeiYang Micro-Emergence model to provide preliminary answers in both languages, generating bilingual hypothetical documents. The original queries and their translations are then combined with the corresponding hypothetical documents for actual retrieval, solving the problem of insufficient professional query expression while ensuring comprehensive coverage across the bilingual knowledge base.

The retrieval mechanism operates in parallel across both languages: for each language version, vector-based semantic search and BM25-based lexical matching are executed simultaneously, resulting in four parallel retrieval paths. All retrieved results from the four paths are then unified through score normalization, duplicate removal via hash identification, and weighted re-ranking, producing the final ranked candidate set that integrates evidence from both languages and both retrieval methods.

Consistency verification serves as a confidence detection mechanism, assessing answer reliability by comparing initially generated answers with answers generated after retrieval augmentation. High consistency indicates high model confidence on the topic, while significant divergence signals uncertainty warranting additional verification. In specific implementation, consistency is compared through the PeiYang Micro-Emergence model. When significant differences are detected, specifically when the consistency score is below 7/10, the system initiates a supplementary retrieval round. Following the HyDE approach [27], the original query maintains primary importance while keywords extracted from the enhanced answer serve as supplementary context, preserving the original query intent while expanding retrieval coverage. The retrieved additional literature is then used to regenerate a more robust answer. This verification mechanism operates with a maximum of two iterations to balance answer quality and response latency. When the consistency score remains below the threshold after reaching the iteration limit, the system outputs the literature-supported answer with an explicit uncertainty indicator, ensuring every query receives a response. This approach is analogous to hypothesis verification in scientific research, seeking literature support after constructing initial answers.

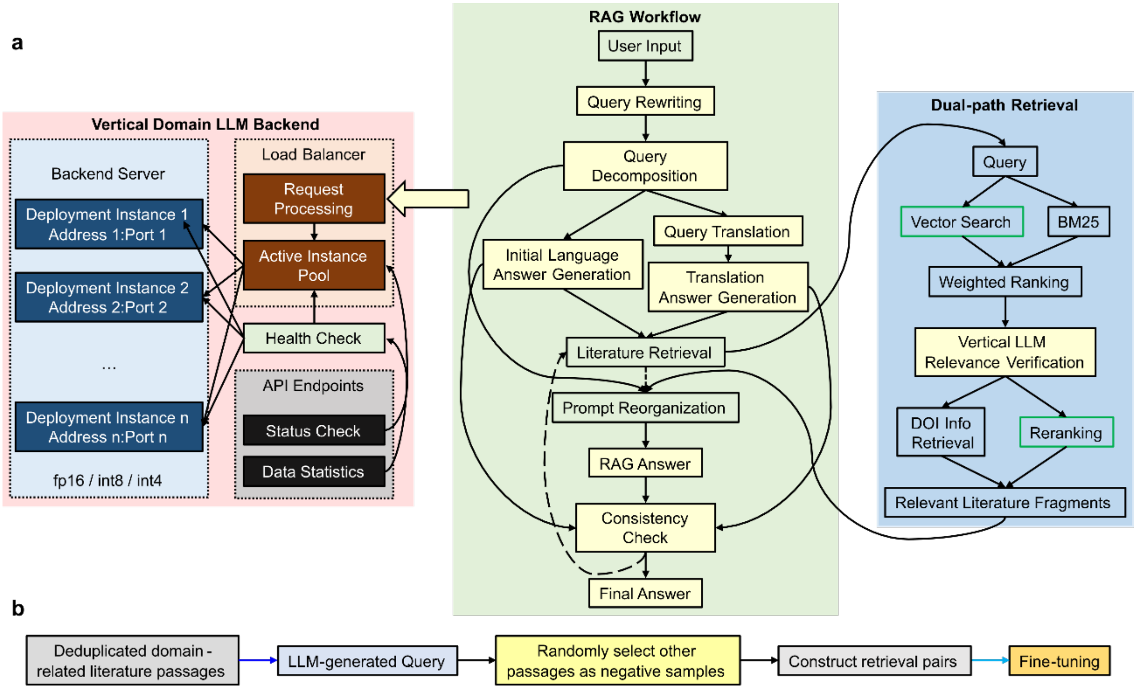


Figure 2. (**a**) Overall design architecture of the intelligent question-answering system based on PeiYang Micro-Emergence, consisting of three main components: (i) domain-specific large model backend; (ii) retrieval augmented generation workflow; (iii) dual-path retrieval and recall module, with components marked in yellow processed by the load-balanced domain-specific large model backend. (**b**) Domain-specific retrieval pair construction workflow.

To improve vector retrieval effectiveness, targeting the special needs of the industrial catalysis professional field, we constructed a systematic catalysis field retrieval dataset for domain fine-tuning of the embedding model. We extracted text paragraphs from the industrial catalysis literature library, used large language models to generate corresponding queries for each text paragraph, constructed retrieval training data pairs, generating 5 queries from different angles in both Chinese and English for each literature paragraph, ensuring dataset diversity and comprehensiveness. The constructed training data pairs include positive and negative examples, using generated queries with original text paragraphs as positive examples, and randomly selecting other paragraphs as negative examples. We finally constructed a catalysis field retrieval dataset containing 2.69 million training retrieval pairs and 680 k test retrieval pairs, totaling 3.37 million retrieval pairs, covering three main domains: propane dehydrogenation, computational catalysis, and thermal catalysis (as shown in Figure 2b). Based on this dataset, we selected the bge-m3 embedding model [25] for fine-tuning, adopting Matryoshka representation learning (MRL) method [26] to achieve multi-granularity representation optimization.

3. Results and Discussion

3.1. Domain Fine-Tuning Effect Evaluation

Through systematic corpus construction and distribution optimization, we finally constructed a high-quality training corpus library totaling 2.3 billion tokens. The corpus is divided by training stage into continuous pre-training corpus (68.75%) and supervised fine-tuning corpus (31.25%), with domain-specific and general content approaching a 1:1 balance. The language distribution presents characteristics of balanced Chinese-English bilingualism, with Chinese content accounting for approximately 50.64%, English content approximately 34.11%, and bilingual mixed content approximately 15.25% (as shown in Figure 3). This balanced corpus allocation strategy achieved good results, significantly improving training efficiency.

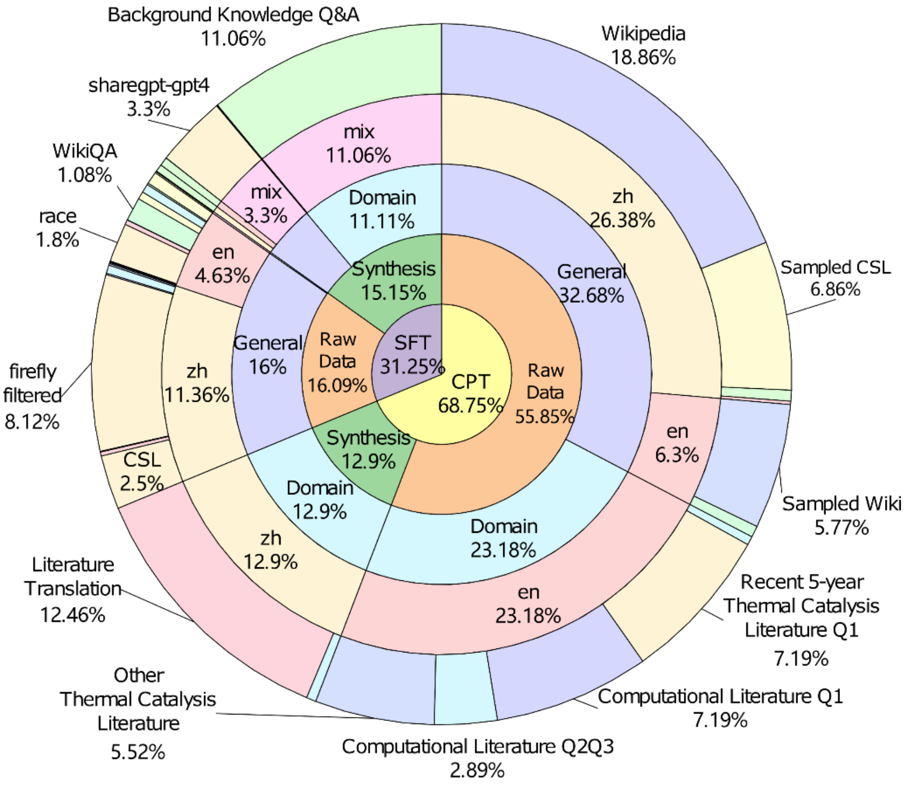


Figure 3. Sunburst chart of training corpus distribution, from inner to outer layers including: (i) continuous pre-training (CPT) and supervised fine-tuning (SFT) data distribution; (ii) instruction source distribution; (iii) general and domain-specific data distribution; (iv) language distribution; (v) dataset source distribution. Zh: Chinese. En: English.

General capability evaluation through 39 standard test sets such as SimpleQA [40,41], RACE [42], LCSTS [43], MMLU [44,45], C-Eval [46], and GSM8K [47] shows that the PeiYang Micro-Emergence model exhibits significant deviation from general models with the same base in capabilities (see Table S1 for detailed benchmark). The domain model performs outstandingly in logical reasoning and language understanding, related to training reinforcement of causal reasoning understanding in the industrial catalysis field. We note that this improvement in causal reasoning represents a secondary outcome of domain training; the decline in mathematical reasoning (GSM8K: 37.45 vs. 76.04) reflects a known trade-off in domain-specific fine-tuning, as industrial catalysis literature emphasizes conceptual reasoning but lacks structured mathematical problem-solving content. Although there are certain gaps in common sense memory and mathematical reasoning, the model still maintains basic general capability levels, effectively suppressing catastrophic forgetting phenomena (as shown in Figures 4a and S3 for detailed benchmark results). Furthermore, the decline in mathematical reasoning reflects a fundamental limitation of current LLM architectures in numerical tokenization rather than a deficiency of domain-specific training. In practical deployment, precise numerical calculations can be reliably delegated to external computational tools through function calling mechanisms.

The PeiYang Micro-Emergence model after domain fine-tuning performs excellently in industrial catalysis field evaluation. Domain evaluation based on 95,206 multiple-choice questions in propane dehydrogenation shows that the fine-tuned model achieves 76.81 points out of 100, significantly exceeding the base model's 27.96 points and the chat model's 63.91 points. To contextualize the domain fine-tuning effectiveness, we benchmarked against Qwen2.5-72B-Instruct [24], selected as the comparison model for its state-of-the-art performance across major benchmarks (MMLU, GSM8K, HumanEval), strong bilingual capabilities essential for our Chinese-English catalysis domain, and full open-source availability enabling reproducible evaluation. Despite having 12 times more parameters, Qwen2.5-72B-Instruct achieves only 65.45 points on our domain evaluation, significantly lower than PeiYang's 76.81 points, demonstrating that domain-specific fine-tuning can enable small models to surpass large general-purpose models in specialized fields (as shown in Figure 4b, see Figures S4 and S5 for qualitative comparison of responses before and after fine-tuning).

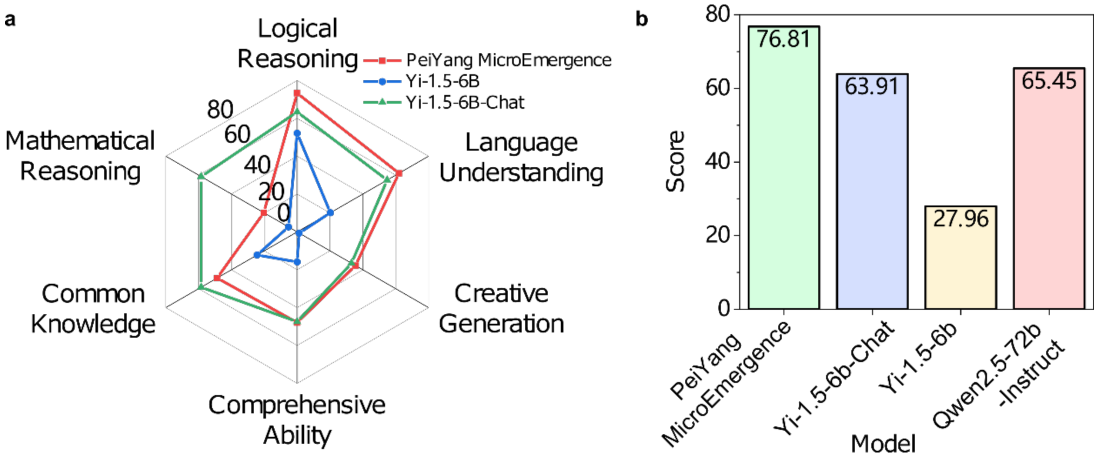


Figure 4. (**a**) Radar chart of general capability evaluation results, with red line representing PeiYang Micro-Emergence model, blue line representing the fine-tuned base Yi-1.5-6B model, and green line representing the baseline Yi-1.5-6B-Chat model. (**b**) Bar chart of evaluation results on domain-specific evaluation datasets.

3.2. Retrieval-Augmented System Performance Evaluation

The domain embedding model optimization effect is significant. We constructed a specialized test set covering multiple aspects such as catalytic materials, reaction mechanisms, process conditions, and application fields, with balanced distribution of Chinese and English documents. The test set design fully considers actual needs of industrial catalysis research, covering different levels of content from basic theory to engineering applications. This comprehensive and balanced test set design ensures objectivity and representativeness of model evaluation, providing a reliable foundation for domain embedding model performance verification (as shown in Figure 5).

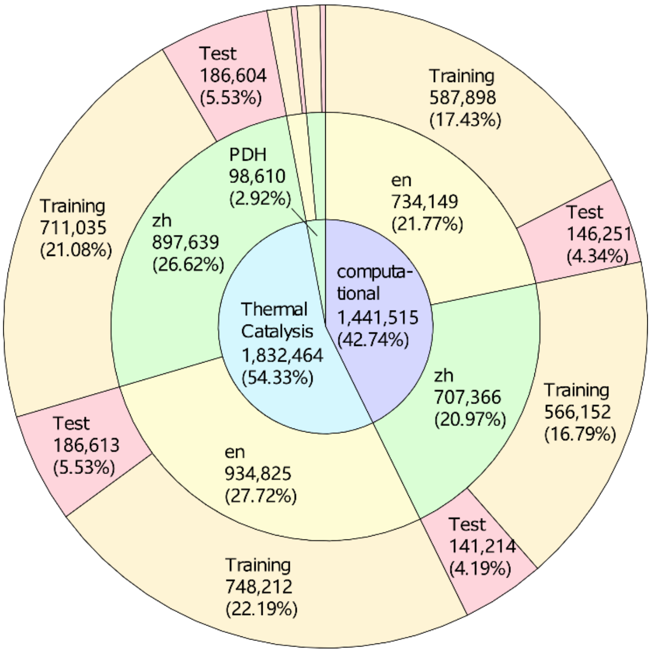


Figure 5. Sunburst chart of domain-specific retrieval dataset distribution, from inner to outer layers including: (i) source literature distribution; (ii) language distribution; (iii) usage distribution.

Comprehensive evaluation results show that the fine-tuned model significantly outperforms the original pre-trained model in catalysis professional field retrieval tasks. The fine-tuned model outperforms the baseline model across all datasets, achieving an average recall@3 (defined as the proportion of queries for which the correct answer is retrieved within the top-k results) of 55.55% on propane dehydrogenation literature, a 3.3% improvement over the baseline's 52.25%; the average recall@10 reaches 66.65% compared to the baseline's 63.95%, representing a 2.70 percentage point improvement. The deployed system uses recall@10 from both vector and BM25 retrieval paths, followed by weighted score fusion and re-ranking, with the top-4 results presented to users;

while maintaining a high recall@3 of 83.5% on general documents, indicating that the model enhances domain-specific capabilities while maintaining general capabilities. For all catalysis field retrieval task datasets, the fine-tuned model's recall@3 improved by an average of 2.87% relative to the baseline model (as shown in Figure 6, see Table S2 for detailed benchmark).

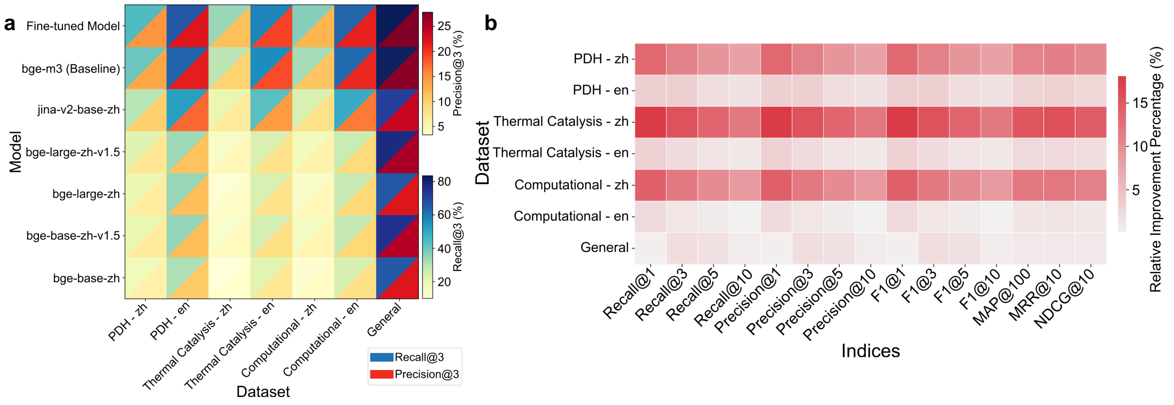


Figure 6. (**a**) Precision@3 and Recall@3 of different models on various datasets, with upper triangles representing recall and lower triangles representing precision. (**b**) Heatmap showing performance improvement of fine-tuned models relative to the bge-m3 baseline. See Table S2 for detailed benchmark.

System performance evaluation results indicate that the technical architecture design achieves expected goals across multiple dimensions. The dual-path retrieval mechanism effectively addresses the complexity and diversity of catalysis field queries by fusing advantages of semantic understanding and exact matching. When processing queries containing precise information such as chemical formulas and reaction conditions, literal matching retrieval can provide accurate term correspondence; while when processing conceptual and mechanistic questions, vector retrieval demonstrates good semantic understanding capabilities.

The application of the consistency verification mechanism further enhances system answer reliability. Through multi-path result comparison, the system can identify and correct potential inconsistencies. Considering that erroneous information in catalysis research may lead to experimental failures or safety hazards, this quality assurance mechanism is particularly important for professional field applications. Additionally, the literature recommendation function composed of retrieved literature achieves effective connection from question-answering to in-depth research, providing researchers with a complete knowledge acquisition path. A practical generation example of the overall workflow is presented in Figure S6.

3.3. System Applicability Analysis

The PeiYang Micro-Emergence model performs best in propane dehydrogenation catalysis research and also demonstrates the ability to understand related concepts in broader thermal catalysis and computational catalysis fields. However, in other catalysis directions or related technical fields without specialized training, the system's professionalism and accuracy may significantly decline. This applicability boundary stems from the coverage scope of training corpora and differences in professional knowledge depth.

The designed technical architecture possesses good universality and scalability, applicable not only to the industrial catalysis field but also transferable to other professional fields. The system adopts a highly modular structure, with functional modules connected through standardized interfaces, allowing flexible adjustment according to target field characteristics. Core technical solutions such as domain model construction, dual-path retrieval fusion, and consistency verification mechanisms address problems common to professional fields, characteristics that universally exist in multiple professional fields such as medicine, law, and materials science. When application to new fields is needed, rapid adaptation can be achieved by supplementing relevant field literature and adopting the same corpus construction, model fine-tuning, and retrieval system construction processes. When stronger general capabilities are needed, the system can also directly interface with other general large language models.

From a broader perspective, this research demonstrates an effective path for constructing professional field artificial intelligence systems, namely that through carefully designed data processing processes, targeted model optimization, and innovative system architecture, high-quality professional services can be achieved under limited resource conditions. This small but specialized development model provides new ideas for deep application of artificial intelligence technology in professional fields, helping to promote the transformation of artificial intelligence technology from generality to specialization, better serving actual needs of scientific research and engineering practice.

4. Conclusions and Outlook

This paper addresses key challenges in large language model applications in the industrial catalysis field, including insufficient professional knowledge understanding, high computational resource requirements, and unreliable answer results, constructing a complete domain-specific fine-tuning and retrieval-augmented generation solution. Through systematic technological innovation, we successfully achieved performance where small professional models exceed large general models in specific fields.

In data construction, the research designed a multi-model collaborative corpus construction process, achieving systematic conversion from domain literature to high-quality training data, including key elements such as literature translation, domain question-answer pair generation and multi-turn logical dialogue synthesis, and domain evaluation set construction. Through establishing complete data cleaning and aggregation mechanisms, we constructed a large-scale training corpus library containing 1.2 billion general tokens and 1.1 billion domain tokens, totaling 2.3 billion tokens.

In model optimization, based on rsLoRA technology, we performed parameter-efficient fine-tuning on the Yi-1.5-6B model. The obtained PeiYang Micro-Emergence model achieved 76.81 points in the propane dehydrogenation field, significantly exceeding the 65.45 points of the general model Qwen2.5-72B-Instruct with 12 times the parameters, while effectively maintaining general capabilities, verifying the feasibility of small models exceeding large general models in specific professional fields through targeted training. The PeiYang Micro-Emergence model has been open-sourced on Hugging Face at: https://huggingface.co/Invalid-Null/PeiYangMe (accessed on 1 January 2025).

In retrieval augmentation technology, we constructed a 3.37 million retrieval pair dataset and fine-tuned the bge-m3 embedding model using Matryoshka representation learning technology, achieving an average improvement of 2.87 percentage points in domain retrieval recall@3 while maintaining general capabilities. The professional retrieval-augmented generation workflow effectively solves core problems such as insufficient professional query expression and answer reliability through integrating bilingual hypothetical document generation, dual-path retrieval, and consistency verification mechanisms. We note that the current domain evaluation is based on questions synthesized from the same literature collection used for training. While the questions are newly generated formulations rather than direct excerpts, this introduces potential knowledge overlap. Future work should incorporate independently curated test sets and temporal-split evaluation to more rigorously assess generalization capabilities.

Although the research has made significant progress, some limitations still exist. Current evaluation is mainly based on multiple-choice question format, with relatively limited assessment of model innovative thinking and complex reasoning capabilities; domain corpora, although carefully screened and constructed, still have room for improvement in coverage and timeliness; the system currently focuses on text modality and has not fully utilized multi-modal information specific to professional fields such as chemical structure diagrams and reaction mechanism diagrams.

Future research will deepen technological development in several key directions: studying differentiated fine-tuning methods for different catalysis subfields to expand professional coverage; developing dynamic knowledge update mechanisms to timely integrate latest research progress; exploring multi-modal information fusion technology to integrate professional field visual information; constructing structured catalysis field knowledge graphs, deeply integrating with retrieval augmentation systems to provide more systematic and interpretable professional knowledge services.

The technical solution proposed in this research provides a systematic solution for large language model applications in the industrial catalysis field. Through innovative architectural design and implementation methods, it effectively solves key challenges in professional field knowledge services, demonstrating that small-parameter models can achieve performance comparable to much larger models in specific domains. Research results not only provide strong support for knowledge acquisition and innovation for research and developments in the catalysis field but also offer replicable technical pathways for artificial intelligence applications in other professional fields, demonstrating the enormous potential of domain-specific large language models in achieving balance between high professionalism and high resource efficiency, contributing to promoting deep application of artificial intelligence technology in professional fields.

Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/2603161343267240/SCE-25120214-SM.pdf. Figure S1: (a) Domain-specific vocabulary construction process (b) Sunburst chart showing proportion of redpajama-refine dataset after domain-specific vocabulary filtering, from inner to outer rings: (i) dataset subsets; (ii) language distribution; (iii) domain entity match distribution. Figure S2: Benchmark comparison of Yi-1.5-6B and Yi-1.5-9B base models against contemporaneous open-source LLMs on English and Chinese benchmarks (as of 12 May 2024). Data source: https://huggingface.co/01-ai/Yi-1.5-6B (accessed on 12 May 2024). Figure S3: Radar chart of evaluation results on general benchmark datasets, with green line representing PeiYang Micro-Emergence model, red line representing fine-tuned base Yi-1.5-6B model, and blue line representing baseline Yi-1.5-6B-Chat model. Figure S4: Question-answering responses from Yi-1.5-6B-Chat model before domain fine-tuning, showing incorrect interpretation of domain-specific terminology (PDH misidentified as 'Partial Differential Hydrogen'). Figure S5: Question-answering responses from PeiYang Micro-Emergence model after domain fine-tuning, demonstrating accurate domain knowledge with correct interpretation of PDH (propane dehydrogenation) and detailed catalyst information. Figure S6: Example responses from the domain-specific retrieval-augmented generation system. Table S1: Comparative evaluation results on general benchmark datasets. Table S2: Extended retrieval performance metrics.

Author Contributions

S.W.: conceptualization, methodology, software, data curation, writing—original draft preparation; X.C.: conceptualization, methodology, software, data curation, writing—original draft preparation; X.M.: software, validation; X.L.: visualization, investigation; R.Z.: data curation, validation; Z.-J.Z.: conceptualization, supervision, funding acquisition, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Natural Science Foundation of China (No. U22A20409, 22121004), the Energy Revolution S&T Program of the Yulin Innovation Institute of Clean Energy (Grant No. E511040817), the Emerging Frontiers Cultivation Program of the Tianjin University Interdisciplinary Center, the Haihe Laboratory of Sustainable Chemical Transformations, the Program of Introducing Talents of Discipline to Universities (BP0618007), and the XPLORER PRIZE. We also thank the High Performance Computing Center of Tianjin University for providing generous computing resources.

Data Availability Statement

The PeiYang Micro-Emergence model has been open-sourced on Hugging Face at: https://huggingface.co/Invalid-Null/PeiYangMe (accessed on 1 January 2025). The training corpora, evaluation datasets and retrieval pair datasets are not publicly available due to copyright restrictions on the source literature and proprietary considerations.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Claude (Anthropic) to assist with manuscript language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Zheng, R.; Liu, Z.; Wang, Y.; et al. Industrial Catalysis: Strategies to Enhance Selectivity. *Chin. J. Catal.* **2020**, *41*, 1032–1038.
2. Wang, Y.; Tian, Y.; Pan, S.Y.; et al. Catalytic Processes to Accelerate Decarbonization in a Net-Zero Carbon World. *ChemSusChem* **2022**, *15*, e202201290.
3. Nørskov, J.K.; Bligaard, T.; Rossmeisl, J.; et al. Towards the Computational Design of Solid Catalysts. *Nat. Chem.* **2009**, *1*, 37–46.
4. Ludwig, J.R.; Schindler, C.S. Catalyst: Sustainable Catalysis. *Chem* **2017**, *2*, 313–316.

5. Abbas, A.; Cross, M.; Duan, X.; et al. Catalysis at the Intersection of Sustainable Chemistry and a Circular Economy. *One Earth* **2024**, *7*, 738–741.
6. Wang, Y.; Shi, J.; Jin, Z.; et al. Focus on the Chinese Revolution of Catalysis Based on Catalytic Solutions for the Vital Demands of Society and Economy. *Chin. J. Catal.* **2018**, *39*, 1147–1156.
7. Bornmann, L.; Haunschild, R.; Mutz, R. Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases. *Humanit. Soc. Sci. Commun.* **2021**, *8*, 224.
8. Wei, J.; Tay, Y.; Bommasani, R.; et al. Emergent Abilities of Large Language Models. *arXiv* **2022**, arXiv:2206.07682.
9. Wang, L.; Chen, X.; Du, Y.; et al. CataLM: Empowering Catalyst Design through Large Language Models. *Int. J. Mach. Learn. Cybern.* **2025**, *16*, 3681–3691.
10. Chen, X.; Gao, Y.; Wang, L.; et al. Large Language Model Enhanced Corpus of CO_2 Reduction Electrocatalysts and Synthesis Procedures. *Sci. Data* **2024**, *11*, 347.
11. Su, Y.; Wang, X.; Ye, Y.; et al. Automation and Machine Learning Augmented by Large Language Models in a Catalysis Study. *Chem. Sci.* **2024**, *15*, 12200–12233.
12. Song, Z.; Yan, B.; Liu, Y.; et al. Injecting Domain-Specific Knowledge into Large Language Models: A Comprehensive Survey. *arXiv* **2025**, arXiv:2502.10708.
13. Bai, G.; Chai, Z.; Ling, C.; et al. Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models. *arXiv* **2024**, arXiv:2401.00625.
14. Wang, M.; Stoll, A.; Lange, L.; et al. Bring Your Own Knowledge: A Survey of Methods for LLM Knowledge Expansion. *arXiv* **2025**, arXiv:2502.12598.
15. Benavides-Hernández, J.; Dumeignil, F. From Characterization to Discovery: Artificial Intelligence, Machine Learning and High-Throughput Experiments for Heterogeneous Catalyst Design. *ACS Catal.* **2024**, *14*, 11749–11779.
16. Li, A.; Cui, P.; Wang, X.; et al. The Artificial Intelligence-Catalyst Pipeline: Accelerating Catalyst Innovation from Laboratory to Industry. *Front. Chem. Sci. Eng.* **2025**, *19*, 55.
17. Tan, Z.; Yang, Q.; Luo, S. AI Molecular Catalysis: Where Are We Now? *Org. Chem. Front.* **2025**, *12*, 2759–2776.
18. Bran, A.M.; Cox, S.; Schilter, O.; et al. Augmenting Large Language Models with Chemistry Tools. *Nat. Mach. Intell.* **2024**, *6*, 525–535.
19. Chattoraj, J.; Hamadicharef, B.; Chang, T.S.; et al. AceWGS: An LLM-Aided Framework to Accelerate Catalyst Design for Water-Gas Shift Reactions. *arXiv* **2025**, arXiv:2503.05607.
20. Lu, W.; Luu, R.K.; Buehler, M.J. Fine-Tuning Large Language Models for Domain Adaptation: Exploration of Training Strategies, Scaling, Model Merging and Synergistic Capabilities. *NPJ Comput. Mater.* **2025**, *11*, 84.
21. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
22. Young, A.; Chen, B.; Li, C.; et al. Yi: Open Foundation Models by 01.AI. *arXiv* **2024**, arXiv:2403.04652.
23. Kalajdzievski, D. A Rank Stabilization Scaling Factor for Fine-Tuning with LoRA. *arXiv* **2023**, arXiv:2312.03732.
24. Yang, A.; Yang, B.; Zhang, B.; et al. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115.
25. Chen, J.; Xiao, S.; Zhang, P.; et al. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings through Self-Knowledge Distillation. *arXiv* **2024**, arXiv:2402.03216.
26. Kusupati, A.; Bhatt, G.; Rege, A.; et al. Matryoshka Representation Learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 30233–30249.
27. Gao, L.; Ma, X.; Lin, J.; et al. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, ON, Canada, 9–14July 2023.
28. Gojare, S.; Joshi, R.; Gaigaware, D. Analysis and Design of Selenium WebDriver Automation Testing Framework. *Procedia Comput. Sci.* **2015**, *50*, 341–346.
29. Conneau, A.; Khandelwal, K.; Goyal, N.; et al. Unsupervised Cross-Lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic*s*, online, 5–10 July 2020; pp. 8440–8451.
30. Zheng, L.; Chiang, W.L.; Sheng, Y.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 46595–4662.
31. Weber, M.; Fu, D.; Anthony, Q.; et al. RedPajama: An Open Dataset for Training Large Language Models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 116462–116492.
32. Li, J.; Du, L.; Zhao, H.; et al. Infinity Instruct: Scaling Instruction Selection and Synthesis to Enhance Language Models. *arXiv* **2025**, arXiv:2506.11116.
33. Meng, F.; Wang, Z.; Zhang, M. PiSSA: Principal Singular Values and Singular Vectors Adaptation of Large Language Models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 121038–121072.
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.

35. Wen, K.; Li, Z.; Wang, J.; et al. Understanding Warmup-Stable-Decay Learning Rates: A River Valley Loss Landscape Perspective. *arXiv* **2024**, arXiv:2410.05192.
36. Wen, C.; Sun, X.; Zhao, S.; et al. ChatHome: Development and Evaluation of a Domain-Specific Language Model for Home Renovation. *arXiv* **2023**, arXiv:2307.15290.
37. Li, S.; Zhao, Y.; Varma, R.; et al. PyTorch Distributed: Experiences on Accelerating Data Parallel Training. *arXiv* **2020**, arXiv:2006.15704.
38. Zhao, Y.; Huang, J.; Hu, J.; et al. Swift: A Scalable Lightweight Infrastructure for Fine-Tuning. *arXiv* **2024**, arXiv:2408.05517.
39. Contributors, O. OpenCompass: A Universal Evaluation Platform for Foundation Models. Available online: https://github.com/open-compass/opencompass (accessed on 4 March 2025).
40. He, Y.; Li, S.; Liu, J.; et al. Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models. *arXiv* **2024**, arXiv:2411.07140.
41. Wei, J.; Karina, N.; Chung, H.W.; et al. Measuring Short-Form Factuality in Large Language Models. *arXiv* **2024**, arXiv:2411.04368.
42. Lai, G.; Xie, Q.; Liu, H.; et al. RACE: Large-Scale Reading Comprehension Dataset from Examinations. *arXiv* **2017**, arXiv:1704.04683.
43. Hu, B.; Chen, Q.; Zhu, F. LCSTS: A Large Scale Chinese Short Text Summarization Dataset. *arXiv* **2015**, arXiv:1506.05865.
44. Hendrycks, D.; Burns, C.; Basart, S.; et al. Measuring Massive Multitask Language Understanding. *arXiv* **2020**, arXiv:2009.03300.
45. Li, H.; Zhang, Y.; Koto, F.; et al. CMMLU: Measuring Massive Multitask Language Understanding in Chinese. *arXiv* **2023**, arXiv:2306.09212.
46. Huang, Y.; Bai, Y.; Zhu, Z.; et al. C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 62991–63010.
47. Cobbe, K.; Kosaraju, V.; Bavarian, M.; et al. Training Verifiers to Solve Math Word Problems. *arXiv* **2021**, arXiv:2110.14168.