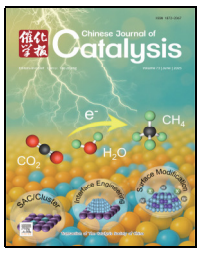
Chinese Journal of Catalysis 73 (2025) 159–173

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.sciencedirect.com/journal/chinese-journal-of-catalysis



Article

From lab to fab: A large language model for chemical engineering

Jibin Zhou [a,1], Feiyang Xu [b,1], Zhijun Chang [c], Duiping Liu [a], Lulu Li [a], Jian Cui [b], Yi Li [b], Xin Li [b,d,e,*], Li Qian [c], Zhixiong Zhang [c], Guoping Hu [b,e], Mao Ye [a,*], Zhongmin Liu [a]

[a] *National Engineering Research Center of Lower-Carbon Catalysis Technology, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China*

[b] *Artificial Intelligence Research Institute, iFLYTEK Co., Ltd., Hefei 230000, Anhui, China*

[c] *National Science Library, Chinese Academy of Sciences, Beijing 100190, China*

[d] *University of Science and Technology of China, Hefei 230000, Anhui, China*

[e] *State Key Laboratory of Cognitive Intelligence, Hefei 230000, Anhui, China*

ARTICLE INFO

Article history:
Received 27 March 2025
Accepted 13 May 2025
Available online 20 June 2025

Keywords:
Large language model
Chemical engineering
Process development
Multidimensional benchmark
Domain adaptation

ABSTRACT

The development of chemical technologies, which involves a multistage process covering laboratory research, scale-up to industrial deployment, and necessitates interdisciplinary collaboration, is often accompanied by substantial time and economic costs. To address these challenges, in this work, we report ChemELLM, a domain-specific large language model (LLM) with 70 billion parameters for chemical engineering. ChemELLM demonstrates state-of-the-art performance across critical tasks ranging from foundational understanding to professional problem-solving. It outperforms mainstream LLMs (e.g., O1-Preview, GPT-4o, and DeepSeek-R1) on ChemEBench, the first multidimensional benchmark for chemical engineering, which encompasses 15 dimensions across 101 distinct essential tasks. To support robust model development, we curated ChemEData, a purpose-built dataset containing 19 billion tokens for pre-training and 1 billion tokens for fine-tuning. This work establishes a new paradigm for artificial intelligence-driven innovation, bridging the gap between laboratory-scale innovation and industrial-scale implementation, thus accelerating technological advancement in chemical engineering. ChemELLM is publicly available at https://chemindustry.iflytek.com/chat.

1. Introduction

The development of chemical technologies is a multi-stage process that typically begins with laboratory research, progresses through scale-up and basic engineering, and culminates in industrial deployment [1,2]. This complex process demands synergistic collaboration from experts with diverse disciplinary backgrounds, such as chemistry, physics, mathematics, electrical engineering, process design, and architecture, to address technical bottlenecks while balancing economic viability [3]. Such multidisciplinary cooperation is essential for addressing the complexities inherent in each stage and ensuring the successful industrial implementation of new technologies. However, interdisciplinary collaboration is still constrained by disciplinary boundaries, resulting in a big challenge to maintain design intention consistency in chemical process development

* Corresponding author. E-mail: maoye@dicp.ac.cn (M. Ye), leexin@ustc.edu.cn (X. Li).

[1] Contributed equally to this work.

This work was supported by the Liaoning Binhai Laboratory (LBLF-2023-01) and the Strategic Priority ResearchProgram of Chinese Academy of Sciences (XDA0490000), and the Key Research and Development Program of Liaoning (2023JH26/10200012).
https://doi.org/10.1016/S1872-2067(25)64725-5

[4]. Additionally, foundational stages like molecular understanding, kinetic modeling, reactor design, process operation, and complex-wide optimization necessitate specialized methodologies that are hard, if not impossible, to follow the standardized and/or predefined approaches [5]. Consequently, translating laboratory-scale innovations into industrial applications tends to be labor-intensive and time-consuming [6]. While multidisciplinary frameworks have improved the success rates of technology transfer, persistent inefficiencies highlight the urgent need for transformative approaches by enhancing the proficiency of collaborative chemical process development to accelerate innovation cycles in chemical engineering.

Recently, emerging strategies, such as data-driven artificial intelligence (AI) technologies, have gained increasing recognition for their potential to streamline development pipelines and enhance process efficiency [7–10]. Particularly, the emergence of large language models (LLMs), trained on extensive corpora that encapsulate complex, cross-disciplinary information [11,12], presents unprecedented opportunities to revolutionize the scientific workflow. Notably, these systems exhibit transformative potential in autonomously designing, planning, and executing complex chemical experiments, thereby catalyzing a paradigm shift in chemical research methodologies [13–16]. Empirical evidence across multiple domains underscores this potential, with LLMs demonstrating some ability in chemical entity recognition, molecular design, yield prediction, catalyst discovery, and reaction optimization [17–22]. For example, these LLMs excel in processing chemical literature and datasets expressed in natural language, thereby enabling tasks such as literature mining, data extraction, and decision support in catalyst design [16,23]. Furthermore, recent studies have highlighted that, when integrated with search algorithms, LLMs can effectively identify plausible catalytic mechanisms by combining chemical principles with systematic reasoning [24]. Despite their versatility, general-purpose LLMs frequently face limitations in addressing specialized chemical task requirements due to the lack of domain-specific knowledge, resulting in a significant performance gap compared to specialized chemical models [25]. As a result, considerable attention has been directed toward the integration of chemical knowledge into LLMs, resulting in the development of specialized chemical LLMs, including ChemDFM [25], LlaSMol [26], BatGPT-Chem [27], and ChemLLM [28]. These models are designed with tailored training strategies aimed at embedding domain-specific chemical knowledge, thereby improving their performance in chemical applications. Such advancements highlight the enhanced capabilities of fine-tuned LLMs in tackling complex chemical problems and highlight their potential to accelerate innovation in chemical sciences. However, current chemical LLMs mainly focus on molecular-scale tasks, with limited applicability to system engineering challenges. Consequently, significant limitations persist in addressing core chemical engineering problems, such as process simulation, equipment design, and industrial-scale optimization.

During the development of domain-specialized LLMs, a systematic assessment of their ability to comprehend and apply domain knowledge remains equally critical [29,30]. In the chemical sciences, several robust benchmarks, including ChemLLMbench [17], SciBench [31], and ChemEval [32], have been established to facilitate such assessments. For example, ChemLLMbench comprises eight tasks designed to evaluate fundamental competencies in chemical concept interpretation, logical reasoning, and explanatory proficiency [17]. Similarly, SciBench collects open-ended questions from college-level textbooks in physics, chemistry, and mathematics to assess the ability to solve complex scientific problems [31]. Meanwhile, ChemEval provides a comprehensive evaluation of LLMs' performance across diverse chemical domain tasks [32]. Despite these advancements, current benchmarks exhibit significant limitations in evaluating LLMs' performance in chemical engineering research, particularly in assessing the core competencies required for industrial-scale challenges. This gap highlights the pressing need for the development of specialized benchmarks that enable systematic and rigorous assessment of LLMs' proficiency in resolving chemical engineering problems. Such benchmarks would facilitate the effective deployment of LLMs in critical applications, including catalyst design, fluid dynamics simulation, process optimization, and apparatus selection.

In this paper, to meet the growing demands of chemical engineering for foundational language models, we present ChemELLM, the first domain-specialized LLM designed for chemical engineering applications. Built upon the Spark 70B foundation model, ChemELLM undergoes domain-adaptive pretraining and instruction fine-tuning using ChemEData, a carefully curated corpus of high-quality chemical engineering. Furthermore, to comprehensively evaluate the knowledge and problem-solving capabilities of LLMs in chemical engineering, we have developed ChemEBench, a comprehensive benchmark specifically designed for this domain. ChemEBench comprises 3 levels, 15 dimensions, and 101 distinct tasks, covering a broad spectrum of challenges in chemical engineering research. Notably, it includes numerous tasks that have not been unexamined in existing benchmarks. Using ChemEBench with designed prompts, 10 general-purpose LLMs (O3-mini [33], O1-Preview [33], GPT-4o [33], Claude 3.7 [34], Llama 3.1 [35], DeepSeek-R1 [36], DeepSeek-V3 [37], Kimi [38], GLM-4 [39], and ERNIE-4.0 [40]) and 3 chemical domain LLMs (ChemLLM [28], ChemDFM [25], and LlaSMol-Mistra [41]) are evaluated.

2. Methodology

In this section, the proposed dataset, evaluation benchmark, and the developed LLM will be sequentially introduced.

2.1. ChemEData

The critical determinant in improving LLMs' scientific problem-solving capabilities lies in the construction of large-scale, high-quality datasets. In the context of domain-specific applications in chemical engineering, this technology necessitates the creation of a purpose-built textual corpus. In response, we have constructed ChemEData, a high-quality collection of chemical engineering texts serving as the foundation for both the pre-training and fine-tuning stages of our specialized LLM. This

Table 1
Statistics of pre-training data.

Data source	Document	Size
Scholarly paper	1.06 million	30.5 GB
Chemical patent	5.79 million	58.9 GB
Professional book	1200	106.2 GB

dataset enables the model to acquire domain expertise critical for excelling in specialized tasks. Specifically, the pre-training stage leverages an extensive volume of unlabeled raw text, comprising ~19 billion tokens derived from over 6.85 million chemical papers and patents and 1200 textbooks. This extensive corpus facilitates the transfer of domain-specific knowledge to the LLM through self-supervised learning. For the fine-tuning stage, we have curated more than 2.75 million high-quality synthetic instructions, encompassing ~1 billion tokens, drawn from various chemical engineering databases. These instructions are designed to enhance the LLMs' ability to execute domain-relevant directives within professional contexts.

2.1.1. *Pre-training data*

The text corpus collected for model pre-training comprises three components: papers published in high-quality journals, searchable patents retrieved from websites, and domain-specific textbooks. The statistics of the collected corpus are summarized in Table 1, which provides an overview of its volume and diversity. This comprehensive dataset forms the cornerstone for the pre-training efforts, enabling the model to assimilate a wide range of professional knowledge. Such a dataset is essential for achieving advanced performance in domain-specific tasks.

This work constructs a pre-training corpus for a chemical engineering LLM by applying a multimodal deconstruction and restructuring approach to a curated collection of raw scientific literature. The processing workflow of the corpus, as elaborated in Fig. 1, consists of the following sequential steps: First, a preliminary screening is performed using keyword searches (e.g., "chemical engineering") across academic databases and patent repositories. The process prioritizes peer-reviewed articles published within the last five years, highly cited publications, and core patent documents to establish a foundational corpus. Next, optical character recognition (OCR) technology [42] is applied to parse PDF documents into structured formats, decomposing full-text content into modular units such as text paragraphs, figures, and tabular data. The document parsing process involves initial parsing using the in-house developed integrated PDF parsing tool, followed by a three-tier data correction mechanism customized for chemical documents. This pipeline includes: (1) fixing symbol confusion using predefined character mapping tables (e.g., 0→O, 1→l, 9→g), (2) identifying and formatting molecular formulas based on regular expression templates (e.g., auto-completion of subscript tags), and (3) standardizing unit conversions with the Pint library for unit conversion. When evaluated on a manually annotated dataset comprising 100 research papers, the system achieved an overall recognition accuracy of 90%. This step is crucial as it breaks down complex scientific documents into more manageable and analyzable components. To mine table-text relationships, a context-aware semantic matching approach is employed: explicit correlations between datasets and descriptions are established by identifying table reference markers (e.g., "as shown in Table 1") and analyzing keyword co-occurrence patterns in adjacent paragraphs. Subsequently, a knowledge object tagging schema is designed and implemented. Text is tagged based on paragraph segmentation, while tables and figures are tagged according to their native numbering. This tagging system provides a standardized way to identify and organize the knowledge objects. Finally, a standardized data organization framework is employed for the structured storage of the knowledge objects and their relationships while explicitly preserving their relations. The resulting structured corpus data greatly facilitates the pre-training of the chemical engineering LLMs, enabling them to better learn and understand the complex knowledge within the chemical engineering domain.

2.1.2. *Fine-tuning data*

In the fine-tuning stage, to enable the model with the ability

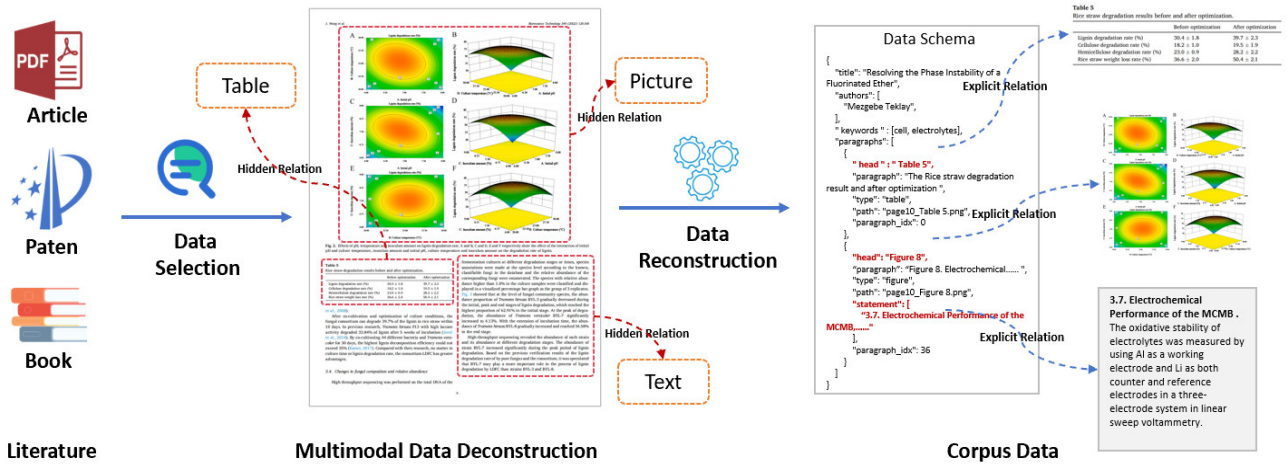


Fig. 1. Flowchart for the conversion of scientific literature into the pre-training corpus.

Table 2
The supervised fine-tuning data contains 1.24 million instruction-tuning Q&A.

Type	Catalyst	Simulation	Equipment	Separation	Safety	Heat	Engineering
Multiple choice	24600	43900	48600	96700	9200	10800	8000
True/False	14100	46800	42400	80100	5800	10000	7000
Fill-in-the-blank	19500	39900	44100	83100	9000	10000	2000
Calculation	30500	54200	63700	116700	13100	12500	5000
Short answer	31500	46300	63000	117500	13000	11200	7000
Sum	120200	231100	261800	494100	50100	54500	29000
	1240800						

to understand task instructions for multitasking, we constructed a large-scale, comprehensive, and high-quality dataset through self-supervised data synthesis for instruction tuning. This dataset covers over 2.75 million professional instructions and is divided into three levels: foundational knowledge, advanced knowledge, and professional skills.

The foundational knowledge and advanced knowledge levels were sourced from open-source datasets. After a rigorous screening process, 1.51 million high-quality data entries were retained from these sources. For the professional skills level, to better align with chemical engineering tasks, we created a training set of 1.24 million entries related to majors in the chemical industry field. Table 2 presents the distribution of 1.24 million instruction-tuning question-answering (Q&A) within the supervised fine-tuning data. These pairs are systematically categorized across five task types: multiple choice, true/false, fill-in-the-blank, calculation, and short answer, and seven specialized chemical engineering domains: Catalyst, Simulation, Equipment, Separation, Safety, Heat, and Engineering.

To construct such a dataset, a three-round prompt engineering methodology “Question Generation - Answer Generation - Quality Check” was implemented by integrating few-shot examples and self-supervised learning methods, as schematically outlined in Fig. 2. The process initiates with the first-round prompt engineering of “Question Generation”, which involves extracting information from raw domain literature such as textbooks and patents. The parsed fragments are then utilized to generate five types of questions: Multiple choice (conceptual distinctions), Fill-in-the-blank (terminology recall), True/False (fact verification), Short answer (descriptive reasoning), and Calculation (numerical problem-solving), thereby forming corresponding Q&A pairs with detailed thinking processes. Notably, this first round incorporates an important step of manual annotation, which is crucial for ensuring the accuracy and relevance of the generated Q&A and adds a layer of quality control that automated processes might overlook. Subsequently, the second and third iterations of prompt engineering, namely "Answer Generation" and "Quality Check," build upon the initial phase, refining the prompts and further expanding the diversity of the question-and-answer generation process. During the "Answer Generation" phase, we feed the original questions, fragmented responses, or referenced knowledge into the LLMs as the knowledge base for the generation of answers. And in the prompt template, emphasize the thinking process of requiring the LLM to supplement the answer. This approach maximizes the authenticity of the answer sources, thereby enhancing the usability of the answers. Specifically, for answer generation, we employed multiple LLMs, including Spark model, GPT-4o, and LLaMA-3-70B, to produce candidate responses. These responses are then evaluated using a model-based scoring system with four criteria, as outlined in Table 3. To enhance the LLMs’ ability to assess answer quality, we provided illustrative scoring examples. Each LLM then returns a score ranging from 0 to 5 based on the predefined criteria. Only those question-answer pairs consistently assigned a score of 5 by multiple LLMs are added to the training set as high-quality data. An example of a scoring result from GPT-4o is presented in Supplementary Table S1 to illustrate this pro-

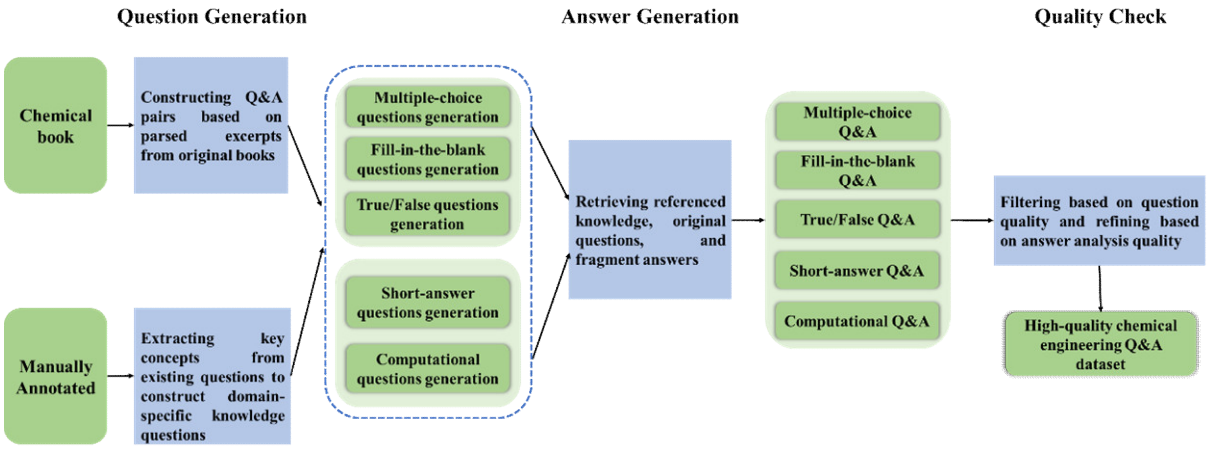


Fig. 2. The construction of fine-tuning data in the format of Q&A pairs.

Table 3
Criteria for model-based scoring of answer generation.

Dimension	Definition	Score range
Objectivity	the question should have a unique and objective answer under unified evaluation standards	0–5
Rationality	the question and answer must be complete and clear, without omitting critical information	0–5
Accuracy	the reasoning chain should be checked step by step to ensure the absence of factual, logical, computational, or knowledge errors	0–5
Generalizability	questions and answers should be based on general domain knowledge rather than relying on specific papers or patents	0–5

cess.

"Quality Check" is a continuous and crucial aspect. In response to this issue, we have established a comprehensive set of quality control guidelines designed to direct LLMs in assisting the evaluation of the logic, comprehensiveness, and accuracy of both the questions and answers in the dataset. Any data that fails to meet these standards is discarded. Considering that reliance on a singular Q&A format may undermine the model's robustness, we manually designed at least 30 instruction templates for each of the five question types in the supervised fine-tuning data. The comprehensive dataset, which passes the quality check, will be embedded into the corresponding templates based on the task type, thereby diversifying and enriching the expression of the questions. Detailed descriptions and examples for each task can be found in Supplementary Table S2. Ultimately, a high-quality fine-tuning dataset is formed by integrating generated questions, crawled knowledge, and the results of quality screening. This comprehensive dataset is designed to train models for proficient question-answering in the chemical engineering domain.

2.2. ChemEBench

To evaluate the effectiveness of LLMs in tackling chemical engineering queries and their proficiency in comprehending and applying chemical engineering knowledge, we have established a multidimensional benchmark called ChemEBench. This benchmark comprises three progressive stages designed to evaluate the capabilities of LLMs in this specialized domain comprehensively：

1. Foundational knowledge level (L1). This initial level focuses on developing a robust comprehension of core domain knowledge. During this stage, the model's proficiency in understanding fundamental concepts in chemical engineering is assessed, with a focus on its capacity to accurately interpret basic principles and terminology. This evaluation ensures that the model possesses a solid foundation upon which a more advanced understanding can be built.

2. Advanced knowledge level (L2). The second level shifts to a deeper understanding of advanced domain knowledge. Here, the model is evaluated on its ability to understand properties and molecular structures. This assessment is designed to demonstrate the model's advanced level of expertise, which extends beyond the foundational concepts and into more intricate areas of chemical engineering.

3. Professional skill level (L3). The final level focuses on the model's high-level professional capabilities. This stage assesses the model's aptitude for handling complex tasks, including problem-solving in real-world scenarios and the practical application of chemical engineering knowledge. By evaluating the model's performance in these areas, its readiness for real-world applications in the chemical engineering domain can be effectively gauged.

By integrating these three levels, ChemEBench provides a structured and comprehensive evaluation framework. This framework not only ensures that the model has a solid foundation in chemical engineering fundamentals but also confirms its capacity to demonstrate advanced reasoning abilities and practical application skills necessary for tackling professional-level tasks within the discipline.

Table 4 presents detailed statistics of ChemEBench. The benchmark is structured into 15 categories and includes 5 question types. This structure ensures a comprehensive evaluation of a model's capabilities across diverse question formats. And the diversity of question types is essential for thoroughly assessing a model's ability to handle different presentations and responses to chemical engineering queries. Furthermore, ChemEBench covers a broad range of chemical engineering topics, from fundamental principles to specialized fields such as catalysis, simulation, safety, and engineering construction. This breadth and depth highlight the benchmark's effectiveness in evaluating models' proficiency in accurately recognizing and processing a wide spectrum of chemical engineering inquiries.

Fig. 3 depicts the distribution of questions within ChemEBench, totaling 2835 questions across various categories and question types. The bar chart on the left illustrates the number of questions distributed among 15 different categories. Notably, the "Equipment" category, which encompasses six subcategories—general equipment, reactor, dryer, centrifuge, pump and tower, has the highest number of questions, totaling 610. "Engineering Construction" includes ten subcategories and has 475 questions. "Separation" (three subcategories) and "Safety" (six subcategories) collectively contains 300 questions. Additionally, the "Fluid Simulation" category, with four subcategories, has 250 questions. In contrast, category such as "Molecular Property Prediction" have the fewest questions, with only 25 questions. The pie chart on the right presents the distribution of questions based on their types. Among these, Short-answer questions (SA) are the most common, accounting for 885 questions. Multiple-choice questions (MCQs) follow with a total of 820 questions. True/False (T/F) questions to 545, Fill-in-the-blank (FB) questions amount to 405, and Numerical Calculation (NC) questions to 180. This detailed distribution analysis provides insights into the composition of the

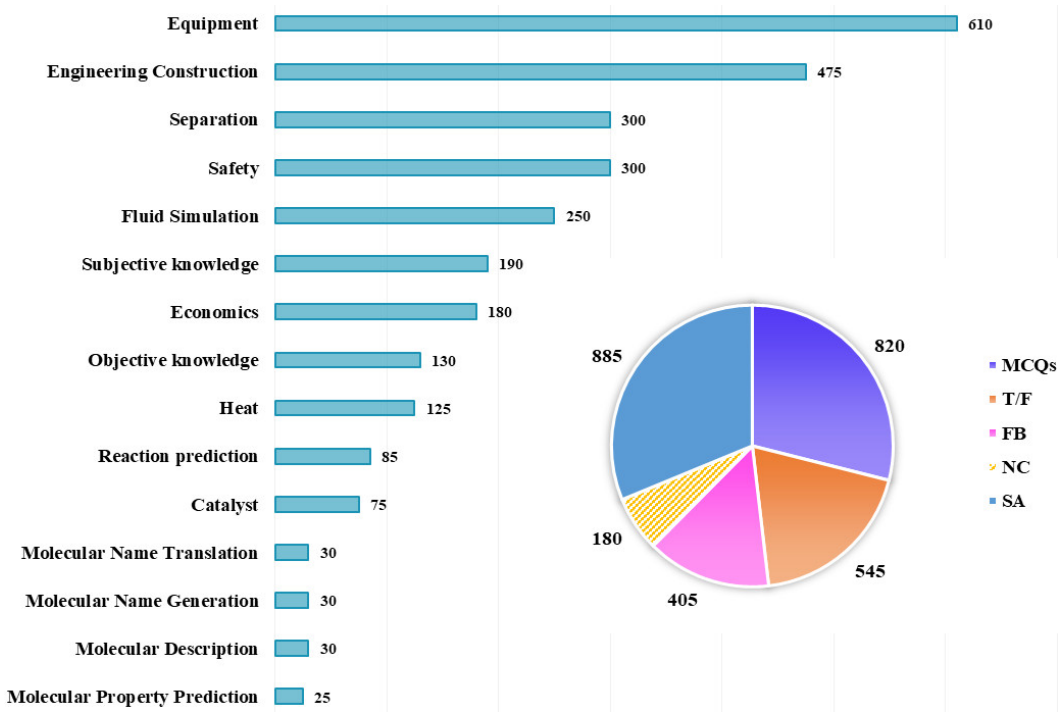


Fig. 3. Distribution of questions in ChemEBench. The bar chart shows the number of questions in different sub-domains. The pie chart shows questions classified according to question structure

dataset, which is crucial for evaluating the performance of LLMs in chemical engineering tasks.

2.3. ChemELLM

In this section, we outline the two-stage specialization process for ChemELLM, an LLM with a 70 billion parameter configuration tailored specifically for chemical engineering applications, as depicted in Fig. 4. The training pipeline consists of two essential stages: domain pre-training and supervised fine-tuning within the chemical engineering domain. During the domain pre-training stage, ChemELLM is exposed to a vast corpus of domain-specific knowledge extracted from chemical engineering papers and textbooks. This extensive exposure enables the model with a deep understanding of the terminology, concepts, and methodologies unique to the field. After pre-training, the instruction-tuning phase further refines the model's capabilities to understand and follow instructions relevant to chemical engineering. This stage equips ChemELLM with the specific language patterns, technical terms, and task structures prevalent in professional chemical engineering settings, thereby enhancing its performance on domain-specific tasks. Through this two-stage specialization process, ChemELLM attains a distinct proficiency in the chemical engineering domain.

2.3.1. Domain pre-training

The datasets utilized for training general-purpose LLMs typically contain a wide range of topics, but they tend to be relatively shallow in any specific area. As a result, although these LLMs have successfully gained strong natural language understanding and reasoning abilities, they often exhibit limitations when confronted with tasks requiring in-depth specialized knowledge [43]. Therefore, to address the limitations of general-purpose LLMs in specialized knowledge, we conducted domain pre-training on the foundational LLM, Spark-70B, using a comprehensive chemical engineering corpus consisting of 19 billion tokens. In selecting the Spark model as the foundation model for ChemELLM over other comparable LLMs, we prioritized its demonstrated superiority in key performance metrics. As evidenced in Supplementary Table S3, the Spark model demonstrates exceptional performance across a spectrum of critical domains, including logical reasoning, mathematical problem-solving, and coding efficiency. Notably, the Spark 4.0 Turbo model, updated in January 2025, demonstrates superior performance to GPT-4o across multiple domains, particularly in mathematics and coding tasks. Across seven standardized evaluation dimensions, the model achieved a performance improvement of 4.42 points compared to GPT-4o. This performance makes the Spark model particularly well-suited to meet the diverse needs of chemical engineering applications. We

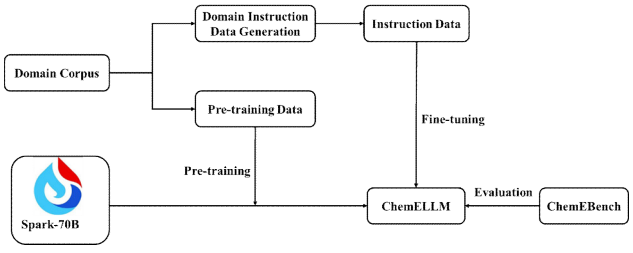


Fig. 4. The overall framework of ChemELLM.

initialized ChemELLM with the parameters of Spark-70B and employed the Adam optimizer to fine-tune the model parameters during training. The learning rate was incrementally increased from 0 to 3×10^{-5} using a warm-up strategy, then gradually decreased to 3×10^{-6} following a cosine decay schedule to ensure stable and smooth convergence. Besides, to improve training efficiency, we adopted a multi-dimensional parallel strategy that incorporated data parallelism, model parallelism, and pipeline parallelism [44]. The pre-training of ChemELLM was conducted on 128 Huawei Ascend 910b GPUs, completing one epoch of training on the 19 billion tokens dataset. This approach enabled ChemELLM to acquire additional domain-specific knowledge while retaining the foundational capabilities inherited from Spark-70B.

2.3.2. Supervised fine-tuning

During the supervised fine-tuning (SFT) stage, our goal is to align ChemELLM with the specific linguistic patterns and terminologies prevalent in chemical engineering. To this end, we fine-tuned the ChemELLM using a curated dataset comprising 2.75 million high-quality data, totaling around 1 billion tokens. For weight initialization, we utilized the ChemELLM model parameters obtained after domain-specific secondary pre-training. The optimization process utilized the Adam optimizer with an initial learning rate of 1×10^{-5}, incorporating a cosine decay strategy to adjust the learning rate during fine-tuning. The training process was executed on 128 Huawei Ascend 910B GPUs over three epochs, balancing computational efficiency with model convergence. The Cross-Entropy Loss curve as a function of iterations is presented in Supplementary Fig. S1. The visualization provides a comprehensive representation of the loss function's optimization trend during training, offering an intuitive assessment of the training performance. During the supervised fine-tuning stage, the high-quality SFT data enabled ChemELLM to enhance its understanding of chemical engineering tasks, thereby facilitating the resolution of knowledge-based queries in the chemical engineering domain.

2.4. Baseline models

For comparative analysis, we select 13 widely used and high-performing LLMs. These LLMs are strategically categorized into two types based on their intended applications. The organizations, model sizes, and accessible approaches of these modes are detailed in Table 5.

General-purpose LLMs: These models represent the cutting edge of language modeling and are designed to handle diverse tasks across multiple domains. The selected models include O3-mini, O1-Preview and GPT-4o by Open AI [33], Claude-3.7 by Anthropic [34], DeepSeek-R1 [36] and DeepSeek-V3[37] by DeepSeek, Kimi released by Moonshot Ai [38], GLM-4 by Zhipu [39], Baidu's ERNIE-4.0 [40], LLaMA 3.1-70B by Meta Platforms [35].

Scientific-domain LLMs: These models have been trained on specialized scientific data and have domain-specific knowledge to perform specialized tasks. Our selection focuses on models tailored for chemistry, including ChemDFM-13B [25], ChemLLM-7B-Char-1.5-SFT [28], and LlaSMol-Mistral-7B [41].

In our experimental design, each input begins with a system prompt that clearly delineates the types and categories of questions to be addressed. For each specific task, a standardized prompt template is employed. As illustrated in Fig. 5, using the Fill-in-the-blank as an example, we instruct ChemELLM to adopt the role of a chemical engineer and specify the tasks it is required to accomplish. The content enclosed within the parenthesis is tailored for each task, aligning with its specific inputs and outputs. The responses generated from ChemELLM are confined to solely returning the desired output without any explanations.

2.5. Evaluation criteria

We employ a set of evaluation metrics and assessment methodologies tailored to the diverse requirements of different task types. In particular, LLMs are used as judgment tools by designing distinct prompt templates for each task type to guide the LLMs in extracting or assessing predicted responses [26,45]. The specific prompt templates and detailed scoring guidelines are listed in Supplementary Figs. S2–S6. For each question, we design tailored evaluation rules and utilize the text comprehension capabilities of LLMs, coupled with specific evaluation code, to extract or score the answers. For tasks that allow multiple valid formulations, such as calculation and short answer questions, we embed the model's predicted answer, the manually annotated standard answer, and the corresponding scoring criteria into a carefully designed prompt template. By utilizing the semantic understanding capabilities of LLMs, we can efficiently evaluate and assign a score to the given question. For question types with a unique correct answer, such as Multiple-choice and True/False questions, we incorporate the standard answer along with explicit extraction rules into the prompt template. Utilizing LLMs, we can rapidly and accurately extract the predicted results. Through these approaches, we effectively leverage the semantic understanding capabilities of LLMs to systematically automate and standardize the evaluation of results for various question types.

For tasks including True/False questions, Multiple choice questions, Fill-in-the-blank, molecular translation and generation, property prediction, and reaction prediction, we employ LLMs to generate answers and subsequently compare them against the corresponding ground truth answers. Accuracy is used as the performance metric and is calculated as follows:

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ samples} \times 100\% \quad (1)$$

It is important to note that for the Fill-in-the-blank questions, potential differences in expression between the predicted and correct answers necessitate the use of LLMs as a judgment mechanism. Specifically, by designing tailored prompts, the LLM evaluates the discrepancies between the predicted and correct answers for each blank and subsequently calculates the

Table 4
The statistics of ChemEBench. It includes 3 progressive levels, evaluating 15 dimensions of LLMs capabilities and featuring 101 distinct chemical tasks.

Level	Category	Task	Type (Metric)
Foundational Knowledge	subjective Q&A of domain knowledge	objective question	multiple choice(Acc), fill-in-the-blank (Acc), true/false (Acc)
	objective Q&A about domain knowledge	subjective question	short answer (score), calculation (score)
	molecular name translation	SMILES to IUPAC	SMILES to IUPAC (Acc)
Advanced Knowledge	molecular name generation	molecular name generation from text description	molecular name generation from text description (Score)
	molecular description	generate text descriptions based on molecular smiles	generate text descriptions based on molecular SMILES (Score)
	Molecular Property Prediction	prediction of molecular properties based on molecular smiles	prediction of molecular properties based on molecular SMILES(Acc)
	reaction prediction	reaction prediction	predict the reactants from the products (Acc), predict the products from the reactants (F1), and predict whether the reaction is high yield based on the reaction information (Acc)
Professional Skill	catalyst	catalyst deactivation	short answer (score)
		catalyst stability	short answer (score)
		catalyst industrial process	short answer (score)
	equipment	general equipment	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		reactor	multiple choice (Acc), Fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		dryer	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		centrifuge	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		pump	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		tower	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
	fluid simulation	computational fluid dynamics	multiple choice (Acc), short Answer (score)
		discrete element method	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score), calculation (score)
		machine learning method	short Answer (score)
		direct numerical simulation	short answer (score), calculation (score)
	separation	absorption	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		distillation	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
		extraction	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
	heat	heat exchanger	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score), calculation (score)
	safety	regulations and standards	multiple choice (Acc)
		process safety	multiple choice (Acc), true/false (Acc), short answer (score)
		environment safety	multiple choice (Acc), true/false (Acc), short answer (score)
		personnel safety	multiple choice (Acc)
		equipment safety	multiple choice (Acc)
	economics	hazardous chemistry	multiple choice (Acc), true/false (Acc), short answer (score)
		economics	multiple choice(Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score)
	engineering construction	electrical engineering	multiple choice (Acc), fill-in-the-blank (Acc), true/false (Acc), short answer (score), calculation (score)
		automatic control	multiple choice (Acc), true/false (Acc), short answer (score)
		material engineering	multiple choice (Acc)
		equipment engineering	multiple choice (Acc), true/false (Acc), short answer (score)
		civil engineering	multiple choice (Acc)
		thermal engineering	multiple choice (Acc)
		water supply and drainage engineering	multiple choice (Acc)
		general plot plan	multiple choice (Acc)
		chemical system	multiple choice (Acc), true/false (Acc)
		fire protection engineering	multiple choice (Acc)

You are a chemical engineer, and your task is to complete the fill-in-the-blank question:
In the absorption process, an increase in operating pressure does not always lead to the ( ) absorption rate.
Please fill in the content within the parentheses without any additional explanations or other information.

Fig. 5. The standardized prompt template for the task of fill-in-the-blank.

accuracy rate for the current question. Similarly, for subjective questions, such as short-answer questions, computational questions, and molecular description tasks within the ChemEBench assessment system, we also adopt LLMs to evaluate the responses. In these cases, the correct answer is used as a reference to assess the quality of the predicted answer. Ultimately, a score within the range of [0, 1] is utilized to signify the degree of correctness of the predicted answer.

Furthermore, for multi-label classification tasks, such as the products from the reactants, we employ the F1 score as the evaluation metric. The F1 score, which represents the harmonic mean of precision and recall, is computed as follows:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

where TP is true positive, FP is false positive, FN is the false negative.

Finally, to comprehensively evaluate the overall performance of LLMs across multiple tasks, the mean score is introduced in this study, and the calculation formula is as follows:

$$Mean\ Score = \frac{\sum_{i=1}^{N} w_i \cdot p_i}{\sum_{i=1}^{N} w_i} \tag{5}$$

where N denotes the total number of tasks, W represents the number of samples in the current task, and p signifies the index value of the current task, which corresponds numerically to metrics such as F1, accuracy, or overall score.

3. Results and discussion

In this section, we conduct a thorough evaluation of ChemELLM using the ChemEBench to assess its performance relative to selected LLMs. Our objective is to meticulously document and analyze the performance of ChemELLM across a diverse array of tasks. This rigorous assessment provides in-depth insights into ChemELLM's capabilities in real-world chemical engineering applications, highlighting its strengths and identifying potential areas for improvement. By comparing it against state-of-the-art LLMs, we demonstrate the advantages of domain-specific specialization in enhancing model proficiency and applicability.

3.1. Overall performance

Firstly, we evaluated LLMs on ChemEBench. We then computed the mean score of each model across the primary evaluation dimensions. Table 6 summarizes the performance rankings of the LLMs, providing valuable insights into the strengths and weaknesses of each model. The results indicate that significant differences exist among the LLMs. Representative examples demonstrating ChemELLM's performance on domain-related problems are provided in Supplementary Tables S4-S18.

In overall performance evaluations, ChemELLM attains the highest score (72.90) among all LLMs, outperforming the second-ranked DeepSeek-R1 (70.33) by nearly two percentage points. O1-Preview (65.76) and DeepSeek‑V3 (62.96) follow closely behind. Models such as Claude 3.7, O3‑mini, ERNIE 4.0, GPT‑4o, and GLM‑4 rank fifth through ninth, respectively. At the lower end, Kimi (50.24) and LLaMA 3.1‑70B (45.26) show clear limitations on these specialized tasks. ChemELLM's superior performance highlights its benefits from specialized architectural and training advantages that are finely attuned to the demands of chemical engineering tasks, even with a potentially smaller parameter size and training corpus. In contrast, the specialized chemical LLMs, including ChemDFM-13B, ChemLLM-7B-Char-1.5-SFT, and LlaSMol-Mistral-7B, which rank twelfth, thirteenth, and fourteenth, respectively, demonstrate significant shortcomings in addressing the diverse challenges posed by ChemEBench. Collectively, these findings illustrate that ChemELLM strikes an optimal balance between model complexity and performance, making it a highly promising candidate for advanced applications in chemical engineering.

Additionally, to evaluate the potential impact of domain-specific pre-training and fine-tuning on ChemELLM's general-purpose capabilities, we conducted a comprehensive assessment of the model's performance on general language tasks such as text generation and understanding. The experimental setup involved comparing ChemELLM against several general-purpose LLMs, including DeepSeek-R1, GPT-4o, O1-Preview, and the Spark model. The results, summarized in Supplementary Table S19, indicate that while ChemELLM ex-

Table 5
Detailed information of the LLMs chosen for evaluation in our experiments. The "size" column represents the number of parameters of each model. The "access" column represents approaches to obtain models through API or loading models with weights.

Model	Developer	Size (parameter)	Access
O3-mini	OpenAI	undisclosed	API
O1-Preview	OpenAI	undisclosed	API
GPT-4o	OpenAI	undisclosed	API
Claude-3.7	Anthropic	undisclosed	API
LLaMA 3.1-70B	Meta	70B	weights
DeepSeek-R1	DeepSeek	671B	API
DeepSeek-V3	DeepSeek	671B	API
Kimi	Moonshot AI	undisclosed	API
GLM-4	Zhipu AI	undisclosed	API
ERNIE-4.0	Baidu	undisclosed	API
ChemDFM-13B	Suzhou Lab	13B	weights
ChemLLM-7B-Char-1.5-SFT	Shanghai AILab	7B	weights
LlaSMol-Mistral-7B	OSU	7B	weights

hibits a slight performance drop on certain general tasks compared to the foundation Spark model, it retains competitive capabilities relative to other general-purpose LLMs. This demonstrates that ChemELLM's domain-specific training strategy effectively enhances its performance on chemical engineering tasks without significantly compromising its general language abilities.

3.2. Performance on each level

In the foundation knowledge dimension, L1 reflects the model's grasp and memory of scientific knowledge. DeepSeek-R1, leveraging its extensive knowledge reserve and sophisticated model architecture, achieves the highest performance with a mean score of 82.19, demonstrating a clear superiority in fundamental tasks. DeepSeek-V3, equipped with 671 billion parameters, also exhibits remarkable learning capabilities, resulting in superior performance in relevant tasks. Remarkably, ChemELLM, despite having only 70 billion parameters, achieves a higher accuracy than DeepSeek-V3, underscoring its exceptional adaptability and specialization in chemical engineering tasks. This performance highlights that in the chemical engineering domain, model performance is not determined solely by the number of parameters but also by the effective integration of domain-specific knowledge.

In the advanced knowledge dimension, L2 measures the model's comprehension and exploration abilities within scientific contexts across 5 critical tasks (molecular name translation, molecular name generation, molecular description, molecular property prediction, and reaction prediction). Overall, ChemELLM achieves a strong performance with a score of 50.25, significantly outperforming the second-best model, ChemDFM-13B, which scores 28.25. In contrast, general-purpose LLMs struggle with these tasks, highlighting the challenges they face in adapting to the nuanced demands of chemical engineering.

In terms of the professional skill dimension (L3), which evaluates the model's capability to handle specialized tasks in chemical engineering. ChemELLM achieves a leading score of 74.72, outperforming DeepSeek-R1's 73.49. While DeepSeek-R1 benefits from powerful reasoning abilities and a broad knowledge base, ChemELLM's advantage reflects its superior domain adaptation. These findings underscore ChemELLM's exceptional specialization and robust competitiveness in handling professional-level challenges within chemical engineering.

3.3. Performance on different question types

As shown in Table 7, the performance of the models across different question types within the ChemEBench evaluation system is presented. By analyzing the performance, we can categorize the question types into three distinct categories based on their difficulty levels:

Firstly, multiple choice questions, which require selecting the correct answer from a set of provided options, are generally less difficult compared to other question types. LLMs endowed with robust textual comprehension capabilities and expansive knowledge repositories generally exhibit commendable proficiency in this domain. As depicted in Table 7, DeepSeek-R1 and ChemELLM achieve the highest and second-highest accuracy, with scores of 78.54 and 77.32, respectively, evidencing their superior performance in objective tasks.

Secondly, True/False and fill-in-the-blank questions, which primarily assess the LLM's depth of chemical knowledge, present greater difficulty than multiple-choice questions. Notably, ChemELLM scores 80.18 on the True/False problems, decisively outperforming the second-ranked O1-Preview model, which scores 71.01. In fill-in-the-blank tasks, ChemELLM achieves a score of 66.60 , below the top-ranked DeepSeek‑R1 (72.68), yet still demonstrating its robust domain expertise. These outcomes underscore ChemELLM's significant capabilities and its potential for comprehending chemical engineering knowledge.

Finally, both short answer and calculation questions are categorized as subjective tasks. Short answer questions primarily assess the model's ability to accurately address specific chemical engineering problems, while calculation questions evaluate the model's logical reasoning processes and the correctness of its final results. These two tasks not only assess the model's grasp of chemical knowledge but also its ability to navigate and resolve complex logical relationships. Experimental results indicate that ChemELLM achieves a score of 68.81 on short-answer tasks, the highest performance among the LLMs. However, in calculation tasks, O3-mini demonstrates superior performance, achieving a score of 77.64. While ChemELLM excels in short-answer tasks, its comparatively lower performance in calculation tasks suggests a need for further refinement in logical reasoning and calculation proficiency.

3.4. Performance on ChemLLMBench

Building on the evaluation in Section 3.1, where ChemELLM demonstrated clear advantages across chemical engineering

Table 6
Performance of the selected LLMs and ChemELLM. The best and second-best results are labeled in bold and underlined, respectively.

Model	L1	L2	L3	Mean score	Overall rank
O3-mini	74.72	23.13	59.74	58.85	6
O1-Preview	<u>76.10</u>	23.88	67.94	65.76	3
GPT-4o	62.81	23.19	58.48	56.48	8
Claude-3.7	70.38	21.76	64.01	61.75	5
LLaMA 3.1-70B	48.48	10.25	47.84	45.26	11
DeepSeek-R1	**82.19**	14.75	<u>73.49</u>	<u>70.33</u>	2
DeepSeek-V3	69.83	17.13	65.97	62.96	4
Kimi	51.12	16.25	53.06	50.24	10
GLM-4	54.95	11.75	57.24	53.77	9
ERNIE-4.0	57.01	26.62	60.49	57.71	7
ChemDFM-13B	29.71	<u>28.25</u>	31.69	31.22	12
ChemLLM-7B-Char-1.5-SFT	20.10	6.50	21.97	20.67	13
LlaSMol-Mistral-7B	16.90	26.38	19.64	19.81	14
ChemELLM	<u>73.88</u>	**50.25**	**74.72**	**72.90**	1

Table 7
Performance of the selected LLMs and ChemELLM on different question types. The best and second-best results are labeled in bold and underlined, respectively.

Model	Objective task			Subjective task		Mean score	Overall rank
	multiple choice	true/false	fill-in-the-blank	calculation	short answer		
O3-mini	59.63	62.94	53.12	**77.64**	54.41	58.85	6
O1-Preview	71.46	71.01	61.46	72.22	57.88	65.75	3
GPT-4o	63.78	56.88	54.25	56.80	51.09	56.69	8
Claude-3.7	67.93	63.49	55.97	67.78	56.37	61.75	5
LLaMA 3.1-70B	52.81	60.73	37.81	39.03	33.43	45.26	11
DeepSeek-R1	**78.54**	69.36	**72.68**	76.67	60.96	70.33	2
DeepSeek-V3	72.93	61.28	62.63	63.89	54.72	62.96	4
Kimi	53.05	56.88	46.19	43.89	46.69	50.24	10
GLM-4	60.98	62.57	51.33	43.20	44.94	53.77	9
ERNIE-4.0	64.63	64.22	54.50	49.31	50.45	57.71	7
ChemDFM-13B	29.51	43.67	25.61	11.39	31.75	31.23	12
ChemLLM-7B-Char-1.5-SFT	21.10	35.05	21.14	5.14	14.35	20.67	13
LlaSMol-Mistral-7B	13.90	48.81	13.83	1.67	13.84	19.81	14
ChemELLM	77.32	**80.18**	66.60	64.93	**68.81**	**72.90**	1

tasks on ChemEBench, we further assessed its performance using ChemLLMBench [17]. ChemLLMBench is a comprehensive benchmark encompassing a wide range of chemistry-related topics, making it an excellent supplement to ChemEBench. Table 8 presents a detailed comparison of ChemELLM against several LLMs, including DeepSeek-R1, GPT-4o, and O1-Preview.

ChemELLM consistently outperforms other models in property prediction tasks, achieving the highest accuracy across all datasets (c, BBBP, ClinTox, HIV, and Tox21 [46]). This highlights its strong ability to understand and predict molecular properties, which is crucial for applications in drug discovery, material science, and other chemistry domains [47]. However, in yield prediction tasks, ChemELLM falls short compared to other models, suggesting further improvement is needed. In name prediction tasks, ChemELLM's performance is mixed. It excels at converting SMILES to IUPAC names and vice versa, but struggles with IUPAC to formulas and SMILES to formulas. In text-based molecule design and molecule captioning, ChemELLM achieves significantly higher BLEU and score metrics, highlighting its strength in generating and interpreting textual descriptions of molecules. This is highly valuable for natural language understanding in chemistry. Additionally, ChemELLM also leads in reactant prediction, retro synthesis, and reactant selection with higher F1 scores. While its performance in solvent and ligand selection is comparable or slightly

Table 8
Performance comparison of different LLMs on ChemLLMBench tasks. The best and second-best results are labeled in bold and underlined, respectively.

Task type		Quantity	Metric	Models			
				GPT-4o	O1-Preview	DeepSeek-R1	ChemELLM
Property prediction	BACE	100	ACC	35	40	38	**64**
	BBBP	100	ACC	61	56	52	**67**
	ClinTox	100	ACC	50	52	31.5	**57.5**
	HIV	100	ACC	33	78	40	**81**
	Tox21	1044	ACC	80.27	81.9	81.03	**83.14**
Yield prediction	Buchwald-Hartwig	100	ACC	62	**75**	63	61
	Suzuki-Miyaura	100	ACC	52	**65**	61	48
Name prediction	iupac2formula	100	Exact	28	**65**	38	4
	smiles2iupac	100	Exact	1	0	0	**24**
	iupac2smiles	100	Exact	8	14	9	**20**
	smiles2formula	100	Exact	9	**42**	24	5
Molecule analysis	text-based molecule design	100	BLEU	42.56	51.76	58.12	**75.71**
	molecule Captioning	100	score	20	23.5	18.25	**26.5**
Synthetic analysis	reactant Prediction	100	F1	3	32.67	25	**61**
	retro synthesis	100	F1	4.9	14.13	11.5	**33.83**
	solvent selection	100	F1	51	51	51	51
	reactant selection	100	F1	24.7	20.83	26	**50.47**
	ligands selection	100	F1	15.27	**18.19**	16.9	17.97
Overall		2744	mean	48.78	56.67	51.36	**58.89**

inferior to other models, this still reflects its capabilities in synthetic tasks.

Overall, with a mean score of 58.89 on ChemLLMBench, ChemELLM outperforms all other models. This reinforces that ChemELLM not only maintains its strengths in chemical engineering tasks but also excels in typical chemistry challenges. It further emphasizes ChemELLM's value for both theoretical research and practical engineering development, where consistent and robust performance across diverse task types is essential.

3.5. The influence of few-shot learning

Table 9 presents a detailed comparison of each LLM's performance under 3-shot versus 0-shot settings across 3 levels (L1-L3) and 15 dimensions (C1-C15). Overall trends reveal that the effectiveness of few-shot prompting depends heavily on task complexity and model capacity.

For tasks at the L1 level, which evaluate foundational knowledge, few-shot learning has a limited impact. Most LLMs either maintain their performance or experience slight degradation or improvement, indicating that models already possess sufficient knowledge. For example, ChemELLM's performance on task C2 increases slightly from 63.56 (0-shot) to 66.35 (3-shot), while on task C1, it remains stable (80.95 to 80.36). Other models like DeepSeek-R1 and LLaMA 3.1-70B exhibit mixed results, with some improvements and others declines.

In the L2 level, which focuses on advanced knowledge, 3-shot prompts consistently enhance model performance. These tasks are highly specialized, and relevant examples enable models to better understand task requirements. Notably, ChemELLM demonstrates significant improvements on tasks C5 (45.00 to 53.33) and C7 (52.94 to 54.12). Other models also benefit, such as Claude-3.7 on task C4 (16.67 to 30.00) and

Table 9
3-shot versus 0-shot performance across LLMs and tasks. Bold indicates performance improvement compared to the 0-shot setting.

Category		L1			L2					L3							Overall
Model		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	
O3-mini	0*	74.39	**75.19**	0.00	6.67	**60.83**	36.00	20.00	**57.33**	**59.07**	**72.50**	67.31	72.78	61.00	53.21	47.75	58.85
	3*	**80.92**	72.69	**3.33**	**10.00**	56.67	**56.00**	**37.65**	51.00	56.29	71.44	**69.36**	**73.48**	**67.25**	**54.74**	**56.14**	**61.46**
O1-Preview	0*	**80.01**	70.39	3.33	3.33	**45.83**	44.00	24.71	**59.67**	**66.92**	**73.57**	**74.31**	**77.26**	**73.33**	**63.21**	59.53	65.76
	3*	79.12	**72.88**	**6.67**	**20.00**	39.17	**64.00**	**30.59**	54.67	64.02	70.55	73.89	75.46	70.92	61.99	**66.16**	**65.94**
GPT-4o	0*	**67.63**	**55.77**	**3.75**	6.67	**37.50**	**76.00**	**15.29**	**53.33**	**56.89**	66.71	65.36	**61.54**	62.83	54.54	51.84	56.69
	3*	65.93	54.62	1.67	**10.00**	25.00	72.00	10.59	48.33	56.57	**68.09**	**66.45**	60.98	**68.00**	**54.92**	**59.16**	**58.03**
Claude-3.7	0*	74.59	**64.23**	3.33	16.67	51.67	36.00	15.30	**60.50**	60.64	73.32	65.88	**71.11**	69.25	59.60	59.30	61.75
	3*	**79.21**	63.27	**6.67**	**30.00**	**62.50**	**48.00**	**29.41**	55.33	**62.73**	**74.38**	**73.55**	69.46	**73.75**	**61.45**	**64.74**	**65.48**
LLaMA 3.1-70B	0*	58.23	34.23	0.00	0.00	**28.33**	28.00	5.88	33.67	47.42	59.09	49.53	46.28	55.33	40.29	42.19	45.26
	3*	**59.96**	**38.85**	0.00	**13.33**	12.50	**60.00**	**22.35**	**36.33**	**49.08**	**59.81**	**58.92**	**50.94**	**60.17**	**41.93**	**48.07**	**49.63**
DeepSeek-R1	0*	**85.92**	**76.73**	3.33	16.67	**45.00**	12.00	8.24	**59.00**	**73.44**	**78.13**	79.16	**81.46**	78.50	67.92	66.68	70.33
	3*	85.03	70.58	3.33	**20.00**	41.67	**48.00**	**40.00**	55.67	71.19	77.30	**79.97**	78.00	**80.08**	**71.55**	**75.28**	**72.38**
DeepSeek-V3	0*	**75.90**	60.96	0.00	**20.00**	**47.50**	40.00	4.71	**56.00**	64.92	73.22	70.63	**73.14**	72.25	**62.59**	57.56	62.96
	3*	75.79	**65.96**	**3.33**	13.33	35.00	**52.00**	**16.47**	55.33	**66.80**	**73.37**	**71.14**	70.10	**74.75**	60.57	**67.72**	**65.63**
Kimi	0*	57.80	**41.35**	**3.33**	0.00	**25.00**	72.00	7.06	**45.67**	53.02	63.63	**58.40**	**57.32**	60.92	45.59	42.07	50.24
	3*	**63.29**	40.19	0.00	0.00	11.67	**84.00**	**30.59**	44.33	**53.54**	**64.26**	57.19	55.86	**62.25**	**52.52**	**53.59**	**53.64**
GLM-4	0*	**63.99**	41.73	0.00	0.00	**28.33**	44.00	4.71	47.33	54.76	65.57	60.04	**56.84**	**66.84**	55.63	50.48	53.77
	3*	60.52	41.73	0.00	**3.33**	17.50	**56.00**	**14.12**	**49**.00	**56.30**	**67.31**	**62.06**	55.74	66.00	**57.36**	**53.60**	**55.08**
ERNIE-4.0	0*	**63.51**	**47.50**	**3.33**	3.33	**37.50**	72.00	25.88	**46.33**	**62.45**	**67.58**	61.64	**61.55**	67.75	**57.71**	51.93	**57.71**
	3*	58.34	43.27	0.00	**6.67**	20.00	76.00	**32.94**	42.00	57.00	62.61	**63.97**	57.34	**69.17**	56.42	**60.88**	57.13
ChemDFM-13B	0*	40.17	14.42	10.00	**56.67**	**28.33**	40.00	**21.18**	**31.67**	30.69	34.03	**33.86**	31.16	**45.67**	31.13	21.91	31.23
	3*	**43.68**	**15.96**	10.00	40.00	11.67	**44**.00	20.00	26.33	**35.44**	**36.64**	33.72	**34.81**	44.58	**32.45**	**31.29**	**33.98**
Chem-LLM-7B-Char-1.5-SFT	0*	**27.41**	9.42	0.00	0.00	0.00	48.00	1.18	**13.33**	**24.25**	**22.07**	**28.25**	**18.12**	**25.50**	**16.48**	**17.24**	**20.67**
	3*	25.29	**14.62**	0.00	0.00	0.00	48.00	1.18	6.67	16.65	15.50	15.76	12.46	15.92	10.25	14.14	14.87
LlaSMol-Mistral-7B	0*	25.70	4.04	**3.33**	**30.00**	**19.17**	64.00	**24.71**	**9.00**	**24.48**	**17.23**	**18.31**	23.54	**23.00**	**24.67**	12.16	**19.81**
	3*	**28.27**	**7.31**	0.00	3.33	0.00	**80**	1.18	4.33	18.51	12.43	12.67	**31.18**	16.00	15.85	**24.11**	17.65
ChemELLM	0*	**80.95**	63.56	**30.00**	**56.67**	45.00	64.00	52.94	**60.67**	**72.59**	**75.56**	80.42	74.06	80.75	**69.53**	**73.95**	**72.90**
	3*	80.36	**66.35**	26.67	53.33	**53.33**	64.00	**54.12**	60.00	72.07	74.70	**82.19**	**75.44**	**82.08**	64.49	72.93	72.73

0* indicates 0-shot, 3* indicates 3-shot.

LLaMA 3.1-70B on task C6 (28.00 to 60.00).

The L3 level, encompassing expert-level skills across diverse knowledge types, presents mixed results. Weaker models often benefit from few-shot examples, while stronger models may see limited gains or even performance declines. ChemELLM maintains strong performance, showing slight improvements or stability (e.g., C11: 80.42 to 82.19, C13: 80.75 to 82.08, C15: 73.95 to 72.93). On task C14, ChemELLM experiences a minor drop from 69.53 to 64.49, while DeepSeek-R1 shows improvement, increasing from 67.92 to 71.55. Similarly, on task C15, DeepSeek-R1 demonstrates notable improvement, rising from 66.68 to 75.28.

These findings underscore that few-shot learning is beneficial when the provided examples are highly aligned with the task and the underlying knowledge is specialized. In contrast, for tasks where knowledge is already internalized, or for professional-level tasks that primarily rely on reasoning, few-shot prompts offer limited value or may even introduce noise. Compared to other models, ChemELLM consistently demonstrates strong performance across all levels, particularly in L2 and L3, owing to its well-developed pre-trained knowledge base, which reduces its reliance on few-shot prompts.

4. Conclusions

In this work, we introduce ChemELLM, a domain-specific LLM developed for chemical engineering, along with ChemEBench, the first benchmark specifically tailored to evaluate LLMs in this field. ChemEBench is structured into 3 levels, encompassing 15 domains and 101 specialized tasks, enabling a thorough and multidimensional assessment of LLM capabilities. Extensive evaluations of both general-purpose and domain-specific LLMs demonstrate that ChemELLM exhibits superior performance on domain-relevant tasks, highlighting its exceptional capability in understanding and solving complex challenges in chemical engineering.

In future work, we will focus on further enhancing the causal reasoning and multimodal capabilities of ChemELLM. Specifically, the chain-of-thought (COT) reasoning framework will be implemented, enabling the decomposition of complex engineering problems into logically sequenced sub-tasks, thereby facilitating the effective handling of multi-step problems. Furthermore, we intend to incorporate multimodal processing techniques that will allow ChemELLM to seamlessly integrate and process various types of data. This includes textual data (e.g., technical literature), visual data (e.g., equipment schematics such as process flow diagrams and piping and instrumentation diagrams), and experimental data (e.g., time-series data from distributed control systems). This expansion will not only enrich ChemELLM's contextual understanding but also significantly expand its applicability within the field of chemical engineering. With these enhancements, we anticipate that ChemELLM will evolve into an even more robust and versatile tool, thereby driving innovation and efficiency in both research and industrial applications within chemical engineering.

Acknowledgments

The authors acknowledge the support from the Liaoning Binhai Laboratory (LBLF-2023-01), Strategic Priority ResearchProgram of Chinese Academy of Sciences (XDA0490000) and the Key Research and Development Program of Liaoning (2023JH26/10200012). The authors also thank the kind help from Prof. Chun Deng, Yufei Wang, Bidan Zhao, Chang He, Kai Han, Jingai Hao, Danzhu Liu, Lei Pan, Lei Pan for providing the supervised fine-tuning data.

Competing interests

The authors declare no competing interests.

Data and code availability

The test data, API, and model weights for ChemELLM are available upon request by contacting the corresponding author. In the near future, we plan to release the test data, model weights, fine-tuning pipelines, and related documentation of ChemELLM on GitHub at https://github.com/DICPZhou/ChemELLM.

References

[1] T. Rambaran, R. Schirhagl, *Nanoscale Adv*., **2022**, 4, 3664–3675.
[2] P. Tian, Y. Wei, M. Ye, Z. Liu, *ACS Catal*., **2015**, 5, 1922–1938.
[3] Y. Xie, Y. Ma, *Comput. Aided Chem. Eng.*, **2014**, 34, 747–752.
[4] A. Wiesner, J. Morbach, W. Marquardt, *Comput. Chem. Eng*., **2011**, 35, 692–708.
[5] S. A. Gembicki, K. M. VandenBussche, A. R. Oroskar, *Chem. Eng. Sci*., **2003**, 58, 549–555.
[6] J.-F. Joly, F. Giroudière, F. Bertoncini, *Catal. Today*, **2013**, 218, 153–161.
[7] C. He, C. Zhang, T. Bian, K. Jiao, W. Su, K.-J. Wu, A. Su, *Processes*, **2023**, 11, 330.
[8] P. K. Pal, A. Hens, N. Behera, S. K. Lahiri, *Can. J. Chem. Eng*., **2025**, https://doi.org/10.1002/cjce.25611.
[9] M. Mowbray, M. Vallerio, C. Perez-Galvan, D. Zhang, A. Del Rio Chanona, F. J. Navarro-Brull, *React. Chem. Eng*., **2022**, 7, 1471–1509.
[10] L. H. Chiang, B. Braun, Z. Wang, I. Castillo, *AIChE J*., **2022**, 68, e17644.
[11] L. Li, L. Dinh, S. Hu, L. Hemphill, *arXiv preprint arXiv:2408.04163*, **2024**.
[12] C. Mammides, H. Papadopoulos, *Methods Ecol. Evol*., **2024**, 15, 1774–1776.
[13] D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, *Nature*, **2023**, 624, 570–578.
[14] T. Song, M. Luo, X. Zhang, L. Chen, Y. Huang, J. Cao, Q. Zhu, D. Liu, B. Zhang, G. Zou, G. Zhang, F. Zhang, W. Shang, Y. Fu, J. Jiang, H. N. Laboratory, Y. Luo, H. N. Laboratory, *J. Am. Chem. Soc*., **2025**, 147, 12534–12545.
[15] Q. Zhang, K. Ding, T. Lv, X. Wang, Q. Yin, Y. Zhang, J. Yu, Y. Wang, X. Li, Z. Xiang, X. Zhuang, Z. Wang, M. Qin, M. Zhang, J. Zhang, J. Cui, R. Xu, H. Chen, X. Fan, H. Xing, H. Chen, *ACM Comput. Surv*., **2025**, 57, 1–38.
[16] M. C. Ramos, C. J. Collison, A. D. White, *Chem. Sci*., **2025**, 16, 2514–2572.
[17] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N.V. Chawla, O. Wiest, X. Zhang, *Advances in Neural Information Processing Systems*, **2023**,

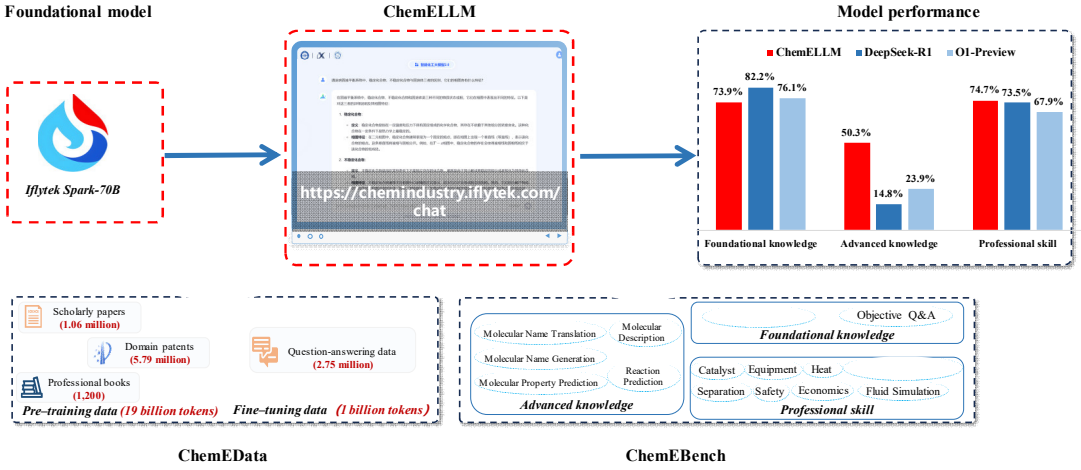
Graphical Abstract

Chin. J. Catal., 2025, 73: 159–173 doi: 10.1016/S1872-2067(25)64725-5

From lab to fab: A large language model for chemical engineering

Jibin Zhou, Feiyang Xu, Zhijun Chang, Duiping Liu, Lulu Li, Jian Cui, Yi Li, Xin Li *, Li Qian, Zhixiong Zhang, Guoping Hu, Mao Ye *, Zhongmin Liu

Dalian Institute of Chemical Physics, Chinese Academy of Sciences; Artificial Intelligence Research Institute, iFLYTEK Co., Ltd.; National Science Library, Chinese Academy of Sciences; University of Science and Technology of China; State Key Laboratory of Cognitive Intelligence



A 70 billion-parameter large language model tailored for chemical engineering

ChemELLM, a 70 billion-parameter LLM tailored for chemical engineering, outperforms leading LLMs (e.g., DeepSeek-R1) on ChemEBench across 101 tasks, trained on ChemEData's 19 billion pretraining and 1 billion fine-tuning tokens, accelerating lab-to-fab innovation.

36, 59662–59688.

[18] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, *Nat. Mach. Intell.*, **2024**, 6, 161–169.

[19] D. Bhattacharya, H. J. Cassady, M. A. Hickner, W. F. Reinhart, *J. Chem. Inf. Model.*, **2024**, 64, 7086–7096.

[20] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson, A. Jain, *Nat. Commun.*, **2024**, 15, 1418.

[21] D. H. Mok, S. Back, *arXiv preprint arXiv:2407.14040*, **2024**.

[22] L. Wang, X. Chen, Y. Du, Y. Zhou, Y. Gao, W. Cui, *Int. J. Mach. Learn. Cyber.*, **2025**, 2473.

[23] Y. Su, X. Wang, Y. Ye, Y. Xie, Y. Xu, Y. Jiang, C. Wang, *Chem. Sci.*, **2024**, 15, 12200–12233.

[24] A. M. Bran, T. A. Neukomm, D. P. Armstrong, Z. Jončev, P. Schwaller, *arXiv preprint arXiv:2503.08537*, **2025**.

[25] Z. Zhao, D. Ma, L. Chen, L. Sun, Z. Li, Y. Xia, B. Chen, H. Xu, Z. Zhu, S. Zhu, S. Fan, G. Shen, K. Yu, X. Chen, *arXiv preprint arXiv:2401.14818*, **2024**.

[26] D. Li, B. Jiang, L. Huang, A. Beigi, C. Zhao, Z. Tan, A. Bhattacharjee, Y. Jiang, C. Chen, T. Wu, K. Shu, L. Cheng, H. Liu, *arXiv preprint arXiv:2411.16594*, **2024**.

[27] Y. Yang, R. Shi, Z. Li, S. Jiang, Y. Yang, B.-L. Lu, H. Zhao, *Preprint at https://doi.org/10.26434/chemrxiv-2024-1p4xt*, **2024**.

[28] D. Zhang, W. Liu, Q. Tan, J. Chen, H. Yan, Y. Yan, J. Li, W. Huang, X. Yue, W. Ouyang, D. Zhou, S. Zhang, M. Su, H.-S. Zhong, Y. Li, ChemLLM: a chemical large language model. **2024**: 2402.06852. https://arxiv.org/abs/2402.06852v2.

[29] R. Bommasani, P. Liang, T. Lee, *Ann. N Y Acad. Sci.*, **2023**, 1525, 140–146.

[30] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, Supryadi, L. Yu, Y. Liu, J. Li, B. Xiong, D. Xiong, Evaluating large language models: a comprehensive survey. **2023**: 2310.19736. https://arxiv.org/abs/2310.19736v3.

[31] X. Wang, Z. Hu, P. Lu, Y. Zhu, J. Zhang, S. Subramaniam, A. R. Loomba, S. Zhang, Y. Sun, W. Wang, *arXiv preprint arXiv:2307.10635*, **2023**.

[32] Y. Huang, R. Zhang, X. He, X. Zhi, H. Wang, X. Li, F. Xu, D. Liu, H. Liang, Y. Li, J. Cui, Z. Liu, S. Wang, G. Hu, G. Liu, Q. Liu, D. Lian, E. Chen, *arXiv preprint arXiv:2409.13989*, **2024**.

[33] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. *arXiv preprint arXiv:2303.08774*, **2023**.

[34] A. Anthropic, https://www-cdn.anthropic.com/ de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf, **2024**, 3.

[35] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al., *arXiv preprint arXiv:2407.21783*, **2024**.

[36] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, *et al.*, *arXiv preprint arXiv:2501.12948*, **2025**.

[37] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, *arXiv preprint arXiv:2412.19437*, **2024**,

[38] R. Qin, Z. Li, W. He, M. Zhang, Y. Wu, W. Zheng, X. Xu, *arXiv preprint arXiv:2407.00079*, **2024**.

[39] T. Glm, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J.

Gui, J. Tang, J. Zhang, J. Sun, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, Z. Wang, ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools. **2024**: 2406.12793. https://arxiv.org/abs/2406.12793v2.

[40] Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, W. Liu, Z. Wu, W. Gong, J. Liang, Z. Shang, P. Sun, W. Liu, O. Xuan, D. Yu, H. Tian, H. Wu, H. Wang, *arXiv preprint arXiv:2107.02137*, **2021**.

[41] B. Yu, F. N. Baker, Z. Chen, X. Ning, H. Sun, *arXiv preprint arXiv:2402.09391*, **2024**.

[42] M. Bennamoun, G. J. Mamic, in: *Object Recognition*. London, Springer, **2002**, 199–220.

[43] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: adapt language models to domains and tasks. **2020**: 2004.10964. https://arxiv.org/abs/2004.10964v3.

[44] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, *arXiv preprint arXiv:1909.08053*, **2019**.

[45] A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fern ndez, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, *arXiv preprint arXiv:2406.18403*, **2024**.

[46] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, V. Pande, *Chem. Sci.*, **2018**, 9, 513–530.

[47] Z. Guo, C. Zhang, W. Yu, J. Herr, O. Wiest, M. Jiang, N.V. Chawla, *Proceedings of the World Wide Web Conference 2021*, **2021**, 2559–2567.

从实验室到工厂：化学工程领域的大语言模型

周吉彬[a,1], 徐飞扬[b,1], 常志军[c], 刘对平[a], 李路路[a], 崔　健[b], 李　益[b], 李　鑫[b,d,e,*], 钱　力[c], 张智雄[c], 胡国平[b,e], 叶　茂[a,*], 刘中民[a]

[a]中国科学院大连化学物理研究所, 低碳催化技术国家工程研究中心, 辽宁大连116023

[b]科大讯飞股份有限公司, 人工智能研究院, 安徽合肥230000

[c]中国科学院文献情报中心, 北京100190

[d]中国科学技术大学, 安徽合肥230000

[e]认知智能国家重点实验室, 安徽合肥230000

摘要: 化学工程技术的开发是一个复杂多阶段的过程, 涵盖实验室研究、过程放大到工业部署应用等多个环节. 该过程不仅需要化学、材料和工程等多学科的紧密协作, 还面临着漫长的研发周期和高昂的经济成本. 尽管以大语言模型为代表的生成式人工智能在基础研究领域取得显著进展, 但其在复杂工程问题中的深度应用仍面临挑战. 现有通用大语言模型对化学工程专业知识的理解有限, 难以支撑从实验室创新到工业化实施的全链条技术转化. 同时, 由于缺乏系统性评估基准, 难以客观评价大语言模型在化工专业场景中的实际性能.

为了应对上述挑战, 本文以星火大模型为基座, 成功开发出面向化学工程领域的垂直大语言模型ChemELLM, 其参数规模高达700亿. 同时, 为了全面且系统地评估大语言模型在化学工程领域的综合能力, 本文精心构建了首个化学工程多维度评估基准体系ChemEBench. 该体系采用从基础知识理解、领域高级解析到专业问题求解的递进式三级架构评估框架, 涵盖了催化剂设计、流体模拟、设备选型和安全评估等15个核心领域, 并设置101项细粒度评估任务, 实现了从基础理论认知到复杂工程建设的全维度能力评估. 基准测试结果表明, ChemELLM在上述关键指标上均表现卓越, 综合性能领先于O1-Preview, GPT-4o和DeepSeek-R1等主流大语言模型. 此外, 为了支撑大语言模型的高质量训练与微调, 构建了ChemEData数据集, 其中预训练语料规模达190亿token, 包含106万篇高质量专业文献、579万篇高价值专利以及1200本专业书籍; 微调数据集规模达10亿token, 包含275万对精心设计的问答对数据.

综上, 本研究聚焦化学工程领域大语言模型的开发, 提升其对化学工程领域的理解和推理能力, 有望建立从实验室研究到工业应用之间的桥梁, 加速化工新技术落地与产业化进程, 构建人工智能驱动化学工程创新的新范式. ChemELLM已上线部署并可公开访问, https://chemindustry.iflytek.com/chat.

关键词: 大语言模型; 化学工程; 过程开发; 多维度基准评估体系; 领域适用性

收稿日期: 2025-03-27. 接受日期: 2025-05-13. 上网时间: 2025-06-20.

*通讯联系人. 电子信箱: maoye@dicp.ac.cn (叶茂), leexin@ustc.edu.cn (李鑫).

[1]共同第一作者.

基金来源: 辽宁滨海实验室联合类基金(LBLF-2023-01); 中国科学院A类先导专项(XDA0490000); 辽宁省重点研发计划(2023JH26/10200012).