

A large language model system for the field of chemical engineering technology

Heng Zhang^{1,a} , Jibin Zhou^{3,a} , Feiyang Xu^{4,5} , Jian Cui^{4,5} , Yi Li^{4,5} ,
Fan Yang^{4,5} , Hao Wang^{1,5,*} , Xin Li^{2,4,5,6,*} , and Mao Ye^{3,*} 

¹ School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

{wanghao3, leexin}@ustc.edu.cn

² School of Information Science and Technology, University of Science and Technology of China, Hefei, China

³ National Engineering Research Center of Lower-Carbon Catalysis Technology, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, China
maoye@dicp.ac.cn

⁴ Artificial Intelligence Research Institute, iFLYTEK Co., Ltd., Hefei, China

⁵ State Key Laboratory of Cognitive Intelligence, Hefei, China

⁶ iFLYTEK (INTERNATIONAL) LIMITED, Hong Kong, China

^a These authors contributed equally to this work.

Abstract. The development of chemical engineering technology is a multi-stage process that encompasses laboratory research, scaling up, and industrial deployment. This process demands interdisciplinary collaboration and typically incurs significant time and economic costs. To tackle these challenges, we have developed a system based on ChemELLM in this work. This system enables users to interact freely with the chemical engineering model, establishing a new paradigm for AI-driven innovation and accelerating technological advancements in the chemical sector. If you would like to experience our system, please visit our official website at: <https://chemindustry.iflytek.com/chat>.

Keywords: chemical engineering · large language model.

1 Introduction

The development of chemical engineering technology is a multi-stage progression, spanning laboratory research, scaling-up, and the advancement of foundational engineering, before ultimately achieving industrial deployment [7, 8]. This intricate process demands collaboration among experts from diverse disciplinary backgrounds—including chemistry, physics, mathematics, electronics, and computer engineering—while balancing economic feasibility with resolving technical challenges. Nevertheless, interdisciplinary collaboration remains hindered by disciplinary boundaries, presenting obstacles to ensuring consistency throughout the development of chemical processes [10].

Recently, emerging strategies such as data-driven AI technologies have gained increasing recognition for their potential to streamline development processes

and enhance efficiency [3,6]. Particularly, the advent of large pre-trained models, trained on vast interdisciplinary corpora, offers greater possibilities for optimizing scientific workflows [1,5]. However, current chemical LLMs focus primarily on molecular-scale tasks and are ineffective in addressing challenges in the system engineering domain. There remain significant limitations in solving core chemical engineering issues, such as process simulation, equipment design, and industrial-scale optimization, thus developing large models for chemical engineering has become an urgent task.

In the development of domain-specific large language models (LLMs), systematically evaluating their ability to understand and apply domain knowledge is crucial. Within the broader field of chemistry, several benchmarks including ChemLLMBench [2], SciBench [9], and ChemEval [4] have been established to rigorously assess LLMs’ performance in chemical concept explanation, logical reasoning, and complex problem-solving. However, these benchmarks have limitations in evaluating LLM capabilities specific to chemical engineering, particularly when it comes to addressing the core competencies required for tackling industrial-scale challenges. Thus, there is an urgent need to develop specialized benchmarks for systematically assessing the practical application capabilities of LLMs in chemical engineering. This will facilitate their effective deployment in key technical areas such as catalyst design, fluid dynamics simulation, process optimization, and equipment selection.

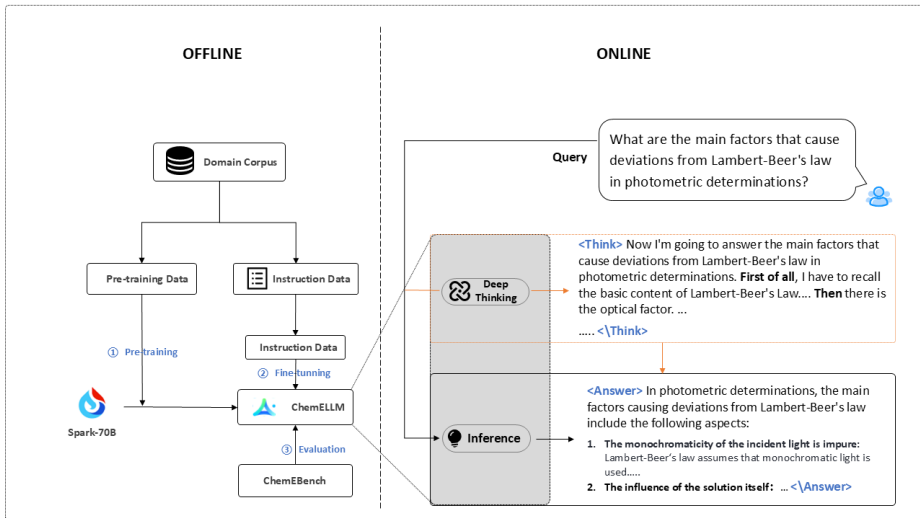


Fig. 1. System architecture diagram

This paper builds upon the work of the first domain-specific LLM designed for chemical engineering applications (ChemELLM) [11], integrating a large chemical model platform. ChemELLM uses the Spark-70B base model, leveraging

chemData (a meticulously curated high-quality chemical engineering corpus) for domain-adaptive pre-training and instruction fine-tuning.

2 System Framework

In this section, we present the framework of the large chemical model, as illustrated in Fig. 1. The entire system is divided into two components: an online component and an offline component. The offline component primarily describes the training process of the model using interdisciplinary corpora, whereas the online component outlines how the chemical model system processes user input data. As depicted in Fig. 2, users can seamlessly access the system via the intuitive interactive interface.

2.1 Online Part



Fig. 2. System interface diagram

User Input In the system described in this paper, users can input chemical engineering queries in the form of text or voice, covering multiple subfields such as catalyst properties, fundamental knowledge, fluid simulation, chemical engineering safety, physical chemistry, and chemical separation. The system supports input of up to 10,000 characters.

Question Inference Question inference, as the core component of ChemELLM in addressing queries in the chemical domain, consists of two closely linked and mutually supportive parts: ChemELLM Inference and Deep Thinking. This architectural design aims to balance the efficiency of the model’s basic operations with the rigor in handling complex problems. Among them, ChemELLM Inference forms the basic operational framework for question response, responsible for completing the technical process from original query to initial answer generation; while Deep Thinking, as an advanced processing mechanism, endows the inference results with higher accuracy, coherence, and interpretability through logical decomposition and self-iteration. Together, they form a complete "basic processing - in-depth optimization" inference loop, ensuring that the model can not only quickly respond to simple queries but also properly handle complex problems involving multi-dimensional knowledge associations.

ChemELLM Inference: After a user inputs a query, the large chemical model first preprocesses the query, including tokenization, labeling, and vector representation, converting it into a vector form that the chemical model can understand. Then, through multi-layer neural networks and techniques such as attention mechanisms, it deeply processes and analyzes the semantic and grammatical information of the input. Finally, in the decoding phase, based on the output of the semantic representation by the encoder, it generates the corresponding answer, calculates the probability distribution of the output tokens using functions like Softmax, and selects the appropriate token sequence as the final output result to present to the user.

Deep Thinking: In contrast to ordinary inference, when confronted with a user’s query, ChemELLM first decomposes the question into multiple interconnected aspects, constructing a logical chain of thought. It systematically identifies all potential sub-questions that the original query might entail, breaks down the complex problem into manageable smaller questions, and addresses each with targeted analysis. Ultimately, the model meticulously reviews the handling of each sub-question and rigorously evaluates the coherence and accuracy of the final integrated outcome. During the problem-solving process, ChemELLM conducts continuous self-assessment, engaging in explicit self-affirmation or critical self-denial to determine whether sub-questions require reprocessing—repeating the cycle until it achieves a internally consistent and convincing conclusion. Throughout this iterative inference process, the model transparently displays its detailed reasoning steps, enabling users to clearly trace the entire chain of logic from initial query to final answer.

Result Return The output content, shown on the right side of Fig. 1, varies based on whether the deep thinking feature is enabled. When activated, the system returns both the step-by-step reasoning process—with clear logical progression to trace the model’s cognitive path—and the corresponding answer, enhancing transparency and aiding users in understanding the derivation.

If the feature is not activated, the system directly outputs the inference result, optimized for efficiency to meet needs for quick answers in time-sensitive

or straightforward scenarios. This dual-mode design balances comprehensiveness and efficiency, adapting to diverse user needs in chemical engineering applications, from in-depth research to practical operations.

Example In this paper, we present the answers provided by the large model to two questions as examples.

In Fig. 3, when a user poses a question about catalytic removal of nitrogen oxides in chemical engineering technology, the large model responds comprehensively. It covers the method’s principle of using catalysts to convert NOx, its application in treating nitric acid production tail gases and fuel combustion emissions, and also mentions key aspects like catalyst selection and process flow, and would provide application examples.

In Fig. 4, users are afforded the flexibility to decide whether to engage in deep thinking according to their specific needs. When the deep thinking option is chosen, the large model showcases a comprehensive and meticulous detailed reasoning process. This process is not merely a simple listing of steps but involves a thorough exploration of the problem from multiple angles. For instance, when addressing the differences between stable compounds, unstable compounds, and solid solutions in a solid - liquid equilibrium system, the model first starts by confirming its understanding of the core concepts. It carefully defines each term, like stable compounds maintaining a fixed composition in the solid state, unstable compounds potentially decomposing or reacting in the solid state, and solid solutions being a uniform single phase formed by complete dissolution of components in the solid state. Then, it delves into the phase diagram characteristics, distinguishing between stable and unstable compounds, and considering how solid solutions manifest in phase diagrams, such as through eutectic points or congruent reactions. Throughout this, the large model continuously reflects on its reasoning steps. It checks if the definitions are accurate, if the connections between concepts and phase diagram features are logical, and if there are any gaps in the explanation. By doing so, it ensures that each part of the reasoning is sound, which in turn enhances the accuracy and reliability of the final answers. This deep thinking process serves as a valuable tool for users, as it not only provides the result but also allows them to follow along the model’s cognitive journey, thereby gaining a much deeper and more insightful understanding of the intricate issue in the field of chemical engineering thermodynamics.

2.2 Offline Part

This section’s work is primarily based on the previous work by [11]. The prior work developed a vertical large language model for the chemical engineering field, ChemELLM, and constructed the first multidimensional evaluation benchmark system for chemical engineering, ChemEBench. Experiments have demonstrated that ChemELLM surpasses mainstream large language models in key chemical engineering tasks. The offline part is introduced in three main sections, including domain pre-training, supervised fine-tuning, and ChemEBench.

Please provide a detailed introduction to what catalytic removal of nitrogen oxides is.





Catalytic removal of nitrogen oxides (NO_x) is a method that utilizes catalysts to convert NO_x in exhaust gases into harmless nitrogen gas (N₂) and water (H₂O). This method is mainly applied to treat tail gases produced during nitric acid production and NO_x emissions from fuel combustion processes. In the nitric acid production process, the nitrogen monoxide (NO) generated from the ammonia oxidation reaction is further oxidized to nitrogen dioxide (NO₂), which then reacts with water to produce nitric acid. During this process, tail gases containing NO and NO₂ are produced. Through catalytic denitration technology, these harmful gases can be effectively converted.

Basic Principle of Catalytic Denitration

Catalytic denitration typically employs Selective Catalytic Reduction (SCR) technology. SCR technology uses ammonia (NH₃) or urea (CO(NH₂)₂) as a reducing agent to reduce NO_x to nitrogen gas and water in the presence of a catalyst. The main chemical reaction equations are as follows:

$$4\text{NO} + 4\text{NH}_3 + \text{O}_2 \rightarrow 4\text{N}_2 + 6\text{H}_2\text{O}$$

$$2\text{NO}_2 + 4\text{NH}_3 + \text{O}_2 \rightarrow 3\text{N}_2 + 6\text{H}_2\text{O}$$

Selection of Catalysts

Common SCR catalysts include vanadium-titanium, iron-based, copper-based, and manganese-based catalysts. These catalysts have the following characteristics:

- High Activity: Capable of achieving efficient NO_x reduction at relatively low temperatures.
- Durability: Able to operate stably over a long period without deactivation.
- Resistance to Poisoning: Tolerant to impurities such as SO₂ and dust in the flue gas.

Process Flow

A typical SCR denitration process mainly includes the following steps:

- Pre-treatment: Removal of particulate matter and sulfides from flue gas to prevent catalyst poisoning.
- Reductant Injection: Uniformly spray liquid ammonia or urea solution into the flue.
- Mixing: Ensure thorough mixing of reductant and flue gas using static mixers or other mixing devices.
- Reaction: Mixed flue gas enters a reactor equipped with catalysts, where reduction reaction occurs at an appropriate temperature.
- Emission: Treated flue gas is discharged into the atmosphere once it meets standards.

Application Examples.....

Fig. 3. ChemELLM's response to the question

Please explain the differences between stable compounds, unstable compounds, and solid solutions in a solid-liquid equilibrium system, and describe the characteristics of their phase diagrams.

 **deep thinking:**

The user is asking about the differences between stable compounds, unstable compounds, and solid solutions in a solid-liquid equilibrium system, as well as their phase diagram characteristics. **First, I need to confirm whether my understanding of these concepts is correct.** A stable compound should be a substance that maintains a fixed composition in the solid state, like water forming ice where the composition is H_2O and does not change. An unstable compound might decompose in the solid state or react with other phases, such as certain intermediate compounds that may decompose at high temperatures. A solid solution refers to two or more components that completely dissolve in each other in the solid state, forming a uniform single phase, like the copper-nickel solid solution in alloys. **Next,** regarding phase diagram characteristics. I might need to further distinguish between stable and unstable compounds. **Additionally,** solid solutions in phase diagrams might have eutectic points or congruent reactions, while stable compounds might form simple eutectic systems, and unstable compounds could lead to more complex phase changes, such as peritectic reactions or incongruent melting. Examples might be needed to illustrate these points.

To summarize, stable compounds have definite composition and melting points in phase diagrams, often represented by vertical lines;...

In a solid-liquid equilibrium system, the differences between stable compounds, unstable compounds, and solid solutions, along with their phase diagram characteristics, are as follows:

Stable Compounds
Definition: These are compounds with a definite stoichiometric ratio that remain stable below their melting point. Upon melting, they transition directly into the liquid phase, without decomposing in the solid state.
Phase Diagram Characteristics: The phase diagram features a highest point (corresponding to the compound's melting point), similar to that of pure components.

Unstable Compounds.....

Fig. 4. Example of Using Deep Thinking

Domain Pre-training Datasets used to train general large language models (LLMs) typically cover a wide range of topics but remain relatively shallow in any specific domain. As a result, while these models have successfully acquired strong natural language understanding and reasoning capabilities, they often reveal limitations when confronted with tasks demanding profound professional knowledge. To address the shortcomings of general language models in specialized fields, we conducted domain-specific pre-training on the foundational language model Spark-70B, utilizing a comprehensive chemical engineering corpus comprising 19 billion tokens. This approach enables ChemELLM to retain the inherent foundational capabilities of Spark-70B while additionally acquiring domain-specific knowledge, thereby effectively overcoming the professional bottlenecks of general models.

Supervised Fine-tuning During the supervised fine-tuning (SFT) phase, our goal is to align the large chemical model with specific language patterns and terminology prevalent in chemical engineering. To achieve this, we employed 2.75 million high-quality data points, totaling 1 billion tokens for fine-tuning. The optimization process utilized the Adam optimizer with an initial learning rate of 1×10^{-5} , and 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 and model integration. At this stage, high-quality SFT data enhances the chemical model’s understanding of chemical engineering tasks, thereby facilitating the resolution of knowledge-based problems in the chemical engineering field.

ChemEBench The ChemEBench benchmark comprises three progressive stages aimed at comprehensively evaluating the capability of LLMs in this specialized domain. The three stages are basic knowledge level, advanced knowledge level, and professional skill level. The basic knowledge level assesses proficiency in understanding fundamental concepts in chemical engineering; the advanced knowledge level aims to demonstrate the model’s advanced professional level, surpassing basic concepts and extending into more complex areas of chemical engineering; the professional skill level evaluates the model’s ability to handle complex tasks, including problem-solving in real-world scenarios and the practical application of chemical engineering knowledge. By integrating these three levels, ChemEBench provides a structured and comprehensive evaluation framework. This framework not only ensures the model has a solid foundation in chemical engineering basics but also confirms its capability to exhibit advanced reasoning and practical application skills required to handle professional-grade tasks within the discipline.

3 Demonstration

We will begin the demonstration by explaining the framework of the large chemical model and introducing its key features. Following this, participants will be

invited to interact with the system firsthand. The demonstration process consists of two primary steps. First, users input a question related to chemical engineering into the system and select whether to enable deep thinking functionality. Second, the large model performs inference based on the user’s input, with the chemical model’s output results then displayed on the webpage.

4 Conclusion and Future works

In this work, we have introduced the large chemical model system—a domain-specific platform developed specifically for chemical engineering applications.

For future work, we aim to enhance the causal reasoning and multi-modal capabilities of the large chemical model system, while integrating online search and knowledge augmentation technologies. These improvements will enable the system to evolve into a more robust and versatile tool, thereby accelerating innovation in both research and industrial applications within chemical engineering. Additionally, the official release of the system is forthcoming.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA0490200), the Science and Technology Innovation Project of Anhui Province (Grant No. 202423k09020010), and the National Natural Science Foundation of China (Grant No. 62441239), which jointly enabled the experimental investigations and engineering analyses reported in this work. We sincerely thank the project team for their valuable resources and technical support, which significantly contributed to the completion of this research.

References

1. Chiang, L.H., Braun, B., Wang, Z., Castillo, I.: Towards artificial intelligence at scale in the chemical industry. *AIChE Journal* **68**(6), e17644 (2022)
2. Guo, K., Nan, B., Zhou, Y., Guo, T., Guo, Z., Surve, M., Liang, Z., Chawla, N., Wiest, O., Zhang, X.: Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems* **37**, 134721–134746 (2024)
3. He, C., Zhang, C., Bian, T., Jiao, K., Su, W., Wu, K.J., Su, A.: A review on artificial intelligence enabled design, synthesis, and process optimization of chemical products for industry 4.0. *Processes* **11**(2), 330 (2023)
4. Huang, Y., Zhang, R., He, X., Zhi, X., Wang, H., Li, X., Xu, F., Liu, D., Liang, H., Li, Y., et al.: Chemeval: A comprehensive multi-level chemical evaluation for large language models. *arXiv preprint arXiv:2409.13989* (2024)
5. Mowbray, M., Vallerio, M., Perez-Galvan, C., Zhang, D., Chanona, A.D.R., Navarro-Brull, F.J.: Industrial data science—a review of machine learning applications for chemical and process industries. *Reaction Chemistry & Engineering* **7**(7), 1471–1509 (2022)

6. Pal, P.K., Hens, A., Behera, N., Lahiri, S.K.: Digital twins: Transforming the chemical process industry—a review. *The Canadian Journal of Chemical Engineering*
7. Rambaran, T., Schirhagl, R.: Nanotechnology from lab to industry—a look at current trends. *Nanoscale advances* **4**(18), 3664–3675 (2022)
8. Tian, P., Wei, Y., Ye, M., Liu, Z.: Methanol to olefins (mto): from fundamentals to commercialization. *Acs Catalysis* **5**(3), 1922–1938 (2015)
9. Wang, X., Hu, Z., Lu, P., Zhu, Y., Zhang, J., Subramaniam, S., Loomba, A.R., Zhang, S., Sun, Y., Wang, W.: Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635* (2023)
10. Wiesner, A., Morbach, J., Marquardt, W.: Information integration in chemical process engineering based on semantic technologies. *Computers & Chemical Engineering* **35**(4), 692–708 (2011)
11. Zhou, J., Xu, F., Chang, Z., Liu, D., Li, L., Cui, J., Li, Y., Li, X., Qian, L., Zhang, Z., et al.: From lab to fab: A large language model for chemical engineering. *Chinese Journal of Catalysis* **73**(6), 159 (2025)