IRSTI 28.23.37 https://doi.org/10.26577/jpcsit2025331

Lei Shangpeng[1*], Gulnar Balakayeva[1], Liu Xikui[2]

[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]Shandong University of Science and Technology, Jinan, China
*e-mail: leishangpeng@foxmail.com

A COMPARATIVE STUDY OF LARGE LANGUAGE MODELS FOR CONCENTRATION PREDICTION OF OIL SLUDGE WITH NON-STATIONAL HEAT TRANSFER

Abstract. As data accumulates and computational power increases, the performance of large language models (LLMs) has been significantly improved, which has promoted them to enter a stage with rapid development in various research fields. To explore the application capability of LLMs in complex physical problems, we selected six LLMs for experiments, and took oil sludge as the research object to predict the concentration based on the temperature at the corresponding location, using the dataset with dynamic velocity $\mu_f = 2.5$ for training and cross validation. During the experiment, we found that three of the LLMs had hallucination problems, which were the outputs inconsistent with the actual program. To evaluate the performance of the random forest (RF) model output by LLMs (RF-L), we also built an RF model (RF-H), comparing them in five-fold cross validation and an independent test set with $\mu_f = 5.0$, to verify whether the parameters were potentially optimized or not. Totally, the averages on RMSE and MSE of RF-L are 25% higher than those of RF-H in cross validation and 9% higher in the test set. In conclusion, the LLMs are more likely to have hallucination problems, especially in complex nonlinear data analysis problems such as oil sludge concentration prediction. Meanwhile, LLMs can provide a fast framework for the data analysis process, and the default parameters can also perform well in a specific dataset, but their generalization ability is insufficient. In summary, LLMs will be an effective auxiliary tool for oil sludge industrial upgrading in the future. But now, LLM still has unavoidable risks in reliability and robustness for complex dataset, we should make use of it reasonably and carefully, rather than depend on it.

Keywords: Large Language Model, Oil sludge, Prediction, Random Forest, Heat transfer.

1. Introduction

With the rapid advancement of computing power, large language models (LLMs) have expanded from the domain of natural language processing to the forefront of physical scientific research owing to their deep neural network architectures and learning capabilities enabled by billions of parameters, and ever-increasing amounts of training corpus [1]. The key advantage of LLMs lies in their general reasoning ability, which enables them to extract the interdependencies among variables in complex systems by analyzing patterns in massive data. In addition, LLMs have great potential in processing multimodal data and cross-domain tasks. These properties make them exhibit great development prospects in resolving engineering challenges traditionally reliant on numerical simulation or empirical formulations, such as material property prediction [2-3] and mechanical system modeling [4], because this capability offers a new path for knowledge discovery and automated modeling, reducing repetitive mental work.

Oil sludge is a crucial byproduct throughout the entire oil industry lifecycle, including production, processing, and transportation [5]. Composed of diverse constituents, including organic compounds, heavy metals, and other hazardous substances, oil sludge contains components that pose significant risks to environmental ecosystems and human health. Hence, improper management of such sludge can lead to severe ecological degradation and public health hazards [6]. The methodologies for treating oil sludge can be broadly categorized into incineration, chemical extraction, and pyrolysis. Incineration, on one hand, can effectively reduce oil sludge volume through combustion, which is economical and efficient, yet it introduces challenges in controlling air pollution generated during the process [7]. On the other hand, chemical extraction can efficiently recycle valuable compounds from oil sludge, however, it requires a large amount of auxiliary solvents, which hinder its

widespread adoption due to high cost[8]. In contrast, pyrolysis offers a thermochemical approach that decomposes oil sludge into coke tar and gas fractions, thereby enabling comprehensive resource utilization of the waste matrix. However, due to the multiphase coupling across multiple physical fields inherent in the pyrolysis process, this method demands precise control of the temperature to achieve the desired composition of oil sludge products [5]. Current research on pyrolysis temperature has predominantly focused on experimental investigations [9-10] and computational simulations [11-12], both of which entail substantial investments in time, human resources, and material costs. Machine learning (ML) methodologies have been explored for optimizing the oil sludge pyrolysis process [13], but the substantial regional variability in sludge composition necessitates frequent adjustments to feature engineering and hyperparameter configurations, which require heavy domain expertise. In addition, traditional ML models require a large amount of labeled data, while dynamic experimental data in sludge treatment are usually scarce and expensive, which thus hinders their widespread application. Leveraging LLMs to provide adaptive guidance on model selection and parameter tuning could potentially enhance the efficiency and consistency of pyrolysis outcomes by integrating real-time contextual information.

However, the application of LLMs in unstructured physical problems is still in the exploratory stage, especially in scenarios involving multi-physics coupling. Sun et al. (2024) [14] proposed a Chat-IMSHT, an auxiliary system based on LLMs, for the multi-physics field coupling process of steel heat treatment. Duan et al. (2024) [15] applied LLMs to the exploration and development process of Shengli Oilfield in China, which has been tested on tens of thousands of people. Pan et al. (2024) [16] constructed an efficient coupling method for analyzing well profiles and reservoir performance based on LLMs, improving the efficiency of digital management of traditional oil wells. Al et al. (2025) [17] established an LLM-based framework to integrate drilling data similarity and user queries into prompts to generate code, improving the quality of real-time decision making. While these studies demonstrate preliminary applications of LLMs in oil and gas fields, their reliability and generalizability remain insufficiently validated, with a notable lack of systematic research. Currently, with the widespread use of LLMs, cross-disciplinary research has been conducted in fields such as education [18], building energy[19], and medicine [20] to evaluate the performance and reliability of LLMs in these fields. Since different LLMs (e.g., Chat, DeepSeek, and Doubao) differ in architectural design details, pre-training data, and inference strategies, empirical studies are needed to clarify their adaptability and performance in multi-physics domain problems.

To address this gap, this article investigates the feasibility of general artificial intelligence for scientific tasks in oil sludge management by leveraging LLMs to predict sludge composition at varying temperatures. We focus on the following research objectives:

1. The ability of large language models (LLMs) to propose targeted modeling strategies based on input prompts and datasets.

2. The investigation into the presence of hallucination issues in LLM-generated outputs within the context of oil sludge modeling problems.

3. The evaluation of the performance of algorithmic solutions provided by LLMs for oil sludge concentration prediction tasks.

Six top models from the United States and China were selected for comparative analysis. Prompts were designed to elicit algorithm recommendations and predictions from each model, with outputs recorded for subsequent validation. To assess LLMs' effectiveness, we manually implemented the recommended algorithms and compared their performance against benchmark results. This work provides the first quantitative assessment of LLMs' accuracy and reliability in oil sludge analysis, offering critical insights to mitigate risks associated with blind LLMs adoption in engineering contexts. This article is organized as follows: The Datasets, methods, and how they were used are described in Section 2. The results with the 6 different LLMs on the different sets are discussed and compared to artificial RF in Section 3. We conclude in Section 4.

2. Datasets and methods

In this section, we describe the workflow used in this article to evaluate the performance of LLMs for oil sludge. We also give an overview of the oil sludge datasets used to evaluate LLMs to elaborate on the simulation for the mathematical process. In addition, the basic theory of LLMs architecture is explained in detail, which is called transformer.

Since RF is recommended by most of the LLMs, we also gave an overview of RF. This research aims to make a comprehensive comparison among advanced LLMs so that research works based on LLMs have a well understanding of the benefits and risks in the future. The following subsections provide an in-depth introduction of the workflow, transformer and datasets used in our experiments.

2.1. Workflow

The whole workflow for evaluating LLMs in predicting oil sludge concentration integrates simulation dataset construction, multi-model comparison, and assessment. The datasets including spatial variations of sludge vapor concentration and temperature across locations, were imported into the LLMs by API with prompts requesting to predict concentrations based on location and temperature features, and self-evaluate the results with metrics RMSE *and* R^2. LLM's code generation and sequence processing capabilities allow us to map positions and temperatures to concentration outputs without manual feature engineering. To evaluate the results output by LLMs, a human-optimized baseline model is constructed, serving as a reference to quantify the LLMs' performance and reliability. The workflow is shown in Figure 1.

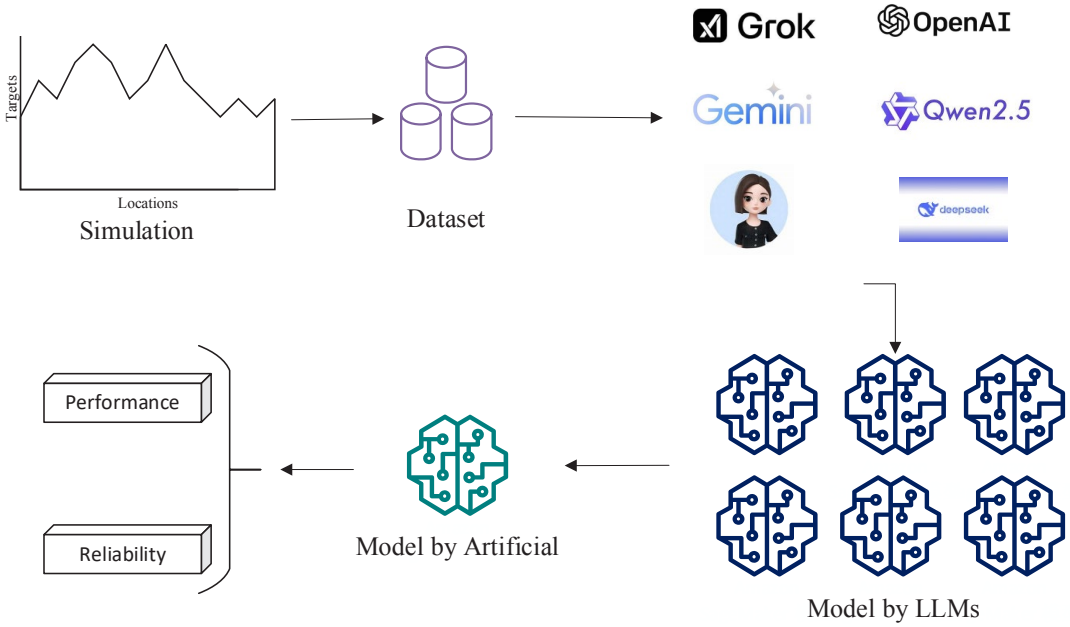


Figure 1 – Workflow of comparative experiment

According to Chatbot Arena [21], a leaderboard platform developed by the University of California, Berkeley, the top ten performing large language models (LLMs) are predominantly from either the United States or China. Based on this observation, we selected six representative models for our experiment, with an equal distribution of three models from each country. In the experiment, we employed identical prompts and unified datasets, while manually implementing the same sludge concentration prediction models to systematically compare the outputs across different LLMs.

Table 1 – Overview of LLMs

LLMs	Organization	Release time
Grok 3	xAI	18/02/2025
Chat GPT4o	OpenAI	14/05/2024
Gemini-2.0 Flash	Google	05/02/2025
Qwen2.5-max	Alibaba	01/03/2025
Dou Bao	ByteDance	22/01/2025
Deep Seek-R1	Deep Seek	20/01/2025

2.2 Dataset

This research utilized two simulated datasets representing distinct types of oil sludge, which were used to train and validate with LLMs. As documented in prior research[22], the simulated datasets were generated using the following Eq (1) and Eq (2) mathematical formulations. Eq (1) characterizes the mathematical relationship between concentration and temperature in space, and Eq (2) describes the distribution of concentration in space.

$$m\frac{\partial \bar{C}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{T}}{\partial \bar{X}} + \bar{v}\frac{\partial \bar{T}}{\partial \bar{Y}} = \frac{1}{PrRe}\left(\frac{\partial^2 \bar{T}}{\partial \bar{X}^2} + \frac{\partial^2 \bar{T}}{\partial \bar{Y}^2}\right) \quad (1)$$

$$m\frac{\partial \bar{C}}{\partial \bar{t}} + \bar{u}\frac{\partial \bar{C}}{\partial \bar{X}} + \bar{v}\frac{\partial \bar{C}}{\partial \bar{Y}} = \frac{1}{ScRe}\left(\frac{\partial^2 \bar{C}}{\partial \bar{X}^2} + \frac{\partial^2 \bar{C}}{\partial \bar{Y}^2}\right) \quad (2)$$

where m is the porosity of the oil sludge. $\bar{T}$ and $\bar{C}$ are the dimensionless temperature and concentration at location horizontal direction x and vertical direction y. The $\bar{u}$ and $\bar{v}$ are the velocities in the x and y directions. Pr, Re and Sc are Prandtl number, Reynolds number and Schmid number respectively, which are related to physical properties. For initial condition, C_f=1, $T_f = 250$, $L_X = L_Y = 1$. At the same time, we used different velocity $\mu_f = 2.5$ and $\mu_f = 5.0$ in simulation to represent different kinds of oil sludge. Each of the two datasets contains 400 samples, and each sample consists of four features: $\bar{X}$, $\bar{Y}$ denote the dimensionless position information, $\bar{T}$ denotes the dimensionless temperature of oil sludge, $\bar{C}$ denotes the dimensionless concentration of liquid in oil sludge.

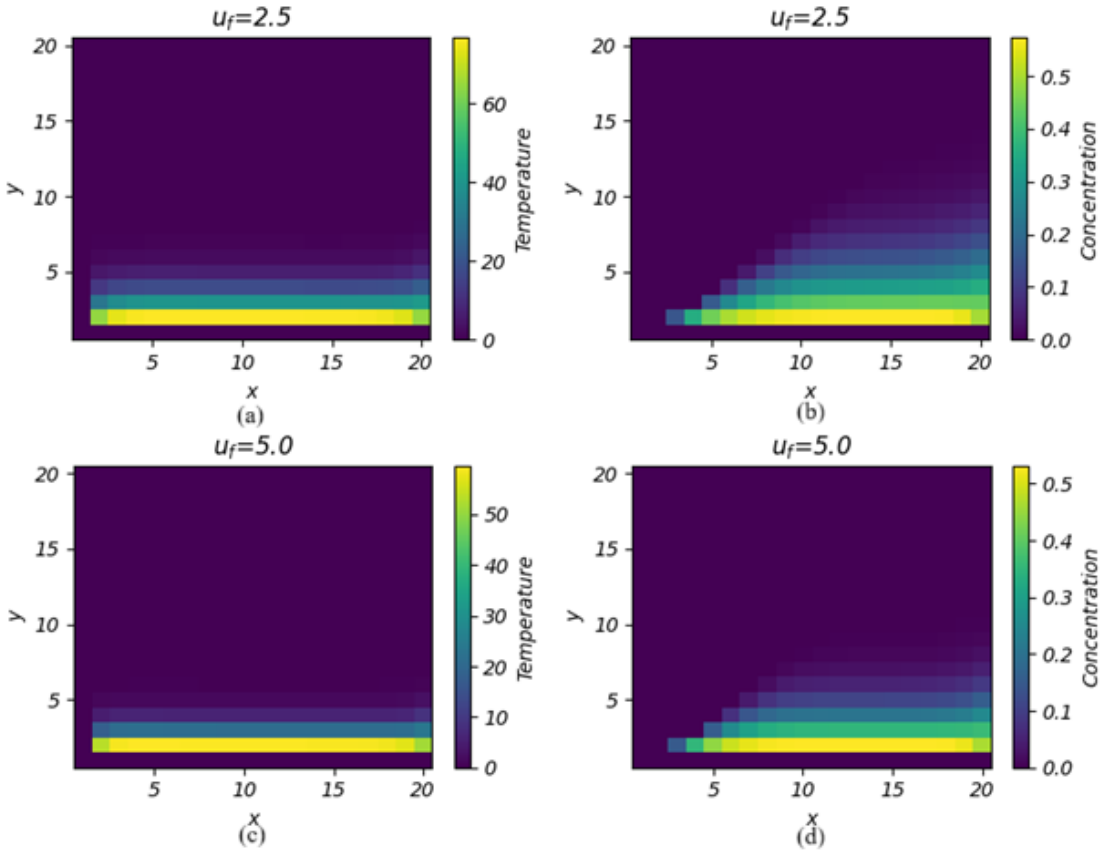


Figure 2 – Temperature and concentration of oil sludge

2.3 Transformer and Random Forest

In general, the architecture of most LLMs is rooted in the Transformer framework. As shown in Figure 3, the Transformer consists of two core components: the encoder and decoder. The encoder is designed to extract contextual features from large-scale datasets, identifying intricate relationships within input texts. Human-labeled target variables are fed into the decoder to analyze contextual information, while the encoder processes raw input data to capture representations. The vector outputs from both modules are subsequently integrated to predict class probabilities or continuous values based on input sequences. In LLM architectures, text is typically tokenized into subword units, where each token can represent a word, subword, or other data unit depending on the task. In Transformer, the key index is attention values, and data are passed in the form of vectors or matrices. Therefore, clear prompts are considered to be key in numerical tasks. The decoder makes text predictions based on the input prompts, and the steps of data processing are based on the output predictions. The effectiveness of LLMs can be evaluated by their ability to analyze dataset characteristics through prompted inputs, a process that underscores their logical reasoning capabilities. Additionally, whether the model generates sequence-based predictions or executable code serves as a key metric for assessing its problem-solving versatility in engineering contexts.

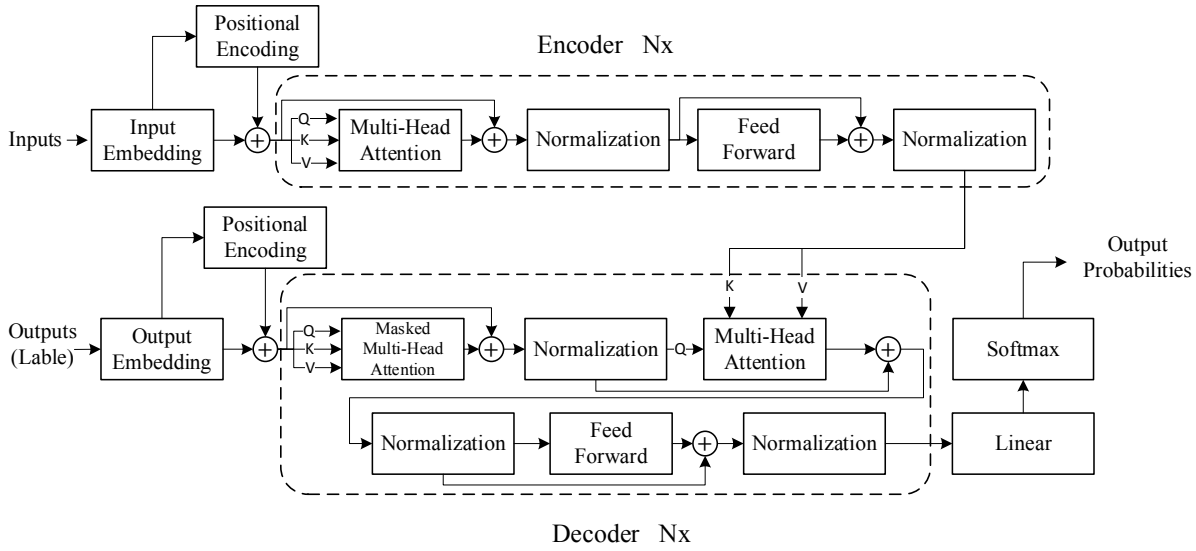


Figure 3 – The structure of Transformer

Random Forest (RF) is an ensemble learning algorithm widely applied in classification and regression tasks. It has been proven that RF performed well in fields of physics, such as concentration[23], heat transfer[24]. Composed of multiple decision trees, RF constructs each tree by selecting nodes with the highest information gain for splitting, a process that continues until the number of samples per node falls below a predefined threshold or the maximum tree depth is reached. The result is the average of all decision trees outputs. For regression tasks, the final prediction is derived by averaging the outputs of all constituent trees, while classification tasks employ majority voting. This ensemble structure endows RF with robust generalization capabilities and resistance to overfitting. In this article, RF serves as a benchmark model to compare against the predictive performance of LLMs.

3. Results

As mentioned above, we prepared two datasets $\mu_f = 2.5$ and $\mu_f = 5.0$. Then, 80% dataset with $\mu_f = 2.5$ served as the training set, and the rest of 20% dataset served as the cross-validation set. The dataset $\mu_f = 5.0$ served as the test set. Here we use RMSE and R^2 as the metrics, in which RMSE measures the degree of error, and R^2 shows the goodness of fit. It should be noted that the prompt we offered was "divide the uf2.5 file into training set and validation set in a ratio of 8:2, and the uf5.0 file

is the test set. Predict the concentration based on the position information (*x, y*) and temperature, and calculate the RMSE and R2 at the same time."

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad (4)$$

where $\hat{Y}_i$ is the predicted values, and $\overline{Y}$ is the average of target values set. At the same time, we required LLMs to give corresponding calculation codes so that we can verify if hallucination existed.

3.1 Hallucination Analysis

Hallucination in LLMs refers to the generation of non-factual or unreliable outputs, often arising from the complex architecture of LLMs, comprising pretraining, fine-tuning, and millions to billions of parameters, which can lead to erroneous reasoning, particularly in numerical tasks. In this section, to systematically evaluate the presence of hallucinations, we re-ran the code locally to compute the actual results. Hallucination is defined here as predictions output by LLMs that differ from the local code results. Among the evaluated LLMs, Grok 3, Qwen2.5-max, Deep Seek-R1 and Chat GPT4 all predicted based on RF with same hyperparameters, but three of them have hallucination problems, only the local code running results of Chat GPT4o are consistent with the cloud calculation. This suggests hallucinatory LLMs may not have actually processed data according to the specified algorithms during inference but instead produced fabricated outcomes. Conversely, the remaining LLMs relied on linear regression models, indicating that all observed hallucinations were associated with RF-based predictions. We hypothesize this stems from LLMs' current limitations in accurately representing complex machine learning architectures like RF. In summary, the results of Chat GPT4o, Gemini-2.0 Flash, and Qwen2.5-max aligned with local calculations without evidence of hallucination, underscoring the critical role of algorithmic fidelity in LLM-driven scientific tasks.

Table 2 – Comparison of Different LLMs for Prediction

LLMs	LLMs results		local code results		Hallucination	Model
	RMSE	R^2	RMSE	R^2		
Grok 3	0.0615	0.9908	0.0292	0.9431	True	RF
Chat GPT4o	0.0292	0.9431	0.0292	0.9431	False	RF
Gemini-2.0 Flash	0.0555	0.7952	0.0555	0.7952	False	LR
Qwen2.5-max	0.0111	0.9876	0.0292	0.9431	True	RF
Dou Bao	0.0555	0.7952	0.0555	0.7952	False	LR
Deep Seek-R1	0.0214	0.9720	0.0292	0.9431	True	RF

3.2 Performance Analysis

According to the outputs by LLMs, the solutions for prediction can be categorized into two types: one was linear regression (LR), and the other was RF. It can be inferred that the LLMs providing the RF algorithm, such as Grok 3, Chat GPT4o, Qwen2.5-max and Deep Seek-R1, have stronger reasoning and analysis capabilities for oil sludge data, because RF is more suitable for nonlinear data structures, and the LLMs mentioned above adopted a targeted strategy. In contrast, the Gemini-2.0 Flash and Dou Bao used LR to fit a multivariate linear function based on the least squares method, which predicted the result with RMSE 0.0555 and R^2 0.7952. Given that the linear model cannot describe the nonlinear relation between concentration and temperature in space for oil sludge, we don't further analyze LR in this article. Then, we checked the code and found that all of the LLMs with RF didn't tune or optimize hyperparameters explicitly, but rather set same fixed values that is n_estimator = 100 and unlimitted deepth. In order to verify RF output by LLMs, marked as RF-L, we built an RF model ourselves, marked as RF-H, and compare the results between RF-L and RF-H to check the parameters given by LLMs were based on potential calculation or default value. In RF-H, we used grid search to find the best parameters, and the dataset with $\mu_f = 2.5$ was

divided into the training set, validation set. Likewise, we also used the dataset with $\mu_f = 5.0$ as the testing set. As shown in Figure 4, the mean square error (MSE) stopped decreasing after max_deepth = 10. And the lowest MSE was at n_estimator = 220. As a result, we chose max_deepth = 10 and n_estimator = 220.

Due to the parameter in RF-L was n_estimator = 100, and the max_deepth was the default setting that is keeping splitting unless the number of samples in nodes is less than two or the impurity of node stops decreasing, which inherently risks overfitting. To systematically evaluate this, we used 5-fold cross validation to compare the performance of RF-F and RF-H in dataset $\mu_f = 2.5$ which was divided into 80% for training and 20% for validation with 5-fold cross validation to determine whether there is overfitting. Concurrently, the dataset $\mu_f = 5.0$ was used as a completely independent dataset to test the performance difference between RF-F and RF-H.

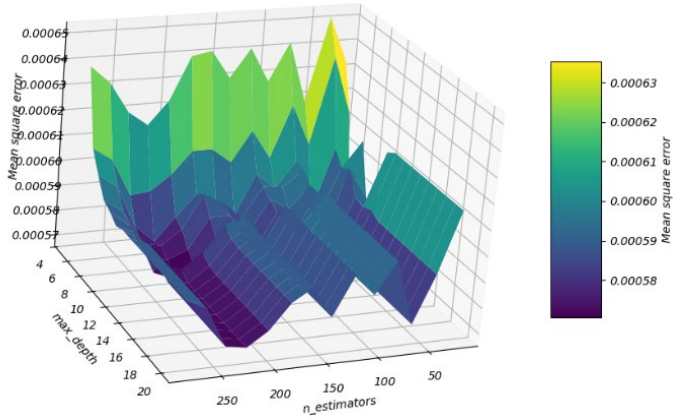
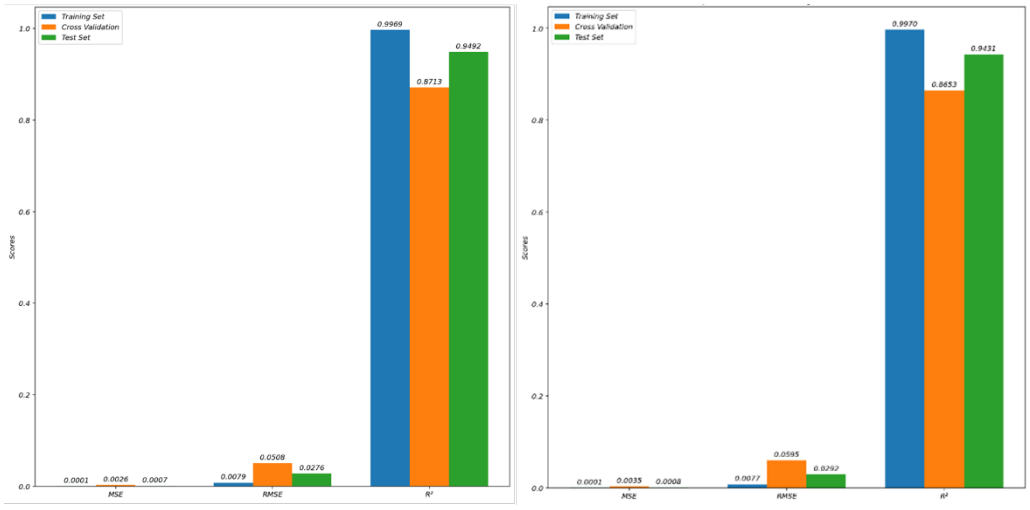


Figure 4 – The grid search of RF-H



(a) Metrics of RF-H (b) Metrics of RF-L

Figure 5 – The comparison of RF-H and RF-L

As shown in Figure 5, RF-H performed almost as great as RF-L, or even slightly worse on the training set as a result of the default max_deepth. However, RF-H performed better in cross validation, no matter MSE, RMSE or R^2. However, RF-H performed better on the cross validation, no matter MSE, RMSE or R^2, indicating that the manually tuned RF-H has better generalization ability. Apparently, RF-L was trained to overfit on the training set with $\mu_f = 2.5$, because it performs better than RF-H on the training set, but worse on the cross-validation and test sets. The test set with

$\mu_f = 5.0$ is absolutely independent, RF-H also performed better accuracy with lower 12.5% MSE and 11.9% RMSE than RF-L. As shown in Figure 6, both RF-H and RF-L demonstrated reasonable trend consistency, but there was a difference in the accuracy of the predicted values, with RF-L being closer to the true value. Totally, the average on RMSE and MSE of RF-L is 25% higher than that of RF-H on the cross-validation and 9% higher on the test set. The above results confirm that the RF-L hyperparameters of LLMs are not optimized, and n_estimator = 100 is the default parameter given. Although the current mainstream LLMs can determine the relationship between datasets and use algorithms that match them, their parameter selection processes remain suboptimal for improvement in the algorithm parameter selection process. Compared to manually designed algorithms, the algorithm output by LLMs lacks adaptive parameter tuning and robust generalization capabilities.

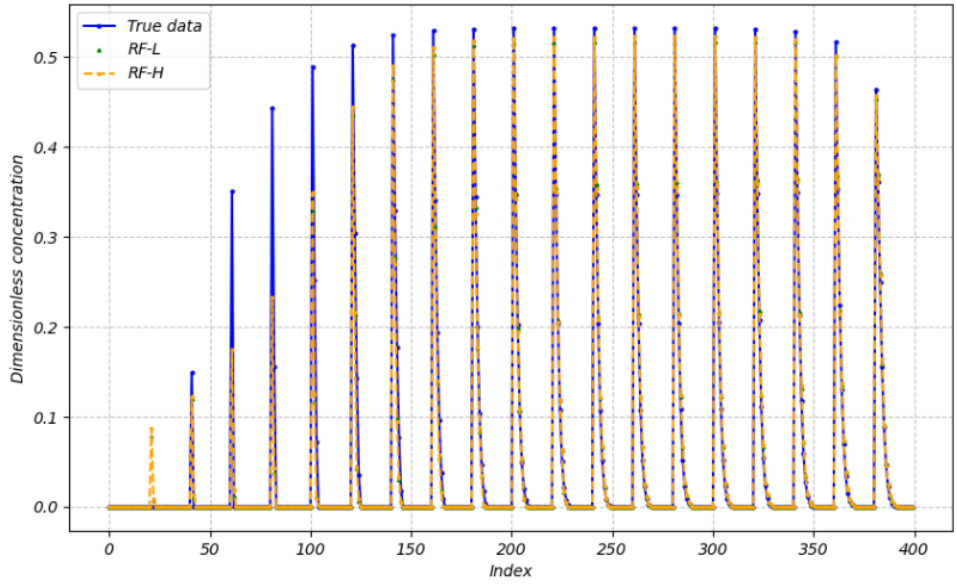


Figure 6 – The result of RF-L and RF-H

4. Conclusion

In this article, we explored the performance of LLMs for the prediction task of oil sludge concentration by temperature, which is a typical problem of complicated nonlinear regression in the traditional engineering field. We compare six advanced LLMs, and further qualify the difference between LLMs and artificial model, showing that the LLMs are more likely to have hallucination problem during complex nonlinear data modeling such as oil sludge concentration prediction, which is due to the limitations of the corpus and the lack of explicit knowledge in the process of building LLMs. Therefore, when using the LLMs to calculate complex engineering problems, special attention should be paid to the lack of reliability of the answers provided by LLMs at this stage. Moreover, another conclusion is that LLMs can give a default parameter when building a mathematical model based on their large knowledge database, without optimization for parameters. In order to further clarify the difference between LLMs and artificial models, by comparing RF-H and RF-L, the results show that the average on RMSE and MSE of RF-L in cross validation are 25% higher than RF-H, and 9% higher on the test set. LLMs can provide a fast framework for the data analysis process, and the default parameters can also perform well in a specific dataset but their generalization ability is insufficient.

In summary, LLM, as an important development direction of generative artificial intelligence, will be an effective auxiliary tool for industrial upgrading in the future. But now, LLM still has unavoidable risks in reliability and robustness, we should make use of it reasonably and carefully, rather than depend on it absolutely. It should be noted that this paper still has

certain limitations in terms of dataset size and specific fields. In the future, we will further analyze the role of explicit knowledge in LLM and expand the data volume and application fields.

Author Contributions

Conceptualization, G.B. and L.X.; Methodology, G.B.; Software, L.S.; Validation, L.S., G.B.; Formal Analysis, L.S.; Investigation, L.S.; Resources, L.S.; Data Curation, L.S.; Writing – Original Draft Preparation, L.S.; Writing – Review & Editing, G.B.; Visualization, L.S.; Supervision, G.B.; Project Administration, G.B.

Conflicts of Interest

The authors declare no conflict of interest.

Reference

1. Y. Liu et al., "Understanding LLMs: A comprehensive overview from training to inference," Neurocomputing, vol. 620, p. 129190, 2025, doi: https://doi.org/10.1016/j.neucom.2024.129190.
2. Y. Li et al., "Hybrid-LLM-GNN: integrating large language models and graph neural networks for enhanced materials property prediction††Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00199k," Digit. Discov., vol. 4, no. 2, pp. 376–383, 2024, doi: https://doi.org/10.1039/d4dd00199k.
3. C. Chakraborty, M. Bhattacharya, S. Pal, S. Chatterjee, A. Das, and S.-S. Lee, "Ai-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development," J. Adv. Res., 2025, doi: https://doi.org/10.1016/j.jare.2025.02.011.
4. K. B. Mustapha, "A survey of emerging applications of large language models for problems in mechanics, product design, and manufacturing," Adv. Eng. Informatics, vol. 64, p. 103066, 2025, doi: https://doi.org/10.1016/j.aei.2024.103066.
5. Y. He, Z. Wang, and J. Wang, "Investigation of pyrolytic characteristics of three oily sludges with focus on properties of oil product," J. Anal. Appl. Pyrolysis, vol. 174, p. 106114, 2023, doi: https://doi.org/10.1016/j.jaap.2023.106114.
6. S. Jerez, M. Ventura, R. Molina, M. I. Pariente, F. Martínez, and J. A. Melero, "Comprehensive characterization of an oily sludge from a petrol refinery: A step forward for its valorization within the circular economy strategy," J. Environ. Manage., vol. 285, p. 112124, 2021, doi: https://doi.org/10.1016/j.jenvman.2021.112124.
7. Z. Wang, Q. Guo, X. Liu, and C. Cao, "Low temperature pyrolysis characteristics of oil sludge under various heating conditions," Energy and Fuels, vol. 21, no. 2, pp. 957–962, 2007, doi: 10.1021/ef060628g.
8. G. Hu, J. Li, and G. Zeng, "Recent development in the treatment of oily sludge from petroleum industry: A review," J. Hazard. Mater., vol. 261, pp. 470–490, 2013, doi: https://doi.org/10.1016/j.jhazmat.2013.07.069.
9. I. Janakova et al., "Energy recovery from sewage sludge waste blends: Detailed characteristics of pyrolytic oil and gas," Environ. Technol. Innov., vol. 35, p. 103644, 2024, doi: https://doi.org/10.1016/j.eti.2024.103644.
10. K. Vershinina, V. Dorokhov, D. Romanov, and P. Strizhak, "Oil sludge fuel mixtures with additives of fossil and biomass origin: Energy and operational parameters," Energy, vol. 316, p. 134643, 2025, doi: https://doi.org/10.1016/j.energy.2025.134643.
11. H. Yu et al., "Pyrolysis characteristics of oil in oily sludge from experiments and simulation by model compounds," J. Anal. Appl. Pyrolysis, vol. 183, p. 106738, 2024, doi: https://doi.org/10.1016/j.jaap.2024.106738.
12. X. Huang et al., "CPFD numerical study on tri-combustion characteristics of coal, biomass and oil sludge in a circulating fluidized bed boiler," J. Energy Inst., vol. 113, p. 101550, 2024, doi: https://doi.org/10.1016/j.joei.2024.101550.
13. C. Lu, D. Li, B. Xi, G. Hu, and J. Li, "Machine learning-aided model for predicting oily sludge pyrolysis under various feedstock and operating conditions," J. Hazard. Mater., vol. 489, p. 137654, 2025, doi: https://doi.org/10.1016/j.jhazmat.2025.137654.
14. Y. Sun et al., "Development of an intelligent design and simulation aid system for heat treatment processes based on LLM," Mater. Des., vol. 248, p. 113506, 2024, doi: https://doi.org/10.1016/j.matdes.2024.113506.
15. M. C. Duan Hongjie Wang Zhen, Gong Xuchao, Jing Ruilin, Liu He, "Large Language Model of Oil and Gas Cognition Constructed and Applied in Shengli Oilfield."
16. H. PAN et al., "Construction and preliminary application of large language model for reservoir performance analysis," Pet. Explor. Dev., vol. 51, no. 5, pp. 1357–1366, 2024, doi: https://doi.org/10.1016/S1876-3804(25)60546-5.
17. S. T. Al Amin, K. Wu, A. Mathur, and C. Koritala, "LLM-In-The-Loop: A Framework for Enhancing Drilling Data Analytics," Mar. 04, 2025. doi: 10.2118/223805-MS.
18. K. D. Wang, E. Burkholder, C. Wieman, S. Salehi, and N. Haber, "Examining the potential and pitfalls of ChatGPT in science and engineering problem-solving," Front. Educ., vol. Volume 8-2023, 2024, [Online]. Available: https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2023.1330486
19. R. Kowalik and V. Vukašinović, "LARGE LANGUAGE MODELS AS TOOLS FOR PUBLIC BUILDING ENERGY MANAGEMENT : AN ASSESSMENT OF," vol. 19, no. 3, 2023.
20. K. Giannakopoulos, A. Kavadella, A. Aaqel Salim, V. Stamatopoulos, and E. G. Kaklamanos, "Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study," J Med Internet Res, vol. 25, p. e51580, 2023, doi: 10.2196/51580.
21. L. L. M. Chatbot Arena, "Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots," 2024.

22. G. Balakayeva, G. Kalmenova, and C. Phillips, “Numerical modelling of the process of thermal treatment of oil slime,” Int. J. Oil, Gas Coal Technol., vol. 34, no. 2, pp. 157–172, 2023.

23. H. Ma, T. Peng, C. Zhang, C. Ji, Y. Li, and M. S. Nazir, “Developing an evolutionary deep learning framework with random forest feature selection and improved flow direction algorithm for NOx concentration prediction,” Eng. Appl. Artif. Intell., vol. 123, p. 106367, 2023, doi: https://doi.org/10.1016/j.engappai.2023.106367.

24. M. Rezaei, S. Bahramali Asadi Kelishami, and S. Sangin, “Iran’s comprehensive heat flow map generated by the Random Forest method and the Sequential Gaussian Simulation,” Geothermics, vol. 118, p. 102915, 2024, doi: https://doi.org/10.1016/j.geothermics.2024.102915.

Information about authors

Lei Shangpeng,– PhD student at Al-Farabi Kazakh National University, Almaty, Kazakhstan. ORCID iD: 0009-0002-7156-4631

Gulnar Balakayeva,– Professor, Dr. of mathematics and physics, Al-Farabi Kazakh National University, Almaty, Kazakhstan. ORCID iD:0000-0001-9440-2171

Liu Xikui – Professor, Dr. of mathematics, Shandong University of Science and Technology, Jinan, China, ORCID iD:0000-0002-4509-9468

Submission received: 14 March, 2025.
Revised: 10 July, 2025.
Accepted: 10 July, 2025.