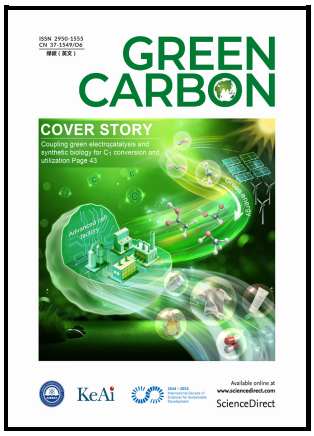
Journal Pre-proof



Evaluation of Large Language Models as an Alternative to Traditional Machine Learning Approaches for Biochar Property Prediction

Jianfeng Peng, Yi Zhang, Yu Fu, Yijing Feng, Nuo Lin, Yeqing Li, Jianping Su, Mianfeng Zhang, Xiaonan Wang

PII: S2950-1555(26)00061-3

DOI: https://doi.org/10.1016/j.greenca.2026.03.016

Reference: GREENC196

To appear in: *Green Carbon*

Received date: 27 September 2025
Revised date: 20 February 2026
Accepted date: 18 March 2026

Please cite this article as: Jianfeng Peng, Yi Zhang, Yu Fu, Yijing Feng, Nuo Lin, Yeqing Li, Jianping Su, Mianfeng Zhang and Xiaonan Wang, Evaluation of Large Language Models as an Alternative to Traditional Machine Learning Approaches for Biochar Property Prediction, *Green Carbon,* (2026) doi:https://doi.org/10.1016/j.greenca.2026.03.016

Evaluation of Large Language Models as an Alternative to Traditional Machine Learning Approaches for Biochar Property Prediction

Jianfeng Peng[a,1], Yi Zhang[b,1], Yu Fu[c], Yijing Feng[c], Nuo Lin[c], Yeqing Li[a, c, d,*], Jianping Su[d, e], Mianfeng Zhang[f], Xiaonan Wang[g,*]

[a] College of Artificial Intelligence, China University of Petroleum (Beijing), Beijing 102249, China

[b] School of Environment, Tsinghua University, Beijing 100084, China

[c] Beijing Key Laboratory of Biogas Upgrading Utilization, College of New Energy and Materials, China University of Petroleum (Beijing), Beijing 102249, China

[d] State Key Laboratory of Heavy Oil Processing, Beijing 102249, China

[e] College of Carbon Neutrality Future Technology, China University of Petroleum (Beijing), Beijing 102200, China

[f] China Aerospace Investment Holdings Ltd., Beijing 100035, China

[g] Department of Chemical Engineering, Tsinghua University, Beijing 100084, China

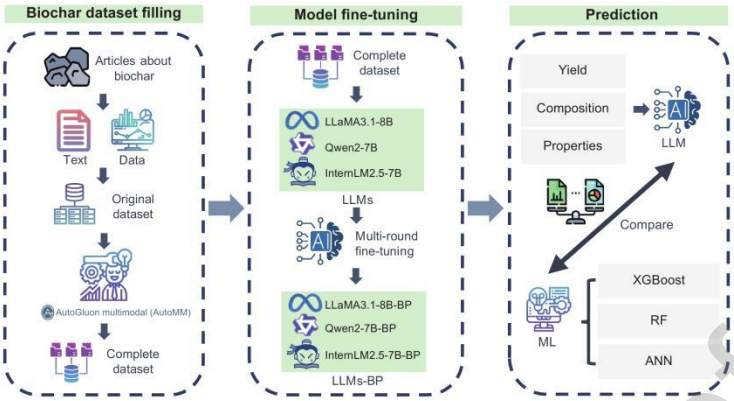
[1] These authors contributed equally to this work.

*** Corresponding authors:**

E-mail address: liyeqingcup@126.com (Yeqing Li);

wangxiaonan@tsinghua.edu.cn (Xiaonan Wang).

Graphical Abstract:



ABSTRACT:

Biochar properties are influenced by various factors, including raw material type and pyrolysis conditions. However, existing methods for predicting biochar properties exhibit significant limitations, particularly in the integration of structured and unstructured textual data. To address these challenges, this study proposes a multimodal data fusion framework based on large language models (LLMs) to improve the accuracy of prediction of biochar properties. First, we constructed a multidimensional biochar dataset incorporating the raw material composition, pyrolysis parameters, and biochar performance indicators. A customized multimodal algorithm was used to impute missing values, achieving an average R^2 value of 0.8863 across multiple features, indicating the effective restoration of data regularity. The data were then converted into question–answer pairs using a specific instruction format and fine-tuned for two rounds to characterize the model's parameter adaptation under a continuous sample supplementation scenario. Among the models, InternLM2.5-7B-BiocharPrediction-v2 (InternLM2.5-7B-BP2) achieved the highest R^2 value of 0.9552, demonstrating its ability to form stable predictive mappings for gradually expanding

in-distribution samples. In addition, compared with traditional machine learning methods, InternLM2.5-7B-BP2 showed clear advantages in handling multidimensional prediction tasks, with an average prediction accuracy improvement of 16.33%. Overall, this study demonstrates the potential of LLMs for biochar property prediction and provides a new framework and theoretical support for optimizing biochar performance prediction.

KEYWORDS: Biochar, Large Language Model, Multimodal Data Fusion, Machine Learning, Fine-Tuning.

Synopsis:

This study integrates large language models with multimodal data to improve prediction of biochar properties, as well as advance sustainable management and environmental remediation.

1. INTRODUCTION

Biochar is a carbon-based material produced by pyrolysis of organic substances under anaerobic conditions. Owing to its superior physicochemical properties and environmental friendliness [1], it is widely used in fields such as environmental protection [2], soil improvement [3], and energy storage [4,5]. The large specific surface area, complex pore structure, and surface chemical characteristics of biochar make it particularly effective for pollutant adsorption [6], heavy metal removal [7], and carbon dioxide sequestration [8,9]. However, the final properties of biochar are strongly

influenced by factors such as the type of raw material used, the preparation process, and the pyrolysis conditions [10,11,12]. Although substantial progress has been made in the application of biochar, systematic studies of the effects of raw material types and production processes on biochar performance have been scarce, with most studies reported in the literature focusing on individual physicochemical properties. For example, Wang *et al.* (2023) observed that biochar derived from different waste materials exhibited varying effectiveness in pollutant removal, with the raw material type playing a critical role [13]. Moreover, previous studies often failed to explore the relationships between raw materials, production processes and biochar's characteristics [14,15]. Additionally, current experimental approaches typically emphasize changes in individual properties and neglect more comprehensive analyses involving multiple influencing factors. As a result, these studies cannot fully capture the overall performance of biochar.

With advancements in computer science and data analysis technologies, machine learning [16] and deep learning [17] have gradually been integrated into biochar research to address these challenges [18]. For example, Khan *et al.* (2022) presented a framework using artificial neural networks (ANNs) and metaheuristic algorithms to optimize biochar yield predictions [19]. Similarly, Pathy *et al.* (2020) employed the eXtreme Gradient Boosting (XGBoost) algorithm to predict the biochar yield from algae [20]. However, none of these studies considered the combined effects of multiple influencing factors. Traditional machine learning methods rely predominantly on structured data (such as experimental data) [21], while a large amount

of textual information in the literature — such as feedstock sources, pretreatment strategies, and process context—has rarely been systematically utilized. In recent years, deep learning–based natural language processing technologies [22], such as the large language models (LLMs) proposed by Sébastien *et al.* (2023) [23], have enabled identification of patterns within unstructured textual data, providing new avenues for biochar prediction. However, the existing predictive models fail to handle the dynamic interplay between diverse raw materials and pyrolysis processes. These models often reduce complex biomass materials into single feature vectors, overlooking the chemical variations among the raw materials and the multi-dimensional nature of the pyrolysis process [24]. For example, different feedstocks under the same pyrolysis temperature can exhibit distinct thermal behaviors and carbon structure evolution pathways, leading to inconsistent trends in the properties such as specific surface area, pH, and distribution of surface functional groups [25]. Consequently, these models struggle to accurately reflect the complexity of engineering practices and biochar performance. This raises a critical question: Can LLMs extend conventional structure-based regression through semantic modeling of multimodal information about biochar preparation toward a multisource semantic inference framework that better reflects real physical processes?

To address this gap, this study proposes a novel multimodal data fusion framework based on LLMs for predicting biochar properties. Unlike traditional machine learning approaches, which rely primarily on structured features, our framework further incorporates the semantic modeling of natural language, enabling experimental parameters, feedstock descriptions, and process context to be unified

within a single predictive framework. We also constructed a multidimensional database linking various raw materials to their corresponding biochar properties. To the best of our knowledge, this is the first application of LLMs to biochar prediction, revealing the unique advantages of processing complex multisource data. Our approach not only achieves unprecedented prediction accuracy, but it also uncovers hidden correlations among biochar characteristics that have been largely overlooked by the conventional methods. This innovative framework provides more precise theoretical support for ecological simulations and biochar applications and provides a robust foundation for advancing the use of LLMs in environmental science. This achievement has profound implications because enhanced biochar prediction can lead to optimized production processes, improved environmental remediation strategies, and the establishment of a new paradigm in sustainable resource management. In summary, our work sets a new benchmark in environmental data analytics, marking a transformative step toward intelligent data-driven innovation in biochar research.

2. METHODS

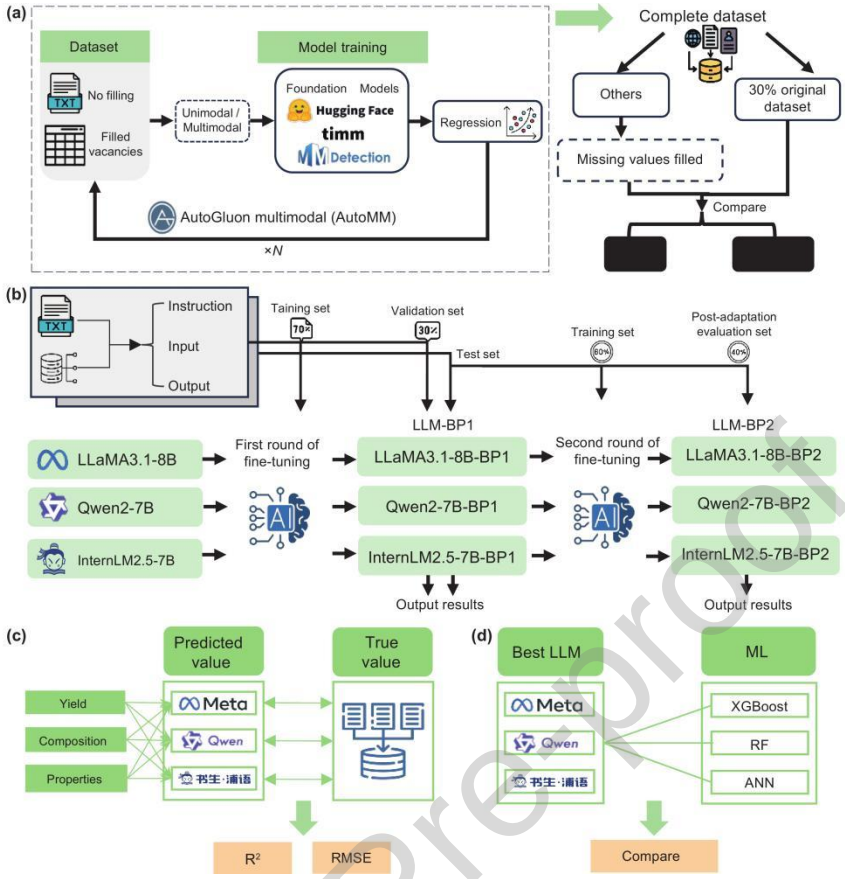


Fig. 1 Overall workflow of this study (a) Missing value imputation and evaluation of imputation effects (b) Two-stage fine-tuning process and data usage strategy (c) Evaluation method for the fine-tuning effects of large language models (d) Prediction performance comparison framework for different modeling paradigms (large language models vs. traditional machine learning).

2.1. Dataset Construction and Preprocessing

In this study, a comprehensive dataset for biochar was developed, encompassing multidimensional data describing the characteristics of the raw materials, pyrolysis conditions, and the resulting biochar properties (**Table S1**). The data were collected through a systematic survey of the literature and experimental records (**Table S3**), including key biochar attributes and corresponding experimental conditions (**Text S1**) such as (1) the basic composition of raw materials (*e.g.*, cellulose and lignin content); (2) pyrolysis parameters (*e.g.*, highest treatment temperature, heating rate, and

residence time); and (3) biochar performance indicators (*e.g.*, yield, specific surface area, and elemental composition). These features serve as the foundational data for the development of subsequent biochar property prediction models.

The dataset was designed for multimodal modeling based on literature-level heterogeneous information, with variable completeness, structural diversity, and feature standardizability as the primary construction principles. A total of 141 literature samples covering diverse feedstock types, pyrolysis processes, and performance indicators were selected to support a systematic evaluation of the proposed framework.

During the data preprocessing phase, the original dataset contained a significant amount of missing data, and addressing these gaps was crucial to ensure the reliability of the model. To enhance dataset completeness and consistency, we used a multimodal algorithm (AutoMM for Text) from the AutoGluon library (**Fig. 1(a)**; **Text S2**). This method not only considers numerical data but also integrates unstructured textual information such as raw material sources and preprocessing methods to fill in the missing values.

To avoid potential data leakage during the imputation process, a "split-first, impute-later" protocol was adopted. Prior to model training and evaluation, the samples were grouped according to their literature sources to ensure that data from the same reference were assigned exclusively to the same subset, thereby preventing cross-set contamination. The AutoMM imputation model was fitted using only the training subset, whereas the missing values in the test subset were predicted using the trained model without access to their true values. By using this procedure, the imputation and model

evaluation remained strictly independent at the data level, effectively reducing both sample- and feature-level information leakage. Subsequently, the feature columns were filled sequentially, and multiple iterations were conducted to update the input feature matrix by leveraging the latent correlations among the variables (**Text S3**) to restore the overall data structure and original characteristics as much as possible.

To evaluate the reliability of the imputation process, a stringent validation procedure was designed (**Fig. 1(a)**). Initially, 30% of the original data were randomly removed, and the missing data were subsequently filled using the imputed values. The imputation effectiveness was verified by comparing the data before and after imputation, confirming the validity of the imputed results.

2.2. Multiple Rounds of Fine-Tuning of Large Language Models

Three widely used LLMs were employed: InternLM2.5-7B [26], LLaMA3.1-8B [27], and Qwen2.5-7B [28]. These models are based on advanced generative pretraining architectures, demonstrate robust language understanding and generation capabilities, and have achieved notable success in various natural language processing tasks (**Text S4**). However, for domain-specific tasks, such as biochar property prediction, further adaptation using domain-relevant data is still required. Therefore, the models were fine-tuned to enhance their performance for prediction of biochar properties (**Fig. 1(b)**).

Fine-tuning involves further training a LLM on a task-specific dataset, which enables it to retain general language knowledge while gradually learning the relationships between domain-specific semantics and target properties, thereby

improving its modeling capability. The fine-tuning process was conducted in two phases. The first phase utilized a training set containing raw materials, features, and target attributes. Supervised fine-tuning techniques [29] were applied to adjust the model parameters and enhance the ability of the models to predict biochar properties. During this phase, a separate test set was held in reserve and fully isolated from the other data to evaluate the performance of the model on unseen data (**Fig. 1(c)**). The remaining dataset was randomly split into 70% training set and 30% validation set to ensure the reliability of the model. Each sample was formatted into specific instructions based on the prediction target (*e.g.*, yield, specific surface area) (**Text S5, Fig. 2**) to guide the model in generating accurate predictions.

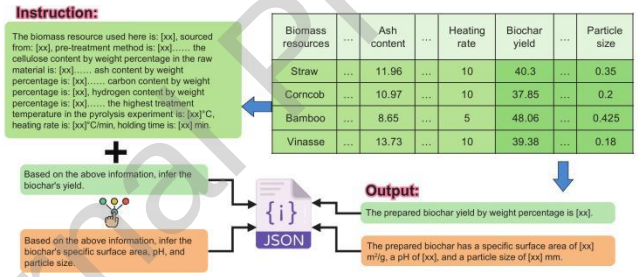


Fig. 2 The process of generating question–answer pairs from the dataset through a specific instruction-based format.

To further characterize the performance evolution of the model under a continuous data accumulation scenario, a second round of fine-tuning was introduced based on the first round. At this stage, the independent test set held in reserve during the first round was treated as a new data source from which 60% of the samples were randomly selected as additional training data and the remaining 40% were used to characterize the changes in the predictive performance after the incorporation of new samples. This setting was designed to simulate a continuous data supplementation

process and analyze the adaptability of the model under gradually expanding data distributions. Under this design, the second round of fine-tuning focused on capturing the parameter adaptation behavior of the model as new in-distribution samples became available, thereby mimicking the updating process of models in practical deployment environments. The detailed fine-tuning methods and parameter settings are described in **Text S6**.

2.3. Traditional Machine Learning

To compare the performance of the fine-tuned LLMs with that of the traditional machine learning methods in predicting biochar properties (**Fig. 1(d)**), we selected three representative machine learning algorithms: XGBoost [30], random forest (RF) [31], and ANN [32] (**Text S7**). These algorithms represent the approaches of ensemble learning, decision trees, and deep learning models, respectively, and have been widely used and demonstrated to be highly effective in chemistry and environmental science, particularly for handling data with strong correlations and complex features (for details on our parameter settings and tuning strategies, see **Text S8** and **Table S2**).

In this study, traditional machine learning models were primarily treated as structured numerical-data-based baseline models. Only numerical input features such as feedstock composition, pyrolysis temperature, and residence time were used. Each model was trained to predict a single biochar property at a time, with the predicted properties including yield, ash content, specific surface area, pH, and particle size.

It should be noted that in principle, RF, XGBoost, and ANN can be extended to multi-output regression or multi-task learning. However, in this work, they were

deliberately configured in a “single-target, numerical-only” setting to establish a standard baseline based solely on structured experimental parameters. This design allows for a clear comparison between different modeling paradigms without introducing additional feature encodings or architectural modifications.

Accordingly, under this experimental configuration, traditional machine learning models were restricted to single-target regression using structured numerical features only and did not incorporate any textual information such as feedstock origin or pretreatment descriptions. In this setting, these models do not explicitly capture the potential correlations among multiple target properties and cannot directly utilize the semantic information contained in the text. By contrast, LLMs can integrate numerical process parameters and textual descriptions within a unified prompting framework, enabling multimodal and multi-output joint prediction, and thus offer an alternative modeling paradigm for complex biochar systems.

The training process for traditional machine learning models was carried out in two phases. In the first phase, we used the same training, validation, and test sets as those used for the fine-tuning of the LLMs but only considered numerical data (such as raw material composition and pyrolysis conditions) and excluded unstructured textual data (such as raw material sources and preprocessing methods).

The dataset was reorganized during the second phase. The training set included the numerical data used in both the first and second rounds of the fine-tuning of the LLMs, whereas the test set consisted only of the data used for the post-adaptation evaluation. This difference arises from the fact that under the implementation

framework adopted in this study, traditional machine learning models do not support cross-round parameter inheritance. Unlike LLMs, they cannot incrementally update their parameters based on prior training and therefore must be retrained from scratch in each round using different data compositions.

Through this comparative framework, the performance of traditional machine learning models and LLMs can be systematically evaluated under different modeling paradigms, providing methodological insights for the selection and optimization of biochar property prediction models.

2.4. Comparative Evaluation

The coefficient of determination (R^2) and root mean square error (RMSE) were used to evaluate the predictive performance of the model [33]. These metrics are commonly employed to assess the performance of regression models and effectively reflect their adaptability and accuracy in the prediction of biochar properties.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

In this study, model performance was characterized from two complementary perspectives. The first is based on the predictive results obtained from the independent data split after the first round of fine-tuning, which reflects the model's performance under an initial, relatively static data scenario. The second is derived from the data expansion setting constructed during the second round of fine-tuning, which is used to analyze how the model performance evolves as new samples are introduced. Together,

these two evaluation settings correspond to the model behavior under static conditions and under continuous data extension, respectively.

3. RESULTS AND DISCUSSION

3.1. Data Imputation Results

Missing data is a key challenge in biochar research, limiting model training and the reliability of results, particularly when dealing with multidimensional datasets such as raw material characteristics, pyrolysis conditions, and biochar properties. The causes of missing data vary and include inconsistencies in experimental records, differences in measurement standards, and limitations in sample selection. For instance, some reports lack detailed descriptions of raw material sources and preprocessing methods, whereas others suffer from incomplete measurements of elemental content and pyrolysis conditions, owing to the use of different equipment and technologies. To address these missing values, we employed a multimodal algorithm from the AutoGluon library (AutoMM for Text), which integrates numerical data and textual information to fill in missing values, thereby significantly improving data completeness.

Evaluation of the imputation results revealed that AutoMM for Text exhibited high accuracy in restoring missing values with an average R^2 value of 0.8863, indicating that the model effectively recovered the underlying data patterns (**Fig. 3(a), Table S4**). For the raw material composition (*e.g.*, ash and sodium content) and biochar properties (*e.g.*, specific surface area and particle size), the imputed values closely matched the actual measurements because of the strong regularity of these features under different materials and pyrolysis conditions. For these properties, the R^2 value approached 1 and

the RMSE was low, suggesting that the imputation was successful. However, the imputation was less accurate for certain key indicators such as the highest treatment temperature and biochar yield, mainly due to the complexity of these features, which are strongly influenced by the experimental conditions. The differences between the experiments made it difficult to recover the original variation, leading to a higher RMSE.

Further statistical analysis showed that the mean, median, and standard deviation of the imputed data closely matched those of the original data, confirming that imputation successfully restored the global data characteristics (**Fig. 3(b)**). For example, the imputation of cellulose and fixed carbon content yielded mean values that were consistent with the original data, and the standard deviation was reduced, indicating that the method restored data regularity while reducing variability, thus validating the accuracy of the imputation. By contrast, the imputation of certain elements, such as potassium and iron, showed fluctuations in the mean values, likely owing to the diverse raw material types and pyrolysis conditions, reflecting local uncertainties in the imputation process.

Overall, AutoMM for Text was effective in leveraging the correlation between numerical data and textual information to restore most of the missing values in the biochar dataset. Although some features still exhibited imputation errors, the resulting data provided a more complete and reliable foundation for the subsequent model training, strengthening the data support for improving the accuracy of the biochar property predictions.

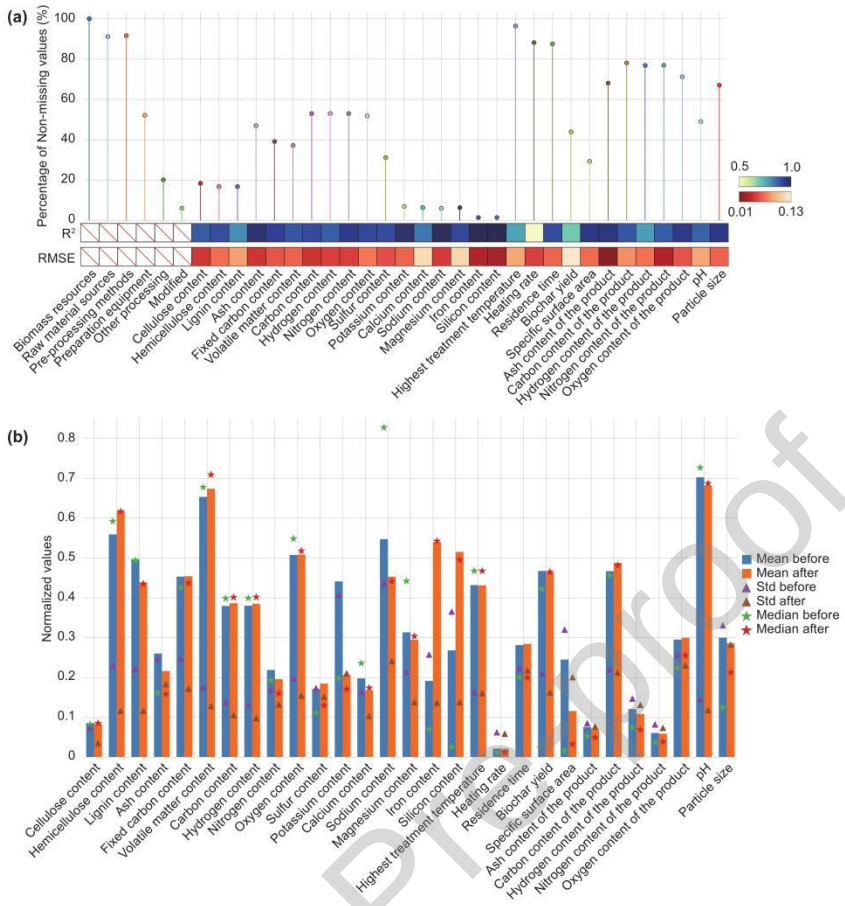


Fig. 3 Analysis of imputation effectiveness (a) Proportion of non-null values in the original dataset and the heatmap of R^2 and RMSE for the evaluation of the imputation effectiveness (b) Mean, standard deviation, and median before and after the imputation of missing values (for each variable, the data were normalized separately prior to imputation).

3.2. Results of the First Round of Fine-Tuning

Using the multimodal imputed dataset, we converted the instructions described in Section 2.2 to generate the corresponding biochar prediction instruction dataset. Three widely used LLMs (InternLM2.5-7B, LLaMA3.1-8B, and Qwen2-7B) were selected for the first round of fine-tuning to enhance their performance in predicting the biochar properties. The fine-tuning was necessary because biochar is a multifunctional

material whose properties are influenced by various factors, including raw material type, pyrolysis conditions, and processing methods.

During the training process, we analyzed four key performance indicators: training loss, training runtime, samples processed per second, and training steps per second (**Fig. S1(a–d), Text S9, Table S5**). The results revealed that all models successfully converged after approximately 1,250 steps (**Fig. 4(b)**). Although Qwen2-7B-BP1 (BiocharPrediction-v1) exhibited the lowest training loss (0.589), its longer training time resulted in a higher computational cost. By contrast, LLaMA3.1-8B-BP1 (BiocharPrediction-v1) demonstrated greater training efficiency with a slightly higher training loss but also the shortest runtime, indicating superior resource utilization. InternLM2.5-7B-BP1 (BiocharPrediction-v1) achieved an optimal balance between training loss and training speed.

Additionally, we comprehensively evaluated the predictive performance of the models. The three models performed reasonably well on the overall validation set (**Fig. 4(a)**), with InternLM2.5-7B-BP1 producing the most accurate predictions. InternLM2.5-7B-BP1 achieved an R^2 value of 0.9357 and an RMSE of 62.69 on the validation set, demonstrating strong fitting ability. However, its performance on the test set declined significantly, with the R^2 value decreasing to below zero and the RMSE increasing, indicating poor generalization to new data. This phenomenon may be attributed to overfitting during fine-tuning where the model overfitted to the specific patterns in the training data, resulting in erroneous predictions (referred to as "hallucinations") when exposed to data that differ from the training set. This issue is

particularly relevant in biochar prediction tasks where the diversity of biochar properties and the influence of multiple factors make it difficult for models to adapt to unseen data.

Further analysis showed that the three LLMs demonstrated high fitting accuracy when predicting five key biochar properties (yield, specific surface area, ash content, pH, and particle size) in the validation set (**Fig. 4(d–h)**). These properties are crucial for the environmental benefits, economic viability, and applications of biochar. The regression slopes were close to 1, indicating a strong linear relationship between the predicted and actual values achieved due to effective capture of the complex correlations between these properties. Among the models, InternLM2.5-7B-BP1 showed excellent performance in predicting multiple properties (yield and pH), with the highest average R^2 of 0.9088. Although Qwen2-7B-BP1 showed slight advantages in predicting specific surface area and particle size, its overall performance was less stable than that of InternLM2.5-7B-BP1.

Prediction speed is another important metric, particularly in biochar production optimization and large-scale applications where fast predictions are crucial for real-time decision-making. According to the prediction speed results (**Fig. 4(a)**), LLaMA3.1-8B-BP1 showed higher prediction speed for both the validation and test sets, processing 1.044 and 1.282 samples per second, respectively; these values were slightly higher than those for InternLM2.5-7B-BP1 (0.781 and 0.9 samples per respectively). However, considering the balance between the prediction accuracy and practical application

requirements, InternLM2.5-7B-BP1, is the optimal model for predicting biochar properties owing to its excellent fitting ability and reasonable prediction speed.

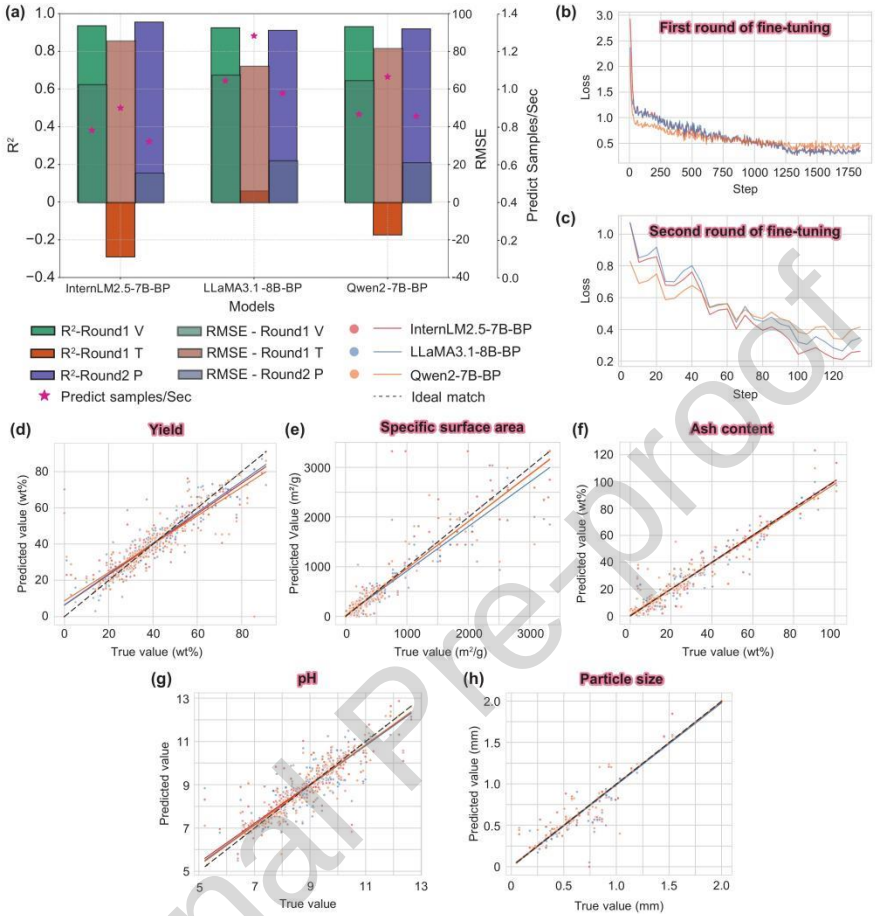


Fig. 4 Analysis of model adaptability under multi-round fine-tuning (a) R^2, RMSE, and average prediction speed for InternLM2.5-7B-BP, LLaMA3.1-8B-BP, and Qwen2-7B-BP at the stages of first-round validation, first-round test, and post-adaptation evaluation (b) Evolution of the training loss during the first round of fine-tuning for InternLM2.5-7B-BP1, LLaMA3.1-8B-BP1, and Qwen2-7B-BP1 (c) Evolution of the training loss during the second round of fine-tuning for InternLM2.5-7B-BP2, LLaMA3.1-8B-BP2, and Qwen2-7B-BP2 (d–h) Comparison between predicted and true values for the five target variables (yield, specific surface area, ash content, pH value, and particle size) and corresponding linear fitting for the first-round validation set for InternLM2.5-7B-BP1, LLaMA3.1-8B-BP1, and Qwen2-7B-BP1.

3.3. Results of the Second Round of Fine-Tuning

After the first round of fine-tuning, although the models performed well on the validation set, their generalization ability remained insufficient when faced with a new test set, showing signs of overfitting. Therefore, to maintain a consistent overall data distribution, a second round of fine-tuning was introduced to characterize the adaptive evolution of the model under a continuous sample supplementation scenario. The core goal of this round was to optimize the adaptability of the model for biochar property prediction, particularly under gradually expanding but in-distribution data conditions.

In the second fine-tuning round, we re-partitioned the data from the first evaluation stage using 60% of the samples for parameter updating and the remaining 40% for post-adaptation performance evaluation. This stage focused on the evolution of the model's performance under continuous in-distribution sample supplementation rather than independent extrapolation to unseen distributions. After the second round of fine-tuning, the model performance in the post-adaptation evaluation stage improved significantly (**Fig. 4(a)**). Specifically, the R^2 value for InternLM2.5-7B-BP2 increased from −0.2907 in the first round to 0.9552, and the RMSE dropped from 85.85 to 15.80, indicating that the model formed a more stable and accurate predictive map under continuous sample expansion. By contrast, LLaMA3.1-8B-BP2 and Qwen2-7B-BP2 exhibited relatively small improvements. These results indicate that multi-round parameter updating helps the model maintain high prediction consistency under gradually expanding but similar data distributions. This outcome indicated that the second fine-tuning round improved both the fitting accuracy and adaptability of the

model within a gradually expanded but in-distribution data regime. In biochar property prediction, characteristics such as yield and specific surface area are influenced by multiple factors such as the raw material type, pyrolysis temperature, and heating rate. Because of the complex relationships between these variables, the models are prone to overfitting or overreliance on specific features from the training set. By introducing additional training samples, the second fine-tuning round enabled the model to learn a broader feature distribution, alleviating the overfitting tendency observed after single-round training and improving prediction accuracy. Moreover, the robust performance of the fine-tuned InternLM2.5-7B model is consistent with previous findings in the chemical domain where its efficacy in understanding and processing complex chemical knowledge was demonstrated in the ChemLLM study by Zhang *et al.* (2024) [34].

During the second fine-tuning round, we analyzed four key performance indicators: training loss, training runtime, number of samples processed per second, and number of training steps per second (**Fig. 4(c), Fig. S1(e–h)**). All models exhibited a decrease in the training loss, with InternLM2.5-7B-BP2 showing the best performance, with a training loss of 0.4946, which was significantly lower than those of LLaMA3.1-8B-BP2 (0.5446) and Qwen2-7B-BP2 (0.5247). This indicates that InternLM2.5-7B-BP2 was more effective at fitting the training data, supporting its high accuracy for subsequent predictions. Although the training runtimes varied among the three models, their computational costs were similar, suggesting similar efficiencies for all models when processing comparable datasets. The analysis of the samples processed per second and training steps per second showed that all models maintained high processing

efficiency and training stability during the second fine-tuning round, further validating the effectiveness of the incremental learning strategy.

Although the prediction accuracy improved, the prediction speed of the models decreased (**Fig. 4(a)**). Specifically, InternLM2.5-7B-BP2's prediction speed decreased from 0.9 samples per second in the first round to 0.721 samples per second, with LLaMA3.1-8B-BP2 and Qwen2-7B-BP2 also showing lower speeds compared to those achieved in the first round. This is likely due to the parameter adjustments and an increase in the model complexity during the second fine-tuning round. Nevertheless, the prediction speed remained within a reasonable range that is sufficient for most real-time biochar production prediction tasks.

In summary, the second fine-tuning round significantly enhanced model performance under continuous sample supplementation conditions, particularly for InternLM2.5-7B-BP2. Although the prediction speed decreased slightly, the improved accuracy provides a higher practical value in application scenarios. InternLM2.5-7B-BP2 demonstrated strong adaptability and prediction consistency under progressive parameter optimization, providing a reliable technical basis for further model development and application.

3.4. Comparison of Machine Learning Results

To examine the advantages of fine-tuned LLMs in biochar property prediction, we compared them with three traditional machine learning algorithms: XGBoost, RF, and ANN. These models are commonly used in environmental and chemical data prediction and demonstrate strong expressiveness and stable performance for

regression tasks involving a single target variable. By comparing LLMs with traditional machine learning methods, this study highlights the differences between the two modeling paradigms in multi-dimensional property prediction, focusing on prediction consistency and model adaptability under distribution-consistent data expansion scenarios.

To predict biochar properties, we selected five key characteristics: yield, ash content, specific surface area, pH, and particle size. The traditional machine learning models performed relatively well on the training and validation sets (**Fig. 5(a), Fig. S2**), particularly for properties such as ash content and specific surface area, for which XGBoost and RF showed strong fitting ability. However, these models exhibited a significant drop in performance on the test set, particularly for the prediction of yield and particle size, with R^2 values decreasing and RMSE increasing, indicating inadequate generalization to new data. Although these models are computationally efficient, their reliance on single-target regression restricts their ability to capture the complex interrelationships among multiple biochar properties.

For each of the five key biochar properties, we selected the best-performing results among the three traditional machine learning models. In practice, XGBoost consistently achieved the highest predictive accuracy across all five properties and was therefore adopted as the representative conventional benchmark for the subsequent comparisons. We then compared these benchmark results with the results obtained by the fine-tuned InternLM2.5-7B-BP2. This analysis revealed that InternLM2.5-7B-BP2 achieved an average improvement in the prediction accuracy of approximately 16.33%

across all five features (**Fig. 5(b–e)**). InternLM2.5-7B-BP2 demonstrated a significant improvement in the R^2 values for the post-adaptation evaluation set, reaching 0.88697 for biochar yield and 0.9999 for particle size, demonstrating its ability to better understand the relationships between complex features. Under the current data scale and multimodal input conditions, this improvement is likely attributable to the semantic representations learned during pre-training as well as the joint modeling of structured features and textual information. During fine-tuning, this mechanism enabled the model to more effectively capture the latent relationships among the feedstock properties, pyrolysis parameters, and biochar characteristics. For instance, the biochar yield is closely related to the raw material composition and pyrolysis conditions. The ability of the fine-tuned model to jointly model these features enhanced its predictive accuracy.

However, for the specific surface area feature, the prediction results of InternLM2.5-7B-BP2 showed a slight drop in accuracy compared with the best results among the machine learning algorithms that were obtained by XGBoost, with a of 2.449% decrease in prediction precision. This result may be attributed to the unique characteristics of the specific surface area, which is highly sensitive to variations in the raw material composition and pyrolysis conditions. This suggests that although LLMs can handle multidimensional data, they may not always outperform traditional methods in predicting certain features, particularly when large differences exist in complex data characteristics. Thus, although LLMs show strong advantages in multidimensional prediction tasks, further optimization of their structure and parameters may be necessary for specific tasks or features.

Further analysis revealed that traditional machine learning models are usually limited to single-target prediction tasks and do not capture the complex interrelationships among biochar characteristics, which limits their effectiveness in handling multi-dimensional prediction tasks. By contrast, LLMs can jointly model multiple characteristics and consider their potential relationships, thereby improving prediction accuracy. Additionally, traditional models cannot incorporate unstructured textual data such as raw material sources or preprocessing methods, whereas LLMs integrate this information to enhance the overall understanding of biochar property prediction.

Although InternLM2.5-7B-BP2 demonstrated significant improvements in prediction accuracy, its training process is complex and computationally demanding. Traditional machine learning models exhibit high efficiency when working with smaller datasets or simpler feature relationships. However, LLMs offer clear advantages when dealing with complex multifactorial data such as biochar.

After fine-tuning, the LLMs demonstrated clear advantages over traditional machine learning algorithms in multidimensional feature prediction tasks, particularly in terms of prediction consistency and adaptive modeling capability. InternLM2.5-7B-BP2 achieved an average improvement of approximately 16.33% in the prediction accuracy across multiple tasks, showing better adaptability and accuracy when handling complex data. This makes it a more accurate and reliable tool for real-time prediction of biochar production.

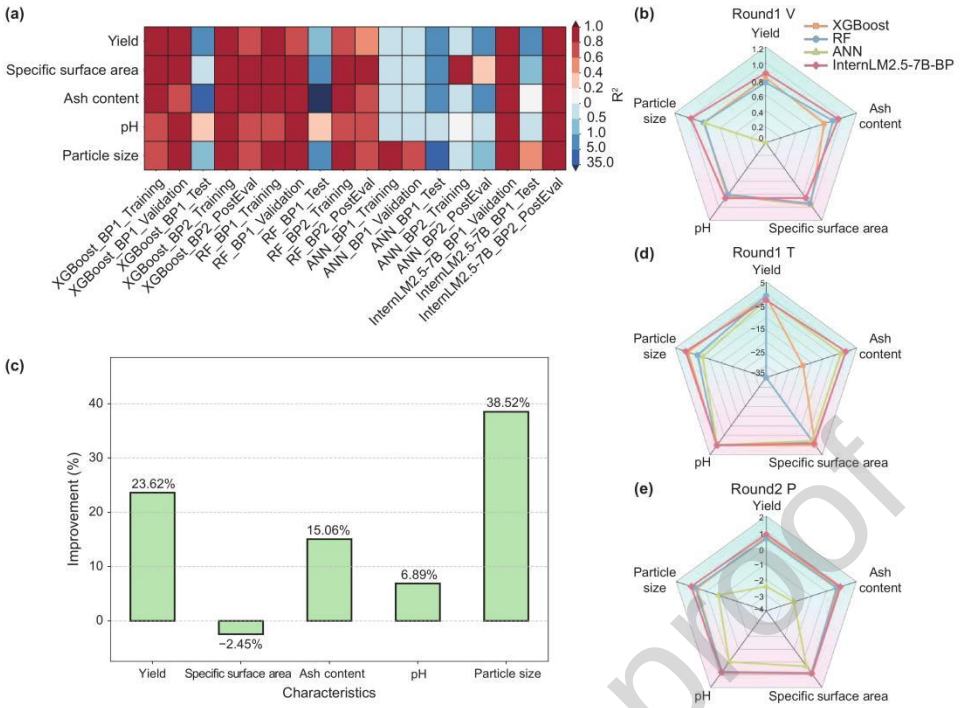


Fig. 5 Comparison of prediction results for fine-tuned large models and traditional machine learning algorithms (a) Heatmap of R^2 prediction results for yield, ash content, specific surface area, pH, and particle size at each stage for XGBoost, RF, ANN, InternLM2.5-7B-BP, LLaMA3.1-8B-BP, and Qwen2-7B-BP (b) Improvement in the prediction accuracy for the best results of fine-tuned large language models compared to traditional machine learning models for the five selected characteristics (c–e) Radar charts of R^2 prediction results for yield, ash content, specific surface area, pH, and particle size at the stages of first-round validation, first-round test, and post-adaptation evaluation for XGBoost, RF, ANN, InternLM2.5-7B-BP, LLaMA3.1-8B-BP, and Qwen2-7B-BP.

3.5. Limitations and Future Work

Although this study successfully demonstrates the potential of a multimodal data fusion framework based on LLMs for biochar characteristic prediction, several limitations remain. First, the process of converting structured tabular data into question–answer pairs for fine-tuning requires careful design to ensure the accurate conveyance of key information. This approach is particularly challenging when dealing

with complex datasets, because the design of prompts directly affects the model's ability to understand the relationships between features. Improperly designed instructions may hinder a model's ability to capture these relationships, thereby limiting its predictive performance. Second, although fine-tuning can enhance the performance of LLMs, their architectures are not specifically tailored to handle biochar-related data. Even though language understanding is the core strength of this type of models, this advantage is constrained by the architecture of the underlying model. Additionally, the model incurs a high computational cost, particularly when processing high-dimensional and complex datasets, creating a bottleneck for practical applications. Moreover, because the framework was constructed on literature-derived data, the current results primarily reflect statistical associations within the existing data system rather than direct causal or engineering-level validation. Under the present data scale and information representation, multimodal semantic modeling provides an effective representation of the relationships between process parameters and product properties; however, further verification of model stability and generalizability is still required.

To address these limitations, future studies should focus on several key research areas. First, the design of the instructions and data conversion processes should be optimized to enhance the model's ability to interpret and process complex datasets. By improving the precision of question–answer pair generation, we can ensure that the model fully utilizes the correlations within tabular data, thereby improving the prediction accuracy. Second, there is a need to develop more specialized model architectures that are better suited to handle the integration of both tabular data and

textual information and to further explore more flexible multimodal fusion and multi-target modeling mechanisms. Introduction of task-specific attention mechanisms tailored to the characteristics of biochar data can further strengthen the ability of the model to process multidimensional data. Finally, future studies should aim to improve the computational efficiency of the model and reduce its reliance on large computational resources. Through further optimization of the model architecture, the feasibility of deployment of LLMs for real-world biochar prediction tasks can be markedly enhanced. In addition, future work may consider selection of representative raw materials and operating conditions and combining larger-scale or cross-database samples with experimental measurements or authoritative datasets to further evaluate the robustness and transferability of the proposed framework.

4. CONCLUSION

This study introduces a multimodal data fusion framework based on LLMs to enhance the accuracy of biochar characteristic predictions. An LLM for biochar prediction was developed by integrating structured experimental data with unstructured textual data. First, the multimodal algorithm from AutoGluon was employed to impute the missing values in the dataset. The imputed data showed high accuracy across various features, with an average R^2 value of 0.8863, confirming that the imputation process successfully restored the regularity of the data and provided a robust foundation for the subsequent modeling. Next, the data were converted into question–answer pairs using a specific instruction format to fine-tune the LLM.

In the first round of fine-tuning, all three LLMs exhibited good predictive performance, with InternLM2.5-7B-BP1 in particular achieving an R^2 value of 0.9357 and an RMSE of 62.69 on the validation set, demonstrating strong fitting capabilities. However, the performance of the model on the test set declined significantly, with the R^2 value decreasing to a negative value and the RMSE increasing, indicating poor generalization to new data. To address this issue, we conducted a second round of fine-tuning. Under a continuous sample supplementation scenario, this incremental updating strategy led to substantial improvements across all models, particularly for InternLM2.5-7B-BP2 for which R^2 value increased from −0.2907 in the first round to 0.9552 in the second round, while the RMSE decreased from 85.85 to 15.80. These results indicate that the model can progressively optimize its parameter mapping and achieve more stable and accurate predictions when new samples are introduced.

Finally, compared with traditional machine learning methods, the fine-tuned LLM InternLM2.5-7B-BP2 demonstrated clear advantages in biochar property prediction tasks, showing an average accuracy improvement of 16.33% across multiple properties. Under the multimodal input setting and modeling framework adopted in this study, these results suggest that LLMs possess strong adaptability and representation capabilities for multi-target and multi-property prediction tasks.

In conclusion, this study presents a novel multimodal data fusion framework that leverages LLMs to enhance the prediction of biochar properties. The insights derived from our work not only advance theoretical understanding but also provide practical guidance for optimizing biochar performance prediction. Moreover, the

versatility of our approach, which integrates structured experimental data with unstructured textual information, makes it highly adaptable to a wide array of applications beyond biochar research. This methodology may lead to transformative impacts in diverse fields—such as soil remediation, sustainable agriculture, energy storage, and materials science—where complex, multifaceted data integration is essential. Ultimately, our work establishes a new paradigm in data-driven predictive analytics, setting the stage for future research efforts that harness the power of LLMs to address challenging problems across a wide range of environmental science and industrial applications.

ASSOCIATED CONTENT

Data Availability Statement

The code and data underlying this study are openly available on Github at https://github.com/SinceraXY/LLMs-BiocharPredict.git

The model weights after the first and second rounds of fine-tuning are publicly accessible at https://huggingface.co/collections/SinceraXY/llms-biocharpredict.

Supporting Information

The Supporting Information (SI) provides detailed information about the dataset preprocessing, model selection, and evaluation methodologies used in this study. It includes descriptions of the data collection and integration protocols (**Text S1**), the imputation process with AutoMM for Text (**Text S2** and **S3**), fine-tuning procedures for

LLMs (**Text S4–S6**), and a comparison with traditional machine learning algorithms, including relevant parameter settings (**Text S7** and **S8**). Additionally, the SI presents the performance metrics from the fine-tuning process (**Text S9**), with specific performance indicators for the two fine-tuning rounds (**Fig. S1**), and a heatmap of the RMSE prediction results for the biochar properties (**Fig. S2**). **Tables S1** and **S3** summarize the dataset variables and data sources. **Table S2** outlines the parameter settings for conventional machine learning models. **Table S4** reports the missing rates and imputation performance (R^2 and RMSE) of the key variables. The training cost and inference efficiency of large language models are summarized in **Table S5a–b**. **Table S6** lists all abbreviations used in the manuscript and Supporting Information. This comprehensive presentation ensured the transparency and reproducibility of the methodology and results of the study.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 22278435) and the Carbon Neutrality Research Institute Fund (Nos. 20220202 and 20230102).

REFERENCES

(1) F.W. Cheng, H.X. Luo, L.M. Colosi, Slow pyrolysis as a platform for negative emissions technology: an integration of machine learning models, life cycle assessment, and economic analysis, Energy Convers. Manage. 223 (2020) 113258.

(2) M. Marcińczyk, P. Oleszczuk, Biochar and engineered biochar as slow- and controlled-release fertilizers, J. Cleaner Prod. 339 (2022) 130685.

(3) Y. Lu, K. Gu, Z.T. Shen, C.S. Tang, B. Shi, Q.Y. Zhou, Biochar implications for the engineering

properties of soils: a review, Sci. Total Environ. 888 (2023) 164185.
(4) R.K. Mishra, K. Mohanty, A review of the next-generation biochar production from waste biomass for material applications, Sci. Total Environ. 904 (2023) 167171.
(5) B. Sajjadi, W.Y. Chen, N.O. Egiebor, A comprehensive review on physical activation of biochar for energy and environmental applications, Rev. Chem. Eng. 35 (2019) 735-776.
(6) S.P.C. Gonçalves, M. Strauss, D.S.T. Martinez, The positive fate of biochar addition to soil in the degradation of PHBV-silver nanoparticle composites, Environ. Sci. Technol. 52 (2018) 13845-13853.
(7) Z.J. Wang, P.Z. Sun, Y.X. Li, T. Meng, Z.P. Li, X. Zhang, R.C. Zhang, H.Z. Jia, H. Yao, Reactive nitrogen species mediated degradation of estrogenic disrupting chemicals by biochar/monochloramine in buffered water and synthetic hydrolyzed urine, Environ. Sci. Technol. 53 (2019) 12688-12696.
(8) L. Luo, J.X. Wang, J.T. Lv, Z.G. Liu, T.R. Sun, Y. Yang, Y.G. Zhu, Carbon sequestration strategies in soil using biochar: advances, challenges, and opportunities, Environ. Sci. Technol. 57 (2023) 11357-11372.
(9) N. Bolan, S.A. Hoang, J.Z. Beiyuan, S. Gupta, D.Y. Hou, A. Karakoti, S. Joseph, S. Jung, K.H. Kim, M.B. Kirkham, H.W. Kua, M. Kumar, E.E. Kwon, Y.S. Ok, V. Perera, J. Rinklebe, S.M. Shaheen, B. Sarkar, A.K. Sarmah, B.P. Singh, G. Singh, D.C.W. Tsang, K. Vikrant, M. Vithanage, A. Vinu, H.L. Wang, H. Wijesekara, Y.B. Yan, S.A. Younis, L. Van Zwieten, Multifunctional applications of biochar beyond carbon storage, Int. Mater. Rev. 67 (2022) 150-200.
(10) Y.B. Jiang, C. Ren, H.Y. Guo, M.X. Guo, W. Li, Speciation transformation of phosphorus in poultry litter during pyrolysis: insights from X-ray diffraction, Fourier transform infrared, and solid-state NMR spectroscopy, Environ. Sci. Technol. 53 (2019) 13841-13849. DOI:10.1021/acs.est.9b03261.
(11) Y. Lee, J. Park, C. Ryu, K.S. Gang, W. Yang, Y.K. Park, J. Jung, S. Hyun, Comparison of biochar properties from biomass residues produced by slow pyrolysis at 500°C, Bioresour. Technol. 148 (2013) 196-201.
(12) W. Wu, S.S. Zhu, X.C. Huang, W. Wei, C. Jin, B.J. Ni, Determination of instinct components of biomass on the generation of persistent free radicals (PFRs) as critical redox sites in pyrogenic chars for persulfate activation, Environ. Sci. Technol. 55 (2021) 7690-7701.
(13) W.Y. Wang, J.C. Huang, T. Wu, X. Ren, X.S. Zhao, Research on the preparation of biochar from waste and its application in environmental remediation, Water 15 (2023) 3387.
(14) S. Adhikari, W. Timms, M.A.P. Mahmud, Optimising water holding capacity and hydrophobicity of biochar for soil amendment – A review, Sci. Total Environ. 851 (2022) 158043.
(15) S. Praveen, J. Jegan, T.B. Pushpa, R. Gokulan, L. Bulgariu, Biochar for removal of dyes in contaminated water: an overview, Biochar 4 (2022) 10.
(16) D.M. Dutton, G.V. Conroy, A review of machine learning, Knowl. Eng. Rev. 12 (1997) 341-367.
(17) Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436-444.
(18) X.Z. Zhu, Y.N. Li, X.N. Wang, Machine learning prediction of biochar yield and carbon contents in biochar based on biomass characteristics and pyrolysis conditions, Bioresour. Technol. 288 (2019) 121527.
(19) M. Khan, Z. Ullah, O. Mašek, S.R. Naqvi, M.N.A. Khan, Artificial neural networks for the prediction of biochar yield: a comparative study of metaheuristic algorithms, Bioresour. Technol. 355 (2022) 127215.

(20) A. Pathy, S. Meher, B. P, Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods, Algal Res. 50 (2020) 102006.
(21) K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547-555.
(22) M. Zhou, N. Duan, S.J. Liu, H.Y. Shum, Progress in neural NLP: modeling, learning, and reasoning, Engineering 6 (2020) 275-290.
(23) S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y.T. Lee, Y.Z. Li, S. Lundberg, H. Nori, H. Palangi, M.T. Ribeiro, Y. Zhang, Sparks of artificial general intelligence: early experiments with GPT-4, arXiv preprint arXiv:2303.12712 (2023).
(24) K. Khandelwal, S. Nanda, A.K. Dalai, Machine learning to predict the production of bio-oil, biogas, and biochar by pyrolysis of biomass: a review, Environ. Chem. Lett. 22 (2024) 2669-2698.
(25) J.A. Ippolito, L.Q. Cui, C. Kammann, N. Wrage-Mönnig, J.M. Estavillo, T. Fuertes-Mendizabal, M.L. Cayuela, G. Sigua, J. Novak, K. Spokas, N. Borchard, Feedstock choice, pyrolysis temperature and type influence biochar characteristics: a comprehensive meta-data analysis review, Biochar 2 (2020) 421-438.
(26) Z. Cai, M.S. Cao, H.J. Chen, K. Chen, K.Y. Chen, X. Chen, X. Chen, Z.H. Chen, Z. Chen, P. Chu, X.Y. Dong, H.D. Duan, Q. Fan, Z.Y. Fei, Y. Gao, J.Y. Ge, C.Y. Gu, Y.Z. Gu, T. Gui, A.J. Guo, Q.P. Guo, C.H. He, Y.F. Hu, T. Huang, T. Jiang, P.L. Jiao, Z.J. Jin, Z.K. Lei, J.X. Li, J.W. Li, L.Y. Li, S.B. Li, W. Li, Y.N. Li, H.W. Liu, J.N. Liu, J.W. Hong, K.W. Liu, K.K. Liu, X.R. Liu, C.Q. Lv, H.J. Lv, K. Lv, L. Ma, R.Y. Ma, Z.R. Ma, W.C. Ning, L.K. Ouyang, J.T. Qiu, Y. Qu, F.K. Shang, Y.F. Shao, D.M. Song, Z.F. Song, Z.H. Sui, P. Sun, Y. Sun, H.Z. Tang, B. Wang, G.T. Wang, J.Q. Wang, J.Y. Wang, R. Wang, Y.D. Wang, Z.Y. Wang, X.J. Wei, Q.Z. Weng, F. Wu, Y.T. Xiong, C. Xu, R.L. Xu, H. Yan, Y.R. Yan, X.G. Yang, H.C. Ye, H.Y. Ying, J. Yu, J. Yu, Y.H. Zang, C.Y. Zhang, L. Zhang, P. Zhang, P. Zhang, R.J. Zhang, S. Zhang, S.Y. Zhang, W.J. Zhang, W.W. Zhang, X.C. Zhang, X.Y. Zhang, H. Zhao, Q. Zhao, X.M. Zhao, F.Z. Zhou, Z.D. Zhou, J.M. Zhuo, Y.C. Zou, X.P. Qiu, Y. Qiao, D.H. Lin, InternLM2 technical report, arXiv preprint arXiv:2403.17297 (2024).
(27) H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
(28) Qwen Team, Qwen2.5 technical report, arXiv preprint arXiv:2412.15115 (2024).
(29) A. Pareja, N.S. Nayak, H. Wang, K. Killamsetty, S. Sudalairaj, W.L. Zhao, S. Han, A. Bhandwaldar, G.X. Xu, K. Xu, L.G. Han, L. Inglis, A. Srivastava, Unveiling the secret recipe: a guide for supervised fine-tuning small LLMs. Proceedings of the 13th International Conference on Learning Representations. Singapore, Singapore: OpenReview.net, 2025.
(30) T.Q. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, arXiv preprint arXiv:1603.02754 (2016).
(31) X.F. Hu, J.H. Belle, X. Meng, A. Wildani, L.A. Waller, M.J. Strickland, Y. Liu, Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach, Environ. Sci. Technol. 51 (2017) 6936-6944.
(32) J. Rodriguez-Perez, C. Leigh, B. Liquet, C. Kermorvant, E. Peterson, D. Sous, K. Mengersen, Detecting technical anomalies in high-frequency water-quality data using artificial neural networks, Environ. Sci. Technol. 54 (2020) 13719-13730.
(33) J.J. Ma, S. Zhang, X.J. Liu, J.Q. Wang, Machine learning prediction of biochar yield based on biomass characteristics, Bioresour. Technol. 389 (2023) 129820.

(34) D. Zhang, W. Liu, Q. Tan, J.D. Chen, H. Yan, Y.L. Yan, J.T. Li, W.R. Huang, X.Y. Yue, W.L. Ouyang, D.Z. Zhou, S.F. Zhang, M. Su, H.S. Zhong, Y.Q. Li, ChemLLM: a chemical large language model, arXiv preprint arXiv:2402.06852 (2024).

Highlights:

Introduces multimodal fusion of tabular and textual biochar data for prediction.

Uses AutoGluon text-informed imputation to recover missing data with R^2 of 0.8863.

Fine-tunes language model in two rounds to reach R^2 of 0.9552 on test data.

Achieves 16.33% higher accuracy than classic methods across five biochar properties.

Offers a framework for smart biochar optimization and broader environmental use.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: