International Journal of Machine Learning and Cybernetics (2025) 16:3681–3691
https://doi.org/10.1007/s13042-024-02473-0

ORIGINAL ARTICLE

CataLM: empowering catalyst design through large language models

Ludi Wang[1] · Xueqing Chen[1,2] · Yi Du[1,2,3] · Yuanchun Zhou[1,2,3] · Yang Gao[4] · Wenjuan Cui[1,2]

Received: 26 June 2024 / Accepted: 14 November 2024 / Published online: 15 January 2025
© The Author(s) 2024

Abstract
The field of catalysis holds paramount importance in shaping the trajectory of sustainable development, prompting intensive research efforts to leverage artificial intelligence (AI) in catalyst design. Presently, the fine-tuning of open-source large language models (LLMs) has yielded significant breakthroughs across various domains such as biology and healthcare. Drawing inspiration from these advancements, we introduce **CataLM** (**Cata**lytic **L**anguage **M**odel), a large language model tailored to the domain of electrocatalytic materials. Our findings demonstrate that **CataLM** exhibits remarkable potential for facilitating human-AI collaboration in catalyst knowledge exploration and design. To the best of our knowledge, **CataLM** stands as the pioneering LLM dedicated to the catalyst domain, offering novel avenues for catalyst discovery and development.

Keywords AI for science (AI4S) · Large language models (LLMs) · Electrocatalytic materials · Catalyst design

1 Introduction

The field of catalysis is crucial to the future of sustainable development. Innovative catalysts can generate clean fuels, reduce the impact of global warming and provide solutions to environmental pollution [1, 2]. Theoretical calculations and simulations can accelerate catalyst screening through activity descriptors that link structure to catalyst activity [3–5]. However, numerous variables exist in the synthesis, composition, structure, and performance of electrocatalysts, with much of this critical knowledge often elusive within scientific literature. This poses challenges in elucidating the intricate correlations from limited experimental data. Artificial intelligence can be used to extract, analyze and understand key information embedded in the vast scientific literature on catalysis that can be dedicated to predicting new catalysts. Natural language processing techniques and generative language models enable the extraction and analysis of textual information from scientific literature and the generation of domain-relevant on-demand text, which has shown potential in recent biological and thermoelectric research. However, language models in the field of catalysis are sparse and limited in scale, which restricts their use in empowering knowledge extraction in catalytic materials research and further enabling the discovery of new catalysts.

Pre-trained models have demonstrated their powerful capabilities in Natural Language Processing (NLP). There are two main types of pre-trained models: (1) BERT-like

Ludi Wang and Xueqing Chen have contributed equally to this work.

✉ Yang Gao
gaoyang@nanoctr.cn

✉ Wenjuan Cui
wenjuancui@cnic.cn

Ludi Wang
wld@cnic.cn

Xueqing Chen
xqchen@cnic.cn

Yi Du
duyi@cnic.cn

Yuanchun Zhou
zyc@cnic.cn

[1] Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] Hangzhou Institute for Advanced Study, UCAS, Hangzhou 310000, China

[4] CAS Key Laboratory of Nanosystem and Hierarchical Fabrication, National Center for Nanoscience and Technology (NCNST), Beijing 100190, China

models [6–8], which are mainly used for language comprehension tasks, and (2) GPT-like models [9–11], which are mainly used for language generation tasks. Currently, large-scale language models (LLMs) such as GPT−4.0 [12] have laid a solid foundation for various applications. Although current large language models are effective in general domains, they often fail to meet the needs of catalytic scientists. Much of this inadequacy is attributed to the lack of reliable knowledge about catalysts, as relevant catalyst structural features and performance analyses are rarely present in commonly used pre-trained text corpora, such as C4 [13] and the Pile [14]. Furthermore, the best performing large language models like ChatGPT are only served through APIs, which creates a barrier to research and progress in external domains. Fine-tuning open-source large language models is an effective way to meet domain-specific needs.

Currently, fine-tuning open-source large language models have reached considerable success in fields such as biology, healthcare, and finance. In biology, a domain-specific pre-trained Transformer language model, BiOGPT [15], has been developed for biomedical text generation and mining. The model can be optimised and enhanced for performance in tasks such as biological named entity tasks and protein molecular design. In healthcare, models like HuatuoGPT [16] and DoctorGLM [17] have been developed to address healthcare challenges, which exhibit a high degree of expertise and provide valuable insights into the healthcare domain. In recent years, researchers have utilized existing databases such as Atomly [18], OQMD [19], MaterialsProject [20] and others. They have successfully explored the complex relationship between material structure and properties [21], addressing the challenges posed by the scarcity of material data by developing more accurate AI optimisation [22] and training methods [23]. With the application of LLMs, materials science researchers have explored the use of these models to address challenges such as chemical reactions and the complex nature of structures. Examples include the MatSciBERT [24] model for the task of materials named entity recognition. MatSciBERT uses a large amount of materials science literature to fine-tune the BERT model [6], demonstrating the ability to automatically extract information from the literature, perform data mining, and construct knowledge graphs. MatChat [25] optimises the LLaMA2-7B model using knowledge of inorganic materials science literature and presents a viable solution for predicting chemical synthesis pathways of inorganic materials, opening up new possibilities for the use of language models in materials science. To the best of our knowledge, there has been no reported utilization of large language models in catalyst science so far.

In this work, we provide **CataLM**, a large language model aligned with knowledge in the field of electrocatalytic materials. This large language model takes advantage of the pre-trained Vicuna-13B model, and is trained on domain literature and data annotated by experts. With this extensive and diverse data, the original LLM is specialized with two phases: `Domain Pre-training`, where the model harvests the chemical knowledge from domain field literature, and `Instruction Tuning`, where the model further understands the requirements of downstream task with the annotation data. We use two tasks to validate **CataLM**, namely entity extraction task and control method recommendation task. In addition to using the constructed knowledge base for validation, we also invited domain experts to evaluate the answers of **CataLM** to verify its generalization ability. Results show that our large language model has potent potential for human-AI collaboration in catalyst knowledge search and design. To the best of our knowledge, **CataLM** is the first LLM that focus on the catalyst domain field, and we believe it can bring new possibilities for the preparation of new catalysts.

2 Related work

ChatGPT was selected as one of Nature's Top 10 Individuals of 2023, marking the unprecedented selection of a computer program-the first non-human entity in history-to receive such recognition. Nature states that this award aims to recognize the role of large language models (LLMs) in scientific development and progress. In the field of materials, numerous studies have utilized language models to address diverse tasks. Chen et al. provide the model MatChat [25], for predicting inorganic material synthesis pathways. Xie et al. [26] use FAIR database to fine-tune LLMs and design a downstream task named SII which aims to extract hierarchical, domain specific material and device information, such as composition, structure, preparation conditions, etc., from unstructured scientific texts. Zheng et al. [27] used prompt engineering to guide ChatGPT in the automation of text mining of metal-organic framework (MOF) synthesis conditions from diverse formats and styles of the scientific literature. InstructMol [28] adopts Vicuna to multiple chemical tasks with task-specific fine-tuning. Meta FAIR researchers launched the Open Materials 2024 (OMat24) dataset and large pre-trained models focusing on structural and compositional diversity. OMat24 dataset [29] is one of the largest publicly available datasets in material fields. Leong et al [30] explored a multimodal large language model (MLLM) based on scientific literature to analyze various data inputs for automated electrosynthetic reaction mining. However, previous works focus on the development of new materials instead of new catalyst designing. Considering the diversity of structural characteristics such as composition, crystal structure, and crystal plane of materials, potential catalysts are very abundant. Secondly, domain fine-tuning data sets

which are consistent with downstream applications are crucial for the capability migration of LLMs, which is lacking in the field of catalyst design. This deficiency results in the model's lack of catalyst knowledge, making it challenging to achieve satisfactory parameters.

To promote the creative utilization of large language models in catalysts science, this study utilizes a meticulously crafted database for question-answering to investigate their capabilities in the field of catalysts science. While building this model, we also refer to the successful experiences in the field of other science domains. For example, DeepGO-SE [31] tries to predict GO functions from protein sequences using a pretrained large language model. MedPaLM2 [32] and PMC-LLaMa [33] attempt to tailor LLMs specifically for the fields of biology and medicine through fine-tuning with domain-specific instructions. In addition, Chen et al construct the xTrimo model [34] based on massive protein sequence data training, achieving sota levels on multiple prediction tasks.

3 CataLM

As shown in Fig. 1, the training of **CataLM** consists of two stages, which are Domain Pre-training and Instruction Tuning respectively. Due to the lack of open-source corpora for recommending catalyst control methods, we utilized expert annotated corpora, as well as the retrieval enhanced corpora generated by large language models for training during the instruction fine-tuning stage.

3.1 Domain pre-training

In this work, the text corpus we used to further pre-train Vicuna-33b-v1.3, including the full text of open-access catalytic papers published in selected high-quality journals in the field of electrocatalytic science. We used Web of Science to find scientific literature on electrocatalytic CO_2 reduction. Specifically, we exported the metadata of more than 22,000 articles from Web of Science using the keywords "CO_2", "Reduction", and "Electro*" as subject indexes. Eventually, we used the full-text PDFs of 12,643 open-access papers to build the text corpus.

PDF parsing. We build an automatic PDF parsing toolkit based on the PyMuPDF library [35]. Since the processed documents contain irrelevant tags, we developed a data cleaning method for parsing article tag strings into consistently formatted text paragraphs while retaining the same section and paragraph structure as the original paper. Finally, we use regular expressions and rule-based scripts to clean the data, removing the text obstructing reading, garbled, and impurity data.

Vector database. Despite the fact that Language Model Models (LLMs) are capable of responding to broad inquiries, they are limited in their ability to provide in-depth, precise, and timely information within specific vertical domain. To tackle this issue, we have employed vector databases to augment the reasoning capabilities of LLMs in vertical domain contexts. Vector databases can transform literature and data into vector representations through the process of embedding vectors. For the establishment of vector

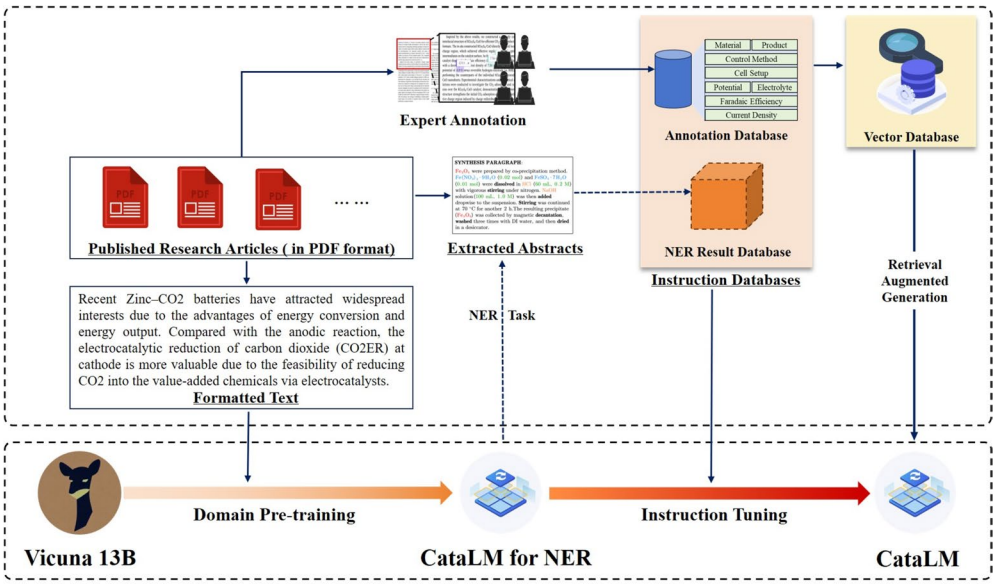

Fig. 1 The training pipeline of **CataLM**. The bottom part illustrates the primary training pipeline of CataLM, while the top part of the figure delineates the entire data preparation process for training

databases, Sci-BERT [36] has been utilized as an embedding model.

The study involved retrieving titles and abstracts from a dataset containing 12,643 documents, manually annotating catalytic reaction processes by domain experts, and then merging and converting these textual elements into vector representations using Sci-BERT as an embedding framework. In the context of a catalytic domain-specific task such as Name Entity Recognition (NER), the embedding model operates by converting the user query into vector form. Relevant articles are then identified by vector distance calculations to facilitate the retrieval of accurate and relevant information.

3.2 Instruction tuning

In order to align pre-trained models with domain user intent, we need to construct instruction tuning datasets. Currently available generic instruction tuning datasets such as Alpaca-GPT4 [37] and ToolBench [38] can only teach models to follow human instructions. For the specialized field of catalytic materials, we need to train models with knowledge-intensive data that can reflect domain knowledge. Considering the relatively small sample of data annotated by the experts, we use the large language model pre-trained in the previous section to expand it by automatically extracting abstracts from 12,643 documents. The entities were extracted based on an expert-constructed system of electrocatalytic reduction systems for literature content, including materials, conditioning methods, products, faradaic efficiency, cell setup, electrolyte, synthesis method, current density and voltage. The specific meanings and dataset formats of these entities can be found in the previous corpus construction work [39].

Firstly, we invite experts in the field of catalysis to perform manual annotation using a well-developed annotation tool, AutoDive [40]. This tool allows annotators to access material literature through a web browser, view sentences for annotation, and interact with predefined entity types and descriptions. Annotators have the flexibility to include new entities, rearrange existing ones, or make edits in a separate view. We end up with a standard corpus [39] in the field of electrocatalytic CO_2 reduction containing 6985 entities, with each record containing the entity extracted from the paper, its corresponding label, and the context sentence in which the entity is located. The standard corpus is provided as a file in CSV format, and the details are shown in Table 1.

Next, we used the pre-trained large language model based on vector database augmentation from the previous section to perform automatic extraction of literature abstracts in the field of catalysis, which extracts a total of 30,283 entities. It is important to highlight that the synthetic method of expert annotation in the dataset is an unstructured text paragraph description. We used a multi-model algorithm combining pattern recognition and neural networks to convert it into a structured synthetic pathway containing information about the prepared and target materials, synthetic operations and operating conditions. This structuring of information enhances the interpretability of domain knowledge by the expansive models.

Table 1 The summary of the standard corpus

Entity type	Benchmark Corpus
Material	1092
Control method	1086
Product (including the second and third product)	1340
Faradaic efficiency (including the Faradaic efficiency of second and third product)	1135
Cell setup	435
Electrolyte	475
Synthesis method	228
Current density	393
Voltage	801
Total	6985

After rigorous filtering, de-duplication and cleaning, we obtained a training set consisting of 13,432 highly reliable catalytic process descriptions, based on the combination of domain entities. Next, this dataset is further pre-processed and integrated into an instruction question-answering format. For example, for a certain catalytic reaction, using the entities provided in the dataset, we can reconstruct it as a recommendation task for catalyst preparation for a given product. As shown in Fig. 2, the prompt involves a specific catalyst material query for a given product, and the answer provides the recommended material and its preparation method.

3.3 Training process

The parameters of the model fine-tuning stage and application stage are listed in Table 2. In named entity recognition task, we set temperature as 0.1 to limit the divergence of CataLM, and we tried various temperature values to verify the recommendation ability of the CataLM. We use NVIDIA A100 GPUs for training, and techniques such as low-rank adaptation [41] is adopted to save storage memory and accelerate the process. Low Rank Adaptation of Large Language Models, also known as LoRA, is a technology developed by Microsoft researchers to address fine-tuning of large language models. The approach of LoRA is to freeze the pre-trained model weight parameters, and then inject trainable layers into each Transformer block. Since there is no need to recalculate the gradient of the model weight parameters, it greatly reduces the computational workload that needs to

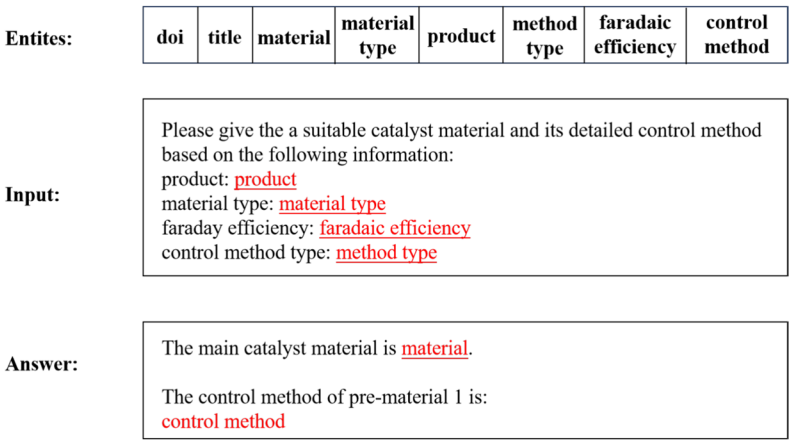

Fig. 2 Catalytic material recommended scenario's command format

Table 2 Parameter set

Parameter	Value
Fine-tuning stage	
Batch size	10
Learning rate	$3 * 10^{-4}$
Lora r	8
Lora alpha	32
Lora dropout	0.1
NER stage	
Temperature β	0.1
Top p	0.75
Top k	40
Beams	4
Max tokens	1248
Recommendation stage	
Temperature β	0.1/0.5
Top p	0.75
Top k	40
Beams	4
Max tokens	1248

be trained. Research has found that the fine-tuning quality of LoRA is comparable to that of full model fine-tuning, thus we chose this method in the training process of **CataLM**. Due to hardware limitations, we will batch size as 10, and the lora *r* was default setting in the work [41]. The alpha and dropout are set refer to PEFT in Huggingface library [42].

4 Evaluation

4.1 Named entity recognition task

The first task is named entity recognition, which aims to extract entity from the abstract of given literature. In this task, we use a dataset of 12,643 abstract from electrocatalytic scientific literature (the full text of these literature also be used in the fine-tuning of **CataLM**) for named entity recognition. We extracted eight types of entity labels, including material, control method, product, faradaic efficiency, cell setup, electrolyte, current density, and voltage. When performing entity recognition, the user first inputs the text to be extracted, and the embedding model transforms it into vectors. Then the similar articles will be obtained by calculating the vector distance, and will be used to generate precise and pertinent information, which be shown in Fig. 3. The prompt will be fed into the fine-tuned LLM for entity recognition.

For the evaluation and validation of the the entity extraction capability of **CataLM**, we randomly select 160 entries and validate the LLM's answers for them by experts, and ensure that each category has 20 test data. The evaluation result is shown in Table 3. The Count represents the total amount of samples from different categories, the Correct represents the number of correctly identified entities, and the Existence represents the number of entities of this type that do exist in the text input to the large language model. It is worth mentioning that if there is indeed no corresponding entity in the text input to the large language model, the situation where the large language model answers empty should also be considered as correct recognition. Therefore, we use Modified Correct to remove the above influence. We have redesigned the validation method, comparing and analyzing the results extracted from the large model with those annotated by experts, and calculating specific indicators based on binary classification. The calculation method is shown in the Fig. 4.

To the best of our knowledge, CataLM is the first work to use LLMs for extracting knowledge content in the field of photocatalysis. However, in the field of materials, some scholars have used supervised intelligent methods to assist in the recognition of material entities. Wang et al. [43] proposed CHEMNER, an ontology guided, distantly-supervised method for fine grained chemistry NER, and achieved 45.96% F1. In the work [44], a supervised entity extraction

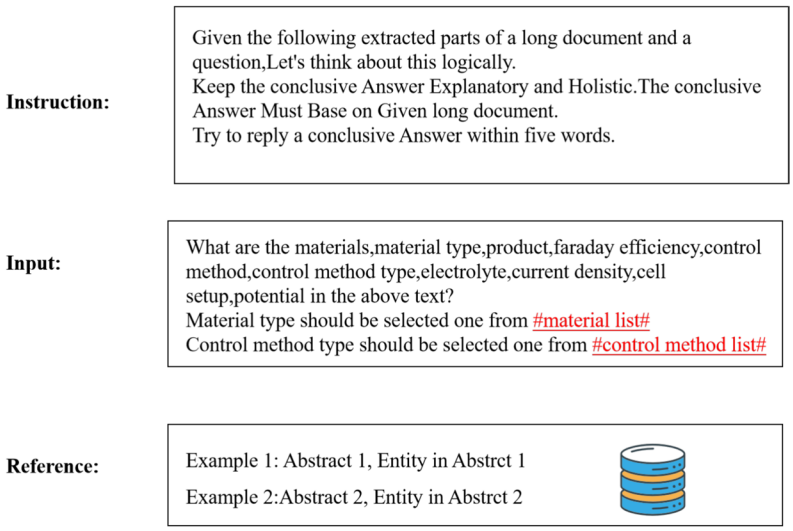

Fig. 3 Prompt in the named entity recognition task

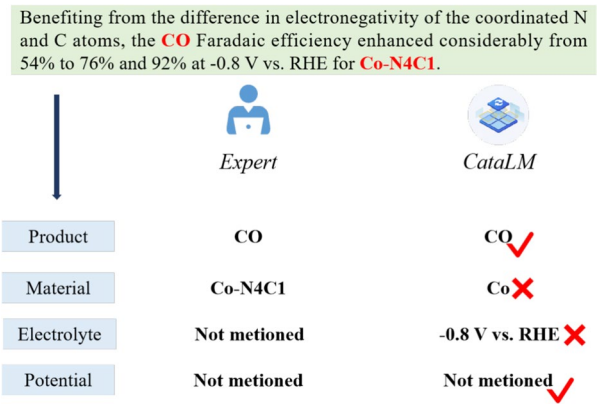

Fig. 4 Metric setting in CataLM

method was proposed in the field of Catalyst (the entity label are material, product, control method and faradaic efficiency), and obtain 57.68% F1. In this work, the ontology is more complex and hard to recognize, in the entity label as catalyst, product and faraday efficiency, CataLM's modified accuracy (which also can be seen as F1) are 75, 85 and 90%, but it performs poorly in entity extraction for control method, cell setup, and the reason may be those entity are more difficult to understand and cause hallucinations in LLMs.

We also conducted ablation experiments in this paper. We decomposed the model into two modules, namely the model Domain pre-training module and the Retrieval-Augmented Generation (RAG) module, and they were combined in pairs to form four possibilities. From Table 4, it can be seen that our method (i.e. Pre-training LLM + Few shot) performs the best. We can also see that both the fine-tuned module and the RAG module contribute to the improvement of model extraction accuracy. To present the recent trends in the development of CO_2 reduction electrocatalysts, we showcased and analyzed the information in the extracted database. Notably, for the sake of clarity, less frequently reported catalyst categories have been excluded. E/C and E-M are recommended choices for producing CO, while E-M and EOx demonstrate the potential for formic acid production. As for C_2 products like C_2H_4 and C_2H_5OH,

Table 3 The evaluation of entity recognition of **CataLM**

Entity	Count	Correct	Existence	Modified correct	Recall	Precision	F1
MATERIAL	20	17	17	15	88.23%	83.33%	75%
CONTROL METHOD	20	19	19	13	68.42%	92.86%	65%
PRODUCT	20	17	17	17	100%	85%	85%
FARADAIC EFFICIENCY	20	11	11	18	100%	84.62%	90%
ELECTROLYTE	20	10	10	10	90%	50%	50%
POTENTIAL	20	7	7	16	100%	63.63%	80%
CURRENT DENSITY	20	7	7	12	100%	46.67%	60%
CELL SETUP	20	6	6	9	100%	35.29%	45%
OVERALL	160	85	94	110	90.43%	67.46%	68.75%

Table 4 Results of ablation experiment

Model	Correct	Modified correct	Recall	Precision	F1
Original LLM + Zero shot	27	59	40.23%	36.05%	36.88%
Original LLM + Few shot	37	66	42.04%	46.25%	41.25%
Pre-training LLM + Zero shot	49	85	50.93%	71.43%	53.12%
Our method	85	110	90.43%	67.46%	68.75%

both E and EO_x are viable alternatives. We also open source the fine-tuning data, and domain experts can conduct more in-depth analysis.

4.2 Control method recommendation task

With the continuous development of big data technology, basic scientific research has shifted from the traditional "random trial and error" to the "data-driven AI" scientific model. Domain experts have also begun to attempt to use large language models to promote scientific innovation, such as literature understanding and summarization, experimental scheme generation, as well as unmanned experimental systems and scientific data sharing platforms, in order to improve scientific research efficiency and promote scientific progress and development. **CataLM** focuses on the scientific problems in the Catalyst Control field, and tries to assist scientists in catalyst design. In Retrieval Augmented Generation(RAG) method, we use Sci-BERT [36] as the embedding method to construct the vector database, and select three papers as the references.

However, how to evaluate the effectiveness of recommended catalyst control methods is a challenge faced by **CataLM**. In this paper, we invite domain experts to evaluate and analyze the recommendation methods generated by **CataLM** and the original LLM. Several representative results are listed in Table 5. The original LLM typically offers vague material selections and often lacks domain-specific expertise. Additionally, its understanding of control methods usually involves mere restatements of definitions, sometimes leading to significant misconceptions. In contrast, **CataLM** consistently integrates domain knowledge to provide more specific and accurate material choices. Its responses are logically coherent, with reasonable explanations of the mechanisms behind the selected materials. As illustrated in Fig. 5, **CataLM** demonstrates its ability to accurately interpret control methods and offer reasonable suggestions, such as the composite system of Cu and carbon, along with clear explanations for the choice of carbon materials. In contrast, the original LLM's misunderstanding of the control method is evident in its incorrect suggestion of Pt, which is not a catalyst for CO_2 reduction. Additionally, we investigated the impact of hyperparameter selection on the responses generated by CataLM. All CataLMs produced reasonable answers, albeit with varying emphases, similar to how different researchers might provide distinct academic perspectives on the same research question.

For the quantitative analysis in recommendation, we choose Faraday efficiency as a measure of the effectiveness of the regulation methods. This index refers to the percentage of actual product and theoretical product, which is as important as the onset potential and the total current density that can be reached at a certain potential. The specific calculation formula is as follows:

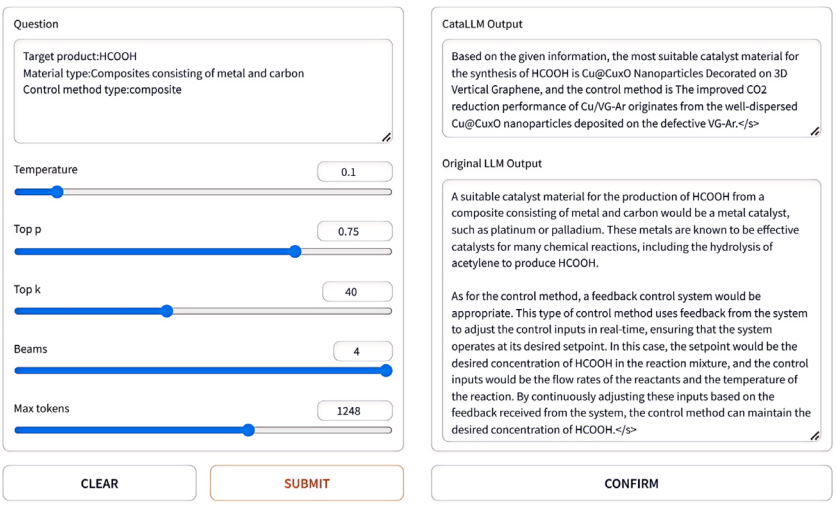

Fig. 5 Answer from **CataLM** and original LLM

Table 5 The input and answer provided by orginal LLM and **CataLM**

Input	Original LLM	**CataLM** (β=0.1)	**CataLM** (β=0.5)
C_2H_5OH, Single metal, structure control	The suitable catalyst material for the production of C_2H_5OH (ethanol) would be a noble metal such as platinum or palladium. These metals are known for their ability to catalyze a wide range of chemical reactions and are often used in the production of ethanol through the hydrolysis of ethylene. As for the control method type, structure control would be the most appropriate	Based on the given information, the most suitable catalyst material for producing C_2H_5OH is Palladium (Pd). The control method that should be used is creating high-facets of wrinkled Pd surrounded by PdO mesh patterns.	Based on the given information, the most suitable catalyst material for producing C_2H_5OH is Cu, and the suitable control method is nanostructured self-supporting Cu electrodes
C_2H_4, Metal oxide, structure control	A suitable catalyst material for the production of C_2H_4 (ethylene) would be a metal oxide such as iron oxide or copper oxide. These materials are known to be effective in the production of ethylene through the cracking of hydrocarbons. As for the control method, structure control would be an appropriate method to use. This type of control method involves manipulating the structure of the catalyst material in order to optimize its performance	Based on the information provided, the most suitable catalyst material for producing C_2H_4 is OxygenCu (OCu) combination. The control method that should be used is anodized copper (AN-Cu) Cu(OH)(2) catalysts by a simple electrochemical synthesis method.	Based on the information provided, the most suitable catalyst material is Cu_2O, and the control method is the selectivity and activity of the C_2H_4 production depend strongly on the crystal facets exposed in Cu_2O NPs
HCOOH, Composites consisting of metal and carbon, composite	Based on the given information, a suitable catalyst material for the production of HCOOH would be a composite consisting of metal and carbon. One example of such a catalyst is platinum-on-carbon (Pt/C). As for the control method, a suitable method would be to use a feedback control system. This type of control system uses a sensor to measure the concentration of HCOOH in the reaction mixture, and adjusts the reaction conditions in real-time to maintain a constant concentration of HCOOH	Based on the given information, the most suitable catalyst material for the synthesis of HCOOH would be a composite of Copper nanoparticles embedded in Nitrogen-doped graphene oxide (Cu/N-GO). The control method used in this composite would be the interaction between the copper nanoparticles and the nitrogen-doped graphene oxide support. This interaction is expected to enhance the electrocatalytic reduction of CO_2 to HCOOH, due to the synergistic effects between the copper nanoparticles and the nitrogen-doped	Based on the given information, the most suitable catalyst material for the synthesis of HCOOH is Cu@CuxO Nanoparticles Decorated on 3D Vertical Graphene, and the control method is The improved CO_2 reduction performance of Cu/VG-Ar originates from the well-dispersed Cu@CuxO nanoparticles deposited on the defective VG-Ar
CO, Alloys/composites of two or more metals, alloy	A suitable catalyst material for producing CO (carbon monoxide) from an alloy or composite of two or more metals would be a metal oxide such as iron oxide or cobalt oxide. As for the control method, a suitable option would be to use a feedback control system	Based on the given information, the most suitable catalyst material for the target product CO is Gold-Copper alloy. The control method for this catalyst material is also Alloy	Based on the given information, the most suitable catalyst material for the target product CO is Copper-Indium (CuIn) and the control method is alloy

Table 6 Results of quantitative analysis

Model	Average predicted faraday efficiency
Original LLM	61.40%
CataLM	64.15%

$$FE = \frac{m \times n \times f}{I \times t} \quad (1)$$

where m is the actual mole number of the product, n is the number of reaction electrons, f is the Faraday constant, that is, the electric quantity contained in one mole of electrons, I is the current, t is the time. In order to quickly and automatically evaluate this indicator, we use the deep learning prediction model based on semantic information (word embedding) and correlation of knowledge triples (graph embedding), which is proposed in work [44]. This method was trained by 533 Faraday efficiency data samples, with the form as [data: (material, product, method, method type, material type), label: Faraday Efficiency], and outperformed other prediction models on Faraday Efficiency prediction. The evaluation result is shown in Table 6.

From the result, we can observe that the control method details generate by CataLM are obtain better performance than original LLM. However, beside the target product and material, the prediction results of the model are also influenced by the category of regulatory methods, which is an input parameter. This means that for the original LLM and CataLM, the regulatory methods are the same, so the difference in the predicted Faraday effect is not significant.

5 Conclusion

In this paper, we introduce **CataLM**, an effective attempt towards catalyst design leveraging the capabilities of large language models. By undergoing domain pre-training and instruction tuning, our large language model has exhibited robust comprehension and reasoning skills in catalyst knowledge and patterns, achieving advanced performance in application tasks like knowledge extraction and recommendation of control methods. We have open sourced the **CataLM** model and fine-tuning data to facilitate further expansion and development by interested researchers, which is available at https://github.com/kg4sci/CataLM. The result of NER task is available at Science Data Bank (ScienceDB), which is a public, general-purpose data repository aiming to provide data services for researchers, research projects/teams, journals, institutions, universities, etc, the link is https://www.scidb.cn/en/detail?dataSetId=3f6204bc48704fac9b64b8e95a904e02.

Currently, CataLM focuses more on catalyst design in the field of carbon dioxide reduction due to limitations in the scope of field literature. In the future, while continuously enhancing the field understanding ability of **CataLM** by incorporating more domain literature, we will also design and develop an auxiliary platform for field researchers based on it, in order to improve the efficiency of catalyst design work in practical applications. We believe that large language models will bring new and infinite possibilities to basic scientific research.

Acknowledgements This work was supported by the National Key R&D Program of China under Grant No.2022YFF0711900 and 2022YFF0712200, the Natural Science Foundation of China under Grant No. T2322027, Youth Innovation Promotion Association CAS. We thank all the domain researchers that help to annotate and verify the corpus.

Author Contributions All authors contributed substantively to the work presented in this paper. Conception and Supervision: W. Cui, Y. Gao; Data acquisition: L. Wang, X. Chen, Y. Du; Data validation: Y. Gao, Y. Zhou, W. Cui; Technical validation: L. Wang, X. Chen; Dataset mining: Y. Gao, W. Cui; Writing and Proof reading: L. Wang, Y. Gao, X. Chen, W. Cui, Y. Du.

Data Availibility No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. De Luna P, Hahn C, Higgins D, Jaffer SA, Jaramillo TF, Sargent EH (2019) What would it take for renewably powered electrosynthesis to displace petrochemical processes? Science 364(6438):3506
2. Seh ZW, Kibsgaard J, Dickens CF, Chorkendorff I, Nørskov JK, Jaramillo TF (2017) Combining theory and experiment in electrocatalysis: Insights into materials design. Science 355(6321):4998
3. Nørskov JK, Bligaard T, Rossmeisl J, Christensen CH (2009) Towards the computational design of solid catalysts. Nat Chem 1(1):37–46

4. Suntivich J, May KJ, Gasteiger HA, Goodenough JB, Shao-Horn Y (2011) A perovskite oxide optimized for oxygen evolution catalysis from molecular orbital principles. Science 334(6061):1383–1385
5. Liu J, Liu H, Chen H, Du X, Zhang B, Hong Z, Sun S, Wang W (2020) Progress and challenges toward the rational design of oxygen electrocatalysts based on a descriptor approach. Adv Sci 7(1):1901614
6. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
7. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692
8. Clark K, Luong M-T, Le QV (2020) Manning, CD Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555
9. Radford A, Narasimhan K, Salimans T, Sutskever I et al (2018) Improving language understanding by generative pre-training
10. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
11. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. In: Advances in neural information processing systems vol. 33, p. 1877–1901
12. OpenAI R (2023) Gpt-4 technical report. ArXiv:2303.08774
13. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res 21(140):1–67
14. Gao L, Biderman S, Black S, Golding L, Hoppe T, Foster C, Phang J, He H, Thite A, Nabeshima N et al (2020) The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027
15. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, Liu T-Y (2022) Biogpt: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinformatics 23(6):409
16. Zhang H, Chen J, Jiang F, Yu F, Chen Z, Li J, Chen G, Wu X, Zhang Z, Xiao Q et al (2023) Huatuogpt, towards taming language model to be a doctor. arXiv preprint arXiv:2305.15075
17. Xiong H, Wang S, Zhu Y, Zhao Z, Liu Y, Huang L, Wang Q, Shen D (2023) Doctorglm: Fine-tuning your chinese doctor is not a herculean task. arXiv preprint arXiv:2304.01097
18. Xie F, Lu T, Yu Z, Wang Y, Wang Z, Meng S, Liu M (2023) Lu-h-n phase diagram from first-principles calculations. Chin Phys Lett 40(5):057401
19. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). Jom 65:1501–1509
20. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G et al (2013) Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater 1(1):011002
21. Liang Y, Chen M, Wang Y, Jia H, Lu T, Xie F, Cai G, Wang Z, Meng S, Liu M (2023) A universal model for accurately predicting the formation energy of inorganic compounds. Sci China Mater 66(1):343–351
22. Liu Z, Guo J, Chen Z, Wang Z, Sun Z, Li X, Wang Y (2022) Swarm intelligence for new materials. Comput Mater Sci 214:111699
23. Guo J, Chen Z, Liu Z, Li X, Xie Z, Wang Z, Wang Y (2022) Neural network training method for materials science based on multi-source databases. Sci Reports 12(1):15326
24. Gupta T, Zaki M, Krishnan NA, Mausam (2022) Matscibert: A materials domain language model for text mining and information extraction. NPJ Comput Mater 8(1):102
25. Chen Z-Y, Xie F-K, Wan M, Yuan Y, Liu M, Wang Z-G, Meng S, Wang Y-G (2023) Matchat: a large language model and application service platform for materials science. Chin Phys B 32(11):118104. https://doi.org/10.1088/1674-1056/ad04cb
26. Xie T, Wan Y, Huang W, Zhou Y, Liu Y, Linghu Q, Wang S, nCG, Zhang W, Hoex B (2023) Large language models as master key: unlocking the secrets of materials science with GPT arXiv:2304.02213
27. Zheng Z, Zhang O, Borgs C, Chayes JT, Yaghi OM (2023) Chatgpt chemistry assistant for text mining and the prediction of mof synthesis. J Am Chem Soc 145(32):18048–18062. https://doi.org/10.1021/jacs.3c05819
28. Cao H, Liu Z, Lu X, Yao Y, Li Y (2023) InstructMol: multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. arXiv preprint arXiv:2311.16208
29. Barroso-Luque L, Shuaibi M, Fu X, Wood BM, Dzamba M, Gao M, Rizvi A, Zitnick CL, Ulissi ZW (2024) Open materials 2024 (omat24) inorganic materials dataset and models. arXiv preprint arXiv:2410.12771
30. Leong SX, Pablo-García S, Zhang Z, Aspuru-Guzik A (2024) Automated electrosynthesis reaction mining with multimodal large language models (mllms). Chemical Science. arXiv preprint arXiv:2311.16208
31. Kulmanov M, Guzmán-Vega FJ, Duek Roggli P, Lane L, Arold ST, Hoehndorf R (2024) Protein function prediction as approximate semantic entailment. Nat Mach Intell 6(2):220–228
32. Qian J, Jin Z, Zhang Q, Cai G, Liu B (2024) A liver cancer question-answering system based on next-generation intelligence and the large model med-palm 2. Int J Comput Sci Inf Technol 2(1):28–35
33. Wu C, Zhang X, Zhang Y, Wang Y, Xie W (2023) Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454
34. Chen B, Cheng X, Li P, Geng Y-a, Gong J, Li S, Bei Z, Tan X, Wang B, Zeng X et al (2024) xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. arXiv preprint arXiv:2401.06199
35. Liu R, McKie J (2018) PyMuPDF. May. http://pymupdf.readthedocs.io/en/latest/
36. Beltagy I, Lo K, Cohan A (2019) Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676
37. Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, Hajishirzi H (2022) Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560
38. Qin Y, Liang S, Ye Y, Zhu K, Yan L, Lu Y, Lin Y, Cong X, Tang X, Qian B et al (2023) Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv preprint arXiv:2307.16789
39. Wang L, Gao Y, Chen X, Cui W, Zhou Y, Luo X, Xu S, Du Y, Wang B (2023) A corpus of co2 electrocatalytic reduction process extracted from the scientific literature. Sci Data 10(1):175
40. Du Y, Wang L, Huang M, Song D, Cui W, Zhou Y (2023) Autodive: An integrated onsite scientific literature annotation tool. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp. 76–85
41. Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) LoRA: low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685
42. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J, Shleifer S, Platen P, Ma C, Jernite Y, Plu J, Xu C, Scao TL, Gugger S,

Drame M, Lhoest Q, Rush AM (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Online . https://www.aclweb.org/anthology/2020.emnlp-demos.6

43. Wang X, Hu V, Song X, Garg S, Xiao J, Han J (2021) Chemner: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing
44. Gao Y, Wang L, Chen X, Du Y, Wang B (2023) Revisiting electrocatalyst design by a knowledge graph of cu-based catalysts for co 2 reduction. ACS Catal 13:8525–8534. https://doi.org/10.1021/acscatal.3c00759

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.