scientific data

OPEN

DATA DESCRIPTOR

Large language model enhanced corpus of CO_2 reduction electrocatalysts and synthesis procedures

Xueqing Chen[1,2,5], Yang Gao[3,5], Ludi Wang[1,5], Wenjuan Cui[1], Jiamin Huang[3], Yi Du[1,2,4] ✉ & Bin Wang[3] ✉

CO_2 electroreduction has garnered significant attention from both the academic and industrial communities. Extracting crucial information related to catalysts from domain literature can help scientists find new and effective electrocatalysts. Herein, we used various advanced machine learning, natural language processing techniques and large language models (LLMs) approaches to extract relevant information about the CO_2 electrocatalytic reduction process from scientific literature. By applying the extraction pipeline, we present an open-source corpus for electrocatalytic CO_2 reduction. The database contains two types of corpus: (1) the benchmark corpus, which is a collection of 6,985 records extracted from 1,081 publications by catalysis postgraduates; and (2) the extended corpus, which consists of content extracted from 5,941 documents using traditional NLP techniques and LLMs techniques. The Extended Corpus I and II contain 77,016 and 30,283 records, respectively. Furthermore, several domain literature fine-tuned LLMs were developed. Overall, this work will contribute to the exploration of new and effective electrocatalysts by leveraging information from domain literature using cutting-edge computer techniques.

Background & Summary

CO_2 electroreduction has garnered significant attention from both the academic and industrial communities, owing to its potential to effectively mitigate greenhouse gas emissions while simultaneously producing fuels and chemicals[1–3]. Its widespread adoption relies heavily on the development of efficient and reliable electrocatalysts. Over the past three decades, scientists have invested substantial efforts in the development of CO_2 reduction electrocatalysts[4,5]; However, this trial-and-error approach has proven to be time-consuming and labor-intensive. Consequently, it becomes pivotal in accelerating catalyst development to establish a comprehensive database for CO_2 electroreduction, which should encompass various information pertaining to the composition, synthesis, regulation, and performance of catalysts. Given the substantial workload involved, the manual annotation method by domain experts is deemed unreasonable. In recent years, emerging artificial intelligence (AI) technologies have exhibited tremendous potential in facilitating the construction of realm-specific datasets[6,7]. Extracting crucial information related to catalysts from domain literature is the initial step toward accelerating catalyst development using AI technologies. Traditionally, Named Entity Recognition (NER) methods have been employed for text mining and information retrieval[8–11]. However, NER often necessitates the establishment of algorithms tailored to specific tasks, which are typically undertaken by scientists or engineers with expertise in coding, data structures, and computer algorithms. Therefore, this approach is labor-intensive. Furthermore, NER algorithms are closely tied to their assigned tasks, lacking generalizable ability and thus making direct transfer to other tasks challenging. Additionally, extracted information tends to be intricate, heterogeneous, and diverse in the field of catalysis, leading to unsatisfied NER performance and reduced accuracy[12]. Therefore, the

[1] Laboratory of Big Data Knowledge, Computer Network Information Center, Chinese Academy of Sciences, Beijing, 100083, China. [2] University of Chinese Academy of Sciences, Beijing, 100049, China. [3] CAS Key Laboratory of Nanosystem and Hierarchical Fabrication, National Center for Nanoscience and Technology (NCNST), Beijing, 100190, China. [4] Hangzhou Institute for Advanced Study, UCAS, Hangzhou, 310000, China. [5] These authors contributed equally: Xueqing Chen, Yang Gao, Ludi Wang. ✉e-mail: duyi@cnic.cn; wangb@nanoctr.cn

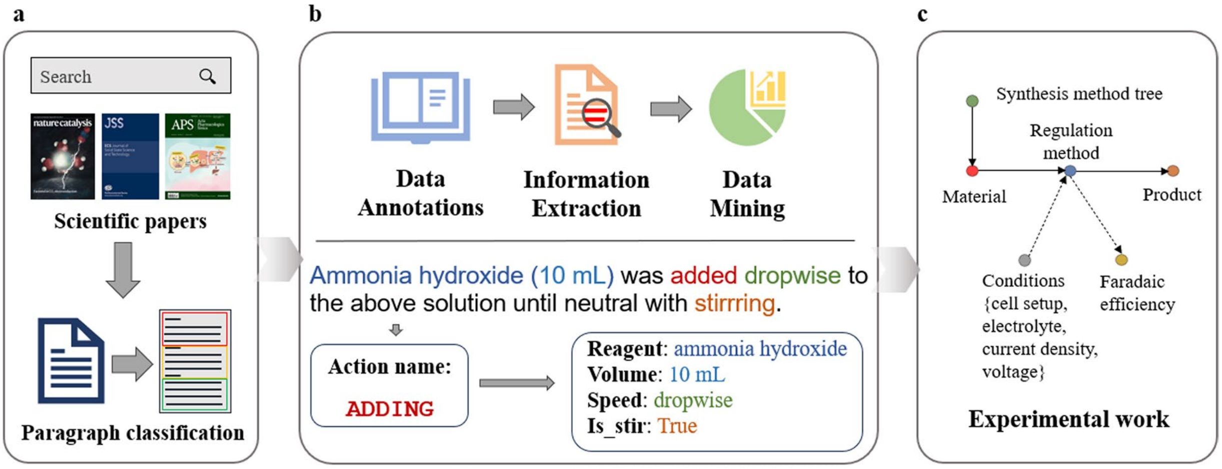


Fig. 1 The schematic overview of dataset construction pipeline. (**a**) The process of literature search filtering and paragraph classification. (**b**) The top panel shows the schematic diagram of the standard text mining process: <i> expert annotation to build a baseline corpus; <ii> extraction of critical information from the literature text and construction of an extended corpus; <iii> store in a database for future data mining. The bottom panel shows an example of converting a synthesis sentence into action sequences. The key components of an action sequence such as starting and target material, synthesis steps and their conditions are found and extracted from the paragraph by different text mining algorithms (see Methods). (**c**) The entity types and their relationships extracted from the literature. The final constructed dataset can provide guidance for practical experimental work.

development and utilization of more general and robust methods for extracting domain knowledge are becoming increasingly imperative.

Recently, the emergence of large language models (LLMs), especially the widely acclaimed ChatGPT, has brought new prospects to the field of NER tasks[13]. It can be effectively operated by domain scientists who may not be well-versed in computer algorithms. However, ChatGPT is susceptible to information hallucinations, a glaring issue that significantly undermines its reliability in scientific domains[14–16]. Prompt engineering has proven to be a potential solution to mitigate the problem of artificial hallucinations[17–19]. For instance, Zheng *et al.* employed prompt engineering to guide ChatGPT in automating text mining for the synthesis conditions of metal-organic frameworks[17]. Nevertheless, the utility of this approach for more diverse and complex tasks within the catalytic science domain remains an area warranting further exploration. Moreover, the high demand for computing resources in LLMs also limits their application in various fields. The training and application of LLMs usually require a tremendous amount of computational power, which are not only expensive to purchase but also consume substantial amounts of electricity.

In recent work, our team has developed a text-mining pipeline to construct a dataset describing the CO_2 reduction process catalyzed by copper-based electrocatalysts, which specifically includes material, regulation method, product, Faradaic efficiency and relevant conditions[12]. In the current work, we built a more advanced extraction pipeline based on the knowledge system of CO_2 electrocatalytic reduction (Fig. 1), which uses various advanced machine learning, natural language processing techniques and large language models (LLMs) approaches to extract relevant information about the CO_2 electrocatalytic reduction process from scientific literature. In addition, for the purpose of providing a more detailed and complete guidance scheme for materials scientists to develop new catalysts, we designed a set of synthesis actions with predefined properties and a deep-learning sequence to sequence model based on the transformer architecture, which converts unstructured experimental procedure text into structured action sequences. By applying the extraction pipeline, we present an open-source corpus for electrocatalytic CO_2 reduction. The database contains two types of corpus: (1) the benchmark corpus, which is a collection of 6,086 records extracted from 1,081 publications by catalysis postgraduates; and (2) the extended corpus, which consists of content extracted from the abstract of 5,941 documents using traditional natural language processing techniques and large language models techniques. Respectively, the Extended Corpus I contains 77,016 records and the Extended Corpus II contains 30,283 records. In addition, we extracted 476 synthesis procedures for catalytic materials from 2,176 full-text documents, and the extracted information includes target and preparation materials, synthesis operations and the quantity of materials involved in them, and operation properties. The Extended Corpus was evaluated and revised by domain experts. This work provides a valuable resource to accelerate research into CO_2 reduction by supplying structured information and datasets ready for further analysis and hypothesis generation. The tools and datasets created could significantly reduce the time and resources required for literature review and data gathering, allowing scientists to focus on innovation and experimentation.

Methods

The schematic overview of the extraction pipeline is shown in Fig. 1. We first searched the literature related to the electrocatalytic CO_2 reduction process following a series of filtering criteria. For scientific article retrieval and preprocessing, the raw archived corpus was parsed and organized in paragraphs. After paragraph classification, the paragraphs related to the concrete synthesis procedures were automatically selected. The extracted information includes the materials, the target products, their quantities as well as the synthesis operations and their attributes. We then constructed action sequences for each synthesis action in a predefined format. Finally based on the the system of knowledge defined by domain experts, we published a manually annotated baseline corpus and an automatically annotated extended corpus. The final generated dataset can be used for domain data mining and further downstream NLP tasks, as well as provide guidance to material domain scientists for practical experimental work.

Content acquisition. Scientific publications used in this work are journal articles published by Elsevier, the Royal Society of Chemistry, American Chemical Society, Wiley, Acta Physico-Chimica Sinica & University Chemistry Editorial Office (Peking University), MDPI, the Electrochemical Society, Springer Nature, etc. For each publisher, the journals relevant to materials science were manually selected. We used regular expression matching[20] to obtain the dois of relevant literature in the field of CO_2 electrocatalytic reduction. Specifically, we searched and exported metadata for more than 27,000 articles by using the keywords "CO_2", "Reduction", and "Electro*" as subject indexes on the Web of Science website. The exported literature metadata was then filtered step by step according to expert-defined rules. The title of every article was queried for words "CO_2", "carbon dioxide" or "CO(2)", which yielded 9,850 articles. The abstract of every article was queried for words "electroc" or "electror", which yielded 6,973 articles. Finally the domain experts performed manual filtration to exclude articles whose titles contained words that were not relevant to the topic, including: "photoc", "light", "visible", "solar", "microbial", "bacteria", "culture", etc. we eventually obtained 5,941 summary texts of the literature related to the work on CO_2 electrocatalytic reduction and scraped the full text of 2,776 papers from the web. We finally acquired the literature in PDF format and used the PyMuPDF tool, a PDF parsing tool[21], to automatically process these literature data to obtain their metadata such as title, authors, abstract, etc. and the full text in json format. Since the processed document contains irrelevant tags, we developed a data cleaning method for parsing the article tag strings into consistently formatted text paragraphs while retaining the same chapter and paragraph structure as the original paper.

Paragraph classification. We used the Transformers Bidirectional Encoder Representation (BERT) model to identify paragraphs containing descriptions of synthesis methods. MatBERT is a BERT model[22] specifically for material science texts, pre-trained on over 2 million papers in a self-supervised manner, i.e. by predicting masked words based on the context around the target sentence. After training the BERT model, we used a paragraph classification method based on semi-supervised learning[23]. First we applied latent Dirichlet allocation (LDA)[24] on the 12,643 articles in the field of photoelectrocatalysis to identify the experimental steps implicit in sentences. Then we collected all the paragraphs from the literature and manually labelled the paragraphs describing the synthesis protocol. The training data ultimately included 760 training examples, with 228 positive examples and 532 negative examples. We applied the random decision forest (RF) algorithm[25], a supervised machine learning method, to binary classify the training data. This step yielded 476 synthesis paragraphs from a total of 2,776 articles.

Entity annotation. In order to improve the quality of the training data based on the automatically extracted models, we generated a higher-quality dataset, also known as a *gold standard corpus*[26], by manually annotating a portion of the sentences from the abstracts and body of literature related to CO_2 electroreduction. We developed an annotation framework based on the doccano annotation tool[27]. Annotators can open the framework in a web browser and browse through the sentences of the material literature. The page displays the sentence to be annotated along with predefined entity types and related descriptions. The annotator can add new entities, reorder them or edit them by opening a separate view. To ensure consistency between annotators, detailed annotation guidelines are provided.

Entity extraction. In our previous study, we extracted nine types of entities in the literature based on the constructed electrocatalytic reduction system, including material, regulation method, product, faradaic efficiency, cell setup, electrolyte, synthesis method, current density, and voltage. Some of these entity labels are provided with more detailed labelling subclasses to ensure that materials scientists have access to more complete information. In the current construction of the CO_2 electrocatalysis literature dataset, We have updated the categories of the tag subcategories according to the new knowledge system. In addition, we added information on the material synthesis process, which converted unstructured scientific paragraphs describing catalytic materials synthesis into pre-defined "coded recipes" of synthesis. The recipes includes not only the starting materials and final target products but also the synthesis actions and their attributes.

Construction of extended corpus. Traditional entity extraction methods follow the pattern of "expert annotation, model training, model application" and use automatic extraction models to build a wider and larger corpus of lower quality, also known as a *silver standard corpus*(SSC)[26]. The Large Language Models (LLMs) such as GPT-3, GPT-3.5, and GPT-4 have been used for this purpose[28–30]. Its emergency provides a new paradigm for natural language processing modelling, i.e., building prompts with a small amount of expert annotation to directly fine-tune GPT models that have been pre-trained on large-scale data. Traditional NER methods are less general, but have higher domain confidence, while large models may produce uncontrollable illusions. Herein, in this paper, we used two model training approaches separately to generate an extended corpus based on the construction standard of the *silver standard corpus*(SSC).

Entity extraction using traditional NER methods. Regarding the hierarchical structure of entity labelling, we designed a two-step entity recognition model which consists of coarse-grained entity recognition and fine-grained entity classification. In the first step, we used the SciBERT model[31] to convert each word token

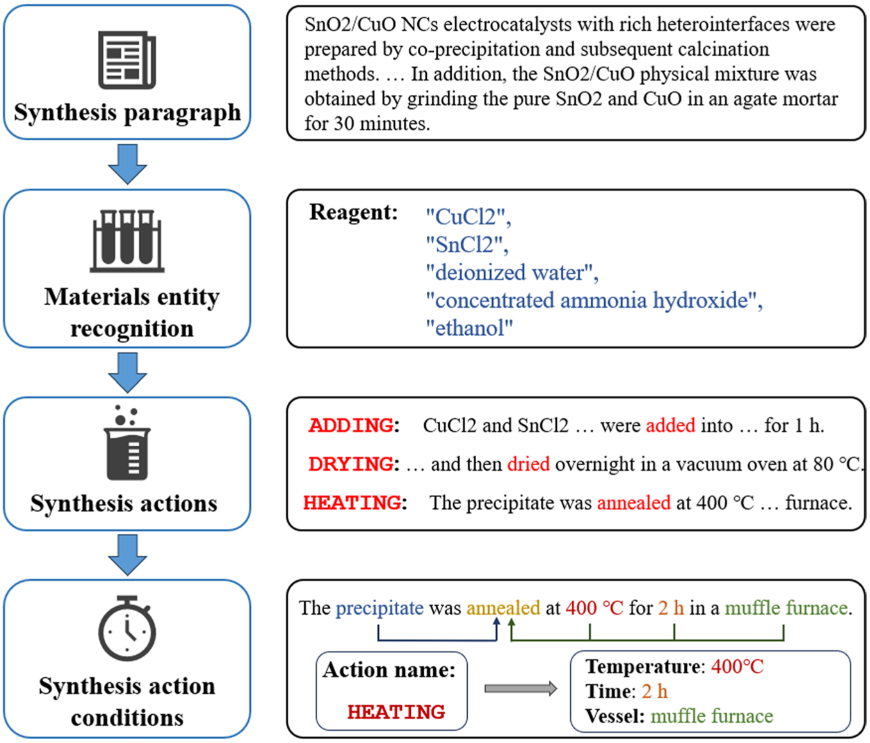

Synthesis paragraph
SnO2/CuO NCs electrocatalysts with rich heterointerfaces were prepared by co-precipitation and subsequent calcination methods. … In addition, the SnO2/CuO physical mixture was obtained by grinding the pure SnO2 and CuO in an agate mortar for 30 minutes.
Materials entity recognition
Reagent: "CuCl2", "SnCl2", "deionized water", "concentrated ammonia hydroxide", "ethanol"
Synthesis actions
ADDING: CuCl2 and SnCl2 … were added into … for 1 h.
DRYING: … and then dried overnight in a vacuum oven at 80 °C.
HEATING: The precipitate was annealed at 400 °C … furnace.
Synthesis action conditions
The precipitate was annealed at 400 °C for 2 h in a muffle furnace.
Action name: HEATING
Temperature: 400°C
Time: 2 h
Vessel: muffle furnace

Fig. 2 Schematic diagram of the process of converting a synthetic paragraph into action sequences.

into an embedding vector. The embedding vector was then passed to a bi-directional long-short-term memory neural network with a conditional random-field top layer(BiLSTM-CRF)[32,33] to identify which class of entity labels the corresponding token was. Considering that the representations of some entities usually have regularities, such as the chemical formula expressions of material entities and the numerical expressions of faradaic efficiency entities, we proposed a regular rule-based approach to assist the deep learning model[34]. The results of the two models were selected using a voting scheme[26]. In the second step, each coarse-grained type entity was classified into finer-grained entity classes using a classification algorithm combining dictionary and maximum entropy model. The dictionary-based recognizers used lists of words built on expert-annotated data[35]. For data that cannot be matched, the word embedding vectors, context vectors, word cluster clustering information and coarse-grained entity category information for each entity were passed through a simple mapping function. The final mapping results were used as entity features for classification probability prediction through a maximum entropy model.

A typical synthesis procedure in the electrocatalytic reduction literature contains information on the prepared and target materials, synthesis operations and operating conditions. These items are organized into material synthesis "recipes" and are extracted from the synthesis paragraph as shown in Fig. 2. Our extraction process consists of multiple algorithms that analyze the passages and identify the relevant materials, the synthesis actions performed, and the condition information associated with those synthetic actions. The method used in each step of the extraction process is described in detail below.

Step 1: Materials entity recognition. The first step is the labelling of the preparation material. The synthesis of the target material involves the names of all the reagents that need to be prepared. We used pattern matching against a database of common reagent names and then used a plain Bayesian classifier to determine whether a candidate phrase is a reagent name, excluding some specific phrases[36]. Through iterative trials, we eventually chose reagent names from the Reaxys database and non-reagent-name texts from the Brown English language corpus to train the classifier.

Step 2: Synthesis actions To identify and classify synthesis actions described in passages, we implemented an algorithm that combines Recurrent Neural Networks (RNN) and rule-based sentence dependency tree parsing[22]. The neural network labelled the sentences in the synthetic passages into nine categories: NOT OPERATION, ADDING, HEATING, CURING, ELECTROCHEMICAL ANODIZATION, FILTERING, DRYING, DIPPING and REACITON, which are the main operations in catalytic materials synthesis. We used ChemDataExtractor's ChemWordTokenizer[37] to tokenize the lemmatized sentences. For each synthesis action obtained, we used the SpaCy library[38] to parse the syntactic information of the dependency subtree for linguistic features of the tokens, such as their lexical properties and their dependency on the root token.

Step 3: Synthesis action conditions For each synthesis action, we used dependency tree parsing and rule-based regular expression methods[39] to extract the relevant attributes of the synthesis action, such as heating time, heating temperature, and potential voltage values. In addition, if there were materials involved such as ADDING and REACTION operations, we used pattern-matching techniques to extract the names and

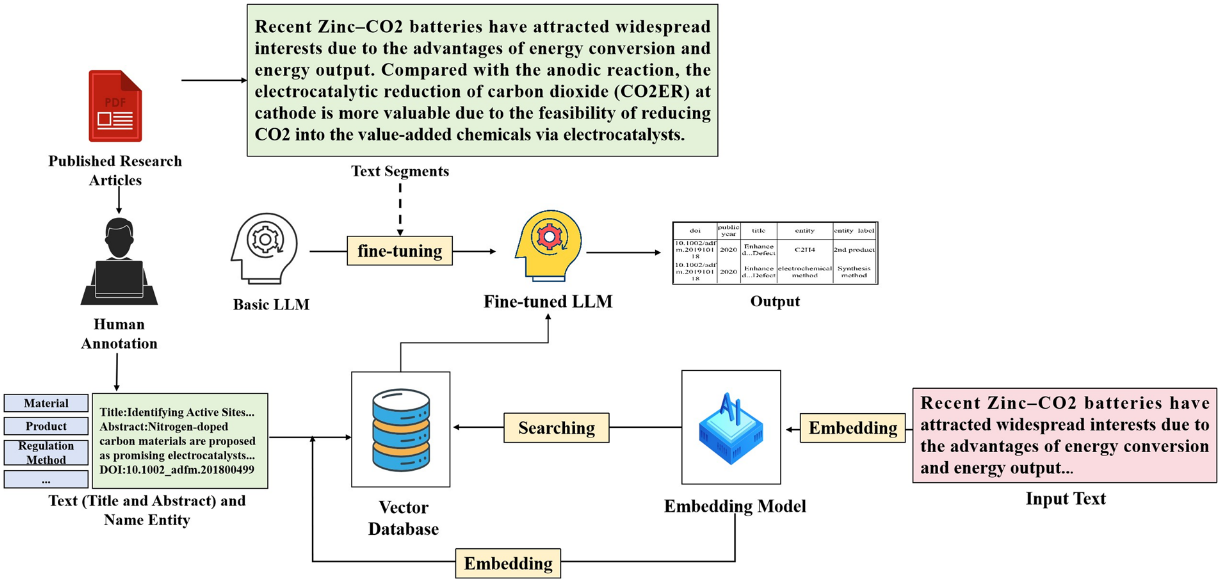

Recent Zinc–CO2 batteries have attracted widespread interests due to the advantages of energy conversion and energy output. Compared with the anodic reaction, the electrocatalytic reduction of carbon dioxide (CO2ER) at cathode is more valuable due to the feasibility of reducing CO2 into the value-added chemicals via electrocatalysts.
PDF
Published Research Articles
Text Segments
fine-tuning
Basic LLM
Fine-tuned LLM
doi | public year | title | entity | entity label
10.1002/adfm.201910118 | 2020 | Enhanced...Defect | C2H4 | 2nd product
10.1002/adfm.201910118 | 2020 | Enhanced...Defect | electrochemical method | Synthesis method
Output
Human Annotation
Material
Product
Regulation Method
...
Title:Identifying Active Sites... Abstract:Nitrogen-doped carbon materials are proposed as promising electrocatalysts... DOI:10.1002_adfm.201800499
Text (Title and Abstract) and Name Entity
Searching
Embedding
Recent Zinc–CO2 batteries have attracted widespread interests due to the advantages of energy conversion and energy output...
Vector Database
Embedding Model
Input Text
Embedding

Fig. 3 The schematic overview of extraction using LLMs and vector database.

corresponding quantities of the reagents involved. For example, one of the patterns used for finding solutions is "a/an XX solution containing Reagent" in which "Reagent" represents a phrase previously tagged as a reagent. An example phrase that would be matched by this pattern is "an aqueous solution containing $HAuCl_4$(10 mol, 125 mL)". The contents of the parentheses are regularly matched to the corresponding quantities of the reagents.

Entity extraction using LLMs. In previous study, we attempted to construct a corpus using an NLP model, but the accuracy of the intelligent model is easily affected by the volume of training data. Herein, we demonstrate that LLMs, including original LLMs and fine-tuned LLMs, can act as assistants to collaborate with human researchers, facilitating entity recognition and text mining to accelerate the research process.

In the realm of catalyst-related tasks, LLM's performance can be significantly enhanced by employing prompt engineering (PE) which can steer LLMs toward generating precise and pertinent information. Although LLMs, including fine-tuned LLMs, can answer general questions, their knowledge depth, accuracy and timeliness are limited in vertical domain filed. To solve this problem, we use vector databases to enhance the reasoning ability of LLMs in vertical domains. Vector databases can transform literature and data into vector representations by embedding vectors. Sci-BERT[31] was used as embedding model for construct the vector database.

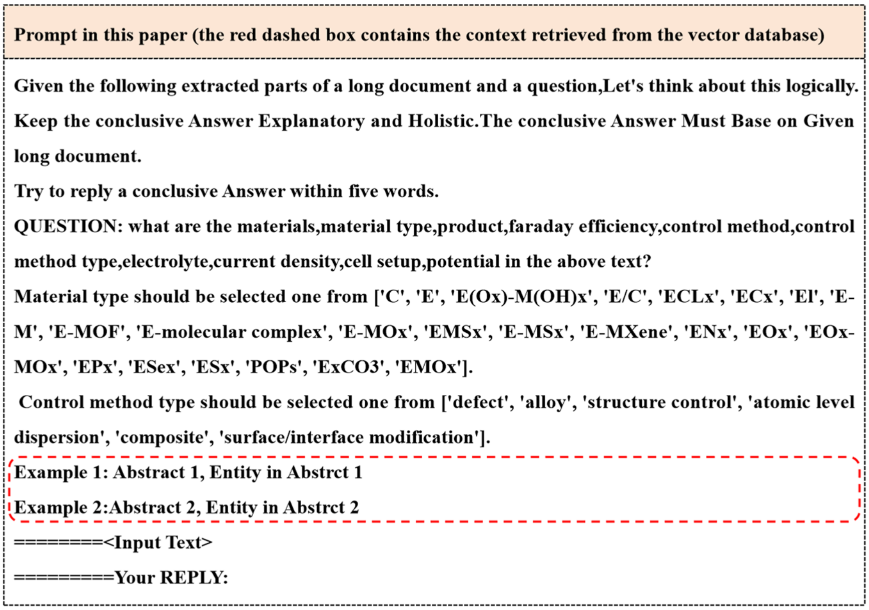
Figure 3 shows the process of knowledge extraction using LLMs and vector database. Firstly, we processed and cleaned the full text of 12,643 photoelectrocatalytic scientific literature, and used them for LLMs fine-tuning. In this step, we chose Vicuna-33b-v1.3 as the basic LLMs. Secondly, we extracted the title, abstract and doi from articles associated with standard corpus, then we use Sci-BERT as the embedding model to transform title and abstract into vector. When performing entity recognition, user first input the text to be extracted, embedding model transform 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. 4. The prompt will be input to the fine-tuned LLMs for entity recognition.

Data Records

The both types of datasets constructed in this work are available in ScienceDB, a public, general-purpose data repository designed to serve data to researchers, research projects/teams, journals, institutions, universities, and others. The metadata contained in the article dataset includes: article DOI, year of publication, and title. Each record corresponds to the process of CO_2 electrocatalytic reduction and its metadata includes: the entity extracted from the paper, the label of the entity, and the sentence in which the entity is located. In addition, the datasets for the catalytic material synthesis methods are available as a single json. Each record corresponds to a synthesis procedure extracted from a paragraph and is represented as a separate json object. The metadata for each reaction includes the DOI of the paper from which the reaction is extracted as well as a fragment of the corresponding synthesis paragraph, the target product, the preparative material used in the reaction, and a tree of seven types of synthesis operations and their corresponding conditions. Table 1 gives extended details of all the dataset format.

The sequence of synthesis steps for the reaction (if specified in a paragraph) is listed as a data structure with the following fields: the original paragraph in the text (`synthesis_paragraph`), its type (`operation_string`) specified by the classification algorithm (see Methods), and the conditions associated with this operation step (`conditions`). We classified the types of operations involved in the synthesis of catalyst materials into eight categories and give detailed descriptions of the types of operations and condition attributes in Table 2.

The corpus is publicly 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,



Prompt in this paper (the red dashed box contains the context retrieved from the vector database)

Given the following extracted parts of a long document and a question,Let's think about this logically. Keep the conclusive Answer Explanatory and Holistic.The conclusive Answer Must Base on Given long document.

Try to reply a conclusive Answer within five words.

QUESTION: what are the materials,material type,product,faraday efficiency,control method,control method type,electrolyte,current density,cell setup,potential in the above text?

Material type should be selected one from ['C', 'E', 'E(Ox)-M(OH)x', 'E/C', 'ECLx', 'ECx', 'EI', 'E-M', 'E-MOF', 'E-molecular complex', 'E-MOx', 'EMSx', 'E-MSx', 'E-MXene', 'ENx', 'EOx', 'EOx-MOx', 'EPx', 'ESex', 'ESx', 'POPs', 'ExCO3', 'EMOx'].

Control method type should be selected one from ['defect', 'alloy', 'structure control', 'atomic level dispersion', 'composite', 'surface/interface modification'].

Example 1: Abstract 1, Entity in Abstrct 1

Example 2:Abstract 2, Entity in Abstrct 2

========<Input Text>

=========Your REPLY:

Fig. 4 The prompt using in the entity extraction.

Data Description	Data Key Label	Data Type
DOI of the original paper	doi	string
Title of the original paper	title	string
Entity extracted from the paper	entity	string
Label of the entity	entity_label	string
Sentence where the entity is located	context	string
Target material data	`target_string`	list of strings
Equipment where the reaction is operated	`hardware`	list of strings
Material data of the preparation process	`reagent`	list of strings
Sequence of synthesis steps and corresponding conditions	`operation`	list of Objects (dict):
		-`string`: string
		-`vessel`: string
		-`reagent`: list of Objects[a]
		-`speed`: string
		-`temp`: list of Objects[b]
		-`time`: list of Objects[b]
		-`potential`: list of Objects[b]
		-`condition`: string
		-`stir`: boolean
		-`reflux`: boolean

Table 1. Format of each data record: description, key label, data type.

universities, etc. The benchmark corpus is publicly available at https://doi.org/10.57760/sciencedb.13290[40]. The extended corpus I and extended corpus II are publicly available at https://doi.org/10.57760/sciencedb.13292[41], where include other extendedcorpuscorpus exacted by LLM model. The two types of Corpus are provided as a file in CSV format, and the details of them are shown in Table 3. A complete dataset of 476 catalytic material synthesis processes is publicly available at https://doi.org/10.57760/sciencedb.13293[42].

Technical Validation

Extraction accuracy. To demonstrate the utility of the extended corpus, we first evaluated the model against other current state-of-the-art traditional entity extraction methods. We selected several generic neural network tagging models, including bi-directional LSTM layers with conditional random field (CRF) layer[33,43,44], bi-directional recurrent neural network Bi-GRU[45], and BERT model with CRF layer. We then chose a multi-feature based maximum entropy machine learning model[46] using two types of features, Parts-of-Speech features generated by GENIA Parts-of-Speech Tagger[47]and lexical features. Table 4 shows the results of the experimental comparison.

Operation Type	Condition attributes	Data description
ADDING	`-left_reagent`	The name and quantity of materials involved in the operation
	`-right_reagent`	The name and quantity of materials involved in the operation
	`-speed`	Speed of adding operations
	`-stir`	Stirring or not during addition operation
HEATING	`-vessel`	Vessel in which the heating operation takes place
	`-temp`	Final temperature for heating operations
	`-time`	Time for heating operations
	`-stir`	Stirring or not during heating
	`-reflux`	Whether the heating process requires reflux
CURING	`-condition`	Curing conditions, deliberately stated in the paragraph
	`-temp`	Temperature during curing operation
	`-time`	Time for curing
ELECTROCHEMICAL ANODIZATION	`-reagent`	Name of electrode solution
	`-potential`	Potential values for anodic oxidation reactions
	`-time`	Time for the electrochemical anodization
FILTERING	`-condition`	Filter conditions, the original sentence text extracted directly
	`-reagent`	Name of the reagent being filtered
DRYING	`-condition`	Dry conditions, the original sentence text extracted directly
DIPPING	`-left_reagent`	Name of material to be dipped
	`-right_reagent`	Name of material immersed in
	`-time`	Time for the immersion
REACTION	`-left_reagent`	Name and quantity of materials involved in the reaction
	`-right_reagent`	Name and quantity of materials involved in the reaction
	`-temp`	Temperature at the time of reaction
	`-time`	Time for the reaction
	`-reflux`	Whether reflux is required for the reaction

Table 2. Format of each synthesis operation record: operation type, condition attributes, data description.

We found that our constructed entity extraction model consistently outperforms other methods, achieving an overall F1 score of 85.16 in recognizing four coarse-grained categories of entities. This also demonstrated an advantage in the subsequent classification of fine-grained entities.

To estimate the quality of the synthesis process dataset, we had a human expert test 100 randomly selected entries. The human expert manually extracted the information provided in the synthesis paragraphs and compared the results with those extracted by the pipeline. Table 5 presents the accuracy statistics, which include the precision, recall, and F1 scores calculated from the test entries.

We also validated the entity recognition results of the LLMs in this paper. We validate the answers of the LLMs by an expert with 160 randomly selected entries, and ensure that each category has 20 test data. The evaluation result is shown in Table 6. The *Count* means the total amount of samples from different categories, the *Correct* means the number of correctly identified entities, and the *Existence* means the number of entities of this type does exist in the text input to the large model. It is worth mentioning that if there is indeed no corresponding entity in the text input to the large model, the situation where the large model answers empty should also be considered as correct recognition. Therefore, we use *Modified Correct* to remove the above influence. Ultimately, we utilize *Modified Correct* and *Count* to calculate the evaluation of LLMs, which is *Modified accuracy*. Using large models for entity recognition also causes significant time loss. We used two NVIDIA A100 GPU graphics processing units for entity recognition, and cost almost 10 hours to process 5,941 literature abstracts.

From the results, we can see that the LLMs perform better in entity extraction for numerical classes (faradaic efficiency, potential, etc.), but perform poorly in entity extraction for descriptive classes. This may be due to the objectivity of data entities, which reduces the possibility of hallucinations in large models.

Dataset mining. To present the recent trends in the development of CO_2 reduction electrocatalysts, we showcased and analyzed the information in the database. Firstly, we demonstrated the publication trends of CO_2 reduction electrocatalysts over the past 30 years. As depicted in Fig. 5a, articles on CO_2 reduction electrocatalysts have experienced a rapid surge since 2010, indicating the burgeoning interest of scientists in this field. Figure 5b illustrates the proportional distribution of various types of CO_2 reduction electrocatalysts. It is evident that the current research predominantly focuses on E (single metal), E/C (metal-carbon composites), E-M (binary or ternary metal systems), and EO_x (metal oxides), with a notable increase in attention toward E/C in recent years.

In addition to the overall development of electrocatalysts, another intriguing aspect lies in the correlation between catalysts and products, which is crucial for product-oriented catalyst design. Figure 6 presents an alluvial plot illustrating the intricate associations between catalysts and products. Notably, for clarity, less reported catalyst categories have not been included. E/C and E-M are favorable choices for generating CO, while E-M and EO_x exhibit the capability for formic acid production. For C_2 products, such as C_2H_4 and C_2H_5OH, both

Corpus Type	Benchmark Corpus	Extended Corpus I	Extended Corpus II
Entity Type			
Material	1,092	18,184	5,977
Regulation method	1,086	35,780	5,488
Product (including the second and third product)	1,340	19,080	6,700
Faradaic efficiency (including the Faradaic efficiency of second and third product)	1,135	3,496	3,152
Cell setup	435	—	170
Electrolyte	475	—	3,919
Synthesis method	228	476	—
Current density	393	—	3,852
Voltage	801	—	1,025
Total	6,985	77,016	30,283

Table 3. Summary of the three corpus.

Entity(freq. in test set)	MaxEnt	BiLSTM-CRF	BiGRU-CRF	BERT-CRF	BERT-BiLSTM-CRF
MATERIAL(92)	40.12	50.43	52.01	59.96	**60.59**
METHOD(97)	38.25	46.89	49.67	57.12	**58.02**
PRODUCT(94)	70.21	82.45	86.12	91.10	**92.34**
FARADAIC EFFICIENCY(62)	88.16	91.18	91.98	94.56	**96.82**
Macro-avg F1	51.26	66.90	68.02	71.48	**73.90**
Micro-avg F1	68.89	81.02	82.33	82.73	**85.16**

Table 4. Compare the F1 scores of entity recognition in various models.

Pipeline Component	Extraction Method	F1:(precision | recall)
Article filtering	Regular match	0.88:(0.84 | 0.93)
Synthesis paragraph classification	BERT classification	0.80:(0.82 | 0.78)
Materials entity recognition	BiLSTM + CRF (BERT embedding) & Regular match	0.96:(0.98 | 0.94) - materials
		0.84:(0.86 | 0.82) - targets
Synthesis actions	BiLSTM (Word2Vec embeddings)	0.89: (0.92 | 0.86)
Synthesis conditions	Rule-based	
-Temperature		0.95: (0.98 | 0.93)
-Time		0.96: (0.98 | 0.94)
-Potential		0.88: (0.91 | 0.86)
Material quantities	Rule-based	0.90: (0.93 | 0.87)

Table 5. Accuracy of synthesis information extraction models.

Entity	Count	Correct	Existence	Modified Correct	Modified accuracy
MATERIAL	20	15	17	15	75%
METHOD	20	13	19	13	65%
PRODUCT	20	17	17	17	85%
FARADAIC EFFICIENCY	20	11	11	18	90%
ELECTROLYTE	20	9	10	10	50%
POTENTIAL	20	7	7	16	80%
CURRENT DENSITY	20	7	7	12	60%
CELL SETUP	20	6	6	9	45%
OVERALL	160	85	94	110	68.75%

Table 6. The evaluation of entity recognition of LLMs.

E and EO_x are viable options. Furthermore, Fig. 6 also reveals some potential research topics that warrant further exploration. For instance, although a few catalysts demonstrate the ability to produce C_3 products, such as n-propanol and acetone, the optimal catalysts have yet to be well-established. While composite systems are gaining increasing attention, their advantages over individual compounds remain to be fully elucidated.

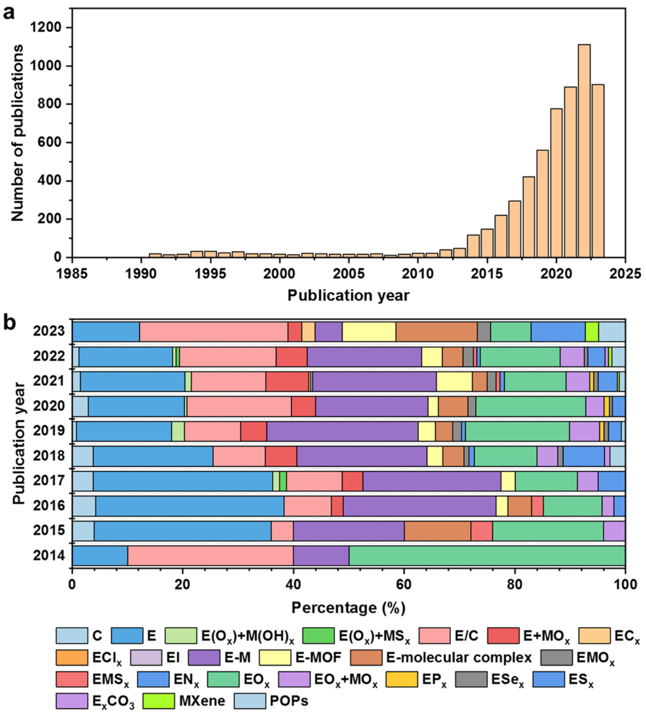


Fig. 5 (**a**) Histograms of the number of publications of CO_2 reduction electrocatalysts over the past thirty years. (**b**) Stacked histograms of the percentage of CO_2 reduction electrocatalysts in the last ten years.

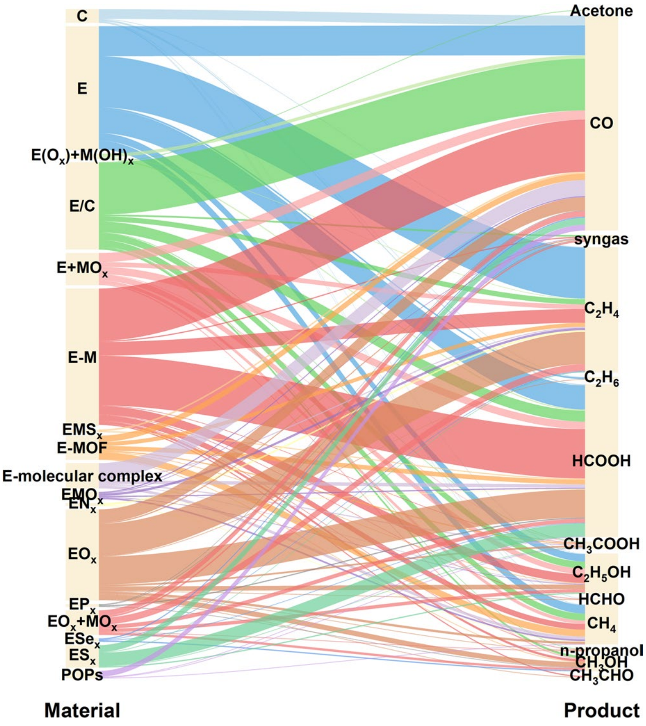


Fig. 6 Alluvial plot illustrating the relationships between catalysts and products.

Moreover, the type of metal, particularly the presence of Cu, is crucial for the performance of catalysts in CO_2 electroreduction. Therefore, we annotated whether the catalysts contained Cu in the database. To illustrate this contrast clearly, we generated doughnut charts to display the percentage of different products from several types of catalysts with or without Cu. As shown in Fig. 7a, the majority of the products for E/C are CO, while Cu/C can generate various C_1 and C_2 products. For single metal systems (Fig. 7b), the primary products of E

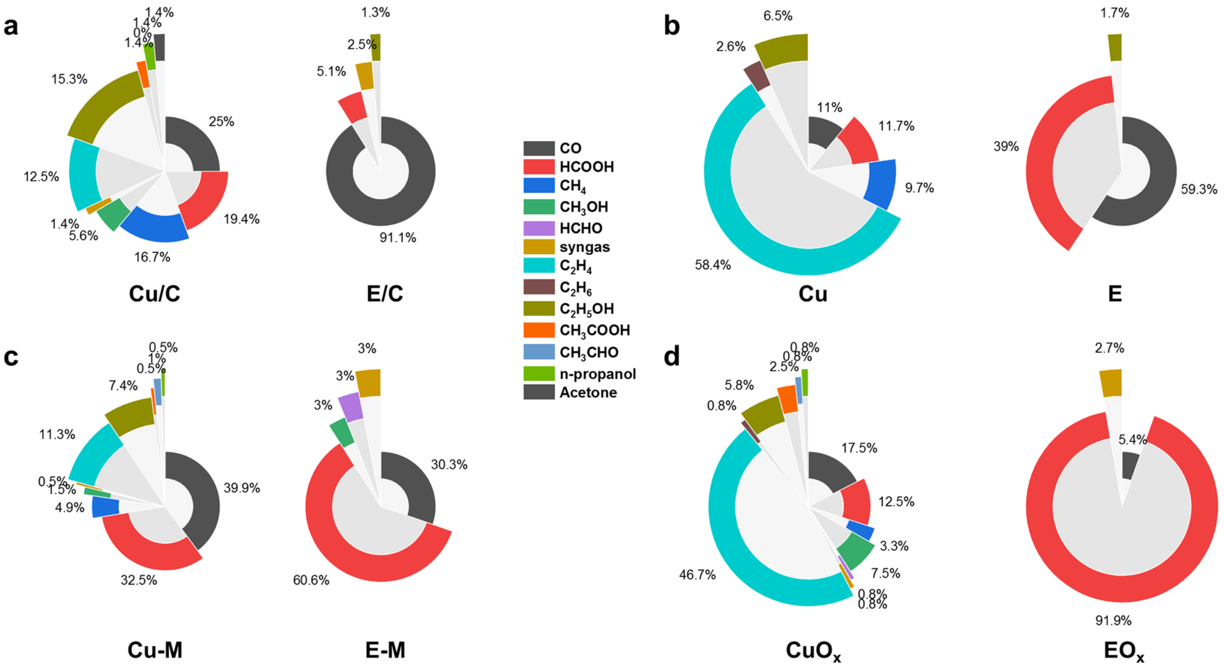


Fig. 7 Doughnut charts showing the percentage of different products of catalysts with or without Cu.

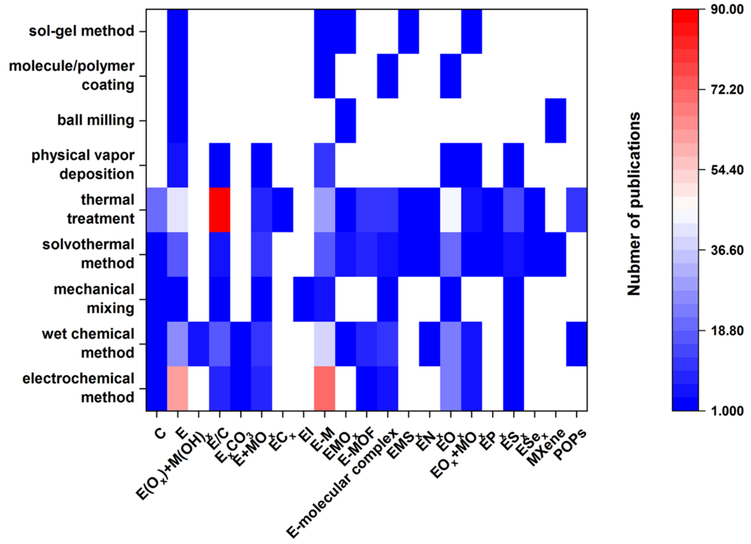


Fig. 8 Heatmap showing the number of publications of CO_2 electrocatalysts with different synthesis methods.

are C_1 products, whereas Cu yields predominantly C_2 products. In the case of binary or ternary metal systems, Cu-M exhibits a stronger capability for producing C_2 products compared to E-M. Regarding metal oxides, the products of EO_x are predominantly formic acid, while CuO_x yields primarily C_2H_4. These findings underscore the significant impact of the presence of Cu on the selectivity of C_2 products for catalysts.

The choice of synthesis method also has a significant impact on the performance of catalysts, so we analyzed the correlation between catalysts and synthesis methods. As shown in Fig. 8, thermal treatment and solvothermal methods are the two most widely used material synthesis methods. In addition, different catalysts also have their conventional synthesis methods. For example, the synthesis of Cu/C, which usually refers to carbon-coated metal nanoparticles or anchored single atoms, is mainly through thermal treatment. The synthesis of E and E-M is mainly electrochemical methods, especially electrochemical reduction treatment. For EO_x and its composites, the solvothermal method, wet chemical method, and electrochemical method are commonly used methods. This analysis is helpful for the screening of target catalyst synthesis methods.

The database encompasses various catalyst types and diverse regulation strategies, which can be utilized to guide the design and optimization of novel catalysts. One feasible approach involves integrating multiple strategies by drawing inspiration from well-performing catalysts and regulation methods in the literature, thus facilitating the development of highly efficient catalysts. For example, CuS serves as a potential efficient catalyst for C_2H_4 production, while nano-sized polymer coatings can enhance the selectivity of C_2H_4. Consequently, CuS nanoparticles coated with an a-few-nm-thick polymer layer represent an effective method for selectively producing C_2H_4. Similarly, coupling Cu_2O nanocrystals with (111) facets with functionalized graphene nanosheets can be employed for C_2H_5OH production. Furthermore, utilizing fine-tuned domain LLMs is also a viable strategy for developing novel catalysts, and further efforts are required in fine-tuning LLMs and prompt engineering.

Code availability

The scripts utilized to parse articles and extract entities are home-written codes which are publicly available at the github repository https://github.com/cxqwindy/CO2_reduction_electrocatalysts_db. The underlying machine-learning libraries used in this project are all open-source: rxn4chemistry(rxn4chemistry), ChemDataExtractor (chemdataextractor.org)[37], gensim (radimrehurek.com)[48], PyMuPDF(PyMuPDF), Pytorch (www.pytorch.org) and scikit-learn (scikit-learn.org)[49].

Received: 18 December 2023; Accepted: 22 March 2024;
Published online: 06 April 2024

References

1. Birdja, Y. Y. *et al*. Advances and challenges in understanding the electrocatalytic conversion of carbon dioxide to fuels. *Nat. Energy* **4**, 732–745 (2019).
2. Zhong, M. *et al*. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
3. Gao, Y., Wang, L., Chen, X., Du, Y. & Wang, B. Revisiting electrocatalyst design by a knowledge graph of Cu-based catalysts for CO_2 reduction. *ACS Catal.* **13**, 8525–8534 (2023).
4. Qiao, J., Liu, Y., Hong, F. & Zhang, J. A review of catalysts for the electroreduction of carbon dioxide to produce low-carbon fuels. *Chem. Soc. Rev.* **43**, 631–675 (2014).
5. Zheng, T., Jiang, K. & Wang, H. Recent advances in electrochemical CO_2-to-CO conversion on heterogeneous catalysts. *Adv. Mater.* **30**, 1802066 (2018).
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
7. Peng, J. *et al*. Human- and machine-centred designs of molecules and materials for sustainability and decarbonization. *Nat. Rev. Mater.* **7**, 991–1009 (2022).
8. He, T. *et al*. Similarity of precursors in solid-state synthesis as text-mined from scientific literature. *Chem. Mater.* **32**, 7861–7873 (2020).
9. Huang, S. & Cole, J. M. A database of battery materials auto-generated using ChemDataExtractor. *Sci. Data* **7**, 260 (2020).
10. Paula, A. J. *et al*. Machine learning and natural language processing enable a data-oriented experimental design approach for producing biochar and hydrochar from biomass. *Chem. Mater.* **34**, 979–990 (2022).
11. Kononova, O. *et al*. Text-mined dataset of inorganic materials synthesis recipes. *Sci. data* **6**, 203 (2019).
12. Wang, L. *et al*. A corpus of CO_2 electrocatalytic reduction process extracted from the scientific literature. *Sci. Data* **10**, 175 (2023).
13. Wang, S. *et al*. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428* (2023).
14. Alkaissi, H. & McFarlane, S. I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus* **15** (2023).
15. Bang, Y. *et al*. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 675-718 (2023).
16. Azamfirei, R., Kudchadkar, S. R. & Fackler, J. Large language models and the perils of their hallucinations. *Crit. Care* **27**, 1–2 (2023).
17. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
18. Kumar, K. Geotechnical parrot tales (GPT): Overcoming GPT hallucinations with prompt engineering for geotechnical applications. *arXiv preprint arXiv:2304.02138* (2023).
19. Polak, M. P. & Morgan, D. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *arXiv preprint arXiv:2303.05352* (2023).
20. Hiszpanski, A. M. *et al*. Nanomaterial synthesis insights from machine learning of scientific articles by extracting, structuring, and visualizing knowledge. *J. Chem. Inf. Model.* **60**, 2876–2887 (2020).
21. Liu, R. & McKie, J. Pymupdf. Available at http://pymupdf.readthedocs.io/en/latest/ (2018).
22. Cruse, K. *et al*. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. *Sci. Data* **9**, 234 (2022).
23. Huo, H. *et al*. Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput. Mater.* **5**, 62 (2019).
24. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
25. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
26. Rebholz-Schuhmann, D. *et al*. The calbc silver standard corpus for biomedical named entities-a study in harmonizing the contributions from four independent named entity taggers. In *LREC 2010-7th International Conference on Language Resources and Evaluation* (CELI Language & Informat Technol; European Media Lab GmBH; Quaero; META, 2010).
27. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y. & Liang, X. doccano: Text annotation tool for human. *Software available from* https://github.com/doccano/doccano 34 (2018).
28. Brown, T. *et al*. Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).
29. Radford, A. *et al*. Language models are unsupervised multitask learners. *OpenAI blog* **1**, 9 (2019).
30. Radford, A. *et al*. Improving language understanding by generative pre-training. (2018).
31. Beltagy, I., Lo, K. & Cohan, A. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*,3615-3620 (2019).
32. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
33. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. & Dyer, C. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,260-270 (2016).
34. Corbett, P. & Murray-Rust, P. High-throughput identification of chemistry in life science texts. In *International Symposium on Computational Life Science*, 107–118 (Springer, 2006).
35. Hettne, K. M. *et al*. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* **25**, 2983–2991 (2009).

36. Vaucher, A. C. *et al*. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **11**, 3601 (2020).
37. Swain, M. C. & Cole, J. M. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.* **56**, 1894–1904 (2016).
38. Honnibal, M. & Johnson, M. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, 1373–1378 (2015).
39. Teller, V. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition* (2000).
40. Wang, L. *et al*. Benchmark corpus of CO_2 reduction electrocatalysts and synthesis procedures, *ScienceDB*, https://doi.org/10.57760/sciencedb.13290 (2023).
41. Wang, L. *et al*. The extended corpus of CO_2 reduction electrocatalysts and synthesis procedures, *ScienceDB*, https://doi.org/10.57760/sciencedb.13292 (2023).
42. Wang, L. *et al*. A complete dataset of 476 catalytic material synthesis processes. *ScienceDB* at https://doi.org/10.57760/sciencedb.13293 (2023).
43. Ma, X. & Hovy, E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. *arXiv preprint arXiv:1603.01354* (2016).
44. Plank, B., Søgaard, A. & Goldberg, Y. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529* (2016).
45. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).
46. Borthwick, A. E. *A maximum entropy approach to named entity recognition* (New York University, 1999).
47. Tsuruoka, Y. & Tsujii, J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In *proceedings of human language technology conference and conference on empirical methods in natural language processing*, 467–474 (2005).
48. Řehřek, R. & Sojka, P. *Software framework for topic modelling with large corpora*. (2010).
49. Pedregosa, F. *et al*. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

This work was supported by the National Key Research and Development Plan of China under Grant No. 2022YFF0712200, 2022YFF0711900 and 2021YFA1202802, the Natural Science Foundation of China under Grant No. T2322027, Information Science Database in National Basic Science Data Center under Grant No.NBSDC-DB-25, the Young Elite Scientists Sponsorship Program by Beijing Association for Science and Technology (BYESS2023410), the CAS Pioneer Hundred Talents Program and Youth Innovation Promotion Association CAS.

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.D. or B.W.

Reprints and permissions information is available at www.nature.com/reprints.

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