Fine-tuning Large Language Models for Chemical Text Mining

Wei Zhang[1,2,#], Qinggong Wang[3,#], Xiangtai Kong[1,2], Jiacheng Xiong[1,2], Shengkun Ni[1,2], Duanhua Cao[1,4], Buying Niu[1,2], Mingan Chen[1,5,6], Runze Zhang[1,2], Yitian Wang[1,2], Lehan Zhang[1,2], Xutong Li[1,2], Zhaoping Xiong[7], Qian Shi[6], Ziming Huang[8], Zunyun Fu[1,*], Mingyue Zheng[1,2,3,*]

[1]Drug Discovery and Design Canter, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China

[2]University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China

[3]Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China

[4]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China

[5]School of Physical Science and Technology, ShanghaiTech University, Shanghai 201210, China

[6]Lingang Laboratory, Shanghai 200031, China

[7]ProtonUnfold Technology Co., Ltd, Suzhou, China

[8]Medizinische Klinik und Poliklinik I, Klinikum der Universität München, Ludwig-Maximilians-Universität, Munich, Germany

[#]Wei Zhang and Qinggong Wang contributed equally to this study.

*Correspondence should be addressed to:

Mingyue Zheng: myzheng@simm.ac.cn

Zunyun Fu: fuzunyun@simm.ac.cn

Abstract

Extracting knowledge from complex and diverse chemical texts is a pivotal task for both experimental and computational chemists. The task is still considered to be extremely challenging due to the complexity of the chemical language and scientific literature. This study explored the power of fine-tuned large language models (LLMs) on five intricate chemical text mining tasks: compound entity recognition, reaction role labelling, metal-organic framework (MOF) synthesis information extraction, nuclear magnetic resonance spectroscopy (NMR) data extraction, and the conversion of reaction paragraph to action sequence. The fine-tuned LLMs models demonstrated impressive performance, significantly reducing the need for repetitive and extensive prompt engineering experiments. For comparison, we guided GPT-3.5 and GPT-4 with prompt engineering and fine-tuned GPT-3.5 as well as other open-source LLMs such as Llama2, T5, and BART. The results showed that the fine-tuned GPT models excelled in all tasks. It achieved exact accuracy levels ranging from 69% to 95% on these tasks with minimal annotated data. It even outperformed those task-adaptive pre-training and fine-tuning models that were based on a significantly larger amount of in-domain data. Given its versatility, robustness, and low-code capability, leveraging fine-tuned LLMs as flexible and effective toolkits for automated data acquisition could revolutionize chemical knowledge extraction.

Introduction

Chemical text mining is a crucial foundation in chemical research. It creates extensive databases that provide access to physicochemical properties and synthetic routes for experimental chemists. Additionally, it accumulates rich data and insights for computational chemists to use for modelling and predicting. More than just extracting information from chemical texts, the rule-based transformation of chemical text is particularly interesting. For instance, synthetic procedures can be converted into action sequences[1,2] or programming languages[3-5]. This allows them to be understood and executed by robotic systems for automated syntheses.

However, converting structured data from intricate scientific literature is a challenging task, especially due to the complexity and heterogeneity of chemical language. As a result, a number of text-mining tools have been developed. For instance, ChemDataExtractor[6,7] was created to extract chemical entities and their associated properties, measurements and relationships from chemical documents, using unsupervised word clustering, conditional random fields, rule-based grammars and dictionary matching. ChemRxnExtractor[8], a BERT-like model, was designed to extract the product and label associated reaction roles such as reactant, catalyst, solvent, and temperature from paragraphs of synthesis experiments. Vaucher et. al.[1,2] developed task-adaptive pre-trained transformers to convert the synthesis protocol paragraphs into action sequences. SynthReader[3] was built to convert literature syntheses to executable XDL formats, containing a series of domain-specific algorithms with predefined rules. Historically, the focus has been on designing models and algorithms specific to certain tasks, requiring extensive domain knowledge and sophisticated data processing These tools, challenging to adapt for diverse extraction tasks, often require complementary collaboration to manage complex information extraction tasks, thus limiting their versatility and practicality.

Recently, large language models (LLMs), represented by ChatGPT released in November 2022, have shown the potential of Artificial General Intelligence (AGI). LLMs, such as GPT-3.5 and GPT-4, can generate logical insights or content that meets requirements based on human instructions. We are entering a new era where AGI and medicinal chemists might work together. There have been some assessments of ChatGPT's chemistry capabilities, including tasks like synonym transformation, property prediction, retrosynthesis, and molecule design[9-11]. However, LLMs tend to "hallucinate", meaning they generate unintended text that misaligns with established facts and real-world knowledge[12,13]. Moreover, objectively evaluating the results of open-ended questions remains a significant challenge.

At this juncture, LLMs may still find it difficult to accurately answer factual and knowledge-based questions. However, using LLMs for knowledge extraction tasks should greatly alleviate hallucination and fully leverage their powerful text comprehension and processing capabilities, making them promising universal tools for chemical text mining. For instance, Zheng et al.[14] used prompt engineering to guide ChatGPT in extracting information about metal-organic framework (MOF) synthesis. Patiny et al.[15] tried to use ChatGPT to extract FAIR (Findable, Accessible, Interoperable, Reusable) data from publications. However, their approach of using LLMs simply based on prompt engineering tend to achieve poor performance in exact accuracy. According to the biomedical benchmark study by Chen et al.[16], ChatGPT performed significantly worse on biomedical text mining compared to existing models. These findings seem contradicts the common belief in the LLMs' superior comprehension abilities. Either way, LLMs have limitations due to their model architecture and memory, including a maximum length of prompt tokens. Additionally, human expressions can be ambiguous, incomplete, vague, and difficult to refine. Outputs may not strictly adhere to formatting requirements, leading to misunderstanding and poor performance in mining complex text, such as patents or scientific literature. Therefore, zero-shot or few-shot prompts are often insufficient to address the diversity of scenarios and cannot guarantee the quality of extracted data.

In this study, we explore the effectiveness of fine-tuning LLMs on five challenging tasks in chemical text mining: compound entity recognition, reaction role annotation, metal-organic framework (MOF) synthesis information extraction, nuclear magnetic resonance spectroscopy

(NMR) data extraction, and conversion reaction paragraphs into action sequences. We found that fine-tuning GPT models significantly enhances performance in chemical text mining tasks, compared to prompt-only version, while also reducing dependency on the repetitive and extensive prompt engineering experiments. Meanwhile, we also evaluated other prevalent generative pre-trained language models, such as Llama2[17], T5[18], and BART[19]. Among these, the fine-tuned ChatGPT (gpt-3.5-turbo) models achieved state-of-the-art (SOTA) performance across all five tasks. Remarkably, it even outperformed models that have been trained specifically for each task and subsequently fine-tuned, based on a significantly larger amount of in-domain data. This study highlights the potential of fine-tuning LLMs to revolutionize complex knowledge extraction with their versatility, robustness, and low code capability. Fine-tuned LLMs can be easily generalizable and can optimize the labor-intensive and time-consuming data collection workflow, even when trained with few data. This will accelerate the discovery and creation of novel substances, making them powerful tools for universal use.

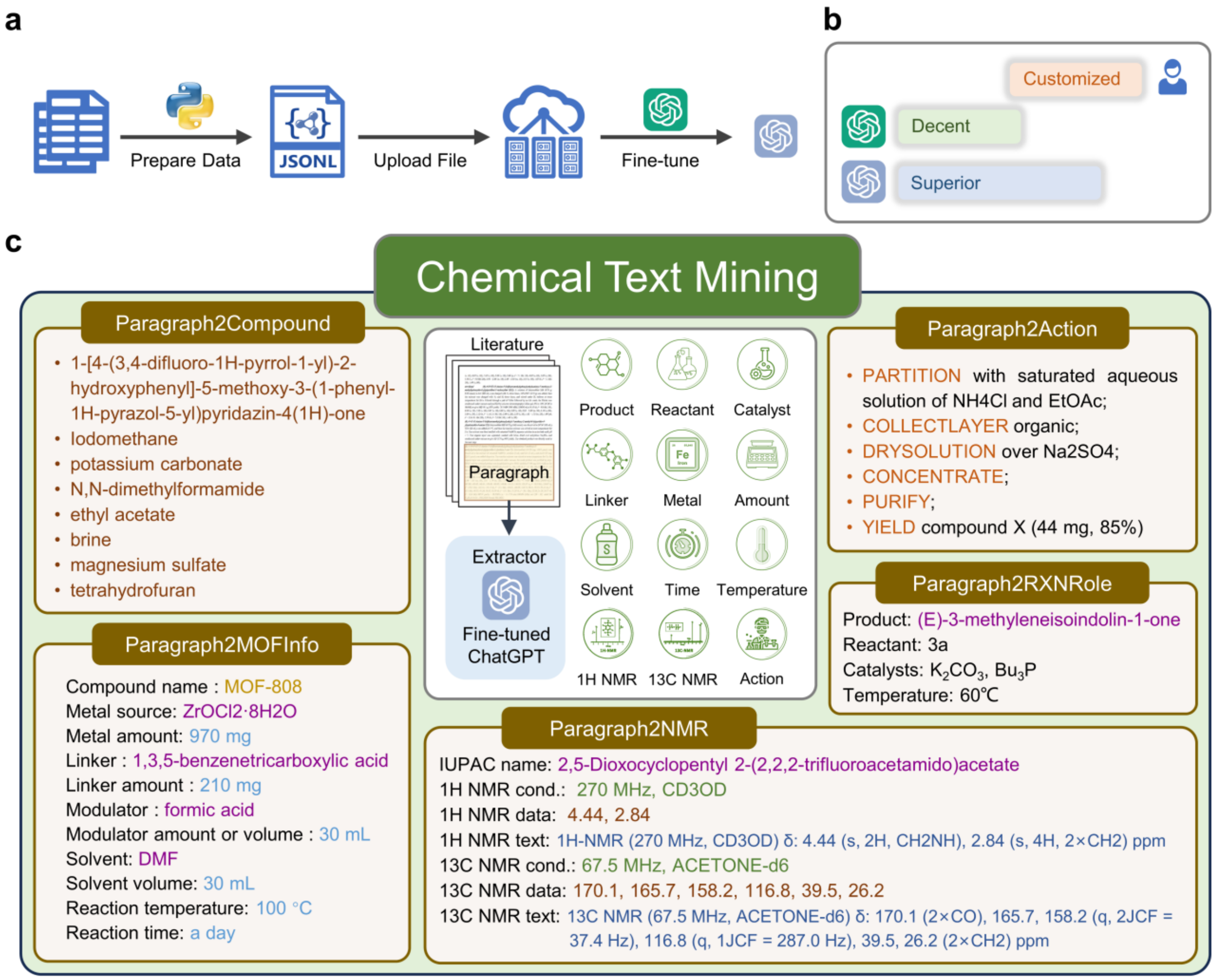

a
Prepare Data
JSONL
Upload File
Fine-tune
b
Customized
Decent
Superior
c
Chemical Text Mining
Paragraph2Compound
• 1-[4-(3,4-difluoro-1H-pyrrol-1-yl)-2-hydroxyphenyl]-5-methoxy-3-(1-phenyl-1H-pyrazol-5-yl)pyridazin-4(1H)-one
• Iodomethane
• potassium carbonate
• N,N-dimethylformamide
• ethyl acetate
• brine
• magnesium sulfate
• tetrahydrofuran
Literature
Paragraph
Extractor
Fine-tuned ChatGPT
Product
Reactant
Catalyst
Linker
Metal
Amount
Solvent
Time
Temperature
1H NMR
13C NMR
Action
Paragraph2Action
• PARTITION with saturated aqueous solution of NH4Cl and EtOAc;
• COLLECTLAYER organic;
• DRYSOLUTION over Na2SO4;
• CONCENTRATE;
• PURIFY;
• YIELD compound X (44 mg, 85%)
Paragraph2RXNRole
Product: (E)-3-methyleneisoindolin-1-one
Reactant: 3a
Catalysts: K_2CO_3, Bu_3P
Temperature: 60℃
Paragraph2MOFInfo
Compound name : MOF-808
Metal source: ZrOCl2·8H2O
Metal amount: 970 mg
Linker : 1,3,5-benzenetricarboxylic acid
Linker amount : 210 mg
Modulator : formic acid
Modulator amount or volume : 30 mL
Solvent: DMF
Solvent volume: 30 mL
Reaction temperature: 100 °C
Reaction time: a day
Paragraph2NMR
IUPAC name: 2,5-Dioxocyclopentyl 2-(2,2,2-trifluoroacetamido)acetate
1H NMR cond.: 270 MHz, CD3OD
1H NMR data: 4.44, 2.84
1H NMR text: 1H-NMR (270 MHz, CD3OD) δ: 4.44 (s, 2H, CH2NH), 2.84 (s, 4H, 2×CH2) ppm
13C NMR cond.: 67.5 MHz, ACETONE-d6
13C NMR data: 170.1, 165.7, 158.2, 116.8, 39.5, 26.2
13C NMR text: 13C NMR (67.5 MHz, ACETONE-d6) δ: 170.1 (2×CO), 165.7, 158.2 (q, 2JCF = 37.4 Hz), 116.8 (q, 1JCF = 287.0 Hz), 39.5, 26.2 (2×CH2) ppm

Fig. 1. | Schematics of fine-tuning ChatGPT for chemical text mining. a, The pipeline of fine-tuning ChatGPT on proprietary data. The green OpenAI logo symbolizes official gpt-3.5-turbo, while the blue one symbolizes fine-tuned gpt-3.5-turbo. **b,** Supervised fine-tuned LLMs outperforms prompt-only LLMs in some customized scenarios. **c,** Illustration of cheminformatics insights to be extracted from paragraph. And illustration of the five practical tasks in chemical text mining with respective example outputs, including Paragraph2Compound, Paragraph2RXNRole, Paragraph2MOFInfo, Paragraph2NMR, and Paragraph2Action.

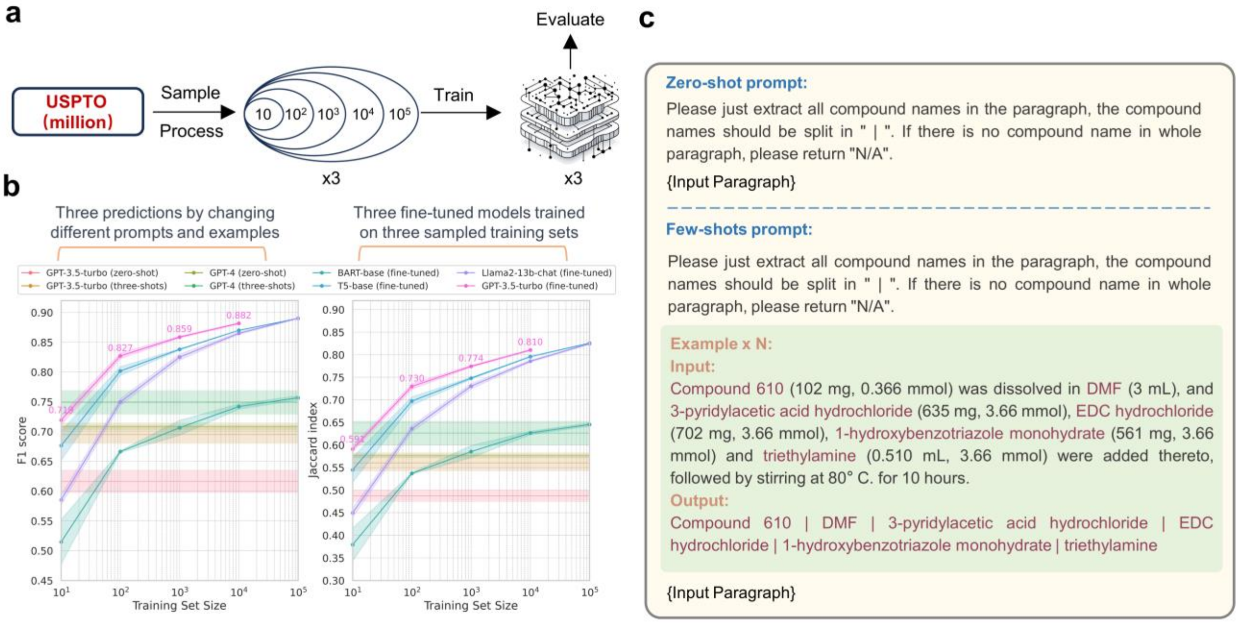
Results & Discussion

Overview of Chemical Text Mining Tasks

Given the complex and diverse information embedded in chemical literature, we designed five extraction tasks to demonstrate the potential and practicality of LLMs in chemical text mining (Fig. 1). Paragraph2Compound task is a relatively simple name entity recognition task, to extract all chemical compound entities from the given paragraph. Paragraph2RXNRole task is to label the reaction roles including product, reactant, catalyst, temperature, solvent, time, and yield in the paragraph. Paragraph2MOFInfo task is to extract all MOF synthesis conditions including compound name, metal source, metal amount, linker, linker amount, modulator, modulator amount or volume, solvent, solvent volume, reaction temperature and reaction time. Paragraph2NMR task is to extract the IUPAC name, experimental condition including frequency and solvent as well as chemical shift data for both 1H NMR and 13C NMR spectra. Paragraph2Action task is to convert experimental procedures to structured synthetic steps (action sequences). All tasks are unified to sequence-to-sequence formats to facilitate the uses of LLMs. More details can be found in the Methods section.

Paragraph2Compound—Extract All Chemical Compound Entities.

Fig. 2a illustrates the process of random sampling from millions of paragraph-entities pairs, which refer to UPSTO annotations. It starts by randomly selecting 100,000 samples, then choosing 10,000 from them, followed by randomly picking 1,000, then 100, and finally 10. This sampling process ensures each smaller subset is included in the larger one, with each subset used for individual training. Fig. 2b demonstrates the performance of prompt-only models and fine-tuned models, which are evaluated on a consistent evaluation set of 1,000 samples across varying training data sizes. These results are obtained from three independent trials. In the case of prompt-only models, randomness is intentionally introduced by altering the prompt and examples (Fig. 2c, Supplementary Fig. 2). Given the task’s straightforward nature and clear instructions, even the prompt-only language models achieved decent F1 scores over 0.6. For fine-tuned models, the sampling and training process for the training set is repeated three times, as depicted in Fig. 2a. As shown in Fig. 2b, all fine-tuned models demonstrate a performance improvement, especially in terms of the F1 score and Jaccard index, proportional to the increase in dataset size. These models outperform the prompt-only models designed for this task. When the training data size is substantial enough, the F1 scores of GPT-3.5-turbo, Llama2, and T5 can reach close to 0.9, and the Jaccard index can approach 0.8. Notably, gpt-3.5-turbo, when fine-tuned, showed minimal fluctuations and superior performance. However, it is essential to emphasize that the cost of fine-tuning gpt-3.5-turbo increased tenfold with each tenfold increase in data volume. Our experimentation with gpt-3.5-turbo were capped at 10,000 training samples for 3 epochs due to OpenAI's limitations, resulting in a nearly 90-dollar expense—a low cost-effective investment in computational resources. In contrast, other fine-tuned language models have displayed notable cost advantages in this simple task.



Zero-shot prompt:

Please just extract all compound names in the paragraph, the compound names should be split in " | ". If there is no compound name in whole paragraph, please return "N/A".

{Input Paragraph}

Few-shots prompt:

Please just extract all compound names in the paragraph, the compound names should be split in " | ". If there is no compound name in whole paragraph, please return "N/A".

Example x N:

Input:

Compound 610 (102 mg, 0.366 mmol) was dissolved in DMF (3 mL), and 3-pyridylacetic acid hydrochloride (635 mg, 3.66 mmol), EDC hydrochloride (702 mg, 3.66 mmol), 1-hydroxybenzotriazole monohydrate (561 mg, 3.66 mmol) and triethylamine (0.510 mL, 3.66 mmol) were added thereto, followed by stirring at 80° C. for 10 hours.

Output:

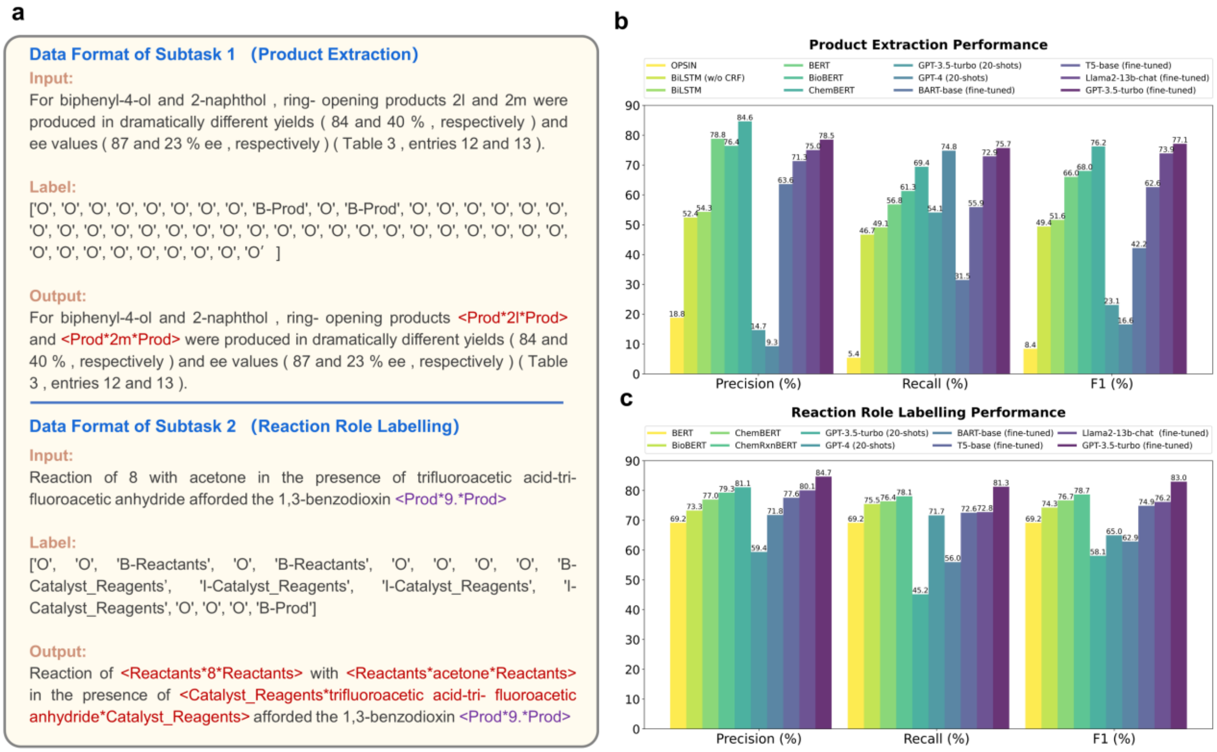
Compound 610 | DMF | 3-pyridylacetic acid hydrochloride | EDC hydrochloride | 1-hydroxybenzotriazole monohydrate | triethylamine

{Input Paragraph}

Fig. 2. | Design and Performance for Paragraph2Compound task. a, The workflow of sampling and training based on USPTO dataset for Paragraph2Compound task. **b**, The performance of different models across varying size of training set. The data point and the shaded areas represent respectively the mean values and standard deviations derived from three independent trials. **c**, Example of the zero-shot and three-shots prompts utilized for Paragraph2Compound task.

Paragraph2RXNRole—Product Extraction and Reaction Role Labelling.

According to Guo et al.[8], the Paragraph2RXNRole task comprises two subtasks. The first is to extract the central product, and the second is to label the associated reaction roles within specified paragraphs (Fig. 3a). For these tasks, Guo et al. developed two-stage BERT-like token-multi-classification models. To enable a fair comparison with generative language models, we converted the data into sequence-to-sequence formats by adding <Role*Compound*Role> annotations to the input paragraphs. We then converted the language models' outputs back into lists of BIO-tags, followed by post-processing to align with the original BIO-tags labels for assessment. Notably, even though utilizing prompt engineering with 20-shots examples (Supplementary Fig. 3, 4), GPT-3.5 and GPT-4 perform poor on two Paragraph2RXNRole tasks, which may result from the complicated syntax cases and limited context length (Fig. 3b, 3c). However, the fine-tuned GPT models perform well. For product extraction, the fine-tuned gpt-3.5-trubo (best over one epoch) achieved a F1 score of 77.1%, slightly surpassing the previous SOTA approach, ChemBERT, which scored 76.2% (Fig. 3b). For reaction role labelling, the fine-tuned gpt-3.5-trubo (best over five epochs) achieved a F1 score of 83.0%, significantly outperforming the previous SOTA approach, ChemRxnBERT, which scored 78.7% (Fig. 3c). It's notable that the fine-tuned gpt-3.5-trubo models, which cost only $1 and $5 respectively, demonstrated extremely high cost-effectiveness with small training datasets. In contrast, ChemBERT was domain-adaptive pre-trained on 9,478,043 sentences from 200,000 journal articles, and ChemRxnBERT was further task-adaptive trained on 944,733 reaction-inclusive sentences. We should also mention that the outputs of fine-tuned GPTs and Llama2 align almost perfectly with the input text, with 100% and 99% post-processing-free ratios respectively. On the other hand, most outputs of fine-tuned T5 and BART require additional alignment due to their tokenization and vocabulary limitations, with a ratio of only 31% that does not require post-processing. Even after post-processing, the F1 scores of T5 and BART were significantly lower than those token-classification BERT-like models or large language models such as GPTs and Llama2.



Data Format of Subtask 1 (Product Extraction)

Input:

For biphenyl-4-ol and 2-naphthol , ring- opening products 2l and 2m were produced in dramatically different yields (84 and 40 % , respectively) and ee values (87 and 23 % ee , respectively) (Table 3 , entries 12 and 13).

Label:

['O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'B-Prod', 'O', 'B-Prod', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O', 'O’]

Output:

For biphenyl-4-ol and 2-naphthol , ring- opening products <Prod*2l*Prod> and <Prod*2m*Prod> were produced in dramatically different yields (84 and 40 % , respectively) and ee values (87 and 23 % ee , respectively) (Table 3 , entries 12 and 13).

Data Format of Subtask 2 (Reaction Role Labelling)

Input:

Reaction of 8 with acetone in the presence of trifluoroacetic acid-tri-fluoroacetic anhydride afforded the 1,3-benzodioxin <Prod*9.*Prod>

Label:

['O', 'O', 'B-Reactants', 'O', 'B-Reactants', 'O', 'O', 'O', 'O', 'B-Catalyst_Reagents', 'I-Catalyst_Reagents', 'I-Catalyst_Reagents', 'I-Catalyst_Reagents', 'O', 'O', 'O', 'B-Prod']

Output:

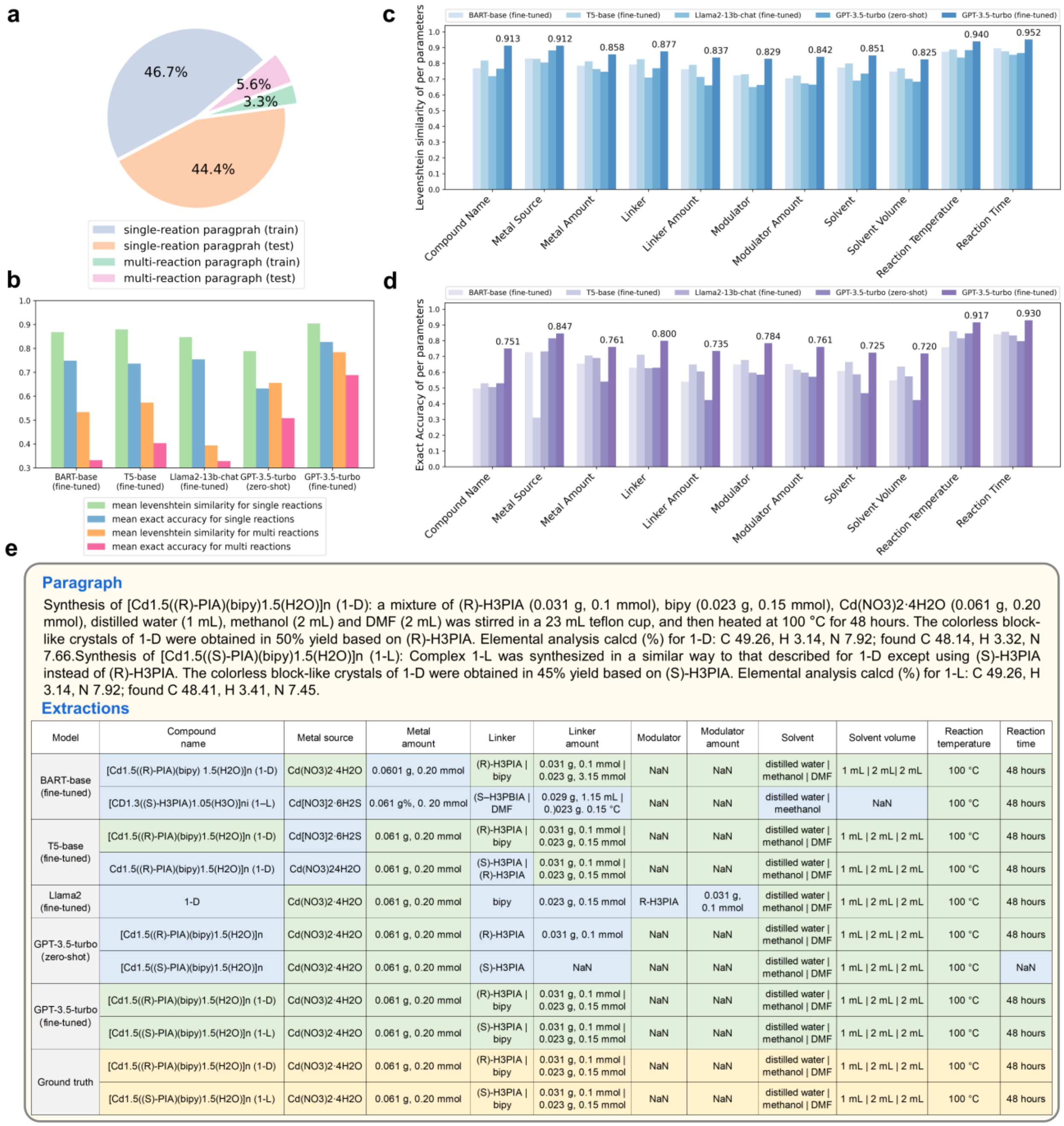
Reaction of <Reactants*8*Reactants> with <Reactants*acetone*Reactants> in the presence of <Catalyst_Reagents*trifluoroacetic acid-tri- fluoroacetic anhydride*Catalyst_Reagents> afforded the 1,3-benzodioxin <Prod*9.*Prod>

Fig. 3. | Design and Performance for Paragraph2RXNRole task. a, Data formats of two subtasks in paragraph2RXNRole task. **b**, Performance of product extraction. **c**, Performance of reaction role labelling.

Paragraph2MOFInfo—Extraction of MOF Synthesis Information.

Our re-annotated dataset for the Paragraph2MOFInfo task displayed in Fig. 4a, mostly contains single reaction paragraphs with a few featuring multiple reactions. We used Levenshtein similarity and exact accuracy as metrics to objectively assess the models' ability to extract formatted data that fully complies with customized requirements in the task. This approach is more objective and accurate with less manual intervention, compared to the manual analysis and evaluation used by Zheng et al.[14]. The fine-tuned gpt-3.5-turbo significantly outperforms the gpt-

3.5-turbo with prompt engineering, improving exact accuracy by over 20% for both single and multiple reactions (Fig. 4b, Supplementary Fig. 5). It also surpasses other fine-tuned models, especially when handling complex multi-reaction paragraphs. Exact accuracy rates for single and multiple reactions are 82.7% and 68.8%, respectively (Fig. 4b). As depicted in Fig. 4c and Fig. 4d, while most models achieve high Levenshtein similarity across the 11 parameters, only a few maintain high exact accuracy, which is the golden metric that we mainly focus on. Considering that some MOF synthesis paragraphs may include multiple reactions, we provide an example of multi-reaction extraction by various models in Fig. 4e. The paragraph includes two reactions, the first with (R)-H3PIA and bipy as linkers, providing all reaction conditions explicitly, and the second with the substitution of (R)-H3PIA with (S)-H3PIA, keeping all other conditions unchanged. Most models successfully interpreted the semantics and extracted two reactions from the MOF synthesis paragraph. However, only the fine-tuned ChatGPT perfectly extracted information that matched our annotated ground truth. Other models showed varying degrees of incompleteness, particularly with items involving multiple components and their quantities.



Paragraph

Synthesis of [Cd1.5((R)-PIA)(bipy)1.5(H2O)]n (1-D): a mixture of (R)-H3PIA (0.031 g, 0.1 mmol), bipy (0.023 g, 0.15 mmol), Cd(NO3)2·4H2O (0.061 g, 0.20 mmol), distilled water (1 mL), methanol (2 mL) and DMF (2 mL) was stirred in a 23 mL teflon cup, and then heated at 100 °C for 48 hours. The colorless block-like crystals of 1-D were obtained in 50% yield based on (R)-H3PIA. Elemental analysis calcd (%) for 1-D: C 49.26, H 3.14, N 7.92; found C 48.14, H 3.32, N 7.66.Synthesis of [Cd1.5((S)-PIA)(bipy)1.5(H2O)]n (1-L): Complex 1-L was synthesized in a similar way to that described for 1-D except using (S)-H3PIA instead of (R)-H3PIA. The colorless block-like crystals of 1-D were obtained in 45% yield based on (S)-H3PIA. Elemental analysis calcd (%) for 1-L: C 49.26, H 3.14, N 7.92; found C 48.41, H 3.41, N 7.45.

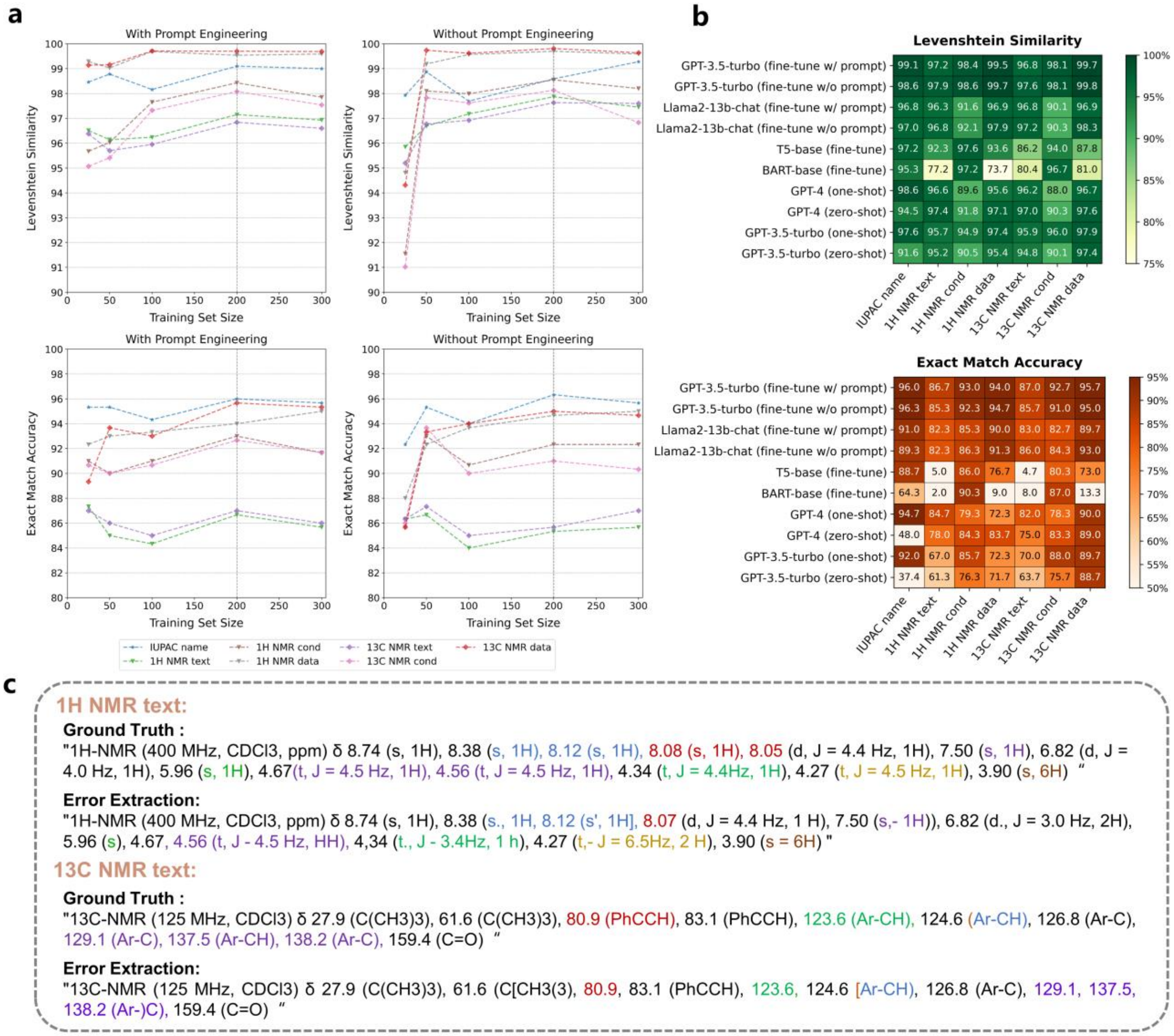
Extractions

Model	Compound name	Metal source	Metal amount	Linker	Linker amount	Modulator	Modulator amount	Solvent	Solvent volume	Reaction temperature	Reaction time
BART-base (fine-tuned)	[Cd1.5((R)-PIA)(bipy) 1.5(H2O)]n (1-D)	Cd(NO3)2·4H2O	0.0601 g, 0.20 mmol	(R)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 3.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL| 2 mL	100 °C	48 hours
	[CD1.3((S)-H3PIA)1.05(H3O)]ni (1–L)	Cd[NO3]2·6H2S	0.061 g%, 0. 20 mmol	(S–H3PBIA | DMF	0.029 g, 1.15 mL | 0.)023 g. 0.15 °C	NaN	NaN	distilled water | meethanol	NaN	100 °C	48 hours
T5-base (fine-tuned)	[Cd1.5((R)-PIA)(bipy)1.5(H2O)]n (1-D)	Cd[NO3]2·6H2S	0.061 g, 0.20 mmol	(R)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
	Cd1.5((R)-PIA)(bipy)1.5(H2O)]n (1-D)	Cd(NO3)24H2O	0.061 g, 0.20 mmol	(S)-H3PIA | (R)-H3PIA	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
Llama2 (fine-tuned)	1-D	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	bipy	0.023 g, 0.15 mmol	R-H3PIA	0.031 g, 0.1 mmol	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
GPT-3.5-turbo (zero-shot)	[Cd1.5((R)-PIA)(bipy)1.5(H2O)]n	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(R)-H3PIA	0.031 g, 0.1 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
	[Cd1.5((S)-PIA)(bipy)1.5(H2O)]n	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(S)-H3PIA	NaN	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	NaN
GPT-3.5-turbo (fine-tuned)	[Cd1.5((R)-PIA)(bipy)1.5(H2O)]n (1-D)	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(R)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
	[Cd1.5((S)-PIA)(bipy)1.5(H2O)]n (1-L)	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(S)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
Ground truth	[Cd1.5((R)-PIA)(bipy)1.5(H2O)]n (1-D)	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(R)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours
	[Cd1.5((S)-PIA)(bipy)1.5(H2O)]n (1-L)	Cd(NO3)2·4H2O	0.061 g, 0.20 mmol	(S)-H3PIA | bipy	0.031 g, 0.1 mmol | 0.023 g, 0.15 mmol	NaN	NaN	distilled water | methanol | DMF	1 mL | 2 mL | 2 mL	100 °C	48 hours

Fig. 4. | Design and Performance for Paragraph2MOFInfo task. a, A statistic of the dataset. **b**, Mean performance of Levenshtein similarity and exact match accuracy by different models. **c**, Levenshtein similarity for 11 parameters in the Paragraph2MOFInfo task. **d**, Exact match accuracy for 11 parameters in the Paragraph2MOFInfo task. **e**, An example of extractions by different models from a multi-reaction MOF synthesis paragraph. The cells in yellow represented the ground truth. The cells in green represented the exact match predictions. The cells in blue represented the incorrect predictions.

Paragraph2NMR—Extraction of Experimental Conditions and NMR Chemical Shifts.

The impact of training set sizes and the use of prompt engineering on the performance of fine-tuning gpt-3.5-turbo in extracting NMR information is illustrated in Fig. 5a. Regardless of the training data size for fine-tuning (ranging from 25 to 300), or the presence of prompt engineering, there are hardly any significant fluctuations in performance. This holds true for metrics such as Levenshtein similarity and exact match accuracy of the fine-tuned gpt-3.5-turbo when the numbers of training samples exceed 50. This demonstrates the strong learning capability and robustness of LLMs. Fig. 5b illustrates the performance of different generative language models using the same 200 training data. In terms of Levenshtein similarity, a metric based on edit distance, almost all fine-tuned language models achieved impressing scores, outperforming GPT models that solely rely on prompt engineering (Fig. 5b, Supplementary Fig. 6). However, when considering the exact match accuracy metric, where each character must perfectly align with the ground truth count, LLMs such as GPTs and Llama2 take the lead. While fine-tuned T5 and BART manage to extract the majority of the text, they often miss or mistakenly copy several characters. This contributes to a significant decrease in their exact match accuracy metric, as shown in Fig. 5c. In this context, the extraction of long complex text by LLMs is more standardized and high-quality, aligning more closely with human expectations. It is worth noting that fine-tuning Llama2 provides an alternative approach for deploying text mining locally, given its exceptional exact match accuracy.



Levenshtein Similarity

	IUPAC name	1H NMR text	1H NMR cond	1H NMR data	13C NMR text	13C NMR cond	13C NMR data
GPT-3.5-turbo (fine-tune w/ prompt)	99.1	97.2	98.4	99.5	96.8	98.1	99.7
GPT-3.5-turbo (fine-tune w/o prompt)	98.6	97.9	98.6	99.7	97.6	98.1	99.8
Llama2-13b-chat (fine-tune w/ prompt)	96.8	96.3	91.6	96.9	96.8	90.1	96.9
Llama2-13b-chat (fine-tune w/o prompt)	97.0	96.8	92.1	97.9	97.2	90.3	98.3
T5-base (fine-tune)	97.2	92.3	97.6	93.6	86.2	94.0	87.8
BART-base (fine-tune)	95.3	77.2	97.2	73.7	80.4	96.7	81.0
GPT-4 (one-shot)	98.6	96.6	89.6	95.6	96.2	88.0	96.7
GPT-4 (zero-shot)	94.5	97.4	91.8	97.1	97.0	90.3	97.6
GPT-3.5-turbo (one-shot)	97.6	95.7	94.9	97.4	95.9	96.0	97.9
GPT-3.5-turbo (zero-shot)	91.6	95.2	90.5	95.4	94.8	90.1	97.4

Exact Match Accuracy

	IUPAC name	1H NMR text	1H NMR cond	1H NMR data	13C NMR text	13C NMR cond	13C NMR data
GPT-3.5-turbo (fine-tune w/ prompt)	96.0	86.7	93.0	94.0	87.0	92.7	95.7
GPT-3.5-turbo (fine-tune w/o prompt)	96.3	85.3	92.3	94.7	85.7	91.0	95.0
Llama2-13b-chat (fine-tune w/ prompt)	91.0	82.3	85.3	90.0	83.0	82.7	89.7
Llama2-13b-chat (fine-tune w/o prompt)	89.3	82.3	86.3	91.3	86.0	84.3	93.0
T5-base (fine-tune)	88.7	5.0	86.0	76.7	4.7	80.3	73.0
BART-base (fine-tune)	64.3	2.0	90.3	9.0	8.0	87.0	13.3
GPT-4 (one-shot)	94.7	84.7	79.3	72.3	82.0	78.3	90.0
GPT-4 (zero-shot)	48.0	78.0	84.3	83.7	75.0	83.3	89.0
GPT-3.5-turbo (one-shot)	92.0	67.0	85.7	72.3	70.0	88.0	89.7
GPT-3.5-turbo (zero-shot)	37.4	61.3	76.3	71.7	63.7	75.7	88.7

1H NMR text:

Ground Truth :

"1H-NMR (400 MHz, CDCl3, ppm) δ 8.74 (s, 1H), 8.38 (s, 1H), 8.12 (s, 1H), 8.08 (s, 1H), 8.05 (d, J = 4.4 Hz, 1H), 7.50 (s, 1H), 6.82 (d, J = 4.0 Hz, 1H), 5.96 (s, 1H), 4.67(t, J = 4.5 Hz, 1H), 4.56 (t, J = 4.5 Hz, 1H), 4.34 (t, J = 4.4Hz, 1H), 4.27 (t, J = 4.5 Hz, 1H), 3.90 (s, 6H) "

Error Extraction:

"1H-NMR (400 MHz, CDCl3, ppm) δ 8.74 (s, 1H), 8.38 (s., 1H, 8.12 (s', 1H], 8.07 (d, J = 4.4 Hz, 1 H), 7.50 (s,- 1H)), 6.82 (d., J = 3.0 Hz, 2H), 5.96 (s), 4.67, 4.56 (t, J - 4.5 Hz, HH), 4,34 (t., J - 3.4Hz, 1 h), 4.27 (t,- J = 6.5Hz, 2 H), 3.90 (s = 6H) "

13C NMR text:

Ground Truth :

"13C-NMR (125 MHz, CDCl3) δ 27.9 (C(CH3)3), 61.6 (C(CH3)3), 80.9 (PhCCH), 83.1 (PhCCH), 123.6 (Ar-CH), 124.6 (Ar-CH), 126.8 (Ar-C), 129.1 (Ar-C), 137.5 (Ar-CH), 138.2 (Ar-C), 159.4 (C=O) "

Error Extraction:

"13C-NMR (125 MHz, CDCl3) δ 27.9 (C(CH3)3), 61.6 (C[CH3(3), 80.9, 83.1 (PhCCH), 123.6, 124.6 [Ar-CH), 126.8 (Ar-C), 129.1, 137.5, 138.2 (Ar-)C), 159.4 (C=O) "

Fig. 5. | Performance for Pargraph2NMR task. a, The performance of fine-tuning gpt-3.5-turbo with and without prompt engineering as it varies with training data size. **b**, Heat map illustrating Levenshitein similarity and exact match accuracy of various models in extracting each NMR information. **c**, Examples of error extractions by T5 and BART, compared with the ground truth.

Paragraph2Action—Action Sequence Extracted from an Experimental Procedure.

The above-mentioned extraction tasks simply require the model to replicate specific information from the paragraph. However, the Paragraph2Action task requires the model to understand and transform the paragraph. Clearly, GPT models with prompt engineering has difficulty with this task, especially when it involves multiple complex conversions and insufficient prompt descriptions (Table1, Supplementary Fig. 7). To gauge the maximum potential of ChatGPT using only prompts, we incrementally increased the number of transformation examples from 6 to 60. Despite encompassing all types of actions at least once and nearly reaching the token limit of 4,096 for GPT-3.5 and 8192 for GPT-4, their performance in the few-shot scenario remains disappointingly poor. The currently best-performing LLM GPT-4 with 60 examples for in-context learning, it achieved only 32.7% full sentence exact accuracy, a BLEU score of 65.0, and a Levenshtein similarity of 72.8. However, fine-tuning pre-trained language models with a small amount of data could yield decent results (Table 1). Remarkably, after 3 epochs of fine-tuning gpt-3.5-turbo on 1,060 hand-annotated training data, we achieved 62.5% full sentence exact accuracy, an 84.8 Modified BLEU score, and an 87.6 Levenshtein similarity. This process took only 1 hour and cost $3 for fine-tuning. These metrics surpass the SOTA results previously reported by Vaucher et al.[1], which used an ensemble of three models, each task-adaptively pre-trained on 2 million rule-based data and refined on 14,168 augmented data. Interestingly, further improvement was achieved by augmenting the training data size to 14,168. This resulted in 69.0% full sentence exact accuracy, an 86.4 Modified BLEU score, and an 89.9 Levenshtein similarity (Table 1). For autonomous robots, it is challenging to generate instructions that follow strict syntax rules. Fine-tuning LLMs plays a crucial role in bridging the gap between fuzzy natural language and structured machine-executable programming languages, significantly improving the accuracy of customization with a small amount of annotated data. In similar tasks involving “fuzzy rules” or hard-to-define extraction, fine-tuning LLMs might offer considerable advantages in tailoring the transformation.

Tabel 1 | Performance on Paragraph2Action task.

Model	Training data strategy	100% accuracy	90% accuracy	75% accuracy	Modified BLEU score	Levenshtein similarity	Cost
GPT-3.5-turbo (6-shots)	No training	8.2	16.8	34.7	38.6	59.4	905 mean tokens
GPT-3.5-turbo (12-shots)	No training	8.8	19.3	42.3	43.1	62.3	1,374 mean tokens
GPT-3.5-turbo (18-shots)	No training	13.1	23.3	42.6	44.4	64.3	1,670 mean tokens
GPT-3.5-turbo (24-shots)	No training	14.8	25.9	45.5	47.0	65.8	2,598 mean tokens
GPT-3.5-turbo (30-shots)	No training	13.9	26.4	47.2	49.5	66.0	3,610 mean tokens
GPT-4 (6-shots)	No training	13.4	23.3	44.9	44.7	54.5	861 mean tokens
GPT-4 (12-shots)	No training	20.7	30.7	51.1	51.4	69.2	1,357 mean tokens
GPT-4 (18-shots)	No training	21.9	33.0	56.5	53.8	63.0	1,631 mean tokens
GPT-4 (24-shots)	No training	22.7	35.8	58.2	56.7	65.1	2,546 mean tokens
GPT-4 (30-shots)	No training	26.1	40.0	61.6	59.8	67.7	3,611 mean tokens
GPT-4 (60-shots)	No training	32.7	43.8	63.3	65.0	72.8	7,010 mean tokens, $ 41
Transformer (single model) *	No task-adaptive pretraining, hand-annotated data (1,060)	13.1	15.1	21.9	22.5	45.9	-
BART-base (fine-tuned)	No task-adaptive pretraining, hand-annotated data (1,060)	51.1	65.9	77.6	73.2	83.9	-
T5-base (fine-tuned)	No task-adaptive pretraining, hand-annotated data (1,060)	57.7	71.6	83.2	81.8	86.8	-
Lama2-13b-chat (fine-tuned)	No task-adaptive pretraining, hand-annotated data (1,060)	56.8	66.8	80.7	80.3	86.0	40 min for training
GPT-3.5-turbo (fine-tuned)	No task-adaptive pretraining, hand-annotated data (1,060)	**62.5**	**72.7**	**82.9**	**84.8**	**87.6**	3 epochs, 1h, $ 3
Transformer (single model) *	No task-adaptive pretraining, augmented data (14,168)	37.8	47.7	62.8	64.7	76.4	-
BART-base (fine-tuned)	No task-adaptive pretraining, augmented data (14,168)	52.0	68.5	80.1	74.4	84.8	
T5-base (fine-tuned)	No task-adaptive pretraining, augmented data (14,168)	59.7	74.1	82.4	84.1	87.1	-
Llama2-13b-chat (fine-tuned)	No task-adaptive pretraining, augmented data (14,168)	60.2	70.4	83.5	81.6	87.9	9 hours for training
GPT-3.5-turbo (fine-tuned)	No task-adaptive pretraining, augmented data (14,168)	**69.0**	**78.1**	**86.9**	**86.4**	**89.9**	5 epochs, 1.5 h, $ 92
Transformer (single model) *	Task-adaptive pretraining (2 million), hand-annotate (1,060)	56.8	67.3	80.4	81.5	85.7	-
Transformer (single model) *	Task-adaptive pretraining (2 million), augmented data (14,168)	59.4	70.5	81.8	84.3	86.7	-
Transformer (ensemble models) *	Task-adaptive pretraining (2 million), augmented data (14,168)	**60.8**	**71.3**	**82.4**	**85.0**	**86.6**	-

The symbol "*" represented the result reported by Vaucher et al.[1] The result in black bold is the best previous performance. The result in red bold is the best new performance.

Promising Performance and Potentials of Fine-tuning LLMs on Chemical Text Mining.

Chemical text mining expedites scientific discovery in chemistry. Previously, tasks involving complex chemical language and sophisticated processing required the development of specific domain-focused models. Now, the fine-tuning of universal LLMs offers a highly generalized and cost-effective solution. We have demonstrated the impressive efficacy, flexibility, and high exact accuracy of fine-tuning LLMs, regarding all kinds of text mining tasks as generative problems. An examination of incorrect predictions revealed that only a small proportion were entirely incorrect, while most were acceptable alternatives to the ground truth or even pointed out the incorrect labels

(Supplementary Fig. 10-14). These errors can be attributed to inconsistent annotation standards and the inherent ambiguity of terms with multiple interpretations or functions. Therefore, improving the formatted data extraction requires continuous efforts, including the refinement of specific rules and the enrichment of annotations prone to misinterpretation during training and inference. With detailed specifications and high-quality formatted data, the fine-tuning method based on LLMs is highly reliable. It can be easily extended to tasks related to extracting information from scientific literature and transforming data into simple user-friendly reaction format[20] that is both human- and machine-readable. This approach will significantly contribute to the development of extensive databases like Open Reaction Database[21,22], SciFinder[23] and Reaxys[24], which gather comprehensive synthesis data through automated curation and expert verification, to make data more findable, accessible, interoperable, and reusable (FAIR).

Nevertheless, leveraging fine-tuned LLMs is still insufficient to extract all synthesis information from chemical literature, which contains extensive complex figure and form contents. Recently, some tools have been developed to recognize molecular images[25,26] and reaction diagrams[27,28] from the literature. Integrating LLMs with these image recognition tools or developing advanced large multimodal models (LMMs) may be a promising unified solution for further chemical data mining. Notably, when extracting large amounts of data from copyrighted literature, it's essential to access the necessary permissions from scientific publications.

In this work, we have scratched the surface of the vast potential of LLMs in chemistry and materials science by fine-tuning LLMs for chemical text mining. We can see that there is still a gap between open-source language models and GPT models, but considering GPTs' closed-source nature, it becomes imperative for researchers and communities to focus efforts on this direction. Technically, advancements like more effective fine-tuning strategies, improved open-source model architectures, faster inference approaches, wider context windows, and lower computational costs in the era of LLMs are anticipated to further enhance text mining. Meanwhile, it’s more essential to consider what else can be achieved with LLMs and how we can develop more effective LLMs for chemistry and materials science. For instance, LLMs have the potential to revolutionize predictive modelling by incorporating the extensive “fuzzy knowledge” encapsulated within scientific literature, especially in chemistry and drug discovery. By combining empirical results with documented knowledge, LLMs could assist chemists identify patterns in experiments that might otherwise be missed, predict properties of compounds and outcomes of reactions, and even generate new chemical hypotheses and theories. Furthermore, the integration of LLMs’ comprehension with specialized tools could substantially lower the barrier of chemists to use these tools throughout the entire workflow, thanks to interactive interfaces in natural language. Future research could investigate how to merge formatted laboratory data with wealth of information in scientific literature and develop the multimodal capability to enrich specific domain knowledge for LLMs. This endeavor will require a sustained, long-term effort.

Conclusion

Here, we have demonstrated the effectiveness of fine-tuning LLMs in chemical text mining. We conducted five complex tasks: compound entity recognition, reaction role labelling, MOF synthesis information extraction, NMR data extraction, and the transformation from reaction procedures to action sequences. Chemical text mining remains a challenging professional domain when leveraging language model mining, even with prompt engineering. However, LLMs that are fine-tuned with appropriate annotations can produce structured outputs that perfectly fulfil human requirements not easily expressed in natural language. This feature fully utilizes their natural language understanding and formatting capability. Using chemical text mining as an example, this study provides guidance on fine-tuning of LLMs to serve as universal knowledge extraction toolkits. These toolkits can be easily extended for automated extraction from documents and rule-based formatted transformations. Our work lays the groundwork for the applications of LLMs in information extraction within the chemical domain, which will catalyze data-driven innovations in chemical and materials science.

Methods

Data Preparation

For the Paragraph2Compound task, we compiled an automatically annotated dataset. This dataset is based on the publicly accessed USPTO subset extracted by Lowe et al. [29,30], and includes millions chemical reaction paragraphs from patents, each paired with compound tags. We used regular expressions to identify compound labels within each paragraph, separating them with "|" symbol based on their sequential occurrence in the paragraph. For the Paragraph2RXNRole task, we used the manually annotated dataset by Guo et al.[8], following the same data partitioning strategy. We transformed the data from the BIO-token classification format to a sequence-to-sequence format using the annotation scheme "<Role*compound*Role>". We processed paragraphs containing multiple central products and related reactions into several input and output pairs. For the Paragraph2MOFInfo task, we manually checked and re-annotated the raw data of Zheng et al.[14], transforming them into a sequence-to-sequence format. This dataset comprises MOF synthesis paragraphs, extraction by ChatGPT, and human-evaluated answers. For the Paragraph2NMR task, we manually curated a dataset of 600 high-quality annotations. These were mainly sourced from various literature on PubMed to ensure a wide diversity. The task is aims to extract information such as IUPAC name, experimental conditions, including frequency and solvent, and chemical shifts data from both 1H NMR and 13C NMR spectra. For the Paragraph2Action task, we utilized the hand-annotated dataset by Vaucher et al., employing the same data partitioning strategy. This dataset is derived from the Pistachio dataset by NextMove software[31]. The details of datasets used for the five chemical text mining tasks are listed in Supplementary Table 1.

Prompt-only ChatGPT

Prompt-only interaction enables users to efficiently communicate with large language models through simple prompts. This guides the model to produce relevant responses without further training. In a zero-shot scenario, the model generates responses using only a descriptive prompt and its pre-trained knowledge. However, in a few-shot approach, the model uses a small number of examples to improve its understandings and responses. To maximize the performance, we selected diverse examples and ensured a large number of tokens. We interacted with ChatGPT using API keys and employed model versions gpt-3.5-turbo-0613 and gpt-4-0613. The zero-shot and few-shot prompts for chemical text mining tasks can be found in Supplementary Fig. 2-7.

Fine-tuning ChatGPT

Since late August 2023, supervised fine-tuning capabilities have been available for the gpt-3.5-turbo model[32]. The aim is to enhance performance in specific scenarios customized based on private data. In this study, we fine-tuned the gpt-3.5-turbo-0613 model for chemical text mining sceneries. We formatted the data into jsonl and uploaded them to OpenAI's cloud servers, then initiated fine-tuning jobs. Once the training was complete, the fine-tuned gpt-3.5-turbo model was ready for inference. API keys were requisite throughout the training and inference procedures. Fine-tuning for the gpt-4-turbo model is expected in the future.

Open-Source Language Models

We selected the most widely used and representative generative pre-trained language models like Llama2,[17] T5[18] and BART[19]. These serve as baselines for a comprehensive comparison with the fine-tuned ChatGPT across five chemical text mining tasks. Considering performance, efficiency, and hardware resource constraints, we used full parameter fine-tuning for BART-base and T5-base. We applied multitask-learning to BART and T5 in the Paragraph2MOFInfo task and

Paragraph2NMR task due to their limitations in generating multi-attribute long sentences (Supplementary Fig. 8, 9), aiming to enhance their performance. This approach significantly improved their performance. For Llama2, we used Q-LoRA[33] to efficiently fine-tune llama2-13b-chat. This method maintains most performance of full parameter fine-tuning while significantly reducing computational demands. We used vllm[34] to speed up the inference of llama2-13b-chat, which is tens of times faster than Hugging Face's pipeline. To ensure optimal performance, we adjusted hyperparameters such as learning rates, lora_r, and lora_alpha during the fine-tuning process of baseline models (Supplementary Table 2). More details of training, pre-processing, and post-processing can be found in the Supplementary Information.

Metrics for Evaluation

Since fine-tuning ChatGPT does not allow for early stopping based on optimal validation loss, we report the performances of all models at the best epoch selected from the evaluation set for fair comparison. Given the task specifics, we use metrics including precision, recall, and F1 score for evaluating entity-level performance. For sentence-level performance assessment, we use Levenshtein similarity, exact match accuracy, partial accuracy, and a modified BLEU score.

Data Availability

All datasets used in this work are available from the authors upon request.

Code Availability

All scrips for training and evaluating can be found on GitHub at https://github.com/zw-SIMM/SFTChatGPT_for_chemtext_mining.

References

1 Vaucher, A. C. *et al.* Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Comm.* **11**, 3601 (2020).

2 Suvarna, M., Vaucher, A. C., Mitchell, S., Laino, T. & Pérez-Ramírez, J. Language models and protocol standardization guidelines for accelerating synthesis planning in heterogeneous catalysis. *Nat. Comm.* **14**, 7964 (2023).

3 Mehr, S. H. M., Craven, M., Leonov, A. I., Keenan, G. & Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **370**, 101-108 (2020).

4 Steiner, S. *et al.* Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **363**, eaav2211 (2019).

5 Ha, T. *et al.* AI-driven robotic chemist for autonomous synthesis of organic molecules. *Sci. Adv.* **9**, eadj0461 (2023).

6 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).

7 Mavracic, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. *J. Chem. Inf. Model.* **61**, 4280-4289 (2021).

8 Guo, J. *et al.* Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model.* **62**, 2035-2045 (2021).

9 Castro Nascimento, C. M. & Pimentel, A. S. Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *J. Chem. Inf. Model.* **63**, 1649-1655 (2023).

10 Clark, T. M., Anderson, E., Dickson-Karn, N. M., Soltanirad, C. & Tafini, N. Comparing the Performance of College Chemistry Students with ChatGPT for Calculations Involving Acids and Bases. *J. Chem. Educ.* **100**, 3934-3944 (2023).

11 Guo, T. *et al.* What indeed can GPT models do in chemistry? A comprehensive benchmark on eight tasks. Preprint at *arXiv* https://arxiv.org/abs/2305.18365 (2023).

12 Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Computing Surveys* **55**, 1-38 (2023).

13 Zhang, Y. *et al.* Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. Preprint at *arXiv* https://arxiv.org/abs/2309.01219 (2023).

14 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).

15 Patiny, L. & Godin, G. Automatic extraction of FAIR data from publications using LLM. (2023).

16 Chen, Q. *et al.* An Extensive Benchmark Study on Biomedical Text Generation and Mining with ChatGPT. *Bioinformatics*, btad557 (2023).

17 Touvron, H. *et al.* Llama 2: Open foundation and fine-tuned chat models. Preprint at *arXiv* https://arxiv.org/abs/2307.09288 (2023).

18 Raffel, C. *et al.* Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* **21**, 5485-5551 (2020).

19 Lewis, M. *et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Preprint at *arXiv* https://arxiv.org/abs/1910.13461 (2019).

20 Nippa, D. F. *et al.* Simple User-Friendly Reaction Format. (2023).

21 Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143**, 18820-18826 (2021).

22 Mercado, R., Kearnes, S. M. & Coley, C. W. Data sharing in chemistry: lessons learned and a case for mandating structured reaction data. *J. Chem. Inf. Model.* **63**, 4253-4265 (2023).

23 SciFinder. https://scifinder-n.cas.org. (accessed August 29, 2023).

24 Reaxys. https://www.reaxys.com. (accessed August 29, 2023).

25 Xiong, J. *et al.* αExtractor: a system for automatic extraction of chemical information from biomedical literature. *Sci. China. Life. Sci.* (2023).

26 Qian, Y. *et al.* MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation. *J. Chem. Inf. Model.* **63**, 1925-1934 (2023).

27 Qian, Y., Guo, J., Tu, Z., Coley, C. W. & Barzilay, R. RxnScribe: A Sequence Generation Model for Reaction Diagram Parsing. *J. Chem. Inf. Model.* **63**, 4030-4041 (2023). https://doi.org/10.1021/acs.jcim.3c00439

28 Wilary, D. M. & Cole, J. M. ReactionDataExtractor 2.0: A deep learning approach for data extraction from chemical reaction schemes. *J. Chem. Inf. Model.* **63**, 6053-6067 (2023).

29 Lowe, D. Chemical reactions from US patents (1976-Sep2016). figshare https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873. (accessed August 29, 2023).

30 Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. Thesis, University of Cambridge, (2012).

31 NextMoveSoftware. Pistachio. https://www.nextmovesoftware.com/pistachio.html. (accessed August 22, 2023).

32 Peng, A., Wu, M., Allard, J., Kilpatrick, L. & Heidel, S. GPT-3.5 Turbo fine-tuning and API updates. https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates. (accessed August 22, 2023).

33 Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. Preprint at *arXiv* https://arxiv.org/abs/2305.14314 (2023).

34 Kwon, W. *et al.* in *Proceedings of the 29th Symposium on Operating Systems Principles* 611–626 (Association for Computing Machinery, Koblenz, Germany, 2023).

Acknowledgements

We thank all contributions of the open-source community on LLMs. We appreciate Yaghi's group for guiding in ChatGPT prompt engineering for chemistry tasks.

This work was supported by National Natural Science Foundation of China (T2225002, 82273855 to M.Y.Z., and 82204278 to X.T.L.), the National Key Research and Development Program of China (2022YFC3400504 to M.Y.Z.), SIMM-SHUTCM Traditional Chinese

Medicine Innovation Joint Research Program (E2G805H to M.Y.Z.), and Shanghai Municipal Science and Technology Major Project.

Contributions

W.Z., J.C.X., Z.Y.F., and M.Y.Z. conceived the idea. M.Y.Z and Z.Y.F designed the research. W.Z., Q.G.W., Z.M.H. implemented the codes. W.Z., Q.G.W., X.T.K, J.C.X, S.K.N., Z.Y.F. collected, annotated, and processed training data. D.H.C., B.Y.N., Q.S., and X.T.L. checked the data. M.A.C., R.Z.Z., Y.T.W., L.H.Z benchmarked the models. W.Z. wrote the initial draft. M.Y.Z., Z.Y.F. and Z.P.X. reviewed and refined the article. All authors contributed to the analysis of the results. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.