An Autonomous Large Language Model Agent for Chemical Literature Data Mining
用于化学文献数据挖掘的自主大型语言模型代理

Kexin Chen[a], Hanqun Cao[a], Junyou Li[b], Yuyang Du[a], Menghao Guo[b], Xin Zeng[b], Lanqing Li[b], Jiezhong Qiu[b], Pheng Ann Heng[a], Guangyong Chen[b]
[a] The Chinese University of Hong Kong, New Territories, Hong Kong SAR
[b] Zhejiang Lab, Zhejiang University, Hangzhou, China

Abstract—Chemical synthesis, which is crucial for advancing material synthesis and drug discovery, impacts various sectors including environmental science and healthcare. The rise of technology in chemistry has generated extensive chemical data, challenging researchers to discern patterns and refine synthesis processes. Artificial intelligence (AI) helps by analyzing data to optimize synthesis and increase yields. However, AI faces challenges in processing literature data due to the unstructured format and diverse writing style of chemical literature. To overcome these difficulties, we introduce an end-to-end AI agent framework capable of high-fidelity extraction from extensive chemical literature. This AI agent employs large language models (LLMs) for prompt generation and iterative optimization. It functions as a chemistry assistant, automating data collection and analysis, thereby saving manpower and enhancing performance. Our framework's efficacy is evaluated using accuracy, recall, and F1 score of reaction condition data, and we compared our method with human experts in terms of content correctness and time efficiency. The proposed approach marks a significant advancement in automating chemical literature extraction and demonstrates the potential for AI to revolutionize data management and utilization in chemistry.
化学合成对于推进材料合成和药物发现至关重要，它影响着包括环境科学和医疗保健在内的各个领域。化学技术的兴起产生了大量的化学数据，对研究人员辨别模式和改进合成过程提出了挑战。人工智能 (AI) 通过分析数据来优化合成并提高产量。然而，由于化学文献的格式非结构化和写作风格多样，人工智能在处理文献数据方面面临挑战。为了克服这些困难，我们引入了一个端到端的 AI 代理框架，该框架能够从大量化学文献中进行高保真提取。该 AI 代理采用大型语言模型 (LLM) 进行快速生成和迭代优化。它充当化学助手，自动收集和分析数据，从而节省人力并提高性能。我们框架的有效性是使用反应条件数据的准确度、召回率和 F1 分数来评估的，并且我们在内容正确性和时间效率方面将我们的方法与人类专家进行了比较。所提出的方法标志着化学文献提取自动化的重大进步，并展示了 AI 彻底改变化学数据管理和利用的潜力。

Index Terms—Chemical synthesis, literature mining, intelligent agent, large language models
化学合成、文献挖掘、智能代理、大型语言模型

I. Introduction

The discipline of chemistry, characterized by its immense potential and practical utility, is deeply intertwined with the synthesis of materials and the discovery of drugs. These sectors, propelled by the investigation of novel materials, contribute to advancements in energy, environmental science, and nanotechnology. They also lay the groundwork for the discovery of new pharmaceuticals and analyses in the life sciences. Such progressions are instrumental in enhancing therapeutic methods for diseases and promoting the growth of the health sector. The technological evolution has led to the accumulation of a wealth of data regarding chemical reactions, most of which is freely accessible. However, the challenge lies in efficiently harnessing this data to unearth intricate patterns, and to facilitate the discovery of novel reaction mechanisms. Addressing this issue could expedite the synthesis of materials, and the development of drugs, thereby fostering innovation in the field of chemistry.
化学学科因其巨大的潜力和实用性而备受关注，它与材料合成和药物发现紧密相连。这些领域在新材料的研究推动下，促进了能源、环境科学和纳米技术的发展。同时，它们也为新药的发现以及生命科学中的分析奠定了基础。这些进展对于改善疾病的治疗方法和促进健康产业的发展至关重要。技术的发展使得化学反应相关的数据大量积累，其中大部分是自由可获取的。然而，如何高效地利用这些数据以挖掘复杂的模式并促进新反应机制的发现仍然是一项挑战。解决这一问题可以加速材料的合成和药物的开发，从而推动化学领域的创新。

Artificial Intelligence (AI) has the potential to identify salient features and patterns of reactions by learning from extant data and predicting outcomes of new reactions [1], [2]. This capability can aid chemists in rapidly screening and evaluating diverse reaction conditions, optimizing synthetic routes, and enhancing synthetic efficiency. Moreover, when AI is combined with algorithms for predicting and optimizing reactions, it can generate a variety of synthetic paths and optimize them based on specific objectives and constraints. This process assists chemists in rapidly identifying efficient, sustainable synthetic routes, thereby

improving the yield and purity of synthetic products.
人工智能（AI）具有通过学习现有数据并预测新反应结果来识别反应的关键特征和模式的潜力 [1], [2]。这一能力可以帮助化学家快速筛选和评估不同的反应条件，优化合成路线，并提高合成效率。此外，当 AI 与反应预测和优化算法结合时，它可以生成多种合成路径，并根据特定目标和约束条件对其进行优化。这一过程有助于化学家迅速识别高效且可持续的合成路线，从而提高合成产物的产率和纯度。

Despite the successes of AI in these areas, gaining a deeper understanding of reaction rules remains essential for analyzing chemical reactions and discovering valuable chemical reactions. Unearthing associations and patterns concealed in data, revealing common characteristics and mechanisms between different reactions, aids chemists in 1) better understanding the underlying principles of reactions and 2) guiding the design of novel reactions.
尽管 AI 在这些领域取得了一定的成功，但深入理解反应规律仍然是分析化学反应和发现有价值反应的关键。从数据中挖掘隐藏的关联和模式，并揭示不同反应之间的共同特征和机制，有助于化学家 1) 更好地理解反应的基本原理，以及 2) 指导新反应的设计。

To accomplish this, the integration and knowledge management of data regarding chemical reactions are of utmost importance, as they form the basis for discovering new reaction rules. Through automated data collection, organization, and annotation, AI can establish a comprehensive database of chemical reactions, enabling chemists to conveniently access and utilize these data. This aids in enhancing the discoverability and reproducibility of data, allowing researchers to better utilize extant knowledge of chemical reactions.
为此，对化学反应数据的整合和知识管理至关重要，因为它们构成了发现新反应规律的基础。通过自动化的数据收集、组织和注释，AI 可以建立一个全面的化学反应数据库，使化学家能够便捷地访问和利用这些数据。这有助于提高数据的可发现性和可重复性，使研究人员能够更好地利用现有的化学反应知识。

However, contemporary AI technologies encounter some challenges when dealing with data from chemical reaction literature. The data lacks uniform organization and processing, and extracting core reaction information from intricate and lengthy literature is a challenging task. This necessitates AI models to possess advanced context analysis capabilities and high standards for pattern recognition in text style and content.
然而，当代 AI 技术在处理化学反应文献数据时仍然面临一些挑战。这些数据缺乏统一的组织和处理方式，从复杂冗长的文献中提取核心反应信息是一项艰巨的任务。这要求 AI 模型具备高级的上下文分析能力，并在文本风格和内容模式识别方面达到较高的标准。

The introduction of large language models (LLMs) like Chat-GPT enables efficient communication between humans and machines by converting instructions into text and integrating solvers for multiple subtasks [3], [4]. This provides vast opportunities for literature mining and opens new avenues for AI exploration in the field of chemistry.
大语言模型（LLMs），如 Chat-GPT 的引入，使得人与机器之间能够通过将指令转换为文本并集成多任务求解器来实现高效通信 [3], [4]。这为文献挖掘提供了广阔的机会，并为 AI 在化学领域的探索开辟了新的途径。

In response to these challenges, we propose an end-to-end framework based on a powerful AI agent that automatically extracts high-fidelity chemical data from the vast amount of literature, which is shown in Figure 4.
针对这些挑战，我们提出了一个端到端的框架，该框架基于一个强大的 AI 代理，能够从大量文献中自动提取高保真化学数据，如图 4 所示。

- We propose a novel approach to create AI agents in chemistry literature information extraction. For the first time, we have linked the concept of AI agents with AI-based chemical research. The agent-based framework for extracting chemical literature has greatly saved manpower and achieved intelligent task automation.
  我们提出了一种新方法来创建化学文献信息提取的 AI 代理。首次将 AI 代理的概念与基于 AI 的化学研究联系起来。基于代理的化学文献提取框架极大地节省了人力，并实现了智能任务自动化。
- We design a novel assessment scheme for evaluating the agent's intelligence in terms of literature text mining through accuracy, recall, and F1 score. This evaluation scheme is one of the most important links to ensure the efficiency of the agent's task execution. Our task-oriented application of chemical professional knowledge can bring a more intuitive performance display of the agent.
  我们设计了一种新颖的评估方案，通过准确率、召回率和 F1 分数来评估代理在文献文本挖掘方面的智能水平。该评估方案是确保代理任务执行效率的关键环节之一。我们基于任务的化学专业知识应用能够更直观地展现代理的性能。

II. Results and discussion 结果与讨论

A. General Results 总体结果

Acting as an efficient helper for chemists, our agent is expected to obtain higher reaction information retrieval quality and less time cost. Thus, quantitatively measuring the performance of AI-aided approaches, as well as effectively comparing its abilities with human experts, are necessary for exploring this new field. To investigate the efficacy of our framework, we devise a novel and comprehensive pipeline for evaluating the proficiency of GPT-based literature mining methods.
作为化学家的高效助手，我们的代理预计能够获得更高质量的反应信息检索结果，并降低时间成本。因此，量化评估 AI 辅助方法的性能，并有效地将其能力与人类专家进行比较，对于探索这一新领域至关重要。为了研究我们框架的有效性，我们设计了一种新颖且全面的流程，以评估基于 GPT 的文献挖掘方法的能力。

In our evaluation process, we place great emphasis on assessing the quality of reactants/reagents, solvents,

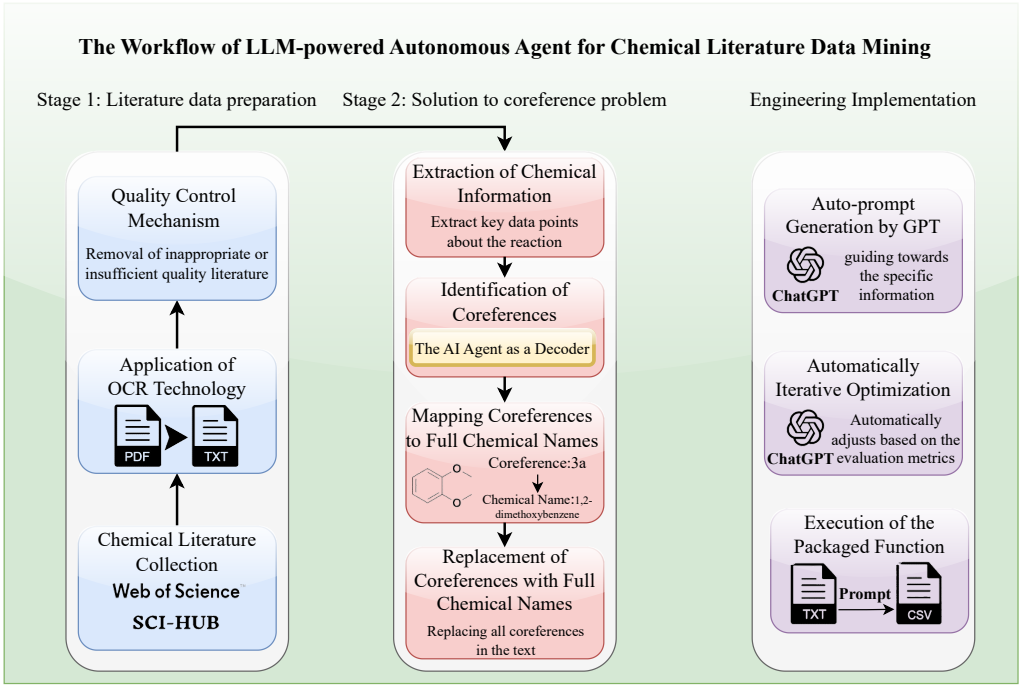

Figure 1. The framework of chemical literature analysis and reaction information extraction agent based on LLMs.

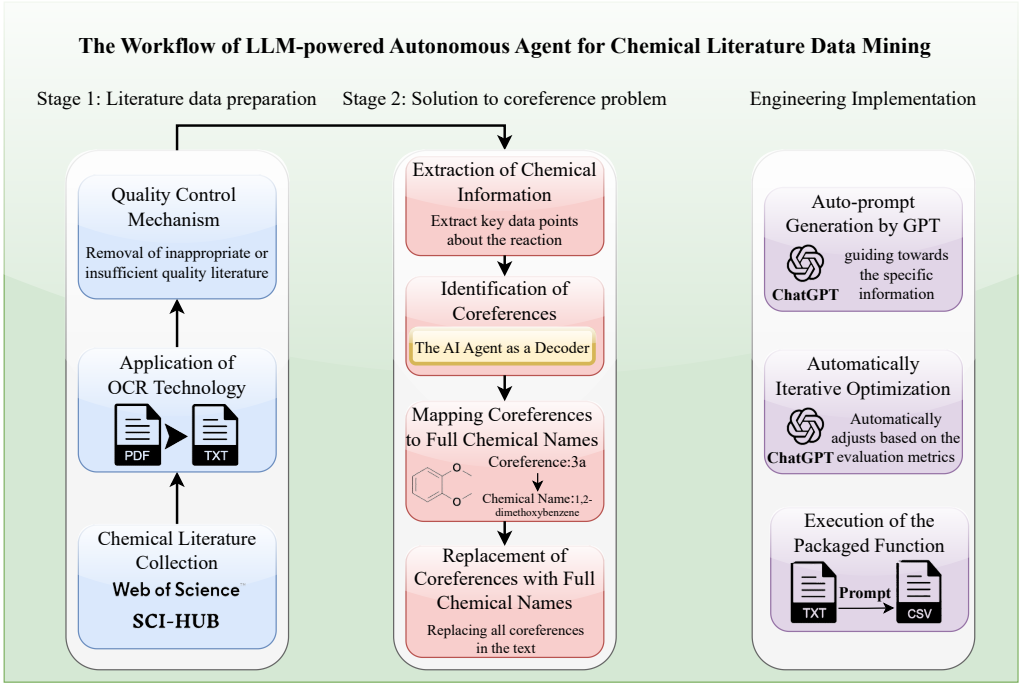

Figure 2. 基于 LLMs 的化学文献分析与反应信息提取代理框架。

products, and yields involved in each Suzuki reaction, which is the primary goal of effective retrieval.
在我们的评估过程中，我们特别关注 Suzuki 反应中涉及的反应物/试剂、溶剂、产物及收率的质量，这是有效检索的主要目标。

To ensure precise quantification of each component, we have employed an evaluation scheme that includes precision, recall, and F1-score. By utilizing these metrics, we can gauge the model's ability to accurately extract pertinent reaction information and conduct exhaustive searches of factors related to reactions.
为了确保对各个组成部分的精准量化，我们采用了包含精准率、召回率和 F1 分数的评估方案。通过使用这些指标，我们可以衡量模型准确提取相关反应信息的能力，并对反应相关因素进行全面搜索。

Table I
The precision, recall, F1-score results of data mining for yield, reactant/reagent, solvent, product.
数据挖掘中收率、反应物/试剂、溶剂和产物的精准率、召回率及 F1 分数结果。

	Precision	Recall	F1-score
Yield	92.19%	78.53%	84.81%
Reactant / Reagent	89.04%	76.00%	82.00%
Solvent	91.90%	75.77%	83.06%
Product	87.45%	78.22%	82.58%

Table II
The quantities of correct, extracted, and total pieces of reaction information.
正确、提取以及总反应信息的数据量。

Table III

	Correct data	Extracted data	Total data
Yield	236	256	326
Reactant / Reagent	203	228	300
Solvent	227	247	326
Product	223	255	326

After obtaining the generation results from ChatGPT, we restore the outcomes for later comparison with the ground truth collected by human experts. We annotate 17 literature and 326 reactions to validate the effectiveness of our agent. On average, the precision, recall, and F1-score reached 90.14%, 77.13%, and 83.11%, respectively. A detailed representation of the precision, recall, and F1-score results can be found in Table 1. We also provide the quantities of correct, extracted, and total pieces of reaction information in Table 2.
在获得 ChatGPT 生成的结果后，我们恢复这些结果，以便稍后与人类专家收集的真实数据进行比较。我们对 17 篇文献和 326 个反应进行了标注，以验证代理的有效性。平均而言，精准率、召回率和 F1 分数分别达到了 90.14%、77.13% 和 83.11%。表 1 详细展示了精准率、召回率和 F1 分数的结果。我们还在表 2 中提供了正确、提取以及总反应信息的数据量。

B. Comparison to Human Experts 与人类专家的比较

As far as we know, there are currently no other open-source tools available for extracting chemical reaction data from academic journals. Thus, this paper primarily validates the effectiveness and performance of the agent in extracting chemical reaction data information through comparison with manual data collection by human chemists. The primary indicators for evaluation are precision, average cost, and average speed. To minimize uncertainty and randomness for human chemists, the comparative study selected ten graduate students specializing in chemistry at the master's or doctoral level to perform manual data collection. The results from these ten chemistry professionals were then averaged to provide a comprehensive comparison with our agent. From Table 3, we can see that our AI agent reached competitive precision performance and much better performance in average cost and average speed.
据我们所知，目前尚无其他开源工具可用于从学术期刊中提取化学反应数据。因此，本文主要通过与人类化学家手动收集数据的对比，验证代理在化学反应数据提取方面的有效性和性能。主要的评估指标包括精准率、平均成本和平均速度。为了最大程度减少人为化学家操作的不确定性和随机性，我们在比较研究中选择了十名硕士或博士阶段的化学专业研究生进行手动数据收集。然后，将这十位化学专业人士的结果进行平均，以与我们的代理进行全面比较。从表 3 可以看出，我们的 AI 代理在精准率上达到了竞争水平，并且在平均成本和平均速度方面表现更优。

III. Method 方法

A. Data Preparation 数据准备

Literature Collection **文献收集**:

In the initial stage of our research, the acquisition of a high-quality literature dataset was paramount. To this end, we embarked on an extensive data collection process, leveraging the vast repository of chemical literature available on SciHub. Our specific focus was on articles about Suzuki reactions, a popular topic in organic chemistry. This search provides a substantial foundation for our subsequent analysis.
在研究的初始阶段，获取高质量的文献数据集至关重要。为此，我们进行了广泛的数据收集，利用 SciHub 上丰富的化学文献资源。我们特别关注有关 Suzuki 反应的文章，这是有机化学中的热门主题。这一搜索为我们的后续分析奠定了坚实的基础。

Application of OCR Technology OCR **技术的应用**:

To convert these articles into a machine-readable format, we employed Optical Recognition (OCR) technology. This enabled us to transform the PDFs into text, thus making them amenable to further computational processing. However, it is noteworthy that the OCR process, while generally accurate, is not infallible. It occasionally struggles with complex layouts and low-quality scans, which can lead to errors in the resulting text.
为了将这些文章转换为机器可读格式，我们采用了光学字

Table IV
The precision, average cost, and average speed of manual data collection and our agent.
手动数据收集和我们的代理的精度、平均成本和平均速度。

	Precision	Average cost (USD)	Average speed (second)
Manual data collection	90%	1.41	288
AI agent	87%	0.0025	0.43

符识别（OCR）技术。这使我们能够将 PDF 转换为文本，从而便于进一步的计算处理。然而，需要注意的是，尽管 OCR 过程通常较为准确，但并非完全无误。它在处理复杂布局和低质量扫描件时可能会遇到困难，从而导致生成文本中出现错误。

Quality Control Mechanism **质量控制机制**:

Given this potential for error, we instituted a quality control mechanism to ensure the reliability of our dataset. Specifically, we analyzed each article for the presence of key phrases such as "General Procedure", "Typical Procedure", or "General Experiment". These phrases are often indicative of the detailed methodological sections that are crucial for our purposes. Articles that did not contain these phrases were deemed to be of insufficient quality and were thus excluded from our dataset. Similarly, we also excluded articles where these phrases appeared more than five times, as this was suggestive of an overly complex or convoluted methodology that may not lend itself well to our extraction process.

鉴于可能存在的错误，我们建立了质量控制机制，以确保数据集的可靠性。具体而言，我们分析了每篇文章是否包含 “General Procedure”、“Typical Procedure” 或 “General Experiment” 等关键词。这些短语通常指示了详细的方法部分，对于我们的研究至关重要。未包含这些短语的文章被视为质量不足，因此被排除在数据集之外。同样，我们也排除了包含这些短语超过五次的文章，因为这通常意味着过于复杂或冗长的方法，可能不适合我们的提取过程。

Final Dataset:

After this rigorous quality control process, we were left with a refined dataset of 1000 articles. These articles formed the basis of our subsequent extraction and performance measurement activities. Our dataset, while significantly reduced in size, was of high quality and well-suited to our research aims.

经过严格的质量控制过程后，我们最终得到了一个包含 1000 篇文章的精炼数据集。这些文章构成了我们后续提取和性能评估活动的基础。尽管数据集的规模大幅缩小，但其质量较高，能够很好地满足我们的研究目标。

B. General Procedure

In our quest to extract chemical reaction conditions from the literature, we employed an AI Agent. This AI Agent, akin to a diligent chemist, navigates through the labyrinth of chemical literature, performing several tasks simultaneously to extract the desired information.

在从文献中提取化学反应条件的过程中，我们使用了 AI 代理。该 AI 代理类似于一名勤奋的化学家，穿梭于庞杂的化学文献中，同时执行多个任务以提取所需信息。

Task 1 (Extraction of Chemical Information) **任务 1（化学信息提取）**：

The final task that our AI Agent undertakes is the extraction of chemical information from the now standardized text. This is akin to a chemist analyzing the detailed notes of an experiment to extract key data points about the reaction.

AI 代理执行的最终任务是从标准化文本中提取化学信息。这类似于化学家分析实验的详细记录，以提取反应的关键数据点。

Prompt:

Answer the question as truthfully as possible using the provided context.

Please summarize the following details with units in a json: yield(include%), reactant/reagent(s), solvent(s), product(s). Please note that the content usually includes a general procedure, followed by the specific description of the reaction. The general procedure provides the overall context, and the specific descriptions of each reaction offers unique details.

Please also note that the keys of a json object should be "yield", "reactant/reagent", "solvent", "product". If there do not exist such information, please tell me 'NaN'.

Example input:

General Procedure for the Preparation of Products. [Ni2(iPr2Im)4(μ-COD)] (0.1 mmol, 83 mg), CsF (2 mmol, 304mg), Ar-Bneop (2 mmol), fluoroarene, and toluene (10 mL) were added to a Schlenk tube equipped with a magnetic stirring bar. The reaction mixture was heated at 100 °C for 18 h, and after that H2O (5 mL) was added. The product was extracted with EtOAc (3 × 20 mL), and then the combined organic layers were dried over Na2SO4 and filtered, and the volatiles were removed in vacuo. The product was purified by column chromatography on silica gel using hexane as the eluent. The solvent of the product-containing fraction of the eluent was evaporated in vacuo. The yields provided are based on Ar-Bneop. Spectroscopic Data of the Products. 2,3,4,5,6-Pentafluoro-1,1′-biphenyl (3aa). Following the general procedure, a white solid in 72% yield (351 mg) was obtained from C6F6 (4 mmol, 462 μL) and C6H5-Bneop (2 mmol, 380 mg). 2,3,4,5,6-Pentafluoro-4′-methyl-1,1′-biphenyl (3ab). Following the general procedure, a white solid in 76% yield (390 mg) was obtained from C6F6 (4 mmol, 462 μL) and 4-CH3-C6H4-Bneop (2 mmol, 408 mg).

Example output:

yield	reactant/reagent	solvent	product
72% (351 mg)	C6F6(4 mmol, 462 μL),C6H5-Bneop(2 mmol, 380 mg),fluoroarene,[Ni2(iPr2Im)4(μ-COD)](0.1 mmol, 83 mg),CsF(2 mmol, 304 mg)	toluene(10 mL)	2,3,4,5,6-Pentafluoro-1,1′-biphenyl
76% (390 mg)	C6F6(4 mmol, 462 μL),4-CH3-C6H4-Bneop(2 mmol, 408 mg),fluoroarene,[Ni2(iPr2Im)4(μ-COD)](0.1 mmol, 83 mg),CsF(2 mmol, 304 mg)	toluene(10 mL)	2,3,4,5,6-Pentafluoro-4′-methyl-1,1′-biphenyl

Figure 3. The prompt, example input, example output in chemical reaction information extraction, demonstrating the in-context learning technique.
化学反应信息提取中的提示词、示例输入和示例输出，展示了上下文学习技术。

The result of this task is a structured dataset containing the yield, reactant, catalyst, solvent, and product information for each reaction described in the text. This dataset serves as the final output of our AI Agent, representing the culmination of its diligent and meticulous work, much like the final report of a chemist after a series of experiments.

本任务的结果是一个结构化数据集，包含文本中每个化学反应的产率、反应物、催化剂、溶剂和产物信息。该数据集作为我们 AI 智能体的最终输出，代表其勤勉细致工作的成果，就如同化学家在一系列实验后的最终报告。

Task 2 (Identification of Coreferences: The AI Agent as a Decoder): **任务 2（共指识别：AI 智能体作为解码器）：**

The task of identifying coreferences is the primary and most crucial step in our AI Agent's operation. In chemical literature, coreferences are typically denoted by a combination of a number and a letter, serving as abbreviations for longer, more complex chemical names. This form of shorthand, while effective for human readers familiar with the context, presents a unique challenge for machine-reading and understanding.
识别共指是我们 AI 智能体运行中最主要也是最关键的步骤。在化学文献中，共指通常由数字和字母的组合表示，作为较长、较复杂化学名称的缩写。这种简写方式虽然对熟悉上下文的人类读者有效，但对机器阅读和理解构成了独特挑战。

To tackle this challenge, our AI Agent is equipped with the capability to recognize these specific patterns within the text. This process is not merely a superficial scan of the document; rather, it involves a deep, context-aware analysis of the text. The AI Agent uses the capabilities of GPT, which is designed to understand the context within the investigated text, to identify these coreferences accurately. It makes use of GPT's transformer-based architecture, which allows it to understand the dependencies between words in a sentence and across sentences. Figure 3 shows the prompt, example input, and example output of the in-context learning process in coreference extraction.
为了应对这一挑战，我们的 AI 智能体具备识别文本中这些特定模式的能力。这个过程不仅仅是对文档的表层扫描，而是进行深入的、具备上下文感知的文本分析。AI 智能体利用 GPT 的能力，GPT 被设计用于理解文本中的上下文，从而准确识别这些共指。它依托 GPT 基于 transformer 的架构，能够理解句子内部及句子之间的词语依赖关系。图 3 展示了用于共指提取的上下文学习过程中使用的提示、示例输入和示例输出。

Upon encountering a potential coreference, the AI Agent validates it against the patterns typically used for coreferences in the chemical literature. This step ensures that the identified coreferences are not false positives, such as a number and a letter appearing together by coincidence in the text.
当 AI 智能体遇到潜在的共指时，会将其与化学文献中常见的共指模式进行比对验证。此步骤确保所识别的共指不是误报，例如文本中数字与字母偶然组合出现的情况。

Once a coreference is validated, the AI Agent records it for the subsequent steps. This record-keeping is meticulous and organized, ensuring that each coreference is accurately linked with its position in the text. This step is crucial as it sets the stage for the mapping of coreferences to their full chemical names in the following task.
一旦共指被验证无误，AI 智能体会将其记录下来以供后续使用。这种记录是细致且有组织的，确保每个共指都能准确地与其在文本中的位置关联。这一步至关重要，因为它为接下来的共指与完整化学名称的映射任务奠定了基础。

In summary, the task of identifying coreferences is a complex process that requires the AI Agent to combine pattern recognition with deep, context-aware text analysis. The accuracy and efficiency of this task are critical to the success of the subsequent steps in the AI Agent's operation. Through this task, the AI Agent demonstrates its ability to navigate and understand the complexities of chemical literature, setting the foundation for the extraction of chemical reaction conditions.
总之，共指识别是一个复杂的过程，要求 AI 智能体将模式识别与深度的、具备上下文感知能力的文本分析相结合。该任务的准确性和效率对于 AI 智能体后续操作的成功至关重要。通过此任务，AI 智能体展示了其驾驭和理解化学文献复杂性的能力，为化学反应条件的提取奠定了坚实基础。

Prompt:

I am providing a paragraph from a piece of chemical literature. I would like you to help me identify instances of coreference, where a full chemical name is immediately followed by a shorthand label or alias.

Please provide the coreference in json format. Pay attention to direct aliases that come immediately after the chemical names. If there do not exist such coreference, please tell me "No coreference". Please check carefully about the full chemical name and shorthand label.

Example input:

Tetraethyl (E)-8,9-Bis((Z)-3-ethoxy-3-oxo-2-phenylprop-1-en-1-yl)hexadeca-1,8,15-triene-6,6,11,11-tetracarboxylate, 7c. It was obtained from 3n (25 mg, 0.06 mmol) following the general procedure for cycloisomerization reactions with Cp*RuCl(cod) and purified by flash column chromatography (Hexane/AcOEt, 19:1). Colorless oil (22 mg, 0.03 mmol, 86%).

Example output:

Coreference	Full chemical name
7c	Tetraethyl (E)-8,9-Bis((Z)-3-ethoxy-3-oxo-2-phenylprop-1-en-1-yl)hexadeca-1,8,15-triene-6,6,11,11-tetracarboxylate

Figure 4. The prompt, example input, example output in the identification of coreferences, demonstrating the in-context learning technique.
提示词、示例输入、示例输出在共指识别中体现了上下文学习技术。

Task 3 (Mapping Coreferences to Full Chemical Names) **任务 3（将共指映射到完整化学名称）：**

The third task undertaken by our AI Agent is the mapping of coreferences to their corresponding full chemical names. This task is crucial, as it transforms the shorthand notations into their full forms, thereby enabling a more comprehensive understanding of the chemical reactions described in the text. In essence, this task serves as a bridge, linking the efficient but context-dependent coreferences with their context-independent full chemical names.
我们 AI 智能体承担的第三项任务是将共指映射到其对应的完整化学名称。这项任务至关重要，因为它将简写形式转换为完整形式，从而使对文中所述化学反应的理解更加全面。本质上，该任务充当了桥梁，将高效但依赖上下文的共指连接到不依赖上下文的完整化学名称上。

Building on the coreferences identified in Task 2, the AI

Agent begins the process of mapping these coreferences to their full chemical names. To do this, the AI Agent makes use of GPT's context-understanding capabilities, scanning the text for instances where the coreference is defined, typically in proximity to its first mention.
在任务 2 识别出的共指基础上，AI 智能体开始将这些共指映射到完整的化学名称。为此，AI 智能体利用 GPT 的上下文理解能力，扫描文本中共指被定义的实例，通常这些定义会出现在共指首次提及时的附近。

The AI Agent is designed to handle the complexity of this task, which often involves navigating intricate sentence structures or piecing together information spread across multiple sentences. It employs advanced natural language processing techniques to understand the grammatical structure of the sentence, identify the subject and object, and distinguish between different clauses.
AI 智能体被设计为能够处理此任务的复杂性，这通常涉及理解复杂句子结构或整合分布在多个句子中的信息。它运用先进的自然语言处理技术来理解句子的语法结构，识别主语和宾语，并区分不同从句。

Once a full chemical name corresponding to a coreference is identified, the AI Agent meticulously records this mapping in a structured format. This data structure is designed for flexibility, allowing for updates if a more accurate or complete definition of the coreference is encountered later in the text.
一旦识别出与某个共指对应的完整化学名称，AI 智能体就会以结构化格式仔细记录该映射关系。这种数据结构具有灵活性，允许在文本后续出现更准确或更完整的定义时进行更新。

In summary, Task 3 is a complex linguistic challenge that requires the AI Agent to act as a linguistic cartographer, drawing connections between different points of reference within the text. This task is vital for the subsequent steps in the AI Agent's operation, ensuring that the replacement of coreferences and the extraction of chemical information are based on accurate and complete data.
总而言之，任务 3 是一个复杂的语言挑战，要求 AI 智能体如同语言地图绘制者一样，在文本中各个引用点之间建立联系。此任务对于 AI 智能体后续操作至关重要，确保共指替换和化学信息提取基于准确完整的数据之上。

Task 4 (Replacement of Coreferences with Full Chemical Names): **任务 4（用完整化学名称替换共指）：**

The fourth task is a critical juncture in the AI Agent's operation, where it begins to transform the raw text into a more analyzable form. This task involves replacing all instances of the identified coreferences in the text with their corresponding full chemical names.
第四项任务是 AI 智能体运行中的一个关键转折点，在此阶段，它开始将原始文本转化为更易分析的形式。该任务涉及将文本中所有识别出的共指实例替换为对应的完整化学名称。

This task is implemented by first creating a dictionary where the keys are the coreferences and the values are the corresponding full chemical names. The AI Agent then iterates over the text, and each time it encounters a coreference, it consults the dictionary for the corresponding full chemical name and replaces the coreference with it. This is accomplished using a string replacement function, which scans the text for the coreference patterns and replaces them with the full chemical names.
该任务首先通过建立一个字典实现，其中键为共指，值为对应的完整化学名称。AI 智能体随后遍历文本，每当遇到一个共指时，便查阅字典获取对应的完整名称并进行替换。这一过程通过字符串替换函数实现，该函数扫描文本中的共指模式，并将其替换为完整化学名称。

This replacement process is akin to a search-and-replace operation in a text editor, but it is performed on a much larger scale and with a higher degree of complexity due to the intricacies of chemical nomenclature. The result is a text where the shorthand notations have been replaced with full chemical names, making the subsequent information extraction task more straightforward and accurate.
该替换过程类似于文本编辑器中的查找与替换操作，但由于化学命名法的复杂性，其执行规模更大，复杂度更高。其结果是一个已将简写形式替换为完整化学名称的文本，从而使后续的信息提取任务更直接、更准确。

Engineering Implementation through GPT **通过 GPT 的工程实现：**

From an engineering perspective, our approach to extracting chemical information from the literature is built upon a multi-task framework. This design was motivated by the complex and multi-faceted nature of the problem at hand. By breaking down the overall task into smaller, more manageable subtasks, we are able to tackle each aspect of the problem with specialized strategies, thereby improving the overall effectiveness and efficiency of our system.
从工程的角度来看，我们从文献中提取化学信息的方法基于一个多任务框架。这一设计动因于所面临问题的复杂性与多面性。通过将整体任务拆分为更小、更易管理的子任务，我们能够针对每个方面采用专门的策略，从而提升系统的整体效果与效率。

The first stage in our process involves the use of GPT to generate prompts. GPT is a state-of-the-art language model that has been trained on a vast corpus of text from the internet. It has the ability to generate human-like text based on a given prompt. We leverage this capability to generate initial prompts for our tasks. These prompts act as the starting point for our AI Agent, guiding its focus toward the specific information we are interested in.
我们流程的第一阶段是使用 GPT 生成提示语。GPT 是一种最先进的语言模型，已在来自互联网的大规模文本语料上进行了训练。它能够基于给定提示生成类人文本。我们利用这一能力为各项任务生成初始提示，这些提示作为 AI 智能体的起点，引导其聚焦于我们所关心的具体信息。

However, not all prompts are equally effective. Therefore, we employ an iterative optimization process to refine these prompts. In each iteration, the AI Agent evaluates the effectiveness of the current prompt in extracting the desired information from the text. This evaluation is based

on a performance metric that we define, which could be accuracy, precision, recall, or asny other metric that is relevant to the specific task. If the performance of the prompt is not satisfactory, the AI Agent modifies the prompt and evaluates its performance again. This process is repeated until we obtain a prompt that meets our performance criteria.
然而，并非所有提示语都同样有效。因此，我们采用迭代优化流程对这些提示进行完善。在每次迭代中，AI 智能体都会评估当前提示在从文本中提取目标信息方面的有效性。该评估依据我们自定义的性能指标进行，这些指标可能是准确率、精确率、召回率，或任何与具体任务相关的指标。如果提示的表现不令人满意，AI 智能体将修改该提示，并重新评估其性能。此过程反复进行，直到得到符合性能标准的提示语为止。

Once we have the optimized prompts, we map them to the API function that calls GPT. This mapping process involves translating the prompts into a format that the GPT API can understand. This is a crucial step as it ensures that our prompts are correctly interpreted by GPT, thereby maximizing the effectiveness of our extraction process.
一旦我们获得了优化后的提示语，就将其映射到调用 GPT 的 API 函数中。该映射过程包括将提示语转换为 GPT API 可识别的格式。这是一个关键步骤，它确保我们的提示被 GPT 正确解析，从而最大化信息提取过程的效果。

The final step in our process is the execution of the packaged function. This function takes the text as input, applies the mapped prompts to extract the desired information, and returns this information as output. The function is packaged in a way that it can be easily integrated into other systems or workflows, making our AI Agent a versatile tool for chemical information extraction.
我们流程的最后一步是执行封装好的函数。该函数以文本作为输入，应用已映射的提示来提取所需信息，并将这些信息作为输出返回。此函数被封装为易于集成至其他系统或工作流程的形式，使我们的 AI 智能体成为一个多功能的化学信息提取工具。

In summary, our engineering approach is a careful orchestration of several components, each designed with a specific purpose and all working together towards the common goal of effective and efficient chemical information extraction. By leveraging the power of GPT and the flexibility of our multi-task framework, we are able to tackle the complex task of extracting chemical information from the literature, paving the way for new possibilities in the field of chemical informatics.
总之，我们的工程方法是对多个组件的精心编排，每个组件都有其特定的目的，并共同致力于实现高效、有效的化学信息提取。通过发挥 GPT 的强大能力以及我们多任务框架的灵活性，我们能够应对从文献中提取化学信息的复杂任务，为化学信息学领域开辟新的可能性。

This AI Agent, through its systematic and rigorous approach, mirrors the meticulous nature of a chemist, making it an apt tool for the task at hand. By breaking down the complex task of chemical information extraction into manageable subtasks, we have created an AI Agent that is not only efficient but also scalable, paving the way for future advancements in the field.
这个 AI 智能体通过其系统化和严谨的方法，展现出化学家一丝不苟的特质，使其成为该任务的理想工具。通过将复杂的化学信息提取任务拆分为可管理的子任务，我们创建了一个不仅高效且具备可扩展性的 AI 智能体，为该领域的未来发展铺平了道路。

IV. Conclusion 结论

This research introduced an innovative AI agent that leverages LLMs to automate the extraction of high-fidelity chemical data from chemical literature. Our system has demonstrated superior performance in terms of accuracy, recall, and F1 score. The agent's ability to act as a chemistry assistant streamlines the data collection and analysis process, leading to significant savings in manpower and enhancements in performance. The agent's iterative optimization and prompt generation capabilities have shown to be particularly effective in dealing with the variegated and unstructured nature of literature. Additionally, the comparison with human experts has validated the AI agent's efficiency and correctness, showcasing its potential to revolutionize chemical data management and utilization.
本研究提出了一种创新的 AI 智能体，它利用大语言模型（LLM）自动从化学文献中提取高保真度的化学数据。我们的系统在准确率、召回率和 F1 分数方面表现出优越性能。该智能体作为化学助手的能力简化了数据收集与分析过程，显著节省了人力成本并提升了性能。其迭代优化与提示语生成能力在应对文献多样性和非结构化特征方面尤为有效。此外，与人类专家的对比验证了该 AI 智能体的效率与准确性，展示了其在化学数据管理与利用方面的变革潜力。

Future research will focus on refining the AI agent's capabilities, expanding its applications, and integrating it with other advanced technologies to further elevate its performance and utility. The current work has laid a solid foundation for AI's role in chemical literature mining, promising to accelerate advancements in material synthesis, drug discovery, and a myriad of other areas within the field of chemistry.
未来的研究将致力于进一步提升该 AI 智能体的能力，拓展其应用范围，并将其与其他先进技术相集成，以进一步提高其性能与实用性。目前的工作为 AI 在化学文献挖掘中的作用奠定了坚实基础，有望加速材料合成、药物发现及化学领域中诸多其他方向的进步。

References

[1] K. Chen, G. Chen, J. Li, Y. Huang, E. Wang, T. Hou, and P.-A. Heng, "MetaRF: attention-based random forest for reaction yield prediction with a few trails," Journal of Cheminformatics, vol. 15, no. 1, pp. 1–12, 2023.

[2] K. Chen, J. Li, K. Wang, Y. Du, J. Yu, J. Lu, G. Chen, L. Li, J. Qiu, Q. Fang et al., "Towards an automatic ai agent for reaction condition recommendation in chemical synthesis," arXiv preprint arXiv:2311.10776, 2023.

[3] H. Cui, Y. Du, Q. Yang, Y. Shao, and S. C. Liew, "Llmind: Orchestrating ai and iot with llms for complex task execution," arXiv preprint arXiv:2312.09007, 2023.
[4] Y. Du, S. C. Liew, K. Chen, and Y. Shao, "The power of large language models for wireless communication system development: A case study on fpga platforms," arXiv preprint arXiv:2307.07319, 2023.