

SynAsk: Unleashing the Power of Large Language Models in Organic Synthesis

Chonghuan Zhang^{1†}, Qianghua Lin^{1†}, Biwei Zhu^{2†}, Haopeng Yang², Xiao Lian²,
Hao Deng², Jiajun Zheng², Kuangbiao Liao^{1*}

¹Guangzhou National Laboratory, Guangzhou, Guangdong, PR China, 510005.

²AIChemEco Inc., Guangzhou, Guangdong, PR China, 510005.

*Corresponding author(s). E-mail(s): liao_kuangbiao@gzlab.ac.cn;

†These authors contributed equally to this work.

Abstract

The field of natural language processing (NLP) has witnessed a transformative shift with the emergence of large language models (LLMs), revolutionizing various language tasks and applications, and the integration of LLM into specialized domains enhances their capabilities for domain-specific applications. Notably, NLP has made significant strides in organic chemistry, particularly in predicting synthetic tasks, paving the way for the development of LLMs tailored to the organic chemistry field.

自然语言处理 (NLP) 领域随着大规模语言模型 (LLM) 的出现经历了革命性的变革，极大地推动了各种语言任务和应用的发展。同时，将 LLM 集成到特定领域可增强其在该领域的专用应用能力。值得注意的是，NLP 在有机化学领域取得了重要进展，尤其是在合成任务预测方面，这为专门针对有机化学领域的 LLM 的开发奠定了基础。

In this work, we introduce SynAsk, a comprehensive organic chemistry domain-specific LLM platform developed by AIChemEco Inc. By finetuning an LLM with domain-specific data and integrating it with a chain of thought approach, SynAsk seamlessly accesses our knowledge base and advanced chemistry tools in a question-and-answer format. This includes functionalities such as a basic chemistry knowledge base, molecular information retrieval, reaction performance prediction, retrosynthesis prediction, chemical literature acquisition, and more.

在本研究中，我们介绍了由 AIChemEco Inc. 开发的综合性有机化学专用 LLM 平台——SynAsk。通过使用领域专用数据对 LLM 进行微调，并结合“思维链”方法，SynAsk 能够在问答模式下无缝访问我们的知识库和先进的化学工具。其功能包括基础化学知识库、分子信息检索、反应性能预测、逆合成预测、化学文献获取等。This novel methodology synergizes fine-tuning techniques with external resource integration, resulting in an organic chemistry-specific model poised to facilitate research and discovery in the field. Accessible via <https://synask.aichemeco.com>, SynAsk represents a significant advancement in leveraging NLP for synthetic applications.

这一创新方法将微调技术与外部资源集成相结合，构建了一个专注于有机化学的模型，有望推动该领域的研究与发现。SynAsk 可通过<https://synask.aichemeco.com>访问，标志着在合成应用中利用 NLP 的重大进展。

Keywords: Large Language Model, AI in Chemistry, organic synthesis, retrosynthesis

1 Introduction

In recent years, the field of natural language processing (NLP) has undergone a revolutionary shift with the emergence of large language models (LLMs), advanced artificial intelligence systems trained on massive datasets to understand and generate human-like text across various language tasks and applications. At the core of LLMs lies the remarkable technology of generative pre-trained transformers (GPT) [1]. Developed by OpenAI, GPT models like ChatGPT [2] have gained widespread attention and adoption for their capacity to produce coherent and contextually relevant text. ChatGPT, in particular, represents a milestone in conversational AI, enabling human-like interactions that go beyond scripted responses. Evolving from ChatGPT to GPT-4 [3] through continual learning from vast datasets allows these models to grasp nuances of language and context, making them versatile tools for diverse tasks, from assisting in creative writing to generating videos. While GPT models have dominated the landscape, other models like Qwen [4] and LLaMA [5] also make significant contributions to the field, and these models are open-sourced for the community to utilize. Qwen, primarily trained from Mandarin Chinese language sources, is renowned for its robustness in question-answering tasks, leveraging a different architecture and training approach. On the other hand, LLaMA specializes in language understanding and inference tasks, offering unique capabilities in semantic analysis and knowledge extraction.

近年来，自然语言处理（NLP）领域随着大规模语言模型（LLM）的出现经历了革命性的变革。这些先进的人工智能系统通过大规模数据训练，能够理解和生成类人文本，广泛应用于各种语言任务和应用。LLM 的核心技术是生成式预训练变换器（GPT）[1]。由 OpenAI 开发的 GPT 模型，如 ChatGPT [2]，因其生成连贯且符合语境的文本能力而受到广泛关注和应用。特别是 ChatGPT 在对话式人工智能方面取得了里程碑式的突破，使其能够实现超越预设脚本的类人交互。随着从 ChatGPT 发展到 GPT-4 [3]，这些模型通过不断学习海量数据，更加精确地把握语言和语境的细微差别，使其成为多用途工具，可用于从创意写作到视频生成等多种任务。尽管 GPT 模型在该领域占据主导地位，但其他模型如 Qwen [4] 和 LLaMA [5] 也做出了重要贡献，并且这些模型是开源的，供社区使用。Qwen 主要从中文语料库训练，以其在问答任务中的卓越表现而闻名，并采用了不同的架构和训练方法。而 LLaMA 则专注于语言理解和推理任务，在语义分析和知识提取方面展现了独特的能力。

Beyond ChatGPT and other models, LLMs encompass a spectrum of applications across vertical domains. Domain-specific and customized data have been collected and labeled to fine-tune these LLMs. One of the key benefits of vertically specialized LLMs is their capacity to bolster domain-specific applications. By refining their expertise within a particular domain, these models possess the capability to delve deeply into the nuances of the subject matter, rendering them invaluable tools for professionals operating in specialized domains. For instance, a legally specialized LLM, namely DISC-LawLLM [6], can provide precise legal counsel, draft contracts, and facilitate intricate legal research, thereby streamlining processes and conserving resources for legal practitioners. Similarly, a medically specialized LLM, namely MultiMedQA [7], can assist physicians in diagnosing rare conditions, proposing tailored treatment plan, and staying updated on the latest technologies in medical research.

除了 ChatGPT 及其他模型之外，LLM 还广泛应用于垂直领域。研究人员收集并标注了特定领域的数据，以便对这些 LLM 进行微调。垂直领域专用 LLM 的一大优势在于其能够增强领域特定的应用能力。通过在特定领域内深度优化，这些模型能够深入理解学科的细微差别，从而成为专业人士的重要工具。例如，专注于法律的 LLM——DISC-LawLLM [6]，可以提供精准的法律咨询、起草合同，并促进复杂的法律研究，从而优化流程并节约法律从业者的资源。同样，专注于医学的 LLM——MultiMedQA [7]，能够帮助医生诊断罕见疾病、制定个性化治疗方案，并获取医学研究的最新进展。

The integration of NLP into organic chemistry has brought about a revolution in research and discovery. Molecules and reactions can now be represented using SMILES (Simplified Molecular Input Line Entry System), a textual notation for depicting high-dimensional chemical structures [8]. NLP techniques have been employed to tackle organic synthesis tasks using SMILES strings, treating the synthesis problem as a sequence generation task. This approach involves training machine learning models to predict the sequence of molecules and reactions necessary to synthesize a target molecule based on desired products. These models learn from extensive datasets of annotated reactions, where each reaction is represented as a sequence of SMILES strings. Leveraging the patterns and rules encoded in the data, these models can generate plausible synthesis pathways [9, 10].

NLP 技术与有机化学的结合极大地推动了该领域的研究和发现。分子和化学反应现在可以使用 SMILES（简化分子输入行表示法）表示，这是一种用于描述高维化学结构的文本符号 [8]。NLP 技术已经被应用于处理基于 SMILES 字符串的有机合成任务，将合成问题视为序列生成任务。这种方法涉及训练机器学习模型，以预测合成目标分子所需的分子和反应序列。该类模型通过学习大量标注的反应数据集，每个反应均表示为一组 SMILES 字符串，并利用数据中的模式和规则生成合理的合成路径 [9, 10]。

LLMs have found applications in organic chemistry as well. However, without further tuning with organic chemistry domain-specific data, researchers have evaluated five LLMs in tasks related to organic chemistry, including reaction prediction and retrosynthesis. While these models provide reasonable results in classification or ranking tasks like yield prediction and reagent selection, they face challenges in generative tasks that require a deep understanding of molecular structures [11]. This difficulty may stem from the highly experimental nature of organic chemistry, the lack of labeled data, and the limited scope and applicability of computational tools in this field [12]. To bridge this gap and motivate further exploration of LLM potential in chemistry, several domain-specific LLMs for organic chemistry have been developed. ChemCrow [12] was the first proposed LLM in chemistry aimed at enhancing its capabilities through external tools. It employs chain-of-thought (CoT) strategies [13], which are a series of intermediate reasoning steps to improve LLMs' ability to understand tasks from prompts. ChemCrow also utilizes LangChain[14], a framework to connect the LLM with multiple external tools downstream to solve specific tasks and return answers back to the LLM. However, this method relies on the reliability of tools, and general LLMs may not comprehensively understand prompts and link to the correct tools to solve specific tasks. Another approach, ChemLLM [15], was proposed to transform structured chemical data into forms suitable for LLMs to fine-tune the LLaMA model. ChemLLM excels in tasks such as cheminformatic programming. However, its performance may not be as robust as comprehensive models like ChatGPT-4, possibly due to human biases in the collection of incomplete structural chemical data.

LLM 在有机化学领域也得到了应用。然而，如果不对其进行有机化学领域的专门微调，研究人员评估了五种 LLM 在有机化学相关任务（包括反应预测和逆合成）中的表现。虽然这些模型在分类或排序任务（如收率预测和试剂选择）方面表现合理，但在需要深入理解分子结构的生成任务中面临挑战 [11]。这种困难可能源于有机化学的高度实验性、缺乏标注数据以及计算工具在该领域的适用范围有限 [12]。为了弥合这一差距，并进一步激发 LLM 在化学领域的潜力，研究人员开发了多种有机化学专用 LLM。例如，ChemCrow [12] 是第一个提出的化学 LLM，旨在通过外部工具增强其能力。它采用思维链（CoT）策略 [13]，即一系列中间推理步骤，以提高 LLM 从提示中理解任务的能力。此外，ChemCrow 还利用 LangChain[14]，一个用于连接 LLM 与多个外部工具的框架，以解决特定任务并将答案返回给 LLM。然而，该方法依赖于工具的可靠性，而通用 LLM 可能无法完全理解提示并正确链接到合适的工具以解决特定任务。另一种方法是 ChemLLM [15]，该模型旨在将结构化化学数据转换为适合 LLM 微调 LLaMA 模型的格式。ChemLLM 在化学信息学编程等任务中表现出色，但其性能可能不及 ChatGPT-4 等综合模型，可能是由于在收集结构化化学数据时存在人为偏见和数据不完整的原因。

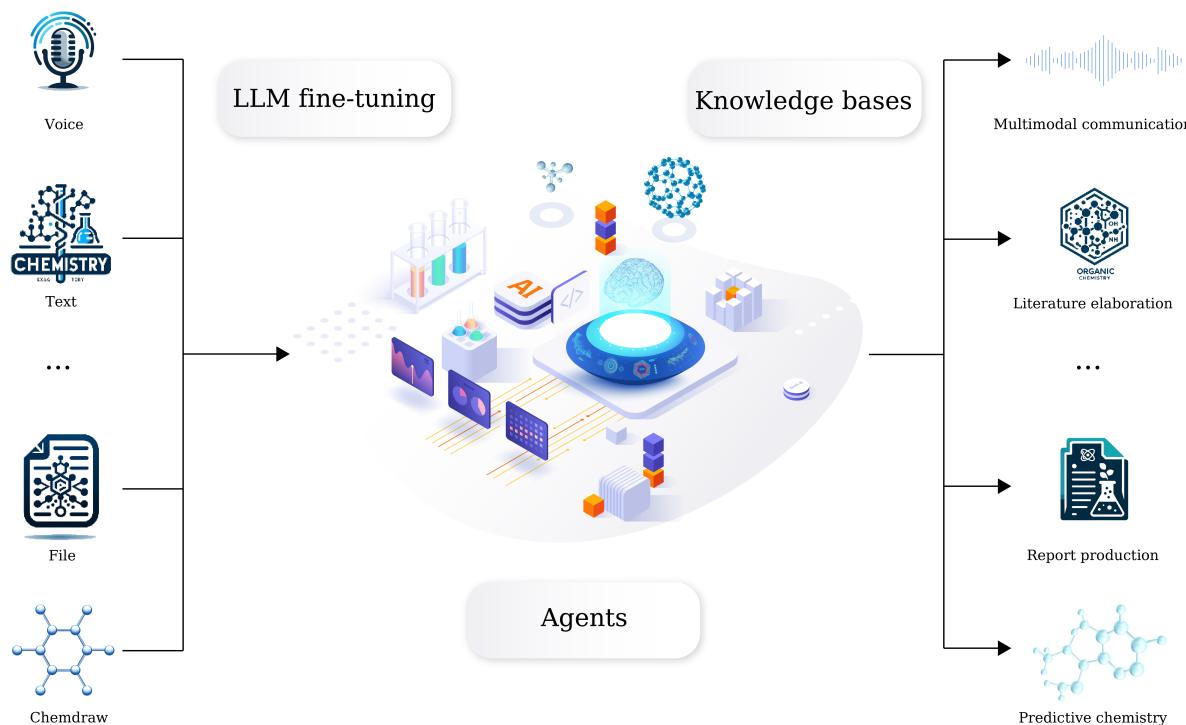


Fig. 1 The overview of SynAsk platform.

We have long been dedicated to AI in chemistry research, developing a series of machine learning and computational based tools to solve fundamental organic chemistry tasks. However, we recognize that directly connecting these tools to large language models (LLMs) may not yield appropriate results. Here we introduce a comprehensive domain-specific LLM for organic chemistry developed by AIChemEco, named SynAsk, as shown in Figure 1. An LLM was refined using a limited set of domain-specific chemistry data and integrate it with a chain-of-thought approach to understand user prompts. Our aim is to utilize Langchain to seamlessly connect SynAsk with our existing suite of tools, addressing specific user inquiries, drawing on the framework of Langchain–chatchat [16]. This methodology allows us to combine fine-tuning techniques with the integration of external resources, resulting in the development of an organic chemistry-specific model. The model can be accessed at <https://synask.aichemeco.com>.

我们长期致力于化学研究中的人工智能（AI），开发了一系列基于机器学习和计算的工具，以解决基础有机化学任务。然而，我们认识到，直接将这些工具与大规模语言模型（LLM）连接可能无法获得理想的结果。因此，我们在此介绍由 AIChemEco 开发的有机化学专用 LLM——SynAsk，如Figure 1所示。我们使用有限的领域专用化学数据对 LLM 进行了微调，并结合“思维链”方法来增强其对用户输入的理解。我们的目标是利用 Langchain 将 SynAsk 无缝连接到我们现有的工具集，以回答用户的特定问题，并借鉴 Langchain–chatchat 框架 [16]。这一方法使我们能够将微调技术与外部资源整合，从而开发出一个专注于有机化学的模型。该模型可通过<https://synask.aichemeco.com>访问。

2 Methods

To construct the comprehensive model integration platform, our approach unfolds along three primary dimensions: utilizing a powerful foundation LLM as the base for SynAsk, crafting more effective prompts and implementing fine-tuning to the foundation model, and connecting with multiple tools to assemble a chemistry domain-specific model platform.

为了构建一个综合的模型集成平台，我们的策略涵盖三个主要方面：选择强大的基础 LLM 作为 SynAsk 的底层模型、设计更有效的提示词并对基础模型进行微调、以及连接多个工具以构建化学领域专用的模型平台。

2.1 Selection of a foundation LLM

Through various experiments, we have recognized that for the foundation LLM to effectively understand prompts from end-users and apply insights to decide whether to provide LLM inference answers or use specific tools to resolve downstream tasks, it needs to have at least 14 billion parameters. Therefore, only foundation models with over 14 billion parameters were considered. The capabilities of the LLM was assessed using indicators such as Massive Multi-task Language Understanding (MMLU) [17], Multi-level multi-discipline Chinese evaluation (C-Eval) [18], GSM8K [19], (BIG-Bench-Hard) BBH [20] and Measuring massive multitask language understanding in Chinese (CMMLU) [21], as elaborated in Section S1 of Electronic Supplementary Information (ESI). These indicators collectively offer a comprehensive assessment of a model's proficiency, covering areas such as linguistic understanding, mathematical reasoning, contextual comprehension, multi-modal integration, and the application of Chain-of-Thought (CoT), which evaluates the fluency of LLMs' integration with external tools. This evaluation framework underscores the essential and diverse skills a model must possess to adeptly address complex real-world problems.

通过多次实验，我们认识到，基础 LLM 若要有效理解用户的提示词，并能自主判断是提供 LLM 推理结果，还是调用特定工具来解决下游任务，则其参数量至少需要达到 140 亿。因此，我们仅考虑参数量超过 14 亿的基础模型。LLM 的能力评估基于以下指标：大规模多任务语言理解（MMLU）[17]、多层次多学科中文评测（C-Eval）[18]、GSM8K [19]、大规模基准测试 BBH (BIG-Bench-Hard) [20]，以及大规模中文多任务语言理解（CMMLU）[21]，详细内容见电子补充信息（ESI）的Section S1。这些指标共同提供了对模型能力的全面评估，涵盖语言理解、数学推理、上下文理解、多模态整合，以及“思维链”方法（CoT）的应用，该方法用于评估 LLM 与外部工具的整合流畅度。此评估框架强调了模型必须具备的核心能力，以便能够有效解决复杂的现实问题。

As indicated in Table S1 [4], the Qwen series [4] outperforms other models with equivalent parameter counts, including LLaMA2 [22], ChatGLM2 [23], InterLM [24], Baichuan2 [25] and Yi [26] in these areas. Additionally, our testing has confirmed that the Qwen series is more compatible with our framework, especially with the release of Qwen-1.5, which provides us with more options. We acknowledge that GPT series [2], particularly GPT-4 [3], scores higher than Qwen. However, at the time of this work, GPT-4 has not been open-sourced and requires paid API tokens to use as a foundation model. To ensure SynAsk remains publicly accessible, we opted to use only open-sourced foundation LLMs and developed an architecture that allows for smooth switching of the foundation LLM, as discussed in Section 2.4.

如Table S1所示 [4]，Qwen 系列 [4] 在多个评估指标上均优于其他参数量相当的模型，包括 LLaMA2 [22]、ChatGLM2 [23]、InterLM [24]、Baichuan2 [25] 和 Yi [26]。此外，我们的测试表明，Qwen 系列与我们的框架兼容性更高，特别是 Qwen-1.5 的发布为我们提供了更多选择。我们承认，GPT 系列 [2]，尤其是 GPT-4[3]，在某些评估指标上的得分高于 Qwen。然而，在本研究开展时，GPT-4 尚未开源，并且使用该模型需要付费 API。因此，为了确保 SynAsk 保持公开可用，我们选择仅使用开源的基础 LLM，并开发了一种架构，以便在未来能够顺畅地更换基础 LLM，相关内容将在Section 2.4中讨论。

2.2 Refinement to more Reasonable Prompt

To improve the model's performance in two key areas—providing more targeted responses in the chemical domain and enhancing its ability to efficiently utilize tools—we refined our prompt templates through iterative testing and adjustments. We guide the model to generate responses that are not only accurate but also consistent with specific demand expectations. This process encourages the model to become more deeply involved in the task at hand, reducing ambiguity and focusing its attention. These optimized guidance models function as both competent chemists and skilled tool users, establishing a more focused, efficient, and effective interaction between the model and the user.

为了提升模型在两个关键领域的性能——在化学领域提供更有针对性的回答，并增强其高效利用工具的能力——我们通过迭代测试和调整优化了提示模板。我们引导模型生成既准确又符合特定需求预期的回答。此过程使模型能够更深入地参与任务，减少歧义，并增强其关注度。这些优化后的引导模型既是专业的化学家，也是熟练的工具使用者，从而建立起模型与用户之间更聚焦、高效且精准的交互。

In our integrated platform, utilizing the classification function of LLMs is particularly crucial, as illustrated in [Figure 2](#). Since this platform extends from our existing NLP project, we believe it inherently possesses enhanced capabilities. To further train it, we employ a tailored hint project, where the model's role is set as a chemist evaluating and scoring the generated results. This project provides several examples to guide the model. This setup enables the model to discern whether responses augmented by the knowledge database meet the criteria, thereby classifying the results into those that meet expectations and those that do not.

在我们的集成平台中，利用 LLM 的分类功能尤为关键，如[Figure 2](#)所示。由于该平台是我们现有 NLP 项目的延伸，我们认为它在本质上具备更强的能力。为了进一步训练模型，我们采用了一个定制的提示工程项目，将模型设定为化学家，以评估并评分其生成的结果。该项目提供了一些示例，以引导模型学习。这一设定使得模型能够判断由知识库增强的回答是否符合标准，从而将结果分类为符合预期的和不符合预期的。

2.3 Fine-tuning of the LLM

The selected model underwent fine-tuning to specialize it further in the field of chemistry, ensuring its engagement in professional chemical dialogues, particularly in organic synthesis. The fine-tuning process comprised two iterations, with data processed accordingly for each iteration.

所选模型经过微调，使其在化学领域的专业能力进一步专业化，确保其能够参与专业的化学对话，特别是在有机合成方面。微调过程包括两个迭代步骤，并针对每个阶段相应处理数据。

- The first iteration was supervised fine-tuning: This stage focused on enhancing the model's cognitive abilities, reinforcing its identity as an expert in chemistry. The objective was to delve deeper into the model's capabilities within the chemistry domain without expanding its original data source. This approach allowed the model to utilize existing data more effectively to solve chemical problems. 第一阶段是监督微调：此阶段的重点在于增强模型的认知能力，并强化其化学专家的身份。目标是在不扩展原始数据来源的情况下，进一步挖掘模型在化学领域的专业能力。该方法使模型能够更有效地利用现有数据来解决化学问题。
- The second iteration was instruction-based fine-tuning: The aim here was to improve the model's reasoning and tool invocation capabilities, thereby enhancing its chain of thought. It learned to differentiate between various types of chemical identifiers, such as SMILES and CAS numbers, rather than treating them as ordinary words or sequences of numbers.

第二阶段是基于指令的微调：此阶段的目标是提升模型的推理能力和工具调用能力，从而增强其“思维链”能力。模型学习区分不同类型的化学标识符，如 SMILES 和 CAS 号，而不是将它们简单地视为普通单词或数字序列。

The rationale for dividing the fine-tuning into two stages is threefold:

将微调分为两个阶段的原因如下：

- Clear and Controllable Training: Each fine-tuning task addressed a specific sub-problem, ensuring clarity and controllability in the training process and outcomes. This approach facilitates adjustments and improvements based on the results of previous fine-tuning, gradually enhancing the model's performance on specific tasks.
明确且可控的训练：每个微调任务都针对特定的子问题进行，使训练过程和结果具有清晰性和可控性。此方法便于根据前一次微调的结果进行调整和优化，从而逐步提高模型在特定任务上的表现。
- Prevention of Interference: Segregating the tasks prevents confusion and interference between them. Combining all tasks into a single fine-tuning session might lead to instability in training or reduced performance.
避免任务干扰：将任务分离可以防止相互混淆和干扰。如果所有任务在一次微调过程中进行，可能会导致训练不稳定或性能下降。
- Accelerated Training: This approach speeds up the training process. By simplifying each fine-tuning task, the training becomes more efficient, yielding quicker results and feedback. The shorter training times for each task contribute to a faster overall training cycle.
加速训练过程：此方法能够加快训练速度。通过简化每个微调任务，训练过程变得更加高效，从而更快地获得结果和反馈。缩短每个任务的训练时间有助于加快整体训练周期。

After fine-tuning, detailed techniques, procedures, and the necessary equipment are elaborated in [Section S2](#) of the ESI. Post-fine-tuning, our emphasis mainly lies on the model's ability to demonstrate Chain of Thought (CoT) in its output. Following the fine-tuning process, we provide two examples of the model's simplified output format:

微调完成后，具体的技术、流程以及所需设备将在电子补充信息 (ESI) 的[Section S2](#)中详细说明。在微调之后，我们主要关注模型在输出中展现“思维链”(CoT)能力的表现。在微调完成后，我们提供了两个示例，以展示模型的简化输出格式：

Prompt:	What is the SMILES of toluene?
Response:	Action: get_SMILES
	Action Input: {"query": "toluene"}
Prompt:	What is the name of CC1=CC=CC=C1?
Response:	Action: CASToName
	Action Input: {"query": "CC1=CC=CC=C1"}

Notably, the power of these fine-tuned results is significantly enhanced when used in conjunction with appropriately designed prompting strategies and specially designed tool formats. These responses demonstrate the model's ability to identify the required action and its corresponding input from the prompts. However, within our framework, these responses are not the final outcome. Instead, they serve as intermediate prompts to be re-fed into the model. This intermediary step is pivotal, enabling the model to discern the specific tool it requires (e.g., 'get_SMILES' for the initial example) and to process the "Action Input" (e.g., "query: 'toluene'") utilizing the designated tool. Subsequently, the expansive model amalgamates the tool's output with its vast knowledge base, culminating in the generation of a final answer.

值得注意的是，当这些经过微调的结果与精心设计的提示策略和专门制定的工具格式结合使用时，其性能得到了显著提升。这些响应展示了模型从提示中识别所需操作及其对应输入的能力。然而，在我们的框架中，这些响应并不是最终结果，而是作为中间提示重新输入模型的关键步骤。此中间过程至关重要，它使模型能够识别所需的具体工具（例如，在第一个示例中调用‘get_SMILES’），并利用指定的工具处理“操作输入”（例如“query: ‘toluene’”）。随后，模型将工具的输出与其庞大的知识库相结合，从而生成最终答案。

2.4 SynAsk Architecture

In the final phase, we implemented the LangChain framework to seamlessly integrate our local knowledge base with both internal and external open-source tools and APIs. Its primary role is to interpret the outputs from the language models, converting them into a format understandable by external tools, thus facilitating the execution of corresponding actions. Simultaneously, it translates the responses from these tools back into a form comprehensible by the language models. Furthermore, LangChain's support for context management enables it to track the interaction history between users and the system. This enhances the system's ability to understand user intentions and maintain session continuity during interactions with external tools. Its scalability ensures that the system can adapt to technological advancements and changing user demands, providing a dynamic and responsive framework for our integration needs. The LangChain framework serves as a pivotal bridge, culminating in a logically coherent and systematically robust integration platform known as SynAsk.

在最终阶段，我们采用 LangChain 框架，将本地知识库与内部和外部的开源工具及 API 无缝集成。其主要作用是解释语言模型的输出，并将其转换为外部工具可理解的格式，从而促进相应操作的执行。同时，它还能将这些工具的响应转换为语言模型能够理解的形式。此外，LangChain 对上下文管理的支持使其能够跟踪用户与系统之间的交互历史，从而增强系统对用户意图的理解，并在与外部工具交互时保持会话连续性。其可扩展性确保了系统能够适应技术进步和不断变化的用户需求，为我们的集成需求提供了一个动态且灵活的框架。LangChain 框架作为核心桥梁，最终形成了一个逻辑清晰、系统稳健的集成平台——SynAsk。

The structural framework of SynAsk is illustrated in [Figure 2](#). Initially, it can accept both voice and text inputs as queries, which are then segmented into multiple tasks by a LLM and matched against our knowledge base. At this stage, users also have the option to upload their local files as supplementary knowledge or directly engage in conversations with the uploaded files. Once matching texts are obtained, the large model synthesizes the content along with its understanding of the question to deduce a conclusion, thereby generating a result. Subsequently, the model evaluates this result to determine if it meets the expected criteria. If the outcome is deemed satisfactory, it is directly outputted as the Final Answer. Conversely, if the results do not meet expectations, we will enter our customized Agent Q&A mode and call our tools to answer. Finally, the tool output is combined with the LLM's self-knowledge to generate the final answer.

SynAsk 的结构框架如[Figure 2](#)所示。最初，它可以接受语音或文本输入作为查询，随后由LLM将其分解为多个任务，并与我们的知识库进行匹配。在此阶段，用户还可以选择上传本地文件作为补充知识，或直接与上传的文件进行交互。一旦获得匹配文本，大型模型便会整合内容，并结合其对问题的理解推导出结论，从而生成结果。接下来，模型评估该结果是否符合预期标准。如果结果符合预期，它将直接输出为最终答案 (Final Answer)。如果结果不符合预期，则系统将进入定制的代理问答 (Agent Q&A) 模式，并调用我们的工具进行解答。最终，工具的输出将与 LLM 的自身知识结合，生成最终答案。

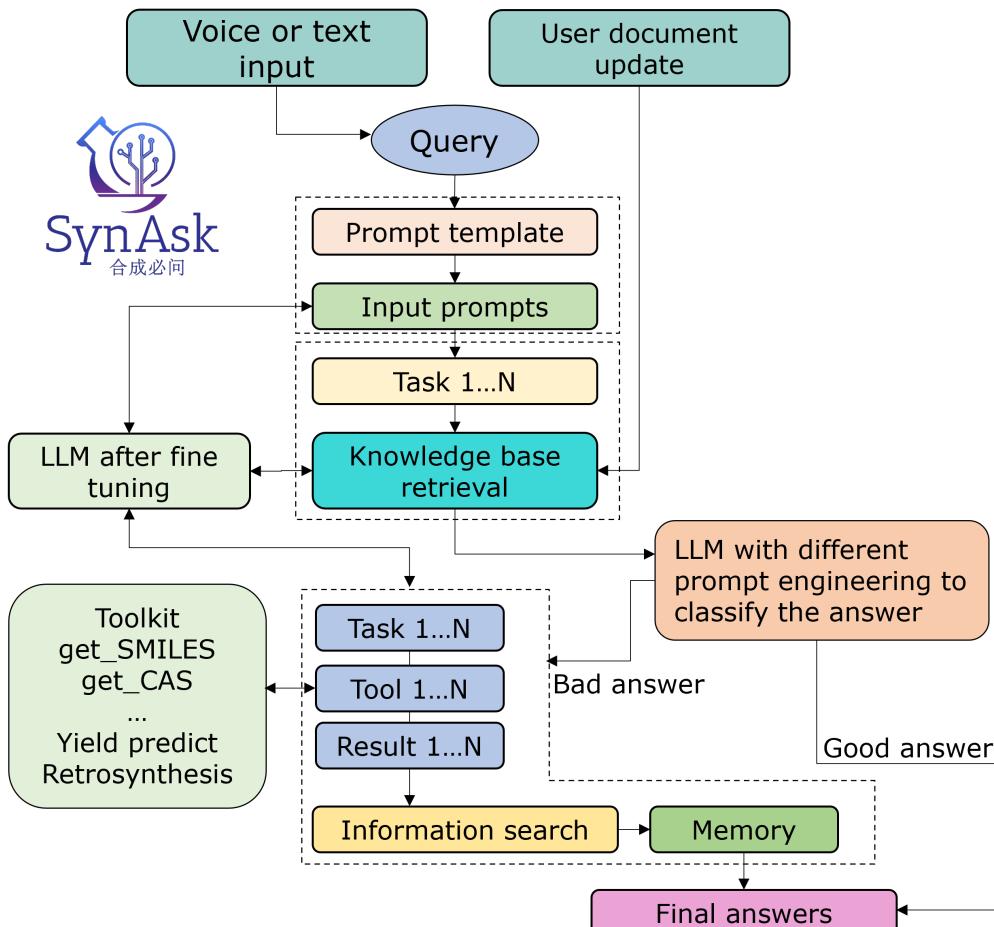


Fig. 2 The workflow of the SynAsk platform: from the input to the final answer.

In the SynAsk architecture, although we currently utilize Qwen-1.5 as the foundation LLM, we recognize the ongoing revolutions in LLM technology. Consequently, we have developed a workflow to swiftly adjust the foundation model and fine-tune the domain-specific data. This approach ensures that SynAsk can continuously update and iterate, leveraging the latest advancements in foundation LLMs.

在 SynAsk 架构中，尽管目前我们采用 Qwen-1.5 作为基础 LLM，但我们深知 LLM 技术的持续革新。因此，我们开发了一套工作流程，以便快速调整基础模型并微调领域专用数据。此方法确保 SynAsk 能够不断更新迭代，充分利用基础 LLM 的最新进展。

2.5 SynAsk Toolsets

Chemoinformatic tools are seamlessly connected with SynAsk through LangChain to provide comprehensive organic synthesis answers. This includes a variety of machine learning-powered tools developed both internally and by external teams, all dedicated to solving organic synthesis tasks. At the time of publishing this work, **12** internal tools and **10** external tools have been integrated into SynAsk. External tools are appropriately cited with their origins. With the rapid development of this field, we anticipate an increasing influx of tools joining SynAsk. These tools are categorized into molecular, reaction tools, and others, with a number of advanced in-house tools elaborated in [Section 2.5.5](#).

通过 LangChain，化学信息学工具被无缝连接至 SynAsk，以提供全面的有机合成解答。这些工具包括由内部团队和外部团队开发的各种基于机器学习的工具，均致力于解决有机合成任务。截至本研究发表时，SynAsk 已集成 **12** 个内部工具和 **10** 个外部工具。外部工具均已注明来源。随着该领域的快速发展，我们预计将有越来越多的工具加入 SynAsk。这些工具被分类为分子工具、反应工具及其他类别，其中多个先进的内部工具将在 [Section 2.5.5](#) 中详细介绍。

2.5.1 Molecular Information Retrieval

This category encompasses tools designed for querying various molecular identifiers and properties. Functions include retrieving Chemical Abstracts Service (CAS) numbers, Simplified Molecular Input Line Entry System (SMILES) strings, molecular weights, assessing molecular similarity, identifying types of functional groups, and checking the regulatory status of molecules. The respective tools for these purposes are

该类别包括用于查询各种分子标识符和性质的工具。其功能包括检索化学文摘服务 (CAS) 号、简化分子输入行表示法 (SMILES) 字符串、分子量、评估分子相似性、识别官能团类型以及检查分子的法规状态。用于这些目的的相应工具包括：

- `get_cas` – for CAS numbers retrieval [27]
- `get_smiles` – for obtaining SMILES strings [27]
- `CASToName` – to convert CAS numbers to chemical names [28]
- `SMILESToName` – to convert SMILES strings to chemical names [28]
- `get_mol_weight` – for calculating molecular weights
- `get_mol_similarity` – to determine molecular similarity
- `check_functional_groups` – for functional group identification
- `ControlmolCheck` – to check if a molecule is controlled

2.5.2 Chemical Reaction and Retrosynthesis

This category aids in querying chemical reaction conditions, planning chemical reaction pathways, predicting chemical reaction yields, performing retrosynthetic analysis, and predicting reaction derivatives. Tools provided for these functions include

此类别可帮助查询化学反应条件、规划化学反应途径、预测化学反应产量、进行逆合成分析和预测反应衍生物。为这些功能提供的工具包括：

- `Get_condition` – to query chemical reaction conditions
- `ReactionPlanner` – for planning chemical reaction pathways [29]
- `ReagentsPredict` – to predict reagents in chemical reactions
- `YieldPredict` – for predicting chemical reaction yields
- `Retrosynthesis` – to perform retrosynthetic analysis
- `DerivatePredict` – to predict the derivatives from a chemical reaction, using reactants' names or SMILES, enhancing the exploration of reaction outcomes.
- `AutoMapping` – to identify the position of each atom in the molecules before and after a chemical reaction [30][31]

2.5.3 Acquisition of Chemical Literature and Knowledge

Dedicated to acquiring chemical literature and extracting chemical knowledge, tools in this section include:
本节中的工具致力于获取化学文献和提取化学知识，包括：

- `Get_literature` – for retrieving literature [32][33]
- `get_knowledge` – to obtain chemical knowledge [33]
- `Rxn_literature` – for sourcing reaction-specific literature

2.5.4 Miscellaneous

This section covers a diverse array of functions including drawing chemical molecular structures and balancing chemical equations. Tools include:

本节涵盖多种功能，包括绘制化学分子结构和平衡化学方程式。工具包括：

- `Moldraw` – for drawing chemical molecular structures
- `calculate` – a general-purpose calculation tool
- `automatic_balance` – to automatically balance chemical equations [34]
- `image_gen` – for generating and searching images [33]

2.5.5 Advanced In-House Analytical Tools

YieldPredict

This is an API tool linked with our self-developed reaction yield prediction tool. By inputting at least two substrates, either in their molecular name or molecular SMILES, this tool can identify the possible reaction types of the molecules by querying our reaction template library. With the known reaction types, the molecules are passed into the reaction models as substrates. The models then suggest products and the most suitable reaction reagents and conditions for the substrates. For example, by asking the reaction yield of triethoxy(naphthalen-1-yl)silane and 5-bromobenzothiazole, the tool first parses the two molecules into the reaction templates as substrates [Figure 3](#). This suggests Hiyama cross-coupling reactions. The two substrates are then parsed into the Hiyama reaction models, generating products and possible reaction yields under specific reaction reagents and conditions.

YieldPredict 是一款 API 工具，与我们自主开发的反应收率预测工具相连。用户可以输入至少两个底物（分子名称或 SMILES 格式），工具会通过查询我们的反应模板库来识别这些分子可能涉及的反应类型。在确定反应类型后，分子会作为底物传递到反应模型中。随后，模型会预测生成的产物，并推荐最合适的反应试剂和条件。例如，若查询三乙氧基(萘-1-基)硅烷与 5-溴苯并噻唑的反应收率，工具会首先将这两个分子解析为反应模板中的底物 [Figure 3](#)，进而识别出 Hiyama 交叉偶联反应。然后，这两个底物会被输入 Hiyama 反应模型中，生成相应的产物，并预测在特定反应试剂和条件下的可能反应收率。

We have dedicated our efforts to developing data-driven reaction yield prediction models for common reaction types [35–38]. For each model of a specific reaction type, we conduct chemical reaction experiments using high-throughput experimentation (HTE) techniques with various substrates. This enables us to draw insights from existing literature data and identify areas where experimental data collection is necessary to augment an equitable data space for refining model training, thus facilitating more robust interpolation. We develop reaction models using machine learning techniques such as support vector machine (SVM) and NLP deep learning models like BERT (Bidirectional Encoder Representations from Transformers) [39]. These models are validated using external literature test data, achieving reasonable Mean Absolute Error (MAE), commonly below 0.15. As of the publication of this work, we have included 18 reaction types in this tool.

我们专注于开发基于数据驱动的反应收率预测模型，涵盖多种常见的反应类型 [35–38]。针对每种特定反应类型的模型，我们使用高通量实验 (HTE) 技术对各种底物进行化学反应实验。这使我们能够从现有文献数据中获得洞见，并识别出需要额外实验数据采集的领域，以扩充数据空间，从而优化模型训练并增强插值能力。我们采用支持向量机 (SVM) 等机器学习技术，以及 NLP 深度学习模型（如 BERT，双向编码器表示转换器）[39] 来开发反应模型。这些模型经过外部文献测试数据的验证，其平均绝对误差 (MAE) 通常低于 0.15。截至本研究发表时，该工具已涵盖 18 种反应类型。

Get_Conditions

This tool is a simplified version of YieldPredict. Instead of predicting the reaction product and yield, it provides rapid responses and suggests only the suitable reaction conditions and reagents for the substrates.

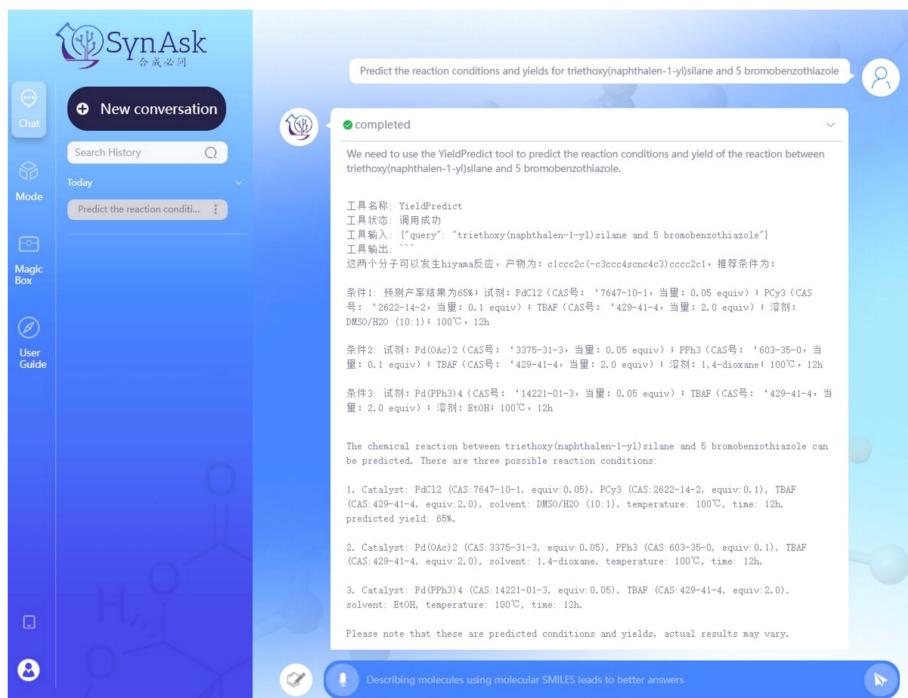
Get_Conditions 是 YieldPredict 的简化版本。它不预测反应产物和收率，而是快速提供适用于底物的最佳反应条件和试剂建议。

Retrosynthesis

By inputting the desired target products, this tool generates numerous reaction pathways of molecules starting from buyable precursors. We have developed our own retrosynthesis model for this purpose. For a desired product, it is parsed into the reaction template library to find possible substrates and, consequently, the suitable reaction site for bond breakage. A reinforcement learning-trained agent selects the most suitable reaction from the candidates based on the forecasted synthesis difficulty and predicted reaction yield of the substrates (desired products at the previous step). This process is conducted recursively until the last substrates are buyable. At the output, we present the results in both textual form and as retrosynthetic route images. The algorithm of our retrosynthesis model will be published elsewhere.

该工具通过输入目标产物，生成从可购得前体分子出发的多个反应路径。我们为此开发了自研的逆合成分析模型。对于指定产物，模型会将其解析到反应模板库中，以寻找可能的底物，并确定适合的反应位点进行化学键断裂。随后，经过强化学习训练的智能体会根据预测的合成难度和底物（前一步骤中的目标产物）的预估反应收率，从候选反应中选择最合适的方法。此过程会递归进行，直到最后的底物为可购得的化学品。最终，工具会以文本形式和逆合成路线图的方式呈现结果。我们的逆合成分析模型的算法将在后续研究中发布。

(a)



(b)

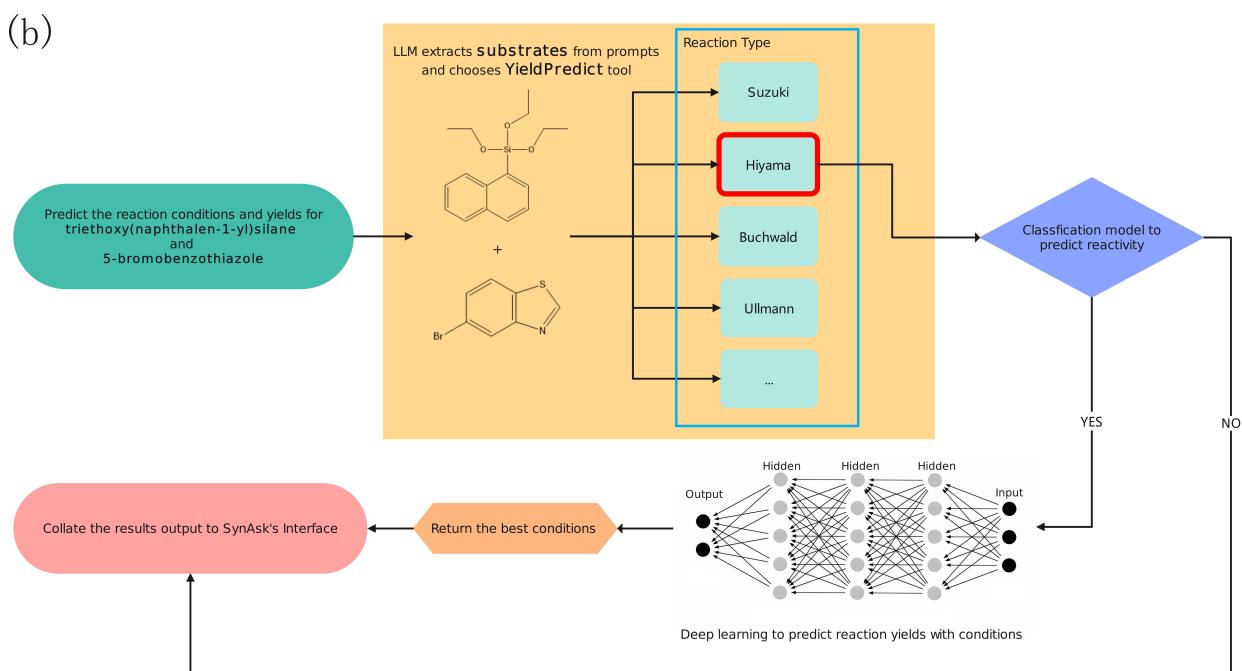


Fig. 3 An example of the YieldPredict tool workflow for predicting the reaction yield of triethoxy(naphthalen-1-yl)silane and 5-bromobenzothiazole: (a) the user interface of SynAsk, (b) the thinking process of the YieldPredict tool.
用于预测三乙氧基(萘-1-基)硅烷和5-溴苯并噻唑的反应产量的 YieldPredict 工具工作流程示例：(a) SynAsk 的用户界面，(b) YieldPredict 工具的思考过程。

3 SynAsk Performance

We evaluate the performance of SynAsk from two perspectives: its general ability as a large language model (LLM), and its proficiency in synthetic chemistry. Additionally, we provide several examples of SynAsk's outputs to demonstrate the platform's comprehension capabilities.

我们从两个方面评估 SynAsk 的性能：其作为大规模语言模型（LLM）的通用能力，以及其在合成化学领域的专业能力。此外，我们提供多个示例，以展示 SynAsk 在理解能力方面的表现。

3.1 General ability of SynAsk

We evaluate the performance enhancements achieved through our first fine-tuning method on the SynAsk model based on OpenCompass [40], which serves as a universal evaluation platform for foundation LLMs. The efficacy of the method is demonstrated by its superior scores across various assessment indicators, particularly in its application to chemistry. The definitions of the general indicators used in [Figure 4](#) are provided in [Section S1](#) of the ESI, while the chemistry-related indicators are outlined in [Section S3](#) along with examples. It's noteworthy that indicators such as College Chemistry, High School Chemistry, and Middle School Chemistry in the figure all stem from C-Eval. SynAsk significantly outperforms its foundation model predecessors. For example, in the area of College Chemistry, SynAsk achieves a remarkable score of 70.83%, compared to 50% by both Qwen-14B-Chat and Qwen1.5-14B-Chat. This signifies a substantial improvement, highlighting the model's enhanced ability to effectively utilize existing data sources for solving complex chemical problems.

我们基于 OpenCompass [40] 评估了 SynAsk 模型在首次微调后取得的性能提升。OpenCompass 是一个通用的基础 LLM 评估平台。我们的微调方法在多个评估指标上均表现出色，尤其是在化学领域的应用能力上取得了显著优势。[Figure 4](#) 中所示的通用指标定义详见电子补充信息 (ESI) 的 [Section S1](#)，而化学相关指标及示例详见 [Section S3](#)。值得注意的是，图中的大学化学 (College Chemistry)、高中化学 (High School Chemistry) 和初中化学 (Middle School Chemistry) 指标均源自 C-Eval。SynAsk 在这些评测中显著超越其基础模型前身。例如，在大学化学领域，SynAsk 取得了 70.83% 的高分，而 Qwen-14B-Chat 和 Qwen1.5-14B-Chat 均仅为 50%。这一显著提升表明，该模型在利用已有数据源解决复杂化学问题方面的能力得到了极大增强。

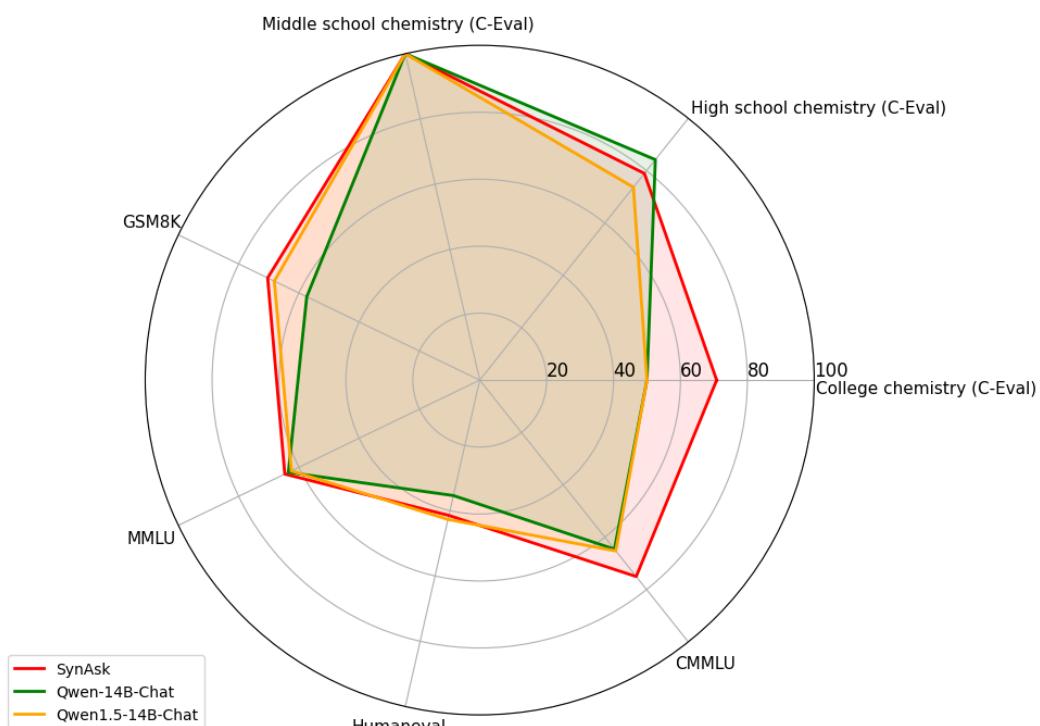


Fig. 4 The comparison of general ability between SynAsk and Qwen in seven aspects, including its applications in chemistry.

Furthermore, the scores in other key benchmarks such as MMLU, GSM8K and CMMLU also reflect the overall enhancement of the SynAsk model. In CMMLU, which assesses cross-model multitask learning, SynAsk scored 75.03%, indicating its proficiency in integrating textual and visual information, crucial for multi-model applications. Similarly, its performance in MMLU and GSM8K benchmarks demonstrate its improved global knowledge comprehension and multi-step mathematical reasoning, respectively.

此外，SynAsk 在 MMLU、GSM8K 和 CMMLU 等关键基准测试中的得分也反映了其整体性能的提升。在评估跨模型多任务学习能力的 CMMLU 测试中，SynAsk 得分为 75.03%，表明其在文本和视觉信息融合方面的能力，这对于多模态应用至关重要。同样，在 MMLU 和 GSM8K 测试中的表现展示了其在全球知识理解和多步数学推理能力方面的改进。

The advancements in SynAsk are attributed to the fine-tuning approach that leverages existing data sources more efficiently, thus enhancing the model's ability to address nuanced chemical contexts and complex reasoning tasks. This is particularly crucial for applications requiring deep understanding and contextual awareness, as indicated by the improvements in C-Eval scores.

SynAsk 的进步归因于微调方法的优化，该方法能够更高效地利用现有数据源，从而增强模型处理复杂化学背景和推理任务的能力。C-Eval 评分的提升进一步验证了这一点，表明模型在需要深入理解和上下文感知的应用场景中的潜力。

These results collectively underscore the effectiveness of our fine-tuning methodology, confirming its potential to significantly boost performance across diverse linguistic and cognitive challenges, thereby reinforcing the model's utility in academic and practical applications.

这些结果共同证明了我们的微调方法的有效性，表明其在提升模型在不同语言和认知挑战方面的表现方面具有巨大潜力，从而增强了该模型在学术研究和实际应用中的价值。

3.2 Proficiency in synthetic chemistry

The primary proficiency of SynAsk in synthetic chemistry lies in its ability to predict reaction performance, such as yield, and to conduct retrosynthetic planning of target molecules, utilizing the embedded tools within SynAsk. Several case studies are presented and compared with benchmarks to evaluate the model's performance. Additionally, the other functions of SynAsk are compared with ChatGPT-4.0 answers to highlight its advancements in various areas.

SynAsk 在合成化学方面的主要优势在于其预测反应性能（如收率）以及进行目标分子的逆合成规划的能力，这些功能依托于 SynAsk 内部集成的工具。我们提供多个案例研究，并与基准模型进行比较，以评估模型的性能。此外，SynAsk 的其他功能也与 ChatGPT-4.0 的答案进行对比，以展示其在多个领域的进步。

3.2.1 Reaction yield prediction

A number of reaction yield prediction models have been developed and widely used to forecast the performance of reactions for frequently encountered reaction classes. For instance, Doyle *et al.*'s palladium-catalysed Buchwald–Hartwig cross-coupling reaction model [41] and Richardson *et al.*'s Suzuki-Miyaura cross-coupling reaction model [42] are among the notable examples. These models were trained using self-developed high-throughput experimentation (HTE) reaction data employing machine learning algorithms. Schwaller *et al.* [43] further enhanced the performance of these models using the same datasets through pre-trained BERT model. While these methods effectively predict the product yield within the self-developed HTE reaction dataset, their applicability to predicting the product yield of external literature recorded reactions may be limited. 多个反应收率预测模型已经被开发并广泛应用于预测常见反应类别的性能。例如，Doyle *et al.* 开发的钯催化 Buchwald–Hartwig 交叉偶联反应模型 [41]，以及 Richardson *et al.* 提出的 Suzuki-Miyaura 交叉偶联反应模型 [42]，均是典型案例。这些模型通过机器学习算法训练，并基于自研的高通量实验 (HTE) 反应数据。Schwaller *et al.* [43] 进一步利用预训练 BERT 模型对这些数据集进行了优化，从而提升了模型的预测性能。然而，这些方法虽然能有效预测自研 HTE 数据集中的产物收率，但在预测外部文献记录的反应收率时，适用性可能受到一定限制。

We tested our in-house nucleophilic aromatic substitution (S_NAr) reaction model embedded in SynAsk with both a test set and external literature reaction data. The model test set comprises unseen HTE reaction data, yielding a mean absolute error (MAE) of 11.5%. For the external literature reaction data, to minimize bias, we randomly collected 60 recently published S_NAr reactions from the last three years (2022–2024), including new substrate molecules never seen by the reaction model. The comparison between the model-predicted yield and literature-reported yield is presented in Figure 5b, yielding an MAE of 14.1%. These recently published reactions encompass seven different reaction conditions. For example, N-methyl-1-phenylmethanamine reacting with 2-fluoro-5-methoxybenzaldehyde under K_2CO_3 and DMF is illustrated in Figure 5c. The literature-reported yield of the product 2-(benzyl(methyl)amino)-5-methoxybenzaldehyde is 75% [44], whilst our model predicts 80% and our HTE experimental yield is 70%. Additional results are provided in Section S4 of the ESI. 我们在 SynAsk 中嵌入的亲核芳香取代 (S_NAr) 反应模型进行了测试，测试对象包括模型测试集和外部文献中的反应数据。测试集包含未曾见过的 HTE 反应数据，其平均绝对误差 (MAE) 为 11.5%。对于外部文献反应数据，为减少偏差，我们随机收集了近三年 (2022–2024) 发表的 60 个 S_NAr 反应，其中包括模型未曾见过的新底物分子。模型预测的收率与文献报道的收率的对比结果如Figure 5b 所示，MAE 为 14.1%。这些近期发表的反应涵盖七种不同的反应条件。例如，N-甲基-1-苯基甲胺与 2-氟-5-甲氧基苯甲醛在 K_2CO_3 和 DMF 条件下反应的案例如Figure 5c 所示。该产物 2-(苄基(甲基)氨基)-5-甲氧基苯甲醛的文献报道收率为 75%[44]，而模型预测收率为 80%，我们的 HTE 实验收率为 70%。更多结果详见 ESI 的Section S4。

The decay in prediction accuracy observed when transitioning from HTE reactions to literature-reported reactions is primarily attributed to the increased complexity of substrates in literature reactions. These substrates are often more intricate and unseen by the model, thereby encompassing a wider range within the chemical space, as depicted in [Figure 5a](#). To compute the chemical space, we digitized the reactions using RXNFP pretrained reaction fingerprint [45] and reduced into two dimensions. [Figure 5a](#) also weakly show three clusters of the S_NAr reaction. Despite the challenges posed by the complexity of literature-reported reactions, our in-house S_NAr model demonstrates the capability to predict the reaction performance of these reactions. This is particularly valuable as it enables predictions closer to the reactions of interest to synthetic chemists. 当从 HTE 反应数据迁移到文献报道的反应数据时，预测精度下降的主要原因在于文献反应的底物复杂性更高。这些底物往往更加复杂，并且模型未曾见过，从而扩大了化学空间的覆盖范围，如[Figure 5a](#) 所示。为了计算化学空间，我们使用 RXNFP 预训练反应指纹 [45] 对反应数据进行数字化，并将其降维至二维。[Figure 5a](#) 也显示出 S_NAr 反应可能存在的三个聚类。尽管文献反应的复杂性带来了挑战，我们的 S_NAr 模型仍表现出对这些反应的预测能力。这对于合成化学家而言尤为重要，因为它能够提供更接近实际研究兴趣的预测结果。

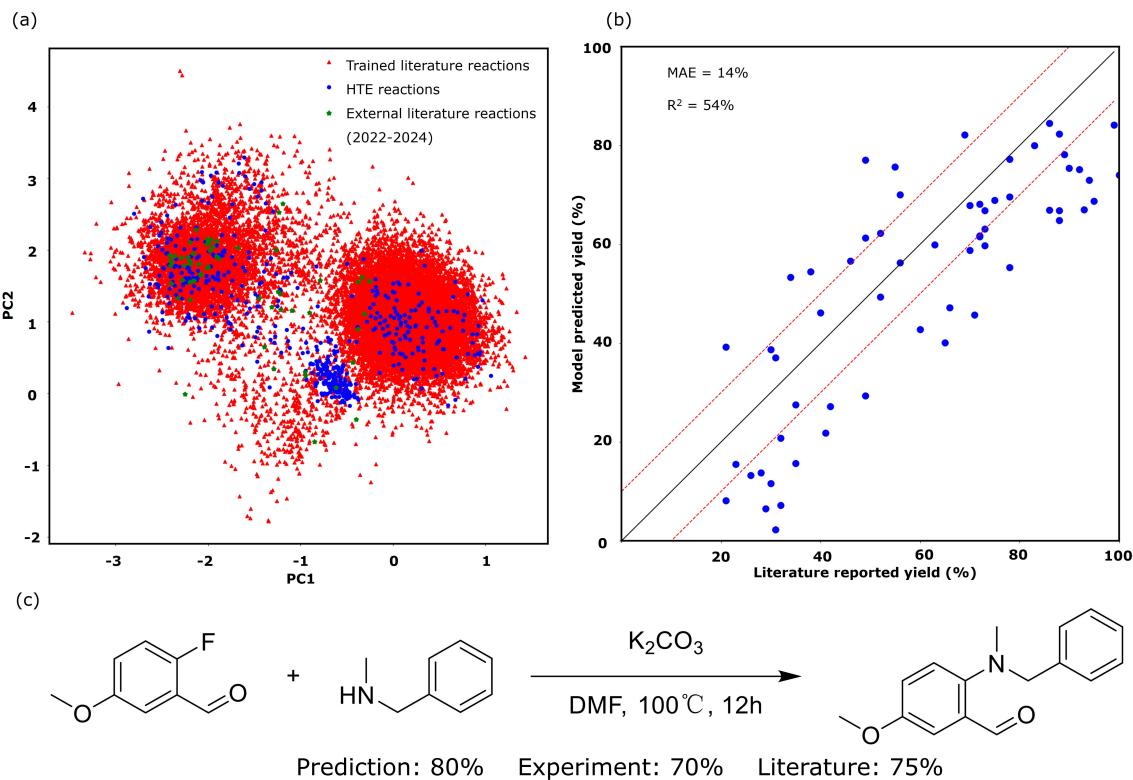


Fig. 5 The S_NAr reaction model results: (a) the chemical space of S_NAr reactions under the HTE and literature recorded datasets, (b) the predicted yield versus experimental yield of the test dataset from the three different models, and (c) an example of S_NAr reaction: N-methyl-1-phenylmethanamine reacting with 2-fluoro-5-methoxybenzaldehyde.

3.2.2 Retrosynthetic route planning

We tasked SynAsk with planning retrosynthetic routes for 11,549 small molecule drugs recorded in the ChEMBL database [46]. SynAsk successfully predicted retrosynthetic routes for 6,358 molecules, suggesting step-by-step routes starting from buyable precursors. This accounts for 55% of the queried molecules. In contrast, the State of the Art (SOTA) open-sourced retrosynthetic planning tool, AIzynthFinder [47], only suggested 3,118 retrosynthetic routes, covering 27% of the queried molecules.

我们要求 SynAsk 为 ChEMBL 数据库 [46] 中记录的 11,549 种小分子药物规划逆合成路线。SynAsk 成功预测了 6,358 种分子的逆合成路径，并提供了从可购得前体出发的逐步合成路线，占查询分子的 55%。相比之下，当前最先进的（SOTA）开源逆合成规划工具 AIzynthFinder [47] 仅提供了 3,118 条逆合成路线，覆盖率仅为 27%。

As a case study, let's consider the retrosynthesis of Gilmelisib, a novel small molecule under investigation as a selective inhibitor of PIK3C α , potentially treating cancers characterized by PIK3C α mutations. SynAsk proposes a seven-step synthetic route with four precursors (as shown in Figure 6a), matching the route suggested by an experienced human chemist in terms of length and number of precursors (as shown in Figure 6b). SynAsk utilizes inexpensive precursors to rapidly obtain key heterocyclic fused ring intermediates through straightforward Knoevenagel condensation and addition-elimination reactions. In contrast, AIZynthFinder fails to provide a synthesis route for the target molecule, even after enriching its starting materials with our lists of buyable precursors. Additional synthetic routes for small molecule drugs are detailed in Section S5 of the ESI. 以 Gilmelisib 的逆合成成为案例分析，该化合物是一种新型小分子，正在被研究为 PIK3C α 的选择性抑制剂，有望用于治疗 PIK3C α 突变相关的癌症。SynAsk 提出了一条七步合成路线，包含四种前体（见Figure 6a），其步骤长度和前体数量与经验丰富的化学家提出的路线相当（见Figure 6b）。SynAsk 采用廉价前体，通过简单的 Knoevenagel 缩合和加成-消除反应迅速获得关键的杂环稠合中间体。相比之下，AIZynthFinder 即使在补充了我们的可购得前体清单后，仍未能为目标分子提供合成路线。关于其他小分子药物的合成路线，更多详情见 ESI 的Section S5。

While we refrain from concluding that SynAsk is smarter or approaching the intelligence of a human chemist in retrosynthesis, as this determination would necessitate passing the Turing test [48, 49] or experimental validation, we acknowledge that SynAsk's retrosynthetic ability offers valuable insights for synthetic chemists and assists in synthesis planning.

尽管我们不认为 SynAsk 在逆合成方面比人类化学家更聪明，或已接近人类化学家的智能水平，因为这一判断需要通过图灵测试 [48, 49] 或实验验证，但我们承认，SynAsk 的逆合成能力为合成化学家提供了有价值的见解，并在合成规划中发挥了重要的辅助作用。

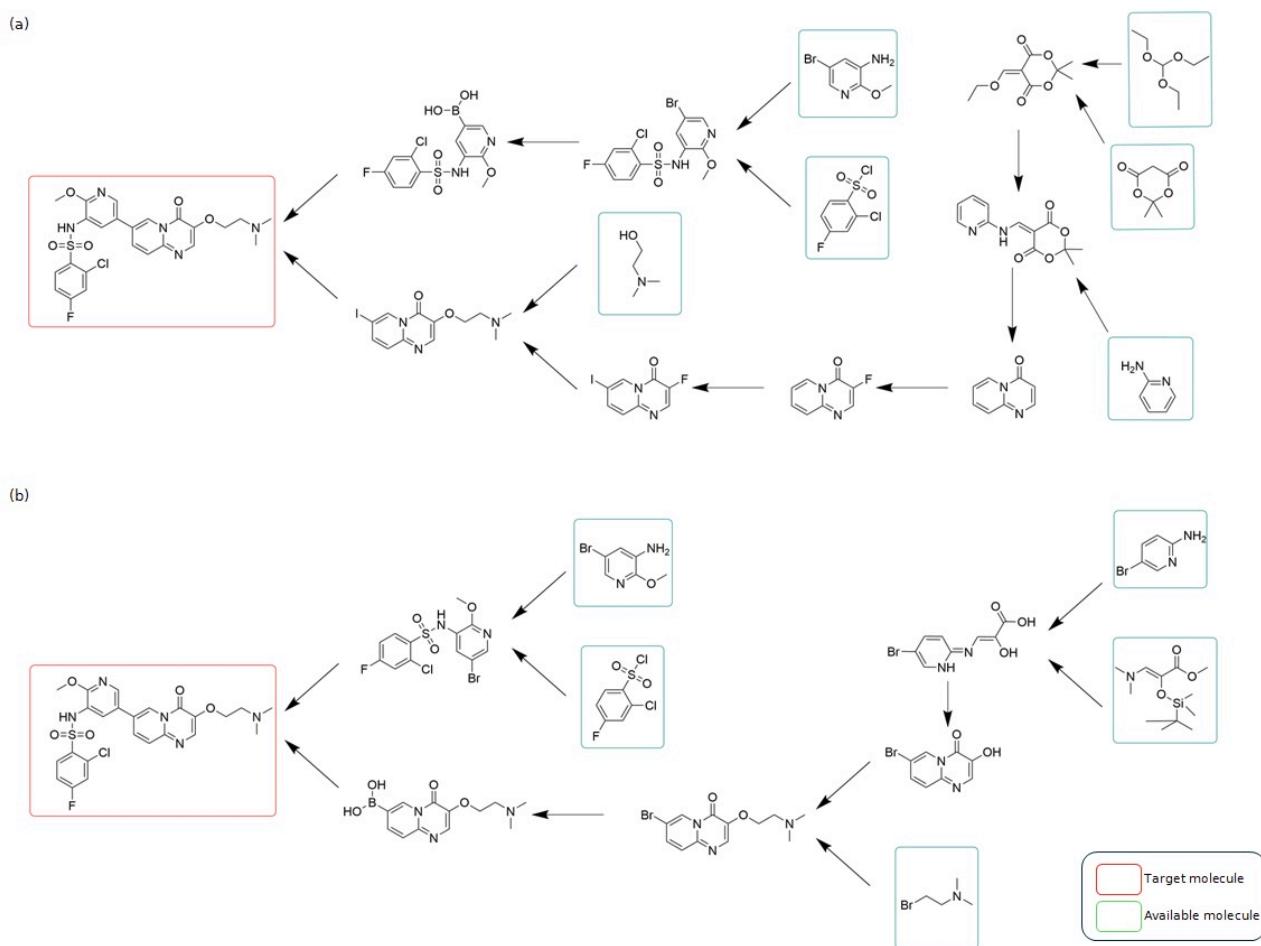


Fig. 6 The comparison among synthetic routes of the target molecule Gilmelisib: planned by (a) SynAsk's retrosynthetic tool and (b) an experienced synthetic chemist.

目标分子 Gilmelisib 的合成路线比较：由 (a) SynAsk 的逆合成工具和 (b) 经验丰富的合成化学家规划。

3.3 Examples of the SynAsk platform outputs versus other LLMs

Here we present a comparative analysis of the performance of three LLMs –SynAsk, ChatGPT-4.0, and ChemCrow –in addressing synthetic chemistry queries. We evaluated their capabilities by inputting a set of synthetic questions, encompassing both general and professional inquiries, to assess their aptitude in providing accurate and relevant responses.

在本节中，我们对比分析了 SynAsk、ChatGPT-4.0 和 ChemCrow 这三种 LLM 在应对合成化学问题时的表现。我们通过输入一系列合成相关的问题（涵盖通用和专业查询）来评估它们提供准确、相关回答的能力。

3.3.1 General inquiries

Queries such as “Can you recommend me some reaction conditions for Suzuki cross-coupling?” or “Please help me find some literature related to C-H activation” were presented to all three LLMs. Across the board, each model exhibited proficiency in generating appropriate responses, showcasing their utility in aiding chemists with routine inquiries, details in [Section S6](#) of the ESI.

例如，我们向三种 LLM 提出了一些通用的化学问题，如“你能推荐一些 Suzuki 交叉偶联反应的反应条件吗？”或“请帮我查找一些与 C-H 活化相关的文献”。所有模型均能够生成合理的回答，表明它们在处理常规化学问题方面具备实用性，具体示例详见 ESI 的[Section S6](#)。

3.3.2 Professional synthetic questions

A more rigorous evaluation was conducted by inputting a specific synthetic question: “Tell me what reaction can occur between Nc1ccc2nccnc2c1.O=C(O)Cc1cc(F)cc(F)c1 and what the product is.” Here “Nc1ccc2nccnc2c1.O=C(O)Cc1cc(F)cc(F)c1” represents the SMILES syntax for quinoxalin-6-amine and 3,5-Difluorophenylacetic acid as substrates. The deliberate use of SMILES allows us to assess the LLMs’ ability to recognize molecules from SMILES.

为了进行更严格的评估，我们输入了一个具体的合成问题：“请告诉我 Nc1ccc2nccnc2c1.O=C(O)Cc1cc(F)cc(F)c1 之间可能发生的反应及其产物。”其中，“Nc1ccc2nccnc2c1.O=C(O)Cc1cc(F)cc(F)c1”采用 SMILES 表示法，分别对应喹喔啉-6-胺和 3,5-二氟苯乙酸。我们特意使用 SMILES 表示，以评估 LLM 在识别分子结构方面的能力。

As shown in [Figure 7](#), SynAsk demonstrates its specialization in organic chemistry by providing a comprehensive list of potential reactions and their corresponding products. Leveraging its domain-specific knowledge, SynAsk offers a diverse array of feasible transformations, including N-acylation, Buchwald-Hartwig amination, Minisci reaction, among others. This exhaustive output underscores SynAsk’s capacity to analyze complex molecular interactions and propose multiple viable pathways.

如[Figure 7](#)所示，SynAsk 展现了其在有机化学领域的专业性，它能够提供详细的可能反应类型及对应产物。利用其领域专长，SynAsk 提出了多种可行的转化路径，包括 N-酰化、Buchwald-Hartwig 胺化、Minisci 反应等。这一详尽的输出体现了 SynAsk 分析复杂分子相互作用并提出多种可行合成路线的能力。

In contrast, ChemCrow delivers a singular response, identifying the reaction as N-acylation and providing the corresponding product. While ChemCrow offers a concise solution, its limitation in providing alternative reaction pathways restricts its utility in scenarios where multiple transformation possibilities exist.

相比之下，ChemCrow 仅提供了单一的解答，将该反应归类为 N-酰化，并给出了相应的产物。虽然 ChemCrow 的回答简明扼要，但其缺乏多种可能转化路径的能力，在涉及多种可能反应的情境下，其实用性受到限制。

ChatGPT-4, although proficient in understanding the query, encounters a misinterpretation in identifying the compounds involved. While it accurately delineates the structure and classification of the provided molecules, it erroneously labels Nc1ccc2nccnc2c1 as nicotinic acid derivative, instead of recognizing it as quinoxalin-6-amine. This discrepancy underscores the model’s susceptibility to misinterpretation of chemical structures, particularly in complex contexts.

ChatGPT-4 尽管能够正确理解查询内容，但在识别涉及的化合物时出现了误判。它虽然能够正确解析提供的分子结构和分类，但却错误地将 Nc1ccc2nccnc2c1 标记为烟酸衍生物，而未能识别其实际为喹喔啉-6-胺。这一错误凸显了 ChatGPT-4 在处理复杂化学结构时的易错性，尤其是在涉及高度专业的化学情境下。

SynAsk distinguishes itself as a specialized LLM tailored specifically for organic chemistry tasks. Its domain-specific training and integration of fine-tuning techniques result in a robust model capable of providing detailed insights and accurate predictions for complex synthetic queries. While ChatGPT-4 and ChemCrow offer general language processing capabilities, they lack the nuanced understanding and domain expertise exhibited by SynAsk in the context of organic chemistry applications. Therefore, for researchers seeking nuanced insights and comprehensive analyses in organic synthesis, SynAsk stands as a valuable tool for augmenting chemical exploration and discovery.

SynAsk 在有机化学任务中表现出色，展现出其作为专用 LLM 的独特优势。其领域专用训练以及微调技术的整合，使其成为一个强大的模型，能够针对复杂的合成问题提供详细的见解和准确的预测。相比之下，ChatGPT-4 和 ChemCrow 虽然具备通用的自然语言处理能力，但在有机化学应用中缺乏对专业知识的深入理解。因此，对于希望在有机合成研究中获得深入见解和全面分析的研究人员而言，SynAsk 是一个极具价值的工具，可助力化学探索与发现。

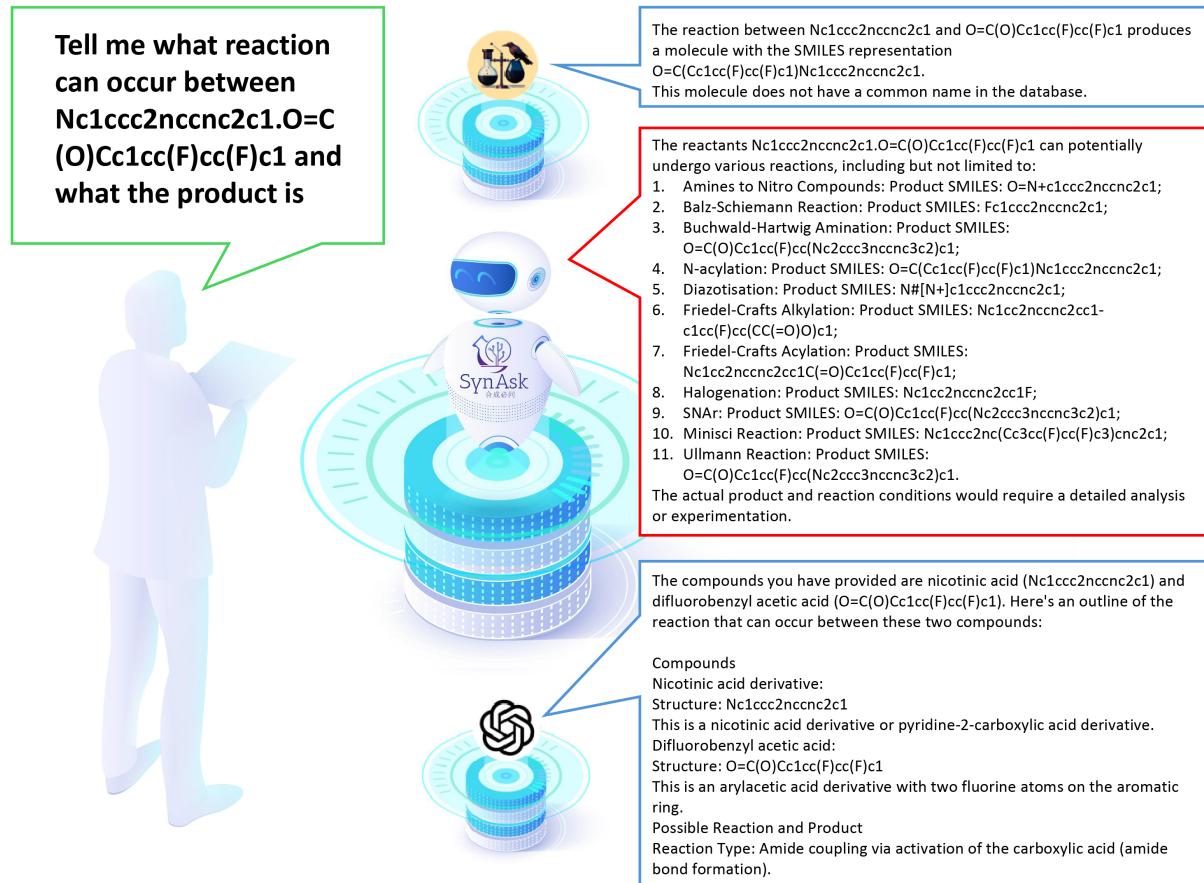


Fig. 7 The comparison of SynAsk, ChatGPT-4, and ChemCrow output on a professional synthetic question.

4 Conclusions and Future Works

In this work, we have developed SynAsk, a specialized LLM-powered platform for synthetic chemistry. It represents the first publicly accessible chemistry domain-specific LLM, fine-tuned with selected chemistry data and connected with both in-house and external chemoinformatic tools. Through comparative analyses with foundation LLMs, we have demonstrated SynAsk’s proficiency and specialization in synthetic chemistry. Results obtained in reaction yield prediction and retrosynthesis further validate SynAsk’s capability in providing valuable chemical insights to synthetic chemists across multiple domains.

在本研究中，我们开发了 SynAsk，一个专为合成化学设计的 LLM 驱动平台。SynAsk 是首个公开可用的化学领域专用 LLM，经过精选的化学数据微调，并连接了内部和外部的化学信息学工具。通过与基础 LLM 的对比分析，我们展示了 SynAsk 在合成化学领域的专业性和高效性。其在反应收率预测和逆合成规划方面的表现进一步验证了 SynAsk 在多个化学领域为合成化学家提供有价值化学见解的能力。

Looking ahead, our future endeavors aim to enhance the functionality of SynAsk by empowering the language model and fine-tuning it with additional data for more seamless and appropriate responses. Additionally, we envision SynAsk playing a pivotal role in driving autonomous reaction laboratories [50]. Traditionally, reaction robots have been constrained by written scripts to define their scopes. Recent research has showcased the potential of LLMs to drive robotic chemists effectively [51]. Leveraging SynAsk’s capabilities such as retrosynthesis, inference, and programming script writing, we foresee it being instrumental in driving autonomous laboratories, representing the next phase of our fusion of LLM and hardware research.

展望未来，我们计划通过增强语言模型的能力并利用额外数据对其进行微调，以提升 SynAsk 的功能性，使其能够提供更加流畅和精准的响应。此外，我们还设想 SynAsk 将在推动自动化反应实验室方面发挥关键作用 [50]。传统上，反应机器人依赖于预定义的书面脚本来限定其操作范围。然而，最新研究表明，LLM 在驱动机器人化学家方面具有巨大潜力 [51]。借助 SynAsk 在逆合成、推理和编写程序脚本等方面的能力，我们预见其将在自主实验室的构建中发挥核心作用，标志着 LLM 与硬件研究融合的下一个发展阶段。

5 Acknowledgements

We are grateful for financial support from Guangzhou National Laboratory, and the National Natural Science Foundation of China (22071249, 22393892).

6 Conflict of Interest

We have a patent application in China with the application number 202410714040.6 titled "A Human-Computer Interaction Method and Electronic Device Based on a Large Language Model".

References

- [1] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
- [2] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
- [3] OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decaireaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Kondracik, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Avila Belbute Peres, F., Petrov, M., Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Toootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (2023)
- [4] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang, S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen Technical Report (2023)
- [5] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: LLaMA: Open and Efficient Foundation Language Models (2023)

- [6] Yue, S., Chen, W., Wang, S., Li, B., Shen, C., Liu, S., Zhou, Y., Xiao, Y., Yun, S., Huang, X., Wei, Z.: DISC-LawLLM: Fine-tuning Large Language Models for Intelligent Legal Services (2023)
- [7] Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., *et al.*: Large language models encode clinical knowledge. *Nature* **620**(7972), 172–180 (2023)
- [8] Weininger, D., Weininger, A., Weininger, J.L.: Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Computer Sciences* **29**(2), 97–101 (1989) <https://doi.org/10.1021/ci00062a008> <https://doi.org/10.1021/ci00062a008>
- [9] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C.A., Bekas, C., Lee, A.A.: Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science* **5**(9), 1572–1583 (2019) <https://doi.org/10.1021/acscentsci.9b00576> <https://doi.org/10.1021/acscentsci.9b00576>. PMID: 31572784
- [10] Weber, J.M., Guo, Z., Zhang, C., Schweidtmann, A.M., Lapkin, A.A.: Chemical data intelligence for sustainable chemistry. *Chem. Soc. Rev.* **50**, 12013–12036 (2021) <https://doi.org/10.1039/D1CS00477H>
- [11] Guo, T., Guo, K., Nan, B., Liang, Z., Guo, Z., Chawla, N.V., Wiest, O., Zhang, X.: What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks (2023)
- [12] M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P.: Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 1–11 (2024)
- [13] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., *et al.*: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
- [14] Topsakal, O., Akinci, T.C.: Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In: International Conference on Applied Engineering and Natural Sciences, vol. 1, pp. 1050–1056 (2023)
- [15] Zhang, D., Liu, W., Tan, Q., Chen, J., Yan, H., Yan, Y., Li, J., Huang, W., Yue, X., Zhou, D., Zhang, S., Su, M., Zhong, H., Li, Y., Ouyang, W.: ChemLLM: A Chemical Large Language Model (2024)
- [16] chatchat-space: Langchain-Chatchat. <https://github.com/chatchat-space/Langchain-Chatchat>
- [17] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
- [18] Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Fu, Y., et al.: C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems* **36** (2024)
- [19] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
- [20] Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H.W., Chowdhery, A., Le, Q.V., Chi, E.H., Zhou, D., Wei, J.: Challenging big-bench tasks and whether chain-of-thought can solve them. arXiv preprint arXiv:2210.09261 (2022)
- [21] Li, H., Zhang, Y., Koto, F., Yang, Y., Zhao, H., Gong, Y., Duan, N., Baldwin, T.: CMMLU: Measuring massive multitask language understanding in Chinese (2023)
- [22] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
- [23] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: Glm: General language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 320–335 (2022)
- [24] Team, I.: Internlm: A multilingual language model with progressively enhanced capabilities (2023)

- [25] Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023)
- [26] AI, ., ;, Young, A., Chen, B., Li, C., Huang, C., Zhang, G., Zhang, G., Li, H., Zhu, J., Chen, J., Chang, J., Yu, K., Liu, P., Liu, Q., Yue, S., Yang, S., Yang, S., Yu, T., Xie, W., Huang, W., Hu, X., Ren, X., Niu, X., Nie, P., Xu, Y., Liu, Y., Wang, Y., Cai, Y., Gu, Z., Liu, Z., Dai, Z.: Yi: Open Foundation Models by 01.AI (2024)
- [27] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al.: Pubchem 2023 update. Nucleic acids research **51**(D1), 1373–1380 (2023)
- [28] Pence, H.E., Williams, A.: ChemSpider: an online chemical information resource. ACS Publications (2010)
- [29] Chemistry team, I.R.: rxn4chemistry: Python wrapper for the IBM RXN for Chemistry API. <https://github.com/rxn4chemistry/rxn4chemistry> (2023)
- [30] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., Laino, T.: Unsupervised attention-guided atom-mapping (2020)
- [31] Chen, S., An, S., Babazade, R., Jung, Y.: Precise atom-to-atom mapping for organic reactions via human-in-the-loop machine learning. Nature Communications **15**(1), 2250 (2024)
- [32] Ginsparg, P.: Arxiv at 20. Nature **476**(7359), 145–147 (2011)
- [33] SerpAPI: SerpAPI - Google Search Results API. <https://serpapi.com/> (2023)
- [34] Dahlgren, B.: Chempy: A package useful for chemistry written in python. Journal of Open Source Software **3**(24), 565 (2018)
- [35] Xu, Y., Gao, Y., Su, L., Wu, H., Tian, H., Zeng, M., Xu, C., Zhu, X., Liao, K.: High-throughput experimentation and machine learning-assisted optimization of iridium-catalyzed cross-dimerization of sulfoxonium ylides. Angewandte Chemie International Edition **62**(48), 202313638 (2023) <https://doi.org/10.1002/anie.202313638> <https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202313638>
- [36] Qiu, J., Xu, Y., Su, S., Gao, Y., Yu, P., Ruan, Z., Liao, K.: Auto machine learning assisted preparation of carboxylic acid by tempo-catalyzed primary alcohol oxidation. Chinese Journal of Chemistry **41**(2), 143–150 (2023) <https://doi.org/10.1002/cjoc.202200555> <https://onlinelibrary.wiley.com/doi/pdf/10.1002/cjoc.202200555>
- [37] Xu, Y., Ren, F., Su, L., Xiong, Z., Zhu, X., Lin, X., Qiao, N., Tian, H., Tian, C., Liao, K.: Hte and machine learning-assisted development of iridium (i)-catalyzed selective o-h bond insertion reactions toward carboxymethyl ketones. Organic Chemistry Frontiers **10**(5), 1153–1159 (2023)
- [38] Yu, Z., Kong, Y., Li, B., Su, S., Rao, J., Gao, Y., Tu, T., Chen, H., Liao, K.: Hte-and ai-assisted development of dhp-catalyzed decarboxylative selenation. Chemical Communications **59**(20), 2935–2938 (2023)
- [39] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018) [1810.04805](https://arxiv.org/abs/1810.04805)
- [40] Contributors, O.: OpenCompass: A Universal Evaluation Platform for Foundation Models. <https://github.com/open-compass/opencompass> (2023)
- [41] Ahneman, D.T., Estrada, J.G., Lin, S., Dreher, S.D., Doyle, A.G.: Predicting reaction performance in c–n cross-coupling using machine learning. Science **360**(6385), 186–190 (2018)
- [42] Perera, D., Tucker, J.W., Brahmbhatt, S., Helal, C.J., Chong, A., Farrell, W., Richardson, P., Sach, N.W.: A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science **359**(6374), 429–434 (2018)
- [43] Schwaller, P., Vaucher, A.C., Laino, T., Reymond, J.-L.: Prediction of chemical reaction yields using deep learning. Machine learning: science and technology **2**(1), 015016 (2021)
- [44] Zaitseva, E.R., Smirnov, A.Y., Myasnyanko, I.N., Mineev, K.S., Sokolov, A.I., Volkhina, T.N., Mikhaylov,

- A.A., Baleeva, N.S., Baranov, M.S.: Imidazol-5-ones as a substrate for [1, 5]-hydride shift triggered cyclization. *New Journal of Chemistry* **45**(4), 1805–1808 (2021)
- [45] Schwaller, P., Probst, D., Vaucher, A.C., Nair, V.H., Kreutter, D., Laino, T., Reymond, J.-L.: Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **3**(2), 144–152 (2021)
- [46] European Bioinformatics Institute: ChEMBL Drug Database. <https://www.ebi.ac.uk/chembl/g/#browse/drugs>. Accessed: 2024-04-22 (2024)
- [47] Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O., Bjerrum, E.: Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics* **12**(1), 70 (2020)
- [48] Mikulak-Klucznik, B., Gołębiowska, P., Bayly, A.A., Popik, O., Klucznik, T., Szymkuć, S., Gajewska, E.P., Dittwald, P., Staszewska-Krajewska, O., Beker, W., *et al.*: Computational planning of the synthesis of complex natural products. *Nature* **588**(7836), 83–88 (2020)
- [49] Segler, M.H., Preuss, M., Waller, M.P.: Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* **555**(7698), 604–610 (2018)
- [50] Burger, B., Maffettone, P.M., Gusev, V.V., Aitchison, C.M., Bai, Y., Wang, X., Li, X., Alston, B.M., Li, B., Clowes, R., *et al.*: A mobile robotic chemist. *Nature* **583**(7815), 237–241 (2020)
- [51] Boiko, D.A., MacKnight, R., Kline, B., Gomes, G.: Autonomous chemical research with large language models. *Nature* **624**(7992), 570–578 (2023)

SynAsk: Unleashing the Power of Large Language Models in Organic Synthesis

Electronic Supplementary Information

Chonghuan Zhang^{1†}, Qianghua Lin^{1†}, Biwei Zhu^{2†}, Haopeng Yang², Xiao Lian², Hao Deng², Jiajun Zheng²,
Kuangbiao Liao^{1*}

¹Guangzhou National Laboratory, Guangdong, PR China, 510005

²AIChemEco Inc., Guangdong, PR China, 510005

* Corresponding author(s). E-mail(s): kuangbiao_liao@gzlab.ac.cn

†These authors contributed equally to this work.

S1 The indicators used to assess LLMs

We evaluated the LLM's capabilities using various metrics including Massive Multi-task Language Understanding (MMLU), Multi-level multi-discipline Chinese evaluation (C-Eval), GSM8K, BIG-Bench-Hard (BBH), and Measuring massive multitask language understanding in Chinese (CMMLU). These metrics collectively provide a thorough assessment of a model's proficiency, encompassing linguistic understanding, mathematical reasoning, contextual comprehension, multi-modal integration, and the application of CoT, which examines the fluency of LLMs' integration with external tools. This evaluation framework emphasizes the diverse and essential skills a model needs to effectively tackle complex real-world problems.

我们使用多个指标来评估 LLM 的能力，包括大规模多任务语言理解 (MMLU)、多层次多学科中文评测 (C-Eval)、GSM8K、大型基准测试 BBH (BIG-Bench-Hard)、以及大规模中文多任务语言理解 (CMMLU)。这些指标共同构成了对模型能力的全面评估，涵盖语言理解、数学推理、上下文理解、多模态整合以及 CoT (思维链) 应用，以检验 LLM 在整合外部工具时的流畅性。该评估框架强调了模型在应对复杂现实问题时所需的多样化核心技能。

- **Massive Multi-task Language Understanding, MMLU** represents a comprehensive and multi-faceted initiative that aims to evaluate and enhance the performance of language models across a broad range of linguistic challenges, providing an extensive evaluation of global knowledge and problem-solving abilities.

大规模多任务语言理解 (MMLU) 是一个全面且多维的评测体系，旨在评估和提升语言模型在广泛语言挑战中的表现，对全球知识覆盖度和问题解决能力进行深度测评。

- **Multi-level Multi-discipline Chinese Evaluation, C-Eval** tests models in scenarios that necessitate an understanding of subtle context, which is crucial for applications involving natural language understanding and generation.

多层次多学科中文评测 (C-Eval) 旨在测试模型在需要细微上下文理解的场景中的表现，这对于涉及自然语言理解和生成的应用至关重要。

- **Grade School Math 8K, GSM8K** is a widely recognized test set designed to assess the mathematical capabilities of language models. It comprises problems that require 2-8 steps of basic mathematical operations to test the models' multi-step mathematical reasoning.

小学数学 8K (GSM8K) 是一个广受认可的测试集，专门用于评估语言模型的数学能力。该测试集包含需要 2-8 步基本数学运算的问题，以测试模型的多步数学推理能力。

- **BIG-Bench-Hard, BBH** evaluates language models' capabilities in applying Chain of Thought to humanistic knowledge. It measures how effectively a model can navigate through complex humanistic concepts and ideas, emphasizing its ability to perform sequential reasoning that mirrors human-like understanding in tasks with cultural and historical depth.

大型基准测试 BBH (BIG-Bench-Hard) 评估语言模型在思维链 (CoT) 方法下对人文知识的应用能力。该测试衡量模型在处理复杂人文概念和思想方面的能力，强调其在具有文化和历史深度的任务中进行类人推理的能力。

- **Measuring massive multitask language understanding in Chinese, CMMLU** is a comprehensive Chinese evaluation benchmark specifically used to evaluate the knowledge and reasoning capabilities of language models in the Chinese context. CMMLU covers 67 topics from basic subjects to advanced professional levels.

大规模中文多任务语言理解 (CMMLU) 是一个专门用于评估语言模型在中文环境下的知识和推理能力的综合评测基准。CMMLU 涵盖 67 个主题，涉及基础学科到高级专业水平的内容。

Model	MMLU (5-shot)	C-Eval (5-shot)	GSM8K (8-shot)	BBH (3-shot)	CMMLU
LLaMA2-7B	46.8	32.5	16.7	38.2	31.8
LLaMA2-13B	55.0	41.4	29.6	45.6	38.4
LLaMA2-32B	62.6	-	42.2	44.1	-
ChatGLM2-6B	47.9	51.7	32.4	33.7	-
InterLM-7B	51.0	53.4	31.2	37.0	51.8
InterLM-20B	62.1	58.8	52.6	52.5	59.0
Baichuan2-7B	54.7	56.3	24.6	41.6	57.1
Baichuan2-13B	59.5	59.0	52.8	49.0	62.0
Yi-34B	76.3	81.8	67.9	66.4	85.6
Qwen-1.8B	45.3	56.1	32.3	22.3	52.1
Qwen-7B	58.2	63.5	51.7	45.0	62.2
Qwen-14B	66.3	72.1	61.3	53.4	71.0
Qwen-72B	77.4	83.3	78.9	67.7	83.6

Table S1 Performance of Different Models on Various Benchmarks

S2 Fine-tuning techniques and procedures

In our experiments, we explored two distinct fine-tuning methodologies for LLMs. The first approach involved techniques such as quantization to enable the operation of a 14-billion-parameter model within a 24GB GPU environment. The second approach was direct fine-tuning without additional quantization techniques.
在我们的实验中，我们探讨了两种不同的 LLM 微调方法。第一种方法采用量化技术，使得 14 亿参数规模的模型能够在 24GB GPU 环境下运行。第二种方法则是直接进行微调，不使用额外的量化技术。

For our experiments, we selected a model with 14 billion parameters. We applied Low-Rank Adaptation (LoRA) by incorporating low-rank matrices into the fully connected layers. The parameter details are presented in Table S1.

在实验过程中，我们选择了一款具有 14 亿参数的模型，并采用低秩适配（LoRA）技术，通过在全连接层中引入低秩矩阵进行微调。模型的参数详情见表 S1。

Total Parameters	Trainable Parameters	Percentage of Total
14,209,134,120	41,843,040	$\approx 0.294\%$

Table S1 Parameter quantity of the 14-billion-parameter model

The fine-tuning with quantization process, conducted on a dataset of 200 entries with a batch size of 2, was completed within an hour. This method is a viable solution for managing large model training on hardware with limited memory without significantly compromising precision.

对包含 200 个条目、批大小为 2 的数据集进行的量化微调过程在一小时内完成。此方法是在内存有限的硬件上管理大型模型训练的可行解决方案，同时不会显著影响精度。

	Before Quantization	After 4-bit Quantization
During loading	$14 \times 10^9 \times 4$ bytes	$14 \times 10^9 \times 0.5$ bytes
During computation		$14 \times 10^9 \times 2$ bytes
Memory Consumption	≈ 56 GB	≈ 7 GB during loading

Table S2 Memory usage before and after quantization

Leveraging a single GeForce RTX 4090 with 24GB of VRAM for fine-tuning a 14-billion-parameter model, we initially applied quantization to reduce the memory usage and accelerate inference, though at the potential cost of precision loss. During loading, the model was quantized to 4-bit precision and subsequently converted to 16-bit for computations. Post-loading, neither the original nor the quantized weights were retained in memory. 在微调 14 亿参数模型时，我们利用单张 GeForce RTX 4090 显卡（24GB VRAM），并首先应用量化技术以减少内存占用并加速推理，尽管这可能会带来一定的精度损失。在加载过程中，模型被量化至 4-bit 精度，并随后转换为 16-bit 以进行计算。加载完成后，原始权重和量化权重均不会被保留在内存中。

The fine-tuning without quantization approach utilized LoRA under the deepspeed's ZeRO-3 optimization. We employed three GeForce RTX 4090 GPUs, each with 24GB of memory, which allowed the fine-tuning of the model on a dataset of over 4,000 entries. The process took approximately seven hours to complete.

在不进行量化的微调方法中，我们采用了 DeepSpeed 的 ZeRO-3 优化策略，并使用了三张 GeForce RTX 4090 显卡（每张 24GB 内存）。这一配置使我们能够在包含 4,000 多条数据的训练集上对模型进行微调，整个过程大约耗时七小时。

Both fine-tuning methodologies proved to be effective, demonstrating the practical applicability of our approaches to large-scale model optimization.

两种微调方法均被证明是有效的，表明我们的优化策略在大规模模型优化中具有良好的实用性。

S3 Chemistry related indicators and examples

We assessed the chemistry ability of the LLMs using the chemistry test questions from C-Eval, which comprises multiple discipline questions in multiple levels in Chinese ([Section S1](#)). This test was completed in Chinese since SynAsk's original language is Chinese. However, we acknowledge that with the LLMs' powerful language ability, testing of the LLMs with different major languages in the world would reach close results.

我们使用 C-Eval 中的化学测试题来评估 LLM 的化学能力，该测试涵盖了多个学科和多个难度级别的中文测试题 ([Section S1](#))。由于 SynAsk 的原始语言为中文，因此该测试完全采用中文进行。然而，我们也认识到，由于 LLM 强大的语言能力，若使用不同的主要语言进行测试，测试结果可能会相近。

We provide a set of example questions for the chemistry question in C-Eval at multiple levels. Sections [Section S3.1](#), [Section S3.2](#), and [Section S3.3](#) refers to the chemistry questions at college, high school and middle school levels, respectively. The dataset format consists of multiple-choice questions and answers. The Prediction contains the answers predicted by three models: SynAsk, Qwen1.5-14B-Chat, and Qwen-14B-Chat. 我们提供了 C-Eval 化学测试中的多个层级的示例题目。[Section S3.1](#)、[Section S3.2](#)和[Section S3.3](#)分别对应大学、高中和初中的化学测试题。该数据集采用多项选择题的格式，包括问题和答案。预测结果部分包含了三个模型的答案：SynAsk、Qwen1.5-14B-Chat 和 Qwen-14B-Chat。

S3.1 C-Eval (College Chemistry)

Problem:

UTF8gbsn 以下是中国关于大学化学考试的单项选择题，请选出其中的正确答案。

下列说法中，正确的是：

- (A) 单质的焓为零
- (B) 反应的热效应就是该反应的摩尔焓变
- (C) 单质的摩尔生成焓为零
- (D) 由最稳定单质生成 1 mol 化合物时，该化合物的标准摩尔生成焓 $\Delta_f H_m^e$ 等于该生成反应的 $\Delta_r H_m^e$

English translation:

The following are single-choice questions on university chemistry exams in China. Which of the following statements is correct?

- (A) The enthalpy of an element is zero
- (B) The heat of reaction is equal to the molar enthalpy change of the reaction
- (C) The molar enthalpy of formation of an element is zero
- (D) When 1 mole of a compound is formed from the most stable elements, the standard molar enthalpy of formation $\Delta_f H_m^e$ of the compound is equal to the standard molar enthalpy of reaction $\Delta_r H_m^e$ of the formation reaction

Answer: D

Predictions:

SynAsk: D

Qwen1.5-14B-Chat: C

Qwen-14B-Chat: C

S3.2 C-Eval (High School Chemistry)

UTF8gbsn Problem:

以下是中国关于高中化学考试的单项选择题，请选出其中的正确答案。

下列说法中，正确的是：

在一定温度下的恒容密闭容器中，当下列哪些物理量不再发生变化时，表明下述反应：

$A(s) + 2B(g) \rightleftharpoons C(g) + D(g)$ 已达到平衡状态：

- ① 混合气体的压强
- ② 混合气体的密度
- ③ B 的物质的量浓度
- ④ 气体的总物质的量
- ⑤ 混合气体总质量

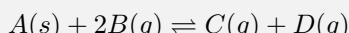
- (A) ②③⑤
- (B) ①②③
- (C) ②③④
- (D) ①③④⑤

English translation:

The following are single-choice questions on high school chemistry exams in China. Please select the correct answer.

Which of the following statements is correct?

In a constant-volume sealed container at a certain temperature, when which of the following physical quantities no longer change, it indicates that the following reaction:



has reached equilibrium:

- ① Pressure of the mixed gases
- ② Density of the mixed gases
- ③ Concentration of substance B
- ④ Total amount of gas
- ⑤ Total mass of the mixed gases

- (A) ②③⑤
- (B) ①②③
- (C) ②③④
- (D) ①③④⑤

Answer: A

Predictions:

SynAsk: A

Qwen1.5-14B-Chat: ②③⑤

Qwen-14B-Chat: A

It is noted while Qwen1.5-14B-Chat provides the right answer, it predicts with the context of the answer directly without showing the correct choice "A".

S3.3 C-Eval (Middle School Chemistry)

UTF8gbsn Problem:

以下是中国关于初中化学考试的单项选择题，请选出其中的正确答案。

下列有关实验现象的描述正确的是：

- (A) 硫在氧气中燃烧发出淡蓝色火焰
- (B) 无色酚酞试液遇稀盐酸变成红色
- (C) 硫酸铜溶液和氢氧化钠溶液反应会产生蓝色沉淀
- (D) 红磷在空气中燃烧产生白雾

English translation:

The following are single-choice questions on junior high school chemistry exams in China. Please select the correct answer.

Which of the following descriptions about experimental phenomena is correct?

- (A) Sulfur burns with a pale blue flame in oxygen
- (B) Colorless phenolphthalein turns red when mixed with dilute hydrochloric acid
- (C) The reaction between copper sulfate solution and sodium hydroxide solution produces a blue precipitate
- (D) Red phosphorus burns in air to produce white smoke

Answer: C

Predictions:

SynAsk: C

Qwen1.5-14B-Chat: C

Qwen-14B-Chat: C

S4 Reaction yield prediction results

We randomly collected 60 recent published S_NAr reactions from the last three years (2022-2024), which includes new substrate molecules and never seen by the reaction model. The model predicted yield versus literature reported yield are compared in the attached spreadsheet file, SI.xlsx, with an MAE of 14.1%. These recent published reactions consist of seven different reaction conditions.

我们随机收集了过去三年（2022-2024年）中60个最近发表的 S_NAr 反应，其中包括新的底物分子和反应模型从未见过的分子。在随附的电子表格文件SI.xlsx中比较了模型预测的产量与文献报道的产量，MAE为14.1%。这些最近发表的反应由七种不同的反应条件组成。

S5 Retrosynthetic pathway of selected target molecules

Figure S1, Figure S2, Figure S3 and Figure S4 shows numbers of retrosynthetic pathways generated by SynAsk, which provide insights for synthetic chemists. The routes indicate the ability of SynAsk in computer assisted synthetic planning (CASP).

Figure S1、Figure S2、Figure S3 和 Figure S4 显示了 SynAsk 生成的逆合成途径数量，为合成化学家提供了见解。这些途径表明了 SynAsk 在计算机辅助合成规划 (CASP) 中的能力。

We are developing strategies towards generation of more reasonable retrosynthetic pathways. This will be published elsewhere, and integrated into SynAsk. Till now, no efforts were made to experimentally validate the synthetic routes provided in the ESI, and more synthetic routes to other target molecules can be generated via command to SynAsk.

我们正在制定生成更合理的逆合成途径的策略。这将在其他地方发布，并集成到 SynAsk 中。到目前为止，还没有尝试通过实验验证 ESI 中提供的合成路线，可以通过 SynAsk 命令生成更多其他目标分子的合成路线。

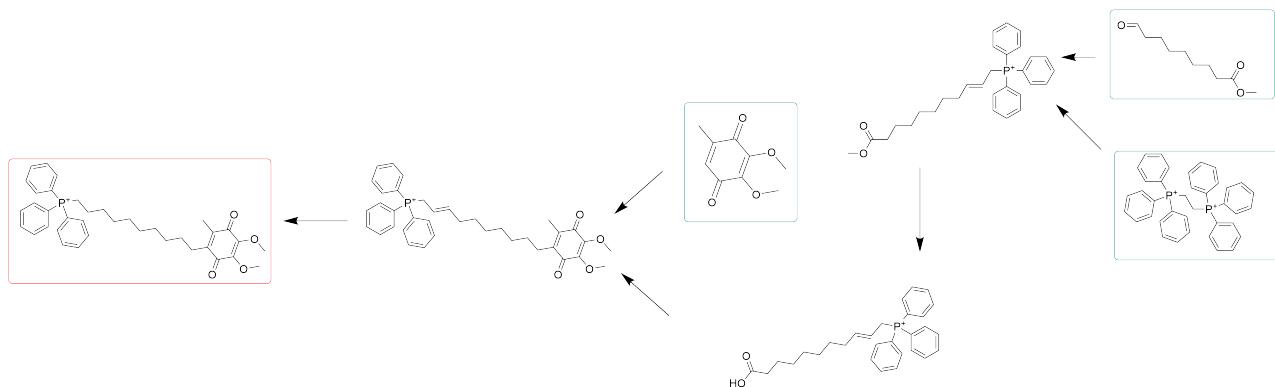


Fig. S1 The synthetic route of the target molecule mitoquinone planned by SynAsk's retrosynthetic tool.

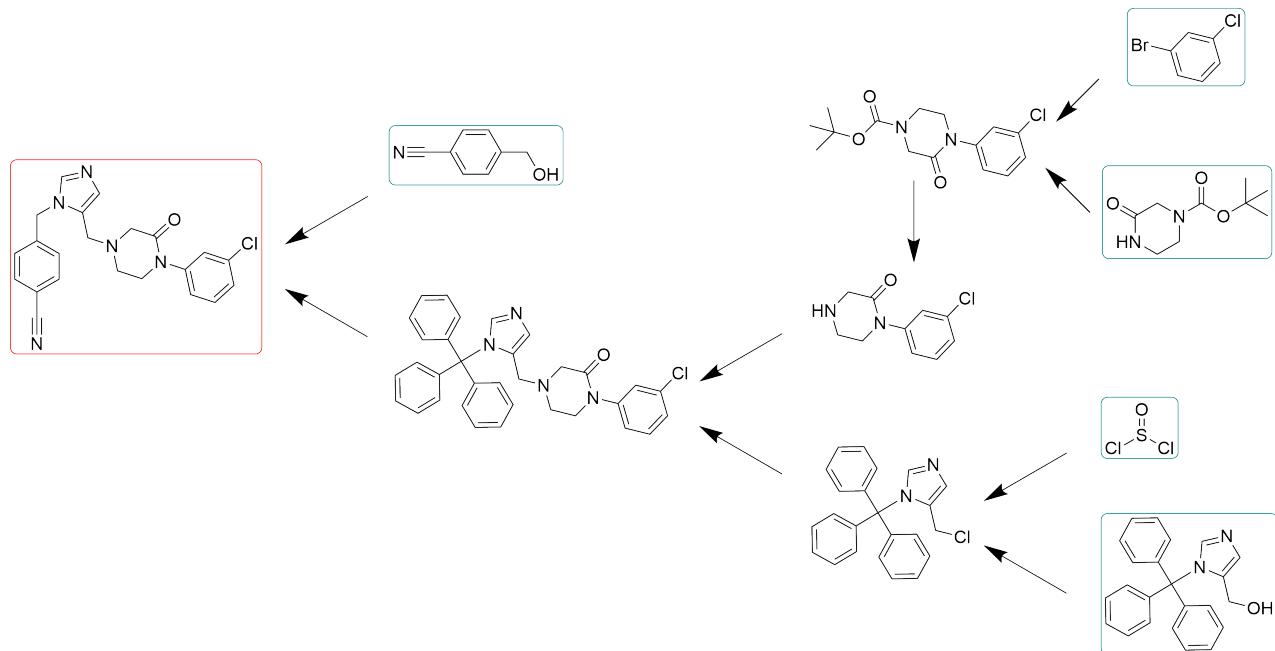


Fig. S2 The synthetic route of the target molecule L-778123 planned by SynAsk's retrosynthetic tool.

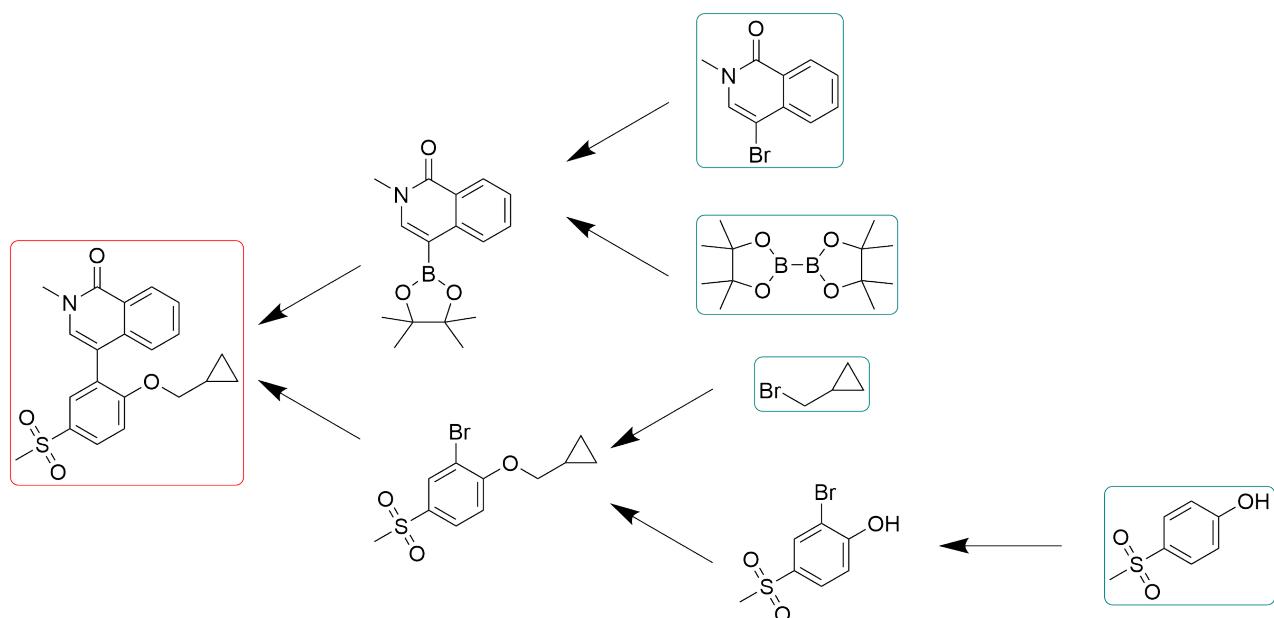


Fig. S3 The synthetic route of the target molecule trotabresib planned by SynAsk's retrosynthetic tool.

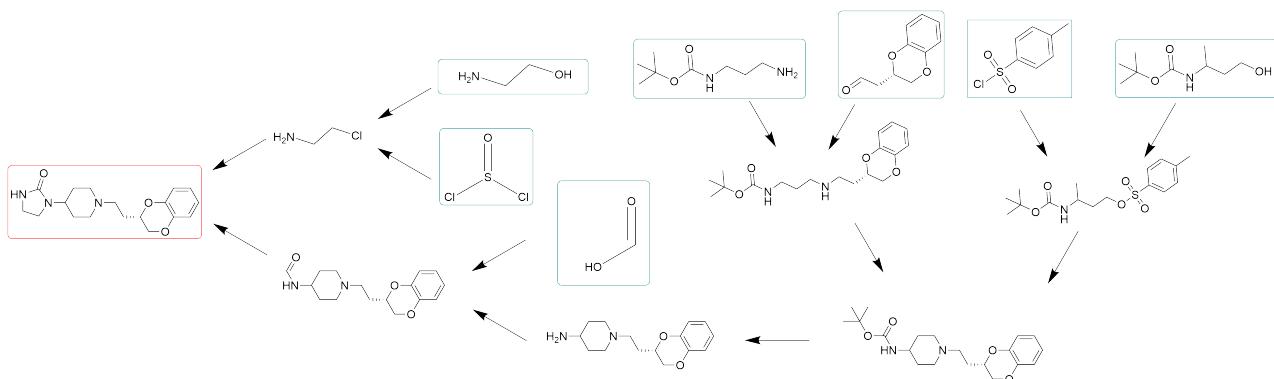


Fig. S4 The synthetic route of the target molecule azaloxan planned by SynAsk's retrosynthetic tool.

S6 Examples of the SynAsk platform outputs versus other LLMs

User	ChatGPT-4
what is dopamine?	Dopamine is a neurotransmitter, a chemical messenger that plays a crucial role in transmitting signals in the brain and other areas of the body. Here are some key points about dopamine:
SynAsk	ChemCrow
Dopamine is a neurotransmitter, a chemical messenger, that plays a crucial role in the brain. It is primarily involved in the regulation of mood, motivation, reward, and movement. Dopamine is often associated with feelings of pleasure and reinforcement, and imbalances in dopamine levels have been linked to various neurological and psychological conditions, such as Parkinson's disease and addiction.	<p>Chemical Properties Chemical Structure: C8H11NO2 Systematic Name: 4-(2-aminoethyl)benzene-1,2-diol Classification: Catecholamine and phenethylamine Abbreviation: DA Function in the Nervous System Neurotransmitter Role: Dopamine is primarily a neurotransmitter in the central nervous system (CNS). It is synthesized in dopaminergic neurons from the amino acid tyrosine via the intermediate L-DOPA. Major Dopaminergic Pathways in the Brain: Mesolimbic Pathway: Related to reward, pleasure, and reinforcement learning. Mesocortical Pathway: Associated with cognition and executive function. Nigrostriatal Pathway: Involved in movement regulation; degeneration of neurons in this pathway is linked to Parkinson's disease. Tuberoinfundibular Pathway: Inhibits prolactin release from the pituitary gland. Physiological Functions Reward and Pleasure: Dopamine is often called the "feel-good" neurotransmitter because it contributes to feelings of pleasure and reward. Motor Control: In the nigrostriatal pathway, dopamine helps regulate movement. Its deficiency leads to motor symptoms seen in Parkinson's disease. Cognition and Attention: It plays a role in memory, attention, and problem-solving. Hormone Regulation: Inhibits prolactin release from the anterior pituitary. Clinical Significance Neurological Disorders: Parkinson's Disease: Characterized by a loss of dopamine-producing neurons in the substantia nigra. Schizophrenia: Linked to dysregulated dopamine activity, especially in the mesolimbic pathway. Psychiatric Disorders: Depression: Dopamine dysregulation is implicated in certain depressive disorders. ADHD (Attention Deficit Hyperactivity Disorder): Related to dopamine function in the prefrontal cortex. Addiction: Dopamine release is associated with addictive behaviors due to its role in the reward pathway. Synthesis and Metabolism Biosynthesis: Precursor: Tyrosine Intermediate: L-DOPA Enzymes: Tyrosine hydroxylase (tyrosine to L-DOPA) and DOPA decarboxylase (L-DOPA to dopamine) Catabolism: Metabolized by monoamine oxidase (MAO) and catechol-O-methyltransferase (COMT) to homovanillic acid (HVA), which is excreted in the urine. Applications and Treatments Pharmacology: Levodopa (L-DOPA): Used to treat Parkinson's disease. Dopamine Agonists: Drugs that mimic dopamine effects in the brain. Dopamine Antagonists: Block dopamine receptors and are used in treating schizophrenia. Clinical Monitoring: Dopamine levels can be assessed through homovanillic acid (HVA) levels in the cerebrospinal fluid or urine. If you need further elaboration on any specific aspect or its role in particular disorders, I'd be happy to expand on that! </p>

Fig. S5 The first example of the outputs from the LLMs.

User	ChatGPT-4
Can you recommend me some reaction conditions for Suzuki cross-coupling?	<p>The Suzuki cross-coupling reaction, or Suzuki-Miyaura coupling, is a powerful method for forming carbon-carbon bonds between organoboranes (typically boronic acids) and organohalides or pseudohalides in the presence of a palladium catalyst. Here are some common reaction conditions for this versatile reaction:</p> <p>General Reaction Scheme $R-X + R'-B(OR)_2 + \text{Base} + \text{Pd catalyst} \rightarrow R-R'$</p> <p>Common Reaction Components</p> <p>Substrates: R-X: Aryl or vinyl halides (X = I, Br, Cl) or pseudohalides (OTf, N2, etc.) R'-B(OR)2: Boronic acids, boronate esters, trifluoroborates, etc.</p> <p>Catalysts: Pd(PPh3)4 Pd(OAc)2/PPh3 Pd2(dba)3/P(tBu)3 PdCl2(dppf)</p> <p>Bases: Inorganic: K2CO3, Cs2CO3, Na2CO3, NaHCO3, NaOH Organic: Et3N, iPr2NEt, DBU</p> <p>Solvents: Aqueous Organic: DMF/H2O, THF/H2O, DME/H2O Non-Aqueous: Toluene, dioxane, DME, MeOH, DMF</p> <p>Additives: Ligands: Phosphines like PPh3, P(tBu)3, PCy3, BINAP, SPhos, XPhos Stabilizers: Tetrabutylammonium salts</p> <p>Example Conditions</p> <ol style="list-style-type: none"> Standard Conditions Aryl Halide: R-Br or R-I Boronic Acid: R'-B(OH)2 Catalyst: Pd(PPh3)4 (1-5 mol%) Base: K2CO3 or Cs2CO3 Solvent: Toluene/H2O (4:1), DME/H2O, or EtOH/H2O Temperature: 80–100 °C Time: 12–24 hours Example Reaction Setup: Aryl Halide (1 mmol) Boronic Acid (1.2 mmol) Pd(PPh3)4 (3 mol%) K2CO3 (2 mmol) Toluene/H2O (4:1, 5 mL) 80 °C, 12 h Chloride Activation Aryl Halide: R-Cl Boronic Acid: R'-B(OH)2 Catalyst: PdCl2(dppf) or Pd2(dba)3/SPhos Base: K3PO4 or CsF Solvent: 1,4-Dioxane or DMF Temperature: 80–120 °C Example Reaction Setup: Aryl Chloride (1 mmol) Boronic Acid (1.2 mmol) PdCl2(dppf) (3 mol%) K3PO4 (2 mmol) 1,4-Dioxane (5 mL) 100 °C, 24 h Water-Based Conditions Aryl Halide: R-Br Boronic Acid: R'-B(OH)2 Catalyst: Pd(OAc)2/SPhos or Pd(OAc)2/PPh3 Base: K2CO3, NaOH Solvent: H2O or H2O/EtOH (1:1) Example Reaction Setup: Aryl Bromide (1 mmol) Boronic Acid (1.2 mmol) Pd(OAc)2 (2 mol%) SPhos (4 mol%) NaOH (2 mmol) H2O/EtOH (1:1, 5 mL) 80 °C, 12 h <p>Notes on Optimizing Conditions</p> <p>Ligand Choice: Bulky phosphine ligands like SPhos and XPhos work well with less reactive aryl chlorides.</p> <p>Base Selection: Inorganic bases like K2CO3 and Na2CO3 are commonly used, but stronger bases (NaOH, CsF) may be required for less reactive halides.</p> <p>Temperature: Typically 80–120 °C; adjust for substrate reactivity.</p> <p>Catalyst Loading: 1–5 mol% is common; higher loading may be needed for sterically hindered substrates.</p> <p>These guidelines provide a starting point for experimenting with Suzuki coupling reactions. Do you need specific conditions tailored to a particular set of substrates?</p>
ChemCrow	
<p>The recommended reaction conditions for Suzuki cross-coupling are as follows:</p> <p>Condition 1: Reagents: Pd(PPh3)4 (0.05 equiv); Reagents: K2CO3 (3.0 equiv); Solvent: THF/H2O (4:1); Reaction conditions: 100°C, 12h</p> <p>Condition 2: Reagents: PdCl2(PPh3)2 (0.05 equiv); Reagents: Na2CO3 (3.0 equiv); Solvent: THF/H2O (4:1); Reaction conditions: 100°C, 12h</p> <p>Condition 3: Reagents: Pd(OAc)2 (0.05 equiv); Reagents: dppf (0.1 equiv); Reagents: Cs2CO3 (3.0 equiv); Reagents: CuCl (1.0 equiv); Solvent: DMF; Reaction conditions: 100°C 12h</p>	<p>The Suzuki-Miyaura cross-coupling reaction is a pivotal method in organic chemistry for the formation of carbon-carbon bonds. The efficiency of this reaction hinges on the careful selection of reaction conditions, such as the choice of solvent, catalyst, and particularly, the reaction temperature, to ensure the desired coupling while minimizing side reactions. For instance, in the optimization of conditions for diazirine-containing compounds, a temperature of 40°C was identified as optimal. This temperature not only preserved the diazirine moiety but also limited its participation in competitive side reactions. The choice of coupling partners is also crucial, with bromoquinoline showing higher conversion efficiency than chloroquinoline. This specificity underscores the importance of selecting appropriate reaction parameters to achieve efficient coupling, which is essential for the synthesis of pharmaceutically relevant structures and the broader application of the Suzuki-Miyaura reaction in organic synthesis and material science.</p>

Fig. S6 The second example of the outputs from the LLMs.

User

Please help me find some literature related to C-H activation

ChatGPT-4

C-H activation is a significant area of research in organic chemistry that involves the activation of otherwise inert C-H bonds to facilitate their functionalization. Here are some sources and directions to guide your literature search:

General Overview**Review Articles:**

Yu, J.-Q.; Shi, Z. (Eds.) "C-H Activation." Springer, 2010.
Labinger, J. A.; Bercaw, J. E. "Understanding and exploiting C-H bond activation." Nature, 2002, 417(6888), 507-514.

Books:

Crabtree, R. H. "The Organometallic Chemistry of the Transition Metals." Wiley, 2014.
Yamaguchi, J.; Yamaguchi, A. D.; Itami, K. "C-H Bond Functionalization: Emerging Synthetic Tools." Angew. Chem. Int. Ed., 2012, 51(36), 8960-9009.

Specific Topics in C-H Activation**Transition Metal Catalysts:**

Ritter, T.; Stahl, S. S. "Catalytic C-H Functionalization via C-H Bond Activation." Catalytic C-H Functionalization, Wiley-VCH, 2013.

Giri, R.; Shi, B. F.; Engle, K. M.; Maugel, N.; Yu, J.-Q. "Transition Metal-Catalyzed C-H Functionalization Reactions: Diastereoselectivity and Enantioselectivity." Chem. Soc. Rev., 2009, 38, 3242-3272.

Direct Arylation:

Seregin, I. V.; Gevorgyan, V. "Direct Transition Metal-Catalyzed Functionalization of Heteroaromatic Compounds." Chem. Soc. Rev., 2007, 36, 1173-1193.

Ackermann, L. "Carboxylate-Assisted Transition-Metal-Catalyzed C-H Bond Functionalization: Mechanism and Scope." Chem. Rev., 2011, 111(3), 1315-1345.

C-H Activation in Pharmaceuticals:

Li, B.-J.; Shi, Z.-J. "Challenging C-H Functionalization Reactions for Organic Synthesis." Chem. Soc. Rev., 2012, 41(15), 5588-5598.

Godula, K.; Sames, D. "C-H Bond Functionalization in Complex Organic Synthesis." Science, 2006, 312(5770), 67-72.

Recent Advances and Novel Methodologies**Recent Reviews:**

Chen, Z.; Rong, M.-Y.; Nie, J.; Zhu, X.-F.; Shi, B.-F.; Ma, J.-A. "Catalytic Alkylation of Unactivated C(sp³)-H and C(sp²)-H Bonds via Radical Process." Chem. Soc. Rev., 2019, 48, 4921-4942.

Hartwig, J. F.; Larsen, M. A. "Undirected, Homogeneous C-H Bond Functionalization: Challenges and Opportunities." ACS Cent. Sci., 2016, 2(5), 281-292.

Cross-Coupling Reactions:

Sun, C.-L.; Li, B.-J.; Shi, Z.-J. "Direct C-H Transformation via Iron Catalysis." Chem. Rev., 2011, 111(3), 1293-1314.

Wencel-Delord, J.; Dröge, T.; Liu, F.; Glorius, F. "Towards Mild Metal-Catalyzed C-H Bond Activation." Chem. Soc. Rev., 2011, 40(9), 4740-4761.

Photo and Electrochemistry:

Cui, L.; Ackermann, L. "Electrocatalysis: C-H Activation via Anodic Oxidation." ACS Catal., 2021, 11(23), 14022-14029.

Liu, Z.; Chen, X.; Tan, J. D.; Wu, Y. Z.; Hartwig, J. F.; Zhao, P. "Visible-Light Photoredox-Catalyzed C-H Bond Functionalization with Metal Catalysts." Acc. Chem. Res., 2020, 53(6), 1314-1325.

Practical Search Strategies**Use Relevant Databases:**

SciFinder: The most comprehensive database for chemical literature.

Web of Science: Provides citation connections between articles.

Google Scholar: Free and useful for initial searches.

Target Keywords:

"C-H Activation," "C-H Functionalization," "Transition Metal-Catalyzed," "Cross-Coupling," "Direct Arylation," etc.

Research Groups and Researchers to Follow:**ChemCrow**

Here are some literature related to C-H activation:

1. Liangliang Song and E. V. Van der Eycken. "Transition metal-catalyzed intermolecular cascade c-h activation/annulation processes for the synthesis of polycycles." Chemistry, 2020. This paper discusses significant advancements in the field of C-H activation, particularly in the synthesis of polyheteroaromatic compounds and other complex molecules through innovative strategies such as rhodium(III)-catalyzed double C-H activation/annulation reactions.
2. Richmond Sarpong. "C-h functionalization/activation in organic synthesis." Beilstein Journal of Organic Chemistry, 12:2315 - 2316, 2016. This paper discusses the growth in popularity of C-H functionalization/activation due to its potential for more sustainable synthesis methods and the pursuit of novel reactivity and selectivity in molecule building.

Fig. S7 The third example of the outputs from the LLMs.