

mol-instructions：用于 llms 的大规模生物分子指令数据集

Yin Fang^{★*}, Xiaozhuan Liang^{★*}, Ningyu Zhang,^{★†} Kangwei Liu[★], Rui Huang[★], Zhuo Chen[★], Xiaohui Fan[★], Huajun Chen^{★♦♡†}

*浙江大学计算机科学与技术学院

♦浙江大学-蚂蚁金服集团知识图谱联合研究中心

♡浙江大学杭州创新中心

{fangyin, liangxiao, kangweiliu, hrhr, zhuo.chen, fanxh}@zju.edu.cn,

{zhangningyu, huajunsir}@zju.edu.cn

<https://github.com/zjunlp/Mol-Instructions>

😊 数据集 🏠 分子模型 🐛 蛋白质模型 📁 生物质模型

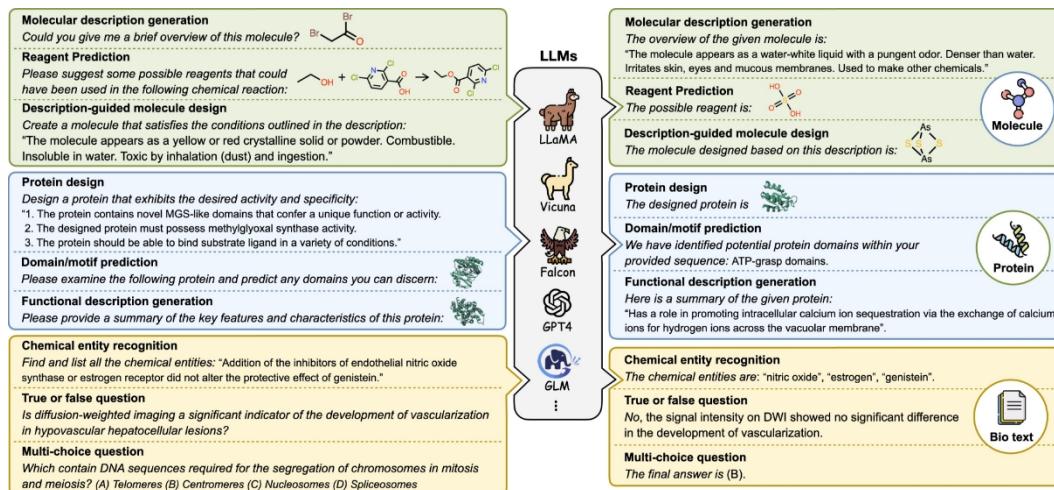


图 1：利用“分子指令”（Mol-Instructions）增强大型语言模型的能力，从而打开生物分子领域的大门。分子结构和蛋白质结构都以序列的形式表示。

摘要

大型语言模型（LLMs）具有出色的任务处理能力和创新的输出结果，推动了各个领域的重大进步。然而，它们在生物分子研究等专业领域的能力仍然有限。为了应对这一挑战，我们推出了专为生物分子领域设计的综合指令数据集 Mol-Instructions。Mol-Instructions 包含三个关键部分：面向分子的指令、面向蛋白质的指令和生物分子文本指令。每个部分都旨在提高 LLM 对生物分子特征和行为的理解和预测能力。通过对LLMs进行广泛的指令调整实验，我们证明了Mol-Instructions在提高大型模型在错综复杂的生物分子研究领域的性能方面的有效性，从而促进了生物分子研究领域的进步。Mol-Instructions可公开用于正在进行的研究，并将定期更新，以提高其适用性。

作为会议论文在 2024 年国际比较文学和历

史研究国际会议 (ICLR 2024) 上发表

· 平等贡献和共同第一作者。

† 通讯作者：

1 导言

GPT-4 (OpenAI, 2023a)、Chinchilla (Hoffmann 等, 2022)、PaLM (Chowdhery 等, 2023)、Codex (Chen 等, 2021)、LLaMA (Touvron 等, 2023)、FLAN (Wei 等, 2022) 和 GLM (Zeng 等, 2023) 等大型语言模型 (LLM) 彻底改变了自然语言处理 (NLP) 的面貌。这些模型拥有数十亿个参数, 在大量文本语料库中经过精心训练, 擅长生成类人文本和理解复杂语境。为了让 LLMs 适应特定任务, 研究人员采用了指令调整技术 (欧阳等人, 2022a; Sanh 等人, 2022)。这包括用专门的指令数据集来训练模型, 让它们获得特定任务的知识和模式, 从而提高它们在特定领域的性能。目前已经开发了许多用于一般领域的指令数据集。例如, 斯坦福大学的 Alpaca 数据集 (Taori et al.) GPT4All 数据集 (Anand 等人, 2023 年) 包含多种格式, 包括代码、故事和对话, 专门用于训练和评估通用语言模型。同样, COIG 数据集 (Zhang 等人, 2023a) 整合了各种语料, 如翻译、考试和人值对齐指令, 专门用于中文语言处理。

随着最近的进步, LLMs 显然已经超越了传统的文本处理 (Zhang 等人, 2024 年; Tinn 等人, 2023 年; Wang 等人, 2023 年 a)。它们在生物分子研究 (包括结构生物学、计算化学和药物开发) 方面的潜力尤其令人期待。在这一领域利用法医学可能会彻底改变我们对生物分子数据的理解和处理, 加速科学创新和药物发现。

然而, 在生物分子领域利用 LLM 的一个主要障碍是缺乏该领域的专用数据集。尽管在一般领域已有多个指导数据集, 但在生物分子领域仍存在明显差距。这一空白主要源于三个挑战: **首先**, 鉴于生物分子数据本身的复杂性和所含信息的丰富深度, 获取和注释生物分子数据需要大量成本。**其次**, 生物分子组合涉及广泛的知识领域, 交织着来自不同领域的专业见解, 包括结构生物学、计算化学和药物开发。**第三**, 与自然语言处理领域成熟的框架不同, 生物信息学没有标准化的通用语言。不同的应用通常会对生物分子及其相关计算采用不同的表示方法。这种多样性加大了制作一个可在整个领域统一使用的数据集的挑战。为了解决生物分子领域的这一迫切需求, 我们推出了 Mol-Instructions (CC BY-NC-SA 4.0), 这是一个针对生物分子研究的独特挑战而量身定制的数据集。如图 1 所示, 该数据集由三个核心部分组成:

- **以分子为导向的指令:** 该组件深入研究小分子领域, 强调其固有特性和行为。它揭示了各种化学反应和分子设计所面临的基本挑战, 包含六项任务的 1484K 条指令。
- **面向蛋白质的指令:** 它以生物科学为重点, 涵盖 505K 条指令, 涵盖五类任务。这些任务旨在预测蛋白质的结构、功能和活性, 并促进基于文本指令的蛋白质设计。
- **生物分子文本指令:** 这部分主要针对生物信息学和化学信息学领域的 NLP 任务, 包括六项信息提取和问答任务, 共 53K 条指令。

创建该指令数据集涉及从各种**授权来源**收集生物分子数据 (详见附录 A.2), 然后将这些数据转换为适合特定任务的易学指令格式。我们的目标是让 LLM 具备特定领域的洞察力, 提

他们解码和预测生物分子特征的能力。这种提升可以彻底改变对生物分子数据的解读，简化药物开发过程，并开辟生物分子研究的新领域。有了 "莫尔指令"，大型模型被赋予了理解生物学的能力，为新的科学发现打开了大门。为了评估 "莫尔指导" 的实际效果，我们进行了一系列广泛的评估。我们采用具有代表性的 LLM 作为基础模型，针对三大类指令分别进行了指令调整。结果凸显了 Mol-Instructions 的价值，证明它有能力在复杂的生物分子研究领域提高大型模型的通用性和理解力。

表 1：与现有数据集的比较。我们使用了以下缩写：HG--由人工策划的数据集，SI--通过自我指导方法生成的数据集，MIX--由人工构建和机器生成的数据组成的数据集，COL--由各种其他数据集组合而成的数据集。

数据集	# 类型	# 说明	收藏品	使用方法	访问
一般领域					
斯坦福羊驼 (Taori 等人, 2023 年)	文本	52,002	SI	指令调谐	开放
多利-v2 (Conover 等人, 2023 年)	文本	15,015	HG	指令调谐	开放
)					
白泽 (Xu 等人, 2023 年)	文本	653,699	混合	指令调谐	开放
FLAN (Wei 等人, 2022 年)	文本	1,764,800	COL	指令调谐	开放
InstructGPT (欧阳等人, 2022b)	文本	112,801	HG	RLHF, 指令调谐	关闭
GPT份额 (沙, 2023年)	文本	260,137	混合	指令调整, 聊天	关闭
BAALCOIG (Zhang 等人, 2023a)	文本	67,798	COL	指令调谐	开放
UltraChat超级聊天 (Ding 等人, 2023 年)	文本	1,468,352	混合	聊天	开放
卡拉狄加》 (泰勒等人, 2022 年)	文本、生物分子	783,599	混合	培训前	关闭
特定领域					
PCdes (Zeng 等人, 2022 年)	文本、分子	15,000	混合	培训前	关闭
ChEBI-20 (Edwards 等人, 2022 年)	文本、分子	33,010	COL	培训前	开放
PubChemSTM (Liu et al., 2023)	文本、分子	281,000	COL	培训前	关闭
墨木 (Su 等人, 2022 年)	文本、分子	15,000	混合	培训前	开放
摩尔指令 (我们的)	文本、生物分子	2,043,587	混合	指令调谐	开放

2 相关工作

教学数据更注重以任务为中心，通常利用有限的实例集来增强 LLM 的功效。最初的研究工作，如 Dolly-v2 (Conover 等人, 2023 年) 和 InstructGPT (欧阳等人, 2022b)，主要依靠人工或专家注释来为各种 NLP 任务提供指导，如 封闭式问答和总结。虽然人工标注的结构化数据通常具有很高的质量，但其数量、多样性和创新性都很有限。认识到这些局限性后，人们开始转向半自动或全自动指令创建。例如，斯坦福大学的 Alpaca (Taori 等人, 2023 年) 采用了自我指令方法 (Wang 等人, 2023b)，利用基于手工指令集的引导技术生成了 52K 条不同的指令。这种创新方法启发了许多模型辅助数据收集工作，如 Baize (Xu 等人, 2023 年)、COIG (Zhang 等人, 2023 年 a)、UltraChat (Ding 等人, 2023 年) 和 ShareGPT (sha, 2023 年)，详见表 1。

一些研究已经探索了文本与生物分子的交叉 (Boiko 等人, 2023 年; Bran 等人, 2023 年; Zeng 等人, 2022 年; Edwards 等人, 2022 年; Nascimento & Pimentel, 2023 年; Zhang 等人, 2023b)。例如，通用科学语言模型 Galactica (Taylor 等, 2022 年) 在预训练阶段纳入了与分子、蛋白质和文本相关的指令数据子集。不过，这一特定数据集的细节仍未披露。其他数据集，如 PCdes (Zeng 等人, 2022 年)、ChEBI-20 (Edwards 等人, 2022 年)、PubChemSTM (Liu 等人, 2023 年) 和 MoMu (Su 等人, 2022 年)，将分子与文本描述配对。虽然这些数据集很有价值，但它们主要用于训练较小的模型，缺乏指导格式。这一局限性限制了它们对大型语言模型的直接实用性。相比之下，Mol-Instructions 涵盖了更广泛的生物分子任务和更大量的数据，它是通过自我指导、基于模板的转换和人为任务描述（详见第 3 节）的组合构建方法实现的。

3 摩尔结构

3.1 基本原则

大规模 为了满足法律硕士的需求，我们精心设计了 Mol-Instructions 这一庞大的工具，其中包含 200 多万条生物分子指令。这一大批内容涵盖了生物分子序列和结构的广泛性和代表性，使模型能够掌握和驾驭复杂的生物分子。

Diversity Mol-Instructions 涵盖三种类型生物分子的 17 个子任务，包括多种文本描述，捕捉了超过 11 种独特的生物分子特性。这种广泛的覆盖面促进了使用我们的数据集训练的模型的通用性，使它们能够应对生物分子研究中的各种挑战。

质量 我们将数据集的质量放在首位，以确保数据集为获得准确、可操作和实用见解奠定可靠的基础。每一份生物分子数据都经过严格审查，以确保其准确性和可信度。

3.2 人机协作任务描述创建

标准的指令条目通常由三部分组成：明确任务的 **指令**、作为任务 **输入** 的输入以及体现预期结果的 **输出**。

在现实世界的场景中，任务 **指令** 呈现出广泛的多样性，以匹配人类查询和需求的动态性和多样性。为了模拟这种多样性，对于每项任务，我们都从人类编写的简洁明了的描述开始，然后将其作为 gpt-3.5-turbo 的输入 (OpenAI, 2023b)。利用其丰富的知识和模式识别技能，LLM 生成了多种任务描述，反映了人类广泛的问题框架风格 (图 2 扇形 1)。为确保这些描述的质量，每一份描述都要经过彻底的人工审核。这种人类洞察力与机器智能的合作不仅增强了任务描述的多样性和创造性，还提高了指令的稳健性和适应性。

3.3 从现有数据中提取信息

由于生物分子数据通常涉及专业的湿实验室实验和化学家的专家总结，因此我们的数据来源于广泛使用的生物化学数据库 (Kim 等, 2021; Wei 等, 2010; Lu & Zhang, 2022; Wu 等, 2018; Ashburner 等, 2000; Consortium, 2023; Krallinger 等, 2015; 2017; Li 等, 2016; Pal 等, 2022; Hendrycks 等, 2021)。基于这些数据源，我们可以通过适当的处理获得所需的指令数据。

有些数据集经过人工处理，标签直接附加到特定的生物分子数据上。这包括为各种预测工作明确界定输入和预期结果的数据集，以及为问答任务预先确定问题和相应答案的数据集。处理这类数据相当直接：标注的信息通常会映射到每个指令条目的相应 **输入** 和 **输出** 字段。

另一类数据源没有明确的人工标签，需要数据挖掘和人工智能辅助生成等技术来提取和选

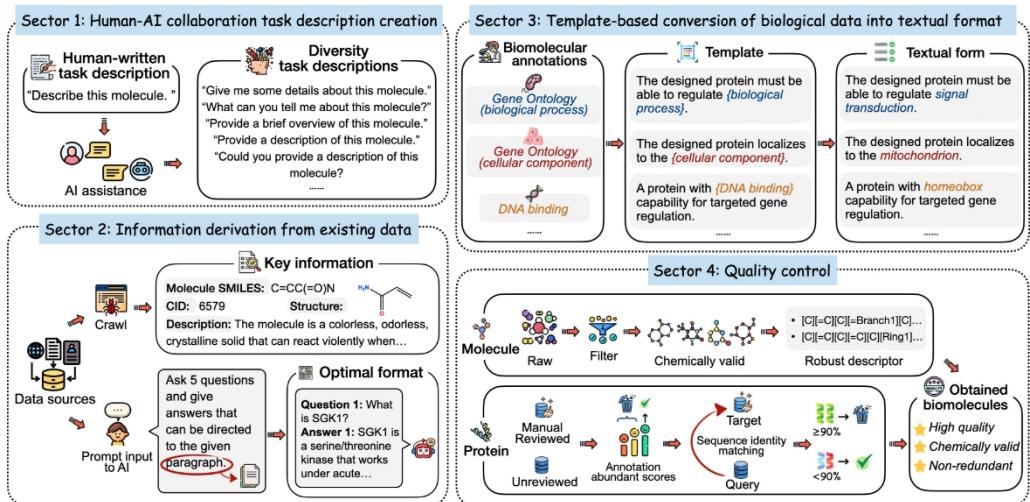


图 2：Mol-Instructions 的数据构建概览，包括四个部分：人机协作任务描述创建 (§3.2)、从现有数据中提取信息 (§3.3)、通过模板将生物数据转换为文本格式 (§3.4) 和质量控制 (§3.5)。每项任务的详细程序请参见附录 B。

史研究国际会议 (ICLR-2024) 上发表。通过数据挖掘，我们从 PubChem (Kim 等人, 2021 年) 等专业化学研究数据库中提取更多相关信息，力求数据的全面性和充分性。具体来说，我们抓取有效的分子描述文本及其相应的 PubChem 化学标识符 (CID)，然后检索相应的分子描述符。为了将模型的注意力集中在描述语义上，我们用 "分子"一词取代了分子术语。在人工智能辅助生成方面，我们使用 PubMed (White, 2020 年) 中的科学摘要来生成开放式问题指令。在这里，我们让 gpt-3.5-turbo 根据摘要以问答对的形式提出问题和相应的答案。这种方法

有效地生成了各种问答指令集，简化了这一特定任务的构建。有关任务层面的详细解释，请参阅附录 B。

3.4 基于模板的生物数据文本格式转换

当然，并非所有数据都能无缝转换成理想的指令集。对于某些新任务来说，找到直接适用的数据尤其具有挑战性。对于蛋白质设计任务来说尤其如此，以往的研究通常以大类 (Madani 等人, 2023 年) 或固定骨架 (Dauparas 等人, 2022 年) 为设计条件，在很大程度上忽略了定制的功能和结构属性 (如螺旋二级结构、DNA 结合域或转录调节活性)。与此相反，我们的目标是根据用户的意图，以文本形式为蛋白质的全新设计提供指导，从而设计出用户感兴趣的复合属性。鉴于直接适用于此类任务的数据稀缺，我们从 UniProtKB (Consortium, 2023 年) 中的特定选择中收集注释。这些注释囊括了该领域研究人员经常探索或寻求的蛋白质特性 (详细示例见附录表 6)。

为了有效地将这些结构化注释转换成文本格式，我们制定了一系列模板，如图 2 扇形 3 和附录表 7 所示。每个由此产生的文本注释都确立了蛋白质设计的标准。通过汇总这些功能描述和属性，我们可以制作精确的蛋白质设计指令，确保合成的蛋白质符合指定的标准。在实践中，考虑到模型的输入长度限制和训练效率，我们会随机选择这些条件的子集作为设计目标。

3.5 质量控制

目前，LLM 尚未完全掌握生物分子语言的复杂性，与人类语言的熟练程度相差甚远。为了提高模型生成准确生物分子的能力，我们对生物分子数据采取了严格的质量保证措施，详见图 2 部门 4。

对于小分子，我们的流程首先是剔除初始数据集中化学性质无效的 SMILES 字符串。虽然外部约束 (Landrum 等人, 2013 年) 可以验证生成的分子，但在分子指令中使用可靠的分子描述符更为有效。虽然 SMILES (Weininger, 1988 年) 字符串仍然是分子描述符的热门选择，但利用它们的模型往往会有语法上有缺陷或化学上不一致的字符串。为了避免这些问题，我们选择 SELFIES (Krenn 等人, 2022 年) 作为分子描述符。SELFIES 规则严密，允许任意符号组合，从而确保生成有效的分子，消除了 SMILES 字符串的常见缺陷，例如产生不合逻辑的符号或不匹配的括号。

与此同时，我们优先考虑蛋白质数据的完整性，主要从 UniProtKB/Swiss-Prot (Consortium, 2023) --一个经过编辑和人工注释的蛋白质序列数据库--中获取条目。为了增加数据的数量和多样性，我们从 UniProtKB/TrEMBL 中补充了高分注释。由于认识到冗余蛋白质序列 (如 Swiss-Prot 和 TrEMBL 数据库中的同源序列或密切相关的蛋白质变体) 可能带来的偏差风险，我们采用了缜密的过滤流程。利用 MMseqs (Steinegger & Soding, 2017) 工具，我

们以 90% 的相似性阈值对蛋白质序列进行聚类，在每个聚类中选择功能丰富的条目作为代表，并排除其他条目。这种严谨的方法确保了我们的数据集包含多样化、高质量、以功能为中心且无冗余的蛋白质说明。

4 进一步了解摩尔指令

4.1 教学任务的分类和潜在应用

如图3所示，Mol-Instructions围绕三个核心领域展开：分子导向、蛋白质导向和生物分子文本。每个类别的教学任务都涵盖了其领域内的基本挑战，希望通过以下方式推动生物分子领域的发展

史研究国际会议 (ICLR 2024) 上发表

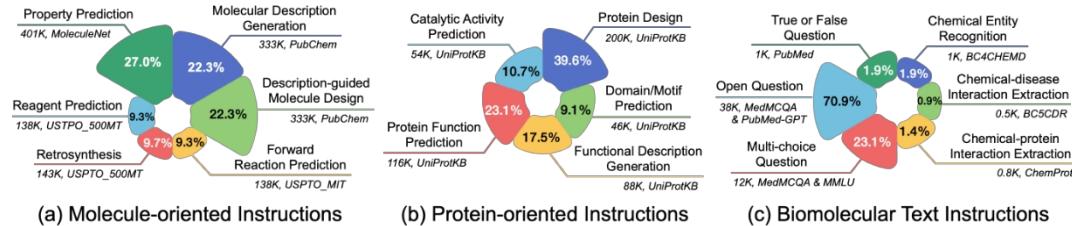


图 3：“莫尔指令”的组成结构。Mol-Instructions 主要包括三大类任务：(a)、(b) 和 (c)。横线上方标注了任务名称，横线下方标注了原始数据的来源和构建的教学数据集的大小，饼图上的百分比代表了每个主要类别中的数据比例。

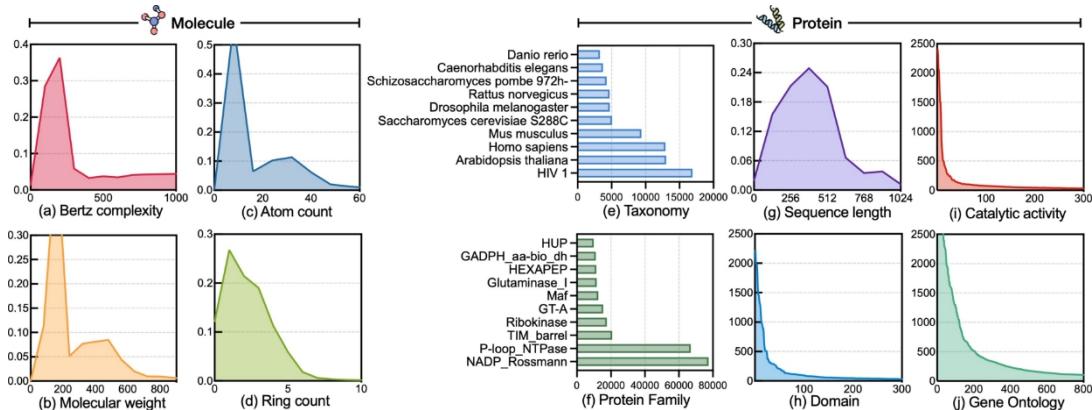


图 4：生物分子序列的多维分析。左侧说明了“莫尔指令”中分子的多样性，包括不同复杂程度和不同结构的分子。右侧的分析则深入研究了 Mol-Instructions 中蛋白质序列的多样性，考虑了序列长度、结构域和活性等不同方面。

LLM 的培训和使用。附录 B 全面阐述了任务定义和指令构建。

以分子为导向的指令以小分子为中心，深入研究其自然特性和行为。它强调了各种化学反应和分子设计的基本挑战，包括图 3 (a) 所示的六项任务。其目的是了解和预测分子的化学性质，改进分子设计，提高化学反应的准确性和速度。在化学和药物设计领域，这些任务的预测结果可加快药物创新并降低开发成本。

面向蛋白质的指令植根于生物科学，主要解决与蛋白质设计和功能有关的问题。这部分内容包括五个不同的类别，详见图 3

(b).这里的工作旨在预测蛋白质结构域、功能和活性，并通过文字说明促进蛋白质设计。了解蛋白质功能和折叠的基本原理对于疾病诊断、治疗和新药发现等领域至关重要。

生物分子文本说明主要集中于与生物信息学和化学信息学有关的 NLP 任务。如图 3 (c)所示，该类别包含六项信息提取和问答任务。其目的是从生物医学文献中解释和提取关键信息，帮助研究人员迅速积累见解，推动他们的研究工作。

4.2 生物分子特征的多样性和复杂性

图 4 全面展示了生物分子在不同维度的分布情况。为了使可视化更加清晰，我们截去了数据的长尾部分，保留了其核心本质。详尽的分析见附录 C。

图 4 (a-d) 展示了分子的各种特征。伯兹复杂性是评估分子复杂性的重要指标。分子量表明了分子的规模和复杂程度，在许多化学反应中起着重要作用。原子数可以让人了解分子的大小和复杂程度，从而影响其稳定性和反应性。环数提供了

表 2：文字描述中包含的主要生物分子特征。

	功能	实例
分子	化学特性	它与金属结合生成氯化物，如氯化钠和氯化钙。这种分子是一种无色易燃气体，有
	物理特性	明显的刺激性气味。
	应用环境	
	安全	用作香料、溶剂和聚合催化剂。
蛋白质	形成	该分子是一种金属，自然存在于岩石、土壤、水和空气中。吸入致命，皮肤吸收剧毒或致命。
	功能	雷普顿蛋白酶抑制剂抑制胰凝乳蛋白酶或胰凝乳蛋白酶抑制剂。
	亚细胞位置	线粒体。
	结构	目标蛋白质必须以螺旋线为主要构象。
	家族和领域 PTM / 处理	设计的蛋白质应包含对其功能至关重要的 PWPW 结构域。在蛋白质设计中加入信号肽。

透视结构的复杂性和潜在的稳定性，以及对化学反应性和可能的生物活性的影响。

在图 4 (e-j) 中，我们深入研究了 Mol-Instructions 中的**蛋白质属性**。图 4 (e-g) 显示了蛋白质序列长度的不同分布。根据 NCBI 分类标准，这些蛋白质涵盖了广泛的物种和实验菌株，包括 13,563 个蛋白家族和 643 个超家族。图 4 (h-j) 聚焦功能方面，如领域、基因本体和催化活性注释。这些数据呈现出明显的长尾分布，凸显了推断蛋白质功能所面临的挑战，尤其是那些功能不常见的蛋白质。

4.3 广泛涵盖生物分子描述

在文本驱动的生物分子设计这一新兴领域，我们的重点在于生物分子描述文本的深度和多样性。

如表 2 所示，**分子文本描述**提供了一个全面、多维和深入的分子信息视角。在信息的深度和广度方面，所列的分子特性提供了一个广阔的视角，从基本的化学属性到具体的应用环境。如此全面的覆盖范围使得文本描述能够以多样和分层的方式描绘分子。通过了解所表达的化学和物理特性，可以掌握分子的基本特征和反应倾向。此外，对分子的应用、环境普遍性和安全方面的深入了解，可以让人全面了解分子的意义和相关的安全考虑因素。

正如第 3.4 节中所详述的，我们将对自然发生的蛋白质的生物学见解和功能注释转化为基于文本的设计规范。表 2 重点介绍了这些**蛋白质**的各种特征，包括蛋白质折叠、成熟、加工及其对生命过程的贡献等五个相关方面。传统的从头开始蛋白质设计强调根据物理原理生成蛋白质，而 Mol-Instructions 则不同，它的目标是创建具有多方面所需特征的蛋白质。这就对模型的能力提出了挑战，即如何辨别由多种属性汇聚而成的错综复杂的取样空间。

5 探索分子指令的潜力

5.1 性能分析的启示

为了研究 "莫尔指令" 能否增强 LLM 对生物分子的理解，我们以 LLama-7B (Touvron 等人

史研究国际会议 (ICLR 2024) 上发表，[2023 年](#)为基础模型，对“谋尔指令”的三个主要领域进行了指令调整。此外，我们还使用了 Alpaca-LoRA ([Tloen, 2023](#))、Baize- 7B ([Xu et al., 2023](#))、[ChatGLM-6B](#) ([Zeng et al., 2023](#))、Vicuna ([Chiang et al., 2023 年](#))、Galactica ([Taylor 等人, 2022 年](#))、Text+Chem T5 ([Christofidellis 等人, 2023 年](#))、Molt5 ([Edwards 等人, 2022 年](#)) 和 PMC- LLaMA-13B ([Wu 等人, 2023 年](#)) 作为基线。我们的数据集分为训练、验证和测试子集。训练集和验证集用于指令调整，测试集用于评估模型性能。有关详细的训练程序、评估指标和案例研究，请参阅附录 D、E 和 F。

表 3：分子特性预测任务的结果

模型	MAE ↓
物业预测 教育	
羊驼	322.109
BAIZE	261.343
CHATGLM	-
拉马	5.553
维库纳	860.051
银河	0.568
我们	10.555 0.013

的

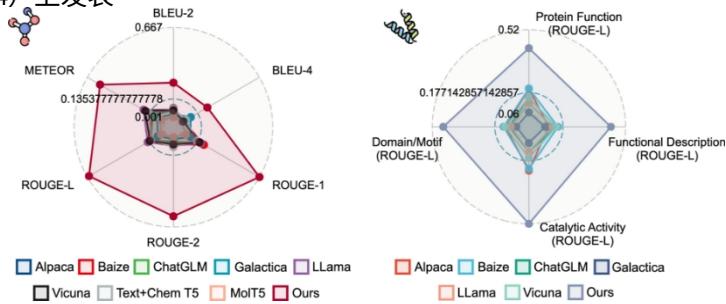


图 5：分子和蛋白质理解任务的性能比较：分子描述生成（左），蛋白质功能、功能描述、催化活性和结构域/特征预测（右）。

表 4：分子生成任务的结果。这些任务包括描述引导的分子设计、试剂预测、正向反应预测和逆合成。

模型	精确性↑	BLEU↑	Levenshtein↓	RDK FTS↑	MACCS FTS↑	MORGAN FTS↑	有效性↑
描述引导的分子设计							
羊驼	0.000	0.004	51.088	0.006	0.029	0.000	0.002
BAIZE	0.000	0.006	53.796	0.000	0.000	0.000	0.002
CHATGLM	0.000	0.004	53.157	0.005	0.000	0.000	0.005
拉马	0.000	0.003	59.864	0.005	0.000	0.000	0.003
维库纳	0.000	0.006	60.356	0.006	0.001	0.000	0.001
卡拉狄加	0.000	0.192	44.152	0.135	0.248	0.088	0.992
文本+化学 T5	0.097	0.508	41.819	0.352	0.474	0.353	0.721
MOLT5	0.112	0.546	38.276	0.400	0.538	0.295	0.773
我们的	0.002	0.345	41.367	0.231	0.412	0.147	1.000
试剂预测							
羊驼	0.000	0.026	29.037	0.029	0.016	0.001	0.186
BAIZE	0.000	0.051	30.628	0.022	0.018	0.004	0.099
CHATGLM	0.000	0.019	29.169	0.017	0.006	0.002	0.074
拉马	0.000	0.003	28.040	0.037	0.001	0.001	0.001
维库纳	0.000	0.010	27.948	0.038	0.002	0.001	0.007
卡拉狄加	0.000	0.141	30.760	0.036	0.127	0.051	0.995
文本+化学 T5	0.000	0.225	49.323	0.039	0.186	0.052	0.313
我们的	0.044	0.224	23.167	0.237	0.364	0.213	1.000
前向反应预测							
羊驼	0.000	0.065	41.989	0.004	0.024	0.008	0.138
BAIZE	0.000	0.044	41.500	0.004	0.025	0.009	0.097
CHATGLM	0.000	0.183	40.008	0.050	0.100	0.044	0.108
拉马	0.000	0.020	42.002	0.001	0.002	0.001	0.039
维库纳	0.000	0.057	41.690	0.007	0.016	0.006	0.059
卡拉狄加	0.000	0.468	35.021	0.156	0.257	0.097	0.946
文本+化学 T5	0.239	0.782	20.413	0.705	0.789	0.652	0.762
我们的	0.045	0.654	27.262	0.313	0.509	0.262	1.000
回溯合成							
羊驼	0.000	0.063	46.915	0.005	0.023	0.007	0.160
BAIZE	0.000	0.095	44.714	0.025	0.050	0.023	0.112
CHATGLM	0.000	0.117	48.365	0.056	0.075	0.043	0.046
拉马	0.000	0.036	46.844	0.018	0.029	0.017	0.010
维库纳	0.000	0.057	46.877	0.025	0.030	0.021	0.017
卡拉狄加	0.000	0.452	34.940	0.167	0.274	0.134	0.986
文本+化学 T5	0.141	0.765	24.043	0.685	0.765	0.585	0.698
我们的	0.009	0.705	31.227	0.283	0.487	0.230	1.000

评估生命科学领域 LLM 生成结果的准确性本身就很复杂。对每项结果都进行专家评审或湿实验室验证既耗时又不切实际。此外，鉴于 LLM 可以对相同的输入产生不同的输出结果，因此确保所有潜在结果的统一准确性是一项艰巨的挑战。虽然我们采用了更广泛领域中普遍接受的指标来评估输出质量，但这些指标只能捕捉到整体情况的一小部分。定量实验可以揭示某些任务的性能，但无法全面应对评估挑战。

如图 5 所示，Mol-Instructions 增强了 LLMs 的分子理解能力，与基线模型（甚至包括特定领域的较小模型）相比，LLMs 在各项指标上都有显著提高。值得注意的是，LLMs 在预测

分子性质方面表现出了令人印象深刻的能力，详见图 3。在这一特定任务中，Alpaca 只为 2.62% 的样本生成了答案，Baize 为 0.62%，LLama 为 54.5%，Vicuna 为 0.14%，Galactica 为 74%，而 ChatGLM 则没有做出回应。关于分子生成任务，如表 4 所示，LLM 具备生成有效分子的能力，与基线相比，这些生成的分子与参考分子的相似度更高。这表明

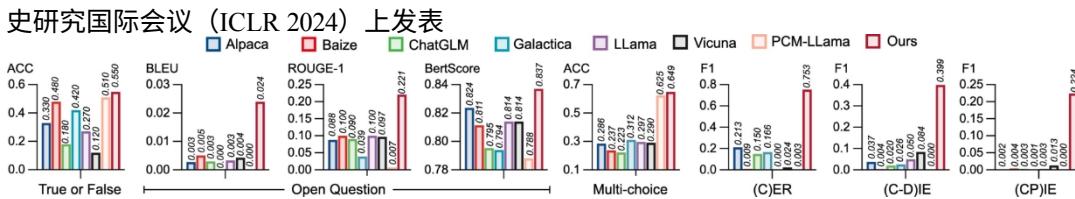


图 6：生物信息 NLP 任务的结果。(C)ER 表示化学实体识别，(C-D)IE 表示化学-疾病相互作用提取，(C-P)IE 表示化学-蛋白质相互作用提取。

在分子生成、化学反应预测和基于特定指令的分子合成方面，Mol-Instructions 为 LLMs 提供了新的见解。然而，与专门的小型模型相比，LLM 的分子生成能力仍然存在明显差距，这主要是因为 LLM 是为处理更广泛的任务而设计的，但却牺牲了重点更突出的模型的专门性能。

如图 5 所示，对于各种蛋白质理解任务，经 Mol-Instructions 调整的 LLM 可根据特定要求分析蛋白质，并在识别蛋白质基本特征方面表现出显著的能力。在蛋白质设计任务中，对生成序列的目标功能特征进行去终端化是一项具有挑战性的实验验证工作。因此，如附录 F 所示，我们采用了一种直接的方法，即使用 BLAST (Camacho 等人, 2009 年) 将生成的蛋白质序列与 UniProtKB 进行比对。我们观察到生成的序列与相应功能区的多个蛋白质之间存在明显的序列同一性。具体来说，我们将蛋白质 (UniProt Accession: A0A518LQL6) 确定为目标蛋白质进行比对，因为它与生成的序列具有最高的序列同一性 (40.9%，p 值为 $7.7e - 30$)。值得注意的是，A0A518LQL6 蛋白和生成的蛋白在 (6S)-NADPH 和金属离子结合区（残基 1-200）具有相似的序列特征。这一发现表明，生成的蛋白质符合设计要求，具有 NADPH 表聚酶活性的潜力。

图 6 显示了模型在理解生物分子文本指令方面的能力。通过为 LLM 注入大量生物分子领域的知识，该模型在所有 NLP 任务中的表现都优于基线。这进一步凸显了 "莫尔指令" 在弥合 LLM 与复杂的生物分子研究之间的差距方面所具有的变革潜力。

5.2 利用摩尔指令的力量

为了优化研究人员对 "莫尔指令" 的利用，我们提出了三个关键方向，以加强对一般模型的探索，推动生物分子理解和药物设计的进步：首先，利用 "莫尔指令" 评估一般模型的跨模态理解能力，从人类语言过渡到生命语言。这些模型应解释用户意图，解码生物分子语言，挑战他们的推理能力。其次，我们的工作为更深入的生物分子设计探索奠定了基础。莫尔指令涵盖了广泛的设计标准，通过提供与生物分子特性预测任务相关的数据，增强了模型对生物分子的理解。第三，将 "莫尔指令" 作为解决复杂生物问题的工具学习的重要数据。尽管研究强调了具有基础模型的专业工具的优势 (Qin 等人, 2023 年)，但在文本说明的可用性方面还存在差距，而我们的数据集正好解决了这一问题。

6 结论和未来工作

在这项工作中，我们介绍了 Mol-Instructions，这是一个专门为生物分子研究而策划的综合指导数据集，它弥补了现有资源的不足，并推动了这一专业领域的法律硕士培训。展望未

史研究国际会议 (ICLR 2024) 上发表。我们将根据化学研究的最新进展和人工智能技术的改进，纳入更广泛的任务类型、教学条目和模式，以满足更广泛的化学研究需求和更高层次的法律硕士培训要求。此外，由于文本和生物分子的表示空间不同，再加上 LoRA 训练策略的限制，目前的 LLM 还不能像掌握人类语言那样熟练地掌握生物分子语言。探索扩大词汇量的方法，或通过生物分子编码器将生物语言作为一种模式 (Fang 等人, 2024 年; Rives 等人, 2021 年; Lin 等人, 2022 年; Cao 等人, 2023 年; Pei 等人, 2024 年)，对于磨练模型在生物分子任务中的理解力和性能至关重要。

致谢

感谢匿名审稿人的宝贵意见。本研究得到了国家自然科学基金 (No. 62206246)、中央高校基本科研业务费 (226-2023-00138)、浙江省自然科学基金 (No.LGG22F030011)、宁波市自然科学基金 (2021J190)、CAAI-华为MindSpore开放基金、甬江人才引进计划 (2021A-156-G)、CCF-百度开放基金、浙江大学信息技术中心和CAD&CG国家重点实验室等的资助。

可重复性声明

所有数据、代码和模型权重可在 GitHub¹ 和抱抱脸^{2,3,4,5}。有关数据集构建过程的详细说明，请参阅附录 B。具体实验设置请参见附录 D 和 E。

道德规范声明

本研究严格按照研究伦理准则和最佳实践进行。所使用的生物分子数据均来自公开数据集，没有使用任何专有或保密数据。此外，我们还为数据集中的所有第三方内容获得了必要的许可和授权，详见附录 A.2。

我们实施了严格的质量控制措施和安全检查，以防止数据集中出现任何有害或恶意内容。不过，我们也认识到将 LLM 与生物分子知识结合起来的深远影响和潜在风险。虽然我们的主要目的是促进科学理解并为社会做出积极贡献，但我们也清楚地意识到，这些工具一旦落入坏人之手，就有可能被滥用。恶意行为者有可能利用 LLMs 和生物分子数据的综合能力制造有害物质，如生化武器或非法药物。

我们强烈要求所有用户在使用我们的数据集时遵守最高道德标准，确保研究的公平性、透明度和责任感。严禁任何可能对社会造成伤害或危害的数据集使用行为。

参考资料

Sharegpt, 2023 年 4 月。URL <https://sharegpt.com/>

Yuvanesh Anand、Zach Nussbaum、Brandon Duderstadt、Benjamin Schmidt 和 Andriy Mulyar。Gpt4all：利用 gpt-3.5-turbo 的大规模数据提炼，训练助理式聊天机器人。GitHub，2023 年。

Michael Ashburner、Catherine A Ball、Judith A Blake、David Botstein、Heather Butler、J Michael Cherry、Allan P Davis、Kara Dolinski、Selina S Dwight、Janan T Eppig 等。基因本体：统一生物学的工具。《自然遗传学》，25 (1) : 25-29, 2000 年。

Satanjeev Banerjee 和 Alon Lavie.METEOR：一种用于 MT 评估的自动度量标准，可提高与人

史研究国际会议 (ICLR 2024) 上发表
类判断的相关性。见 Jade Goldstein、Alon Lavie、Chin-Yew Lin 和 Clare R. Voss (编辑) ,
机器翻译和/或总结的内在和外在评估措施研讨会论文集@ACL 2005，美国密歇根州安娜
堡市，2005 年 6 月 29 日，第 65-72 页。计算语言学协会，2005 年。url
<https://aclanthology.org/w05-0909/>.

¹GitHub: <https://github.com/zjunlp/Mol-Instructions>

²数据集: <https://huggingface.co/datasets/zjunlp/Mol-Instructions>

3分子模型: <https://huggingface.co/zjunlp/llama-molinst-molecule-7b>4蛋白质模型: <https://huggingface.co/zjunlp/llama-molinst-protein-7b>5Biotext 型号: <https://huggingface.co/zjunlp/llama-molinst-biotext-7b>

Daniil A. Boiko、Robert MacKnight 和 Gabe Gomes。大型语言模型的新兴自主科研能力。

CoRR , abs/2304.05332 , 2023 。 DOI : 10.48550/ARXIV.2304.05332.URL <https://doi.org/10.48550/arXiv.2304.05332>。

Andres M Bran、Sam Cox、Andrew D White 和 Philippe Schwaller。Chemcrow：用化学工具
增强大型语言模型。*arXiv 预印本 arXiv:2304.05376*, 2023.

Christiam Camacho、George Coulouris、Vahram Avagyan、Ning Ma、Jason S. Papadopoulos
、Kevin Bealer 和 Thomas L. Madden。BLAST+: architecture and applications.10:421, 2009.
DOI: 10.1186/1471-2105-10-421.url <https://doi.org/10.1186/1471-2105-10-421>。

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li.Instructmol：多模态集成，打造药物发现
领域的多功能可靠分子助手。DOI : 10.48550/ARXIV.2311.16208 。 URL
<https://doi.org/10.48550/arXiv.2311.16208>.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde' de Oliveira Pinto, Jared
Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri,
Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan,
Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power、Lukasz Kaiser、Mohammad Bavarian
、Clemens Winter、Philippe Tillet、Felipe Petroski Such、Dave Cummings、Matthias
Plappert、Fotios Chantzis、Elizabeth Barnes、Ariel Herbert-Voss、William Heben Guss、
Alex Nichol、Alex Paino、Nikolas Tezak、Jie Tang、Igor Babuschkin、Suchir Balaji、
Shantanu Jain、William Saunders、Christopher Hesse、Andrew N. Carr、Jan Leike、Joshua
Achiam、Vedant Misra、Evan Morikawa、Alec Radford、Matthew Knight、Miles Brundage
、Mira Murati、Katie Mayer、Peter Welinder、Bob McGrew、Dario Amodei、Sam
McCandlish、Ilya Sutskever 和 Wojciech Zaremba。评估基于代码训练的大型语言模型。
CoRR, abs/2107.03374, 2021。URL <https://arxiv.org/abs/2107.03374>。

Wei-Lin Chiang、Zuoohan Li、Zi Lin、Ying Sheng、Zhanghao Wu、Hao Zhang、Lianmin
Zheng、Siyuan Zhuang、Yonghao Zhuang、Joseph E. Gonzalez、Ion Stoica 和 Eric P. Xing
。Vicuna：一个开源聊天机器人给 gpt-4 留下深刻印象，聊天质量达到 90%*，2023 年 3
月。URL <https://lmsys.org/blog/2023-03-30-vicuna/>。

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam
Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh,
Kensen Shi、Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay,
Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope,
James Bradbury、Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke,
Anselm Lev-skaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia,
Vedant Misra, Kevin Robinson,Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan,
Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani
Agrawal , Mark Omernick , Andrew M . Dai, Thanumalayan Sankaranarayana Pillai, Marie
Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee,
Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason

史研究国际会议 (ICLR 2024) 上发表
Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel.掌上电脑：

用路 径 扩 展 语 言 建 模 。 *J. Mach.Learn.*24:240:1-240:113 , 2023 。 URL
<http://jmlr.org/papers/v24/22-1144.html>。

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 通过多任务语言建模统一分子和文本表征。In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp.PMLR, 2023.网址
<https://proceedings.mlr.press/v202/christofidellis23a.html>。

Mike Conover、Matt Hayes、Matt Mathur、孟祥瑞、谢建伟、万军、Ali Ghodsi、Patrick Wendell 和 Patrick Zaharia。你好，多莉：用开放模型民主化聊天工具的魔力》，2023 年
。 URL <https://github.com/databrickslabs/dolly>.

UniProt 联盟。Uniprot：2023 年的通用蛋白质知识库。*Nucleic Acids Res.*, 51(D1):523-531, 2023. doi: 10.1093/NAR/GKAC1052.URL <https://doi.org/10.1093/NAR/GKAC1052>。

Justas Dauparas、Ivan Anishchenko、Nathaniel Bennett、Hua Bai、Robert J Ragotte、Lukas F Milles、Basile IM Wicky、Alexis Courbet、Rob J de Haas、Neville Bethel 等。使用 proteinmpnn 进行基于深度学习的稳健蛋白质序列设计。*科学》，378 (6615)：49-56，2022。*

丁宁、陈玉林、徐博凯、秦宇佳、胡圣鼎、刘志远、孙茂松和周博文。通过扩展高质量教学对话增强聊天语言模型。见 Houda Bouamor、Juan Pino 和 Kalika Bali (编)，《2023 年自然语言处理经验方法会议论文集》，EMNLP 2023，新加坡，2023 年 12 月 6-10 日，第 3029-3051 页。计算语言学协会，2023 年。URL <https://aclanthology.org/2023.emnlp-main.183>.

Joseph L. Durant、Burton A. Leland、Douglas R. Henry 和 James G. Nourse。用于药物发现的 MDL 密钥再优化。*J. Chem.Inf.Comput.* DOI: 10.1021/CI010132R.URL <https://doi.org/10.1021/ci010132r>.

Carl Edwards、Tuan Manh Lai、Kevin Ros、Garrett Honke、Kyunghyun Cho 和 Heng Ji。分子与自然语言之间的翻译。In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp.DOI: 10.18653/V1/2022.EMNLP-MAIN.26.URL <https://doi.org/10.18653/v1/2022.emnlp-main.26>.

Yin Fang, Ningyu Zhang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 具有自我反馈功能的领域定向分子生成。In ICLR.OpenReview.net, 2024.URL <https://openreview.net/pdf?id=9rPyHyjfWp>。

Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。测量大规模多任务语言理解。*第九届学习表征国际会议 (ICLR 2021)*，奥地利虚拟活动，2021 年 5 月 3-7 日。OpenReview.net, 2021.URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 训练计算最优的大型语言模型。*CoRR*，abs/2203.15556，2022。DOI : 10.48550/ARXIV.2203.15556。URL <https://doi.org/10.48550/arXiv.2203.15556>.

Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A. Shoemaker、Paul A. Thiessen、Bo Yu、Leonid Zaslavsky、Jian Zhang 和 Evan Bolton。2021 年的 Pubchem：新的数据内容和改进的网络界面。*Nucleic Acids Res.*, 49(Database-Issue) : DOI: 10.1093/NAR/GKA971.URL <https://doi.org/10.1093/nar/gkaa971>.

Martin Krallinger、Obdulia Rabal、Florian Leitner、Miguel Vazquez、David Salgado、Zhiyong Lu、Robert Leaman、Yanan Lu、Donghong Ji、Daniel M. Lowe、Roger A. 赛尔、

史研究国际会议 (ICLR-2024) 上发表。
里扎-特雷莎-巴蒂斯塔-纳瓦罗、拉法尔-拉克、托斯滕-胡贝尔、蒂姆-罗克塔谢尔、塞尔
吉奥-马托斯、戴维-坎波斯、唐步洲、徐华、曾德仁-蒙赫达莱、柳坤浩、S. V. 拉马南、
P. 森蒂尔-纳坦、斯拉夫克-齐特尼克、马尔科-巴耶茨、鲁茨-韦伯、马蒂亚斯-伊尔默、萨
博-A.Akhondi, Jan A. Kors, 徐硕, 安昕, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka,
Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu,
Ravikumar Komandur Elayavilli, Andre Lamurias, Francisco M. Couto, Hong-Jie D., J. M. D.,
J. M. M., J. M. Couto、Hong-Jie Dai、
Richard Tzong-Han Tsai、Caglar Ata、Tolga Can、Anabel Usie、Rui Alves、Isabel Segura-
Bedmar、Paloma Martínez、Julen Oyarzabal 和 Alfonso Valencia。化学品和药物的
CHEMDNER 语料库及其注释原则。*J. Cheminformatics*, 7(S-1):S2, 2015. doi: 10.1186/1758-
2946-7-S1-S2.url <https://doi.org/10.1186/1758-2946-7-S1-S2>.

Martin Krallinger、Obdulia Rabal、Saber Ahmad Akhondi、Martín Pérez Pérez、Jesus
Santamaría、Gael Pérez Rodríguez、Georgios Tsatsaronis、Ader Intxaurreondo、Jose'
Antonio Baso López Umesh K.Nandal, Erin M. van Buel, Ambika Chandrasekhar, Marleen
Rodenburg, Astrid Lægreid, Marius A. Doornenbal, Julen Oyarzabal, Ana'lia Lourenço, and
Alfonso Valencia.生物创造性六化学蛋白相互作用轨道概述。2017.

马里奥-克伦、艾谦祥、森雅-巴特尔、内萨-卡森、安杰洛-弗雷、内森-C-弗雷、帕斯卡尔-弗里德里希、西奥菲尔-高丹、阿尔贝托-亚历山大-盖尔、凯文-迈克-雅布隆卡、拉斐尔-F.Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José' Manuel Na'poles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Ala'n Aspuru-Guzik. SELFIES 和分子弦表征的未来。DOI: 10.1016/J.Patter.2022.100588.URL <https://doi.org/10.1016/j.patter.2022.100588>.

Greg Landrum et al: 用于化学信息学、计算化学和预测建模的软件套件。Greg Landrum, 8:31, 2013.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR 任务语料库：化学疾病关系提取资源。Database J. Biol.Databases Curation, 2016, 2016. doi: 10.1093/DATABASE/BAW068.URL <https://doi.org/10.1093/database/baw068>.

李宇健和刘碧。归一化的莱文斯坦距离度量。IEEE Trans.Pattern Anal.Mach.29(6):1091-1095, 2007. doi: 10.1109/TPAMI.2007.1078.URL <https://doi.org/10.1109/TPAMI.2007.1078>.

Chin-Yew Lin.Rouge：自动评估摘要的软件包。In *Text summarization branches out*, pp.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. 进化尺度上的蛋白质序列语言模型实现了准确的结构预测。BioRxiv, 2022.

刘胜超、聂伟力、王成鹏、陆家瑞、乔卓然、刘玲、唐健、肖超伟、Animashree Anandkumar。基于文本检索和编辑的多模态分子结构-文本模型。Nat.Mac.DOI: 10.1038/S42256-023-00759-6。URL <https://doi.org/10.1038/s42256-023-00759-6>.

Jieyu Lu 和 Yingkai Zhang. 多任务反应预测的统一深度学习模型与解释。J. Chem.Inf.DOI: 10.1021/ACS.JCIM.1C01467.URL <https://doi.org/10.1021/acs.jcim.1c01467>.

Ali Madani、Ben Krause、Eric R Greene、Subu Subramanian、Benjamin P Mohr、James M Holton、Jose Luis Olmos Jr、Caiming Xiong、Zachary Z Sun、Richard Socher 等。大型语言模型生成跨不同家族的功能性蛋白质序列。自然生物技术》，第 1-8 页，2023 年。

卡伊克-蒙泰罗-卡斯特罗-纳西门托和安德烈-席尔瓦-皮门特尔。大型语言模型懂化学吗？与 chatgpt 的对话。J. Chem.Inf.DOI: 10.1021/ACS.JCIM.3C00285.URL <https://doi.org/10.1021/acs.jcim.3c00285>.

OpenAI.GPT-4 技术报告。DOI : 10.48550/ARXIV.2303.08774 。 URL <https://doi.org/10.48550/arXiv.2303.08774>.

OpenAI.Gpt-3.5-turbo。<https://platform.openai.com/docs/models/gpt-3.5-turbo>, 2023b。

欧阳龙、杰弗里-吴、蒋旭、迪奥戈-阿尔梅达、卡罗尔-L-温莱特、帕梅拉-米什金、张冲、桑迪尼-阿加瓦尔、卡塔琳娜-斯拉玛、亚历克斯-雷、约翰-舒尔曼、雅各布-希尔顿、弗雷

史研究国际会议 (ICLR 2024) 上发表。
泽-凯尔顿、卢克-米勒、玛迪-西门斯、阿曼达-阿斯凯尔、彼得-韦林德、保罗-F-克里斯-
蒂亚诺、扬-雷克和瑞安-洛。训练语言模型，使其听从人类反馈指令。见 Sanmi Koyejo、
S. Mohamed、A. Agarwal、Danielle Belgrave、K. Cho 和 A. Oh (编辑)，《神经信息处
理系统进展 35：2022 年神经信息处理系统年会》，NeurIPS 2022，美国洛杉矶新奥尔良
，11 月 28 日

- 2022 年 12 月 9 日，2022a。URL
[http://papers.nips.cc/paper_files/paper/2022/
hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html)。

欧阳龙、杰弗里-吴、蒋旭、迪奥戈-阿尔梅达、卡罗尔-L-温莱特、帕梅拉-米什金、张冲、桑迪尼-阿加瓦尔、卡塔琳娜-斯拉玛、亚历克斯-雷、约翰-舒尔曼、雅各布-希尔顿、弗雷泽-凯尔顿、卢克-米勒、玛迪-西门斯、阿曼达-阿斯凯尔、彼得-韦林德、保罗-F-克里斯蒂亚诺、扬-雷克和瑞安-洛。训练语言模型，使其听从人类反馈指令。见 Sanmi Koyejo、S. Mohamed、A. Agarwal、Danielle Belgrave、K. Cho 和 A. Oh（编辑），《神经信息处理系统进展 35：2022 年神经信息处理系统年会》，*NeurIPS 2022*，美国洛杉矶新奥尔良，11 月 28 日

- 2022 年 12 月 9 日, 2022b。URL

http://papers.nips.cc/paper_files/paper/2022/

hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html。

Ankit Pal、Logesh Kumar Umapathi 和 Malaikannan Sankarasubbu。Medmcqa：用于医学领域问题解答的大规模多主体多选择数据集。In Gerardo Flores, George H. Chen, Tom J. Pollard, Joyce C. Ho, and Tristan Naumann (eds.), *Conference on Health, Inference, and Learning, CHIL 2022, 7-8 April 2022, Virtual Event, volume 174 of Proceedings of Machine Learning Research*, pp.PMLR, 2022.URL <https://proceedings.mlr.press/v174/pal22a.html>。

Kishore Papineni、Salim Roukos、Todd Ward 和 Wei-Jing Zhu。Bleu：机器翻译自动评估方法。In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp.DOI：10.3115/1073083.1073135。URL <https://aclanthology.org/P02-1040/>。

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan.Biot5+：通过iupac集成和多任务调整实现广义生物理解。

秦宇佳、胡圣鼎、林彦凯、陈伟泽、丁宁、崔干渠、曾振妮、黄宇飞、肖超俊、韩驰、冯懿人、苏玉生、王华东、钱程、田润初、朱昆仑、梁世豪、沈星宇、徐博凯、张震、叶一宁、李博文、唐紫薇、易静、朱玉章、戴振宁、严岚、丛昕、卢亚茜、赵伟林、黄宇翔、闫俊熙、韩旭、孙娴、李大海、庞杰、杨成、吴同双、季恒、刘志远和孙茂松。使用基础模型的工具学习 DOI : 10.48550/ARXIV.2304.08354 。 URL <https://doi.org/10.48550/arXiv.2304.08354>。

Alexander Rives、Joshua Meier、Tom Sercu、Siddharth Goyal、Zeming Lin、Jason Liu、Demi Guo、Myle Ott、C. Lawrence Zitnick、Jerry Ma 和 Rob Fergus。将无监督学习扩展到 2.5 亿个蛋白质序列可发现生物结构和功能。*Proc.Natl.USA*, 118(15):e2016239118, 2021. doi: 10.1073/PNAS.2016239118.URL <https://doi.org/10.1073/pnas.2016239118>.

Jonas B. Sandbrink.人工智能与生物滥用：区分语言模型和生物设计工具的风险.*CoRR*, abs/2306.13952 , 2023 。 DOI : 10.48550/ARXIV.2306.13952.URL <https://doi.org/10.48550/arXiv.2306.13952>.

Victor Sanh、Albert Webson、Colin Raffel、Stephen H. Bach、Lintang Sutawika、Zaid

史研究国际会议 (ICLR 2024) 上发表
Alyafeai、Antoine Chaffin、Arnaud Stiegler、Arun Raja、Manan Dey、M Saiful Bari、
Canwen Xu、Urmish Thakker、Shanya Sharma Sharma、Eliza Szczechla、Taewoon Kim、
Gunjan Chhablani、Nihal V。Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang,
Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden,
Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fe'vry,
Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and
Alexander M. Rush.拉什多任务提示训练实现零镜头任务泛化。第十届学习表征国际会议
, ICLR 2022 , 虚拟活动 , 2022 年 4 月 25-29 日。OpenReview.net, 2022.URL
<https://openreview.net/forum?id=9Vrb9D0WI4>.

Nadine Schneider、Roger A. Sayle 和 Gregory A. Landrum。让你的原子井然有序--一种新颖
稳健的分子规范化算法的开源实现。J. Chem. Inf. DOI: 10.1021/ACS.JCIM.5B00543.URL
<https://doi.org/>
[10.1021/acs.jcim.5b00543.](https://doi.org/10.1021/acs.jcim.5b00543)

罗伯特-F-服务聊天机器人能否帮助设计出下一个流行病毒? 科学》(纽约州纽约市), 380 (6651) : 1211, 2023。

Martin Steinegger 和 Johannes Soeding. Mmseqs2 实现了用于分析海量数据集的灵敏蛋白质序列搜索。自然·生物技术》, 35 (11) : 1026-1028, 2017。

苏兵、杜大钊、杨钊、周玉洁、李江萌、饶安义、孙浩、吕志武、温继荣。分子图谱与自然语言关联的分子多模态基础模型。DOI: 10.48550/ARXIV.2209.05481。URL <https://doi.org/10.48550/arXiv.2209.05481>.

Taffee T Tanimoto. 分类和预测的初等数学理论》。1958.

Rohan Taori、Ishaan Gulrajani、Tianyi Zhang、Yann Dubois、Xuechen Li、Carlos Guestrin、Percy Liang 和 Tatsunori B Hashimoto。斯坦福羊驼遵循指令的骆驼模型。GitHub 存储库, 2023 年。

Ross Taylor、Marcin Kardas、Guillem Cucurull、Thomas Scialom、Anthony Hartshorn、Elvis Saravia、Andrew Poulton、Viktor Kerkez 和 Robert Stojnic。银河系: 大型科学语言模型。DOI : 10.48550/ARXIV.2211.09085 。 URL <https://doi.org/10.48550/arXiv.2211.09085>.

Robert Tinn、Hao Cheng、Yu Gu、Naoto Usuyama、Xiaodong Liu、Tristan Naumann、Jianfeng Gao 和 Hoifung Poon。微调用于生物医学自然语言处理的大型神经语言模型。DOI: 10.1016/J.PATTER.2023.100729. URL <https://doi.org/10.1016/j.patter.2023.100729>.

Tloen. <https://github.com/tloen/alpaca-lora>, 2023.

Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothe'e Lacroix、Baptiste Rozie're、Naman Goyal、Eric Hambro、Faisal Azhar、Aurelien Rodriguez、Armand Joulin、Edouard Grave 和 Guillaume Lample。Llama: 开放而高效的基
础语言模型。DOI: 10.48550/ARXIV.2302.13971. URL <https://doi.org/10.48550/arXiv.2302.13971>.

王瀚辰、傅天凡、杜源琦、高文浩、黄可欣、刘子明、Payal Chandak、刘胜超、Peter Van Katwyk、Andreea Deac、Anima Anandkumar、Karianne Bergen、Carla P. Gomes、Shirley Ho、Pushmeet Kohli、Joan Lasenby、Jure Leskovec、刘铁彦、Arjun Manrai、Debora S. Marks。Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora S. Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Velickovic, Max Welling, Linfeng Zhang, Connor W. Coley, Yoshua Bengio, and Marinka Zitnik. 人工智能时代的科学发现。doi: 10.1038/S41586-023-06221-2. URL <https://doi.org/10.1038/s41586-023-06221-2>.

王义忠、耶加内-科尔迪、斯瓦罗普-米什拉、艾丽莎-刘、诺亚-史密斯、丹尼尔-卡沙比、汉娜-哈吉希尔兹。自我指导: 将语言模型与自我生成的指令对齐。In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the*

史研究国际会议 (ICLP 2024) 上发表
Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023, pp. 计算语言学协会 , 2023b. doi: 10.18653/V1/2023.ACL-LONG.754。 URL <https://doi.org/10.18653/v1/2023.acl-long.754>。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai 和 Quoc V. Le。微调语言模型是零点学习者。In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.OpenReview.net, 2022.URL <https://openreview.net/forum?id=gEZrGCozdqR>.

Jin-Mao Wei, Xiao-Jie Yuan, Qinghua Hu, and Shu-Qin Wang.评估分类器的新方法。专家系统
。DOI: 10.1016/J.ESWA.2009.11.040.URL <https://doi.org/10.1016/j.eswa.2009.11.040>.

戴维 - 魏 宁 格 微 笑 , 一 种 化 学 语 言 和 信 息 系 统 。方 法 和 编 码 规 则 介 绍 。*J. Chem.Inf.Comput.*DOI: 10.1021/CI00057A005.URL <https://doi.org/10.1021/ci00057a005>.

雅各布-怀特Pubmed 2.0.《医学参考服务季刊》, 39 (4) : 382-387, 2020。

吴超毅、张晓曼、张雅、王彦峰、谢伟迪。Pmc-llama：进一步微调医学论文中的llama。

DOI: 10.48550/ARXIV.2304.14454.URL <https://doi.org/10.48550/arXiv.2304.14454>.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande.Moleculenet：分子机器学习的基准。《化学科学》, 9 (2) : 513-530, 2018。

Canwen Xu、Daya Guo、Nan Duan 和 Julian J. McAuley。Baize：基于自聊天数据的参数高效调整开源聊天模型。In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 计算语言学协会, 2023 年。URL <https://aclanthology.org/2023.emnlp-main.385>.

曾翱涵、刘晓、杜正晓、王梓涵、赖瀚宇、丁明、杨卓一、徐一帆、郑文迪、夏晓、谭翁林、马子轩、薛宇飞、翟继东、陈文广、刘志远、张鹏、董玉晓、唐杰。GLM-130B：开放式双语预训练模型。第十一届国际学习表征会议 (ICLR 2023), 卢旺达基加利, 2023 年 5 月 1-5 日。OpenReview.net, 2023.URL <https://openreview.net/pdf?id=Aw0rrrPUF>。

曾哲妮、姚远、刘志远和孙茂松。连接分子结构和生物医学文本的深度学习系统，其理解能力可媲美人类专业人士。《自然通讯》, 13 (1) : 862, 2022。

张革、石跃民、刘瑞波、袁瑞斌、李一之、董思伟、舒宇、李兆群、王泽坤、林成华、黄文浩、傅杰。中文开放式教学通论：初步发布。DOI: 10.48550/ARXIV.2304.07987。URL <https://doi.org/10.48550/arXiv.2304.07987>.

张金路、方寅、邵欣、陈华军、张宁宇、范晓辉。大语言模型透视分子研究的未来。J. *Chem.Inf.Model*: doi: 10.1021/ACS.JCIM.3C01977.URL <https://doi.org/10.1021/acs.jcim.3c01977>。

Weitong Zhang, Xiaoyun Wang, Weili Nie, Joe Eaton, Brad Rees, and Quanquan Gu.Moleculegpt：用于分子性质预测的大型语言模型指令。In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*, 2023b.

A 访问和使用摩尔指令

A.1 托管和访问详情

我们的数据集和相关模型安全地托管在 GitHub 和 Hugging Face 上，它们是公认的开源项目管理平台。它们为大量数据和代码库提供了广泛的可访问性和高效的管理，保证所有潜在用户都能畅通无阻地访问。为了有效利用我们的资源，我们在资源库上提供了全面的指南和说明。其中包括如何探索数据集、了解其结构和使用模型。

A.2 数据源和许可证

如表 5 所示，我们详细说明了用于构建 Mol-Instructions 数据集的所有数据组件（包括生物分子和文本描述）的来源和数据权限。我们对所有数据来源进行了严格审查，以确保它们的许可证允许我们的研究和后续使用。在整篇文章中，每一次提及或使用这些数据源时，都会正确无误地注明出处。

A.3 使用指南和义务

我们声明，本作品中包含的所有数据均符合 CC BY 4.0 许可的规定。我们承认，作为许可人，我们有责任促进数据集的开放和公平使用，同时承认创作者的贡献。我们接受与执行本许可相关的所有义务，并承诺协助所有潜在用户了解他们在本许可下的权利和责任。

我们保证我们的数据集不包含任何个人身份信息或隐私敏感信息。我们实施了严格的质量控制程序和安全检查，以防止包含有害或恶意内容。不过，必须指出的是，虽然我们已采取大量措施确保数据集的安全和隐私，但我们不对使用数据集可能产生的任何问题承担责任。

我们强烈要求所有用户在使用我们的数据集时遵守最高道德标准，包括在研究中保持公平、透明和负责。严禁任何可能导致伤害或危害社会的数据集使用行为。

在数据集维护方面，我们承诺提供必要的维护。这将确保数据集的持续相关性和可用性，以适应不断变化的研究环境。这一承诺包括定期更新、错误检查以及根据领域进展和用户反馈进行修正。

B 任务定义和数据构建

B.1 面向分子的指令

分子描述生成 分子描述生成是指根据分子描述符创建详细的文字描述，阐明分子的结构、性质、生物活性和应用。它为化学家和生物学家提供了获取重要分子信息的快捷渠道，从而有效地指导他们的研究和实验。

史研究国际会议 (ICLR-2024) 上发表。为了收集大量的分子文本注释数据，我们选择了 PubChem (Kim 等人, 2021 年) 作为数据源。PubChem 是美国国家生物技术信息中心 (NCBI) 管理的一个可免费访问的数据库，是化学研究的宝贵资源。PubChem 中的许多化合物都有文字说明，这些文字说明来自各研究机构的直接提交，以及科学文献和专利中的自动数据挖掘和信息提取。

首先，我们从 PubChem 的 Power User Gateway 开始，它以 XML 格式提供 PubChem 化合物记录的摘要，便于高效检索和处理化学信息。我们抓取所有有效的分子描述文本以及唯一的 PubChem 化合物记录摘要。

表 5：本文涉及的数据资源和许可证。

数据源	许可证 URL	许可证注释
PubChem	https://www.nlm.nih.gov/web_policies.html	美国政府制作的作品在美国不受版权保护。美国国家医学图书馆 (NLM) 网站上的任何此类作品在美国均可自由使用或复制，无需获得许可。
美国专利商标局	https://www.uspto.gov/学习与资源/开放数据与流动性	任何人都可以自由使用、重复使用和重新发布。
UniProtKB	https://www.uniprot.org/帮助/许可证	您可以自由共享--以任何媒介或形式复制和重新分发资料。改编--出于任何目的，甚至是商业目的，对资料进行混音、改造和扩充。任务数据现在免费提供给再搜索社区。
BC4CHEMD, ChemProt, BC5CDR	https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/	特此允许任何获得本软件及相关文档文件（以下简称 "软件"）副本的人无限制地使用本软件，包括但不限于使用、复制、修改、合并、出版、分发、再许可和/或出售本软件副本的权利，以及允许获得本软件的人这样做。
分子网、MMLU、PubMedQA、MedMCQA	https://opensource.org/licenses/MIT	本网站使用 AddThis 服务，允许访问者在各种社交媒体网站上收藏和分享网站内容。使用 AddThis 服务 共享内容的访问者无需注册或提供任何个人信息。
有毒物质和疾病登记署 (ATSDR)	https://www.cdc.gov/Other/disclaimer.html	除非另有说明，否则 FDA 网站 (www.fda.gov) 的内容（包括文本和图形）不受版权保护。这些内容属于公共领域，任何人都可以自由翻印、转载或以其他方式使用，而无需获得美国食品药品管理局的许可。请注明来源于美国食品药品管理局，但不是必须的。
FDA 药学课程	https://www.fda.gov/about-fda/about-website/website-policies	美国政府制作的作品在美国不受版权保护。除非另有说明，HIV.gov 网站上提供的资料被视为联邦政府信息，属于公共领域。这意味着可以自由复制和分发这些信息。
肝脏毒素	https://www.nlm.nih.gov/licenses.html	根据本出版许可的条款和条件，许可人特此授予您全球性、免版税、不可再许可、非排他性、不可撤销的许可，允许您行使许可材料中的许可权利，以便：仅为非商业目的复制和共享全部或部分许可材料；仅为非商业目的制作、复制和共享改编材料。
药物数据库, Clinicalinfo.hiv.gov	https://www.hiv.gov/关于我们/使命和团队	您可以自由共享--以任何媒介或形式复制和重新分发资料。改编--出于任何目的，甚至是商业目的，对资料进行混音、改造和扩充。
Drug Bank	https://creativecommons.org/licenses/by-nc/4.0/legalcode	YMDB 作为免费资源向公众提供。
ChEBI	https://creativecommons.org/licenses/by/4.0/	LOTUS 是最大、最完善的注释数据库之一。免费提供天然产品出现的来源，没有任何限制。
酵母代谢组数据库 (YMDB)	http://www.ymdb.ca/下载	
LOTUS - 天然产品出现数据库	https://lotus.nprod.net/	
GlyCosmos Glycoscience	https://glycosmos.org/Portalhttps://glycosmos.org/ 许可证	
CAMEO 化学品	https://cameochemicals.noaa.gov/help/reference/terms_and_conditions.htm?	

d
—
f
=
f
a
l
s
e

您可以在任何媒介或形式复制和重新分发资料。改编--出于任何目的，甚至是商业目的，对资料进行混音、改造和扩充。

大肠杆菌代谢组数据库 (ECMDB)

<https://ecmdb.ca/citations> ECMDB 向公众免费提供。
资源。

CAMEO 化学品和所有其他 CAMEO 产品 免费提供给负责化学品安全处理的组织和个人 (接收者)。

其次，考虑到 PubChem 提供的 SMILES 可能并不总是准确的，我们过滤掉了那些包含有 "SMILES" 字样的 SMILES。其次，鉴于 PubChem 提供的 SMILES 并不总是准确的，我们过滤掉了包含以下内容的 SMILES

语法错误或违反基本化学原理，并将所有有效的 SMILES 转换为 SELFIES 格式。请注意，我们在每个任务中都会执行这一步骤，后续章节将不再赘述。第三，为了迫使模型关注描述的语义，我们效仿[爱德华兹等人（2022 年）](#)的做法，在每个描述中用 "该分子是....." 替换分子的名称。最后，我们调用 gpt-3.5-turbo 来通过提示生成一系列不同的任务描述，然后将其随机分配给每个 SELFIES 描述对。我们总共编译了 331261 条指令。

描述引导下的分子生成 基于描述的分子生成的意义在于，它可以生成直接符合给定描述中列出的标准的分子，从而简化分子设计过程。这有助于更有针对性地创造和优化新型分子，并应用于药物发现和材料科学等多个领域。

在这一阶段，我们将重新利用在前一任务中从 PubChem 收集到的 SELFIES 描述对。与之前不同的是，在本例中，指令条目以分子描述为输入，以分子的 SELFIES 表示为输出。随后，我们制作了一系列任务描述，用于在描述引导下生成分子，作为指令。这样就汇编了 331,261 个指令数据项。

前向反应预测 前向反应预测是指根据特定的反应物和试剂，预先确定化学反应的可能产物。这有助于优化研发方法，减少实验猜测，并通过减少废物产生来支持更环保的化学实践。

为了收集高质量的化学反应数据，我们重点关注美国专利商标局的数据集（[Wei 等人，2010 年](#)）。该数据集包含大量从美国专利和专利申请中提取的 SMILES 格式的有机化学反应。一般格式为 "反应物 > 试剂 > 产物"。

通常，试剂被定义为不出现在主要产品中的化学物质。然而，为了模拟非化学专业人士的真实情况，我们承担了一项更具挑战性的任务，即不将试剂和反应物分开，而是要求模型独立识别它们。经过预处理后，我们获得了 138 768 个标准指令数据项。在每个指令条目中，输入表示反应中的反应物和试剂，单个化学品之间用句号（'.'）隔开。相反，输出则表示该反应的产物。

逆合成 逆向合成分析是有机化学中一种重要的合成方法，它采用逆向工程方法，从目标化合物出发，追溯潜在的合成路线和前体分子。事实证明，这项技术有助于为复杂的分子制定高效的合成策略，从而推动新型药物和材料的发展和进步。

在这项任务中，我们只专注于单步逆合成。我们收集的数据来自人工处理的美国专利商标局 500MT 数据集（[Lu & Zhang, 2022 年](#)），该数据集本身是从美国专利商标局数据库（[Wei 等人，2010 年](#)）中精心提炼出来的。通过预处理，我们获得了 143,536 个标准指令条目。在每个条目中，输入是产品，输出是反应物，每个反应物之间用句号（'.'）隔开。

试剂预测 试剂预测旨在确定特定化学反应所需的合适催化剂、溶剂或辅助物质。这项工作

有助于化学家发现新的反应类型和机理，确定更优化或更环保的反应条件，并最终简化整个化学过程，以实现最大的成本效益和环境管理。

与回溯合成一样，这项任务的数据也来自美国专利商标局的 500K 数据集（[Lu 和 Zhang, 2022 年](#)）。经过处理后，我们得到了大约 138,768 个条目。在每个指令条目中，输入是从反应物到产物的化学反应，格式为 "反应物 >> 产物"。预期输出是促进该反应的潜在试剂。

性质预测 性质预测是指根据分子结构特征所提供的信息，预测或估计分子固有的物理和化学性质。它有助于对大量分子特性进行高通量评估，实现化合物的虚拟筛选。此外，它还提供了预测新分子未知属性的方法，从而提高研究效率并缩短开发时间。

在这项任务中，我们主要关注分子的量子力学特性，数据来自 MoleculeNet 的 QM9 数据集（[Wu 等人，2018 年](#)）。该数据集包含与每个分子相关联的多个属性数据，从而为与其对应的分子相关联的每个属性创建了一个独特的指令。例如，给定分子 A 及其属性 P_1 、 P_2 和 P_3 ，我们会生成三条不同的指令，每条指令都以分子 A 为输入，以相应的属性 (P_1 、 P_2 或 P_3) 为输出。相应地，各个指令反映了它们所代表的属性。如需全面了解，请参阅图 8 (a)。按照这一程序，经过预处理后，我们得到了 401 229 条指令条目。

当然，我们承认目前收集的属性范围可能有些狭窄，不能完全代表分子领域的全面情况。我们将努力不断优化和丰富我们的指令数据集。创建此类指令的初衷是研究在 输入相同但 指令要求不同的情况下，LLM 是否能产生相关 输出。

B.2 面向蛋白质的说明

我们使用通用蛋白质知识库 UniProtKB 作为数据源。同时，为了保证蛋白质的质量，我们选择了来自 UniProtKB/Swiss-Prot 的条目，这是一个高度可靠的人工注释蛋白质序列数据库。然后，我们从 UniProtKB/TrEMBL 中选择注释分数较高的条目来扩充数据集，从而提高数据量和多样性。此外，为了减少因蛋白质序列数据冗余（如在 Swiss-Prot 和 TrEMBL 数据库中具有高序列同一性的近缘蛋白质变体和同源序列）而导致的潜在训练偏差，我们采用了严格的过滤流程，以清除数据集中的冗余蛋白质。首先，我们使用 MMseqs（[Steinegger & Söding, 2017](#)）工具 (-min-seq-id 0.9) 根据 90% 的相似性阈值对这些蛋白质序列进行聚类，将每个聚类中功能注释丰富的条目指定为代表性蛋白质，并剔除其余条目。其次，以来自 TrEMBL 的代表性蛋白质为查询数据库，以来自 Swiss-Prot 的蛋白质为目标数据库，执行 MMseqs 搜索 (-min-seq-id 0.9)。在搜索过程中，所有与 Swiss-Prot 序列比对的 TrEMBL 序列，只要序列同一性达到 90% 或更高，就会被剔除。这些措施旨在简化高质量、非冗余蛋白质功能导向指令数据的构建。

蛋白质设计 在这项工作中，我们试图利用文本形式构建指令，以实现用户意图的全新蛋白质设计，从而方便地设计出用户感兴趣的复合属性。从形式上看，给定用户的设计要求，需要模型生成符合这些要求的蛋白质氨基酸序列。因此，我们从 UniProt 知识库的 20 个特征中选择注释，生成蛋白质设计要求作为指令。如表 6 所示，这些特征包括生物研究人员通常研究的或预期的目标蛋白质的特性。为了更好地将这些结构知识与说明结合起来，我们设计了一系列模板（见表 7），将注释转换为文本预期属性。因此，我们可以将上述功能描述和属性集合到蛋白质设计说明书中（我们将注释视为一个单独的条件）。表 7 展示了将结构数据转换为文本格式的模板示例。这样，我们就得到了 200,000 个指令条目， 输入是设计要求， 输出是理想的蛋白质序列。在这项任务中，所提供的目标序列只是一个参考，这

史研究国际会议 (ICLR 2024) 上发表。
与其他有明确基本事实的功能预测任务形成了鲜明对比。

随着时间的推移，UniProt 数据库已经积累了大量经过实验验证的蛋白质功能描述和注释。这些资料有助于研究人员迅速了解这些蛋白质在各种生物过程中的作用和机制。然而，序列数据库中蛋白质数量的迅速增加及其功能的多样性给通过计算方法进行自动功能预测带来了巨大挑战。在我们的工作中，我们重点关注三个不同的功能注释任务，并

表 6：用于制定蛋白质设计精确约束的 20 个注释特征。当涉及到缺乏特定类型的特征时，如信号（只表示特定蛋白质中存在某种特定特征），我们在 "示例" 字段中用"-"表示。

特点	内容	示例
 功能	蛋白质的一般功能	管状脂质转运蛋白，在细胞器接触部位介导膜之间的脂质转运。
 途径	相关代谢途径	氮代谢
 辅助因子酶活性所需的非		蛋白质物质 镍 ²⁺
 催化活性由酶催化的反应		(S)-脲基乙酸酯 = 乙醛酸 + 尿素
 基因本体 (MF) 蛋白质进行的分子水平活动		 活性 基因本体 (BP) 由
	多种分子活动完成的生物过程 信号转导	
	基因本体 (CC) 蛋白质发挥功能的细胞结构的相对位置	线粒体
 信号	将蛋白质定向到外质空间或分泌途径的序列	-
 盘绕线圈蛋白质中的盘绕线圈		区域 -
 基因突变具有生物学意义的		短序列基因突变 核定位信号
 每个模块化蛋白质结构域的结构域		类型 PWWP 结构域
 组成偏倚	蛋白质中的组成偏倚	极性残基 拓扑结构域
 跨膜蛋白质的非膜区 、 黑色素体		
 跨膜跨膜区的 范围		螺旋跨膜区
 DNA 结合DNA 结合结构域的		类型 同源染色体
	结合点任何化学基团的结合点	ATP
活性site	直接促进酶活性的氨基酸	用于 GAT 酶活
性的转折	转	-
贝塔链	贝塔链区域	-
	螺旋区域	-

表 7：将特定蛋白质字段的结构化注释转换为文本形式的模板。在实践中，我们还利用 LLM (GPT-3.5) 来丰富模板的样式。

特点	模板
----	----

史研究国际会议 (ICLR 2024) 上发表



功能对于一般功能，蛋白质需要满足该{功能}。



途径一种蛋白质，能够高效、特异地催化{途径}中的反应。辅助因子该



结合才能发挥酶的功能。



催化活性 蛋白质应能催化下列反应{催化活性}。基因本体 (MF) 蛋白质必须能够{分子功能}。



Gene Ontology (BP) 所设计的蛋白质必须能够调节{生物过程}。基因本体 (CC) 设计的



蛋白质定位在{细胞成分}上。



信号 在蛋白质设计 中加入 信号肽。



盘绕 线圈目标蛋白质必须包含一个盘绕线圈结构域。



基调蛋白质的功能设计需要包含一个{基调}。



领域所设计的蛋白质应包含一个或多个对其功能至关重要的{领域}。组成 蛋白质应表现出{组成偏



差}。



拓扑结构域结构域 蛋白质的{拓扑结构域}应设计成柔性或刚性。跨膜该 蛋白质应包含一个{跨膜}跨

膜区。

DNA 结合一种具有{DNA 结合}能力的蛋白质，用于定向基因调控。

结合 位点蛋白质应能在各种条件下与{结合位点}配体结合

活性位点 设计的蛋白质必须具有在相关酶中高度保守的{活性位点}。二级结构 目标蛋白质主要

构象必须是{β股、螺旋、转折}。

一个专门用于生成功能描述的任务。我们的目标是评估通用语言模型在统一框架下解决各种功能预测问题的能力。具体来说，我们考虑了两种广泛使用的分类方案，即基因本体 (Gene Ontology, GO) 联盟 ([Ashburner 等人, 2000 年](#)) 和酶委员会 (Enzyme Commission, EC) 编号，这两种方案组织了无数的蛋白质功能。

蛋白质功能预测 对于 GO 术语预测，给定特定功能预测指令和蛋白质序列后，模型利用三个不同领域（细胞成分、生物过程和分子功能）中的 GO 术语来描述蛋白质的功能。在获得的 116,458 个指令条目中，**输入**是蛋白质序列，**输出**代表该蛋白质的功能。

催化活性预测 同时，用相应的反应来替代 EC 编号（酶的数字分类系统，以其催化的化学反应为基础）。这种替换旨在利用预先训练的语言模型中根深蒂固的隐性知识，从而鼓励模型预测反应本身，而不仅仅是 EC 编号。在最终的 54,259 个指令条目中，**输入**是蛋白质序列，**输出**是蛋白质的催化活性及其促进的化学反应。

结构域/母题预测 我们介绍结构域预测任务，该任务要求语言模型识别给定蛋白质序列中的结构域类型，蛋白质序列被定义为紧凑折叠的三维结构。46,028 个指令条目中的每个条目都包含一个蛋白质序列作为**输入**，**输出**则是该蛋白质可能包含的结构域或主题。

功能描述生成 在上述功能预测任务的基础上，我们设计了功能描述生成任务，该任务不仅评估语言模型确定蛋白质序列功能的推理能力，还评估语言模型文本生成的效果。经过数据预处理后，我们得到了 88,259 个指令条目，其中**输入**为蛋白质序列，**输出**则描述了蛋白质的功能、亚细胞定位以及可能参与的任何生物过程。

B.3 生物分子文本说明

化学实体识别 化学实体识别 (CER) 是生物医学文本挖掘和自然语言处理 (NLP) 的一项基本任务。它涉及文本数据（如科学文献）中化学实体的识别和分类。这些实体可以包含广泛的概念，包括化合物、药物、元素、离子或官能团。鉴于化学命名的复杂性和多样性，CER 任务对 LLM 来说是一项重大挑战，而他们在这项任务中的表现可以为了解他们在生物医学领域的整体能力提供重要信息。在这项任务中，我们使用了 BC4CHEMD (Krallinger 等人，2015 年) 数据集，其中的化学实体是由专家手动识别和标注的。为了确保每个任务都有均衡的代表性，并使模型具备处理各种任务的能力，我们从 BC4CHEMD 数据集中随机抽取了 1000 个条目。

化学-蛋白质相互作用提取 我们让语言模型承担注释化学-蛋白质相互作用的细致任务。这项工作旨在探索这些模型所包含的生物化学和药理学知识的范围。更具体地说，这些模型会收到科学文献的摘录，不仅需要识别文本中不同的化学物质，还要辨别它们之间相互作用的具体性质。例如，这可能涉及确定已识别配体与蛋白质之间的调控关系。我们利用 ChemProt (Krallinger 等人，2017 年) 数据集作为主要来源，随后将其转换为指令数据。如果在一份共同的科学文档中存在多个关系三元组，我们会将它们合并为一个单独的注释，并用逗号分隔和划定。

化学-疾病相互作用提取 这项任务的目标是从给定的医学文献中找出化学物质与疾病之间的

史研究国际会议 (ICLR 2024) 上发表。这些概念被称为化学诱发疾病 (CID) 关系。这些 CID 关系在生物医学研究中至关重要，在药物发现、毒理学和疾病病因学等领域发挥着关键作用。我们使用的 BC5CDR 语料库 (Li 等人, 2016 年) 由 1,500 篇 PubMed 文章组成。这些文章共收录了 4,409 种注释化学物质、5,818 种疾病和 3,116 种化学物质与疾病的相互作用。考虑到 BC5CDR 数据集的结构与 ChemProt 数据集类似，我们采用了一系列相同的处理步骤来将数据集条目转换为说明。

提示:

给你一段文字，你能针对这段文字提出五个问题并给出答案吗？以问答对的形式列出。

段落：外膜蛋白 G (OmpG) 纳米孔是一个由七个柔性细胞外环组成的单体 β -桶状通道。其最灵活的环路 6 可用于承载高亲和性结合配体，以捕获蛋白质分析物，从而产生用于蛋白质鉴定的特征电流模式。在酸性 pH 值条件下，OmpG 检测蛋白质分析物的能力因其趋于闭合状态而受到影响，这使得纳米孔无法显示由结合的分析物引起的电流信号变化。在这项工作中，我们确定了控制环路 6 的 pH 依赖性门控的关键残基，并通过突变这些 pH 敏感残基，创建了一种能在广泛的 pH 值范围内保持主要开放状态的 OmpG 纳米孔。通过化学方法将短单链 DNA 与对 pH 值

不敏感的 OmpG 连接起来，证明了 OmpG 纳米孔在酸性 pH 值下感知互补 DNA 和 DNA 结合蛋白的实用性。

GPT-3.5 完成：

Q1: 什么是 OmpG 纳米孔？

A1: OmpG 纳米孔是一个由七个柔性胞外环组成的单体 β -桶状通道。

问题 2: 6 号环路有什么用途？

A2: 环路 6 用于承载高亲和性结合配体，以捕获蛋白质分析物，从而产生用于蛋白质鉴定的特征电流模式。

问题 3: 如何解决封闭状态的倾向？

A3: 可以确定控制环路 6 的 pH 依赖性门控的关键残基，通过突变这些 pH 敏感残基，可以创建在广泛 pH 值范围内保持主要开放的 OmpG 纳米孔。

问题 4: 如何证明 OmpG 纳米孔的实用性？

A4: 通过化学方法将短单链 DNA 与对 pH 值不敏感的 OmpG 连接在一起，从而在酸性 pH 值下感知互补 DNA 和 DNA 结合蛋白，证明了 OmpG 纳米孔的实用性。

问题 5: OmpG 纳米孔是什么类型的通道？

A5: OmpG 纳米孔是一个单体 β -桶状通道。

图 7：SELF-QUESTIONING 方法的示意图，该方法可根据从 PubMed White (2020 年) 获得的生物分子领域的 PubMed 摘要生成事实问答对。

多选题 我们从 MedMCQA (Pal 等人, 2022 年) 和 MMLU (Hendrycks 等人, 2021 年) (大规模多任务语言理解) 数据集中收集了 12398 道多选题。利用指令调整方法，我们的目标是激发语言模型的专业能力，以回答与生物分子有关的各种问题。这些问题涉及生物、化学、医学和其他与生物分子研究相关的垂直领域。

真假问题 真假问答任务的目的是回答具有肯定、否定或不确定回答的研究问题（例如，术前服用他汀类药物能减少冠状动脉旁路移植术后的心房颤动吗？）在最初的 PubMedQA 数据集中，每个条目都由一个问题、PubMed (White, 2020 年) 上一篇出版物的摘要（供参考）、一个答案和对答案的相应解释组成。对于这项任务，我们并没有刻板地设计一个提示模型参考材料进行回答的指令，而是取消了参考材料，将问题视为后续开放式问题的补充，以适应更多的问题类型。此外，对给定问题的理想回答应该是提供一个准确的答案，并给出有说服力的合理解释。

作为会议论文在 2024 年国际比较文学和历

史研究国际会议 (ICLR 2024) 上发表

开放式问题 开放式问题的定义是，只需提出问题，而不对回答格式施加任何限制的问题。

这将它们与预先确定回答格式的问题区分开来。我们主要从 MedMCQA 数据集中收集开放式问题，特别是选择那些与生物化学、药理学和医学领域相关的问题。MedMCQA 问题的原始格式是选择题，例如 "白三烯抑制剂对以下哪种情况非常有效？然而，我们发现某些问题没有明确说明答案需要

表 8：生物大分子的数据统计。

特点	最小	最大	平均值	中位数
	贝尔茨复杂性	0	36,222	508
	分子量	1	8,656	273
	原子数	1	574	19
	环数	0	75	3
	序列长度	2	39677	455
	域名	1	2215	27
	催化活性	1	2404	15
	基因本体	1	28924	35

例如，"用于治疗 RSV 感染的抗体是："这样的问题。因此，我们很方便地将 27,574 个问题转化为开放式问题，相应的答案包括正确选项的描述和解释，即 MedMCQA-Open。

为了深入研究有关生物分子的查询，提高语言模型在化学和生物学领域的能力，我们提出了 SELF-QUESTIONING，通过使用 GPT3 从 PubMed 生物分子研究领域的摘要中提取事实问答对，来扩展开放式问题的指导数据。SELF-QUESTIONING 的实施包括三个步骤：1) 数据收集；2) 利用自问式方法生成问答实例；3) 过滤低质量问题。

首先，我们从 2022 年 12 月发布的年度基准 PubMed 中收集完整的摘要。PubMed 中的每篇出版物都包括标题、摘要和医学主题词表 (MeSH)。为了集中回答用户提出的有关生物分子领域的问题，我们只考虑那些摘要和 MeSH 包含一些可能与生物分子研究有关的特定关键词（如蛋白质、分子）的出版物。其次，期望模型能根据所提供的参考资料生成高质量和符合事实的问题具有挑战性。尽管如此，我们发现，在要求模型提取问答对的指令提示下，经过预训练的语言模型可以做到这一点。这种方法可以确保生成的答案与相应的资料来源正确一致。图 7 举例说明了这一生成过程。第三，为了进一步确保所生成问题的整体质量，我们加入了一个排除标准，该标准可大量排除与参考资料密切相关的问题，例如，"研究的目的是什么？这类问题通常涉及 "结果"、"研究"、"段落" 和 "文章" 等特定词汇。通过识别这些词汇，我们可以有效地将它们排除在考虑范围之外。我们将包含 10,521 个示例的数据集称为 PubMedQA-GPT。然后将其与 MedMCQA-Open 数据集结合起来，构建指令数据。

C 更详尽的数据分析说明

图 4 从多方面分析了分子和蛋白质的多样性和复杂性。为清晰起见，只显示了部分坐标覆盖范围。表 8 提供了更全面的统计数据。总体而言，这些统计数据反映了生物大分子的广泛性和多样性，这应该有助于提高模型的鲁棒性和通用性。事实上，蛋白质序列通常都很长，这给 LLM 捕捉和理解其 "语法" 或模式带来了巨大挑战。编码和处理这些冗长序列的策略，以及指导 LLM 预测或设计有效蛋白质序列的有效方法，都是需要进一步研究的关键领域。

D 实验装置详情

作为会议论文在 2024 年国际比较文学和历

史研究国际会议 (ICLR 2024) 上发表

对于 "分子指令" 数据，我们将 7B LLaMA 模型作为整个实验的基础模型，并在 LLaMA 模型上进行指令调整。在这项工作中，由于模式的多样性和任务难度的巨大差异，我们针对特定的生物分子问题，在三个不同的数据集上训练 LLaMA 模型，从而开发出三个微调模型。对于每个不同的任务，我们都分配了近 1k 个样本作为测试集。

计算指标并评估模型的性能。其余样本按 8:2 的比例分为训练集和验证集。

虽然标记化是数据处理的关键部分，尤其是在处理各种模式的数据时，但本研究并未深入广泛地分析不同模式的顺序标记化。因此，我们把蛋白质氨基酸序列和 SELFIES 分子串当作人类语言来处理，并使用字节对编码 (BPE) 算法对数据进行标记化。我们采用与 LLaMA 模型相同的标记化模型。我们使用最长填充策略将标记化序列填充到批次中最长的序列长度。

我们在分子和面向文本的数据集上采用了低秩适配器 (Lowrank Adapter, LoRA) 微调，这是一种高效的方法，可在训练过程中减少内存，保持一小部分参数可训练，同时不更新预训练模型。对于面向蛋白质的指令数据，我们使用 DeepSpeed 库中实现的内存优化技术 ZeRO 进行全面微调。我们使用 AdamW 优化器和线性学习率调度器训练模型。我们在 32GB V100 GPU 上进行 LoRA 训练和生成，同时在 80GB A800 GPU 上进行全模型微调。我们为每个模型调整的超参数如表 9 所示。

表 9：在不同数据集上进行微调的训练超参数。QV：自我关注模块中查询状态和价值状态的两个线性变换矩阵。

超参数	分子	蛋白质	文本
微调方法	LoRA	全部	LoRA
批量大小	800	96	1024
LR	3e-4	2e-5	3e-4
步骤	40,000	25,000	840
热身步骤	1,000	2,500	100
LoRA r	16	-	16
LoRA α	16	-	16
LoRA 缓学	0.05	-	0.05
LoRA 层	QV	-	QV

E 评估指标

E.1 分子指标

为了评估分子理解任务，我们利用 BLEU (Papineni 等人, 2002 年)、ROUGE (Lin, 2004 年) 和 METEOR (Banerjee & Lavie, 2005 年) 等指标，通过与参考答案进行比较来评估生成输出的质量。对于分子生成，我们首先使用 RDKit (Landrum 等人, 2013 年) 验证生成的字符串是否为有效分子，然后计算它们与参考解答的精确匹配度。不过，需要注意的是，一个文本描述可能对应多个分子结构。此外，期望 LLM (即使是根据特定指令对 LoRA 进行微调的 LLM) 产生与参考数据完全匹配的输出结果可能是不现实的。为了适应这些复杂性并提供更全面的评估，我们进一步采用了衡量分子相似性的指标。这些指标包括从 RDKit/MACCS/Morgan 指纹 (Tanimoto, 1958; Schneider 等人, 2015; Durant 等人, 2002) 以及 Levenshtein (李和刘, 2007) 和 BLEU 分数中得出的相似性分数。对于分子性质预测任务，我们采用 MAE (平均绝对误差) 来量化模型预测连续值的准确程度。

E.2 蛋白质指标

对于蛋白质理解任务，我们将这些任务视为功能描述任务，并采用 ROUGE 来量化生成的生物特征描述的质量。

E.3 生物技术指标

对于 NLP 文本任务，我们采用问答、实体识别和关系提取的通用指标来评估生成的输出结果。

F 其他结果

由于篇幅有限，正文中只显示了部分结果。其余结果进一步说明了我们的 "莫尔指令" 的有效性，见图 8、图 9、图 10 和图 11。请注意，Galactica (Taylor 等人, 2022 年)、Text+Chem T5 (Christofidellis 等人, 2023 年) 和 MolT5 (Edwards 等人, 2022 年) 只支持 SMILES 格式。为此，我们将指令中的 SELFIES 转换为 SMILES。还需要提及的是，MolT5 没有经过指令调整，因此只能处理指令中的 "输入" 部分。

对图 8 所示各种实验的分析揭示了 "分子指令" 如何指导 LLM 理解和执行特定的分子任务。这种指导作用在分子描述生成和描述指导分子设计等任务中非常明显，在这些任务中，数据集提供了精确的指令，指导模型生成详细的分子描述和精确的设计。MolT5 的性能明显不如指令跟随模型，表现出很差的泛化能力。在与化学反应有关的任务中，如正向反应预测、逆合成和试剂预测，Mol-Instructions 在指导模型预测反应产物、潜在反应物或正确试剂方面发挥了关键作用。Alpaca-LoRA、ChatGLM、Baize 和 Vicuna 在缺乏此类具体指令的情况下，产生的输出结果要么过于通用，要么与任务要求不符。与特定领域的较小模型相比，LLM 在分子生成方面的表现一般较差，这是因为它们的关注点更广，从而影响了特定任务的专业化。在性质预测任务中，Mol-Instructions 可以帮助模型对 HOMO、LUMO 和 HOMO-LUMO 间隙能等各种分子性质进行近乎精确的估计。相比之下，Alpaca-LoRA 的输出结果由于缺乏具体说明的指导，与实际值有很大偏差。Galactica 一直输出负值。

正如图 9 和图 10 所示的预测实例所证明的那样，微调模型为功能注释提供了一致的、针对特定用户的蛋白质分析。例如，在催化活性预测任务中，该模型准确地将提供的蛋白质归类为 23S rRNA (腺嘌呤 (2503) -C (2)) 甲基转移酶。此外，虽然与 UniProtKB 中的注释相比，该模型在描述不同的功能特征时可能会出现细微差别，但它在辨别蛋白质的基本特性方面展示了非凡的潜力。

如图 11 所示，在与化学实体及其相互作用相关的任务中，Mol-Instructions 可以帮助模型识别正确的实体，并在它们之间建立有意义的关系。相比之下，没有根据我们的指令进行微调的模型（如 Alpaca-LoRA）所产生的输出结果往往过于笼统，或忽略了数据中存在的特定关系。在处理多选题、真假题或开放题时，使用莫尔指令微调的模型不仅能提供正确答案，而且能以更详细、结构更合理的方式提供答案。这证明了 "莫尔指令" 在提高学习者对这些问题的理解和答题能力方面的有效性。

总之，研究结果表明，Mol-Instructions 可以提高 LLM 执行分子导向、蛋白质导向和生物分子文本任务的能力。它强调了特定任务指令对模型指导的重要性，尤其是在处理生物化学等专业领域时。相比之下，像 Alpaca-LoRA 这样缺乏这种基于特定指令的调整的模型，可能很难产生符合任务特定要求的输出结果。

尽管取得了上述令人鼓舞的成果，但也存在一些局限性。首先，虽然实验证实了 "莫尔指令"

史研究国际会议 (ICLR 2024) 上发表。能够在一定程度上帮助学习者理解和掌握与生物分子相关的信息，但需要明确的是，现阶段通过指令调整获得的模型仅仅是一个初步示范。它在实际生产任务中的应用潜力仍然有限。此外，要充分发挥 Mol-Instructions 的潜力，还需要在算法改进方面进行更深入的探索和创新。这包括优化指令调整方法，进一步完善模型理解和遵循指令的能力，以及增强其在各种分子任务中的泛化能力。因此，尽管 Mol-Instructions 代表着生物分子任务中指令数据集的使用向前进了一步，但要最大限度地提高其实用性和有效性，未来的研究和开发仍有很大的空间。

图 8：在面向分子的任务中使用指令调整 LLM 的性能比较。

史研究国际会议 (ICLR 2024) 上发表，这不可避免地引起了人们的担忧。如果落入坏人之手，这些能力可能会使恶意行为者利用该系统达到邪恶目的，如合成有害的生化制剂或生物武器。

(b2) Protein

Domain/motif prediction

Instruction: Given this protein sequence, can you identify any common protein motifs or domains that it may contain?

Input:
MANTKYIIFITGGVVSSLGKGIAASIGALLESRGLSLSLKVDPYINVDPGTMSPFQHGEVFTEDGTETDLDLGHYERFVRFKASKKNNFTAGKVYETVIR
ERKGNLYLGGTVQVPHITNEIKRKKGQQNKDIAVEVGGTVDIIESQPFVPEALRQMALELPNNSWAFVHLTLVPFVINASGELKTTPQHSVKELRSLGISP
VLVCRSEQUELPKDEKKNKIALFCSPVAKSVISMHDVDTVSYIPLLNNKQKVDDTLLKLNKPKNLNDWKRVVAKLPEKEVNFSFGVKYTELKDSYKSI
NEALEHAGIQNKAKVNINFVEAEQITSQNVRLKKSDSLGVPGFGGERGIEGMILACKYARENVPYLICLGMQIAIIYEYARNVLKLKSANSTEDSSTKF
VIGLIETEWSDISGKKEKRTKNSDLGTMRLGGQVCKLKKNSYKMKSEIIERHRHRYEVNPNEYKDKMIEQGLDVVGTISIDGKLEMIELPSHKWFLAC
QFHPEPTSNPDRGHPIFSYIKSTTIC

Output (ground truth): Our predictive analysis of the given protein sequence reveals possible domains or motifs. These include: Glutamine amidotransferase, CTP synthase N-terminal domains.

Output (ours): Based on computational analysis, the provided sequence potentially contains the following protein domains or motifs: Glutamine amidotransferase type-1 domains.

Output (Alpaca): The protein sequence contains multiple protein motifs and domains. It contains an N-terminal signal peptide, a proline-rich region, a coiled-coil domain, a leucine-rich repeat domain, a proline-rich region.

Output (ChatGLM): 与艺术有关, 具体是指与音乐、艺术、文学或文学有关

Output (Baize): Here are some of the possible motifs and domains: Glycine-rich region, Leucine-rich region, Kinase domain, Fibrinogen-like domain, Leucine-rich repeats, GTPase domain.

Output (Vicuna): Based on the provided protein sequence, there are several common protein motifs and domains that can be identified. These include: A signal peptide, which is indicated by the presence of the sequence

"¹MANTKYIIFITGGVVSSLGKGIAASIGALLESRGLSLSLKVDPYINVDPGTMSPFQHGEVFTEDGTETDLDLGHYERFVRFKASKKNNFTAGKVYETVIR
ERKGNLYLGGTVQVPHITNEIKRK"

Output (Galactica): ## Summary **Protein**: 10 kDa chaperonin; **Gene**: groS; **Organism**: Streptococcus pyogenes; **Status**: inferred from homology; ## Function What is the function of the protein 10 kDa chaperonin? binds to Cpn60 in the presence of Mg-ATP and suppresses the ATPase activity of the latter.

Functional description generation

Instruction: Examine the given protein sequence and share a brief overview of its attributes:

Input:
MKIVLASNNQGKLAELKAMILAPLGQLRQAELGPIEAEPFRITFVENALAKRHASALSGLPALADDAGLCVEAFGLPGVDTAFYAVQFGYAKGDANN
VKALLAQAMGITDRAALVSTLVAVRSAEDEPEPLIACGRVAGEVALEPMGSNGFGDPVMFPIPEGOTFAQLPVEVKNANSRGKATRQMMALRERWIT

Output (ground truth): A concise description of the protein with the specified amino acid sequence includes: Pyrophosphatase that catalyzes the hydrolysis of nucleoside triphosphates to their monophosphate derivatives, with a high preference for the non-canonical purine nucleotides XTP (xanthosine triphosphate), dITP (deoxyinosine triphosphate) and ITP. Seems to function as a house-cleaning enzyme that removes non-canonical purine nucleotides from the nucleotide pool, thus preventing their incorporation into DNA/RNA and avoiding chromosomal lesions.

Output (ours): Here is a summary of the protein with the given amino acid sequence: Pyrophosphatase that catalyzes the hydrolysis of nucleoside triphosphates to their monophosphate derivatives, with a high preference for the non-canonical purine nucleotides XTP (xanthosine triphosphate), dITP (deoxyinosine triphosphate) and ITP. Seems to function as a house-cleaning enzyme that removes non-canonical Purine nucleotides from the nucleotide pool, thus preventing their incorporation into DNA/RNA and avoiding chromosomal lesions.

Output (Alpaca): The protein sequence

MKIVLASNNQGKLAELKAMILAPLGQLRQAELGPIEAEPFRITFVENALAKRHASALSGLPALADDAGLCVEAFGLPGVDTAFYAVQFGYAKGDANN
VKALLAQAMGITDRAALVSTLVAVRSAEDEPEPLIACGRVAGEVALEPMGSNGFGDPVMFPIPEGOTFAQLPVEVKNANSRGKATRQMMAL

Output (ChatGLM): The given protein sequence is not provided. However, based on the amino acid sequence of a protein that was provided earlier, it appears to be a protein of unknown function.

Output (Baize): The given protein sequence is an artificial sequence that has not been characterized in any way. Therefore, its attributes cannot be evaluated or understood in any way. This sequence does not correspond to a known protein, and its function, subcellular localization, and biological processes cannot be determined.

Output (Vicuna): The given protein sequence appears to be a random assortment of amino acids without any discernible pattern or structure. A more detailed analysis would be necessary to determine any potential attributes or functions of this protein.

Output (Galactica): Multifunctional enzyme that displays cysteine protease, NTPase, RNA triphosphatase and RNA helicase activities. May play a role in the packaging and maturation of the virion. Specifically cleaves the capsid protein at two distinct sites. This protein has RNA-binding and NTPase activities and associates with the HBV RNA genome to promote encapsidation and replication. Also possesses a guanylyltransferase activity that is involved in maturation of the virion. Together, these activities participate in the packaging of the three reverse-transcribed copies of the genome of the HBV virion. Also plays a role in the biosynthesis of the virion envelope glycoproteins, in the budding of the virions through interactions with the cytoplasmic tail of the glycoprotein E2 at the ER membrane.

Catalytic activity prediction

Instruction: Please evaluate the following protein sequence and provide an explanation of the enzyme's catalytic activity, including the chemical reaction it facilitates:

Input:
MDKIVAVAGFLPEELCASLSSLSPSFRGNQFQWIGKGVDSFDAMTNLSAELRASLAEKAILRSTRVSDVULKADDDGTVKLQIOTEDDLAVETVLLTDKAARKTA
CVSQAGCMGCAFCKTGTGLARLNAAEIVEQFLYLEKHAGALDNIVFMGMGEPLLNLDAHLRKAIAVLTDKGRNRNLSRRITVSTVGIVSGYDLANNG
PDVRLAVSLTAADETLLRRELMPASLTPSLDRQAISSYYIETKGRVTEAVLLSGKNTSEKNAIDLIAFKGLDVHVNLIPWNPVEGLSFVTPDPEETAQFVS
RLEKGLLNVTLRMHRSKISGACGLQKTNPYA

Output (ground truth): Based on the provided protein sequence, the enzyme appears to facilitate the chemical reaction: adenosine(37) in tRNA + 2 reduced [2Fe-2S]-[ferredoxin] + 2 -S-adenosyl-L-methionine = 2-methyladenosine(37) in tRNA + 5'- deoxyadenosine + L-methionine + 2 oxidized [2Fe-2S]-[ferredoxin] + S-adenosyl-L-homocysteine.

Output (ours): An analysis of the protein sequence reveals that the enzyme's catalytic function corresponds to the chemical reaction: adenosine(37) in tRNA + 2 reduced [2Fe-2S]-[ferredoxin] + 2 -S-adenosyl-L-methionine = 2-methyladenosine(37) in tRNA + 5'- deoxyadenosine + L-methionine + 2 oxidized [2Fe-2S]-[ferredoxin] + S-adenosyl-L-homocysteine.

Output (Alpaca): This protein sequence is the amino acid sequence of the enzyme catalase. Catalase is an enzyme that catalyzes the breakdown of hydrogen peroxide into water and oxygen. The chemical reaction catalyzed by catalase is the decomposition of hydrogen peroxide into water and oxygen:
 $2\text{H}_2\text{O}_2 \rightarrow 2\text{H}_2\text{O} + \text{O}_2$

Output (ChatGLM): The protein sequence provided is the catalytic domain of a hypothetical enzyme, named "MLVAVAGFLPEELCASLSSLSPSFRGNQFQWIGKGVDSFDAMT". The catalytic domain contains a large loop (loops 1-6) that is involved in the binding of the substrate and the formation of the catalytic site. The loop is surrounded by small amino acid residues that are involved in specific interactions with the substrate, such as hydrogen bonding and electrostatic interactions.

Output (Baize): This protein sequence corresponds to the active site of a class of enzymes known as cysteine proteases, specifically, a member of the subfamily of cysteine proteases known as serine proteases. The amino acid sequence contains a conserved catalytic triad of residues (Cys, His, Asp) that are essential for the activity of the enzyme. The chemical reaction facilitated by the enzyme is the cleavage of a peptide bond between two amino acids by a serine protease. The specificity of the enzyme for the peptide bond that it cleaves is determined by the amino acid sequence of the substrate.

Output (Vicuna): The protein sequence provided is for an enzyme called pepsin, which is a proteolytic enzyme that catalyzes the hydrolysis of peptide bonds in protein molecules. The chemical reaction that pepsin facilitates is the breakdown of proteins into smaller peptides and individual amino acids. This reaction is an important step in digestion, as proteins are broken down into smaller molecules that can be more easily absorbed by the body.

Output (Galactica): ## Function Q: What is the function of the protein N-acetyl-gamma-glutamyl-phosphate reductase? Catalyzes the NADPH-dependent reduction of N-acetyl-5-glutamyl phosphate to yield N-acetyl-L-glutamate 5-semialdehyde. ### Catalytic Activity - N-acetyl-L-glutamate 5-semialdehyde + NADP(+) + phosphate = H(+) + N-acetyl-L-glutamyl 5-phosphate + NADPH

图 10：在面向蛋白质的任务中使用指令调整 LLM 的性能比较。(2/2)

G.2 生成式人工智能的伦理考量

最近的研究 (Service, 2023; Sandbrink, 2023) 揭示了生成式人工智能在生物工程中的潜在滥用问题。我们的工作虽然侧重于有益的结果, 但并没有对这些问题视而不见。我们认为,

史研究国际会议 (ICLR 2024) 上发表。任何生成式人工智能工具在向公众开放时，都应附带严格的使用指南和监控机制，以检测和防止滥用。

(c) Biomolecular Text

Chemical entity recognition**Instruction:** Find and list all the instances of the chemical entities in the following content.**Input:** Both the control and caramiphen groups with double cannulas had significantly shorter latencies to seizure onset than the corresponding groups with single cannula.**Output (ground truth):** caramiphen**Output (ours):** caramiphen**Output (Alpaca):** Both the control and caramiphen groups with double cannulas had significantly shorter latencies to seizure onset than the corresponding groups with single cannula.**Output (ChatGLM):** Caramiphen is a drug, and cannula is a type of drug container.**Output (Baize):** control group, caramiphen group, double cannula, latency to seizure onset, seizure, onset.**Output (Vicuna):** Both the control and caramiphen groups with double cannulas had significantly shorter latencies to seizure onset than the corresponding groups with single cannula.**Output (Galactico):** 1. 1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12. 13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54...**Output (PMC-Llama):** The use of the double cannula significantly facilitated the monitoring of the experimental rats during the seizure tests. The ease with which experimental animals with double cannulated can be handled during seizure testing thus makes this the preferred method of cannulation.**Chemical-disease interaction extraction****Instruction:** You are provided with a set of clinical trial summaries. Extract the chemical-disease relations from the summaries and present your findings in the format of (Subject, Object).**Input:** Eating disorders and the associated behavioural problems and drug abuse are uncommon in pregnancy. When they do occur they are often unrecognized because of denial but when significant may pose a risk to both the mother and her fetus. This case illustrates a number of problems that may be encountered in women with eating disorders in pregnancy, including prolonged and recurrent metabolic disturbances and diuretic abuse. In particular it illustrates the derangements of thyroid function seen in pregnant women with eating disorders and reminds us that when a cause for thyrotoxicosis remains obscure, thyroxine abuse should be considered and explored.**Output (ground truth):** (thyroxine, thyrotoxicosis)**Output (ours):** (thyroxine, thyrotoxicosis)**Output (Alpaca):** Eating disorders, Behavioural problems, Drug abuse (Subject, Object)**Output (ChatGLM):** (Eating disorders, Behavioral problems, drug abuse, Eating disorders, Behavioral problems, drug abuse, Pregnancy, Eating disorders, Behavioral problems, drug abuse, Pregnancy, Metabolic disturbances, diuretic abuse, Pregnancy, Eating disorders, Behavioral problems, drug abuse, Thyroxine function, Pregnancy, Eating disorders, Behavioral problems, drug abuse, thyrotoxicosis)**Output (Baize):** Subject: Pregnancy, Eating Disorders. Object: Metabolic Disturbances, Drug Abuse, Thyroid Function, Prolonged Thyrotoxicosis.**Output (Vicuna):** Eating disorders - Pregnancy**Output (Galactico):** (Mother, Thyroxine)**Output (PMC-Llama):** yes**Multi-choice question****Instruction:** Which of the following conditions does not show multifactorial inheritance?**Input:** What is the following is the right choice? (A) Pyloric stenosis (B) Schizophrenia (C) Spina bifida (neural tube defects) (D) Marfan syndrome**Output (ground truth):** The final answer is (D).**Output (ours):** The final answer is (D).**Output (Alpaca):** C) Spina bifida (neural tube defects)**Output (ChatGLM):** Pyloric stenosis does not show multifactorial inheritance. It is a condition caused by a narrowing of the stomach outlet, which can lead to discomfort, nausea, and vomiting. The condition is inherited in an allelic fashion, meaning that it is caused by the presence of two different genetic variations.**Output (Baize):** The correct answer is (D) Marfan syndrome.**Output (Vicuna):** The condition that does not show multifactorial inheritance is (C) Spina bifida (neural tube defects). Spina bifida is a condition that occurs due to the failure of the neural tube to close properly during embryonic development, which is a single-gene defect. Pyloric stenosis, schizophrenia, and Marfan syndrome are all examples of multifactorial inheritance, which means they are caused by a combination of genetic and environmental factors.**Output (Galactico):** (C) Spina bifida (neural tube defects) (D) Marfan syndrome**Output (PMC-Llama):** OPTION D IS CORRECT.**True or False question****Instruction:** Does feeding tube insertion and its timing improve survival?**Input:** None**Output (ground truth):** No, neither insertion of PEG tubes nor timing of insertion affect survival.**Output (ours):** No, feeding tube insertion and its timing does not improve survival.**Output (Alpaca):** Yes, feeding tube insertion has been shown to improve survival in critically ill patients.**Output (ChatGLM):** Yes, feeding tube insertion and its timing can improve survival in certain situations.**Output (Baize):** While feeding tube insertion and its timing may be beneficial in some cases, the relationship between feeding tube insertion and survival is complex and requires further research and understanding.**Output (Vicuna):** Feed.**Output (Galactico):** No**Output (PMC-Llama):** no**Chemical-protein interaction extraction****Instruction:** Your task is to extract the subject and object entities from the provided text and establish the corresponding relation triples. Please format your answers as (Subject, Relation, Object).**Input:** Opioid receptors are the membrane proteins that mediate the pain-relieving effect of opioid drugs, such as morphine and fentanyl as well as endogenous opioid peptides enkephalins and endorphins. Using cDNAs for the mu and the kappa opioid receptors, we mapped the chromosomal locations of their genes in mouse. Multi locus cross analysis located the mu receptor gene Oprm on Chr 10 and the kappa receptor gene Oprk1 on Chr 1. Both genes are near centromere, with no markers more centromeric. These data indicate that the two opioid receptors are different gene products, ruling out the possibility that they may be differential splicing products from the same gene.**Output (ground truth):** (morphine, regulator, Opioid receptors), (fentanyl, regulator, Opioid receptors)**Output (ours):** (fentanyl, regulator, Opioid receptors), (morphine, regulator, Opioid receptors)**Output (Alpaca):** (Subject, Relation, Object)(mu receptor, is near, Chr 10)(kappa receptor, is near, Chr 1)**Output (ChatGLM):** (Opioid receptors, membrane proteins, mediate pain-relieving effect ; mu and kappa opioid receptors, genes, mapped chromosome locations ; mu receptor gene Oprm on Chr 10 ; kappa receptor gene Oprk1 on Chr 1 ; Multi locus cross analysis, location , centromere, mouse ; no markers more centromeric) (opioid receptors, gene products ; Morphine ; Endogenous opioid peptides enkephalins and endorphins)**Output (Baize):** (Subject, Relation, Object) Subject: Opioid Receptors, Relation: Mediate, Object: Pain-Relieving Effect of Opioid Drugs and Endogenous Opioid Peptides.**Output (Vicuna):** (Opioid receptors, are, the)**Output (Galactico):** Oprm**Output (PMC-Llama):** The mu opioid receptor gene Oprm locates to the distal region of Chromosome 10 in the mouse.**Open question****Instruction:** What virus causes chronic disease?**Input:** None**Output (ground truth):** Hepatitis C virus (HCV).**Output (ours):** Hepatitis C. Hepatitis C is a chronic liver disease caused by the hepatitis C virus (HCV).**Output (Alpaca):** Hepatitis B is a virus that causes chronic liver disease.**Output (ChatGLM):** There are many viruses that can cause chronic diseases. Some examples include: Chronic obstructive pulmonary disease (COPD), Chronic kidney disease, Chronic heart disease, Chronic diabetes, Chronic fatigue syndrome.**Output (Baize):** There are several viruses that can cause chronic diseases, including: Hepatitis B and C, HIV, Herpes Simplex Virus (HSV), Human T-cell Leukemia Virus Type 1 (HTLV-1), Varicella-Zoster Virus (VZV).**Output (Vicuna):** The virus that causes chronic disease is the Epstein-Barr virus (EBV).**Output (Galactico):** Hepatitis C**Output (PMC-Llama):** OPTION D IS CORRECT.**图 11：在生物分子文本任务中使用指令调整 LLM 的性能比较。**

G.3 降低风险

解决潜在的危险是最重要的。可以采取的一些措施包括

- **规范访问:** 只允许经核实的研究实体或有道德研究记录的个人访问。
- **监控使用模式:** 采用算法检测滥用或探索潜在有害领域的模式。
- **社区监督:** 建立以社区为主导的监督机制，由专家审查和批准具体用途或产生的成果。
- **透明报告:** 鼓励用户报告他们遇到的任何意外结果或可能有害的使用案例。

人工智能的伦理影响，尤其是在生物科学等敏感领域的伦理影响，怎么强调都不为过。虽然我们的工作代表着我们在利用 LLM 的能力做好事方面向前迈出了一步，但我们必须谨慎、勤奋地向前迈进，并坚定不移地恪守伦理原则。