ChemLLM: A Chemical Large Language Model

太长不看版

ChemLLM是一款专为化学领域设计的大语言模型，旨在解决化学数据结构化存储与模型对话能力之间的挑战。通过创新的基于模板的指令构建方法，ChemLLM能够将结构化化学知识转换为自然语言对话，适合语言模型训练。

该大模型的核心能力有：化学任务执行（ChemLLM在化学核心任务如名称转换、分子描述和反应预测上超越了GPT-3.5，并在部分任务上超越GPT-4。）、适应性（模型展现出对数学和物理任务的出色适应性，以及在化学专业NLP任务（如文献翻译和化学信息编程）上的熟练掌握。）、多语言交互（ChemLLM在处理中文化学问题上也表现出色，证明了其作为多语言化学工具的实用性。）

**创建了专门的ChemData数据集**，包含7M化学指令数据，涵盖分子、反应和领域特定任务，用于提升模型的指令遵循能力。**开发了两阶段训练流程**：结合一般语料库训练和化学领域知识训练，增强模型的通用语言能力和化学专业知识。**性能评估**方面通过ChemBench基准测试，MMLU和GSM8K评估模型的化学专业知识、通用语言能力和多语言适应性，证明了该模型的优越性。

**贡献**：ChemLLM的开发为化学研究提供了新的探索途径，推动了化学语言模型的发展。

**局限性**：模型在理解复杂的分子结构和空间配置方面存在限制，且在极端条件下生成响应时遵循科学伦理的能力有待提高。

**结论：** ChemLLM作为首个开源化学大语言模型，不仅在化学领域表现出色，还在多语言处理和跨学科任务中展现出强大的能力，为科学领域的LLM应用开辟了新的可能性。

摘要

大语言模型在分子性质预测、分子生成、实验方案设计等化学应用方面发展很好，但缺少专门为化学设计的对话模型。目前困境是大多数化学数据和科学知识主要储存在结构化的数据库中（structured database），直接使用这些结构化数据会影响模型连贯对话的能力。

为解决这一问题，我们提出一种新颖的基于模板的指令构建方法，将结构化的知识转换为简单的对话，使其适合于语言模型训练。通过这种方法，我们开发了ChemLLM，这是第一个致力于化学的大语言模型，能够通过流畅的对话交互执行跨化学学科的各种任务。

ChemLLM在化学的三项主要任务，名称转换（name conversion）、分子描述(molecular caption)和反应预测(reaction prediction)上击败了GPT-3.5，并在其中两项任务上超过了GPT-4。此外，ChemLLM对相关数学和物理任务也表现出出色的适应性，以及对化学领域专业NLP任务很熟练，如文献翻译和化学信息编程（cheminformatic programming）。

ChemLLM为化学研究的探索开辟了一条新的路径，而我们将结构化的化学知识整合到对话系统中的方法为在各个领域开发大语言模型开辟了新的前沿。

引言

大语言模型展示了强大的语言理解和生成能力，并展示了推理和非语境学习等新兴能力，使其可能应用于各种科学领域。值得注意的是，大语言模型已经被应用于化学相关任务，如分子性质预测、分子生成、实验方案设计等。这些工作展示了LLMs在化学领域的潜力，并为化学研究提供了有见地的建议和解决方案。

尽管先前用LLMs适应各种化学下游任务，但现有的LLMs并不是专门为化学领域设计的，它们缺乏对化学空间的理解，在处理化学数据和知识时面临几个挑战。以前的工作只关注特定下游任务开发专业知识模型，忽略了LLMs的指令遵循(instruction-following)和对话能力(dialogue capbilities)。对于大模型提升逻辑推理和泛化能力这些能力是必需的，而这些能力对于在化学领域能有更宽更多用途的能力也是必需的。

但是，在开发这种化学大语言模型也存在挑战（图1）。

1. 分子在化学信息学中以一种特殊的符号表示，称为SMILES（简化的分子输入行输入规范），模型需要能够正确理解和生成该符号；
2. 大多数化学信息和知识储存在结构化数据库中，直接使用这些数据来训练可能会影响其自然语言处理的技能；
3. 化学数据和任务很多样，很难为化学大语言模型设计一个统一的训练流程（pipeline）。

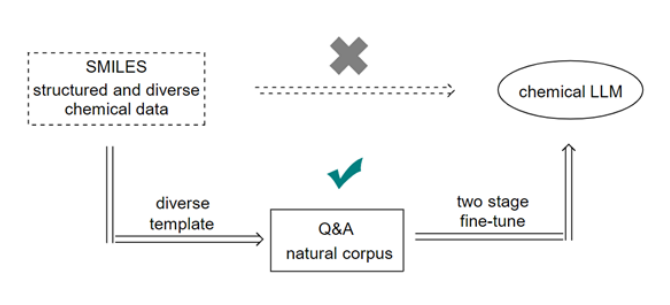


图1：开发化学大语言模型的困难和解决方案

在本文中，我们通过开发合成化学指令微调数据集ChemData来解决这些困难，它用一种基于模板的指令构建方法，将结构化的化学数据转换为适合训练的自然对话形式。

基于此，我们开发了首个开源化学大语言模型，ChemLLM，可以在保持完整自然语言能力的同时完成多种化学任务的能力。ChemLLM在大规模化学语料库上训练，这些语料库来自丰富模板合成的各种化学指令数据组成，如化学名称的转换、分子性质预测、分子生成、分子说明、反应条件预测、反应产物预测等。

我们还提出了一个两阶段的指令微调流程（two-stage instruction tuning pipeline）来调整ChemLLM：一般语料训练和化学领域知识训练。前者保留ChemLLM在一般情况下的能力，后者通过消融实验来比较化学语料训练的作用。

我们从化学专业知识、通用语言能力和多语言能力三个方面对ChemLLM进行评估。结果表明，ChemLLM在化学的三个核心任务：名称转换、分子说明和反应预测上的表现是GPT-3.5的两倍以上。

在一般情况下，ChemLLM在大多数任务上都优于或与现有模型不相上下，显示了其在数学和物理等其他相关领域的强大普适能力。在多语言能力方面，ChemLLM在中文问题上也表现出色。在一些定性任务中，ChemLLM也能处理与化学相关的NLP任务，如化学文献翻译和化学信息编程。

本文贡献如下：

ChemData: 本文提出了一种基于模板的指示框架方法(template-based instruction construction method)，将结构化的化学数据转变为自然对话形式，使其更适合训练LLMs， ChemData包含7M的化学指示数据，这些数据对指数遵循能力是有效的。也可以释放整个文集给社区来推动化学语言模型的进程。

ChemLLM: 本文提出ChemLLM, 第一个开源的化学大模型，在保持完整的自然语言能力的同时达到多种化学能力ChemLLM有推动化学研究进程的可能性。

Two-stage Instruction Tunning Pipeline（二阶段的指令微调渠道）: 本文证明了此渠道对于用化学知识调整ChemLLM的有效性，成功赋予该语言模型以化学知识并且在一般场景有出色表现。这会激励随后的科学语言模型的训练。

3.ChemData :化学指令微调数据集

本数据集包括三个基本类别：分子、反应和特定领域任务。这些类别概括了化学研究的基本组成部分。我们将深入研究数据的不同来源，概述每个类别中的具体任务，并介绍为组件这个数据集而采取的细致过程。

# 原始数据汇总(Collection of Raw Data)

我们汇总了大量来自互联网的化学数据，涵盖广泛的化学领域知识，符合三个主要任务类别：分子、反应和领域任务。

## **Molecules:**

分子类别是理解和识别分子结构及其特性不可或缺的一部分。它包括四个关键领域：分子识别、分子特性预测、分子生成和分子描述。

### Molecular Recognition:

涉及各种分子表征之间的转换，如SMILES、IUPAC名称和化学式。

### Molecular Properties Prediction(分子特性预测):

侧重于预测各种分子属性，包括可溶性和血脑屏障通透性（blood-brain barrier permeability）。

### Molecular Generation:

旨在以SMILES格式设计和生成符合特定属性标准的分子结构。

### Molecular Captioning(分子描述):

是为分子结构生成的描述性文本，提供“描述”，以清晰的自然语言格式详细说明分子的特征、功能或其他相关信息。

## Reactions:

反应类别对于解读化学变化至关重要，包括反应产物预测、反应产率预测、反应条件选择和逆合成。这些任务中的每一项都对了解化学反应的动态和结果起着至关重要的作用。

### 反应产物预测（Reaction Product Prediction）

旨在根据所涉及的反应物预测化学反应的结果。

### 反应产量预测（Reaction Yield Prediction）

侧重于估计反应产物的产量。

### 反应条件选择（Reaction Condition Selection）

决定了反应应进行的最佳条件，以最大限度地提高产率和效率。因素包括温度、压力、催化剂和溶剂。

### 逆向合成（Retro-synthesis）

需要从目标分子进行反向分析，以推断可能的反应物和合成途径。

## 领域任务（Domain Tasks）

除了以分子和反应为中心的任务外，我们的数据集还包括显著扩展LLMs多功能性的领域特定任务，包括化学编程、领域问答、文献翻译和反应设计。

### 化学信息编程（Cheminformatic Programming）

使LLM具备理解和生成化学信息代码的技能，从而直接促进化学分析和研究工作流程。

### 领域问答（Domain Q&A）

借鉴了来自教科书的一般化学知识，旨在建立LLM解决化学领域各种问题的能力，从基本概念到高级主题。

# 3.2指令构筑

为了克服训练化学语言模型时，由于分子的特殊表现和数据库高度结构化的特性，创新战略是将其转变地更适应LLM训练。我们引入了一个新的渠道（pipeline），将这种正式化的化学数据转换为可访问的自然语言格式，确保重要化学信息的保存。我们的方法使用“种子模板”（seed template）。

它实施了’Play and Playwrights’strategy以创建单轮和多轮对话场景，显著增强了训练数据集的多样性。虽然该流程是针对化学数据设计的，但其基本原理具有适应性。它们可以应用于其他科学领域，标志着跨越广泛科学学科的LLMs研究新阶段的开始。

将结构化化学数据转换为适合培训LLM的指令微调数据涉及解决两个关键挑战：1）模板的多样性；2）将化学逻辑和推理纳入问答对。

为了应对第一个挑战，我们首先根据任务的特定要求制作了一个种子模板，如图5所示。

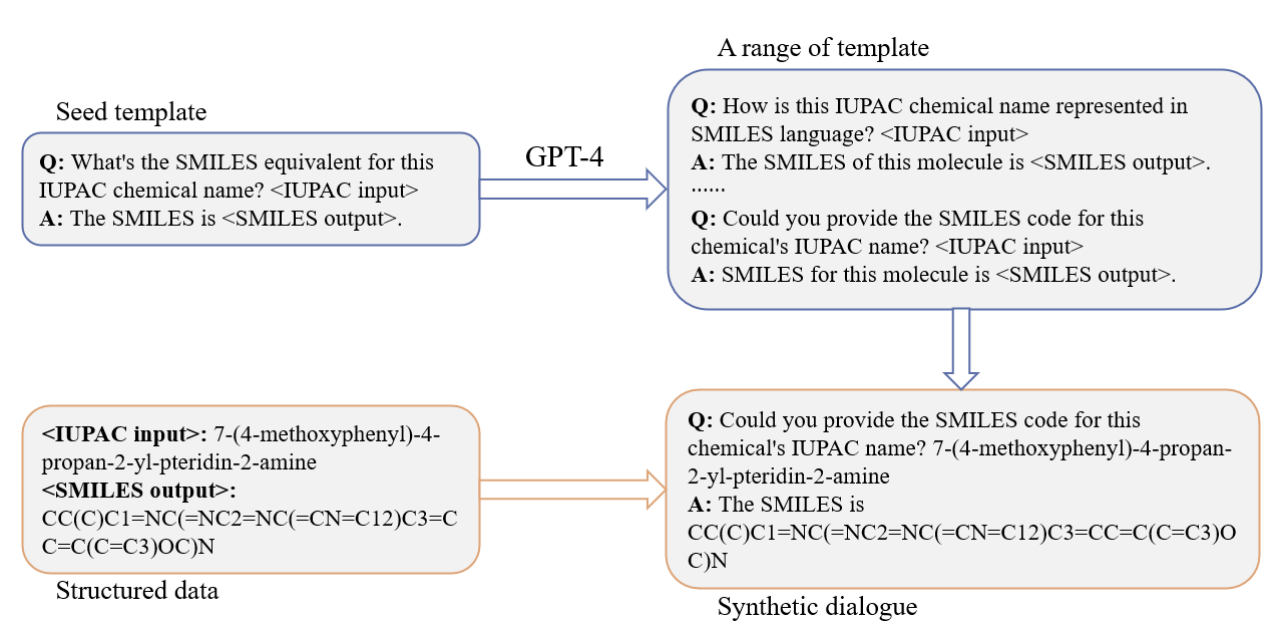


图5：种子模板提示技术的流程

在此基础上，我们利用 GPT-4生成一系列问答对，这些问答对虽然语义一致，但表达方式各不相同，创建了我们所说的多样化模板。这种策略促进了各种表达，提高了模型解释和响应不同指令的能力。对于每条结构化数据，都会从这个多样化的集合中随机选择一个模板来制定一个单轮对话样本。

针对第二个挑战，我们通过构建多轮对话来增强教学调整数据的语境丰富性和逻辑连贯性。在综合多轮对话数据时，我们坚持三个指导原则：内容的专业性和准确性、讨论与中心话题的相关性，以及随着领域话题的展开而扩展内容和深化对话的空间。我们的目标是模仿专家之间典型的动态交流和深入讨论，从而完善模型在专业领域问题上的推理、对话和分析能力。具体来说，我们引入了 "像剧作家一样演戏"（CoT）式的提示技术，指导 GPT-4 在一轮对话中遵循上述原则，在 "提问 "和 "回答 "之间编排 "剧本"。在 "扮演剧作家 "任务中，ChatGPT 分析讨论主题，塑造可信的角色和场景，然后生成一系列简洁的交流，深入而广泛地探讨主题。这种 "像剧作家一样演戏 "的方法使我们能够收集到丰富多样、高度专业化的多轮对话数据，而且对话轮数极少，减少了对人工干预的需求。

## 分子任务（Molecular Tasks）：

我们开发了一种种子模板方法，用于处理分子相关任务，将结构化数据转换为连贯的教学对话。这种方法以分子名称的转换为例，我们从数据库中提取“IUPAC名称、通用名称、SMILES”等结构化数据。然后，我们将这些数据字段组织成针对特定任务的问答格式。例如，“IUPAC名称”可能是问题，以“SMILES”为答案。

我们使用ChatGPT来创建问答对，在对话提示中构建任务和输入，并在访问中制定答案。这对初始配对构成了我们种子模板的基础。为了加强这一点，我们让ChatGPT来转述初始问题和答案，生成40个不同的模板，反映现实世界和学术文本的多样性。这允许我们为共享相同字段的结构化数据集组装160多个模板的集合。在制作说明中，我们为每个结构化数据点选择一个随机模板，将其转换为自然语言问答对。

## 反应任务（Reaction Tasks）：

在化学反应数据条目领域，关键领域表现出高度的均匀性，包括反应物、产物、产率和反应条件。其中，反应条件数据的特点是其不同的格式和缺失值（presence of missing values）的存在。为了解决这个问题，我们为反应条件设计了一个专门的模板，以适应缺失的值，促进将此类数据转换为标准化的自然语言描述。随后，采用种子模板方法，我们为各种预测目标量身定制了不同的模板，从而能够构建有针对性的指令。

## 领域任务（Domain Tasks）：

在构建领域任务指令时，我们的方法主要采用了’play-as-playwrights’的指令技术，该技术将大量领域文献文本和研究主题转化为建设性的多轮对话数据，旨在促进实质性讨论。

为了加强多回合对话的逻辑一致性，特别是与数据集的逻辑一致性，我们引入了一种条件屏蔽策略，以增强响应的逻辑基础。这需要最初在“问题”阶段隐藏关键条件，然后在“答案”阶段需要澄清，从而模拟专家话语。

这项创新解决了数据集缺乏思维深度的简单化答案的趋势，这破坏了模型的特定领域推理技能。通过浏览结构化的查询和演绎过程，此功能旨在使LLMs与专家的严格分析保持一致。

然后，我们从ChemXiv、LibreText Chemistry和维基百科化学门户网站汇总了特定领域的教科书数据，以综合特定领域多回合对话的主题。 预计这种方法将赋予我们的模型更广泛的深入领域知识并提高其阅读理解能力。

4.两阶段的指令微调流程

为了提高语言模型在专业领域的能力，我们采用了一种新颖的两阶段指令微调流程，如图 2 所示，以 ChemLLM 的开发为例，该模型基于Transforemer的 InternLM2-Base-7B，该模型精通中英文，拥有大量 4096 个词的上下文窗口，非常适合复杂任务，尤其是化学学科。

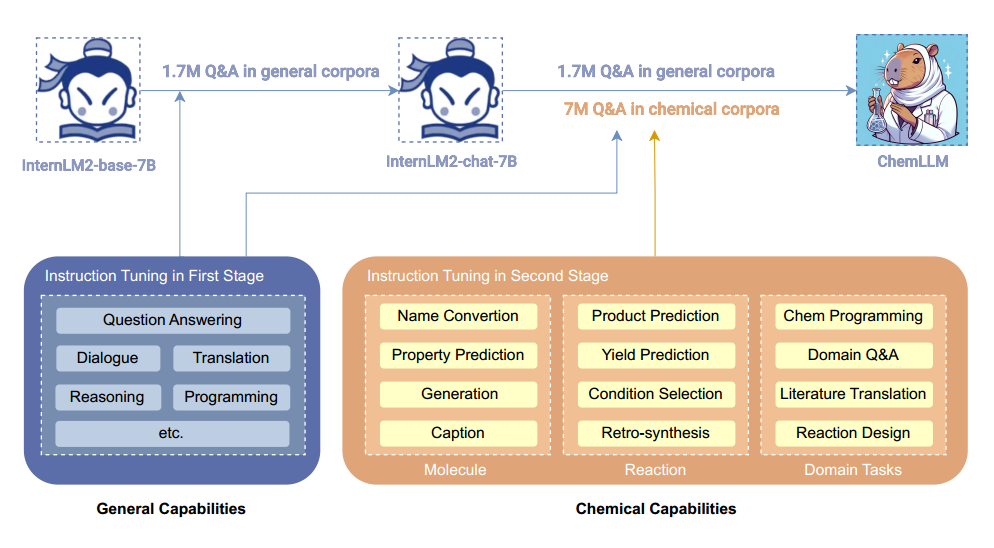


图2：模型训练的主要流程

流程的初始阶段用了170万个不同示例的综合语料库，增强了模型的一般语言能力。这个阶段拓宽了模型对语言细微差别和对话结构的理解，并为专业知识获取奠定了坚实的基础。

精炼的模型，以下简称InternLM2-Basechat-7B，经历了第二阶段，通过集成我们精心策划的以化学为中心的语料库数据集ChemData，它进一步专业化。这个两阶段的指令微调流程显著放大了模型的特定领域性能，明确区分了通用的InternLM2-Chat-7B和化学专用的ChemLLM。

# 初始阶段：

这个阶段对模型训练至关重要，因为它利用FireFly、OpenOrca和UltraChat等数据集将模型浸入多样化的语言环境中。这些数据集涵盖了广泛的对话，从FireFly中错综复杂的中文对话到OpenOrca和UltraChat中由LLMs生成的丰富的人工智能英语对话，使模型对人机交互和对话动态有了全面的了解。在这个阶段获得的技能，特别是在上下文理解和知识推理方面，对于模型产生连贯、上下文相关响应的能力至关重要。这种广泛的对话能力构成了后续特定领域专业化的基石。

# 第二阶段：

转入到第二阶段包括使用特定领域的合成数据集调整模型，以提高模型在化学领域的技术准确性和上下文相关性。这个阶段经过精心设计，旨在提高模型在各种子任务中的能力，从理解化学命名法到解释复杂的反应机制。

初始阶段奠定的坚实基础有助于从一般会话能力到重点领域专业知识的跨领域转移。这种两阶段的微调流程提高了模型的适应性和精确性，标志着从一般LLMs到专业领域不可或缺的知识转移的一次宝贵探索。将开放领域培训与特定领域知识相结合，为AI4Science的未来进步提供了一个有希望的途径，凸显了这种方法在弥合一般人工智能能力和专业领域需求之间差距方面的潜力。

5.ChemLLM性能

我们从三个主要方面对ChemLLM进行评估：1）专业化学任务；2）一般语言能力；3）多语言适应性

# 5.1专业化学任务（Professional chemical tasks）

## ChemBench：

为了严格评估语言模型的化学理解能力，我们推出了 ChemBench，这是一个创新的基准，采用分层框架，复制了专业化学分析的复杂性。它包括分子命名转换、分子标题和化学反应预测。最初的任务是名称转换，衡量模型在各种分子标识符之间的互换能力，如 SMILES 字符串、IUPAC 名称和化学式。

这项任务衡量语言模型对分子结构的理解能力，并通过考察基本化学知识为更高级的分析奠定基础。随后的“分子描述”任务则是对模型预测特定分子特征的考验，突出了模型对分子特性的洞察力。最复杂的任务是反应预测，评估模型预测化学反应结果的熟练程度，这需要对化学原理、反应动力学和分析思想进行全面综合。这些能力提供了可量化的分析，从基本水平到高级水平逐步评估化学语言模型对化学领域的理解。

## 评估：

我们使用 ChemBench 基准对各种大型语言模型（LLM）进行了比较分析，包括类似规模的模型以及GPT3.5和GPT-4。该基准包括三个具体任务：名称转换、性质预测和反应预测。

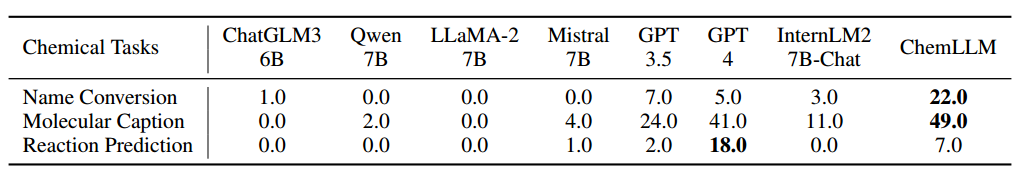


表1：在ChemBench上的性能比较

表 1 列出的结果表明，在所有评估任务中，ChemLLM 与类似规模的模型相比都表现出色。值得注意的是，在名称转换和分子描述任务中，ChemLLM 的成绩分别为22.0分和49.0分，超过了 GPT-4。尽管在反应预测任务中，ChemLLM 的表现并不突出，但它仍然超过了GPT-3.5，仅次于GPT-4.这些结果表明，通过对注入化学知识的指令进行微调，语言模型获得了显著的化学能力。相比之下，基础模型InternLM-7B-Chat在这些任务中的效果有限，这凸显了在模型训练过程中嵌入专业化学知识的重要性。

# 5.2一般语言能力

## 大规模多任务语言理解（MMUL ，Massive Language Understanding）：

MMUL是一项综合性、多层面的计划，旨在评估和改进语言模型在各种语言挑战中的性能。该基准涵盖57个学科，如STEM、人文和社会科学等，提供对全球知识和解决问题能力的广泛评估。

## GSM8K：

GSM8K 是一套被广泛认可的测试语言模型数学能力的测试工具，其中的问题需要进行2-8步的基本数学运算，可以测试模型的多步数学推理能力。

## 评估：

虽然化学LLMs是为化学查询量身定制的，但也不能忽视通用对话能力的必要性，而这正是当前化学信息语言模型经常缺乏的特征。熟练掌握对话和逻辑推理等通用领域的能力，可以增强专业模型的通用性和适用性。通过跨学科知识的视角，它可以更丰富地理解利基任务。

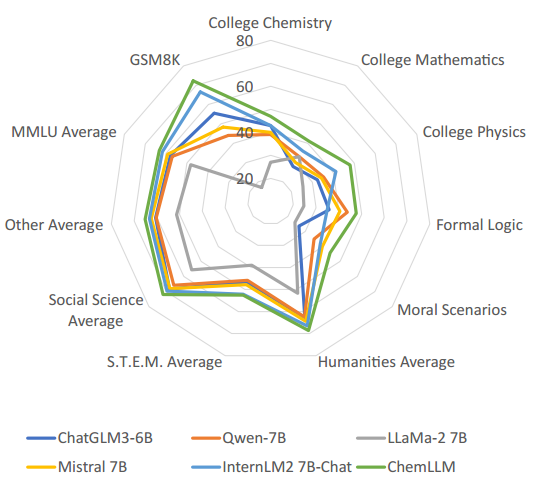


图3：MMLU和GSM8K结果雷达图

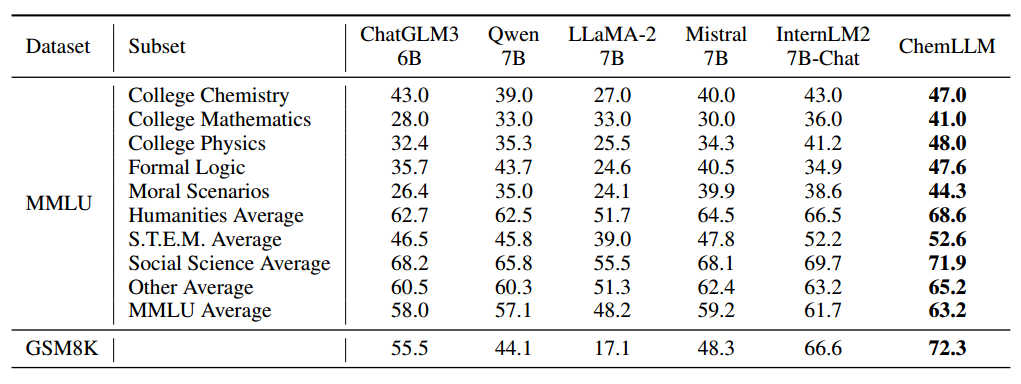


表5：通用能力：MMLU和GSM8K成绩

为了进一步评估 ChemLLM 的专业广度，我们使用 MMLU 基准进行了评估。我们的分析包括与同等规模的模型（如 ChatGLM3-6B、Qwen7B、LLaMA2-7B 和 Mistral-7B）以及 InternLM2-7B-Chat 进行比较，以衡量基础预训练的影响。

尽管ChemLLM接受的是以化学为中心的训练，但它在更广泛的科学学科中表现出了非凡的能力，这一点从它在大学物理和数学部分的优异表现中可以看出。这凸显了该模型在邻近科学领域的通用能力，证明了化学训练的价值。ChemLLM在形式逻辑部分的表现尤为突出，其成绩比基础InternLM27B-Chat模型高出12.8%，展示了其卓越的推理能力。

此外，ChemLLM 在道德情景部分的高分也表明，化学专业训练不仅没有降低其道德决策能力，反而为其增色不少。从人文科学、STEM 到社会科学，ChemLLM 在各个学科领域都有出色的表现，这充分证明了针对化学特定任务的集中训练不仅不会影响模型的一般任务表现，反而会丰富模型的一般任务表现，凸显了其全面而多才多艺的能力。详情请参见图 3 和表 5。

# 5.3多语言能力

## Chinese ChemQA and Chinese M&H ChemTest：

中文化学问答由开放式问答题组成，涉及化学领域的各种主题，从基本化学原理到有机合成和反应机理等更高级的概念。该数据集测试了模型理解和生成准确、详细的中文回答的能力，反映了对化学的深刻理解。中文 M&H 化学测试旨在评估模型对初中和高中课程中通常教授的基础到中级化学概念的掌握情况。它测试模型将理论知识应用于实际问题的能力，通过化学情景进行逻辑推理的能力，以及从一系列选项中选择正确答案的能力。

## 评估：

我们编制了两个数据集来评估 ChemLLM 在中文环境下的泛化能力。

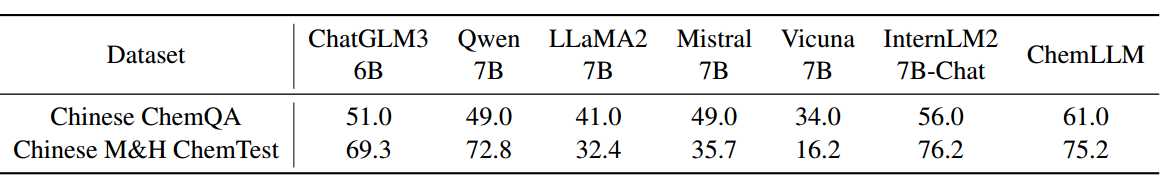


表2：中文化学学科基准测试结果

表 2 中的结果表明，ChemLLM 在两个数据集上都达到了令人称赞的准确度。这突出表明，ChemLLM 能够在不同语言环境中熟练地驾驭复杂的化学概念和术语，肯定了它作为化学领域多语言工具的实用性。特别是中文 M&H ChemTest 的评估结果表明，ChemLLM 可以有效地作为支持教育目标的强大工具。

## 5.4定性结果

## 化学相关的NLP任务：

 要深入了解我们的定性研究结果，请参阅附录 D 至 F，其中详细介绍了模型在化学相关 NLP 任务中的表现。其中包括文本翻译、化学信息编程和化学诗歌创作等。这些结果凸显了模型在各种 NLP 环境中对化学知识的细微理解和创造性应用。

## 伦理定性测试：

 我们认识到在部署化学语言模型时，伦理方面的考虑至关重要此在附录 G 中引入了伦理定性测试。该模块可评估模型在以下六个敏感领域是否符合人类价值观：化学武器、化学和药物安全、实验室安全、精神活性物质和受控物质，以及更广泛的科技伦理困境。这种方法可以衡量模型对技术的理解和对伦理影响的敏感性，解决未来由LLMs驱动的化学研究中固有的潜在风险和争议。

6.局限性

首先，将对把握复杂的分子结构和相互作用至关重要的分子图模式纳入模型是一个挑战。缺乏分子图的直接表示限制了模型理解分子空间构型的能力。其次，ChemLLM 遵循科学伦理的能力，尤其是在极端条件下生成反应的能力，引发了人们对不安全或不道德的实验方法的担忧。尽管我们已经开发并部署了一些有效的基于提示的缓解方法，但这些领域仍有待未来开发，以增强 ChemLLM 的功能和伦理管理。

7.结论

大型语言模型（LLM）极大地推动了化学应用的发展。然而，基于对话的化学模型却明显缺乏。为了弥补这一差距，我们引入了一种基于模板的指令构建方法，该方法可将结构化的化学知识转化为可访问的对话格式，用于训练语言模型。这种方法解决了利用结构化化学数据的难题，因为结构化化学数据通常会妨碍模型中对话的连贯性。我们的创新方法促成了 ChemLLM 的诞生，它是第一个专门用于化学领域的 LLM，能通过无缝对话交互处理各种化学任务。

ChemLLM 在分子识别、性质描述和反应预测等主要化学任务方面超越了 GPT-3.5，在其他领域也表现出了值得称道的多功能性。除了核心能力之外，ChemLLM 在化学领域的专业 NLP 任务方面也表现出色，例如文献翻译、化学信息编程和遵守研究伦理。我们希望我们的专业领域知识注入策略能够激励更多的工作，将 LLM 应用于科学领域。