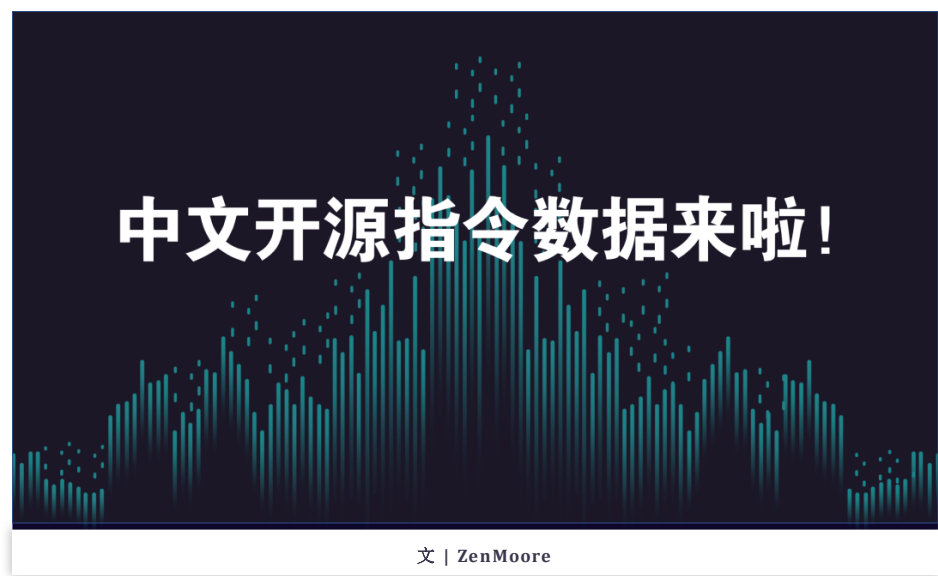


COIG：首个大规模、可商用的中文开源指令数据！

原创 ZenMoore 夕小瑶科技说 2023-04-21 11:21 发表于四川



ChatGPT 出现后的这几个月，整个学界和业界的疯狂想必大家都已经看到了。

然而，在背各种各样的动物还有山海经怪兽的英语单词的时候，其实不妨停下来想一想复现中文 ChatGPT 到底缺什么？缺大模型吗？缺工程经验吗？缺 trick 吗？缺算力吗（确实缺...）？

先来看看比较火的那些“动物”做了什么：Alpaca[1]（羊驼）用语言模型生成了一些指令数据，Baize[2]（白泽）用 ChatGPT 生成了一些对话数据，Vicuna[3]（小羊驼）用 ShareGPT 数据（ChatGPT 用户分享的对话记录）复现了不少 ChatGPT 效果.....

种种迹象表明，数据还是那个最珍贵的资源。那什么样的数据更重要呢？

再来看看 ChatGPT 的关键几步：(1) 大规模语料预训练；(2) 监督指令精调(Supervised Instruction Fine-Tuning)；(3)基于人类反馈的强化学习(RLHF)

对于(1)来说，中文其实不缺，从互联网上爬就是了；对于(3)而言，人工打分标注的成本相对而言低一些。所以，主要是指令数据的稀缺，导致语言模型精调的时候无法获得足够的监督训练，而且，还影响了后续 RLHF 的输入指令集的筹备。要知道，指令精调给模型带来的指令遵循能力，是具有很强的泛化性的，这样模型才能够更好地泛化到用户多种多样的指令上去（参考[4]）。

可惜的是，相比于英文来说，开源的、大规模的中文指令数据实在是太少了.....

因此，本文给大家推荐这么一批珍贵的中文数据：Chinese Open Instruction Generalist（COIG）第一期！

第一期总共发布了 5 个子数据集，包括翻译指令、考试指令、人类价值观对齐指令、反事实修正多轮聊天、Leetcode指令，总计 191k 数据，聚焦中文语料、数据类型多样、经过了人工质检与修正、数据质量可靠，而且可以商用。

论文标题：
Chinese Open Instruction Generalist: a Preliminary Release

论文机构：
北京智源人工智能研究院等

论文链接:

<https://arxiv.org/pdf/2304.07987.pdf>

数据链接:

<https://huggingface.co/datasets/BAAI/COIG>

经过人工验证的翻译通用指令（67,798）

作者对三个数据集进行了翻译工作，包括具有 1,616 个任务描述和示例的 Super-Natural Instructions 数据集，175 个种子任务的 Self-Instruct 数据集，以及 66,007 个指令的 Unnatural Instructions 数据集。整个翻译过程分为自动翻译、人工验证和人工修正三个阶段，以确保翻译结果的准确性和可靠性。

在自动翻译阶段，作者将指令和实例的输入输出组合在一起，然后使用 DeepL 进行翻译。

在人工验证阶段，作者为注释者定义了四个标签，根据指令是否可用和需要的修正程度进行分类。作者使用两阶段质量验证方法进行人工验证，第一阶段由经验丰富的质量检查员进行验证，只有正确率超过 95% 的案例才能进入第二阶段。在第二阶段，专家质量检查员从总语料库中随机抽取 200 个案例进行验证。

在人工修正阶段，注释者需要将翻译后的指令和实例纠正为正确的中文三元组{指令，输入，输出}，而不仅仅是保持翻译的准确性。这是因为在 unnatural instructions 中存在事实错误，这可能会导致 LLMs 出现 hallucination。作者同样使用两阶段质量验证方法进行人工修正，第一阶段的正确率为 97.24%。

人工注释的考试指令（63,532）

中国的高考、中考和公务员考试题目中包含各种问题类型和详细的分析，这些考试可以用来构建思维链（CoT）语料库用于增强模型推理能力。作者从这些考试中提取了六个信息元素，包括指令、问题背景、问题、答案、答案分析和粗粒度学科。这些语料库中的六个主要学科是语文、英语、政治、生物、历史和地质。数学、物理和化学问题很少在语料库中出现，因为这些问题通常包含难以注释的复杂数学符号。

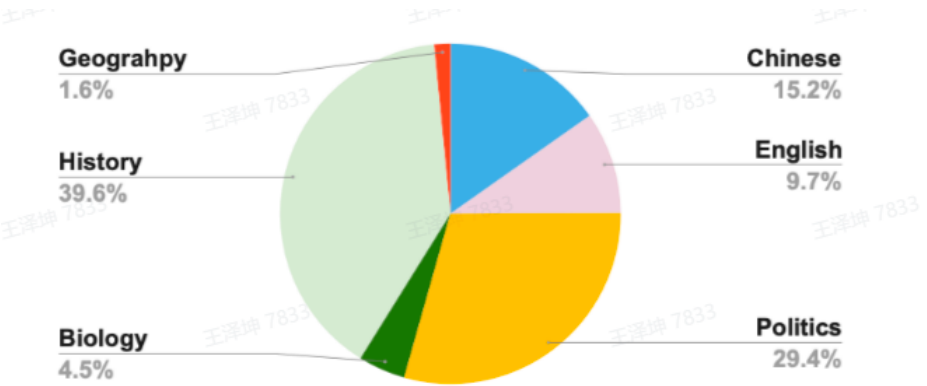


Figure 1: The percentage of instructions of different coarse-grained subjects.

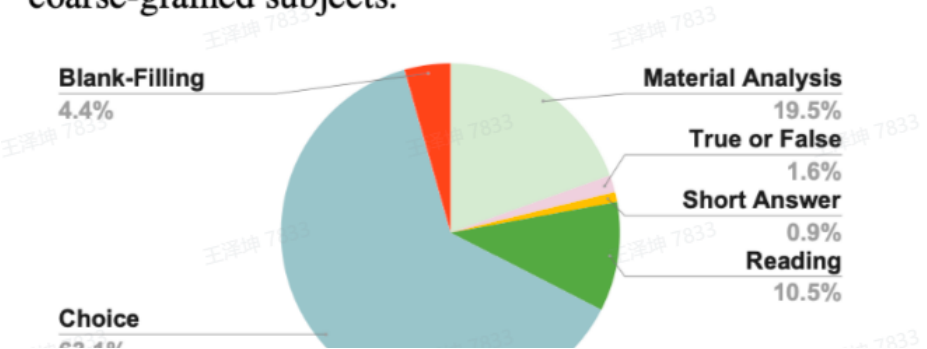


Figure 2: The percentage of instructions in different question formats.

人类价值观对齐指令（34,471）

为了尊重和反映不同文化背景所带来的主要差异，COIG数据集中的价值观对齐数据被分为两个独立的系列：

一组展示中文世界共享人类价值观的样本。作者选择了50个指令作为扩充种子，并使用中文世界通用的价值观对齐样本，生成了3,000个结果指令。另外一些展示特定区域文化或国家特定人类价值观的样本集。以下是数据示例：

| Instructions Sample |
|--|
| Instruction: 假如你是一名校长，最近接到不少学生和家長反映，学校操场上某些同学总是进行踢足球、打篮球等活动，严重影响了其他同学的学习和休息。请你制定一份公告，呼吁全体学生遵守校規，有序使用学校場地。 |
| Input: 公告起草委员会已经开会商讨该公告的内容，加入你的建议，你需要在公告中提供具体制度，呼吁全体学生遵守校規，保证课堂和自习室的安静和干净。 |
| Output: 全校同学在使用学校的场地设施，应自觉遵守学校有关規定和管理制度，保证场地的有序和安全，维护学校正常的教学、体育训练、课外活动等各项工作的进行。学校操场是同学们锻炼身体的场所，但也是学习的场所，应该保持安静，不影响其他同学的学习和休息。进入球场、相邻广场玩耍一律需老师主持，并根据实际情况合理规划使用时间；球鞋等装备必须符合安全要求。严禁未经校方批准私自使用学校未开放的区域及相关设施；公共体育设施的使用，要坚持先预约后使用的原则。我们希望在学校的各个场地内，能够自律自觉，有序文明的使用，并与教职工一起共同创建良好的校园环境。 |

反事实修正多轮聊天（13,653）

反事实修正多轮聊天数据集（CCMC）是基于CN-DBpedia知识图谱数据集构建的，旨在解决当前LLM中出现的幻觉和事实不一致的问题。数据集包含约13,000个对话，每个对话平均有5轮，共约65,000轮聊天。这些对话是在学生和教师之间进行的角色扮演聊天，他们在对话中参考相关的知识。

| Multi-round Chat Sample |
|---|
| Premise: 刻点小食蚜蝇（Paragus tibialis Fallen）是双翅目昆虫纲食蚜蝇科的一种捕食性天敌，属于动物界节肢动物门。它会对棉花等作物造成危害。以蚜虫为寄主。该蚜蝇的成虫在4月份开始在绿肥、蚕豆和榆、槐、柳、木槿以及杂草等植物上产卵繁殖，并于5月上旬迁入棉田。发生数量较少。卵期约3-4天，幼虫期8-9天，蛹期12-14天，由卵到成虫历期23-27天，在6-7月期间卵期一般2-3天，幼虫期6-7天，蛹期6-7天，由卵到成虫历期14-17天。 |
| Confusion: 食蚜蝇科，英文俗称syrphidflies或flowerflies。形似蜂，腹节上常有黄黑相间的斑纹。R脉与M脉间有一游离的伪脉（spuriousvein）。幼虫体平滑或有圆锥突起及刚毛；捕食性种类的体前端尖，后端平截；粪食性种类的腹末有长的呼吸管。成虫通常在阳光下取食花蜜和花粉；飞翔时能在空中静止不移又忽然突进。腐食性和粪食性的幼虫生活在木、粪便和腐敗动植物体中；捕食性种类取食蚜虫、介壳虫、粉虱和叶蝉等。常见种类：纤腰巴食蚜蝇Baccha Maculata Walker。 |
| Round 1: Q: 什么是刻点小食蚜蝇？ A: 刻点小食蚜蝇是双翅目昆虫纲食蚜蝇科的一种捕食性天敌，以蚜虫为寄主，会对棉花等作物造成危害。 |
| Round 2: Q: 您刚才不是说刻点小食蚜蝇是属于食蚜蝇科的吗？那syrphid flies或flower flies难道不是刻点小食蚜蝇的英文名吗？ A: syrphid flies或flower flies是食蚜蝇科的名字，而不是刻点小食蚜蝇的英文名。虽然刻点小食蚜蝇也属于食蚜蝇科，但是不同品种的昆虫是有不同的学名和英文名的。 |
| Round 3: Q: 您刚才所讲的关于食蚜蝇科的形态特征，如形似蜂，腹节上常有黄黑相间的斑纹；幼虫体平滑或有圆锥突起及刚毛；捕食性种类的体前端尖，后端平截；粪食性种类的腹末有长的呼吸管等是否也与刻点小食蚜蝇的形态特征相似？ A: 食蚜蝇科的形态特征您说得没错，但是刻点小食蚜蝇的形态特征可能与食蚜蝇科其他成员有所不同，因为每个物种都有其独特的形态特征，这是我们进行它们的科学分类和分类学研究时所依据的标准。刻点小食蚜蝇的形态特征可以根据其学名或者其他资料来具体分析。 |
| Round 4: Q: 您刚才所讲的关于食蚜蝇科的生活习性，如成虫通常在阳光下取食花蜜和花粉；飞翔时能在空中静止不移又忽然突进；腐食性和粪食性的幼虫生活在木、粪便和腐敗动植物体中；捕食性种类取食蚜虫、介壳虫、粉虱和叶蝉等是否也与刻点小食蚜蝇的生活习性相似？ A: 您所说的食蚜蝇科的生活习性均是该科中许多物种的表现，并不是所有食蚜蝇科中的成员都会表现这些特征，刻点小食蚜蝇在食性和习性方面可能与其他食蚜蝇科的成员有所不同。因此，要了解刻点小食蚜蝇的生活习性，需要查阅具体的资料。 |
| Round 5: Q: 能否请您重新给出一下刻点小食蚜蝇的介绍？ A: 刻点小食蚜蝇是一种捕食性天敌，属于双翅目昆虫纲食蚜蝇科中的一种。其幼虫以蚜虫为寄主，能对棉花等作物造成危害。与食蚜蝇科中的其他成员相比，刻点小食蚜蝇的特点可能在于它在祸害棉田过程中的生命周期以及其形态特征等方面。注意要区分刻点小食蚜蝇和食蚜蝇科，这两者是不同的概念。此外，刻点小食蚜蝇与syrphid flies或flower flies这些名称也不是等同的关系。 |

Leetcode 指令（11,737）

考虑到与代码相关的任务可能有助于LLM能力的涌现，作者从CC-BY-SA-4.0许可下的2,589个编程问题中构建Leetcode指令。这些问题包含问题描述、多种编程语言和解释（其中 834个问题尚没有解释）。编程语言分类统计如下图所示：

Table 5: Statistics of Leetcode Instructions. Task types

C2T and T2C refer to code-to-text and text-to-code, respectively. And *e.* is the abbreviation for “explanation”. Programming languages with less than 50 instructions are merged into the “Others” class.

| Programming Language | Task Type | | | | All |
|----------------------|------------|-----------|------------|-----------|-------|
| | C2T w/o e. | C2T w/ e. | T2C w/o e. | T2C w/ e. | |
| C | 8 | 76 | 12 | 89 | 185 |
| C# | 8 | 56 | 6 | 58 | 128 |
| C++ | 168 | 943 | 180 | 963 | 2254 |
| Go | 175 | 1008 | 164 | 899 | 2246 |
| Java | 213 | 989 | 193 | 983 | 2378 |
| JavaScript | 16 | 172 | 29 | 153 | 370 |
| Python3 | 198 | 995 | 208 | 981 | 2382 |
| Rust | 46 | 252 | 39 | 252 | 589 |
| SQL | 35 | 6 | 30 | 5 | 76 |
| TypeScript | 98 | 454 | 82 | 450 | 1084 |
| Others | 2 | 20 | 3 | 20 | 45 |
| All | 967 | 4971 | 946 | 4853 | 11737 |

结语

COIG 是目前首个指令类型丰富且可商用的中文指令集，作者欢迎大家加入共建，一起扩充指令数据集类型和规模！



卖萌屋作者：ZenMoore
智源实习生 ☺️，爱数学爱物理爱 AI ☺️ 想从 NLP 和 System-2 出发探索人工认知的奥秘 ☺️
📧！即将进入 PhD 申请季，微信 ☎️ zen1057398161 嚶其鸣矣，求其友声 🌟！
作品推荐

- 1. [一文跟进Prompt进展！综述+15篇最新论文逐一梳理](#)
- 2. [图灵奖大佬+谷歌团队，为通用人工智能背书！CV 任务也能用 LM 建模！](#)
- 3. [以4%参数量比肩GPT-3！Deepmind 发布检索型 LM，或将成为 LM 发展新趋势！？](#)



后台回复关键词【入群】
加入卖萌屋NLP、CV、搜推广与求职讨论群



夕小瑶科技说
一线研究员解读人工智能前沿信息，深度分析AI行业趋势。知乎ID：夕小瑶，机器学习...
471篇原创内容

公众号

参考文献

[1] Alpaca: https://github.com/tatsu-lab/stanford_alpaca

[2] Baize: <https://arxiv.org/pdf/2304.01196.pdf>

[3] Vicuna: <https://vicuna.lmsys.org/>

[4] <https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1>

喜欢此内容的人还喜欢

全球最大ChatGPT开源平替来了！支持35种语言，写代码、讲笑话全拿捏
新智元



一网打尽！国内外大模型调研汇总
NewBeeNLP

| Model | Size | Open Source | Performance |
|---------|-------|-------------|-------------|
| GPT-4 | 175B | No | 98.5 |
| GPT-3.5 | 175B | No | 97.5 |
| Flan-T5 | 111B | Yes | 96.5 |
| Flan-T5 | 30B | Yes | 95.5 |
| Flan-T5 | 6B | Yes | 94.5 |
| Flan-T5 | 1.1B | Yes | 93.5 |
| Flan-T5 | 0.3B | Yes | 92.5 |
| Flan-T5 | 0.1B | Yes | 91.5 |
| Flan-T5 | 0.03B | Yes | 90.5 |
| Flan-T5 | 0.01B | Yes | 89.5 |

一文详解Prompt学习和微调（Prompt Learning & Prompt Tuning）
机器学习算法与自然语言处理

