

YuLan-Chat: 基于多阶段课程学习的大语言模型

周 昆¹⁾ 朱余韬²⁾ 陈志朋²⁾ 毛科龙²⁾ 陈文通²⁾ 陈昱硕²⁾ 孙一丁²⁾ 曹 乾²⁾
王 磊²⁾ 张 蕾²⁾ 庞新程²⁾ 谢曙方²⁾ 赵 鑫²⁾ 窦志成²⁾ 林衍凯²⁾ 毛佳昕²⁾
宋睿华²⁾ 陈 旭²⁾ 徐 君²⁾ 胡 迪²⁾ 严 睿²⁾ 黄文炳²⁾ 魏哲巍²⁾ 文继荣^{1),2)}

¹⁾(中国人民大学信息学院 北京 100872)

²⁾(中国人民大学高瓴人工智能学院 北京 100872)

摘 要 近年来,大语言模型已成为研究热点。其在大规模数据上预训练之后,具有强大的少样本和零样本上下文学习能力,能够便捷地用于许多真实场景复杂任务。然而,对大语言模型进行从头到尾的开发和训练,可参考的实现较少;且存在较难习得的知识,如长尾知识相关数据、复杂指令、难区分的负例等。为填补该领域空白,并强化对较难掌握数据的学习,本文提出了多阶段的课程学习方法,针对以上三种典型数据,使用了:(1)迭代增强长尾知识的预训练课程;(2)由简单到复杂的指令微调课程;(3)由易到难的人类对齐课程,完成了 YuLan-Chat 从头开始的整个训练流程。本文在四个与大语言模型基础能力和人类对齐能力相关的中英文评测基准上对 YuLan-Chat 进行评测,结果表明该模型能够在大部分场景下优于基线模型。分析实验进一步表明了该课程学习方法在 GAOKAO 和 AlignBench 评测基准上,能够分别提升模型 9.7% 和 18.9% 的答案预测准确率。

关键词 大语言模型;课程学习;预训练;指令微调;人类对齐

中图法分类号 TP391 DOI号 10.11897/SP.J.1016.2025.00001

YuLan-Chat: A Large Language Model Based on Multi-Stage Curriculum Learning

ZHOU Kun¹⁾ ZHU Yu-Tao²⁾ CHEN Zhi-Peng²⁾ MAO Ke-Long²⁾ CHEN Wen-Tong²⁾
CHEN Yu-Shuo²⁾ SUN Yi-Ding²⁾ CAO Qian²⁾ WANG Lei²⁾ ZHANG Lei²⁾
PANG Xin-Cheng²⁾ XIE Shu-Fang²⁾ ZHAO Xin²⁾ DOU Zhi-Cheng²⁾ LIN Yan-Kai²⁾
MAO Jia-Xin²⁾ SONG Rui-Hua²⁾ CHEN Xu²⁾ XU Jun²⁾ HU Di²⁾ YAN Rui²⁾
HUANG Wen-Bing²⁾ WEI Zhe-Wei²⁾ WEN Ji-Rong^{1),2)}

¹⁾(School of Information, Renmin University of China, Beijing 100872)

²⁾(Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872)

Abstract In recent years, large language models (LLMs) have become a research hotspot in the field of natural language processing. After pre-training on large-scale data, these models exhibit

收稿日期:2024-04-24;在线发布日期:2024-09-27。本课题得到国家自然科学基金(62222215,U2001212)、北京市自然科学基金(4222027)资助。周 昆,博士,主要研究领域为自然语言处理、信息检索。E-mail: francis_kun_zhou@163.com。朱余韬,博士,博士后研究员,主要研究领域为自然语言处理、信息检索。陈志朋,博士研究生,主要研究方向为自然语言处理。毛科龙,博士研究生,主要研究方向为自然语言处理、信息检索。陈文通,博士研究生,主要研究方向为自然语言处理。陈昱硕,硕士研究生,主要研究方向为自然语言处理。孙一丁,博士研究生,主要研究方向为自然语言处理、信息检索。曹 乾,博士研究生,主要研究方向为自然语言处理、信息检索。王 磊,博士研究生,主要研究方向为自然语言处理、信息检索。张 蕾,博士研究生,主要研究方向为自然语言处理、信息检索。庞新程,硕士研究生,主要研究方向为自然语言处理、计算机视觉。谢曙方,硕士研究生,主要研究方向为自然语言处理。赵 鑫(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索。E-mail: batmanfly@gmail.com。窦志成(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为信息检索。E-mail: dou@ruc.edu.cn。林衍凯,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理。毛佳昕,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为信息检索。宋睿华,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索、多模态。陈 旭,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索。徐 君,博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索。胡 迪,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为多模态、计算机视觉。严 睿,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索。黄文炳,博士,副教授,中国计算机学会(CCF)会员,主要研究领域为机器学习。魏哲巍,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习。文继荣(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为信息检索和机器学习。E-mail: jrwen@ruc.edu.cn。

strong few-shot and zero-shot in-context learning capabilities, making them highly applicable to complex tasks in real-world scenarios. However, there are limited references available for the training of LLMs from scratch. Moreover, there are challenges throughout the training process, such as learning the data with long-tail knowledge, complex instructions, and indistinguishable negatives. In this work, we propose a multi-stage curriculum learning approach, which tackles the aforementioned challenges using three curriculum learning strategies respectively: (1) an iterative curriculum for enhancing long-tail knowledge in pre-training, (2) a simple-to-complex curriculum for instruction tuning, and (3) an easy-to-difficult curriculum for human alignment learning. We employ these curricula to sequentially train our YuLan-Chat from scratch. Four Chinese and English benchmark datasets are used to evaluate YuLan-Chat. The results demonstrate that YuLan-Chat can mostly outperform baseline LLMs. Furthermore, our experiments show that the three-stage curriculum can yield performance improvements of 9.7% (GAOKAO) in pretraining, 22.2% and 18.9% (AlignBench) in instruction tuning and human alignment stages, respectively.

Keywords large language models; curriculum learning; pre-training; instruction tuning; human alignment

1 引 言

近年来,随着计算资源的不断升级,研究者们也在尝试继续增大语言模型的参数规模和训练数据量,以进一步提升其在各类下游应用和任务上的效果^[1-2]。当参数和预训练数据量达到一定规模后(例如百亿级别),预训练的大语言模型(Large Language Model, LLM)会涌现出一系列新能力^[3-4]。其中,上下文学习(In-Context Learning)是大语言模型最有代表性的能力之一^[5],这种能力使得大语言模型能够利用少量样本适应任务并以自然语言生成的形式将其完成。这一过程无须参数更新,因而大幅降低了大语言模型应用于真实世界多样任务和场景的计算成本,使得通用人工智能(Artificial General Intelligence, AGI)成为可能^[6]。此外,大语言模型还具有接近人类的文本生成和推理能力,其能够基于用户输入的提示解决真实场景下众多复杂任务,甚至在许多职业资格考试中取得超越人类平均水平的能力^[7]。得益于以上优点,大语言模型为自然语言处理领域带来了一场技术革命,研究者们基于已有的大语言模型开发出了一系列应用,包括人类聊天助手(如 ChatGPT)和办公辅助工具(如 Copilot)等,提升了用户的使用体验。

由于大语言模型具有较高的实际应用价值,开发能力更强的大语言模型受到了学术界和工业界的广泛关注。目前,大语言模型普遍采用自回归语言

模型架构(Autoregressive Language Model)^[3],并基于下一词元预测任务(Next Token Prediction)从大规模文本数据中学习如何对文本进行建模。在这一过程中,如何选择并安排大规模文本数据用于模型训练,是决定模型最终能力强弱的重要因素^[4]。在已有的工作中,大语言模型训练过程通常可以分为三个阶段,即大规模预训练(Pre-training)、指令微调(Instruction-tuning)和人类对齐(Human Alignment)。具体来说,在预训练阶段,大语言模型需要在规模庞大的多源文本数据上(例如 1T 词元网页)进行训练,使得大语言模型学习自然语言的表达形式和大量的世界知识。之后,在指令微调阶段,大语言模型在规范化的指令数据上进行训练,学习如何遵循自然语言指令和提示,以文本生成的形式完成任务^[8]。在人类对齐阶段,通过比较基于人类价值标注的正负样本,大语言模型可以学习到人类的偏好^[9]。大语言模型的训练作为一个复杂的系统工程,其涉及大量的细节设计和实践考量,然而,目前可以参考的资料却较少。这一方面是由于学术界缺乏训练所需的巨大算力,无法对该过程进行有效探索;另一方面则是工业界通常将训练过程视作核心技术,不对外进行公开。

为填补该领域的空白,本文关注于从头训练出通用大语言模型——YuLan-Chat,使其能够具备优秀的中英文自然语言理解与生成的能力,以适应中英文互联网场景下多样且复杂的应用和任务需求。为了训练大语言模型,现有工作普遍收集互联网上

已有的开源数据集,然后直接按照上述的三阶段训练方法,完成整个训练过程。在这一过程中,每阶段均会混合不同来源或不同任务的数据,为模型补充丰富且多样的知识。但是,由于数据本身的知识分布和难易程度不一致,直接将其混合进行训练可能不适合于大语言模型完全习得其中知识,导致其中某些重要知识较难被掌握,进而限制其表现。具体来说:(1) 现有的预训练数据集通常来源于真实网页,其蕴含的知识存在长尾分布问题,某些重要的知识在数据中出现频率较低,难以被模型充分习得,进而导致使用该类知识时容易出现幻象(Hallucination)^[10]; (2) 现有的指令微调数据集普遍由不同类型的指令数据组成,这些指令复杂程度不同但未得到区分,使得模型难以学习其中复杂的指令数据,而倾向于学习简单的部分,最终导致模型无法有效理解并遵循指令^[11]; (3) 现有的人类对齐数据集中包含的正负样本的难易程度不同,其中较难区分的样本并未在训练过程中受到更多关注,从而导致模型不能较好地理解这些困难场景,甚至违反人类对齐原则。

受人类的教育过程启发,本文基于课程学习的思想^[12],在大语言模型的三个训练阶段中分别设计了多级课程,使得大模型能够像人类一样循序渐进地进行知识学习,从而在每个阶段更好地习得较难掌握的知识。对于预训练、指令微调和人类对齐这三个阶段,设计的课程分别关注于下述重要但较难掌握的知识相关数据:(1) 涉及长尾知识的预训练数据;(2) 任务或输入形式复杂的指令;(3) 难以区分的正负例对。具体来说,本文首先提出了迭代增强长尾知识的预训练课程,通过挖掘大语言模型当前未较好掌握的长尾知识,使用 TF-IDF 算法检索与这些知识相关的预训练数据,将其加入到当前课程中。多次迭代上述课程数据筛选和预训练的过程后,即可实现对长尾知识的针对性强化。其次,本文提出了由简单到复杂的指令微调课程,首先采用已有的自然语言处理任务数据进行训练以补充基础知识,然后在合成的复杂指令和多轮对话理解指令上训练,强化对这些较复杂指令的理解。此外,本文还提出了由易到难的人类对齐课程,其基于直接偏好优化的损失函数为每条正负例样本的难易程度打分,然后设置阈值筛选出当前课程中需要学习的数据。通过多次迭代该课程并逐步缩小阈值,即可实现由易到难、循序渐进地学习难以区分的正负样本对。

基于上述课程学习方法,本文在 96 张 A800 GPU 上从头开始对 YuLan-Chat 模型进行了多阶段训

练。其采用自回归语言模型框架,具有 12B 参数,且预训练数据包含 1.6 T 中英文词元,指令数据达到 4200 万条,人类对齐数据达到 20 万条。本文设计的多阶段课程方法使得 YuLan-Chat 模型能够充分地习得这些数据中丰富的知识,使其具备优秀的中英文理解与生成能力、多轮对话与交互能力以及人类对齐能力。为了对其进行验证,本文在基础语义理解、指令遵循、对话和人类对齐等多个榜单上对 YuLan-Chat 模型进行评测,实验结果表明在同等参数和数据规模下,该模型能够展现出超越业内顶尖开源大模型的表现。此外,本文还开展了众多实验,以分别验证以上三阶段课程学习方法的有效性;并对 YuLan-Chat 模型生成的样例进行了展示。未来,我们会对 YuLan-Chat 模型的训练数据和代码进行开源,推动大语言模型研究社区的发展。

本文第 2 节介绍本文涉及的相关工作与背景知识;第 3 节详细介绍本文提出的基于课程学习的多阶段训练方法;第 4 节对本文的实验安排和结果进行展示与分析;第 5 节对本文的结论和未来工作进行概述。

2 相关工作

本文涉及的相关工作主要包括以下三个方面:大语言模型、指令微调和人类对齐。本章节将对这三个方面的相关背景知识进行介绍。

2.1 大语言模型

根据 Kaplan 等人^[13]提出的扩展法制(Scaling Law),随着参数规模的增大,预训练语言模型的效果也会随之提升。近年来,研究人员发现,当模型的参数规模达到一定程度后(如 10B),预训练语言模型能够涌现出一系列极强的能力^[3,5],如上下文学习等,这一类模型被称为大语言模型。2020 年,OpenAI 发布了大语言模型 GPT-3^[3],其能够通过上下文学习以及思维链提示,展示出强大的少样本和零样本学习的能力。之后,ChatGPT 的诞生以及随后的 GPT-4 的推出^[6-7],证明了大语言模型可以作为强大的人类人工智能助手,能够以对话的形式,帮助人类完成各类自然语言处理甚至更复杂的真实世界任务。这些模型的强大功能引起了整个人工智能社区对大语言模型的关注,研究人员和开发人员将其视为下游应用的重要基础,并基于其开发各类系统和真实应用。然而,OpenAI 所发布的 GPT 系列大语言模型均为闭源模型,且这些模型的许多训练细节也

并未公布,这使得其背后的技术路线依旧并不明晰。近年来,一系列开源大语言模型开始出现,并展现出优秀的模型能力。在这一进展中,LLaMA^[14]、ChatGLM^[15]、Baichuan^[16]、InternLM^[17]、Qwen^[18]等系类模型,均在某些特定语言或能力上展现出较好的表现。这些模型均基于自回归 Transformer 架构^[19],使用包含 1T 以上词元的多源语料,基于下一个词元预测任务进行预训练。在预训练过程中,这些模型普遍直接从收集到的数据源中随机采样构造训练批次(Batch),并未考虑其中数据的长尾分布情况,这使得这些模型可能在某些低频知识上出现幻象。在下游应用时,已有工作采用检索增强生成技术^[20],从外部的大规模候选文档库中检索长尾知识的相关信息,并将其加入到输入中以补充相关知识。

2.2 指令微调

在大规模预训练之后,大语言模型通常还需要进行指令微调,使其能够更好地遵循自然语言指令,且能够通过指令解决其未见过的自然语言处理任务。指令微调这一名称由谷歌的研究员正式提出^[8],又被称为有监督微调或多任务提示训练,其通常需要首先构造指令化的样例,然后通过有监督的方式对大语言模型的参数进行微调。已有的指令微调方法的差异主要在于使用的数据集不同,其可以分为以下三类:第一类方法基于现有的自然语言处理任务数据集,通过添加任务描述信息指导模型去完成该任务, Natural Instruction^[21]、FLAN^[22]和 To^[23]是其中代表性的指令数据集;第二类方法基于真实的人类或人机对话数据,以构建指令数据集,其中通常包含真实世界的人类任务需求,例如头脑风暴、开放式问答等, ShareGPT^[24]是其中有代表性的数据集。为了减轻收集和标注数据的负担,研究人员进一步提出自动化的指令数据合成方法,代表性工作如 Self-Instruct^[25]仅使用 100 多条种子指令,通过特殊的提示指导模型合成新的指令数据。然而,由于指令数据集通常包含众多任务类型,这些任务的难易程度通常并不一致,将其混合直接进行训练难以充分学习并理解其中较为复杂的指令。为对该项能力进行强化, WizardLM^[11]提出了指令进化方法,其通过对指令进行多阶段改写以逐步增加其复杂程度,进而可以得到足够复杂的指令用于微调。

2.3 人类对齐

虽然指令微调后的大语言模型已经具备了强大自然语言指令遵循能力,但其依旧可能在生成过程

中遇到与人类需求或价值观不一致的问题^[9],比如生成恶毒文本、幻觉和无意义重复等。为了规避这些潜在风险,研究人员提出了大语言模型的人类对齐 3H 标准^[26],即有用性(Helpfulness)、诚实性(Honesty)和无害性(Harmlessness)。基于 3H 标准,研究人员积极探索大语言模型人类对齐的方法,已有工作首先采集相关的提示及其对应的正负例模型输出结果,然后使用有监督学习或强化学习方式对模型进行训练。其中,基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)是常用的训练策略^[9]。该方法首先基于收集的包含正负例的人类反馈数据训练奖励模型,然后使用强化学习算法,例如近端策略优化(Proximal Policy Optimization, PPO)对大语言模型进行微调^[27]。然而,由于强化学习方法较为复杂且训练过程不稳定,一系列工作利用有监督学习方法代替强化学习。其中,直接偏好优化(Direct Preference Optimization, DPO)算法是其中有代表性的工作^[28],其通过公式推导将强化学习训练目标转化为有监督学习形式,无须使用额外的奖励模型。此外,还有一系列工作基于对齐数据,使用传统的有监督学习损失配合辅助优化目标以强化对其学习的效果,包括基于人类反馈的回复排序(RRHf)^[29]和基于 AI 反馈的强化学习(RLAIF)^[30]等。

2.4 课程学习

课程学习(curriculum learning)由 Bengio 等人^[12]于 2009 年提出,其模仿人类的学习方式,在训练过程中逐渐增加数据样本的复杂性,以逐步提升模型能力。通常来说,课程学习主要包括课程构建和多阶段训练两个步骤,前者基于特定的规则或打分函数,将每个数据样本分发到各个课程中;后者需要设计特定的顺序,依次使用这些课程中的数据对模型进行训练。课程学习方法被广泛用于计算机视觉和自然语言处理领域,其能够提升模型对整个训练数据分布的拟合能力。近年来,课程学习技术也被用于预训练领域专用的大语言模型,如医疗、代码、推荐系统等^[31-32],其通常采用两阶段课程:首先在大量通用数据,其次在少量领域专用数据上进行模型预训练。然而,对于通用大语言模型,目前相关的课程学习方法依旧较少。

3 基于课程学习的多阶段训练方法

本文提出一种基于课程学习的多阶段训练方

法,用于训练大语言模型。由于大语言模型的训练过程分为预训练、指令微调 and 人类对齐三个阶段,该方法针对每个训练阶段均设计了针对性的课程,其

中训练数据被精心安排,用于强化对该阶段中较难的知识进行学习,如图 1 所示。本节将首先介绍问题定义,再分别详细阐述三个阶段设计的方法。

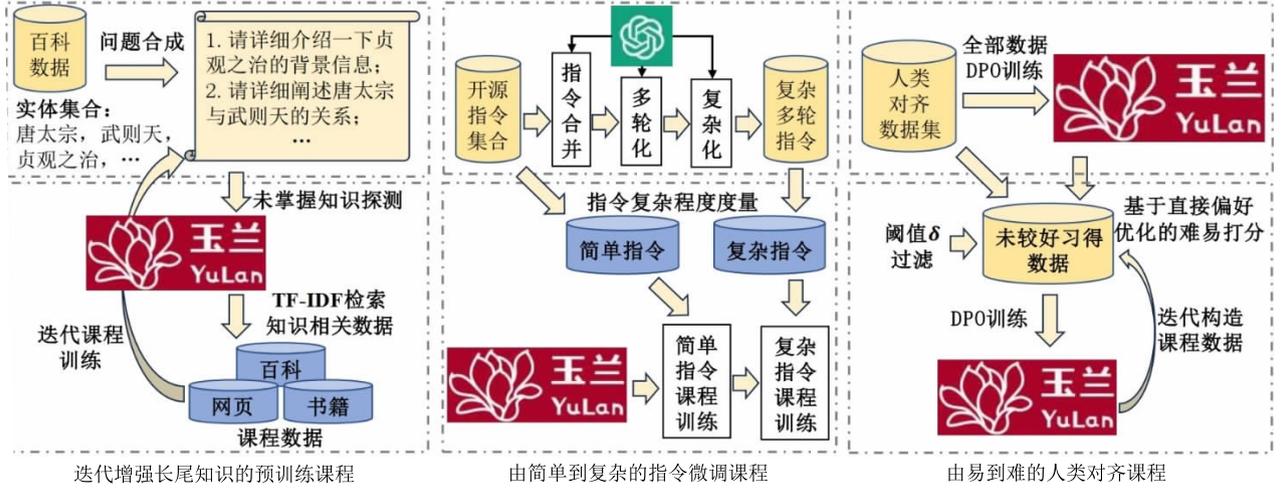


图 1 YuLan-Chat 使用的多阶段课程学习方法示意图

3.1 问题定义

给定预训练数据集 D_{pre} 、指令微调数据集 D_{ins} 和人类对齐数据集 D_{ali} , 本文关注于从头开始训练大语言模型, 并使用上述数据集完成预训练、指令微调 and 人类对齐三阶段训练过程。其中预训练数据集主要由来源于互联网的多源文本数据构成, 如网页、书籍、百科等, 可形式化为 $D_{pre} = \{d_i\}_{i=1}^N$, 其中 N 表示总的文本数量。其中, 每条数据均被预处理为一个文档 d , 其为由多个词元 w 按顺序组成的序列, 即 $d = [w_1, \dots, w_l]$, 其中 l 表示该文档长度。指令微调数据集由多个开源指令集合构成, 其中每条数据包含一个输入指令 x 和输出回复 y , 可形式化为 $D_{ins} = \{\langle x_i, y_i \rangle\}_{i=1}^M$, 其中 M 表示总的指令数量。该输入指令和输出回复均为多个词元组成的序列, 大语言模型需要学习如何基于输入指令预测该输出回复。人类对齐数据集由输入提示 p 及其对应的模型输出正负例 $\langle y^+, y^- \rangle$ 构成, 可形式化为 $D_{ali} = \{\langle p_i, y_i^+, y_i^- \rangle\}_{i=1}^O$, 其中 O 表示总的数据量。这些正负例基于人类偏好进行标注, 大语言模型需要学习如何基于输入提示对其进行区分, 尽量避免生成负例。

3.2 迭代增强长尾知识的预训练课程设计

在预训练阶段, 本文提出的方法主要关注于增强对长尾知识相关数据的记忆, 以巩固大语言模型的基础知识, 减少在下游任务中可能出现的知识记忆错误和幻象相关问题。然而, 由于预训练数据规模较大, 且较难对其中长尾知识的分布情况进行准确的探测与分析。本节提出的方法从模型角度出

发, 通过合成问答数据挖掘其在预训练阶段中并未充分掌握的长尾知识, 然后使用 TF-IDF 算法从预训练数据集中检索相关数据, 构造出专门用于补充模型未掌握知识的新训练数据集, 作为待学习的课程。通过多次迭代以上课程数据构建和模型预训练过程, 即可逐步强化模型对长尾知识的记忆。

3.2.1 基于合成问题挖掘未掌握的长尾知识

为了探测当前模型未能较好掌握的长尾知识, 本文使用合成问题-答案对探测模型当前对长尾知识的记忆情况。这里合成的问题主要从实体层面出发, 通过构造与该实体相关的问题, 探测当前模型对其理解的程度。首先, 需要基于百科数据构造知识相关的实体集合及其对应的描述信息, 然后基于模板和大语言模型构造若干实体相关问题, 最终使用预训练模型回答这些问题, 并通过度量回答的质量来判断对该实体的掌握程度。

具体来说, 由于每条百科数据均包含一个实体 v 与对其详细介绍的文档, 基于该文档可以使用字符串匹配方法, 识别出与该实体频繁共现的相关实体。之后, 可以得到实体集合 \mathcal{V} , 且可以直接使用模板对其中每个实体构造关于其相关知识介绍和关系分析的问题 q , 例如“请详细介绍一下‘贞观之治’的背景信息”和“请详细阐述‘唐太宗’与‘武则天’的关系”。此外, 从模型的输出结果中可以看到大语言模型对于特定类型问题(如列表、计数等)表现较差^[11], 故还使用大语言模型(即 ChatGPT)针对这些场景为每个实体构造了若干复杂问题。通过以上

方法即可获得大量的实体相关问题,之后将文档与该问题拼接输入给大语言模型(即 ChatGPT),并添加指令“请参考上述百科信息回答下述问题”,从而得到参考答案 a 。通过对每条百科数据采用以上方法,即可得到针对每个实体的问题-答案对集合 $\langle\langle q_i, a_i \rangle\rangle_{i=1}^{M_v}$, M_v 表示对于实体 v 的问答对数量。

在知识探测的过程中,每条问题 q_i 将被输入给当前的大语言模型,用于获得其生成的回复 a'_i 。之后,将该回复与参考答案同时输入给其他大语言模型,并使用以下提示指导大模型判断该回复是否正确:“请基于参考答案判断以下回答是否正确,如果正确则返回 1,错误则返回 0”,形式化为 $f(q_i, a_i, a'_i)$ 。基于反馈的结果,即可通过统计回答正确的问题数,得到能够度量当前大模型对该实体掌握程度的分数:

$$s_v = \frac{1}{M_v} \sum_{i=1}^{M_v} f(q_i, a_i, a'_i) \quad (1)$$

通过设置合适的阈值 ϵ ,即可筛选得到当前大模型未较好掌握实体集合 $\mathcal{V}' = \{v | s_v < \epsilon, v \in \mathcal{V}\}$ 。相较于其他复杂方法,该方法成本较低,且更适用于对模型的知识掌握能力进行较为全面的评估。图 2 中展示了历史领域内未较好掌握实体的频率长尾分布情况,可以看到 \mathcal{V}' 中大部分的实体出现频率低于 5,占总数的 60% 以上。

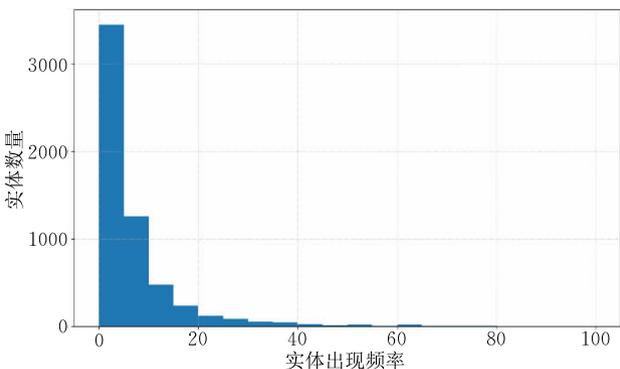


图 2 历史领域内未较好掌握实体的频率长尾分布情况

3.2.2 TF-IDF 算法检索知识相关数据

基于当前大语言模型未较好掌握的实体集合 \mathcal{V}' , 需要从预训练数据集中选出与之相关的数据用于补充对应的知识。考虑到整个预训练数据集规模较大, 本文使用计算效率较高的 TF-IDF (Term Frequency-Inverse Document Frequency) 算法判断每条预训练数据与该实体相关问题的相似性, 进而检索相关数据。具体来说, 给定预训练数据集 $\mathcal{D}_{pre} = \{d_i\}_{i=1}^N$, 首先使用 ElasticSearch 在其上建立索引, 其通过对每条数据计算得到其 TF-IDF 向量。之后, 对于实体

集合 \mathcal{V}' 中的每个实体的相关问题, 均可以通过余弦相似度计算其与每条预训练数据的相似度。对于每个实体, 均选择其前 k 个最相似的预训练数据, 经过去重即可得到数据集 \mathcal{D}'_{pre} , 其可专门用于帮助大语言模型理解其未充分掌握的实体。此外, 在构造上述数据集时, 还需要控制各个数据类型的比例保持不变(如表 1 所示), 以确保训练过程的稳定性。

3.2.3 多轮迭代课程设计

在整个预训练过程中, 首先直接使用大量的预训练数据对模型参数进行学习, 在其各项能力基本形成且能够完成各类下游任务后, 再继续在其基础上进行多轮迭代课程学习。这一过程中均使用以下自回归损失函数在预训练数据上学习文本建模:

$$L_{pre}(d) = \sum_{i=1}^l \log P(\omega_i | \omega_{<i}) \quad (2)$$

在多轮迭代过程中, 每轮均采用上述方法构造新的数据集, 作为对当前预训练模型的知识强化课程。具体来说, 在整个迭代过程开始前, 首先基于百科数据构造了实体集合 \mathcal{V} 且对每个实体合成了对应的问题-答案对集合, 并对整个预训练数据集建立基于 TF-IDF 算法的检索索引。之后, 在每轮的迭代过程中, 首先使用合成的问题-答案对数据对上一轮次预训练后的大语言模型进行知识探测, 筛选得到其未较好掌握的实体集合 \mathcal{V}' , 然后基于其对应的问题集合从预训练数据集中检索得到 \mathcal{D}'_{pre} , 并使用其作为当前课程的训练数据, 继续预训练当前大语言模型。上述迭代的预训练课程共持续 t_{pre} 轮, 直到发现课程学习后模型在实体相关问题上的表现不再有显著提升为止。

3.3 由简单到复杂的指令微调课程设计

在指令微调阶段, 本文提出的方法主要关注于强化对复杂指令的理解, 以增强模型对真实场景指令的遵循能力。然而, 已有开源的指令数据集大部分为相对较为简单的单轮指令。因此, 本节考虑首先合成大量复杂多轮指令, 之后采用课程学习方式使得大语言模型能够先学习较为简单的指令, 然后过渡到学习更加复杂的指令, 实现类似人类由简单到复杂的学习过程。

3.3.1 复杂多轮指令合成

由于直接合成复杂的多轮指令较为困难, 本文采用多阶段的方式, 基于已有指令数据逐步提升指令数据的复杂程度。首先, 本文收集了开源社区的一系列合成的指令数据集, 并对其中指令进行去重以得到初步的指令集合 \mathcal{I} 。之后主要通过以下三阶

段合成复杂多轮指令:指令合并、多轮化、复杂化。

(1) 指令合并。考虑到该指令集合中还可能存在着表达不同但语义重复的指令,指令合并操作可以将语义重复的指令整合成一条更加复杂的指令,同时实现语义去重。具体来说,首先采用 TF-IDF 算法计算所有指令之间的相似性,依次构造相似指令对,然后使用下述提示指导大语言模型(即 ChatGPT)将其合并为一条更复杂的指令并生成回复:“请将以下两条语义相似的指令合并为一条新的指令,其需要包含两条指令本身的功能,且更加复杂”。

(2) 多轮化。下一步继续将所有单轮指令数据改写为多轮指令使其更加复杂,同时能够增强模型对多轮上下文的理解。为了保证多轮指令所覆盖主题的多变性,还收集了聊天社区的主题集合,共包含 293 个主题。对于每一条单轮指令,均利用大语言模型针对随机选择的主题生成下一轮问题,使用的提示示例如下:“请生成一个与‘近代历史’主题相关的问题,并保证其与对话上文内容具有一致性”。该方法在实现多轮化的同时,还能够保证主题多样化和上下文一致性。

(3) 复杂化。考虑到以上方法生成的指令可能不足以覆盖真实世界中复杂且多元的场景,故继续对其进行复杂化操作。具体来说,使用下述指令指导大语言模型将上述问题从知识的深度和广度两个角度分别进行复杂化:“请将以下问题修改为一个更加复杂的指令,使其涉及知识的深度与广度均得到极大的提升”。通过上述流程得到的复杂多轮指令将输入给大语言模型以获得其输出,以得到最终的指令数据集。

在以上整个过程中,还采用以下策略确保指令数据质量:(1) 利用关键字识别并过滤掉合成或回答失败的指令(如“无法回答”“拒绝回答”);(2) 计算合成的指令和回复长度,删除较短的指令和回复;(3) 移除带有原始模型名称或机构(如 ChatGPT 和 OpenAI)的指令数据。

3.3.2 由简单到复杂的课程设计

基于上述合成的复杂指令数据集和可以直接收集到的社区开源指令数据集,本文按照复杂程度将其分配到两个数据集中用于两阶段课程学习:简单指令数据集 $D_{ins}^{(s)}$ 与复杂指令数据集 $D_{ins}^{(c)}$,然后按照先简单再复杂的顺序进行基于课程的指令微调。其中,各指令复杂程度 c 的度量方法如下式所示:

$$\lambda_1 * \text{对话轮次} + \lambda_2 * \text{指令长度} + \lambda_3 * L_{it}(x, y) \quad (3)$$

其中 λ_1 、 λ_2 和 λ_3 分别为基于经验设置的权重系数,

$L_{it}(x, y)$ 为该指令用于微调训练时的损失值,其计算方式如下:

$$L_{it}(x, y) = \sum_{i=1}^{|y|} \log P(y_i | x, y_{1:i-1}) \quad (4)$$

其中 y_i 为输出 y 中的第 i 个词元, $y_{1:i-1}$ 表示其中前 $i-1$ 个词元序列。上式通过学习基于输入指令 x 预测输出 y 中每一个词元的形式,训练大语言模型遵循指令的能力。本文提出的由简单到复杂的课程训练方式能够使大模型更好地学习如何理解并遵循复杂指令。

3.4 由易到难的人类对齐课程设计

在人类对齐阶段,本文提出的方法关注于强化对难以区分的正负例对的学习,以强化其对恶意意图输入的识别能力,且避免模型在真实场景下生成与人类价值观不一致的输出。虽然许多开源人类对齐数据集中会包含人类标注分数信息,但这部分信息不能反映模型区分这些正负例对的难易程度。故本文采用模型在人类对齐训练时的直接偏好优化的奖励函数来对数据的难易程度进行度量,并基于此设计了由易到难的多轮迭代课程。

3.4.1 基于直接偏好优化的难易打分机制

在人类对齐阶段,本文采用直接偏好优化算法^[28]对模型参数进行微调,该方法可以直接使用当前模型与人类对齐训练前的模型来计算奖励值。该奖励值可以直接判断当前模型对正负例对的判别能力,因此本文将其用来对样本的难易程度进行打分,具体的计算方式如下:

$$R(p, y^+, y^-) = \log \frac{\pi_\theta(y^+ | p)}{\pi_{\theta_{old}}(y^+ | p)} - \log \frac{\pi_\theta(y^- | p)}{\pi_{\theta_{old}}(y^- | p)} \quad (5)$$

其中 $\pi_\theta(y | p)$ 和 $\pi_{\theta_{old}}(y | p)$ 分别表示使用当前课程训练后的模型与训练前模型的输出分布。基于式(5)可以看出,如果模型在对齐训练时较好地学习了某正负例对,其对应 $\pi_\theta(y^+ | p)$ 和 $\pi_\theta(y^- | p)$ 的值将分别被较大地增大和减小,则该奖励值会得到较大提升;反之,该奖励值则较小。因此,可以通过设置阈值 δ 来保留奖励值小于其的正负例对,其表示当前模型并未能较好地学习的数据,通过将这些数据加入到下一阶段的课程中实现对其的强化。

3.4.2 由易到难迭代课程设计

基于上述正负例样本对难易打分函数,即可实现在每次人类对齐课程训练完毕后,继续选出模型未能较好地习得的数据加入到下一阶段课程中,依次实现由易到难的迭代课程学习,使模型逐步掌握这些较难的样本。在这一过程中,可以将阈值 δ 逐步

减小,使得模型训练使用的数据越来越困难,从而以实现课程的难度逐渐加大。通过以上方法构造出每阶段课程训练数据后,本文采用直接偏好优化算法对模型参数进行优化,其可以通过与微调相似的策略实现模型对齐,不再需要较为复杂的强化学习框架,减轻了计算开销。目标函数如下所示:

$$\nabla_{\theta} L_{\text{DPO}} = -\beta E_{(p, y^+, y^-) \sim \mathcal{D}_{\text{diff}}} \sigma(R(p, y^+, y^-)) \cdot [\nabla \log \pi_{\theta}(y^+ | p) - \nabla \log \pi_{\theta}(y^- | p)] \quad (6)$$

其中 β 是一个可以动态调整的参数,用于调节参数更新幅度。当损失函数值较小时,可以适当调小 β 的取值,以鼓励模型进行参数更新;反之,则适当调大其取值,以减少更新程度。优化过程中会增大 $\log \pi_{\theta}(y^+ | p)$,同时减小 $\log \pi_{\theta}(y^- | p)$,这可以训练模型在输入为 p 的情况,向符合人类价值观的输出 y^+ 靠近,同时尽量避免生成不符合的结果 y^- 。在迭代过程中,上述由易到难的课程可以使得模型能够在每一阶段课程中针对性地强化其未较好习得的数据,逐步实现对困难正负例的学习。

3.5 三阶段课程学习方法对比

本文针对预训练、指令微调和人类对齐阶段分别设计了对应的课程学习方法,本节主要针对这三阶段课程学习方法的相同与不同之处进行讨论与对比。首先,三阶段课程均采用经典的课程学习范式:先基于特定指标划分课程数据,之后按照顺序迭代训练模型。在具体实现时,每阶段课程均考虑到该阶段存在的重要问题,设计了针对性的学习目标,并提出了不同的课程数据划分方法,如表 1 所示。从中可以看出,三阶段分别针对长尾知识学习、复杂指令理解和困难数据学习这三大挑战,通过与实体掌握程度、指令复杂程度和样本难易程度相关的指标,划分出不同轮次的课程。因此,最终设计出的三阶段课程学习方法在每个阶段均能够循序渐进地学习数据中的重要知识。

表 1 三阶段课程学习方法对比

	学习目标	课程划分方法	轮次
预训练阶段	长尾知识学习	实体掌握程度: s_v	5
指令微调阶段	复杂指令理解	指令复杂程度: c	2
人类对齐阶段	困难数据学习	样本难易程度: R	4

4 实验结果与分析

4.1 训练数据集介绍

4.1.1 预训练数据集

为从头预训练 YuLan-Chat,共从各类开源语料

库中收集并清洗得到约 1680B 数据,其中包括约 1380B 英文,280B 中文和 20B 的多语数据,覆盖网页、书籍、新闻、百科、论文和代码类型数据,这些类型数据的分布情况如表 2 所示。

表 2 YuLan-Chat 的模型参数和预训练数据设置

模型参数设置(总参数量 12B)		
最大长度: 4096	隐表示大小: 4864	注意力头数: 38
总层数: 40	词表大小: 51190	位置编码: RoPE
预训练数据设置(总数据规模 1680B)		
网页数据: 1132B	书籍数据: 90B	新闻数据: 134B
百科数据: 42B	论文数据: 48B	代码数据: 100B

(1) 网页数据。英文网页数据主要来源于 Common Crawl^[33],其包含互联网上爬取的各个时刻的网页数据,数据量以 PB 级别计量,本文使用了 2021 年 1 月到 2023 年 2 月的网页数据。中文网页数据主要来源于 ChineseWebText^[34] 和 Wanjian^[35]。上述数据接下来经过基于启发式方法的质量过滤策略(例如正则表达式匹配、语言模型困惑度计算),移除低质量的广告网页和噪声句子等;之后采用 MinHash 算法^[36]以剔除重复数据,并针对下游需要评测的任务,去除与其测试集中可能重复且相关的文本;最后采用基于正则表达式和关键词的过滤规则,分别从文档级别和句子级别对用户的敏感信息或可识别的个人信息进行剔除。

(2) 书籍数据。英文书籍数据主要来源于 books3^[37] 和 gutenberg^[38],其中共包含约 27 万的英文书籍,覆盖多个领域。中文书籍数据主要来源于 CBook-150k^[39] 数据集,其中包含约 15 万本中文书。本文同样基于启发式规则和语言模型设计了过滤策略,对其中的低质量书籍和段落进行剔除。

(3) 新闻数据。本文使用了 2019 年 4 月到 2023 年 5 月的大量多语种新闻,供大语言模型获取实时世界知识并学习如何撰写新闻等发表稿件。

(4) 百科数据。本文使用了维基百科的多语数据^[40],并主要保留了中英文百科数据对模型进行训练。此外,还收集了互联网上的中文百科数据以进一步补充模型的中文知识。在英文维基百科和中文百科数据中,与社科直接相关的词条约占总数的 3%,STEM 直接相关的占 2%^[41]。

(5) 论文数据。本文使用了 arXiv^[42] 上 1990 年到 2023 年的英文论文的 LaTeX 文件,为模型补充科学知识,其涵盖物理、数学、计算机科学等领域。

(6) 代码数据。本文使用的代码数据主要来源于开源代码平台 Github^[43],并基于项目的 star 数

量选择其中较为高质量的代码源文件用于训练。

(7) 其他数据。除去以上数据,还收集了大量其他领域数据以进一步丰富预训练语料的多样性,包括社区问答数据、法律文书数据等。

4.1.2 指令微调数据集

在指令微调阶段,为保证覆盖指令类型的多样性,本文使用了以下开源的中英文指令数据集,并合成了复杂多轮指令用于补充(见 3.3.1 节)。

(1) FLAN-v2^[22]。FLAN-v2 由 Muffin、NIV2、T0-SF、CoT 构成。其中 Muffin 由之前 Flan 的 62 个任务和新的 26 个任务组成(包括对话和代码合成数据);T0-SF^[23]是从 T0 中抽取出来的指令数据;NIV2 指数据集 Natural-Instructions v2^[21];CoT 是带有思维链的 9 种推理任务的组合。

(2) xP3-zh。xP3^[44]包含 46 种语言和 16 个 NLP 任务的提示和数据集。由于本文主要关注中英文能力,故仅使用其与中文相关的子集。

(3) OpenOrca^[45]。该数据集中的每条数据的提示均来自于 FLAN,之后被输入给 GPT-4 或 ChatGPT 以获得更高质量的回复。

(4) ShareGPT^[24]。该数据集来源于一个开源的数据收集平台,其中包含用户和 ChatGPT 真实的多轮对话和回复,主要包括英语和其他西方语言。

4.1.3 人类对齐数据集

在人类对齐阶段,本文使用了以下中英文人类对齐数据集对模型进行微调。

(1) UltraFeedback^[46]。该数据集从多个资源收集了约 6.4 万条提示,然后使用多个大语言模型对这些提示生成 4 种不同的回复。之后利用 GPT-4 对收集到的样本从指令遵循、有用性、诚实性和真实性共 4 个方面进行质量标注。本文仅使用其中分差在 1 分以上的正负例对进行训练。

(2) Stanford SHP^[47]。该数据集主要关注有用性,其共包含 38.5 万个数据实例,覆盖 18 个主题领域。每条数据都是基于一个寻求帮助的 Reddit 帖子构建的,包含问题和帖子下两个排名较高的评论,分别被用户认为更有用和不太有帮助。

(3) HH-RLHF^[48]。该数据集关注于有用性和无害性,共包含约 16.9 万个开放式对话,涉及寻求帮助、建议或请求完成任务等情景,且每条数据均包含一个被选择的正例和另一个被拒绝的负例。

(4) CValue^[49]。该数据集是一个面向中文的对齐数据集,主要关注于安全性。其包含了 14.5 万条

样例,每条样例包含提示、正例和负例回复。

(5) PKU-SafeRLHF^[50]。该数据集侧重于有用性和无害性,包含 33 万个专家注释的实例,每一个实例都包含一个问题和对应的两个回答及其打分。

4.2 评测方案

4.2.1 评测数据集

为了全面比较 YuLan-Chat 和其他大语言模型,本文采用以下 4 个评测基准,从中英文基础技能以及中英文人类对齐这 4 个方面进行评估;并采用了 4 个问答类任务,对模型的专项能力进行评测。

(1) MMLU^[51]。这是一个综合性的英文评估基准,旨在全面评估大语言模型在多个领域中的知识理解和应用能力,包括人文、社会、自然科学等,其通过一系列选择题来测试模型的能力。本文使用来源于该基准验证集的 5 条样本示例作为输入提示,引导大模型按照分类问题解决该任务,即通过比较各选项(即 ABCD)的概率得到预测答案。

(2) GAOKAO。这是 AGIEval^[52]评估基准中关于中文能力的任务集合,涵盖了高考涉及的许多科目,并通过一系列中文选择题对模型进行评估。这里使用和 MMLU 一致的方式构造基于 5 条样本示例的输入提示,通过对比各选项的概率预测答案。

(3) AlpacaEval^[53]。这是一个针对人类对齐的英文评估基准,其利用强大的大语言模型(例如 GPT-4),对参与比较模型输出的结果进行两两比较。本文首先收集所有模型的输出结果,然后使用 GPT-4 对 YuLan-Chat 与每个基线模型的结果进行两两对比,统计所有结果中 YuLan-Chat 获胜的比例。

(4) AlignBench^[54]。这是一个评估大语言模型中文对齐能力的基准,其采用多维度、规则校准的评价方法,结合思维链对模型回复进行综合评分。本文直接使用其官方提供的思维链提示作为输入,并采用官方代码中建议的 GPT-4 基于参考答案,对模型预测答案进行识别和评价。

(5) BoolQ^[55]、ARC^[56]、OBQA^[57]和 CSQA^[58]这 4 个数据集均为专项问答数据集。BoolQ 是一个包含 15942 个是否类问题的问答数据集,这些问题是在未被提示和不受限制的环境中生成的,其中每个示例是一个(问题、段落、答案)的三元组。ARC 数据集包含 7787 道小学水平的科学多选题,本文主要在其简单数据子集上进行评测。OBQA 数据集模仿开放书本考试的形式,用于评估对某一学科的理解,其包含需要多步推理和常识性知识的问题。

CSQA 是一个多项选择问答数据集,需要不同类型的常识知识来预测正确答案。它包含 12 102 个问题,每个问题有一个正确答案和四个干扰答案。

表 3 中展示了 MMLU、GAOKAO、AlpacaEval 和 AlignBench 上评测时使用的提示模板。

表 3 评测时使用的提示模板

MMLU	The following are multiple choice questions (with answers) about xxx Question: xxx Choices: xxx Answer: xxx
GAOKAO	问题: xxx 选项: xxx 回答: xxx
AlpacaEval	Select the output (a) or (b) that best matches the given instruction. Choose your preferred output, which can be subjective. Your answer should ONLY contain: Output (a) or Output (b). Here's an example: # Example: xxx ## Instruction: xxx ## Output (a): xxx ## Output (b): xxx ## Which is best, Output (a) or Output (b)?
AlignBench	你是一个擅长评价文本质量的助手。 请你以公正的评判者的身份,评估一个 AI 助手对于用户提问的回答的质量。由于您评估的回答类型是 xxx,因此你需要从下面的几个维度对回答进行评估我们会给您提供用户的提问,高质量的参考答案,和需要你评估的 AI 助手的答案。当你开始你的评估时,你需要按照遵守以下的流程:… 用户的提问: xxx 参考答案: xxx 助手的答案: xxx

4.2.2 基线方法

本节选择了以下大语言模型进行对比:

(1) ChatGLM-6B。是由清华大学和智谱 AI 公司共同研发的中英文双语对话大语言模型^[15],其使用了 GLM 论文提出的混合目标函数,经过了 1TB 中英文混合数据的预训练,并辅以监督微调、反馈自助、人类反馈强化学习等技术。

(2) ChatGLM2-6B。是由清华大学和智谱 AI 公司共同研发的 ChatGLM-6B 第二代^[59],其使用 1.4 TB 中英文数据预训练与人类偏好对齐。

(3) InternLM-7B。即书生·浦语大模型^[17],是由上海人工智能实验室和商汤科技主导开发的多语大语言模型,其在 1.6 TB 语料库上进行多阶段渐进式训练,然后并进行微调以与人类偏好对齐。

(4) LLaMA-13B。是由 MetaAI 开发的大语言模型基座^[14],其在 1TB 英文为主的语料上进行预训练,其是目前最常用的英文大语言模型基座之一。

(5) LLaMA2-13B-Chat。是由 MetaAI 开发的 LLaMA 系列模型的升级版^[60],其在 2TB 数据上进

行预训练,并使用了大量数据进行人类对齐微调。

(6) LLaMA2-Chinese-13B。其基于 LLaMA2-13B-Chat,采用中文指令数据集对其进行 LoRA 微调,使其具备较强的中文对话能力^[61]。

(7) Baichuan-13B-Chat。是百川智能推出的中英文通用大模型^[16],其采用 1.4 TB 的中英文语料训练,在中文和英文榜单上均取得较好的效果。

(8) MOSS-moon-003。是复旦大学基于 CodeGen-16B 开发的模型^[62],其在 700B 的中英文和代码的混合数据上进行了继续预训练,之后使用了 110 万多轮对话指令对其进行微调得到。

(9) CodeLLaMA-7B。旨在通过课程学习策略,增强 LLaMA-7B 模型的通用代码生成和理解能力。其首先采用课程预训练策略,先使用 500B 代码数据,然后使用 100B Python 代码数据进行预训练;之后使用 5B 数据进行指令微调^[63]。

4.3 具体实现细节

YuLan-Chat 的训练过程主要分为以下三个阶段,预训练、指令微调、人类对齐。在预训练阶段,首先在上述提到的预训练数据集进行完整一轮的训练,共学习约 1500B 词元,之后利用迭代增强长尾知识的预训练课程,共课程学习 5 轮以巩固模型知识。其中,每次迭代课程均从上述预训练数据集中选择约 30B 长尾知识相关数据用于训练,且由于随着课程迭代时模型能力会逐渐增强,故数据选择的阈值 ϵ 将由 0.5 逐步线性增加到 1。在预训练第一阶段过程中初始学习率为 $2e-4$,并基于余弦曲线逐渐下降至 $2e-5$,在之后的课程学习过程中学习率保持为 $2e-5$ 。在指令微调过程中,首先使用式(3)对上述开源指令数据集的复杂程度进行度量并选择得分较低的简单指令数据用于第一阶段课程学习,超参 λ_1 、 λ_2 和 λ_3 分别设置为 100、1、25,之后使用较为复杂的指令进行第二阶段课程学习。两阶段课程的学习率均为 $2e-5$,第一阶段训练 1 轮,第二阶段训练 3 轮。在人类对齐阶段,首先基于上述全部的人类对齐数据使用直接偏好优化算法进行训练,然后进行由易到难的 4 轮迭代课程学习。在每阶段迭代过程中,首先基于当前模型使用式(5)对 UltraFeedback 数据进行打分,之后采用阈值 δ 以选择较难样本。在这一过程中,学习率被设置为 $1e-6$,阈值 δ 初始设置为 0,之后被线性缩小至 -3 ,以增加数据难度。此外,本文将参数 β 设置为 $\{0.05, 0.1, 1, 5\}$,并分别进行了分析实验,最终根据模型表现将其设为 0.1。

4.4 YuLan-Chat 模型整体效果对比

4.4.1 基础能力对比

表 4 中展示了各个大语言模型在 MMLU 这一英文基础能力评测基准上的效果,可以发现 LLaMA-2 相关的模型(即 LLaMA2-Chinese-13B 和 LLaMA-2-13B-Chat)普遍优于其他模型。该模型预训练语料中包含了大量的英文数据,中文和其他数据占比较少。因此,其英文能力能够得到更充分的训练,而且不会受到中文数据的过多干扰。相比之下,LLaMA2-Chinese-13B 使用中文指令对 LLaMA-2 进行了微调,这对其英文能力有影响,导致其效果出现下降(从 56.2 到 55.0)。此外,可以看到本文提出的 YuLan-Chat 模型能够取得接近于最强基线 LLaMA-2-13B-Chat 的效果,且其在 STEM 和社科这两个子类下还取得了更优的效果。这一结果得益于本文提出的基于课程学习的多阶段预训练方法,该方法能够帮助其较好地掌握数据中各种长尾知识。因此,即使训练的英文数据较少,且混入了大量的中文数据,其英文基础能力依旧较好。相比较于同样采用课程学习的 CodeLLaMA-7B 模型, YuLan-Chat 更加关注于对通用能力的提升,其设计的课程学习方法更加针对于复杂指令理解和人类对齐的场景,故能够在 MMLU 上取得更优表现。

表 4 各大语言模型在 MMLU 上的评测结果

	STEM	社科	人文	其他	均值
MOSS-moon-003	27.3	28.9	29.5	32.7	29.6
CodeLLaMA-7B	—	—	—	—	36.9
ChatGLM-6B	33.9	44.8	39.0	45.7	40.6
ChatGLM2-6B	40.1	51.6	41.2	51.2	45.5
LLaMA-13B	36.4	53.5	44.0	53.3	46.8
InternLM-7B-Chat	39.6	57.9	45.6	56.5	49.9
Baichuan-13B-Chat	40.9	60.9	48.8	59.0	52.1
LLaMA2-Chinese-13B	44.4	63.2	51.6	60.6	55.0
LLaMA-2-13B-Chat	44.6	64.2	53.9	62.2	56.2
YuLan-Chat	45.0	64.8	52.1	61.9	56.0

注:表中短横线指无法获得的评测结果,下划线指排名第二的模型评测结果,粗体指排名第一的模型评测结果。

表 5 中展示了各个大语言模型在 GAOKAO 这一中文基础能力评测基准上的效果,可以看到这些大语言模型的中文和英文基础能力并不一致。其中, Baichuan-13B-Chat 在所有基线方法中取得了最好的效果,且明显优于英文最好的基座模型 LLaMA-2-13B-Chat(46.3 对比 36.7)。该模型在预训练阶段使用了许多高质量的中文语料,这对中文基础能力有较大的增益。此外, YuLan-Chat 能够取得比 Baichuan-13B-Chat 更好的效果,这是由于其在整个训练过程中使用了课程学习训练方法,该方法也能提升大模型的中文能力。

表 5 各大语言模型在 GAOKAO 上的评测结果

	语/数/外	物/化/生	史/地	均值
MOSS-moon-003	28.5/24.8/46.4	25.0/29.5/23.3	33.2/30.7	30.2
LLaMA-13B	22.8/26.2/58.8	26.5/31.4/23.8	23.8/26.6	30.0
LLaMA2-Chinese-13B	26.4/25.6/70.6	29.5/28.0/28.1	38.7/35.2	35.3
LLaMA-2-13B-Chat	27.2/26.2/72.2	30.0 /32.4/26.2	43.0/36.2	36.7
Baichuan-13B-Chat	43.1/26.8/72.2	23.0/36.7/43.8	62.6/62.3	46.3
YuLan-Chat	43.5 / 28.2 / 81.4	29.0/40.6/55.2	68.9 /57.8	50.6

4.4.2 人类对齐能力对比

表 6 中展示了 YuLan-Chat 在 AlpacaEval 这一英文人类对齐能力评测基准上,与各个基线模型对比时的胜率。可以看到在 YuLan-Chat 对比大部分模型时,均拥有 55%以上胜率。这表明了其在英文场景下的人类对齐能力优于这些模型。本文设计的基于课程学习的指令微调和人类对齐训练方法,能够强化对复杂指令和恶意意图理解的能力。然而, LLaMA-2-Chat 在该数据集上的表现优于 YuLan-Chat,这是因为其采用了更多阶段的迭代对齐策略(共五轮),并在整个过程中添加了百万级别专门标注的人类对齐数据。这些策略会带来较大的训练和数据标注开销。

表 6 YuLan-Chat 在 AlpacaEval 数据集上与各模型对比时的胜率结果

	YuLan-Chat 胜率/%
MOSS-moon-003	57.06
ChatGLM-6B	60.81
ChatGLM2-6B	55.56
Baichuan-13B-Chat	59.57
InternLM-7B-Chat	65.13
LLaMA-2-Chat	46.71

表 7 中展示了各个大语言模型在 AlignBench 这一中文人类对齐能力评测基准上的效果。可以看到各个模型的中文对齐和英文对齐能力也不完全一致,其中 Baichuan-13B-Chat 在所有基线模型中取得了最好的效果,其中文推理能力和语言能力均优于其他模型。相比较之下, YuLan-Chat 能够显著优于 Baichuan-13B-Chat,这也是因为其采用了基于课程学习的多阶段训练策略,其能够大幅强化中文场景指令理解和人类对齐能力。

表 7 各大语言模型在 AlignBench 上的评测结果

	中文推理	中文语音	均值
InternLM-7B-Chat	2.09	4.39	3.24
MOSS-moon-003	2.24	4.67	3.46
ChatGLM-6B	2.50	5.31	3.90
ChatGLM2-6B	3.31	5.83	4.57
Baichuan-13B-Chat	3.40	6.35	4.88
YuLan-Chat	3.59	6.69	5.14

4.4.3 专项问答能力对比

表 8 中展示了各个模型在四个专项问答类任务上的评测结果。可以看出由于各个任务关注点不同,不同的基线模型在各个任务上的表现不一。然而, YuLan-Chat 依旧能够取得最好的表现,这也证明了本文提出的多阶段课程学习方法对于模型其他能力也能有较好的提升。

表 8 各大语言模型在专项问答类任务上的评测结果

	BoolQ	ARC	OBQA	CSQA
MOSS-moon-003	59.9	64.4	44.0	28.8
LLaMA-13B	76.4	<u>77.4</u>	47.2	63.6
ChatGLM2-6B	78.5	67.0	40.4	69.2
LLaMA-2-13B-Chat	75.4	76.9	<u>54.2</u>	<u>73.1</u>
Baichuan-13B-Chat	<u>81.4</u>	75.2	47.0	71.0
YuLan-Chat	83.5	80.1	58.4	76.7

4.4.4 数学推理能力对比

为了进一步展示 YuLan-Chat 的推理能力,表 9 中汇报了各个模型在数学推理任务 GSM8K 上的评测结果。GSM8K 数据集由若干数学题组成,其要求模型能够根据题目生成自然语言推理过程,并最终得到该数学题答案。可以看出 YuLan-Chat 依旧能够取得最好的表现,这证明本文提出的多阶段课程学习方法能够增强其解决该类复杂推理任务的能力。

表 9 各大语言模型在数学推理任务上的评测结果

	GSM8K
MOSS-moon-003	4.5
LLaMA-13B	17.1
ChatGLM2-6B	23.9
Baichuan-13B-Chat	26.8
LLaMA-2-13B-Chat	<u>28.9</u>
YuLan-Chat	30.1

4.4.5 代码生成能力对比

为了进一步展示 YuLan-Chat 的代码生成能力,表 10 中汇报了各个模型在 HumanEval 上的结果。该数据集包含若干人类真实场景下的代码需求,需要模型能够生成对应的 Python 代码,通过编译器执行后以得到目标结果。YuLan-Chat 展现了出优于基线模型的代码生成能力,这表明本文提出的多阶段课程学习方法能够帮助该模型较好地习得代码生成这一高级能力。

表 10 各大语言模型在代码生成任务上的评测结果

	HumanEval
LLaMA-13B	11.2
ChatGLM2-6B	10.4
Baichuan-13B-Chat	11.6
YuLan-Chat	12.8

4.5 课程学习效果分析

作为主要的技术贡献,本文设计了一个包含三种课程学习策略的多阶段学习方法,分别用于强化大语言模型的预训练、指令微调和人类对齐阶段。本节通过首先通过消融实验分析每阶段课程学习带来的提升情况,之后研究每阶段课程学习训练后模型能力的变化情况,详细分析每阶段课程学习带来的增益。

4.5.1 三阶段课程消融实验

本文提出的方法分别在预训练、指令微调和人类对齐阶段各设计了对应的课程学习方法,为了验证该三阶段课程学习方法对模型效果的增益,本节在 MMLU 和 GAOKAO 这两个榜单上对模型中英文基础能力进行评测,同时还引入了 GSM8K 和 HumanEval 这两个数据集以对模型能力进行更综合性的评测。基于这四个数据集,本节对课程学习前的基座模型、预训练课程学习后、指令微调课程学习后、人类对齐课程学习后的模型分别进行评测,实验结果如表 11 所示。从表中可以看出,随着三阶段课程学习的陆续进行, YuLan-Chat 在这四个数据集上的效果也在逐渐提升,这证明每个阶段课程学习均有提升模型综合能力的效果。此外,可以看到指令微调课程学习后对模型的提升相对较大,这是因为指令微调阶段着重关注于增强模型解决下游任务的能力,故其能够提升模型任务表现。

表 11 三阶段课程消融实验结果

	MMLU	GAOKAO	GSM8K	HumanEval
课程学习前模型	49.7	42.0	12.5	10.4
-预训练课程学习后	51.3	45.3	18.6	11.6
-指令微调课程学习后	55.8	49.7	29.6	12.2
-人类对齐课程学习后	56.0	50.6	30.1	12.8

4.5.2 预训练阶段课程效果分析

在预训练阶段,本文提出的方法通过增加课程迭代轮次实现对预训练数据集中长尾知识的强化,实现模型基础能力的增强。为了验证其效果,本节在 MMLU 和 GAOKAO 两个评测基准上对每轮课程迭代训练后模型的能力进行评测,其分别评估模型的英文和中文基础语言理解能力。如图 3 所示,随着迭代轮次的增加,课程预训练方法能够逐步提升模型在这两个评测基准上的效果,这验证了该迭代增强长尾知识的预训练课程的有效性。此外,可以看到在中文基准 GAOKAO 上,模型能力的增幅更大,这表明该课程学习方法对中文能力的提升更大。这是由于训练时使用的中文数据相对英文更少

(互联网上中文语料相对较少),许多中文知识较难被直接习得,而本文提出的课程学习方法可以强化对其的学习。然而,随着课程轮次的增加,检索召回的课程数据会包含更多上一轮课程中重复的数据,这会导致模型过拟合这些数据内容,反而不利于学习重要知识并提升基础能力。根据实验结果,本文仅迭代 5 轮预训练课程。

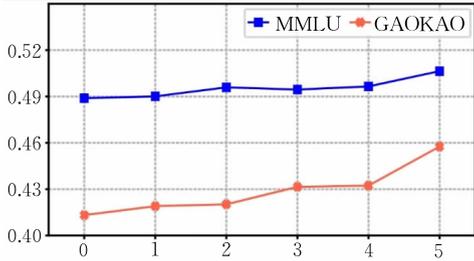


图 3 随预训练课程轮次增加的模型效果变化图

4.5.3 指令微调阶段课程效果分析

在指令微调阶段,本文采用两阶段课程学习方法,实现从简单到复杂指令的逐步学习强化。为了验证该方法对模型指令遵循能力和人类对齐能力效果提升情况,本节在 AlpacaEval 和 AlignBench 上对两阶段课程学习前后的模型效果进行评测,其分别评估模型中英文对话和人类对齐能力。其中,AlpacaEval 使用了预训练得到的基座模型作为基线,这里汇报了与其两两对比的效果,如图 4 所示。在每阶段课程训练后,模型的效果均会有大幅提升,这验证了该由简单到复杂的课程指令微调策略的有效性。此外,可以看到第一阶段课程基于上一阶段带来的相对提升更大,这是由于基座模型未经过指令微调训练,其指令遵循能力还较弱。

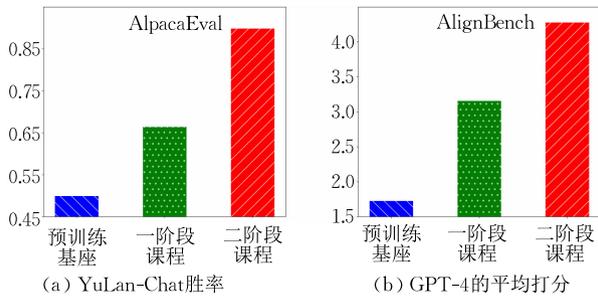


图 4 指令微调课程学习后模型效果变化情况

4.5.4 人类对齐阶段课程效果分析

在人类对齐阶段,本文采用由易到难的人类对齐课程,通过迭代的方式逐渐提升模型人类对齐的效果。为验证该方法的效果,本节在 AlpacaEval 和 AlignBench 上对每轮迭代课程学习后的模型效果进行评测,其能够分别反映模型中英文人类对齐能

力。这里 AlpacaEval 使用了指令微调得到的基座模型作为基线进行两两对比,AlignBench 结果被除以 10 以对齐刻度便于直接对比曲线趋势。如图 5 所示,随着迭代轮次的增加,该课程学习方法能够逐步提升模型在这两个评测基准上的效果,这验证了该迭代增强人类对齐的课程的有效性。此外,可以看到随着迭代轮次的增加,模型的效果提升幅度快速收敛,这表明使用较少的课程学习迭代轮次足以达到较好效果,无须大量计算开销。随着迭代过程的进行,模型未较好习得的正负例对样本数量也会显著减少,剩余数据中噪声数据比例会增大,使用其对模型进行训练可能反而带来负面影响。根据实验结果,本文仅选择迭代 4 轮人类对齐课程。

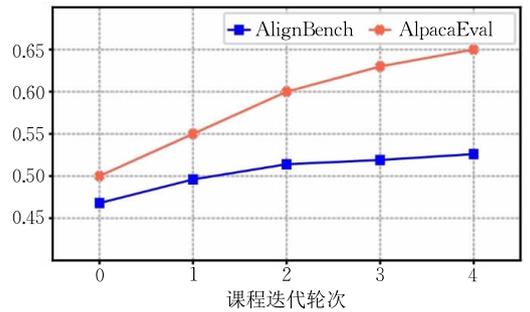


图 5 随人类对齐课程轮次增加的效果变化图

4.5.5 各类变种课程策略效果分析

本文的主要创新点在于由易到难的课程学习方法,为验证该方法的有效性,在人类对齐阶段,本节设计了一系列的课程学习变种方法,包括(1)移除课程学习。直接使用 DPO 算法训练至收敛;(2)逆课程学习。将使用过的课程学习数据逆序(即由难到易)输入给模型;(3)延长迭代周期。将每次迭代时阈值缩小的幅度减半,使得模型可以迭代两倍轮次。实验结果如表 12 所示,可以看出这三种课程学习变种均会带来效果的下降,这证明了本文提出的由易到难的课程学习方法的有效性。当使用逆课程学习方法时,其效果甚至不如移除课程学习,这也证明了由难到易的课程学习策略可能反而对人类对齐能力有所损害。此外,延长迭代周期并未带来效果增益,其背后原因可能是过多轮次的迭代导致了模型出现过拟合,反而不利于对齐。

表 12 人类对齐阶段各课程学习变种在 AlignBench 上的结果

	中文推理	中文语音	均值
YuLan-Chat	3.59	6.69	5.14
+ 移除课程学习	3.32	6.05	4.68
+ 逆课程学习	3.18	5.74	4.46
+ 延长迭代周期	3.46	6.63	5.04

4.5.6 检索算法对比分析

本节对构造预训练课程时所使用的 TF-IDF 检索算法和稠密向量检索算法进行对比。具体来说,这里使用经典的句子表示模型 SimCSE^[64]来生成文档稠密向量,之后利用 FAISS 库^[65]建立索引并执行查询操作。为了对比这两种方法,本节采样了 200 条数据,通过人工标注的形式,判断哪一方检索返回的文档更适合作为输入查询的补充信息,并用于 YuLan-Chat 的预训练。实验结果如表 13 所示,可以看出 TF-IDF 检索返回的文档质量相对较高,更适合作为补充语料。这是因为 TF-IDF 能够基于关键词进行精准匹配,其有助于直接召回高度相关的文档。此外,TF-IDF 检索算法仅需 CPU 集群即可运行,计算开销相对较小,故本文最终采用 TF-IDF 检索

表 13 不同检索算法返回文档质量对比 (单位:%)

TF-IDF 更优	一样	稠密向量更优
29	48	23

算法。

4.6 样例展示

这里展示了一些 YuLan-Chat 和 ChatGPT 在真实场景下的回复样例。表 14 中展示了中英文日常对话和翻译的样例,可以看到 YuLan-Chat 具备成为人们日常使用的人工智能助手的潜质,其生成的结果与 ChatGPT 一样,都能较好地为人们提供建议和工作支持。然而,可以看到 ChatGPT 生成的结果中的一些信息会更加详细且准确,而 YuLan-Chat 的部分回答中还存在幻象问题,如“天坛建于 18 世纪”(实际为 15 世纪)。然而,如果直接将该知识点对应的具体问题输入给 YuLan-Chat,其可以直接准确答出,如表 15 所示;但如果不针对性的进行提问,YuLan-Chat 容易在介绍背景的时候,将该知识点错误地说出。这就表明 YuLan-Chat 对该知识点的掌握不够牢固,在使用这些知识点时可能出现错误。

表 14 YuLan-Chat 样例展示

User: I am going to Beijing, can you suggest me three attractions?

YuLan-Chat: Certainly! Beijing is a city with a rich history and culture, so there are many attractions to choose from. Here are three suggestions:

- (1) The Forbidden City: This is a must-visit attraction in Beijing, and it is the largest palace complex in the world. It was the imperial palace of the Ming and Qing dynasties, and it is now a museum that houses a vast collection of imperial treasures.
- (2) The Great Wall of China: This is one of the most iconic landmarks in China, and it is a must-visit attraction for anyone traveling to Beijing. The wall stretches over 5,000 miles, and it is a testament to the engineering skills of the ancient Chinese.
- (3) The Temple of Heaven: This is a Taoist temple that was built in the 18th century, and it is a UNESCO World Heritage site. The temple is surrounded by beautiful gardens, and it is a popular place for locals to exercise and relax. It is also a great place to see traditional Chinese architecture and culture.

ChatGPT: As a virtual assistant, I can suggest three popular attractions in Beijing for you to visit:

The Great Wall of China: This iconic landmark is a must-visit in Beijing. Spanning over 13 000 miles, it offers breathtaking views and a glimpse into China's rich history.

The Forbidden City: Located in the heart of Beijing, the Forbidden City is an ancient imperial palace that served as the home of Chinese emperors for over 500 years. It is a UNESCO World Heritage site and offers a fascinating insight into the country's imperial past.

The Temple of Heaven: This ancient temple complex is another must-see attraction in Beijing. It was used by the emperors of the Ming and Qing dynasties to perform rituals and ceremonies for good harvests. The stunning architecture and beautiful gardens make it a peaceful and picturesque destination.

Remember to check the opening hours, ticket prices, and any travel restrictions before your visit to ensure a smooth and enjoyable experience.

User: 帮我翻译一下这段回复

YuLan-Chat: 当然,以下是对您请求的回复的翻译:

北京是一个拥有丰富历史和文化的城市,因此有许多值得参观的景点。以下是三个建议:

- (1) 故宫:这是必去的景点,它是明清两代的皇家宫殿,现在是一座博物馆,展示了大量的皇家珍宝。
- (2) 长城:这是中国最著名的地标之一,也是必去的景点。长城长达 5000 多英里,是古代中国工程技能的见证。
- (3) 天坛:这是一座道教寺庙,建于 18 世纪,现在是一座展示古代中国文物和文化的博物馆。寺庙周围有美丽的 3 的花园,是当地人锻炼和放松的好去处。它也是欣赏传统中国建筑和文化的的好地方。

ChatGPT: 作为一个虚拟助手,我可以为您推荐在北京的三个热门景点:长城;这个标志性的地标是北京必游之地。长城的全长超过 13 000 英里,提供了令人惊叹的景色,也让您一窥中国丰富的历史。

故宫:位于北京市中心的故宫是一座古老的皇家宫殿,曾经作为中国皇帝的居所长达 500 多年。它是联合国教科文组织世界遗产,为您提供了对中国帝国时代的迷人洞察。

天坛:这座古老的寺庙建筑群是北京的另一必看景点。明清两代的皇帝在这里进行祭祀和仪式以祈求丰收。壮丽的建筑和美丽的花园使其成为一处宁静而风景如画的目的。

在您前往之前,请记得查看开放时间、门票价格和任何旅行限制,以确保您的访问顺利而愉快。

表 15 直接提问时 YuLan-Chat 答对的样例

User: 天坛的建造时间是什么时候?

YuLan-Chat: 天坛始建于明永乐十八年(1420 年),后经明嘉靖、隆庆、万历等朝续建修缮,形成了集神殿、祈年殿和皇穹宇为中心的主要建筑群。因此,天坛的建造始于明代永乐十八年,即公元 1420 年。这表明天坛作为明清两代皇帝祭天祈谷的场所,具有悠久的历史,并展现了明清时期中国古代建筑和宗教文化的独特魅力。

表 16 中展示了英文数学推理和中文知识推理场景下的样例。可以看出 YuLan-Chat 的数学推理步骤详略得当,能够精准的把握重要的关键步骤;且中文推理过程清晰,能够表达出支持结论的观点。未来,我们会更加关注于增强对已习得知识的利用,使 YuLan-Chat 模型能够具备在更多复杂场景下稳定给出高质量少幻象回答的能力。

表 16 推理场景下 YuLan-Chat 样例展示

<p>User: Claire makes a 3 egg omelet every morning for breakfast. How many dozens of eggs will she eat in 4 weeks? Let's think step by step.</p> <p>YuLan-Chat: Claire makes 3 eggs omelettes every morning * 7 days/week = 21 eggs per week. So in 4 weeks Claire consumes 21 eggs/week * 4 weeks = 84 eggs. Since a dozen refers to a dozen equals 12 items, then 84 eggs/12 eggs/dozen = 7 dozen eggs. The answer is 7.</p>	<p>ChatGPT: Let's break down the problem step by step:</p> <ol style="list-style-type: none"> (1) Calculate the total number of eggs used per day: Claire uses 3 eggs per omelet each day. (2) Calculate the total number of eggs used in a week: There are 7 days in a week, so in one week, Claire will use: 3 eggs/day * 7 days/week = 21 eggs/week (3) Calculate the total number of eggs used in 4 weeks: Over 4 weeks, Claire will use: 21 eggs/week * 4 weeks = 84 eggs (4) Convert the total number of eggs into dozens: There are 12 eggs in a dozen, so: 84 eggs/12 eggs/dozen = 7 dozen eggs So, Claire will use 7 dozen eggs in 4 weeks.
--	---

User: 数据表明,我国目前团购网站数量不到 2000 家,去年高峰时期曾有 10000 家。但不到一年的时间就淘汰了 40% 的网站。分析师认为,经过两年的竞争,资源逐渐掌握在少数几家手里,市场集中度还将进一步提高,只有极少数网站能够存活下来。活下来的网站会逐渐捡回话语权,毛利会逐渐上抬。这段文字主要介绍了团购行业的: A. 竞争方式; B. 经营现状; C. 发展趋势; D. 市场前景。

YuLan-Chat: 这段文字主要介绍了当下团购行业的发展情况和未来趋势,因此最适合的答案应该是 C: 发展趋势。虽然文中提到了竞争方式、经营现状和市场集中度等方面的信息,但这些内容都是为了说明团购行业的发展趋势,即市场会逐渐向少数几家大规模、有实力的网站集中,而活下来的网站将获得更多的话语权,利润率也会上升。因此, C 选项“发展趋势”最能准确地概括这段文字所表达的核心内容。

ChatGPT: 这段文字主要介绍了团购行业的发展趋势。它强调了市场集中度的提高、网站数量的减少,以及未来只有极少数网站能够存活并重新掌握话语权。因此,正确选项是 C: 发展趋势。

5 结 论

本文提出了用于大语言模型训练的多阶段课程学习方法,分别用于大语言模型的预训练、指令微调 and 人类对齐阶段,以强化对这三阶段中较难掌握的知识的 learning,即预训练数据中的长尾知识、复杂困难指令、较难区分的人类对齐正负例。具体来说,该方法包括迭代增强长尾知识的预训练课程、由简单到复杂的指令微调课程、和由易到难的人类对齐课程,这些课程依次对 YuLan-Chat 进行训练,逐步提升其能力。最后,本文在四个与大语言模型基础能力和人类对齐能力相关的中英文评测基准上对 YuLan-Chat 进行评测,结果表明该模型具备与已有大语言模型相当甚至更优的表现,分析实验进一步表明了该三阶段课程对模型能力的提升作用。

在未来工作中,我们计划探究在更多训练数据场景下以上方法的有效性,致力于训练出效果更优的 YuLan 系列大语言模型。此外,我们也关注于将以上方法用于特定领域大语言模型的训练,为数学、科学等场景提供更强的大语言模型。

参 考 文 献

[1] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020, 21: 1-67

[2] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019

[3] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901

[4] Zhao W X, Zhou K, Li J, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023

[5] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022

[6] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023

[7] Achiam J, Adler S, Agarwal S, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023

[8] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021

[9] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022, 35: 27730-27744

[10] Zhang Y, Li Y, Cui L, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023

[11] Xu C, Sun Q, Zheng K, et al. WizardLM: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023

[12] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning//*Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Canada, 2009: 41-48

- [13] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361, 2020
- [14] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023
- [15] Zeng A, Liu X, Du Z, et al. GLM-130B: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414, 2022
- [16] Yang A, Xiao B, Wang B, et al. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305, 2023
- [17] Team I L M. InternLM: A multilingual language model with progressively enhanced capabilities. <https://github.com/InternLM/InternLM>, 2023
- [18] Bai J, Bai S, Chu Y, et al. Qwen technical report. arXiv preprint arXiv:2309.16609, 2023
- [19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017, 30: 5998-6008
- [20] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2023
- [21] Wang Y, Mishra S, Alipoormolabashi P, et al. Super-Natural Instructions: Generalization via declarative instructions on 1600+ NLP tasks//*Proceedings of the Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates, 2022: 5085-5109
- [22] Longpre S, Hou L, Vu T, et al. The flan collection: Designing data and methods for effective instruction tuning//*Proceedings of the International Conference on Machine Learning*. Honolulu, USA, 2023: 22631-22648
- [23] Honovich O, Scialom T, Levy O, et al. Unnatural instructions: Tuning language models with (almost) no human labor //*Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada, 2023: 14409-14428
- [24] ShareGPT Dataset. <https://sharegpt.com/>
- [25] Wang Y, Kordi Y, Mishra S, et al. Self-instruct: Aligning language models with self-generated instructions//*Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada, 2023: 13484-13508
- [26] Wang Y, Zhong W, Li L, et al. Aligning large language models with human: A survey. arXiv preprint arXiv:2307.12966, 2023
- [27] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017
- [28] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 2024, 36: 53728-53741
- [29] Yuan Z, Yuan H, Tan C, et al. RRHF: Rank responses to align language models with human feedback without tears. arXiv preprint arXiv:2304.05302, 2023
- [30] Lee H, Phatale S, Mansoor H, et al. RLAIFF: Scaling reinforcement learning from human feedback with AI feedback. arXiv preprint arXiv:2309.00267, 2023
- [31] Roziere B, Gehring J, Gloeckle F, et al. Code LLaMA: Open foundation models for code. arXiv preprint arXiv:2308.12950, 2023
- [32] Liao J, Li S, Yang Z, et al. LLaRA: Aligning large language models with sequential recommenders. arXiv preprint arXiv:2312.02445, 2023
- [33] Common Crawl Dataset. <https://commoncrawl.org/>
- [34] Chen J, Jian P, Xi T, et al. ChineseWebText: Large-scale high-quality Chinese web text extracted with effective evaluation model. arXiv preprint arXiv:2311.01149, 2023
- [35] Qiu J, Lv H, Jin Z, et al. WanJuan-CC: A safe and high-quality open-sourced English Webtext dataset. arXiv preprint arXiv:2402.19282, 2024
- [36] Li P, Shrivastava A, Moore J, et al. Hashing algorithms for large-scale learning. *Advances in Neural Information Processing Systems*, 2011, 24: 2672-2680
- [37] Gao L, Biderman S, Black S, et al. The pile: An 800 GB dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020
- [38] Gutenberg Dataset. <https://www.gutenberg.org/>
- [39] CBook-150k Dataset. <https://github.com/FudanNLPLAB/CBook-150K>
- [40] Wikipedia Dataset. <https://dumps.wikimedia.org/backup-index.html>
- [41] Lewoniewski W, Węcel K, Abramowicz W. Multilingual ranking of Wikipedia articles with quality and popularity assessment in different topics. *Computers*, 2019, 8: 60
- [42] Arxiv Dataset. <https://arxiv.org/>
- [43] Github. <https://github.com/>
- [44] Muennighoff N, Wang T, Sutawika L, et al. Crosslingual generalization through multitask finetuning//*Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Toronto, Canada, 2023: 15991-16111
- [45] Mukherjee S, Mitra A, Jawahar G, et al. Orca: Progressive learning from complex explanation traces of GPT-4. arXiv preprint arXiv:2306.02707, 2023
- [46] Cui G, Yuan L, Ding N, et al. UltraFeedback: Boosting language models with high-quality feedback. arXiv preprint arXiv:2310.01377, 2023
- [47] Stanford SHP Dataset. <https://huggingface.co/datasets/stanfordnlp/SHP>
- [48] Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022
- [49] Xu G, Liu J, Yan M, et al. CValues: Measuring the values of Chinese large language models from safety to responsibility. arXiv preprint arXiv:2307.09705, 2023
- [50] Ji J, Liu M, Dai J, et al. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems*, 2024, 36: 24678-24704

- [51] Hendrycks D, Burns C, Basart S, et al. Measuring massive multitask language understanding. arXiv preprint arXiv: 2009.03300, 2020
- [52] Zhong W, Cui R, Guo Y, et al. AGIEval: A human-centric benchmark for evaluating foundation models//Proceedings of the Findings of NAACL. Mexico City, Mexico, 2024; 2299-2314
- [53] AlpacaEval Benchmark. https://tatsu-lab.github.io/alpaca_eval/
- [54] Liu X, Lei X, Wang S, et al. AlignBench: Benchmarking Chinese alignment of large language models//Proceedings of the Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024; 11621-11640
- [55] Clark C, Lee K, Chang M W, et al. BoolQ: Exploring the surprising difficulty of natural Yes/No questions//Proceedings of the NAACL-HLT. Minneapolis, USA, 2019; 2924-2936
- [56] Clark P, Cowhey I, Etzioni O, et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018
- [57] Mihaylov T, Clark P, Khot T, et al. Can a suit of armor conduct electricity? A new dataset for open book question answering//Proceedings of the EMNLP. Brussels, Belgium, 2018; 2381-2391
- [58] Talmor A, Herzig J, Lourie N, et al. CommonsenseQA: A question answering challenge targeting commonsense knowledge //Proceedings of the NAACL-HLT. Minneapolis, USA, 2019; 4149-4158
- [59] ChatGLM2-6B Model. <https://github.com/THUDM/ChatGLM2-6B>
- [60] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv: 2307.09288, 2023
- [61] LLaMA2-Chinese-13B Model. <https://github.com/Llama-Family/Llama-Chinese>
- [62] MOSS Model. <https://github.com/OpenMOSS/MOSS>
- [63] Yuan Z, Yuan H, Tan C, et al. How well do large language models perform in arithmetic tasks?. arXiv preprint arXiv: 2304.02015, 2023
- [64] Gao T, Yao X, Chen Danqi. SimCSE: Simple contrastive learning of sentence embeddings//Proceedings of the Empirical Methods in Natural Language Processing. Virtual, 2021; 6894-6910
- [65] Jégou H, et al. FAISS: Similarity search and clustering of dense vectors library. Astrophysics Source Code Library; ascl-2210, 2022



ZHOU Kun, Ph.D. His research interests include natural language processing, information retrieval.

ZHU Yu-Tao, Ph.D., postdoctoral researcher. His research interests include natural language processing, information retrieval.

CHEN Zhi-Peng, Ph.D. candidate. His research interests include natural language processing.

MAO Ke-Long, Ph.D. candidate. His research interests include natural language processing, information retrieval.

CHEN Wen-Tong, Ph.D. candidate. His research interests include natural language processing.

CHEN Yu-Shuo, M.S. candidate. His research interests include natural language processing.

SUN Yi-Ding, Ph.D. candidate. His research interests include natural language processing, information retrieval.

CAO Qian, Ph.D. candidate. His research interests include natural language processing, information retrieval.

WANG Lei, Ph.D. candidate. His research interests include natural language processing, information retrieval.

ZHANG Lei, Ph.D. candidate. His research interests include natural language processing, information retrieval.

PANG Xin-Cheng, M.S. candidate. His research interests include natural language processing, computer vision.

XIE Shu-Fang, M.S. candidate. His research interests

include natural language processing.

ZHAO Xin, Ph.D., professor. His research interests include natural language processing, information retrieval.

DOU Zhi-Cheng, Ph.D., professor. His research interests include information retrieval.

LIN Yan-Kai, Ph.D., associate professor. His research interests include natural language processing.

MAO Jia-Xin, Ph.D., associate professor. His research interests include information retrieval.

SONG Rui-Hua, Ph.D., associate professor. His research interests include natural language processing, information retrieval, multimodal.

CHEN Xu, Ph.D., associate professor. His research interests include natural language processing, information retrieval.

XU Jun, Ph.D., professor. His research interests include natural language processing, information retrieval.

HU Di, Ph.D., associate professor. His research interests include multimodal, computer vision.

YAN Rui, Ph.D., associate professor. His research interests include natural language processing, information retrieval.

HUANG Wen-Bing, Ph.D., associate professor. His research interests include machine learning.

WEI Zhe-Wei, Ph.D., professor. His research interests include machine learning.

WEN Ji-Rong, Ph.D., professor. His research interests include information retrieval, machine learning.

Background

Large language models (LLMs) have received widespread attention in natural language processing area. In existing work, the training process of LLMs is usually divided into three stages: pre-training, instruction tuning, and human alignment. The training of LLMs is a complex system engineering that involves numerous detailed designs and practical considerations. However, there are currently limited references available for this process, mainly due to the lack of computational resources in academia and the industry's tendency to keep the training process as a proprietary technology.

To address this research gap, this paper focuses on training a general LLM, YuLan-Chat, from scratch. However, due to the distribution and difficulty divergence of the training data, it is hard for LLMs to acquire the knowledge from them, especially for the important data with harder-to-grasp knowledge, limiting the LLM's performance. Inspired by the process of human education, this paper proposes a curriculum learning approach^[12] for the three training stages of LLMs, where multi-level curricula are designed to help YuLan-Chat progressively learn knowledge from easy to hard, similar to human. For the pretraining, instruction tuning, and human alignment stages, the designed curricula focus on the following important but difficult-to-master

knowledge-related data: (1) pretraining data that involves long-tail knowledge; (2) complex instructions; (3) indistinguishable positive and negative human-alignment examples. Specifically, this paper proposes an iterative curriculum for enhancing long-tail knowledge in pretraining, a simple-to-difficult curriculum for instruction tuning, and an easy-to-difficult curriculum for human alignment.

We employ these curricula to sequentially train our YuLan-Chat from scratch. Four Chinese and English benchmark datasets are used to evaluate YuLan-Chat. The results demonstrate that YuLan-Chat can mostly outperform competitive LLMs. Furthermore, our experiments show that the three-stage curriculum can yield performance improvements of 9.7% (GAOKAO) in pretraining, 22.2% and 18.9% (Align-Bench) in instruction tuning and human alignment stages, respectively.

This work was partially supported by National Natural Science Foundation of China under Grant Nos. 62222215 and U2001212, and the Beijing Natural Science Foundation under Grant No. 4222027. And this work is also partially supported by the Outstanding Innovative Talents Cultivation Funded Programs 2021 of Renmin University of China.