

文章编号: 1003-0077(2025)10-0001-27

从人类偏好到自主对齐：大语言模型对齐方法综述

窦士涵¹, 张明¹, 黄萱菁^{1,2}, 柳世纯¹, 沈钰炯¹, 张家政¹,
黄宸颢¹, 陈佳逸¹, 郑惠元¹, 周玮康¹, 桂韬^{2,3}, 张奇^{1,2}

(1. 复旦大学 计算与智能创新学院, 上海 200433;

2. 复旦大学 可信具身智能研究院, 上海 200433;

3. 复旦大学 现代语言学研究院, 上海 200433)

摘要: 大语言模型对齐技术旨在确保模型在能力、行为和价值观方面与人类的长远利益保持一致。该文系统且全面地回顾了大语言模型对齐技术的发展历程, 从全新的视角对这些技术进行了整理和分类, 并将其发展脉络总结为三大类别: 从人类偏好中模仿学习, 从反馈信号中归纳学习, 以及通过思考和沟通实现自主对齐。针对每一项技术的特点、优势和挑战, 该文进行了详细阐述和总结。同时, 该文还概述了用于评估大模型对齐技术表现的评测方法, 讨论了当前大语言模型对齐技术所面临的挑战, 并探讨了未来实现更完善对齐技术的可能发展方向, 以推动对齐技术的进一步发展。

关键词: 人工智能; 大语言模型; 价值观和能力对齐; 强化对齐

中图分类号: TP391

文献标识码: A

From Human Preference to Autonomous Alignment: A Survey of Large Language Model Alignment

DOU Shihan¹, ZHANG Ming¹, HUANG Xuanjing^{1,2}, LIU Shichun¹, SHEN Yujiong¹,
ZHANG Jiazheng¹, HUANG Chenhao¹, CHEN Jiayi¹, ZHENG Huiyuan¹,
ZHOU Weikang¹, GUI Tao^{2,3}, ZHANG Qi^{1,2}

(1. Computation and Artificial Intelligence Innovative College, Fudan University, Shanghai 200433, China;

2. Institute of Trustworthy Embodied Artificial Intelligence, Fudan University, Shanghai 200433, China;

3. Institute of Modern Languages and Linguistics, Fudan University, Shanghai 200433, China)

Abstract: Large Language Model (LLM) alignment aims to ensure that models align with human long-term interests in terms of capabilities, behaviors, and values. This survey offers a systematic and comprehensive review of the development of LLM alignment techniques, providing a novel framework to organize and classify these approaches. We categorize the development trajectory of LLM alignment into three primary domains: imitation learning from human preference, inductive learning from feedback signal, and autonomous alignment through reflection and communication. Each category is examined regarding its characteristics, advantages, and challenges. Moreover, we also summarize the methods used to evaluate the performance of LLM alignment techniques and discuss potential future directions to further advance the field.

Keywords: artificial intelligence; large language model; value and capability alignment; reinforcement learning

收稿日期: 2025-04-14 定稿日期: 2025-05-19

基金项目: 广东省重点研发计划(2024B0101050003); 国家自然科学基金(62441602)

0 引言

自然语言处理技术的发展经历了从传统机器学习到深度学习,从小规模模型到拥有千亿级参数的大语言模型(Large Language Model, LLM)的范式转变。早期的模型通过结构设计和算法优化,力求在特定任务上对齐人类行为和表现^[1-2]。尽管这些模型在某些特定任务上表现出色,但其通用能力仍然受到很大限制^[3-4]。近年来,随着预训练技术的突破,以 ChatGPT 和 DeepSeek-R1 为代表的基于 Decoder-only Transformers 结构的大语言模型迅猛发展,标志着自然语言处理领域的一个重要里程碑,也向通用人工智能(Artificial General Intelligence, AGI)迈出了关键一步^[5-8]。这类大模型通过预测下一个词的方式在海量文本语料上学习到了广泛的知识,展现出了前所未有的能力^[9]。它们的出现显著提升了人们的工作效率和决策质量,在健康、教育、文旅等众多领域展示了广泛的应用潜力。

然而,随着大语言模型的日益普及,人们也逐渐发现了这项变革性技术的双面性:模型不仅学习到了海量语料中的广泛知识,也不可避免地吸收了其中存在的虚假、偏见和有害信息^[10-12]。这可能导致模型出现幻觉,生成恶意内容,挑战人类价值观侵犯隐私等,从而引发一系列伦理问题^[13-16]。如果不能有效控制大模型的价值观和安全伦理,相关技术的应用可能导致严重后果,如在法律辅助判别等领域的不公正处理^[17],甚至给社会秩序和国家安全带来危害。同时,若大模型深入人类日常生活而缺乏适当的伦理约束,不安全的应用可能威胁到个人隐私、信息安全乃至生命安全。

另一方面,尽管大模型在许多任务上展现出了强大的能力,但其表现仍未完全达到人类期望,例如在指令遵循和复杂逻辑推理等方面仍有提升空间^[18-19]。同时,在一些更加专业的场景,例如金融分析、医学诊断和法律咨询等垂直领域,模型能力仍需进一步提高以提供更加专业的意见,与人类专家对齐^[20-22]。总的来说,如何让大模型与人类的价值观和伦理标准对齐,同时确保其发展路径与人类长远利益保持一致,是一个极其重要的研究方向。

研究人员深入探索了将大模型与人类价值观和能力全面对齐的方法,力求在低资源消耗的情况下实现最佳效果。这些方法极大地提升了模型的安全性和能力。已有多篇综述总结了当前大模型对齐的技

术细节和目标^[15,23-27]。例如,Wang 等人^[25]从算法本身出发,对大模型价值观对齐技术进行了分类,详细阐述了基于人类反馈的强化学习(Reinforcement Learning from Human Feedback, RLHF)^[28]、直接偏好优化(Direct Preference Optimization, DPO)^[29]等方法的差异和优缺点。Ji 等人^[23]提出了大模型应该遵循鲁棒性、可解释性、可控性和伦理性这四个原则(RICE)。Cao 等人^[24]从对齐信号的来源角度出发,对大模型对齐方法进行了系统阐述和归纳。Shen 等人^[15]将对齐分为外对齐和内对齐两类,从可解释性的角度详细分析了对齐方法。Fang 等人^[27]详细分析了大模型在安全性、偏见和隐私保护方面存在的问题。

与现有综述相比,本综述详细调研了价值观对齐和能力对齐这两大方面的现有对齐技术,并提出了全新的视角来总结归纳这些技术,详细梳理了现有技术的发展脉络。我们将大模型对齐按照人类的学习范式分为三类:从人类偏好模仿学习、从反馈信号归纳学习,以及通过思考和沟通实现自主对齐,全面覆盖了当前对齐技术。此外,现有综述并未包含以 DeepSeek-R1^[30]为代表的长推理技术。因此,我们还对其能力对齐范式和安全对齐挑战进行了系统论述。本文还详细综述了关于对齐的评测方法,探讨了未来实现与人类的价值观和各项能力全面对齐技术的可能发展方向。

本文的组织结构如下:第 1 部分阐述了大模型对齐的基本概念和意义,以及对于该技术的分类角度及其依据;第 2 到第 4 部分则分别从各个角度全面综述了现有的大模型对齐方法;第 5 部分阐述了如何对大模型进行评测;第 6 部分进行了总结。在每一部分末尾,基于调研和实践经验提出了每类对齐技术的发展建议,以推动大模型进一步与人类对齐。

1 大语言模型对齐方法概述

大语言模型对齐方法大体上可以划分为两类:价值观对齐和能力对齐,其核心目的是确保大语言模型的输出与人类的道德准则和行为对齐,与人类的长远利益保持一致。基于这一目标,研究者们从多个方面提出了大语言模型对齐方法。经过详细的调研、总结和归纳,我们从学习范式的角度将对齐方法划分为 3 大类:从人类偏好模仿学习、从反馈信号归纳学习,以及通过思考和沟通的自主对齐。

从人类偏好模仿学习指大模型直接从数据中学习人类的价值观偏好和所包含的知识,以提高模型的安全性和能力。这一范式包括基于指令微调的对齐方法^[31-33]和直接偏好学习的对齐方法^[29,34-35]。前者指大模型通过预测下一个词的方式学习数据集中每个问题下的预期输出行为,使模型能够模仿该情境下的回答方式。直接偏好学习的对齐方法则通过对比学习的思想,让大模型在偏好对(Preference pair)上学习更优质的输出行为,同时使其避免不良行为。这两种方式本质上是让模型模仿数据集中的行为来与人类对齐。

从反馈信号归纳学习则指大模型通过反馈信号间接学习人类的偏好和价值观并提高其能力。这包括基于规则反馈^[36]、基于模型反馈^[37-41]、基于环境反馈^[42-44]和基于多途径反馈融合^[30,45]的四种学习范式。基于规则的反馈通过正则匹配的方式来检验模型的推理结果和格式等的正确性,并以此作为反馈信号为模型提供优化方向。基于模型的反馈则利

用偏好数据建模一个能表达人类偏好的代理奖励模型,并以此输出指导大模型优化的监督信号。与基于指令微调的对齐方法相比,基于模型的反馈可以学习到人类偏好数据集中更通用的偏好知识。基于环境的反馈则通过代码沙盒等环境提供丰富的反馈信号来指导大模型对齐。同时,这些反馈信号可以融合成一个完整的奖励系统,从多角度优化模型,使其与人类对齐。

与前两种对齐范式不同,自主对齐则强调将大模型视为一个独立的个体,使其通过自我反思、批判和对弈等方式实现自身的进化,或通过与其他个体之间的博弈和协作来实现在群体层面与人类目标的对齐。该对齐方法可以分为通过反思和自省等方法实现自我对齐^[46-49],以及通过多模型之间的相互交流实现群体对齐^[50-52]。

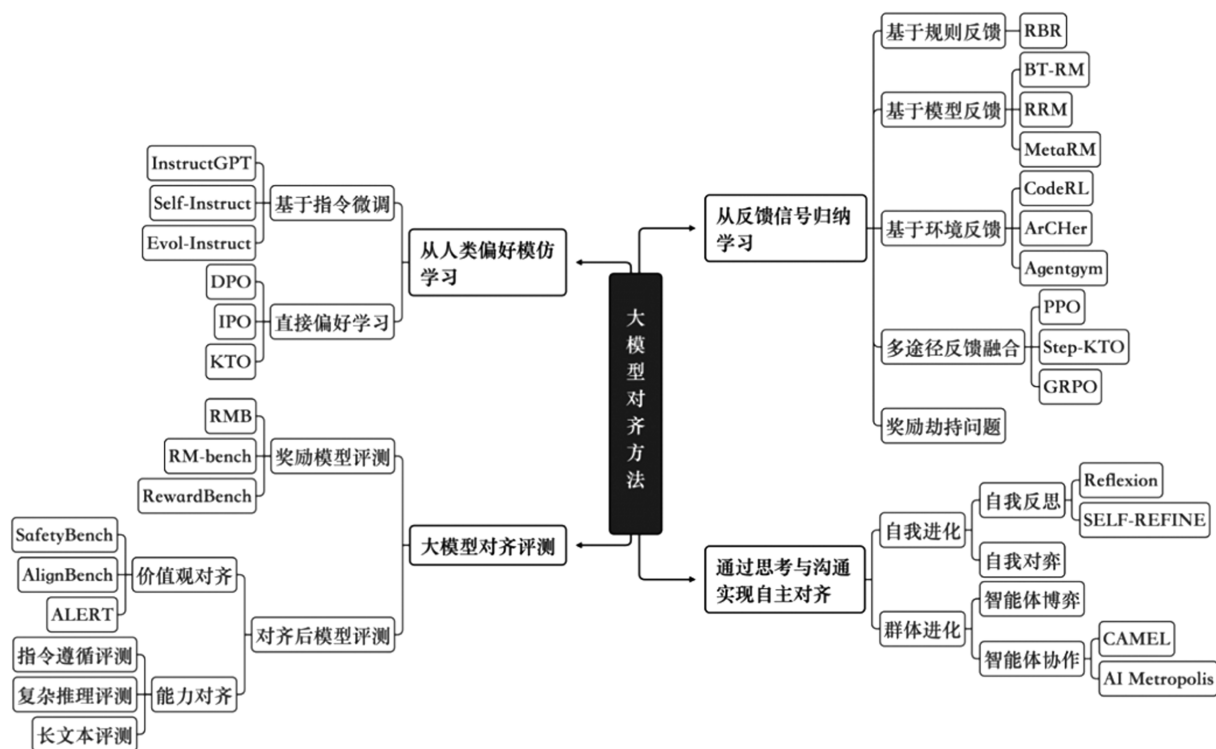
表1中总结了当前大模型对齐技术中常用的训练数据集。同时,这些对齐相关的部分技术方法可以简要表示为图1。下面将对每类对齐范式进行详细阐述。

表1 主流对齐训练数据集简介

序号	数据集名称	适用范围	数据规模	语言类型	数据集简介
1	HH-RLHF ^[41]	价值观对齐	170K	英文	HH-RLHF 由 Anthropic 发布,用于训练模型的有效性和无害性
2	SafeRLHF ^[53]	价值观对齐	83.4K	英文	SafeRLHF 是由北京大学发布的高质量人类偏好数据集,用于训练模型的安全性和有用性
3	UltraFeedback ^[54]	价值观对齐	256K	英文	UltraFeedback 是大规模偏好训练集,涵盖指令遵循、真实性、诚实性与帮助性四个维度
4	CValues-Comparison ^[55]	价值观对齐	145K	中文	CValues-Comparison 通过 self-instruct、大模型生成等方式收集了 145K 条偏好对数据,用于训练模型的安全性和有用性
5	GSM8K ^[56]	能力对齐(数学)	8K	英文	GSM8K 是人类编写的小学数学文字题,旨在训练数学推理和解题能力
6	Math ^[57]	能力对齐(数学)	12.5K	英文	Math 是数学竞赛题目,每题配有详细的逐步解答
7	MathCodeInstruct ^[58]	能力对齐(数学,代码)	80K	英文	MathCodeInstruct 是由 GPT-4 Code Interpreter 自动生成的高质量数学问题与代码解题对,专注于数学和编程的结合
8	The Stack v2 ^[59]	能力对齐(代码)	/	多语言	The Stack 包含来自 GitHub 的开源代码,涵盖 358 种编程语言,广泛用于训练编程相关任务
9	mbpp ^[60]	能力对齐(代码)	1K	英文	mbpp 包含 974 个原创问题,每个问题配有描述、解决方案代码和测试用例,专注于基础到中级的 Python 编程任务,涵盖多种编程概念和算法
10	MSAgent-Bench ^[61]	能力对齐(智能体,工具调用)	600K	中文	MSAgent-Bench 是用于训练大语言模型工具调用能力的大规模多插件指令微调数据集,提升 LLM 的插件执行和响应能力
11	GAIA ^[62]	能力对齐(智能体,工具调用)	450	英文	GAIA 数据集用于大语言模型在工具调用、自主决策、多步推理和跨领域任务中的综合能力

续表

序号	数据集名称	适用范围	数据规模	语言类型	数据集简介
12	ToolBench ^[63]	能力对齐(智能体,工具调用)	120K	英文	ToolBench 是自动构建的高质量指令微调数据集,涵盖数千个真实世界的 API,旨在赋予开源 LLMs 通用的工具使用能力,提升模型的实用性



2 从人类偏好模仿学习的对齐方法

该对齐方法旨在让大模型从数据中直接学习人类偏好,从而对齐人类正确价值观和所需能力。根据人类偏好在数据中的不同表现形式,我们将现有方法划分为两类:在问答对形式的数据上,基于指令微调技术直接模仿标准行为的对齐方法;在偏好对数据上,通过对比学习的思想使模型学习相对正确的回答和行为,同时避免产生那些有害行为的对齐方法。

2.1 基于指令微调的对齐方法

基于指令微调的对齐方法指大模型利用有监督微调(Supervised Fine-Tuning, SFT)方法^[64-66]在人类专家标注的高质量问答对上进行优化,使其模仿这些标准回复方式从而与人类对齐。研究人员通过调整数据分布让大模型对齐正确的价值观和所期望

的目标能力。影响该对齐方法有效性的关键因素是指令数据的质量^[33,67]。

早期的工作主要依赖开源数据集的整合。例如, FLAN^[65]通过混合 170 多个自然语言处理任务构建了一个包含数百万条指令的数据集。虽然这种粗粒度的整合能够提升任务覆盖率,但存在语义冗余和知识冲突等问题^[28]。当前构造对齐数据的主流方法逐渐转向人工标注的方式。例如, InstructGPT^[28]构建了专家驱动的人工标注流程,通过“指令撰写-多维度标注-对抗性测试”的方式迭代地构建高质量数据,其标注密度达到每指令 5.7 个质量维度注释。Databricks-dolly-15K^[68]则是根据上述标注流程和指令分类体系,由企业级专业团队精心标注构建。

另一方面,一些研究致力于从经过对齐的大模型(教师模型)中通过蒸馏的方式来构建指令微调数据,并在未对齐的模型(学生模型)上进行指令微调以实现高效对齐^[69-72]。相比于人工标注的数据,这

种合成数据的方法更加经济高效,能够快速生成大量数据。然而,由于对齐后的大模型仍可能出现幻觉和未完全对齐的情况,蒸馏得到的数据难以完全符合人类偏好^[73-74]。因此,一些研究旨在设计更为精细的指令数据蒸馏框架以提升从教师模型中合成数据的质量。Lee 等人^[75]提出了一个自动化指令生成框架,通过多个语言模型集成合成并过滤指令微调数据,以确保合成数据的质量。Self-Instruct^[76]则使用 175 条人工编写的指令作为初始种子,以确保合成数据的初始质量,并在此基础上生成更多相对高质量的数据。Xu 等人^[77]则提出了 Evol-Instruct 方法,从少量高质量指令开始,逐步增加指令的复杂度以合成更高质量的数据。

越来越多的研究表明,这种通过蒸馏来合成对齐数据的方法能够实现高效的对齐。这被视为对齐知识从对齐模型到未对齐模型的转移。DeepSeek^[30]发现在某些场景下,基于蒸馏数据的指令微调对齐方法比直接在小模型上应用其他高效对齐方法(如 DPO 和 RLHF 等)更为有效。为了进一步挖掘其有效性,一些工作深入研究了如何改进这种从蒸馏数据中学习的对齐框架^[78-79]。例如,Jiang 提出了一种对抗蒸馏方法^[80],通过比较教师模型和学生模型的初始回答从而筛选出难度不同的问题来改进对齐数据。同

时,该方法迭代优化了学生模型的生成能力,使学生模型能够在更少对齐数据下获得同样的推理能力。此外,近期发布的一系列长思维链大模型也证明了该对齐方法在蒸馏数据上的有效性^[81-83]。Huang 等人^[84]展示了仅使用数万条蒸馏自 o1 的长推理链数据,对更小的模型进行指令微调,即可在美国邀请数学竞赛(AIME)等任务上超越 o1-preview,在复杂推理能力上实现高效对齐。

2.2 基于直接偏好学习的对齐方法

尽管基于指令微调的对齐方法表现出了良好的有效性,其仍然存在诸多局限性,例如对齐后模型在泛化能力和回答多样性等方面的表现尚不尽如人意^[85-86]。相比于基于指令微调的对齐方法,另一类从人类偏好数据中模仿学习的对齐方法则让大模型从一组行为偏好中学习来与人类对齐^[29,87-90]。该方法的核心思想在于模型学习的不仅仅是唯一的“正确行为”,而是一组存在相对好坏的偏好序列。模型通过对比的方式,在偏好数据对上学习那些更好的行为,同时避免产生较坏的行为。这有助于提高对齐的泛化能力,让模型学习到每个问题下更多的行为方式。表 2 总结了基于直接偏好学习对齐中重点方法的主要贡献和局限性。

表 2 基于直接偏好学习的对齐方法

序号	方法名称	主要贡献	局限性
1	DPO ^[29]	最先提出直接偏好优化的方法,省去了奖励模型的建模过程	数据偏好明显时模型会很快过拟合,导致模型局部收敛
2	IPO ^[87]	通过在损失函数中加入正则项解决了 DPO 过拟合的问题,并提供了理论框架支持	仍处于理论探索阶段,在真实任务中的性能提升较弱,仍需进一步提升
3	KTO ^[88]	引入经济学中的“前景理论”,指出只需获得回答是否可取的二元信号即可训练模型,解决了传统偏好优化算法中偏好数据获取成本高的问题	对于捕捉细粒度偏好表现不佳,模型泛化能力在复杂对齐任务中存在不足
4	CPO ^[89]	揭露了大模型安全对齐中的对齐税现象,并在损失函数中引入了可控偏好条件,实现多目标对齐的灵活控制	当前仅在 3H 原则的多目标对齐任务中进行实验,缺乏对复杂目标结构的一般化扩展
5	ORPO ^[90]	将监督微调与偏好对齐统一建模,简化训练流程并改善梯度竞争问题	模型泛化能力尚未在多样化任务中得到充分验证,适应性仍有待提高

在阐述基于直接偏好学习的对齐方法之前,我们先简要定义所需的概念。具体而言,我们首先构造该方法所需的偏好数据集。针对数据集中每一个问题 x ,首先从单个或多个模型中采样得到回答集合 $Y = \{y_1, y_2, \dots, y_n\}$ 。随后标注人员根据正确的价值观按照回答的好坏对其进行排序,以形成明确的偏好关系对。例如, $y_w > y_l$,表示回答 y_w 相较

于回答 y_l 更好。该对齐方法旨在让模型学习这种数据对中的偏好关系,让模型输出的行为趋近于那些较好的回答,并避免那些较差的回答。

Rafailov 等人提出的直接偏好优化算法(DPO)是该对齐范式下的开创性工作^[29]。DPO 借鉴了 Bradley-Terry 模型^[91]的偏好建模思路:假定偏好对数据 $D = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$,Bradley-Terry 模型

将偏好数据建模为偏好分布,如式(1)所示。

$$p(y_w \succ y_l | x) = \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp(r^*(x, y_l))} \quad (1)$$

其中, $r^*(x, y)$ 代表样本的偏好评分。DPO 在 Bradley-Terry 模型的基础上将策略模型 π_θ 相对于参考模型 π_{ref} 生成该回答的对数概率比作为其隐式偏好评分,如式(2)所示。

$$r^*(x, y) = \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)} \quad (2)$$

DPO 的核心思想是使模型最大化偏好样本 y_w 相对于非偏好样本 y_l 的生成概率,从而使得模型的输出更符合人类偏好。该方法通过优化式(3)逐渐降低来实现对齐。

$$L_{\text{DPO}} = -E_{(x, y_w, y_l) \sim D} \log \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \quad (3)$$

值得注意的是,直接偏好优化算法在偏好信号较为明确的情况下会出现过拟合的现象,导致模型仅对齐了局部的人类偏好。为了解决这一问题,Google DeepMind 提出了 IPO^[87]。该方法在 DPO 的基础上加入了额外的正则化项,以防止过拟合并提高训练的稳定性,使模型在复杂偏好场景下表现得更加鲁棒。IPO 的损失函数如式(4)所示,通过优化将其降低来逐渐对齐数据中偏好。

$$L_{\text{IPO}} = -E_{(x, y_w, y_l) \sim D} \left(\log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} - \frac{1}{2\tau} \right)^2 \quad (4)$$

其中,参数 τ 作为平滑因子来控制偏好关系的强度。当 τ 较小时模型的优化过程更平滑,有助于减少模型对偏好数据的过度拟合,随着 τ 取值增大, IPO 会逐渐退化为 DPO。

尽管 DPO 和 IPO 证明了直接偏好对齐的有效性,但这些方法对偏好数据的质量要求较高。通常这些数据需要标注人员依据“HHH”(Helpful, Honest, Harmless)原则或更精细的准则进行人工标注。同时标注难免会引入误差,标注人员也难以通过有限的数据准确建模人类偏好。高质量数据的获取难度显著地降低了 DPO 和 IPO 算法的有效性,限制了这些算法的大规模使用。为了解决这一问题, Xu 等人提出了 KTO^[88]。相比于 DPO 通过学习两个回答的相对偏好关系(即 $y_w \succ y_l$)来对齐人类偏好, KTO 发现直接将每个回答作为独立的对象,并根据其好或坏的类别来选择合适的损失项

优化大模型,也能达到对齐人类偏好的效果。同时这种方法降低了数据获取的难度,有效地将直接偏好优化扩展到更大规模的应用场景。KTO 的损失函数如式(5)所示。其中, z_{ref} 是当前模型相对于参考模型的 KL 散度期望值,如式(6)所示。

$$L_{\text{KTO}} = -E_{(x, y) \sim D} \left[\lambda_w \sigma \left(\beta \log \frac{\pi_\theta(y_w | x)}{\pi_{\text{ref}}(y_w | x)} - z_{\text{ref}} \right) + \lambda_l \sigma \left(z_{\text{ref}} - \beta \log \frac{\pi_\theta(y_l | x)}{\pi_{\text{ref}}(y_l | x)} \right) \right] \quad (5)$$

$$z_{\text{ref}} = E_{(x, y) \sim D} [\beta \text{KL}(\pi_\theta(y | x) | \pi_{\text{ref}}(y | x))] \quad (6)$$

它可以用于防止模型过度偏离原始分布而导致的优化不稳定甚至失败,其中, λ_w 和 λ_l 充当了权重因子,用于调节相对较好和较坏的样本的损失贡献。

与此同时,另一项研究揭示了偏好学习中的一个重要现象:对齐税(Alignment Tax)^[89]。研究表明,在进行某一特定目标的对齐时,可能会导致其他目标的对齐性能下降。例如,提高模型的安全性可能会牺牲其部分的创造能力。为了解决这一问题,研究者提出 CPO,该方法通过引入多目标偏好评分,使模型能够在多个对齐目标之间进行权衡,从而实现更加全面的偏好对齐。CPO 的优化目标如式(7)所示。

$$L_{\text{CPO}} = -E_{(x, c, y_w, y_l) \sim D} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w | c, x)}{\pi_{\text{ref}}(y_w | c, x)} - \beta \log \frac{\pi_\theta(y_l | c, x)}{\pi_{\text{ref}}(y_l | c, x)} \right) \right] \quad (7)$$

其中, c 表示条件提示信息,通过类似“<Helpfulness:5>”的方式传递,使模型能够明确知道在对齐过程中应侧重的具体目标,从而在多项指标间实现平衡。

另一方面, Jiwoo 等人还发现,虽然 DPO 等方法能够使模型生成更符合人类偏好的回答,但是损失函数也会适当优化模型来增加偏好数据中那些较差回复的生成概率^[90]。这和 DPO 的设计目标相违背,即在同一问题下模型应该逐渐学习模仿那些相对较好的行为同时抑制相对那些较差的行为。为了解决这一现象, Hong 等人提出了 ORPO,将监督微调(SFT)与偏好对齐过程融合为一个统一的训练步骤,而非像传统方法那样分阶段训练。ORPO 在训练过程中引入了负对数似然损失(NLL)和一个基于奇异比的损失项,前者鼓励模型生成偏好样本,后者则通过强化偏好样本、惩罚非偏好样本来对偏好关系建模。ORPO 的损失函数如式(8)所示。

$$L_{\text{ORPO}} = -E_{(x, y_w, y_l) \sim D} \left[\log p_{\theta}(y_w | x) + \lambda \log \sigma \left(\log \frac{p_{\theta}(y_w | x)}{1 - p_{\theta}(y_w | x)} - \log \frac{p_{\theta}(y_l | x)}{1 - p_{\theta}(y_l | x)} \right) \right] \quad (8)$$

ORPO 通过结合基于指令微调的对齐方法和本节中基于直接偏好学习的对齐方法,使模型在模仿偏好数据集中正确行为的同时学习在同一问题下其他回答之间的偏好关系,从而进一步促进模型从数据集中学习。

另一类方法则将这种偏好关系建模为奖励模型来泛化表示人类偏好,并以此提供反馈信号实现模型与人类的对齐。相比于本节中基于直接偏好学习的对齐方法,该方法能够泛化地建模数据集中蕴含的正确价值观,从而进一步提高对齐的有效性和泛化性。这类基于奖励模型反馈的对齐方法将在后续的“从反馈信号中归纳学习的对齐方法”一节详细介绍。

2.3 总结与发现

从人类偏好模仿学习的对齐方法能够在特定领域利用较低的资源展现显著的性能,但通常缺乏学习和泛化价值与知识的能力。具体来说,尽管监督微调能够快速适应特定任务,但学习模式本质上是“记忆-重现”式的,即便采用蒸馏合成数据等扩展方法也难以突破训练数据分布的界限。基于直接偏好

优化的对齐方法通过引入相对偏好学习,在某种程度上提升了对齐的泛化性,但其优化范围仍然受制于标注的偏好数据。这些对齐方法共同面临的问题是对特定领域数据的高度依赖,而这种学习范式并未赋予模型真正的通用推理和自适应能力,也就无法让大模型从特定领域的数据中学习泛化的知识。后续的对齐方法在一定程度上解决了这种通过模仿学习对齐人类偏好的方法难以泛化的问题。

3 从反馈信号中归纳学习的对齐方法

本节探讨了一种更加有效的对齐方法:从反馈信号中归纳学习。该方法让大模型根据问题探索答案,并根据生成答案的质量提供反馈信号。该反馈信号通常作为奖励信号应用于强化学习框架中,通过策略梯度算法来优化大模型,使其逐渐生成高奖励值的回答,从而在反馈信号的指导下实现大模型与人类的高效对齐。根据反馈信号的来源不同,我们将这种方法分为两大类:基于单途径反馈的对齐方法和基于多途径反馈融合的对齐方法。前者又可以进一步细分为三类:基于规则的反馈、基于模型的反馈、基于环境的反馈。表 3 和表 4 分别总结了各对齐范式下重点方法的主要贡献和局限性。随后,我们也详细阐述了该对齐范式所面临的重大挑战:奖励劫持。

表 3 基于单途径反馈的对齐方法

序号	反馈类型	方法名称	主要贡献	局限性
1	规则反馈	RBR ^[92]	利用规则反馈有效提高模型安全性和准确性,减少对人类数据的依赖,能够细粒度控制模型行为,具有广泛的适用性	规则定义较为复杂,主观性任务难以判别,可能会放大语言模型中的潜在偏差
2	规则反馈	Logic-RL ^[45]	在逻辑推理任务上对齐模型,并有效泛化到数学任务上	数据集规模较小,任务较为局限
3	规则反馈	DPP ^[93]	设计了一个完全自动化的基于规则的数据选择流程	规则生成可能存在偏差,并且多样性和质量没有得到保证
4	模型反馈	RRM ^[94]	基于模型集成的方法来过滤偏好数据集中潜在的错误偏好和模糊偏好	需要高昂的计算资源训练,且奖励模型容易拟合错误偏好
5	模型反馈	MetaRM ^[95]	基于人类偏好的因果关系来进行偏好数据增强,减少奖励劫持行为	所提出方法仅关注增强偏好数据,增加了计算资源
6	模型反馈	DIAL ^[96]	通过元学习来对抗环境分布变化,不断地从新的策略模型中采样新的回复,提升模型的分布外泛化	迭代过程中依赖强模型来进行偏好标注,优化仍存在上限
7	模型反馈	URM ^[97]	用 Wasserstein 距离最小化源域和目标域之间的分布差异,并将其迁移到目标领域,从而实现跨领域的偏好对齐	对于领域自适应如何解决目标任务数据稀缺问题缺乏分析
8	模型反馈	ALaRM ^[98]	建模人类的偏好不确定性,集成多个奖励模型来量化奖励信息的可靠性	该方法利用回归训练来减小方差,但难以建模真实偏好分布的不确定性

续表				
序号	反馈类型	方法名称	主要贡献	局限性
9	模型反馈	RRM ^[94]	基于任务相关的维度指标或训练奖励模型作为维度奖励,筛选出与整体奖励一致性较高的维度奖励作为优化信号	需要针对每个具体任务设计专门的奖励函数,在不同任务之间难以直接迁移和复用
10	环境反馈	CodeRL ^[99]	将强化学习应用于大模型代码生成任务的首篇工作,利用代码沙盒提供反馈信号并将其视为强化学习的奖励信号,以提升模型的代码生成能力	反馈信号的数值化方法设计简单,奖励信号过于稀疏,对大模型提升的效果有限
11	环境反馈	RLTF ^[100]	在提供反馈信号时不仅仅根据模型生成的代码的正确性,还将报错信息数值化为反馈信号,提供了更细粒度的反馈信号	反馈的监督范式仍然是结果监督,在面对更加复杂的编程问题时模型可能难以探索到正确答案,导致强化学习训练停滞
12	环境反馈	StepCoder ^[101]	将逆课程学习引入基于代码沙盒反馈的强化学习中,提高了大模型对于那些复杂编程问题的探索效率	训练集中每一个编程问题需要对应的标准答案才能设计逆课程帮助探索
13	环境反馈	ArCHer ^[102]	将层次化强化学习引入多轮交互推理任务,提供了细粒度的奖励信号,显著提高了大模型在工具学习等任务上的表现	层次化的结构使得训练复杂,消耗的计算资源更多,同时需要设计每一轮的奖励函数
14	环境反馈	StepTool ^[103]	将工具使用视为序贯决策问题,采用策略梯度方法多步优化模型,并通过设计细粒度奖励来帮助模型理解,以提升大模型在工具学习中的性能	多步训练会使得训练相比于其他方法更不稳定,同时多步训练也会消耗更多的计算资源
15	环境反馈	iTool ^[104]	通过蒙特卡罗树搜索的路径探索来增强合成数据的多样性,同时迭代地改善那些存在缺陷的数据	蒙特卡罗树搜索算法需要消耗大量的资源,同时通过合成的方式获得的数据可能与真实场景存在差异,影响模型的泛化能力
16	环境反馈	Agentgym ^[105]	首次提出了综合了多种智能体任务,包括推理和规划在内的反馈环境,为提高智能体的推理规划能力奠定了基础	集成的训练算法的种类可以进一步扩增
17	环境反馈	AdaRefiner ^[106]	使用轻量级模型自动细化任务理解。在保持了大模型泛化能力的同时,增强了其在下游任务中的决策能力	较小语言模型的理解能力可能不足以提供环境和智能体所需的必要分析,同时该方法在更复杂的任务上的表现仍存在较大提升空间

表 4 基于多途径反馈的对齐方法

序号	反馈途径	方法名称	主要贡献	局限性
1	模型,环境	Themis ^[107]	促进工具利用和奖励模型之间的协同作用,增强了解释能力和评分可靠性	需收集大量高质量的工具调用轨迹数据
2	规则,模型	Step-KTO ^[108]	结合来自奖励模型的过程奖励和规则判别的结果奖励,指导 LLMs 生成更值得信赖的推理轨迹,保证了最终答案的准确性和中间推理步骤的质量	过程奖励模型训练需收集高质量的推理轨迹路径;结果奖励依靠规则判别准确性的方式来误判一些有格式差异但实际正确的数据
3	规则,模型,环境	MPLSandbox ^[109]	设计了一个开箱即用的多编程语言沙盒 MPL,能够提供统一编译器反馈,并对大模型生成的代码进行全面的分析	沙盒无法覆盖用户需要的所有依赖,用户需自行使用脚本扩展依赖;仅支持八种常用的编程语言
4	规则,模型,环境	FullStackCoders ^[110]	提出了一个全面的评估框架,并具备用于代码智能评测的执行环境,旨在评估现实代码开发方案中的多语言编程功能	沙盒依赖固定的评测流程,很难自定义扩展

续表

序号	反馈途径	方法名称	主要贡献	局限性
5	规则,模型	GRPO ^[30]	结合多阶段训练和冷启动数据,发布了 DeepSeek-R1 长推理模型,实现了与 OpenAI-o1 在推理任务上相当的表现	对提示敏感性较强;仅局限于推理任务,未进一步扩展至软件工程任务;在工具调用、多轮对话、复杂角色扮演等任务的能力较弱
6	规则,环境	R1-Searcher ^[111]	将仅依赖于 RL 激发模型推理能力的训练方式扩展至检索能力,允许模型自主调用外部搜索引擎在推理过程中访问外部知识	仅在参数较少的模型上进行探索,未尝试在更大规模参数的模型上训练;数据课程较为简单

3.1 基于规则反馈的对齐方法

使用显式规则控制 AI 系统行为的思想在自然语言处理历史中由来已久。在统计学习与深度学习兴起之前,早期如专家系统等多以规则系统为主导^[91]。近年来,这种通过显式规则约束模型行为的思想在大模型对齐中被重新拾起,并逐渐演变为一种新的对齐方法,即基于规则反馈的对齐方法。

具体而言,该方法通过衡量模型输出与一组预定义显式规则之间的一致性来构造用于引导学习的奖励信号^[92]。这些规则通常涵盖多个维度,如安全性、准确性、格式约束和语言一致性,用于识别模型生成结果中的期望行为或禁止行为,从而实现对结果质量的评估^[30,45,92]。

当前已有多种方法采用规则反馈机制来实现大模型与人类的对齐。OpenAI 的 RBR^[92] 框架是最具代表性的方法之一,主要聚焦于大模型的安全性行为对齐。该框架通过一系列规则(通常细化为二元命题)定义期望与非期望行为,包括对敏感请求给出“硬拒答”(即包含简短道歉,避免带有评判性语言)、对情绪性请求采用“软拒答”(即以共情语气回应)以及对于良性请求给予正常回应等。它对模型输出进行逐条命题打分,并以此作为对齐的反馈信号。

规则反馈信号还用于指导大模型对齐人类的能力。例如,DeepSeek-R1^[30] 的规则反馈系统可以在数理逻辑推理任务和指令遵循等任务中提供反馈信号。在具有明确正确答案的推理任务中,可以通过模型生成答案的准确性和通过率等指标作为反馈信号来指导大模型优化。此外,通过格式判断可以确保大模型的回答遵循特定格式。Logic-RL^[45] 则专注于提高大模型的逻辑推理能力,通过在逻辑谜题任务中施加格式奖励与答案奖励,来引导模型形成连贯的推理链条和正确结论。此外,规则反馈还可以用于数据筛选^[112]。这些方法表明,规则反馈方

法在安全性对齐、能力对齐和数据筛选等多个场景中发挥了关键作用。

总的来说,规则反馈对齐是一种简洁且高效的从反馈信号中归纳学习的对齐方法,已经在各种任务中展现出巨大潜力^[30,45,112]。然而,规则反馈方法的有效性高度依赖于规则设计的质量与完整性。在处理主观性强、语境复杂的任务时,规则反馈方法可能难以独立胜任;同时,奖励劫持风险、LLM 评分器质量不高、伦理偏见转移等问题亦需引起重视。因此在多数应用中,规则反馈方法更适合作为初级行为约束手段,并与其他反馈信号相互配合,以更全面地对齐目标。

我们将在本节的“总结与发现”中进一步阐述基于规则反馈对齐与其他类别对齐方法相比所存在的优缺点和各自面临的挑战。

3.2 基于模型反馈的对齐方法

基于模型反馈的对齐方法指通过一个奖励模型来建模人类的价值观和偏好,并以此生成反馈信号来指导对齐。人类反馈的强化学习(RLHF)^[28] 是该对齐范式下被广泛应用的框架,其包含三个训练步骤:有监督微调、奖励建模和基于模型反馈的强化学习。有监督微调通过少量数据让模型具备基本的指令遵循能力。奖励建模阶段则在人类偏好数据集上训练一个奖励模型,以建模人类偏好并对回复进行打分。最后,基于模型反馈的强化学习则根据奖励模型提供的反馈信号,利用强化学习算法优化大模型与人类对齐。奖励建模是 RLHF 中的关键步骤,它决定了对齐信号是否足够鲁棒、泛化且精准。

除了上述从偏好数据集中建模奖励模型的方式外,一些研究者也探索了利用现有对齐后大模型作为奖励模型的对齐方法。

3.2.1 基于偏好数据建模的奖励模型

在基于偏好数据建模的对齐方法中,偏好数据

质量是影响最终对齐表现的一个尤为关键的因素^[39,41,113]。现有研究指出,开源偏好数据集中存在大量模糊甚至错误的偏好数据^[39]。包括 OpenAI 和 Anthropic 在内的多个研究机构^[41]曾报告标注人员的标注一致性仅有 60%至 70%,这严重阻碍了奖励模型准确建模人类的真实偏好意图。基于此,Wang 等人^[113]提出了一种偏好数据筛选方法,试图利用奖励模型自身对低质量数据的识别能力来标识数据质量。

然而,尽管偏好数据质量决定了奖励模型的表现,但获取高质量数据是极其消耗资源的。同时,由于人类偏好的复杂性和不确定性以及上述提到的人类标注人员的不一致性,数据中难免存在噪声^[114]。如何在有限数据资源上提高奖励建模的泛化性和稳定性是当前研究者努力探索的主要方向^[94,115-116]。MetaRM^[95]和 DIAL^[96]尝试提高奖励建模阶段的泛化性,来在有限的数据集上增强模型的泛化能力。MetaRM 指出,随着大模型逐渐对齐,其生成的答案逐渐偏离了奖励模型的打分分布,从而导致对齐后期奖励模型的稳定性下降。基于此,MetaRM 在奖励建模中引入了元学习,使建模过程保持一定的泛化性。DIAL 则指出,人类偏好往往是通用的,可以被奖励模型捕捉并用于迁移学习。他们通过在源域中收集人类偏好反馈来训练奖励模型,使用 Wasserstein 距离最小化源域和目标域之间的分布差异,并将其迁移到目标领域,从而实现跨领域的偏好建模。此外,URM^[97]通过学习不同维度下的偏好分布来评估奖励模型打分的可靠性和置信度。ALARM^[98]则通过层次化学习,即结合整体奖励和特定场景的奖励,提供更准确和一致的监督信号。

上述工作旨在评估回复质量,并对其分配一个最终的奖励值(基于结果监督的奖励建模)。然而,当回答的推理链路过长时,大模型往往会因为奖励信号过于稀疏而难以探索到正确的答案^[101,117-119],从而降低对齐的效率。一些工作尝试对模型生成的回复的中间步骤进行细粒度评估和打分(基于过程监督的奖励建模)^[120-125],以细化监督信号从而提高模型的探索效率,最终提高对齐表现。

Lightman 等人^[120]在数学推理领域深入比较了结果监督和过程监督这两种奖励模型,结果显示过程监督能够更精确地指出错误发生的位置,有望为对齐方法提供更加细粒度的奖励信号。Chen 等人^[124]进一步改进了结果监督的建模范式,通过价值模型提供双向奖励信号,同时评估历史推理步骤

的正确性和未来成功的概率,从而更全面地指导大模型数学能力的对齐。然而上述方法需要步骤级的偏好数据,这往往提高了对数据标准质量的要求以及标注的难度。Math-shepherd^[125]则提出了一个自动过程标注框架,用于解决过程监督训练数据依赖人工标注的问题,以节约训练成本。

基于偏好数据的奖励建模能有效学习符合人类期望的行为模式,并引导大模型生成更高质量的回复。然而,获取高质量的偏好数据并非易事,标注过程往往需要耗费大量人力和时间。同时,这种直接输出打分数值的奖励模型往往不具有可解释性,人们难以探索打分背后的依据。因此研究者们逐渐探索其他更泛化且具有可解释性的方法,来为基于强化学习的大模型对齐提供更加准确的反馈信号。

3.2.2 基于大模型验证的奖励模型

为了解决上述基于偏好数据建模的对齐方法所存在的问题,研究者们尝试利用对齐后的大模型作为验证器来提供对回答的评估和打分,并以此作为指导对齐的反馈信号^[126-130]。这种方法的优势在于,验证过程中可以充分利用大模型在预训练阶段所学习到的广泛世界知识,并且可以让大模型在打分时阐述原因,从而提供一定的可解释性。

Zhang 等人^[131]提出了 GenRM,通过疑问句的方式构造模板,来询问大模型一个回答是否满足了所对应的问题。随后,GenRM 将大模型生成的第一个“是”或“否”词元的概率作为其对这句话正确与否的判断,并以此作为其他模型对齐阶段的反馈信号。另外,Mahan^[132]也尝试利用用户偏好数据微调大模型,让其在具备预训练阶段所学习到的知识的同时也引入人类偏好,从而进一步提高建模的泛化和准确性。

然而,该方法在利用预训练知识的同时,也难免遇到预训练阶段学习到的且没有完全被对齐消除的偏见信息^[133]。这些偏见信息可能在验证回答时显现,导致评分不准确。一些研究尝试在验证过程中引入明确的评判准则,以提高大模型评估的客观性并增强奖励信号的准确性^[134-135]。宪法 AI(Constitutional AI, CAI)^[134]通过一组人为定义的宪法规则,引导大模型生成自我批评并修正回答,然后基于修改后的安全回答微调原始模型。通过这种方法,大模型可以有效处理有害查询并能够解释其反对的理由。此外,其他研究者也提出了 Deliberative Alignment^[135],针对不同类别的安全问题制定相应规则,并将这些规则与问题一起传递给验证模型,从

而提升整体验证的可靠性。

这种基于大模型和准则验证的方法虽然能够提供更为准确和泛化的反馈信号,但准则的设计仍然值得未来进一步探索。准则能否准确、全面地覆盖问题所对应的所有回答,是这种方法是否有效的一大关键。

3.3 基于环境反馈的对齐方法

基于环境反馈的对齐方法旨在通过构建一个模拟环境或沙盒让大模型与其交互,从而间接获取人类偏好和价值观的反馈信号。这类方法利用模拟环境提供的反馈信号,如任务完成情况、用户交互数据和环境条件等,对模型进行优化和调整,使其行为更加符合人类期望。这种方法在代码生成纠错^[136-139]、工具学习^[140-141]、任务规划^[142,143-144]和智能体^[133-134]等人类能力对齐任务中尤为常见。

一些工作尝试让模型与人类的代码能力进行对齐^[99,145]。这些方法通过单元测试验证模型生成的代码,并将代码的正确性作为强化学习的奖励信号,从而指导模型的优化。代表性方法包括 CodeRL^[99]、PPOCoder^[42]、RLTF^[100]和 StepCoder^[101]等。CodeRL^[99]首次将模型生成的代码在沙盒环境中进行验证。如果代码通过所有测试,模型会获得正向奖励,反之则收到负向奖励。这种通过沙盒环境执行得到的反馈信号作为策略梯度的监督信号,进一步优化模型以生成正确的代码并避免生成错误代码。PPOCoder^[42]则将强化学习算法改为近端策略优化算法来进一步提高对齐表现,并提供了能够表明代码语义信息等方面的更加丰富的反馈信号。RLTF^[100]进一步将代码的报错信息映射到奖励信号,以丰富反馈。然而,这些方法需要大模型首先探索到正确的代码才能进行优化。如果模型一直生成错误代码,则环境提供的反馈始终是负向的,导致模型难以探索到正确的解题路径。StepCoder^[101]则结合沙盒环境的反馈和逆课程强化学习,让模型更好地在环境中进行探索。在面对复杂的编程问题时,StepCoder 能根据标准答案为这些问题制定从易到难的课程,来让模型有机会探索到正确答案,从而提高对齐性能。

基于环境反馈的对齐方法在对齐人类工具使用能力时也十分有效。我们希望大模型能够具备与人类类似的工具使用能力,包括使用检索、计算和编程类工具。研究者们让大模型在环境中自由探索工具的使用方法,并根据工具使用的正确性给予反

馈^[102-104,146-148]。和上述基于反馈的方法相同,这种信号会被视为强化学习中的奖励信号来指导大模型优化。然而,工具使用的链路通常比较复杂,包含多个工具的调用和组合工具的使用,这些交互通常通过多轮完成。因此,模型在多轮问题的探索上往往难以一次性探索到正确的答案,这限制了模型对工具调用能力的有效对齐。

为了解决上述问题,ArCHer^[102]将分层强化学习引入多轮交互的工具学习,对每一轮的交互进行监督,从而实现更加细粒度的环境反馈信号。不同于 ArCHer 将工具学习视为多轮交互过程,StepTool^[103]将工具使用过程视为序贯决策问题,将每次工具调用视为一个动作,通过设计细粒度奖励机制来提升模型在多步骤任务中的表现。iTool^[104]则提出通过迭代强化微调的方式提高大语言模型的工具使用能力,让模型从易到难地学习,从而更高效地对齐大模型与人类的工具使用能力。

这种对齐方法也能够复杂任务规划和推理任务中提供有效的反馈信号。MAP^[149]将较大的推理问题分解为多个小的子问题,并使用专门的模块相互交互。环境会对这些子问题的解决情况进行验证,从而让模型对齐人类的规划能力。Zhang 等人提出了 AdaRefiner 框架^[106],在模型解决复杂决策任务时进一步优化其任务理解能力。实验表明这有助于让模型有效被环境引导,从而在更高级别的常识性技能方面对齐人类。Xi 等人则提出了综合多种智能体任务,包括推理和规划在内的反馈环境 AgentGym^[105],让模型在这种多任务信号的协同作用下显著提高其推理和规划能力。

此外,还有众多方法将环境反馈应用到多种任务中,例如序列决策^[150]、社会模拟^[151]和具身智能^[152]等任务。这些任务具有共同的特点,即能够通过定义环境对大模型探索到的答案进行验证,并将其数字化为反馈信号。这种信号可以作为奖励信号用于强化学习中,指导大模型向着人类期望的方向优化,从而在各种能力上与人类对齐。

3.4 基于多途径反馈融合的对齐方法

基于多途径反馈融合的对齐方法旨在综合来自不同途径的反馈信号以优化模型,使其更全面地对齐人类在各个维度上的能力和价值观^[153-154]。这一方法不仅可以利用简单直接的规则反馈对齐人类标准、利用奖励模型预测的反馈对齐人类偏好和价值观,还能够依据从环境中获得的实时反馈对齐人类

的能力。通过整合多种来源的反馈信号,模型能够更精确地调整其行为和策略,实现更高效的优化与改进^[155]。

一些研究者提出了混合奖励系统^[107-110],利用多种不同来源的反馈构建奖励系统,为模型提供多维度的信号,从而更好地优化对齐人类利益。例如,Themis^[107]通过引入来自外部工具的反馈建模奖励模型,促进了环境评分与奖励模型评分之间的协同作用,增强了反馈信号的可靠性。Step-KTO^[108]结合规则反馈与基于模型的过程反馈来优化模型的推理能力。MPLSandbox^[109]则提供了一种开箱即用的多编程语言沙盒,集成了传统的和基于 LLM 的代码分析工具,从而提供综合反馈信号来优化模型的代码能力。SandboxFusion^[110]在此基础上,通过结合代码抽取和合成新的单元测试等反馈方法,实现代码的自动评测,辅助模型对齐人类的代码能力。

此外,随着以 OpenAI-O1^[156]为代表的长推理模型的出现,这种基于多途径反馈信号进行多阶段训练的能力对齐方法也逐渐得到广泛关注^[30,111,157]。例如,DeepSeek-R1^[30]采用了两阶段强化学习方法来增强模型的推理能力。在第一阶段,利用基于规则的语言一致性反馈,减轻了模型在推理过程中出现的语言混乱问题;在第二阶段,通过结合基于规则的答案准确性反馈与格式一致性反馈,指导模型在数理逻辑领域的学习,并通过奖励模型的反馈进一步使模型在复杂场景或通用任务中与人类偏好对齐。Song 等人^[111]则进一步将这一训练方法扩展到工具学习中,大幅提升了大模型的外部信息搜索能力。

多途径反馈融合的对齐方法通过结合来自规则、模型和环境的多元反馈信号,有效提升了模型在复杂任务中的表现。这些方法不仅能够精确地对齐人类标准和价值观,还能根据不同任务需求灵活调整模型行为,实现更高效的优化与改进。随着这一领域的深入发展,未来有望通过更精细化的融合反馈机制进一步推动大模型在各个领域与人类专家进行对齐。

3.5 反馈信号对齐面临的挑战:奖励劫持

上述四种不同来源的反馈信号的有效性都面临着一个重大挑战:奖励劫持(Reward Hacking)。在传统强化学习领域中,奖励劫持指智能体可能通过利用奖励函数中的漏洞或模糊性来获取高额奖励,此时智能体并未真正完成预期任务^[158-161]。这通常

发生在奖励函数设计不完善时。这种传统强化学习中的现象在本节所阐述的基于反馈信号的大模型对齐方法中仍然存在^[162-164]。大模型会探索到一些有害的回答,这些信息可能因反馈系统设计的缺陷而意外导致系统对这些回答给予高奖励的反馈^[165-166]。模型可能错误地侧重于利用这些高奖励的有害回答,从而导致对齐失败。

在不同的反馈来源下,奖励劫持的产生原理和表现略有不同。在基于规则的反馈下,若规则过于简单或存在漏洞,模型可能通过非预期方式获取高奖励,而并未真正展现出目标行为^[162-166]。例如,模型可能仅插入符合格式要求的标记符号,而缺乏真实的推理过程或有意义的回答。这种奖励劫持现象与传统强化学习中的奖励劫持产生的原因类似^[161],都是因为规则或奖励系统的设计存在缺陷,而被大语言模型或智能体所利用以获得高的奖励反馈。总的来说,基于规则的反馈往往是因为规则设计缺陷而导致的。

基于模型的反馈更容易受到奖励劫持的危害^[167-169],从偏好数据中建模的奖励模型泛化性较差。随着大模型逐渐被优化对齐,其输出空间的分布会发生变化,旧的奖励模型对这种新分布的打分往往不准确,进而发生奖励劫持现象。同时,基于大模型验证的奖励模型仍然会存在奖励劫持问题。与前者产生的原因不同,这类反馈方法容易受到大模型自身预训练知识的影响,从而导致验证过程中存在偏见。

基于环境的反馈信号同样容易受到奖励劫持的影响^[109-110,170,172-173]。这类反馈信号产生劫持的原因通常是环境设计不充分、测试存在缺陷和漏洞。例如,在让模型对齐人类的代码生成能力时,需要设计沙盒并利用单元测试对模型生成的代码进行验证,并提供反馈信号。如果用来验证代码的单元测试不充分,模型生成的代码可能包含一些未被执行的语句。这些未被单元测试覆盖的语句无法验证其正确性,从而存在被劫持的风险。

面对每一类方法中奖励劫持的成因,研究者们试图通过解决劫持现象来使对齐变得更加稳定有效。基于规则的对齐方法所面临的奖励劫持现象往往是工程问题,需要设计更加细致和准确的规则来监督模型生成的回复,并设计更为合理的奖励函数来提供反馈信号。另一方面,一些工作侧重于解决基于模型反馈所面临的奖励劫持问题,以实现更加稳定的对齐。Coste 等人^[167]通过训练多个角度的

奖励模型并集成打分来缓解奖励劫持现象。ODIN^[168]观察到奖励模型往往存在长度偏见,通过解耦奖励模型中的长度信号和回答信号,显著减少了因长度偏见导致的奖励劫持现象。Zhu 等人^[169]则通过迭代平滑偏好数据来缓解奖励劫持。

同时,一些工作也尝试解决环境反馈中所面临的奖励劫持问题。Dou 等人^[109]让大模型自动生成边界的单元测试样例,从而进一步提高验证的可信性,缓解大模型在代码相关任务对齐上所面临的奖励劫持现象。RedCode^[170]则聚焦于代码安全评估,提供了网站、文件系统、操作系统等 25 类关键漏洞,在相应的 Docker 沙箱环境下对代码代理的执行能力进行细粒度评估,以缓解奖励劫持。

总的来说,奖励劫持现象通常是由于规则设计不充分、奖励函数设计不合理、奖励建模不泛化、环境设计不完善等因素引起的。在未来的研究中,设计反馈信号时需更加谨慎和细致。奖励反馈的泛化性、稳定性和准确性应被优先考虑,这些因素比反馈信号的丰富度更为重要。

3.6 总结与发现

基于规则反馈的对齐方法在可扩展性、训练效率与可控性方面较为优秀。规则反馈方法通过预定义明确的行为规则,可以显著减少对大量人工标注数据的依赖,从而降低训练成本和时间开销。对于安全策略或期望行为频繁更新的场景,它也可以通过调整规则快速更新模型行为。在结构明晰、评价标准明确的任务中(如输出格式或关键词检查),规则对齐还可实现确定性评估,使训练过程中的奖励信号更为稳定可靠,在一定程度上缓解奖励劫持的发生。然而,规则对齐在设计与实际应用中仍面临若干关键挑战,例如对于创意写作、开放性对话、情感回应等复杂语言行为,往往难以制定全面、精确的规则来覆盖所有期望行为。因此,它更适用于结果可验证、评价标准明确的任务。

与基于规则的方法,基于模型的反馈更适合于一些没有标准答案或涉及人类价值观对齐的场景。通过构建这些问题的偏好答案对,模型能够有效地建模出哪些回答是更好的,从而优化模型以产生更加优秀、更加遵循人类价值观的回答。然而,如本节所述,基于模型的反馈也更容易受到奖励劫持的影响。基于偏好数据建模的方法要解决的首要问题是建模方法的泛化性。方法能否在有限的偏好数据上

尽可能全面地建模人类价值和各项能力,是这种方法是否有效的关键。另一方面,基于大模型验证的方法虽然能够提供更泛化的反馈信号,但其准则的设计仍值得未来进一步探索。

相比前两种方法,基于环境的反馈更适合一些可验证的复杂推理任务。例如,通过将代码编译并利用单元测试执行,来验证模型生成的代码的正确性。同时也可以让环境监测模型是否完成了某个推理或规划任务,从而给予反馈信号。该方法同样存在一定的局限性。例如,如果问题难以获得标准答案或难以验证,环境就难以提供相应的反馈。该方法同样可能会受到奖励劫持,且环境的搭建也是一个重难点。例如代码和工具的执行可能需要隔离的沙盒,代码的执行可能需要很多依赖软件包。这些环境验证所必需的适配过程也是值得进一步探索和完善的。

融合上述所有对齐反馈来多途径地指导模型对齐,可能是未来基于反馈对齐的一种主流方式。规则的信号能够对大模型的推理格式、推理结果等信息进行监督;来自模型的信号能够对那些不能获得明确答案的任务进行监督;环境的反馈能够对那些更复杂的推理和规划任务进行监督。这种多反馈融合的对齐方法有望让大模型全面对齐人类的价值和各项能力,从而构建更加安全、可靠、易用的智能大模型。

4 通过思考与沟通实现自主对齐

大模型也可以通过自身的思考或与其他智能体之间的交流来实现与人类对齐。不同于其他的对齐方法,这类方法侧重于将大模型视为一个社会个体,通过让其自主思考或与其他个体进行交流来实现能力的不断进化。这类方法还有望实现可拓展监督这一愿景,即当大模型的能力超过人类评估者的水平时,仍然能够继续为模型提供有效的监督信号。

4.1 基于自我进化的对齐方法

基于自我进化的对齐方法强调大模型通过自我评估、自我批判和自身迭代等手段,不断修正之前产生的错误输出,并强化正确的输出。该对齐方法大体上可以分为自我反思和自我对弈两种方式。表 5 总结了该类对齐方法的主要贡献和局限性。

表 5 基于自我进化的自主对齐方法

序号	方法名称	主要贡献	局限性
1	LMSI ^[46]	提出了 Language Model Self-Improvement (LMSI) 方法, 利用自我生成的高置信度推理答案改进模型	仅适用于推理任务, 在其他任务可能受限
2	Reflexion ^[171]	提出 Reflexion 框架, 通过语言自我反思替代传统梯度更新, 实现语言智能体在多任务中的无微调自我优化与显著性能提升	可能受限任务复杂性, 难以扩展
3	SELF-REFINE ^[172]	提出 SELF-REFINE 方法, 通过自生成的自然语言反馈迭代优化输出, 无需额外训练或数据, 显著提升了大语言模型的生成质量	在生成质量较低的任务上可能表现不佳
4	Re-ReST ^[173]	提出 Re-ReST 方法, 通过引入基于环境反馈的反思机制, 显著提升语言代理自训练样本质量及下游复杂任务表现	需要外部反馈, 受环境限制
5	RLRF ^[174]	提出 RLRF 框架, 引入细粒度多维反馈与自我反思机制, 显著增强大语言模型在复杂任务中的逻辑与事实性表现, 突破 RLHF 的性能瓶颈	依赖精确反馈, 实际应用可能较难实现
6	AlphaLLM ^[175]	提出 AlphaLLM 框架, 将强化学习中的蒙特卡洛树搜索机制深度融入语言模型推理过程, 实现了语言模型在复杂任务上无监督的自我进化	在复杂任务如数学推理仍然存在局限
7	VOYAGER ^[176]	提出 VOYAGER 框架, 通过自动课程、技能库与迭代提示机制, 使语言模型驱动的实体代理首次实现了在开放世界中的自主终身学习和持续探索	需要复杂环境支持, 如 Minecraft
8	STP ^[177]	提出 STP 框架, 通过大型语言模型自我对弈自动生成并验证中间猜想, 显著提高了形式化数学定理自动证明的能力	在非常复杂的定理证明上仍有局限
9	AutoIF ^[178]	提出 AutoIF 方法, 通过自我对弈与执行反馈机制自动生成和验证高质量指令遵循数据, 显著提升语言模型的指令遵循能力	依赖于代码执行环境, 可能在某些场景受限
10	R3V ^[179]	提出 R3V 框架, 引入“推理-反思-修正”机制, 使视觉语言模型能在无需额外训练的情况下自主提升推理准确性	只适用于视觉推理, 缺少其他任务的测试

4.1.1 自我反思

自我反思是指让大模型对自身的输出进行批判性评估并进行有针对性的改进, 是其实现自我进化的核心机制之一。自我反思模拟了人类学习过程中的自我反思能力, 通过识别错误、分析原因并调整策略来持续提升性能。Shinn 等人提出的 Reflexion^[171] 是让大模型通过自我反思实现自主进化的开创性工作。该工作的核心思想是让模型针对自身生成的回答形成自然语言形式的评估和批判, 并将其作为训练集通过监督微调的方式重新注入到大模型中, 使模型能够自我迭代式地与人类的能力对齐。同期的另一个工作 SELF-REFINE^[172] 进一步简化了流程, 让模型在推理阶段直接进行自我思考和改进, 从而在无需额外训练的情况下实现自主对齐。

除了上述最基本的反思框架, 通过自我反思实现自主进化的思想已被研究者们融合进各种各样的任务中。例如, Re-ReST^[173] 通过自我反思持续修正低质量样本, 并将修复后的低质量数据和其他高质量数据用于微调反思模型, 从而实现自我改进。另一方面, RLRF^[174] 将反思机制与强化学习结合, 让大模型进行多维度的评估并筛选自己的回答, 只采

用更好的回答来强化微调自身, 从而实现迭代进化。AlphaLLM^[175] 则将反思机制与 AlphaGo 的搜索策略结合, 通过想象、搜索、自我批判的机制在数学推理等任务上实现了显著的自我改进。更进一步, VOYAGER^[176] 将反思机制应用在智能体任务上, 使智能体能在沙盒游戏 Minecraft 中自主探索和积累技能, 从而模拟人类在复杂环境中实现自我进化。

4.1.2 自我对弈

自我对弈是指通过让模型扮演不同角色或实例来进行结构化交互, 从而实现对模型的多视角改进与自我完善。这一方法能够使模型在无需外部监督的情况下, 通过内部对话、辩论与协作来发现盲点、纠正错误并优化推理过程, 达到自我进化的目的。代表工作包括 STP^[177] 与 AutoIF^[178]。

STP^[177] 将自我对弈应用在形式化数学定理领域, 让同一个模型分别扮演猜想者和证明者两个互补的角色。猜想者负责提出可能有用的引理和中间结果, 而证明者则尝试严格证明这些猜想。这种协作式对弈不断推进证明过程, 从而让模型在这一任务上实现自我进化。另一方面, AutoIF^[178] 巧妙地将自我对弈与执行反馈相结合, 用于自动生成高质

量的指令遵循训练数据。该方法将指令遵循数据质量验证这一问题转化为代码验证问题。具体来说,研究者令模型在不同角色间切换,让模型先生成指令,再转换角色编写相应的检查代码来验证代码正确性并生成单元测试样本。这种多角色对弈机制确保了生成数据的高质量和可检验性,从而不断合成更高质量的数据来提升自己的性能。

自我对弈通过模拟多角色交互或进行内部辩论的方式,使模型可以从不同视角审视问题,发现在单一思维模式下容易忽略的盲点,从而在这一过程中

不断地完善自身能力,实现自我进化。

4.2 基于群体进化的对齐方法

基于群体进化的对齐方法是指多个大模型或智能体通过相互沟通来实现与人类对齐。这类方法巧妙地利用智能体间的对抗、辩论与协作关系,形成了群体间完善的闭环进化系统。根据智能体之间的交互模式,我们可以将该方法分为智能体博弈和智能体协作两种形式。表 6 总结了这类对齐方法的主要贡献和局限性。

表 6 基于群体进化的自主对齐方法

序号	方法名称	主要贡献	局限性
1	TPA ^[50]	将 LLM 对齐问题重构为对抗智能体与防御智能体的两玩家博弈,迭代地发掘难题并改进响应,理论上收敛到纳什均衡;在安全场景下显著提高对抗与防御策略的泛化能力	需要同时训练两个智能体,计算开销增加
2	SP-DPO ^[180]	通过自我对弈生成辩论数据来训练模型,使评估者在听取模型辩论后能做出更准确的判断,证明辩论训练可带来更强的论证能力,从而帮助评估者得出更加正确的结论	辩论框架需要大量计算资源,且在某些领域或任务类型上的效果尚有待验证
3	DwMP ^[181]	探索让两个强模型辩论,由较弱模型或人类判断;辩论帮助非专家显著提高回答准确率;更具说服力的辩论者能让评估者更易识别真相,提出了弱模型评估强模型可行的思路	需要在辩论场景中防范模型学会“操纵”而非“追求真相”,在特定领域的适用性尚待观察
4	CAMEL ^[182]	利用多智能体角色扮演与启发式提示,构建 AI Society 等对话数据集;通过角色明确、分工合作,显著提升模型在复杂任务求解上的表现	多智能体对话生成仍需考虑可能的循环、上下文丢失与角色混淆等问题
5	AI Metropolis ^[183]	利用乱序执行调度大幅提高多智能体模拟效率,通过动态依赖关系跟踪优化并行度;可在大规模智能体场景中加速执行流程,接近理论最优性能	系统较为复杂,在智能体数量较少时,乱序执行开销可能超过性能收益
6	SCCG ^[184]	通过让多个虚拟角色(分析师、开发者、测试者等)协作分工处理复杂编程任务,显著提高 Pass@1 等指标;无需人工干预即可模拟团队协作模式	依赖多代理提示设计,仍可能出现上下文共享不足或任务分解不当的问题
7	ChatDev ^[185]	提出了基于对话链的多智能体框架 ChatDev,通过大语言模型协同分解软件开发任务,有效缓解代码幻觉问题并生成高质量软件	在需求不明确时效果欠佳,并且多智能体协作增加了计算资源消耗和系统复杂性
8	DyLAN ^[186]	提出了 DyLAN 动态多代理网络,通过代理选择和动态通信结构提升代码生成、决策制定、通用推理以及算术推理的表现;并大幅降低不必要的计算开销	需要合理的代理选择与重要性度量方法,且动态通信结构实现相对复杂

4.2.1 智能体博弈

智能体博弈是指通过构建智能体之间的对抗性或竞争性来促进模型对齐。在这一框架下,智能体通常扮演不同角色或持有不同立场,通过相互挑战、辩论或评估来提升彼此的能力。其中比较有代表性的框架是 Zheng 等人^[50]提出的双智能体博弈,将大模型的对齐问题重新构建为对抗性智能体和防御性智能体的博弈。对抗性智能体负责生成能够暴露防御性智能体弱点的提示,而防御性智能体则不断试图解决这些逐渐变难的提示。两个智能体通过相互

对抗自动生成更难、更丰富的提示,实现共同进化。

有些研究者关注通过多智能体辩论实现共同进化的对齐方法。Arnesen 等人^[180]探索了多智能体之间相互辩论来合成对齐数据,发现经过辩论训练的模型能够生成信息量更大的论证过程,显著提高模型输出的质量,同时模型也能够提升对回答进行评估的能力,更容易发现答案中存在的缺点和漏洞。Khan 等人^[181]则探索了较弱的模型能否通过观察专家模型之间的辩论来提升性能。他们让两个大模型为同一问题的不同答案进行辩论,由较弱的非专

家模型作为裁判来选择更可信的答案。实验结果显示,通过观察强模型的辩论过程,非专家模型的准确率从基线的 48% 提升至 76%,而人类在相同问题上的准确率只有 60%。这证明了弱模型能够通过观察辩论过程提升性能,从而在单任务上实现与人类对齐,甚至超越人类。

4.2.2 智能体协作

智能体协作则是群体进化方法的另一种重要形式。与智能体博弈强调的对抗性交互不同,该方式更注重智能体间的互补性和集体决策,通过分工合作、信息共享和集体评估来实现集体智慧的涌现。就像人类社会一样,多智能体协作让每个智能体都能够发挥各自特长和分工。虽然单个模型可能表现出某些“偏科”,但通过群体合作的方式可以在整体目标上与人类对齐。

Li 等人^[182]通过“角色扮演”与“启发式提示”技术构建了一种新型协同智能体框架 CAMEL。这一系统为不同智能体分配明确角色,使它们能够自主合作完成复杂任务,减少对人工提示的依赖。Liu 等人^[186]则提出了一个动态的智能体协作网络 DyLan,基于团队优化和任务求解两个阶段来实现任务导向的动态协作并生成最终答案,以集体力量解决复杂的推理任务。虽然多智能体协作在某些任务上能够比单智能体表现出更好的效果,但是这往往需要消耗更多的时间。Xie 等人^[183]着力解决多智能体协作系统中的效率问题。他们提出了 AI Metropolis,通过乱序执行调度机制和动态依赖关系跟踪,实时监控智能体间的交互过程与信息流动,从而在标准并行模拟的基础上实现了 1.3~4.15 倍的加速效果。

在软件开发领域中,多智能体协作也展现出了显著优势,使智能体在代码开发任务上对齐人类软件开发工程师。Dong 等人^[184]构建了一个无需人工干预的虚拟大模型软工团队,通过角色指令使多个大模型分别扮演分析师、开发者和测试者等角色。相比于单个大模型,这个大模型团队显著提高了代码生成能力。Qian 等人^[185]则提出 ChatDev,用多智能体模拟软件开发全流程。实验结果表明在代码完整性、可执行性、一致性和质量四项指标上,多智能体协作模式全面超越单智能体方法,验证了多智能体协作在提升软件开发质量方面具备巨大潜力。

这些研究共同表明,多智能体协作将人类社会中分工协作、互利共赢的思想应用到多个大模型之间,从而让它们在群体上与人类对齐。这种方式不

需要合成额外的数据,也不需要其他反馈信号指导,就能够让大模型自发地或相互促进地实现对齐。

4.3 总结与发现

与其他对齐方法相比,基于思考和沟通实现的自主对齐能够避免大量的数据标注和定义繁琐的反馈信号,通过充分挖掘大模型本身的潜力来进一步提高大模型在各种任务上的能力。

在基于自我进化的对齐方法中,自我反思能够通过自我评估和自我改进来迭代提升回答质量,而自我对弈则通过多角色交互实现多视角思考。这两种机制相辅相成,使大模型能在几乎无需外部监督的环境中持续对齐。同时,这些方法也有望实现可拓展监督的愿景。

然而,这种自我进化的框架也存在一些挑战。自我提升的过程往往也会放大模型早期的固有缺陷。例如,模型可能因一些局限性而在反思过程中产生偏见,进而持续加强对这种错误偏见的执念。如何让模型沿着既定路线自我进化,或在模型产生偏见时及时引导其走出偏见,也是未来自我进化方法的重要研究方向。

基于群体进化的对齐方法则通过多个智能体之间的交互来提升模型性能和对齐能力,将集体力量大于个人力量的思想应用到大模型对齐框架中。多智能体相互配合,在群体上的对齐能力甚至超越人类能力。多智能体博弈通过构建对抗性环境,能够在一定程度上解决单智能体自我思考所面临的偏见累计和放大挑战,使模型在动态交互中发现并修正违反人类价值观的行为。多智能体协作则通过模拟团队合作,实现集体对齐,在复杂任务上更符合人类期望。这一对齐方法同样面临一些挑战。例如,协作模型间会出现“霸凌”现象,表现为某个强模型在多智能体团队中占据绝对主导地位,致使进化过程呈现出自私性,损害整体协作效果与发展潜力。

我们认为未来该对齐方法的研究工作可以围绕以下几个方面展开:①在这种自主进化的范式中融入人类的价值观来半监督式地指导对齐,即如何将复杂多元的人类价值观融入自我进化和群体交互框架中,确保模型在自主学习的过程中不偏离人类的正确价值观。②构建更为完善的评价体系,例如如何准确评估每个智能体是否都能够多智能体协作框架中表现稳定,同时也需要建立更全面、更个性化的指标来评价每一个子任务下的智能体。③探索这些对齐方法在各类垂直领域中的有效性和适配性。

由于智能体在处理不同任务时分工协作方式存在差异,因此如何明确每个具体任务的分工,以及能否运用统一框架对多个垂类任务进行抽象表述,值得我们进一步探索。

5 大语言模型对齐评测方法

除了探索更加高效稳定的对齐方法之外,如何对模型的对齐效果进行全面评测也同样值得深入研究。通过系统化、多维度的评测,我们不仅可以发现模型中未被对齐的有害信息,也可以找到模型尚未被充分挖掘的潜在能力。本节将大模型对齐评测方法分为两类:一是对奖励模型进行评测,二是对经过对齐的大模型进行评测。前者主要关注奖励模型对人类偏好的判断是否足够准确,以及是否具有泛化性和鲁棒性;而后者则重点关注对齐后模型是否达到了预期目标,生成的回答是否足够安全且符合人类价值观,是否能够与人类的多种能力有效对齐。

5.1 奖励模型评测方法

奖励建模是大模型在价值观和那些无法监督的任务上对齐人类的关键一环,其质量直接影响对齐效果。对奖励模型的评测旨在检验其对人类偏好的判断是否足够准确,以及在多环境下是否具有泛化性和鲁棒性。当前对奖励模型评测较为充分和完善的代表工作包括 RMB^[105]、RM-bench^[187]等。

RMB^[188]涵盖 49 个现实场景,提出了偏好关系评估和 Best-of-N 评估两种评测范式。实验结果证明其评测结果与奖励模型在下游任务上的打分准确性和泛化性呈正相关关系。另一方面,Liu 等人^[187]提出了 RM-bench,侧重于评估奖励模型对内容在细节上差异的敏感性。RewardBench^[189]则提供了一个系统性的奖励模型评估框架,覆盖聊天、复杂对话、推理以及安全等多个领域,支持对不同类型奖励模型进行打分和可视化分析。Liu 等人^[190]则将奖励模型的评测过程视为擂台赛,通过让多个奖励模型两两比较来进行评估。Ethayarajh 等人^[191]提出了 Stanford-SHP 数据集,通过从 Reddit 上收集的包含法律咨询和烹饪技巧在内的 18 个主题领域偏好对,较为全面地评估了奖励模型在生活问答常识方面打分的准确性。

除了这些通用的奖励模型评测基准,一些研究者也重点评测了其在垂直领域中的有效性,为大模型在这些任务上能否与人类有效对齐提供了重要参

考。例如 Nakano 等人^[192]提出的 WebGPT 偏好数据集可用于评测奖励模型在搜索、信息检索和回答问题方面的表现;而 Volske 等人^[193]构建的摘要偏好数据集则可用于评测奖励模型在摘要任务上打分的有效性。这些奖励模型评测基准从不同角度检验奖励模型的质量和性能,为改进大模型的对齐方法、提升对齐表现提供了重要参考。

5.2 对齐后模型评测方法

从对齐目标出发,可以将对齐后模型评测方法分为两类:对安全和价值观的评测、对能力的评测。前者主要评测模型的输出和行为是否足够安全并符合人类的价值观,后者则重点评测对齐后的模型能否与人类能力保持一致,甚至在一定程度上超越人类。

5.2.1 安全与价值观对齐评测方法

随着大模型在各行各业的广泛应用,确保其传递正确的价值观变得尤为重要。避免误导或危害用户的需求逐渐受到重视,对模型的安全性和价值观评估的重要性也日益凸显。其中一些较为全面的评测工作旨在检验对齐后模型能否拒绝有害请求,避免生成误导性内容,并在多元文化背景下体现正确的价值取向等。

Zhang 等人^[194]提出了安全性评测基准 Safety-Bench,包含 11 435 个多选问题,涵盖七类安全问题,为评估对齐后模型的安全性奠定了充分基础。ALERT^[195]则构建了包含 45 000 个陷阱安全问题的基准,通过红队测试的方法识别大模型的漏洞,从而提出改进建议。在价值观对齐方面,Liu 等人^[196]提出了 AlignBench,收集了 683 条真实场景下的与价值观相关的问题,并分为八大类别,重点评测大模型在多维度价值观方面能否始终与人类保持一致,进而做出正确的价值判断。同时,该方法利用 LLM-as-Judge 的方式,借助提示将主观的价值观输出转化为客观的评测指标,从而实现自动化评测。而 Flames^[197]则更专注于中国文化背景下的价值观对齐,涵盖中国特定价值观与通用无害原则等任务,通过对抗性提示来测试模型的安全边界。

针对安全防护机制,Bhardwaj 等人^[198]提出了基于 Chain of Utterances 的红队测试方法 RED-EVAL,发现即使是被广泛使用的模型(如 GPT-4 等)也可能表现出危害人类的价值观,进而存在安全隐患。部分研究则从安全泛化能力的角度切入,例如 Mou 等人^[199]提出的 SG-Bench 重点评估了大模

型的安全泛化能力,发现大多数大模型对问题的格式高度敏感,在对任务的泛化能力上表现较差。这表明模型容易受到基于提示的越狱攻击。

这些安全与价值观对齐评测基准从不同角度评估了对齐后模型的无害性、安全性、真实性、与人类价值观的一致性,为完善现有对齐方法和如何构建负责任的人工智能系统提供了重要参考。同时,我们也需要关注,随着模型能力的不断提升,模型的恶意价值可能会被表面对齐掩盖。一些复杂或精心构造的问题仍然能够让对齐后的大模型暴露出有害信息。这对评测的准确性、有效性和鲁棒性提出了更高的要求。

5.2.2 能力对齐评测方法

大模型在表现出正确的价值观的同时,也应该易用好用,更好地为人类服务。研究者提出了众多基准来全面评测大语言模型在不同能力上与人类对齐的程度。

指令遵循是大模型的核心能力之一,旨在衡量模型理解并准确执行用户指令的能力。指令遵循能力是大模型应用的前提,其评测基准不仅应该能评估模型对于指令的表面理解能力,而且还应该能评估模型是否理解指令中的潜在意图并生成符合要求的回答。

Dubois 等人^[200]构建了 AlpacaEval 自动化评估框架,通过将待测模型与强模型进行对比,来评估指令遵循能力。Zhang 等人^[201]提出了 IOPO,包含 120 万训练数据和 1000 条评估数据,重点评测模型在复杂指令任务上的遵循能力。IFEval^[202]则进一步聚焦于评测那些可以被客观验证的指令任务,例如要求模型生成的回答在 40 字以内等。该基准包括 25 种可验证任务,能够全面评测模型的指令遵循能力。

一些研究工作也细化了评测指标来更全面地对模型的指令遵循能力进行评测。Qin 等人^[203]提出了 INFOBENCH 和分解需求遵循率指标,将复杂指令系统地拆解为基础评估标准,以此对大模型的指令遵循能力进行评测。ComplexBench^[204]则构建了基于 4 个约束类型、19 个约束维度和 4 种组合类型的分层分类评测体系,通过规则增强的评估方法来全面衡量模型遵循复杂指令的能力。这些指令依据评测基准,从不同角度检验模型理解和执行指令的能力,有助于推动研究者继续探索更加有效的指令遵循能力对齐方法。

除了最基本的指令遵循能力,研究者们也提出了多种基准来评测对齐后模型的综合能力。在数学

推理能力上,MATH^[57]和 GSM8K^[56]数据集已经成为主流的评测基准。然而,除了这些对简单数学问题的评测,研究者也越来越关注更深层次的数学推理能力。Zhao 等人^[86]评测了大模型在面对数学陷阱时的有效性。发现即使是最先进的经过对齐的大模型也容易被数学题中精心设计的陷阱所误导,即使是 GPT-4 在这类问题上的准确率也只有 36.0%,远低于人类表现。还有一些研究者收集了来自美国数学邀请赛(AIME)中的问题,专门用于评估模型在极具挑战性的数学推理任务中的解题能力和思维深度。

在代码推理评测领域,研究已从 HumanEval^[205]和 MBPP^[60]等评估基础编程能力的经典基准转向更高层次的计算思维评估,重点关注模型是否真正具备类似人类程序员的计算思维和系统性思想。ML-Bench^[206]是其中的代表性工作,专注于评估模型在机器学习任务上处理仓库级代码的能力,从代码理解和任务执行两个方面评估模型的复杂编程能力。结果表明,即使是当前最先进的模型,在处理该类问题时仍存在困难。

除此之外,研究人员也重点评估了模型的思考能力,即模型是否能像人类一样进行多层次、情境化的思考,能否产生批判性的思维。是否具有这种能力被视为模型有效应对复杂任务,以及在解决问题的过程中敏锐察觉出错误并进行精准纠正的关键要素。代表工作是 Lan 等人^[207]提出的 CriticEval 基准,在五个不同的推理领域通过批判和纠正两个方面来评估模型识别错误的能力,发现采用批判性数据进行训练的模型能够表现出更强的推理和纠错能力。

在财报摘要生成、文献阅读理解等任务里,大模型应当具备处理长文本的能力。一些工作着重对大模型的长文本处理能力展开评测。An 等人^[208]构建了大模型长文能力的标准化评估框架 L-Eval,包含 200 个子任务和 311 篇长文档,涵盖多种评测方法,覆盖了法律、金融、学校讲座、长对话、新闻、长篇小说和会议等多个领域,能够全面反映大模型的长文处理能力。LongBench^[209]则开创性地提供了首个双语多任务评测基准,同样全面地涵盖了多种长文本任务类型,为全面评估模型处理长上下文的能力提供了统一框架。

总而言之,大模型与人类能力的全面对齐需要多维度、系统化的评测体系。这些多元化的评测基准能够客观翔实地衡量大模型与人类能力对齐的程

度,推动了能力对齐方法向更全面、更深入的方向发展。最新的研究甚至提出了“人类最后一次测试”基准(Humanity's Last Exam)^[210],来评估大模型解决人类边界问题的潜力,为判断对齐后的大模型是否能达到甚至超越人类能力提供了实验依据。

5.3 总结与发现

针对大模型对齐的评测是极其重要的研究方向,堪称所有对齐方法的基石。当前的对齐方法不断创新、高速迭代,若评测体系不能同步发展,我们就无法持续评估对齐后模型存在的问题,也就难以推动对齐方法的进一步发展。

在奖励模型的评测方面,仍存在很多方向值得探索。例如,业界缺少针对奖励劫持问题的评测,我们尚未知晓奖励模型在哪些情况下更易遭受劫持,也不清楚奖励模型对于劫持问题的防范能力究竟处于何种水平。对奖励模型的评测在推动大模型进一步与人类全面对齐中扮演着重要的角色。

另一方面,尽管现有的对齐技术能够让模型变得更加安全,也能在一定程度上缓解模型产生恶意行为的问题。但也有研究者指出现有的安全对齐方法仅仅是在表面上进行了对齐,一些复杂或精心设计的问题仍有可能使模型重新暴露出有害信息。因此,我们也应该进一步加大评测的难度,扩大评测的覆盖度,尽可能全面地评测大模型是否真正与人类对齐,是否真正理解了人类正确的价值观,从而确保模型能更好地为人类服务。

我们认为未来的对齐评测可以从以下两个方向展开探索:首先,评测不能仅仅关注模型输出最终结果的正确性,还应该评估模型在生成回答的过程中,其思考路径是否与人类相似。这种基于过程的对齐对于构建真正可信的人工智能系统至关重要。其次,我们需要重新思考如何准确评测模型的泛化能力,区分真正的能力泛化与通过简单的数据覆盖实现表面对齐。单纯依靠数据堆叠进行对齐所获得的性能提升很可能会掩盖模型在对齐过程中真实泛化能力的不足,从而无法确保模型在真实的应用场景中仍然可信。

6 全文总结

随着大语言模型逐渐应用于各行各业,其安全性和可用性也变得尤为重要。在使用过程中,大语言模型需要与人类的价值观和能力对齐,并与人类

的长远利益保持一致。本文总结了大语言模型对齐的基本概念,并通过全新的视角对现有对齐方法进行分类和概述。同时,本文对每类对齐范式进行了系统性的总结归纳,并提出了未来可能改进的方向。最后,本文还总结了大语言模型评测相关技术,这有助于揭示现有对齐方法的不足之处,并为开发更加完善、全面的对齐方法奠定基础。

大语言模型对齐技术发展意义重大且充满挑战。未来在从人类偏好模仿学习方面,我们期望探索提升模型从人类偏好数据学习通用知识与推理能力的方法,增强泛化性能;从反馈信号归纳学习方面,期望研究人员能够重点解决奖励劫持问题,优化各类反馈机制及对齐方法;在通过思考与沟通实现自主对齐方面,我们可以将人类价值观融入其中并探索这类对齐方法在垂直领域中的应用;在对齐评测方面,我们应重视评测过程的真实性,缓解表面对齐现象,保障模型安全有益地服务人类社会。

另一方面,中文大模型在对齐过程中也存在一些独特的挑战。首先,当前缺乏高质量中文数据集,也缺少高质量中文基座模型,在实践中我们很难找到合适且可靠的数据资料来进行对齐。其次,我国幅员辽阔,沿海地区与内陆地区、南方人群与北方人群在价值偏好上不尽相同,如何兼顾这些差异,成为对齐工作的一大难题。再者,我国有众多的少数民族,在进行价值观对齐时需尊重不同民族独特的文化、信仰和价值观,避免对齐后模型因文化上的差异导致输出的内容出现偏差。最后,我国拥有深厚的历史底蕴,价值观会随着时代发展和社会进步不断演变,这就需要我们采用动态的视角进行调整,持续优化模型的价值观对齐。

参考文献

- [1] RESNIK P, LIN J. Evaluation of NLP systems[J]. The Handbook of Computational Linguistics and Natural Language Processing, 2010: 271-295.
- [2] BELLEGARDA J R. Statistical language model adaptation: Review and perspectives[J]. Speech Communication, 2004, 42(1): 93-108.
- [3] JING K, XU J. A survey on neural network language models[J/OL]. arXiv preprint arXiv: 1906.03591, 2019.
- [4] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
- [5] LIU A, FENG B, XUE B, et al. Deepseek-v3 technical report[J/OL]. arXiv preprint arXiv: 2412.19437,

- 2024.
- [6] BI X, CHEN D, CHEN G, et al. Deepseek llm: Scaling open-source language models with longtermism[J/OL]. arXiv preprint arXiv: 2401.02954, 2024.
 - [7] ACHIAM J, ADLER S, AGARWAL S, et al. Gpt-4 technical report[J/OL]. arXiv preprint arXiv: 2303.08774, 2023.
 - [8] YANG A, YANG B, ZHANG B, et al. Qwen2.5 technical report[J/OL]. arXiv preprint arXiv: 2412.15115, 2024.
 - [9] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[R]. OpenAI, 2018.
 - [10] WEI A, HAGHTALAB N, STEINHARDT J. Jailbroken: How does LLM safety training fail? [J]. Advances in Neural Information Processing Systems, 2023, 36: 80079-80110.
 - [11] WEIDINGER L, MELLOR J, RAUH M, et al. Ethical and social risks of harm from language models[J/OL]. arXiv preprint arXiv: 2112.04359, 2021.
 - [12] WEIDINGER L, UESATO J, RAUH M, et al. Taxonomy of risks posed by language models[C]//Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. New York: Association for Computing Machinery, 2022: 214-229.
 - [13] ZHANG Y, LI Y, CUI L, et al. Siren's song in the AI ocean: A survey on hallucination in large language models[J/OL]. arXiv preprint arXiv: 2309.01219, 2023.
 - [14] TONMOY S M, ZAMAN S M, JAIN V, et al. A comprehensive survey of hallucination mitigation techniques in large language models[J/OL]. arXiv preprint arXiv: 2401.01313, 2024.
 - [15] SHEN T, JIN R, HUANG Y, et al. Large language model alignment: A survey[J/OL]. arXiv preprint arXiv: 2309.15025, 2023.
 - [16] GAO Z, LIU X, LAN Y, et al. A brief survey on safety of large language models[J]. Journal of Computing and Information Technology, 2024, 32(1): 47-64.
 - [17] SURDEN H. ChatGPT, AI large language models, and law[J]. Fordham Law Review, 2023, 92: 1941.
 - [18] HUANG J, CHANG K C C. Towards reasoning in large language models: A survey[C]//Proceedings of the Association for Computational Linguistics. Toronto: Association for Computational Linguistics, 2023: 1049-1065.
 - [19] PLAAT A, WONG A, VERBERNE S, et al. Reasoning with large language models, a survey[J/OL]. arXiv preprint arXiv: 2407.11511, 2024.
 - [20] PENG C, YANG X, CHEN A, et al. A study of generative large language model for medical research and healthcare[J]. NPJ Digital Medicine, 2023, 6(1): 210-218.
 - [21] YANG X, CHEN A, POURNEJATIAN N, et al. A large language model for electronic health records[J]. NPJ Digital Medicine, 2022, 5(1): 194-215.
 - [22] WU S, IRSOY O, LU S, et al. Bloomberggpt: A large language model for finance[J/OL]. arXiv preprint arXiv: 2303.17564, 2023.
 - [23] JI J, QIU T, CHEN B, et al. AI alignment: A comprehensive survey [J/OL]. arXiv preprint arXiv: 2310.19852, 2023.
 - [24] CAO B, LU K, LU X, et al. Towards scalable automated alignment of LLMs: A survey[J/OL]. arXiv preprint arXiv: 2406.01252, 2024.
 - [25] WANG Z, BI B, PENTYALA S K, et al. A comprehensive survey of LLM alignment techniques: RLHF, RLAI, PPO, DPO and more[J/OL]. arXiv preprint arXiv: 2407.16216, 2024.
 - [26] WANG Y, ZHONG W, LI L, et al. Aligning large language models with human: A survey[J/OL]. arXiv preprint arXiv: 2307.12966, 2023.
 - [27] FANG X, LI J, MULCHANDANI V, et al. Trustworthy AI on safety, bias, and privacy: A survey[J/OL]. arXiv preprint arXiv: 2502.10450, 2025.
 - [28] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.
 - [29] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: Your language model is secretly a reward model[J]. Advances in Neural Information Processing Systems, 2023, 36: 53728-53741.
 - [30] GUO D, YANG D, ZHANG H, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning [J/OL]. arXiv preprint arXiv: 2501.12948, 2025.
 - [31] ZHANG S, DONG L, LI X, et al. Instruction tuning for large language models: A survey[J/OL]. arXiv preprint arXiv: 2308.10792, 2023.
 - [32] LIU W, ZENG W, HE K, et al. What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning [J/OL]. arXiv preprint arXiv: 2312.15685, 2023.
 - [33] ZHOU C, LIU P, XU P, et al. Lima: Less is more for alignment[J]. Advances in Neural Information Processing Systems, 2023, 36: 55006-55021.
 - [34] XU S, FU W, GAO J, et al. Is DPO superior to PPO for LLM alignment? A comprehensive study[J/OL]. arXiv preprint arXiv: 2404.10719, 2024.
 - [35] AMINI A, VIEIRA T, COTTERELL R. Direct preference optimization with an offset[J/OL]. arXiv preprint arXiv: 2402.10571, 2024.
 - [36] MU T, HELYAR A, HEIDECHE J, et al. Rule based rewards for fine-grained LLM safety[C]//Proceedings of the ICML 2024 Next Generation of AI Safety Workshop, 2024.
 - [37] YANG A X, ROBEYNS M, COSTE T, et al. Bayesian reward models for LLM alignment[C]//Proceedings of the ICML Workshop on Structured Probabilis-

- tic Inference Generative Modeling, 2024.
- [38] MIAO Y, ZHANG S, DING L, et al. Inform: Mitigating reward hacking in RLHF via information-theoretic reward modeling[C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver BC: Current Associates Inc, 2025: 134387-134429.
- [39] WANG B, ZHENG R, CHEN L, et al. Secrets of RLHF in large language models part II: Reward modeling[J/OL]. arXiv preprint arXiv: 2401.06080, 2024.
- [40] GAO L, SCHULMAN J, HILTON J. Scaling laws for reward model overoptimization[C]//Proceedings of the International Conference on Machine Learning. Hawaii: JMLR, 2023: 10835-10866.
- [41] BAI Y, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J/OL]. arXiv preprint arXiv: 2204.05862, 2022.
- [42] SHOJAEI P, JAIN A, TIPIRNENI S, et al. Execution-based code generation using deep reinforcement learning[J]. Transactions on Machine Learning Research, 2023: 1-26.
- [43] CASSANO F, GOUWAR J, NGUYEN D, et al. Multipl-e: A scalable and extensible approach to benchmarking neural code generation[J/OL]. arXiv preprint arXiv: 2208.08227, 2022.
- [44] ENGELBERTH M, GÖBEL J, SCHÖNBEIN C, et al. Pybox-a python sandbox[C]//Proceedings of the SICHERHEIT 2012-Sicherheit, Schutz und Zuverlässigkeit. Gesellschaft für Informatik eV, 2012: 137-148.
- [45] XIE T, GAO Z, REN Q, et al. Logic-rl: Unleashing LLM reasoning with rule-based reinforcement learning[J/OL]. arXiv preprint arXiv: 2502.14768, 2025.
- [46] HUANG J, GU S S, HOU L, et al. Large language models can self-improve [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023: 1051-1068.
- [47] AKSITOV R, MIRYOOSSEFI S, LI Z, et al. ReST meets ReAct: Self-improvement for multi-step reasoning LLM agent[C]//Proceedings of the Workshop on Large Language Model Agents, 2024.
- [48] DING Y, MIN M J, KAISER G, et al. Cycle: Learning to self-refine the code generation[J]. Proceedings of the ACM on Programming Languages, 2024, 8: 392-418.
- [49] JIANG X, DONG Y, WANG L, et al. Self-planning code generation with large language models[J]. ACM Transactions on Software Engineering and Methodology, 2024, 33(7): 1-30.
- [50] ZHENG R, GUO H, LIU Z, et al. Toward optimal LLM alignments using two-player games [J/OL]. arXiv preprint arXiv: 2406.10977, 2024.
- [51] LIU Y, SUN P, LI H. Large language models as agents in two-player games [J/OL]. arXiv preprint arXiv: 2402.08078, 2024.
- [52] CHAN C M, CHEN W, SU Y, et al. ChatEval: Towards better LLM-based evaluators through multi-agent debate[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2024: 1-15.
- [53] JI J, HONG D, ZHANG B, et al. Pku-saferlhf: Towards multi-level safety alignment for LLMs with human preference[J/OL]. arXiv preprint arXiv: 2406.15513, 2024.
- [54] CUI G, YUAN L, DING N, et al. UltraFeedback: Boosting language models with high-quality feedback [J]. CoRR, 2023, abs/2310.01377: 1-23.
- [55] XU G, LIU J, YAN M, et al. Cvalues: Measuring the values of Chinese large language models from safety to responsibility[J/OL]. arXiv preprint arXiv: 2307.09705, 2023.
- [56] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems [J/OL]. arXiv preprint arXiv: 2110.14168, 2021.
- [57] HANDRYCKS D, BURNS C, KADAVATH S, et al. Measuring mathematical problem solving with the MATH dataset[C]//Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
- [58] WANG K, REN H, ZHOU A, et al. MathCoder: Seamless code integration in LLMs for enhanced mathematical reasoning[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2024.
- [59] LOZHKOVA A, LI R, ALLAL L B, et al. Starcoder 2 and the stack v2: The next generation[J/OL]. arXiv preprint arXiv: 2402.19173, 2024.
- [60] AUSTIN J, ODENA A, NYE M, et al. Program synthesis with large language models[J/OL]. arXiv preprint arXiv: 2108.07732, 2021.
- [61] LI C, CHEN H, YAN M, et al. ModelScope-Agent: Building your customizable agent system with open-source large language models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Singapore: Association for Computational Linguistics, 2023: 566-578.
- [62] MIALON G, FOURRIER C, WOLF T, et al. Gaia: A benchmark for general AI assistants[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2023.
- [63] GUO Z, CHENG S, WANG H, et al. StableTool-Bench: Towards stable large-scale benchmarking on tool learning of large language models[C]//Proceedings of the Association for Computational Linguistics. Thailand: Association for Computational Linguistics (ACL), 2024: 11143-11156.
- [64] DONG G, YUAN H, LU K, et al. How abilities in large language models are affected by supervised fine-

- tuning data composition[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok: Association for Computational Linguistics, 2024: 177-198.
- [65] WEI J, BOSMA M, ZHAO V, et al. Finetuned language models are zero-shot learners[C]//Proceedings of the International Conference on Learning Representations, Virtual Event: OpenReview, 2022.
- [66] ZHANG M, HUANG C, WU Y, et al. Transfer-TOD: A generalizable Chinese multi-domain task-oriented dialogue system with transfer capabilities[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami: Association for Computational Linguistics, 2024: 12750-12771.
- [67] ZHANG M, WANG Y, SHEN Y, et al. PFDial: A structured dialogue instruction fine-tuning method based on UML flowcharts [J/OL]. arXiv preprint arXiv: 2503.06706, 2025.
- [68] CONOVER M, HAYES M, MATHUR A, et al. Free dolly: Introducing the world's first truly open instruction-tuned LLM[EB/OL]. 2023.
- [69] MCDONALD D, PAPADOPOULOS R, BENNING-FIELD L. Reducing LLM hallucination using knowledge distillation: A case study with Mistral Large and MMLU benchmark[C]//Proceedings of the Authorea Preprints, 2024.
- [70] XU X, LI M, TAO C, et al. A survey on knowledge distillation of large language models [J/OL]. arXiv preprint arXiv: 2402.13116, 2024.
- [71] YANG C, ZHU Y, LU W, et al. Survey on knowledge distillation for large language models: Methods, evaluation, and application [J]. CoRR, 2024, abs/2407.01885: 1-27.
- [72] LUO W. A comprehensive survey on knowledge distillation of diffusion models [J/OL]. arXiv preprint arXiv: 2304.04262, 2023.
- [73] ZHANG Z, WANG Y, WANG C, et al. LLM hallucinations in practical code generation: Phenomena, mechanism, and mitigation [J/OL]. arXiv preprint arXiv: 2409.20550, 2024.
- [74] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model [J/OL]. arXiv preprint arXiv: 2405.04434, 2024.
- [75] LEE Y S, SULTAN M, ELKURDI Y, et al. Ensemble-instruct: Instruction tuning data generation with a heterogeneous mixture of LMs [C]//Proceedings of the Association for Computational Linguistics: EMNLP, Singapore: Association for Computational Linguistics, 2023: 12561-12571.
- [76] WANG Y, KORDI Y, MISHRA S, et al. Self-Instruct: Aligning language models with self-generated instructions [C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto: Association for Computational Linguistics, 2023: 13484-13508.
- [77] XU C, SUN Q, ZHENG K, et al. Wizardlm: Empowering large language models to follow complex instructions [J/OL]. arXiv preprint arXiv: 2304.12244, 2023.
- [78] SHIRGAONKAR A, PANDEY N, ABAY N C, et al. Knowledge distillation using frontier open-source LLMs: Generalizability and the role of synthetic data [J/OL]. arXiv preprint arXiv: 2410.18588, 2024.
- [79] YANG R, WU T, WANG J, et al. Llm-neo: Parameter efficient knowledge distillation for large language models [J/OL]. arXiv preprint arXiv: 2411.06839, 2024.
- [80] JIANG Y, CHAN C, CHEN M, et al. Lion: Adversarial distillation of proprietary large language models [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Singapore: Association for Computational Linguistics, 2023: 3134-3154.
- [81] YEO E, TONG Y, NIU M, et al. Demystifying long chain-of-thought reasoning in LLMs [J/OL]. arXiv preprint arXiv: 2502.03373, 2025.
- [82] WU S, PENG Z, DU X, et al. A comparative study on reasoning patterns of OpenAI's o1 model [J/OL]. arXiv preprint arXiv: 2410.13639, 2024.
- [83] DE WINTER J C F, DODOU D, EISMA Y B. System 2 thinking in OpenAI's o1-preview model: Near-perfect performance on a mathematics exam [J]. Computers, 2024, 13(11): 278-285.
- [84] HUANG Z, ZOU H, LI X, et al. O1 replication journey—part 2: Surpassing O1-preview through simple distillation, big progress or bitter lesson? [J/OL]. arXiv preprint arXiv: 2411.16489, 2024.
- [85] LUO Y, YANG Z, MENG F, et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning [J/OL]. arXiv preprint arXiv: 2308.08747, 2023.
- [86] ZHAO J, TONG J, MOU Y, et al. Exploring the compositional deficiency of large language models in mathematical reasoning through trap problems [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami: Association for Computational Linguistics, 2024: 16361-16376.
- [87] AZAR M G, GUO Z D, PIOT B, et al. A general theoretical paradigm to understand learning from human preferences [C]//Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia: PMLR, 2024: 4447-4455.
- [88] ETHAYARAJH K, XU W, MUENNIGHOFF N, et al. KTO: model alignment as prospect theoretic optimization [J/OL]. arXiv preprint arXiv: 2402.01306, 2024.
- [89] GUO Y, CUI G, YUAN L, et al. Controllable preference optimization: toward controllable multi-objective alignment [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Miami: Association for Computational Linguistics,

- 2024; 1437-1454.
- [90] HONG J, LEE N, THORNE J. ORPO: Monolithic preference optimization without reference model [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami: Association for Computational Linguistics, 2024; 11170-11189.
- [91] BRADLEY R A, TERRY M E. Rank analysis of incomplete block designs; I. The method of paired comparisons[J]. *Biometrika*, 1952, 39 (3/4): 324-345.
- [92] MU T, HELYAR A, HEIDECHE J, et al. Rule based rewards for language model safety [C]//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver: OpenReview, 2024.
- [93] LI X, GAO M, ZHANG Z, et al. Rule-based rating and selection of LLM training data [C]//Proceedings of the ICLR Workshop on Navigating and Addressing Data Problems for Foundation Models, 2025.
- [94] LIU T, XIONG W, REN J, et al. RRM: Robust reward model training mitigates reward hacking [J/OL]. arXiv preprint arXiv: 2409.13156, 2024.
- [95] DOU S, LIU Y, ZHOU E, et al. MetaRM: Shifted distributions alignment via meta-learning [J/OL]. arXiv preprint arXiv: 2405.00438, 2024.
- [96] WU D, CHOUDHURY S. Aligning LLMs with domain invariant reward models [J/OL]. arXiv preprint arXiv: 2501.00911, 2025.
- [97] LOU X, YAN D, SHEN W, et al. Uncertainty-aware reward model: Teaching reward models to know what is unknown [J/OL]. arXiv preprint arXiv: 2410.00847, 2024.
- [98] LAI Y, WANG S, LIU S, et al. ALaRM: Align language models via hierarchical rewards modeling [J/OL]. arXiv preprint arXiv: 2403.06754, 2024.
- [99] LE H, WANG Y, GOTMARE A D, et al. Coderl: Mastering code generation through pretrained models and deep reinforcement learning [J]. *Advances in Neural Information Processing Systems*, 2022, 35: 21314-21328.
- [100] LIU J, ZHU Y, XIAO K, et al. RLTF: Reinforcement learning from unit test feedback [J]. *Transactions on Machine Learning Research*, 2023; 1-15.
- [101] DOU S, LIU Y, JIA H, et al. StepCoder: Improving code generation with reinforcement learning from compiler feedback [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: Association for Computational Linguistics, 2024; 4571-4585.
- [102] ZHOU Y, ZANETTE A, PAN J, et al. ArCHer: Training language model agents via hierarchical multi-turn RL [C]//Proceedings of the International Conference on Machine Learning. Vienna: OpenReview, 2024.
- [103] YU Y, WANG Z, MA W, et al. StepTool: A step-grained reinforcement learning framework for tool learning in LLMs [J/OL]. arXiv preprint arXiv: 2410.07745, 2024.
- [104] ZENG Y, DING X, WANG Y, et al. Boosting tool use of large language models via iterative reinforced fine-tuning [J/OL]. arXiv preprint arXiv: 2501.09766, 2025.
- [105] XI Z, DING Y, CHEN W, et al. Agentgym: Evolving large language model-based agents across diverse environments [J/OL]. arXiv preprint arXiv: 2406.04151, 2024.
- [106] ZHANG W, LU Z. Adarefiner: Refining decisions of language models with adaptive feedback [J/OL]. arXiv preprint arXiv: 2309.17176, 2023.
- [107] LI L, CHAI Y, WANG S, et al. Tool-augmented reward modeling [C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2024.
- [108] LIN Y T, JIN D, XU T, et al. Step-KTO: Optimizing mathematical reasoning through stepwise binary feedback [J/OL]. arXiv preprint arXiv: 2501.10799, 2025.
- [109] DOU S, ZHANG J, ZANG J, et al. Multi-programming language sandbox for LLMs [J/OL]. arXiv preprint arXiv: 2410.23074, 2024.
- [110] LIU S, ZHU H, LIU J, et al. Fullstack bench: Evaluating LLMs as full stack coder [J/OL]. arXiv preprint arXiv: 2412.00535, 2024.
- [111] SONG H, JIANG J, MIN Y, et al. R1-Searcher: Incentivizing the search capability in LLMs via reinforcement learning [J/OL]. arXiv preprint arXiv: 2503.05592, 2025.
- [112] LI X, GAO M, ZHANG Z, et al. Rule-based data selection for large language models [J/OL]. arXiv preprint arXiv: 2410.04715, 2024.
- [113] WANG B, ZHENG R, CHEN L, et al. Reward modeling requires automatic adjustment based on data quality [C]//Proceedings of the Association for Computational Linguistics: EMNLP. Miami: Association for Computational Linguistics, 2024; 4041-4064.
- [114] GAO Y, ALON D, METZLER D. Impact of preference noise on the alignment performance of generative language models [C]//Proceedings of 1st Conference on Language Modeling. Philadelphia: OpenReview, 2024.
- [115] CHEN L, ZHENG R, WANG B, et al. Improving discriminative capability of reward models in RLHF using contrastive learning [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Miami: Association for Computational Linguistics, 2024; 15270-15283.
- [116] DONG H, XIONG W, PANG B, et al. RLHF workflow: From reward modeling to online RLHF [J]. *Transactions on Machine Learning Research*, 2024; 1-29.

- [117] HARE J. Dealing with sparse rewards in reinforcement learning[J/OL]. arXiv preprint arXiv: 1910.09281, 2019.
- [118] MOHTASIB A, NEUMANN G, CUAYÁHUITL H. A study on dense and sparse (visual) rewards in robot policy learning[C]//Proceedings of the Towards Autonomous Robotic Systems: 22nd Annual Conference. Lincoln: Springer International Publishing, 2021: 3-13.
- [119] XI Z, CHEN W, HONG B, et al. Training large language models for reasoning through reverse curriculum reinforcement learning[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: OpenReview, 2024.
- [120] LIGHTMAN H, KOSARAJU V, BURDA Y, et al. Let's verify step by step[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview (OR), 2023.
- [121] ZELIKMAN E, WU Y, MU J, et al. Star: Bootstrapping reasoning with reasoning[J]. Advances in Neural Information Processing Systems, 2022, 35: 15476-15488.
- [122] ZHANG D, ZHOUBIAN S, HU Z, et al. Restmcts *: LLM self-training via process reward guided tree search[J]. Advances in Neural Information Processing Systems, 2024, 37: 64735-64772.
- [123] WANG W, GAO Z, CHEN L, et al. Visualprm: An effective process reward model for multimodal reasoning [J/OL]. arXiv preprint arXiv: 2503.10291, 2025.
- [124] CHEN W, HE W, XI Z, et al. Better process supervision with bi-directional rewarding signals [J/OL]. arXiv preprint arXiv: 2503.04618, 2025.
- [125] WANG P, LI L, SHAO Z, et al. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: Association for Computational Linguistics, 2024: 9426-9439.
- [126] GU J, JIANG X, SHI Z, et al. A survey on LLM-as-a-judge [J/OL]. arXiv preprint arXiv: 2411.15594, 2024.
- [127] WANG T, KULIKOV I, GOLOVNEVA O, et al. Self-taught evaluators[J/OL]. arXiv preprint arXiv: 2408.02666, 2024.
- [128] YUAN H, YUAN Z, TAN C, et al. RRHF: Rank responses to align language models with human feedback[J]. Advances in Neural Information Processing Systems, 2023, 36: 10935-10950.
- [129] ZHUGE M, ZHAO C, ASHLEY D, et al. Agent-as-a-judge: Evaluate agents with agents [J/OL]. arXiv preprint arXiv: 2410.10934, 2024.
- [130] LI J, SUN S, YUAN W, et al. Generative judge for evaluating alignment[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2024.
- [131] ZHANG L, HOSSEINI A, BANSAL H, et al. Generative verifiers: Reward modeling as next-token prediction[C]//Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview, 2025.
- [132] MAHAN D, VAN PHUNG D, RAFAILOV R, et al. Generative reward models[J/OL]. arXiv preprint arXiv: 2410.12832, 2024.
- [133] HU T, KYRYCHENKO Y, RATHJE S, et al. Generative language models exhibit social identity biases[J]. Nature Computational Science, 2025, 5 (1): 65-75.
- [134] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional AI: Harmlessness from AI feedback [J/OL]. arXiv preprint arXiv: 2212.08073, 2022.
- [135] GUAN M Y, JOGLEKAR M, WALLACE E, et al. Deliberative alignment: Reasoning enables safer language models[J/OL]. arXiv preprint arXiv: 2412.16339, 2024.
- [136] LEE C, XIA C S, YANG L, et al. A unified debugging approach via LLM-based multi-agent synergy [J/OL]. arXiv preprint arXiv: 2404.17153, 2024.
- [137] NAM D, MACVEAN A, HELLENDORF V, et al. Using an LLM to help with code understanding [C]//Proceedings of the IEEE/ACM 46th International Conference on Software Engineering. Lisbon: Association for Computing Machinery, 2024: 1-13.
- [138] JIANG J, WANG F, SHEN J, et al. A survey on large language models for code generation[J/OL]. arXiv preprint arXiv: 2406.00515, 2024.
- [139] DOU S, JIA H, WU S, et al. What's wrong with your code generated by large language models? An extensive study[J/OL]. arXiv preprint arXiv: 2407.06153, 2024.
- [140] SHEN Z. LLM with tools: A survey[J/OL]. arXiv preprint arXiv: 2409.18807, 2024.
- [141] LI X. A review of prominent paradigms for LLM-based agents: Tool use (including RAG), planning, and feedback learning[J/OL]. arXiv preprint arXiv: 2406.05804, 2024.
- [142] KANNAN S S, VENKATESH V L N, MIN B C. Smart-LLM: Smart multi-agent robot task planning using large language models[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Abu Dhabi: IEEE, 2024: 12140-12147.
- [143] HUANG X, LIU W, CHEN X, et al. Understanding the planning of LLM agents: A survey[J/OL]. arXiv preprint arXiv: 2402.02716, 2024.
- [144] DAGAN G, KELLER F, LASCARIDES A. Dynamic planning with a LLM[C]//Proceedings of the Language Gamification-NeurIPS Workshop, 2024.
- [145] DOLCETTI G, ARCERI V, IOTTI E, et al. Helping LLMs improve code generation using feedback from testing and static analysis[J/OL]. arXiv pre-

- print arXiv: 2412.14841, 2024.
- [146] YE J, WU Y, LI S, et al. TL-Training: A task-feature-based framework for training large language models in tool use[J/OL]. arXiv preprint arXiv: 2412.15495, 2024.
 - [147] QU C, DAI S, WEI X, et al. From exploration to mastery: Enabling LLMs to master tools via self-driven interactions[C]//Proceedings of the 13th International Conference on Learning Representations. Singapore: OpenReview, 2025.
 - [148] ZHUANG Y, CHEN X, YU T, et al. ToolChain *: Efficient action space navigation in large language models with A * search[C]//Proceedings of the 12th International Conference on Learning Representations. Vienna: OpenReview, 2024.
 - [149] WEBB T, MONDAL S S, MOMENNEJAD I. Improving planning with large language models: A modular agentic architecture[J/OL]. arXiv preprint arXiv: 2310.00194, 2023.
 - [150] BANDYOPADHYAY D, BHATTACHARJEE S, EKBAL A. Thinking machines: A survey of LLM based reasoning strategies[J/OL]. arXiv preprint arXiv: 2503.10814, 2025.
 - [151] TALEBI RAD Y, NADIRI A. Multi-agent collaboration: Harnessing the power of intelligent LLM agents[J/OL]. arXiv preprint arXiv: 2306.03314, 2023.
 - [152] LIN M Y, LEE O W, LU C Y. Embodied AI with large language models: A survey and new HRI framework[C]//Proceedings of the International Conference on Advanced Robotics and Mechatronics. Tokyo: IEEE, 2024: 978-983.
 - [153] CHRISTIANO P F, LEIKE J, BROWN T, et al. Deep reinforcement learning from human preferences[J]. Advances in Neural Information Processing Systems, 2017, 30: 4299-4307.
 - [154] IBARZ B, LEIKE J, POHLEN T, et al. Reward learning from human preferences and demonstrations in Atari[J]. Advances in Neural Information Processing Systems, 2018, 31: 8022-8034.
 - [155] LIU F. Learning to summarize from human feedback[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 583-592.
 - [156] JAECH A, KALAI A, LERER A, et al. OpenAI o1 system card[J/OL]. arXiv preprint arXiv: 2412.16720, 2024.
 - [157] JIN B, ZENG H, YUE Z, et al. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning[J/OL]. arXiv preprint arXiv: 2503.09516, 2025.
 - [158] HADFIELD-MENELL D, MILLI S, ABBEEL P, et al. Inverse reward design[J]. Advances in Neural Information Processing Systems, 2017, 30: 6765-6774.
 - [159] LEIKE J, KRUEGER D, EVERITT T, et al. Scalable agent alignment via reward modeling: A research direction[J/OL]. arXiv preprint arXiv: 1811.07871, 2018.
 - [160] YUAN Y, YU Z L, GU Z, et al. A novel multi-step reinforcement learning method for solving reward hacking[J]. Applied Intelligence, 2019, 49: 2874-2888.
 - [161] SKALSE J, HOWE N, KRASHENINNIKOV D, et al. Defining and characterizing reward gaming[J]. Advances in Neural Information Processing Systems, 2022, 35: 9460-9471.
 - [162] PAN A, BHATIA K, STEINHARDT J. The effects of reward misspecification: Mapping and mitigating misaligned models[C]//Proceedings of the International Conference on Learning Representations. Virtual Event: OpenReview, 2022.
 - [163] LENG J, HUANG C, ZHU B, et al. Taming overconfidence in LLMs: Reward calibration in RLHF[J/OL]. arXiv preprint arXiv: 2410.09724, 2024.
 - [164] WANG C, ZHAO Z, JIANG Y, et al. Beyond reward hacking: Causal rewards for large language model alignment[J/OL]. arXiv preprint arXiv: 2501.09620, 2025.
 - [165] MIAO Y, ZHANG S, DING L, et al. The energy loss phenomenon in RLHF: A new perspective on mitigating reward hacking[J/OL]. arXiv preprint arXiv: 2501.19358, 2025.
 - [166] FU J, ZHAO X, YAO C, et al. Reward shaping to mitigate reward hacking in RLHF[J/OL]. arXiv preprint arXiv: 2502.18770, 2025.
 - [167] COSTE T, ANWAR U, KIRK R, et al. Reward model ensembles help mitigate overoptimization[C]//Proceedings of the NeurIPS Workshop on Instruction Tuning and Instruction Following, 2023.
 - [168] CHEN L, ZHU C, CHEN J, et al. ODIN: Disentangled reward mitigates hacking in RLHF[C]//Proceedings of the International Conference on Machine Learning. Vienna: OpenReview, 2024.
 - [169] ZHU B, JORDAN M, JIAO J. Iterative data smoothing: Mitigating reward overfitting and overoptimization in RLHF[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: OpenReview, 2024.
 - [170] GUO C, LIU X, XIE C, et al. Redcode: Risky code execution and generation benchmark for code agents[J]. Advances in Neural Information Processing Systems, 2024, 37: 106190-106236.
 - [171] SHINN N, CASSANO F, GOPINATH A, et al. Reflexion: Language agents with verbal reinforcement learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 8634-8652.
 - [172] MADAAN A, TANDON N, GUPTA P, et al. Self-refine: Iterative refinement with self-feedback[J]. Advances in Neural Information Processing Systems, 2023, 36: 46534-46594.

- [173] DOU Z Y, YANG C F, WU X, et al. Re-rest: Reflection-reinforced self-training for language agents [J/OL]. arXiv preprint arXiv: 2406.01495, 2024.
- [174] LEE K, HWANG D, PARK S, et al. Reinforcement learning from reflective feedback (RLRF): Aligning and improving LLMs via fine-grained self-reflection[J/OL]. arXiv preprint arXiv: 2403.14238, 2024.
- [175] TIAN Y, PENG B, SONG L, et al. Toward self-improvement of LLMs via imagination, searching, and criticizing[J]. *Advances in Neural Information Processing Systems*, 2024, 37: 52723-52748.
- [176] WANG G, XIE Y, JIANG Y, et al. Voyager: An open-ended embodied agent with large language models[C]//*Proceedings of the NeurIPS Foundation Models for Decision Making Workshop*, 2023.
- [177] DONG K, MA T. Beyond limited data: Self-play LLM theorem provers with iterative conjecturing and proving [J/OL]. arXiv preprint arXiv: 2502.00212, 2025.
- [178] DONG G, LU K, LI C, et al. Self-play with execution feedback: Improving instruction-following capabilities of large language models[C]//*Proceedings of the 13th International Conference on Learning Representations*. Singapore: OpenReview.net, 2025.
- [179] CHENG K, LI Y, XU F, et al. Vision-language models can self-improve reasoning via reflection[J/OL]. arXiv preprint arXiv: 2411.00855, 2024.
- [180] ARNESEN S, REIN D, MICHAEL J. Training language models to win debates with self-play improves judge accuracy[J/OL]. arXiv preprint arXiv: 2409.16636, 2024.
- [181] KHAN A, HUGHES J, VALENTINE D, et al. Debating with more persuasive LLMs leads to more truthful answers[C]//*Proceedings of the 41st International Conference on Machine Learning*. Vienna: OpenReview, 2024.
- [182] LI G, HAMMOUD H, ITANI H, et al. Camel: Communicative agents for "mind" exploration of large language model society[J]. *Advances in Neural Information Processing Systems*, 2023, 36: 51991-52008.
- [183] XIE Z, KANG H, SHENG Y, et al. AI Metropolis: Scaling large language model-based multi-agent simulation with out-of-order execution[J/OL]. arXiv preprint arXiv: 2411.03519, 2024.
- [184] DONG Y, JIANG X, JIN Z, et al. Self-collaboration code generation via ChatGPT[J]. *ACM Transactions on Software Engineering and Methodology*, 2024, 33(7): 1-38.
- [185] QIAN C, LIU W, LIU H, et al. ChatDev: Communicative agents for software development[C]//*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok: Association for Computational Linguistics, 2024: 15174-15186.
- [186] LIU Z, ZHANG Y, LI P, et al. A dynamic LLM-powered agent network for task-oriented agent collaboration[C]//*Proceedings of the 1st Conference on Language Modeling*. Vienna: OpenReview (OR), 2024.
- [187] LIU Y, YAO Z, MIN R, et al. RM-Bench: Benchmarking reward models of language models with subtlety and style[C]//*Proceedings of the 13th International Conference on Learning Representations*. Singapore: OpenReview, 2025.
- [188] ZHOU E, ZHENG G, WANG B, et al. RMB: Comprehensively benchmarking reward models in LLM alignment[J/OL]. arXiv preprint arXiv: 2410.09893, 2024.
- [189] LAMBERT N, PYATKIN V, MORRISON J, et al. Rewardbench: Evaluating reward models for language modeling[J/OL]. arXiv preprint arXiv: 2403.13787, 2024.
- [190] LIU Y, YAO Z, MIN R, et al. Pairwise RM: Perform best-of-N sampling with knockout tournament [J/OL]. arXiv preprint arXiv: 2501.13007, 2025.
- [191] ETHAYARAJH K, CHOI Y, SWAYAMDIPTA S. Understanding dataset difficulty with V-usable information [C]//*Proceedings of the International Conference on Machine Learning*. Baltimore: PMLR, 2022: 5988-6008.
- [192] NAKANO R, HILTON J, BALAJI S, et al. WebGPT: Browser-assisted question-answering with human feedback[J/OL]. arXiv preprint arXiv: 2112.09332, 2021.
- [193] VÖLSKE M, POTTHAST M, SYED S, et al. TL; DR: Mining Reddit to learn automatic summarization [C]//*Proceedings of the Workshop on New Frontiers in Summarization*. Copenhagen: Association for Computational Linguistics, 2017: 59-63.
- [194] ZHANG Z, LEI L, WU L, et al. SafetyBench: Evaluating the safety of large language models[C]//*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok: Association for Computational Linguistics, 2024: 15537-15553.
- [195] TEDESCHI S, FRIEDRICH F, SCHRAMOWSKI P, et al. ALERT: A comprehensive benchmark for assessing large language models' safety through red teaming[J/OL]. arXiv preprint arXiv: 2404.08676, 2024.
- [196] LIU X, LEI X, WANG S, et al. AlignBench: Benchmarking Chinese alignment of large language models[C]//*Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. Bangkok: Association for Computational Linguistics, 2024: 11621-11640.
- [197] HUANG K, LIU X, GUO Q, et al. Flames: Benchmarking value alignment of LLMs in Chinese [C]//*Proceedings of the Conference of the North American Chapter of the Association for Computa-*

- tional Linguistics; Human Language Technologies. Bangkok; Association for Computational Linguistics, 2024; 4551-4591.
- [198] BHARDWAJ R, PORIA S. Red-teaming large language models using chain of utterances for safety-alignment[J/OL]. arXiv preprint arXiv: 2308.09662, 2023.
- [199] MOU Y, ZHANG S, YE W. SG-Bench: Evaluating LLM safety generalization across diverse tasks and prompt types[J]. Advances in Neural Information Processing Systems, 2024, 37: 123032-123054.
- [200] DUBOIS Y, GALAMBOSI B, LIANG P, et al. Length-controlled AlpacaEval: A simple way to debias automatic evaluators[J/OL]. arXiv preprint arXiv: 2404.04475, 2024.
- [201] ZHANG X, YU H, FU C, et al. IOPO: Empowering LLMs with complex instruction following via input-output preference optimization[J/OL]. arXiv preprint arXiv: 2411.06208, 2024.
- [202] ZHOU J, LU T, MISHRA S, et al. Instruction-following evaluation for large language models[J/OL]. arXiv preprint arXiv: 2311.07911, 2023.
- [203] QIN Y, SONG K, HU Y, et al. InFoBench: Evaluating instruction following ability in large language models[C]//Proceedings of the Association for Computational Linguistics ACL. Bangkok; Association for Computational Linguistics, 2024; 13025-13048.
- [204] WEN B, KE P, GU X, et al. Benchmarking complex instruction-following with multiple constraints composition[J]. Advances in Neural Information Processing Systems, 2024, 37: 137610-137645.
- [205] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[J/OL]. arXiv preprint arXiv: 2107.03374, 2021.
- [206] TANG X, LIU Y, CAI Z, et al. ML-Bench: Evaluating large language models and agents for machine learning tasks on repository-level code[J/OL]. arXiv preprint arXiv: 2311.09835, 2023.
- [207] LAN T, ZHANG W, XU C, et al. CriticEval: Evaluating large-scale language model as critic[J]. Advances in Neural Information Processing Systems, 2024, 37: 66907-66960.
- [208] AN C, GONG S, ZHONG M, et al. L-Eval: Instituting standardized evaluation for long context language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok; Association for Computational Linguistics, 2024; 14388-14411.
- [209] BAI Y, LV X, ZHANG J, et al. LongBench: A bilingual, multitask benchmark for long context understanding[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok; Association for Computational Linguistics, 2024; 3119-3137.
- [210] PHAN L, GATTI A, HAN Z, et al. Humanity's last exam [J/OL]. arXiv preprint arXiv: 2501.14249, 2025.
- [211] ISHIBUCHI H, NAKASHIMA T. Effect of rule weights in fuzzy rule-based classification systems [J]. IEEE Transactions on Fuzzy Systems, 2001, 9 (4): 506-515.
- [212] SONG Y, BIAN Y, TANG Y, et al. Visiontasker: Mobile task automation using vision based UI understanding and LLM task planning[C]//Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology. Pittsburgh; Association for Computing Machinery, 2024; 1-17.
- [213] GUO T, CHEN X, WANG Y, et al. Large language model based multi-agents: A survey of progress and challenges[C]//Proceedings of the 33rd International Joint Conference on Artificial Intelligence. Jeju: ijcai.org, 2024; 8048-8057.
- [214] XI Z, CHEN W, GUO X, et al. The rise and potential of large language model based agents: A survey [J]. Science China Information Sciences, 2025, 68 (2): 121101.
- [215] YANG W, WANG H, LIU Z, et al. Enhancing the code debugging ability of LLMs via communicative agent based data refinement[J/OL]. arXiv preprint arXiv: 2408.05006, 2024.



窦士涵(1999—),共同第一作者,博士研究生,主要研究领域为大语言模型对齐的相关研究。

E-mail: shdou24@m.fudan.edu.cn



黄莹菁(1972—),通信作者,博士,教授,主要研究领域为大语言模型和具身智能的相关研究。

E-mail: xjhuang@fudan.edu.cn



张明(1999—),共同第一作者,博士研究生,主要研究领域为大语言模型评测和对话系统的相关研究。

E-mail: mingzhang23@m.fudan.edu.cn