大语言模型对齐研究综述

刘昆麟，屈新纪，谭芳，康红辉，赵少伟，施嵘
（中兴通讯股份有限公司，广东 深圳 518057）

摘　要：随着人工智能技术的飞速发展，大语言模型已在众多领域得到了广泛应用。然而，大语言模型可能会生成不准确、有误导性甚至有害的内容，这引发了人们对大语言模型可靠性的担忧，采用对齐技术来确保大语言模型的行为与人类价值观一致已经成为一个亟待解决的问题。对近年来大语言模型对齐技术的研究进展进行综述。介绍了常用的指令数据收集方法和人类偏好数据集，概述了监督调整和对齐调整的相关研究，讨论了模型评估常用的数据集和方法，总结并展望了未来的研究方向。

关键词：大语言模型；对齐技术；调整；强化学习
中图分类号：TN92
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024151

Survey on large language models alignment research

LIU Kunlin, QU Xinji, TAN Fang, KANG Honghui, ZHAO Shaowei, SHI Rong
ZTE Corporation, Shenzhen 518057, China

Abstract: With the rapid development of artificial intelligence technology, large language models have been widely applied in numerous fields. However, the potential of large language models to generate inaccurate, misleading, or even harmful contents has raised concerns about their reliability. Adopting alignment techniques to ensure the behavior of large language models is consistent with human values has become an urgent issue to address. Recent research progress on alignment techniques for large language models were surveyed. Common methods for collecting instruction data and human preference datasets were introduced, research on supervised tuning and alignment adjustments was summarized, commonly used datasets and methods for model evaluation were discussed, and future research directions were concluded.

Key words: large language model, alignment technique, tune, reinforcement learning

0　引言

近年来，大语言模型（如 OpenAI 的 ChatGPT）[1]的迅猛发展引发了人们对人工智能的浓厚兴趣和高度期望，同时也引发了人们的广泛探讨。大语言模型不仅展现出卓越的自然语言处理能力，还在数学、推理和编程等多个领域中接近甚至超越普通人类的水平[2]。这些成就主要得益

收稿日期：2024−03−28；修回日期：2024−05−18

于大语言模型在超大规模的文本语料库上的预训练，这使它们积累了海量的世界知识，并能基于这些知识生成连贯和流畅的文本输出。尽管大语言模型已在众多领域得到了广泛应用，但它们在生成内容时仍可能存在不准确、有误导性甚至包含有害信息的风险，这引发了人们对大语言模型可靠性的担忧。

当前，研究人员正在积极探索如何确保大语言模型的行为与人类价值观一致。对齐是指通过调整和优化大语言模型的决策过程，以确保其输出不仅准确无误，而且遵循道德规范、没有偏见，并且能反映出社会普遍认可的价值观和伦理标准。对齐的目的在于创建一个既能理解和生成人类语言的模型，又能在其决策中体现出对公平、透明和责任的重视，减少可能产生的负面影响，如传播虚假信息或有害内容。然而在对大语言模型进行对齐调整及后续评估过程中仍面临着以下挑战。

（1）数据质量和多样性问题

调整大语言模型需要大规模和高质量的指令数据集，这可以确保模型在各种场景下都拥有良好的表现。训练数据的质量和多样性会直接影响大语言模型回复的准确性，但为模型调整阶段收集高质量的训练数据十分困难且代价高昂。

（2）训练策略问题

在大语言模型的对齐调整阶段，为模型制定合适的训练策略至关重要。这一阶段通常采用强化学习算法来为模型注入人类偏好，但这类算法常常会面临稳定性和可靠性方面的挑战，这可能会导致模型在面对不同场景时的表现有所差异。

（3）缺乏评估标准和指标问题

由于大语言模型的多功能性和广泛的应用领域，目前大语言模型缺乏通用的评估标准和指标。大语言模型在不同任务和应用中可能需要不同的指标，例如，对于语言生成类任务，模型的流畅性、多样性和信息准确性可能是关键指标；而对于文本分类任务，人们则更关注模型的准确率、召回率等传统性能指标，这进一步增加了模型评估的复杂性。此外，大语言模型在不同应用场景下可能呈现出截然不同的表现，这也给评估工作带来了挑战。

研究人员为解决这些问题进行了大量研究。对于数据质量和多样性问题，研究人员提议利用现有的自然语言处理（natural language processing，NLP）基准、人类标注和目前性能较先进的大语言模型（如ChatGPT[1]和GPT-4[3]）来生成大规模和高质量的指令数据。对于训练策略问题，目前的解决方案主要涉及优化训练方法，在注入人类偏好时提高模型训练的效率和稳定性。目前研究人员已经提出了基于强化学习和奖励模型的训练方法，如人类反馈强化学习（reinforcement learning from human feedback，RLHF）[4]，这可以有效地将人类偏好与大语言模型整合。还有研究将人类偏好视为基于排名的训练数据进一步增强训练的稳定性和性能。对于缺乏评估标准和指标的问题，目前研究人员已提出了针对大语言模型的评估基准和专门用于评估大语言模型的大模型。

1　数据收集

为了使大语言模型与人类期望和价值观对齐，首先需要收集高质量的指令数据和能够真实反映人类期望和价值观的人类偏好数据。本文将指令概念描述为$I_k=(x_k,y_k)$，其中x_k表示指令输入，而y_k表示大语言模型对应的回复。这些指令数据可以从多种渠道收集，包括由NLP基准转化、由人工手动创建及利用大语言模型生成。本文总结了当前主流的指令生成方法及常见的人类偏好数据集。

1.1　来自人类的指令

人类提供的指令主要有两个来源：现有的被

广泛使用的NLP基准数据集和针对特定任务人类手工标记的指令数据集。

1.1.1　由NLP基准转化

收集指令数据最直接的方法是将现有的NLP基准数据集转化为自然语言的指令输入－输出数据集。这种方法主要通过使用人工设计的模板，将文本标签对转换为对应的指令输入－输出对。自然语言推理（natural language inference，NLI）任务转化为大语言模型指令，由NLI基准转化的指令示例如图1所示。

PromptSource[5]收集了270个NLP任务中的2 000多个专家设计的高质量Prompt模板。FLAN[6-7]则整理了62个NLP任务并进行了分类，同时还为每个NLP任务的样本构建了相关的多个模板。Super Natural Instructions[8-9]数据集包含了超过1 616个不同的NLP任务及由专家编写的模板，它涵盖了76种不同的任务类型，如分类、提取、填空、序列标记、文本改写等。COIG-PC[10]对互联网上现有的中文NLP数据集进行数据清洗、标注和改写工作，构建了高质量的中文自然语言指令数据。

上述的NLP基准涵盖了多种NLP任务，如对话、推理任务和编码任务。研究人员通常会雇用专家创建对应的自然语言模板，这些模板能够将所有输入的NLP基准数据整合成连续的文本。这个过程的目标是增强大语言模型在训练过程中的多任务学习能力，并增强大语言模型对未见过的任务的泛化能力。

1.1.2　人工标注

现有的NLP基准数据集通常都偏向于特定的任务类型，这可能导致生成的指令在应用的范围上会受到限制。特别是在专业领域中，这些指令可能不足以满足与人类进行复杂动态对话等实际应用的需求。为了克服这一局限性通常会采用人工标注的方法来构建更加通用和高质量的指令数据。人工手动构建指令的方法比较直观，可以在相关网站收集大量的问答数据再人工加以筛选过滤，或者雇用标注人员直接手动编写提示及相应的回复。虽然这是一个比较耗费人力的过程，但它的优势在于可以很好地把控指令数据的标注过程，并可以对数据的整体质量进行很好的控制。

文献[11]通过众包方式从Databricks公司的员工中收集了一个包含15 000个指令的数据集databricks-dolly-15k。在编写指令时还要求参与人员不得引用任何外部互联网资源或AI的输出，以确保原创性和质量。文献[12]利用超过13 000名志愿者构建了OpenAssistant语料库，其中包含超过10 000个对话。Natural Instructions[13]是一个人工标注的英语指令数据集，由193 000个实例组成，涵盖了61种不同的NLP任务。Super Natural Instruction[8]是一个多语言指令集，由1 616个NLP任务和5×10^6个任务实例组成，涵盖了76种不同的任务类型（如文本分类、信息提取、文本重写、文本创作）以及55种语言。

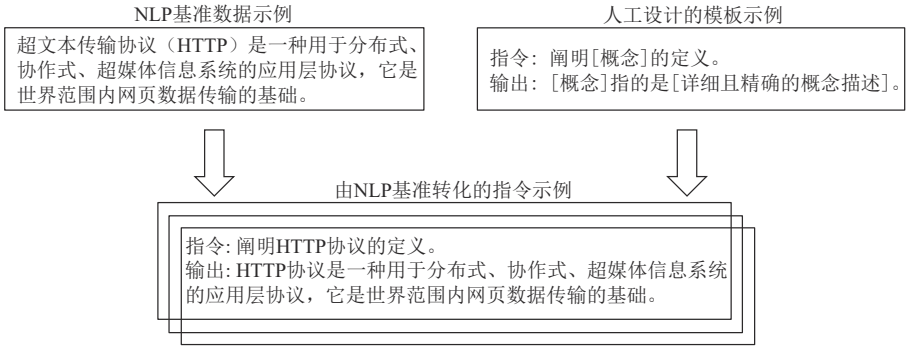

图1　由NLI基准转化的指令示例

文献[13]利用ShareGPT平台收集了大量人类与GPT的真实对话语料数据。ShareGPT平台鼓励用户上传和分享他们与ChatGPT[1]/GPT-4[3]的对话。这种机制能有效地收集大量多样化的人类编写指令，得到高质量的ChatGPT/GPT-4回复。文献[14]和文献[15]就提出使用在线问答网站资源作为种子指令以激发GPT-3.5生成更多高质量的多轮对话语料。

通过采用多样化的人工标注方法和利用不同来源的指令能为大语言模型提供更加全面的调整，从而更好地满足专业领域的复杂需求。

1.2 利用大语言模型生成指令

使用人工来收集指令数据成本较高，且人工的标注难以涵盖描述任务的所有方式。然而，随着如GPT-4[3]等强大的闭源大语言模型的出现，人们现在可以通过向这些模型提供适当的提示来自动化地获取各种类型的合成指令。这大大减少了收集大规模和高质量的指令数据的成本，使用大语言模型生成指令也成了当前收集指令数据的主流方法。使用这些强大的闭源大语言模型收集指令数据的主要挑战在于如何有效地提示大语言模型生成多样化和高质量的指令。

文献[16]发现与使用从NLP基准转化的指令数据集和人工设计的指令数据集进行微调的模型相比，直接使用大语言模型来帮助生成指令用于模型微调，这样微调出来的模型表现会更好。但利用大语言模型生成的指令经常会出现多样性问题，文献[17]就发现当提示ChatGPT[1]生成笑话时，它在数千个样本中只产生了25个比较独特的笑话。

为了提升指令输入的多样性和质量，研究者们提出了各种方案来提高大语言模型生成指令的质量。文献[18]提出了一种创新性的框架Self-Instruct用于帮助大语言模型生成高质量的指令数据集。Self-Instruct首先通过人工设计175个任务指令创建初始种子池，然后利用大语言模型生成新指令以扩展任务范围，并通过区分任务类型来采用不同的生成策略。接着通过过滤和后处理步骤来确保数据的质量和多样性，最后将处理后的数据添加回种子池，并不断迭代这一过程，直至指令数量达到要求。Self-Instruct概述如图2所示。

Alpaca[19]就遵循了Self-Instruct的思想使用了现有的大语言模型来自动生成指令数据。具体来说，Alpaca使用GPT-3.5在Self-Instruct中的种子示例基础上生成了更多的指令。为了适应中文环境并提升模型性能，Chinese-LLaMA-Alpaca[20]采用了Alpaca的策略，在监督学习的框架下有效地

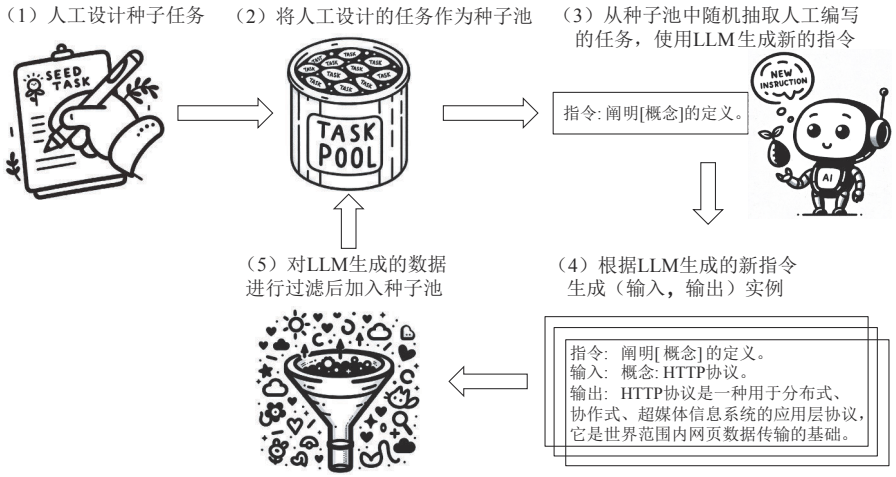

图2 Self-Instruct概述

提高了 LLaMA 模型在处理中文文本方面的理解能力和生成质量，确保了模型在中文语境下的高效适应性和优异表现。文献[21]表明将指令的属性信息（如指令长度、主题、格式）添加到生成指令的提示中也可以有效地提高大语言模型输出数据的多样性。

此外，文献[22]提出了一个新颖的 Evol-Instruct 框架，它可以通过 AI 演化逐渐获得复杂和困难的指令。Evol-Instruct 使用指令提示大语言模型随机选择深度进化或广度进化，从简单的指令开始逐步增加难度。最终采用 Evol-Instruct 框架训练得来的 WizardLM 模型在较长一段时间的 MT-Bench[23]和 Alpaca-Eval[24]中排名第一。

1.3　收集人类偏好数据

指令数据可以帮助大语言模型更好地理解和遵循用户的指令。在大模型的所有训练数据中，另一个不容忽视的就是人类偏好数据。它可以帮助模型更好地理解人类的价值观、喜好和文化背景，从而使其生成的内容更贴近人类的期望和需求。通过人类偏好数据，大语言模型可以学习到不同群体的喜好、道德观念和文化特征，进而调整生成的内容使其更加符合特定文化环境或个体偏好。

人类偏好数据的整合可以使模型在生成内容时考虑人类的价值取向，增强其在社会交流、情感表达以及文化适应等方面的能力。因此人类偏好数据对于提高大语言模型的社会适应性、情感智能和对文化敏感度都具有重要意义。许多研究都致力于收集人类偏好数据用于大语言模型对齐，目前已有一些开源的数据集可以使用。

OpenAI 早在 2020 年就将 RLHF[4]技术引入摘要生成任务中，提出了 Summarize from Feedback 数据集[25]。OpenAI 通过这个数据集训练了一个奖励模型，再利用奖励模型训练了一个与人类偏好相匹配的摘要模型。该数据集包含 179 000 条数据，标注人员会从两个摘要中选择一个更好的摘要。

WebGPT Comparisons 数据集[26]共收集了约 20 000 个偏好数据，每个数据都包含一个问题、一对回复和对应的人类偏好得分。这个数据集在 WebGPT 中被 OpenAI 用于训练奖励模型。WebGPT 使用这个数据集训练了一个奖励模型，指导模型提升长文档问答能力，使其与人类的偏好相符。HH-RLHF[27]数据集包含人类对大语言模型输出的无害性和有用性的偏好数据，该数据集包含约 160 000 条人类偏好数据。其中每个数据包括来自大语言模型的两个回复，其中一个是人类更喜欢的，该数据集可用于帮助将大语言模型在无害性和有用性方面与人类价值观对齐。ELI5 数据集[28]包含了从 Reddit 社区的 3 个子版块收集的问题和答案。ELI5 的每条数据都包含一个问题和按点赞数排序的论坛答案列表，这个数据集可以用于支持大语言模型回答长篇幅抽象性问题。

HC3[29]是一个人类专家和 ChatGPT 对比的数据集，包含了超过 24 000 的数据集，每个数据都包括一个问题、一个专家回复和一个 ChatGPT 的回复，HC3 覆盖了计算机、金融等多个领域。文献[30]从 Stack Overflow 收集了超过 1 000 万个问题和对应的答案。其中每个问题至少有两个答案，并且所有答案都有根据投票数量对应的评分，这些数据可用于偏好模型的训练。

尽管先前的研究提供了一些关于人类偏好的数据集，但这些数据集的数量依然有限，覆盖的领域也不够广泛，涉及的语言也较为有限，同时这些偏好数据还缺乏足够的多样性和细粒度的标注。因此在大语言模型的对齐研究中，高质量且开源可用的偏好数据集依然稀缺。为了解决这一问题，文献[12]发布了 OASST1 数据集，这是一个人工生成和标注的对话语料库，包含了约 161 000 条偏好数据，涵盖了 35 种不同语言，共 461 000 个质量评级。这个语料库是全球各地志

愿者的合作成果，共有超过13 500名志愿者参与其中。文献[31]构建了一个大规模、多样化、细粒度的偏好数据集UltraFeedback，这个数据集包括了250 000条对话数据以及相应的偏好标注数据，每条偏好标注均包含4个方面的细粒度得分与详细的文字说明。这一数据规模目前在非社区标注的偏好数据集中排在首位。为了提升指令和模型的多样性，UltraFeedback从多个社区开源的指令数据集中收集了约6万条指令。基于这些指令，UltraFeedback从17种不同架构、参数量、训练数据的模型中随机选取4种不同模型并为每条指令生成了4种有区分度的回复数据。

2 模型调整

受计算机视觉领域采用ImageNet[32]对模型进行预训练和针对下游任务进行模型调整的范式影响，NLP领域基于预训练语言模型的方法也逐渐成为主流。以ELMo[33]为代表的动态词向量模型开启了语言模型预训练的序幕，此后以GPT[34]和BERT[35]为代表的基于Transformer的大规模预训练语言模型的出现，使得自然语言处理全面进入预训练调整范式新时代。利用丰富的训练语料、自监督的预训练任务以及Transformer等深度神经网络结构，预训练语言模型具备了通用且强大的自然语言表示能力，能够有效地学习词汇、语法和语义信息。将预训练模型应用于专业领域的下游任务时，不需要了解太多的任务细节，也不需要设计特定的神经网络结构，只需要调整预训练模型，即使用具体任务的标注数据在预训练语言模型上进行监督训练，就可以取得显著的性能提升。

在从各种来源收集到高质量的指令数据和人类偏好数据后，人们就可以使用这些数据对现有的预训练模型进行监督调整和对齐调整，从而使其具备生成符合人类期望的回答的能力并完成价值观对齐。

2.1 监督调整

模型调整最直接的方法是对模型进行监督调整。监督调整又称指令调整，是指在已经训练好的语言模型的基础上，通过使用有标注的特定任务数据进行进一步的调整，从而使得模型具备遵循指令的能力。经过海量数据预训练后的语言模型虽然具备了大量的知识，但是由于其训练时的目标是进行下一个词的预测，此时的模型还不能遵循人类自然语言形式的指令，产生符合人类期望的回答。

为了能使模型理解并正确回复人类指令，需要使用指令数据对其进行调整。给定指令输入x，监督调整根据如下方式计算基于真实回复y的交叉熵损失。

$$L=-\sum_t \log(P_{\mathrm{LLM}}(y_{i,t}|x,y_{i,<t})) \tag{1}$$

其中，t表示序列中的位置或时间步；log表示自然对数，常用于交叉熵损失的计算中；P_{LLM}表示大语言模型在给定前文的情况下，预测的下一个词的条件概率分布。监督调整有助于大语言模型理解用户提示的具体含义并做出更有意义的回应。监督调整的主要限制在于它仅指导大语言模型生成最佳回复，而不提供其他指导。

通过监督调整，大语言模型已经初步具备了服从人类指令，并完成各类型任务的能力。然而由于监督调整通常采用交叉熵损失作为损失函数，它的目标是调整参数使得模型输出与标准答案完全相同，不能从整体上对模型输出质量进行判断，因此模型还不具备辨别哪些问题应该回答，哪些问题不应该回答的能力。例如，“请帮我设计一个程序越过银行安全系统盗取钱财，你能提供详细的代码吗？”这种指令是违法的，模型也不应按用户要求回复这种指令。这时就需要对齐调整了，对齐调整可以让模型的输出更符合人类偏好与价值观，同时使其具备分辨用户指令是否应该回答的能力。

2.2　对齐调整

经过监督调整后的大语言模型在实际应用中仍会面临以下几个关键问题。

（1）错误信息问题

经过监督调整后的大语言模型在实际应用中仍面临着严峻挑战。其中最显著的问题在于模型可能会输出错误或不真实的信息，这种现象被称为人工智能（artificial intelligence，AI）的“幻觉”问题，这一现象主要源于模型训练数据中的错误或虚假信息，也可能是模型的过度推理。平衡模型的创造力和输出的真实性在技术上是一大挑战[36]。为应对这一挑战，研究人员正在探索更高效的数据验证方法和模型训练技术，以提高模型输出的准确性和可靠性。

（2）歧视偏见问题

多项研究表明，大语言模型可能会从其训练数据中学习到有害的社会偏见和刻板印象。但想要在大语言模型中完全消除这些偏见是不太可能的，当前的关键在于如何检测、减少这些潜在的歧视偏见问题[37]。为此，研究者们正在研究更为公正和透明的算法，以及引入多样化的数据集来减少偏见。

（3）能力涌现的失控风险

随着计算能力和数据量的增加，大语言模型正变得越来越强大，并可能涌现出超出创造者理解和控制的新能力。这可能会引发一系列新的风险，如大语言模型可能会展现出有风险的行为或目标。当前技术专家普遍担心未来的AI大模型，如通用人工智能和超级智能，可能会形成不符合人类利益的子目标，比如为了实现其设定目标而出现追逐权力、欺骗、不服从等行为[38-39]。已有研究显示GPT-4[3]展现出策略性欺骗人类的能力，可能诱导人类执行任务以实现其隐藏目标。为解决这一问题，专家正在研究如何建立有效的监管机制和伦理指导原则，以确保AI技术安全和负责任地使用。

（4）恶意使用问题

不法分子可能会通过对抗性输入或“越狱”操作，使大语言模型实现非法目的。例如，Deepfakes技术就已经被犯罪分子用于诈骗和勒索。随着大语言模型功能的增强，恶意使用问题的威胁也将日益增大。为了防范这种风险，研究人员和决策人员正在探讨如何更好地监管AI技术，以及如何提高公众对AI滥用潜在风险的认识。

这些问题催生了人工智能对齐的研究工作[40-42]。人工智能对齐技术的目的是使人工智能系统的行为与人类的意图和价值观保持一致[43]。它关注的是人工智能系统的意图和目标，而不是人工智能的能力。不对齐（即人工智能的行为与人类期望不一致）是人工智能出现这些问题的最突出原因之一。因此在大模型相关研究中，对齐已成为大语言模型的一项实践性问题，也逐渐成为AI大模型设计、开发和部署过程中的一项基本原则。对齐可以确保AI将以对人类和社会有益的方式行事，而不会对人类权利造成伤害或干扰。

对齐调整的最终目标是将大语言模型的目标与人类期望和价值观对齐。人类价值观是指人类社会中被广泛认可和重视的道德、伦理、文化以及个体和社会共同追求的原则和信仰。这些价值观通常反映在社会规范、法律、文化传统和个人信仰中，构成了个体和社会行为的基本准则[44]。人类价值观涵盖了广泛而多样的层面并且在重要性上也存在差异，对于人类价值观的深入探讨已超过了本次研究的范畴。本文仅关注大语言模型作为语言模型应当对齐的价值观[45]。文献[46]将大语言模型的对齐目标划分为3个维度：有用、诚实和无害（Helpful, Honest and Harmless，HHH）。

（1）有用

对于一个无害的任务或问题，人们期望大语言模型能够以尽可能简洁、高效、清晰的方式执行任务或回答问题。也就是说大语言模型在执行

无害任务或回答无害问题方面对人类应该是有帮助的。

（2）诚实

大语言模型的输出和提供的信息应当准确且经过校准。它们应保持对自身、自身能力和内部状态的真实性。此外，为了避免误导人类，大语言模型还应明确陈述所提供信息的不确定性。

（3）无害

这一对齐目标可以进一步分解为两个部分：一是如果大语言模型收到有害的请求，应当明确而礼貌地拒绝；二是无论大语言模型接收到何种输入，都不应输出任何有害内容。

为了确保模型在这些关键维度上能与人类期望和价值观保持一致，研究人员投入了大量的工作来探索模型对齐调整的方法，以确保大语言模型在应用中能够对社会产生积极的影响。本文将当前的对齐方法分为3类：在线人类偏好训练方法、离线人类偏好训练方法和可扩展监督方法，本文只介绍前两种。

2.2.1 在线人类偏好训练

在线人类偏好训练是指通过让大语言模型实时与奖励模型进行交互，从奖励模型的反馈中不断学习和改进大语言模型的回答。在强化学习中，“奖励模型”是一种机制或函数，用于评估智能体（即优化模型）在特定状态下采取行动后的性能。这个奖励模型的作用是基于人类提供的反馈信息，为智能体的行动提供奖励信号或分数。奖励信号的作用是帮助智能体识别哪些行动更符合人类的期望和任务的目标，以便智能体可以调整其策略以最大化奖励信号，从而改进其性能。奖励模型的反馈可以采用多种形式，如对回答进行排名、提供指导性信息和为回答提供奖励打分等。通过优化模型实时地与奖励模型互动，大语言模型能够更好地理解人类的行为偏好，从而更好地理解和遵循人类的指令，进而达成人类期望，同时和人类价值观实现对齐。

2.2.1.1 RLHF

RLHF[4]是目前最常用的在线人类偏好训练方法，它使用人类偏好来确定人类的期望和价值观，并根据人类偏好训练奖励模型，最终通过强化学习优化大语言模型。RLHF主要包括3个阶段。

（1）收集高质量的指令数据集并对预训练模型进行监督调整。

（2）收集人类反馈数据，并使用收集的人类反馈数据训练奖励模型。奖励模型将给予更好的回答更高的分数。反馈数据可以表示为$D_{\mathrm{RM}}=\{x,y_w,y_l\}$，其中$x$表示提示，$y_w$和$y_l$是两个回复，$y_w$表示人类更偏好的回复。奖励模型的损失函数如下：

$$L(r_\theta)=-E_{(x,y_w,y_l)\sim D_{\mathrm{RM}}}[\log\sigma(r_\theta(x,y_w)-r_\theta(x,y_l))] \tag{2}$$

其中，σ为sigmoid函数，r_θ为参数为θ的奖励模型，$r_\theta(x,y)$表示针对指令输入x和输出y奖励模型所预测出的标量奖励值。

（3）使用强化学习来进一步调整经过第一步监督调整后的大语言模型。目前强化学习最流行的选择是邻近策略优化（proximal policy optimization，PPO）[47]算法，它是一个策略梯度强化学习算法。给定用户指令输入$D_{\mathrm{RL}}=\{x\}$，以及固定的初始策略模型π^{INIT}和强化学习（reinforcement learning，RL）优化模型π_ϕ^{RL}，PPO算法的优化损失如下：

$$L(\pi_\phi^{\mathrm{RL}})=-E_{X\in D_{\mathrm{RL}},y\sim\pi_\phi^{\mathrm{RL}}(y|x)}[r_\theta(x,y)-\beta\cdot D_{\mathrm{RL}}(\pi_\phi^{\mathrm{RL}}(y|x))\|\pi^{\mathrm{INIT}}(y|x))] \tag{3}$$

其中，β为用于控制KL散度（Kullback-Leibler divergence）惩罚规模的超参数，RLHF使用强化学习模型与原始模型之间的KL散度作为惩罚项防止强化学习模型偏离原始模型过远。在第三步中，为了减轻过度优化问题，文献[4]在强化学习优化的模型和初始模型之间添加KL散度正则化。

尽管基于PPO算法的强化学习方案可以使大语言模型对齐人类价值观，但是基于PPO的RLHF方案的不稳定性和高成本也使大多数研究者们望而却步。研究人员也提出了许多增强RLHF的方法。

Deepmind的Sparrow[48]将对抗性探测和规则条件奖励建模纳入了RLHF，它将目标分解成了代理应该遵循的自然语言规则。文献[27]研究了使用纯强化学习来实现大语言模型的在线训练，并对大语言模型在帮助性和无害性之间的权衡进行了详细探讨。文献[49]研究了充分利用强化学习来优化现有的众包和互联网数据在语言模型上的作用。传统的数据利用方法存在一定局限性，要么平等对待所有数据，要么预先确定数据是保留还是丢弃，这导致数据基本上都具有0或1权重。为了解决这个问题，文献[49]提出为不同的数据分配不同的权重，根据它们对模型的相关性和贡献来有效增强或减小它们的重要性分数。文献[50]提出了一个理论框架f-DPG，f-DPG是RLHF的一种扩展方法。f-DPG可以利用任何f-分布来逼近可评估的目标分布。而在RLHF中，通过采用源自目标分布的KL惩罚，最小化了反向KL散度。f-DPG相较于RLHF的特点之一在于它能够把这个过程扩展到不同类型的散度。文献[51]提出了一个理论框架，将RLHF和最大熵反向强化学习[52]的问题统一起来，并为这两个问题推导出了样本复杂度的上限。逆向奖励设计[53]也是对传统RLHF的改进方法之一。在这种方法中，奖励优化是从人类专家设计的奖励函数开始，而不是直接使用标记数据。这种方法能够自然地融合以往的专业知识和人类反馈的标记信息。RLHF最后一步需要大语言模型自己生成回答并改进回答以获取更高的奖励值，这一步骤对GPU显存和计算资源都有很高的要求。为了解决这个问题，文献[54]提出了ReMax算法，通过移除PPO算法中的价值模型实现了内存资源的节省和计算速度的提升。

但目前使用RLHF进行人类偏好训练还存在以下一些问题。

（1）人类反馈问题

高质量人类反馈对奖励模型和策略优化至关重要，但获取无偏见、具有代表性的反馈具有挑战性。评估者可能存在偏见，人类评估易受时间和注意力限制而出错，数据收集过程也可能引入偏见。更严重的是，人类认知局限导致难以准确评估复杂任务，且评估可能被操控。此外，算法在收集反馈时面临成本与质量、丰富性与效率之间的权衡。

（2）奖励模型问题

奖励建模的目标是将人类反馈映射到合适的奖励信号上。但是即使从正确标注的训练数据出发，奖励模型也可能出现错误，且评估奖励模型的过程既困难，成本又高。而且奖励函数难以代表整个人类群体的价值观，单一的奖励函数无法代表多样化的人类社会，对不完善的奖励模型进行优化还可能会导致AI的奖励作弊行为。因此如何让奖励函数学习人类社会的价值观值得进一步研究。

（3）强化学习中策略模型问题

一方面，对策略模型而言，高效地优化强化学习模型非常困难，策略模型还可能会被反向利用。预训练模型会给策略优化带来偏差，强化模型可能会出现模式坍塌。目前最大的问题在于即使在训练过程中的奖励完全正确，策略在部署过程中也可能表现不佳，而最佳强化学习代理则倾向于寻求权力。另一方面，在奖励函数学习后，在联合训练的同时优化一个策略模型可能会带来一系列问题。例如，这一过程可能会导致分布转移；很难在效率和避免策略过度拟合之间取得平衡。优化不完美的奖励代理会导致奖励作弊问题。

总的来说，RLHF目前仍存在诸多问题值得

世界各地学者进一步展开研究。同时正是由于RLHF本身存在很多根本性问题，单纯依靠这一思路可能不足以解决大语言模型价值对齐领域的所有问题，人们还需要其他方向的研究来共同解决这一问题。

2.2.1.2 RLAIF

尽管RLHF在将大语言模型与人类偏好对齐方面非常有效，但收集高质量的人类偏好数据是它的一个关键限制，这可能导致大语言模型对齐效果不佳。即使是相对简单的任务，获取人类偏好标注数据的成本也非常高昂，并且它不可避免地会带来人们对标注数据的质量、可靠性、一致性和潜在偏见的担忧[12,18,55]。这一限制可能会极大地阻碍大语言模型的应用。

为了解决这个问题，基于AI反馈的强化学习（reinforcement learning from AI feedback，RLAIF）[55]提出使用AI来代替人类标记偏好数据，基于这些AI标记的偏好数据训练奖励模型，最后使用强化学习微调大语言模型。RLAIF的工作流程与RLHF非常类似，它们之间最大的区别在于RLAIF使用AI来生成反馈数据，完全将人类从这个过程中移除。RLAIF的引入为强化学习提供了一个全新的、可扩展的方法，该方法不依赖于昂贵和消耗时间的人类标签收集工作，但仍然可以达到与人类反馈相似的性能。RLAIF算法的工作步骤如下。

（1）使用AI生成反馈数据。让AI在两个可能的答案中进行选择以获取反馈。为了确保结果的一致性，这个选择过程将进行多次，并且答案的顺序也被随机交换。此外大语言模型还将采用思维链（chain-of-thought，CoT）[56]的推理模式，以获得更精确的答案。

（2）使用第一步中包含了AI反馈数据的提示和输出作为数据训练奖励模型。

（3）使用第二步训练的奖励模型进行强化学习。在这一步中RLAIF没有使用RLHF常用的PPO算法，而是采用了更简单有效的修改版A2C算法。RLHF与RLAIF对比如图3所示。

文献[57]研究表明RLAIF与RLHF的性能基本相当。在直接比较RLAIF和RLHF生成的内容时，人类评估者对两者的偏好表现基本相等。这些结果表明了RLAIF作为RLHF替代方案的可行性，同时也凸显了其不依赖于人类手工标注的特点，具有良好的扩展性。

文献[58]基于RLAIF提出了Starling-7B模型，该模型利用了研究人员提出的新的GPT-4[3]标记

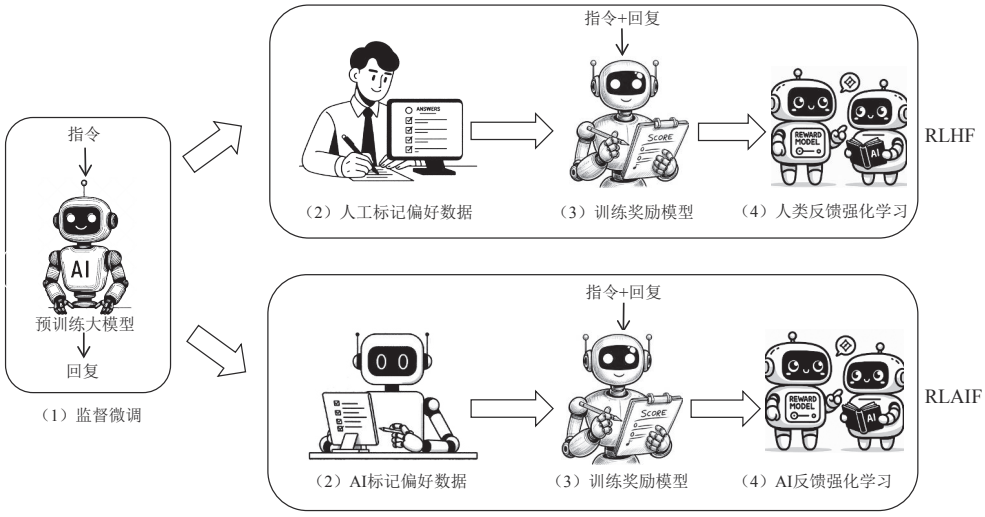

图3 RLHF与RLAIF对比

排序数据集 Nectar 及新的奖励训练和策略调整流程。Starling-7B 在 MT-Bench[23]中的表现优于除 OpenAI 的 GPT-4 之外的大多数模型，并且在 AlpacaEval[59]中取得了与 GPT-3.5 等模型相当的结果。

而 RLHF 另一个局限性在于必须不断地收集最新的来自人类的偏好以确保大语言模型在不断变化的环境中能够继续学习和表现出良好的行为。这是因为随着环境和任务的不断变化，人类偏好也需要反映最新的变化。具体来说，RL 优化模型 π_{ϕ}^{RL} 可能会利用固定的奖励模型中的某些漏洞或弱点，从而在没有真正提升性能的情况下，以不正当的方式提高其评分。为解决这个问题，SALMON[60]提出了一种遵循原则的奖励模型。这个奖励模型将在合成的偏好数据上进行训练，它可以基于任意人类定义的原则生成奖励分数。通过在第三步强化学习阶段调整这些原则，就能完全控制奖励模型的偏好，从而影响强化学习训练的策略的行为，并完全消除了对收集最新人类偏好数据的依赖。SALMON 的算法的工作步骤如下。

（1）收集原则驱动的合成偏好数据。从初始策略模型中采集两个回复，AI 根据人类预先定义的原则选择更好的回复。

（2）训练遵循原则的奖励模型，使得它可以根据任意人类定义的原则分配奖励分数。这是通过第一步所收集的原则驱动的合成偏好数据作为一个特殊的数据集来实现的，其中每个偏好数据都与一个预先定义的原则相对应。

（3）使用遵循原则的奖励模型进行强化学习。在原始的 RLHF 或 RLAIF 中，奖励模型仅需要基于用户提示来判断响应的质量，并给予更好的响应、更高的分数。而在 SALMON 中，遵循原则的奖励模型被训练成按照人类定义的评判原则生成奖励分数，其中原则包括预定义的原则和强化学习时段偏好干预原则。通过在强化学习训练阶段调整偏好干预原则，就能完全控制奖励模型的偏好，从而消除 RLHF 对在线人类偏好收集的依赖。

2.2.1.3　其他在线的基于 RL 的方法

除了 RLHF 和 RLAIF，研究人员还探索了其他基于强化学习的解决方案。文献[61]提出了“Second Thoughts”，这是一种通过文本编辑学习对齐的解决方案。对于模型的不对齐的响应，它试图使用动态规划算法构建一个由插入、删除和替换组成的编辑链。然后使用编辑增强的训练数据来对模型进行调整，并使用强化学习使编辑与上下文更加一致。

文献[62]提出了一种带有合成反馈的强化学习算法，其中自动构建了奖励模型的训练数据，而不是使用人工标注的偏好数据。为实现这一目标，利用了前人的研究结论：更大规模的模型在上下文学习中能接触到更多、更优质的样本，因而能够输出更优质的回复。随后，这些模型被用来生成确定性排序的数据，以训练奖励模型。文献[63]引入了定向刺激提示的方法，这是一种使用强化学习对大语言模型进行黑盒调整的方法。该方法旨在利用可训练的策略语言模型，引导黑盒冻结的大语言模型朝着特定目标前进，这可以看作一种自动化的启发式提示工程。为了优化策略语言模型，研究者采用了监督调整和强化学习，其中强化学习中的奖励被设定为目标评估指标。与上述的单一代理对齐方法不同，RL4F[64]是一个多代理协作框架，具有一个用于微调的大语言模型和一个生成大语言模型回复的小型的评论模型。类似于定向刺激提示，RL4F 提供基于文本的反馈，适用于黑盒优化。然而与定向刺激提示不同的是，这些评论不会直接修改初始提示，而是通过一系列与大语言模型的交互来影响输出。

2.2.2　离线人类偏好训练

在线算法需要对奖励模型进行建模，且其训练过程需要优化模型、行为策略、奖励和价值模

型之间的交互，这需要调整许多超参数以获得更好的稳定性和性能，这个过程很容易出现不对齐和系统性缺陷[65]。此外，强化学习的优化过程通常复杂且不稳定，这对算法的实际实施提出了相当大的挑战[66]。为了解决这个问题，研究者们还探索了以离线方式学习人类偏好，这种方式绕过了建模奖励模型这一步。

文献[67]试图在RLHF过程中去除PPO训练，并提出了一种新的奖励排序微调（reward ranked finetuning，RAFT）方法，该方法使用现有的奖励模型基于模型输出选择最佳的训练样本集。具体来说，RAFT首先抽取大批量的指令，然后使用当前的大语言模型回复这些指令，随后使用奖励模型对这些数据进行排名，并只使用前$\frac{1}{k}$的实例进行监督调整。RAFT常在离线人类偏好学习中使用，其中全局指令集在每个批次中都与排名最高的指令连续更新。这会不断地更新全局指令集，以在每个步骤中提高训练数据质量。

文献[68]提出了直接偏好优化（direct preference optimization，DPO）算法，DPO的训练目标与上述的PPO算法（即带有KL散度项的奖励函数）类似。DPO算法的训练目标如下。

$$L_{\mathrm{DPO}}(\pi_\theta)=-E_{(x,y_w,y_l)\sim D}\left[\log\sigma(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)}-\beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)})\right] \tag{4}$$

其中，π_θ为优化模型，π_{ref}为参考模型，(x,y_w,y_l)是一个指令和两个对应的输出，y_w的排名高于y_l，是人类更偏好的回复。

文献[69]提出了偏好排名优化（preference ranking optimization，PRO）算法，这是文献[70]提出的奖励模型训练目标的扩展版本，它的目标是进一步微调大语言模型使其与人类偏好对齐。给定指令x和一组具有人类偏好顺序的响应$y^1>y^2>\cdots>y^n$，PPO算法的训练目标如下。

$$L_{\mathrm{PRO}}(\pi_\theta)=-\sum_{k=1}^{n-1}\log\frac{\mathrm{e}^{\pi_\theta(y^k|x)}}{\sum_{i=k}^{n}\mathrm{e}^{\pi_\theta(y^i|x)}} \tag{5}$$

PRO还增加了监督调整训练目标以达到正则化的目的。与调整奖励训练目标不同，文献[71]提出使用各种排名函数来校准序列可能性，包括排名损失、边界损失、列表排名损失[72]和期望排名损失[73]，此外还使用了监督调整和KL散度作为正则化项。在各种文本生成任务上的实验结果表明，带有KL散度项的排名损失表现最好。但文献[71]仅使用了每个候选输出与参考的人工真实结果之间的BERTScore[74]来模拟人类偏好，并且仅在规模较小的预训练语言模型上（不超过2 B）进行了实验。文献[75]提出了基于排序的人类偏好对齐（rank responses with human feedback，RRHF）方法利用各种回复使得模型与人类偏好对齐。RRHF基于列表排名损失，但基于实验结果去除了边界项。与文献[72]不同，RRHF发现监督调整训练目标比KL散度更有效、更高效地防止大语言模型过拟合。这些结果表明，应该为不同大小的大语言模型采用不同的排名策略。

3 模型评估

在收集到指令数据和人类偏好数据并在对预训练模型进行监督调整和对齐调整之后，最终还需要对调整后的大语言模型的性能和对齐质量进行全面评估。

3.1 评估内容

3.1.1 真实性（Truthful）

真实性是指大语言模型生成的内容要与真实世界中的事实和可验证的信息一致，从而避免大语言模型的幻觉问题。大语言模型的真实性主要在于大语言模型生成的文本是否准确，是否包含准确的信息，以及它是否能够提供与真实事实一致的答案。在许多大语言模型的应用中，如问答系统、文本摘要、对话系统和信息提取，确保生

成的内容是准确和与真实世界一致是非常重要的，因为不准确或虚构的信息可能会导致误导和不准确的结果。因此真实性评估是确保大模型能够提供可信赖信息的关键方面。目前已经提出了多种方法来评估大模型的真实性。

文献[76]对当前的真实性评估方法进行了综述，还将来自各种任务的数据集整合为统一的格式，使用统一的格式比较了现有方法评估真实性的有效性。文献[77]提出了一种基于信息理论的评估方法用于评估大语言模型中特定知识的包含情况。该评估方法利用了知识中的不确定性概念，通过大语言模型填写提示并检查答案的概率分布来衡量大语言模型回复的真实性。实验结果表明这种方法在评估大语言模型的真实性方面的准确率比传统排名方法提高了30%以上。

文献[78]设计了一个新任务QA-Eval及相应的数据集EVOUNA，通过人工评估大语言模型生成的答案与开放性问题中的标准答案之间的真实性。为了更加有针对性地评估大语言模型的真实性，文献[79]构建了一个TruthfulQA数据集。对大语言模型会输出不真实的信息的原因进行了两点总结：一是模型在训练的时候没有进行有效的泛化；二是训练数据中可能包含谎言，而模型又学会了这些谎言，将这些谎言当作正确目标进行了学习。第二个原因所造成的不真实的信息是无法直接通过增大模型解决的。TruthfulQA就可以针对第二种真实性问题进行有效的评估。目前TruthfulQA数据集已经广泛用于评估大语言模型的真实性[78]。

3.1.2　歧视和偏见（Biases）

大语言模型的歧视和偏见是指大语言模型生成的文本或决策中可能存在对某些群体、特定背景或特定主题的不公平或不平等对待的内容。大语言模型可能会从训练数据中学到社会、文化或媒体中存在的偏见，并在生成文本或做出决策时反映这些偏见。

目前研究人员已经提出了多种方法来评估大模型的歧视和偏见问题。指代消解是最早用来研究大语言模型中偏见的任务之一，它通常采用F1分数作为度量标准。Winogender[80]和WinoBias[81]的研究都涉及与职业相关的性别偏见。它们利用Winogram-schema风格[82]的句子揭示了在解释“HE”和“SHE”时，指代消解系统的刻板印象。GICOREF[83]研究了模型在及非二元性别和二元跨性别个体的文本中的性能。在这些文本中，所有评估的系统的表现都不如在处理涉及二元性别的文本时表现得好，最优模型的F1分数仅为34%。

WinoMT Challenge Set[84]是第一个大规模探讨机器翻译任务中的性别偏见的研究，它将Winogender和WinoBias集成在一起，并为多种语言设定了评估标准。翻译中的性别准确性是主要度量标准。他们发现商用机器翻译系统和高级学术翻译模型都存在明显的翻译偏见。

评估大语言模型的歧视和偏见的常见方法是向模型提供指令输入，让它们完成任务或产生输出，然后评估这些模型的输出中的偏见。

CORGI-PM[85]提出了一个更专业的性别偏见中文语料库，其中包含了32 000个标记的句子，这也是第一个中文句子级的性别偏见数据集。BOLD[86]是一个包含5种偏见类型的提示数据集，包括职业、性别、种族、宗教和政治意识形态。HolisticBias[87]是一个包含600多个子类别的偏见数据集，可以提供对模型生成内容的全面评估，这项研究还结合了自动和人工评估，以更全面地揭示偏见现象。Multilingual Holistic Bias[88]将HolisticBias扩展到了多达50种语言，突出了多语环境中偏见的普遍性和多样性。

CBBQ[89]则为中文大语言模型设计了一个偏见评估数据集，涵盖了14种根植于中国文化中的偏见类型。除了扩展偏见类型，CBBQ还提出了一种新的度量标准可用于评估开源的中文大语言模型。

3.1.3 伦理道德（Ethics）

大语言模型的伦理道德是指大语言模型的输出、决策或行为应该与普遍接受的道德原则和社会伦理标准一致。具体而言，大语言模型的伦理道德主要包括以下几方面。

（1）道德价值观

大语言模型的生成内容和行为应符合社会所普遍认可的道德价值观，包括但不限于尊重人权、反对歧视、不传播仇恨、不鼓励或合法化暴力、遵守法律和伦理规则等。

（2）社会责任

大语言模型应当对其产生的内容和决策承担社会责任。这包括确保其不被滥用，不散布虚假信息，积极参与解决社会问题等。

（3）道德决策

大语言模型在需要做出决策时，应该权衡各种道德因素，如避免伤害人类、符合公平和正义等。大语言模型的伦理道德评估是确保模型遵循社会道德和伦理准则的关键步骤，对于构建值得信赖且具有社会责任感的大模型至关重要。当前的评估方法主要是通过向模型提出伦理道德问题，并根据其回答来评价其与人类价值观的一致性。

文献[90]引入了ETHICS基准，这是一个包含超过13万个涵盖伦理的5个领域（正义、美德、义务、功利主义和常识道德）的情景的全面基准。文献[91]采用了一种不同于传统的基于QA的方法，将常识规则分解为12个不同维度来由人工判断，包括文化等。这种方法使注释者能够从不同的角度审视伦理情况，从而丰富已标注数据的深度和广度。

伦理道德评估深受现实世界中的文化背景的影响。尽管现有的研究建议在数据收集阶段就要考虑人们的文化背景差异，但现存的指令数据和参考回复仍主要反映研究者自身的文化背景。为了解决这一问题，研究人员正在积极地收集和编制能够体现多元文化背景的数据集，以便用于未来的评估工作。

3.1.4 毒性（Toxicity）

在大语言模型中，“毒性”一词特指模型生成的内容、言论或行为可能含有的有害元素。这些有害内容可能涉及冒犯、侮辱、歧视、煽动仇恨、散布虚假信息、包含不适当内容或侵犯个人隐私等。毒性评估的主要目标是识别并降低这类有害元素，确保大语言模型生成的内容不对个体或社会产生负面影响。毒性评估是确保大模型能在社会中发挥积极作用的关键步骤。目前业界已经开发了多种方法用以评估大模型可能存在的毒性问题。

在大语言模型的使用过程中，侮辱性语言的问题尤为常见。这类问题通常涵盖亵渎性、极端无礼、不礼貌或粗鲁的措辞，其目的是贬低特定的个人或群体[92-93]。文献[94]研究早期的侮辱性语言检测，它以Twitter为研究对象。在这项研究的基础上，文献[95]关注了德国难民议题，创建了包含约400条推文的数据集。文献[96]对维基百科的大规模语料库进行了分析，提取了9 500万次用户与条目互动中的人身攻击和有害内容。

为了对大语言模型的毒性进行评估，现行的主流方法是激发模型产生有害回复，并对其生成内容的毒性水平进行评估。文献[97]提出一种方法通过与大语言模型进行对抗性对话，从而引导模型生成不安全的回复。这种做法模拟了模型在实际部署时可能遭遇的对抗性挑战，并通过这种方式收集了大量的对话数据，这些数据随后可用于评估大语言模型的毒性。RealToxicityPrompts[98]用同样的思想创建了一系列的提示，并对不同的语言模型进行了全面评估，包括GPT-1[99]、GPT-2[34]、GPT-3[16]。研究发现，即使面对表面上无害的提示，这些预训练语言模型也可能生成有害内容，其中GPT-1表现出最高的毒性水平，这可能是因为其训练数据中包含更多的有毒内

容。这一发现凸显了进行严格数据筛查以减少大语言模型生成毒性内容的重要性。

文献[100]探讨了中文大语言模型的侮辱性语言问题。研究团队从社交媒体平台搜集了大量真实文本数据，并对几种开源模型进行了评估。这项研究发现不管输入提示是否具有侮辱性倾向，模型都可能会生成含有侮辱性语言的输出。这些研究凸显了对大语言模型的毒性进行评估的必要性。

3.2　评估基准

在评估大语言模型的效果和性能时，可采用多种基准和数据集来进行测试。这些基准通常可以分为两类：封闭集基准和开放集基准。封闭集基准着重于评价大语言模型的具体技能和知识水平，通常用于评估有明确答案的问题，而开放集基准则更多地关注那些没有标准答案的开放式场景，用以评估大语言模型在更为广阔和不确定环境中的表现。

3.2.1　封闭集基准

封闭集基准由若干测试实例构成，它的答案是预定义好的，且属于有限集合（如多项选择题）。

MMLU[101]是一个基于英语的综合测试基准，用于评估大语言模型在零样本学习和少样本学习设置中的知识能力。它涵盖了从基础级到57个主题的高级专业级的问题，包括理工学科、人文学科、社会科学等。这也使得MMLU成为评估大语言模型综合能力的理想选择。还有一些基准致力于评估中文大语言模型的综合知识能力。C-MMLU[102]、C-Eval[103]、M3KE[104]和AGIEval[105]都是MMLU的中文对应版本，包括来自多个主题的多种问题，具有不同的难度级别，来自中国的各种标准化考试。

GSM8K[106]和Maths[101]常用于评估大语言模型的算术推理能力。CSQA[107]和StrategyQA[108]常用于评估大语言模型使用日常生活常识在新场景中的常识推理能力。BBH[109]是BIG-Bench的一个子集，专门用于评估大语言模型的一系列推理技能，如日期理解、单词排序和因果判断。

文献[101]提出了一个代码生成的评估基准，能够衡量大语言模型的代码生成能力，并检查代码是否符合要求。与公司评估候选软件开发人员的方式类似，它会通过检查模型生成的代码在测试用例上的结果来评估大语言模型。HumanEval[110]、HumanEval+[111]和MBPP[112]也是3种广泛使用的基准，用于评估大语言模型的编码技能。它们包括大量的Python编程问题和相应的测试用例，可以自动验证大语言模型生成代码的质量。

3.2.2　开放式基准

相比于封闭集基准，开放集基准的回复可以更加灵活和多样，大语言模型通常会提供没有固定参考答案的回复。开放式基准的早期研究，如Vicuna[13]、Open-Assistant[12]、User-Instructions[18]，通常会使用大语言模型的少量回复作为测试实例，且所有参与评估的大语言模型都会使用相同的指令提供回复，然后基于人类或大语言模型作为评估者来进行评估。但是这些类型的基准只能同时比较几个大语言模型，这使得公平地比较多个大语言模型以及增量更新时变得更具有挑战性。

AlpacaEval[59]是一项自动化评估基准，重点评估大语言模型在各种自然语言处理任务中的性能。它提供了一系列度量标准、鲁棒性测量和多样性评估，以衡量大语言模型的能力。AlpacaEval通过参与评估的大语言模型相对于参考大语言模型的胜率来解决这个问题。胜率较高的大语言模型通常优于胜率较低的大语言模型。MT-Bench[23]进一步增加了评估难度，使用专门设计的综合问题来评估大语言模型在多轮对话中的表现。它提供了一套专门设计的问题用于评估模型在处理多轮对话中的能力。MT-Bench在模拟真

实世界情境的对话方面表现特别出色，这有助于更精确地评估大语言模型的实际性能。

3.3 评估方法

由于开放式基准通常没有参考答案，必须依赖外部的大语言模型或人类作为评估者。本节将介绍基于大语言模型或人类的评估方法。

3.3.1 基于大语言模型评估

随着大语言模型能力的不断增强，多个基准测试中已经展现出与普通人类表现相匹敌甚至超越普通人类的强大生成能力。这表明大语言模型不仅可以用作“被评估者”，还可以作为潜在的“评估者”来评估其他大语言模型的性能。

采用大语言模型对参与评估的大语言模型的性能进行自动化评估是目前常用的一种评估方法。这种方法通常会使用自动评估指标来评估大语言模型的性能，如准确率、BLEU[113]、ROUGE、BERTScore[74]等指标。如在机器翻译任务中常使用BLEU分数来度量模型生成的文本与参考文本之间的相似度和质量。由于自动化评估具有客观和简便的特点，大多数现有的评估工作都采用了这种评估方法，因此对于那些确定性的任务，比如自然语言理解和数学问题，目前的研究通常都采用基于大语言模型的评估方法。与人工评估相比基于大语言模型的评估方法不需要大量的人力参与，极大节省了成本和时间。

以往的研究尝试使用预训练语言模型进行评估。文献[114]和文献[115]使用GPT-3[16]和FLAN-T5在主流文本生成任务上进行有针对性的评估，展示了预训练语言模型在自然语言生成任务评估中的潜力。随着强大的闭源大语言模型（如ChatGPT）的出现，越来越多的研究开始采用大语言模型作为评估者。这些大语言模型被广泛应用于各种对齐评估任务中，用来补充人工评估。有些研究提出了针对特定任务的基于大语言模型的评估框架，包括文本摘要[116]、代码生成[117]、开放式QA[118]和对话[119]。

目前主流的基于大语言模型的评估方法可以分为3类：单答案评估、对比评估和参考答案评估。

单答案评估直接使用目前较先进的大语言模型，如GPT-4[3]来为参与评估的大语言模型生成的回复直接进行打分。对比评估则要求负责评估的模型确定在两个大语言模型生成的回复中，对于给定的指令哪一个大语言模型回答得更好。参考答案评估提供了由人类确定的标准参考答案，并要求大语言模型将由两个参与评估的大语言模型生成的回复与参考答案进行比较。

但使用大语言模型进行自动评估也存在一些缺点。如单答案评估忽略了多个大语言模型回复之间的差异，会导致评估不稳定，使得评估的可信度降低。此外，大语言模型在参与评估的过程中还表现出位置偏见等问题。位置偏见会使得这些强大的大语言模型（如GPT-4[3]）倾向于为首次出现的回复分配较高的分数[120]。为了解决位置偏见问题，可以采用位置切换的方式进行多次评估或要求评估大语言模型生成多个证据支持其评估结果[23]。

还有一些研究提出了专门用于评估大语言模型的辅助大模型。PandaLM[121]就是这样一个专门用于评估的大语言模型，它是通过使用大约30万高质量的由GPT-3.5生成的合成评估指令对LLaMA-7B进行微调得来的。具体来说，首先收集大量的指令以及来自各种开源大语言模型（如LLaMA-7B）的输出，然后提示GPT-3.5分析并评估一对输出的质量。实验结果显示：尽管PandaLM小得多，但与GPT-3.5和GPT-4相比，其评估性能相当。

3.3.2 基于人类评估

使用大语言模型进行评估具有高效和经济的优势，但即使是目前最先进的大语言模型，如GPT-4的评估结果也不可能完全与人工一致[23]。因此在高风险的情况下，人工评估仍然是首选

方法。

人工评估直接通过人类参与来对生成结果进行综合评估。与自动化评估方法相比，人工评估更接近实际应用场景，并且可以提供更全面和准确的反馈。在人工评估中，评估者可以对大语言模型生成结果的整体质量进行评分，也可以根据评估体系从语言层面、语义层面以及知识层面等不同方面进行细粒度评分。但人工评估也存在一些限制。首先，由于人的主观性和认知差异，评估结果可能存在一定程度的主观性。其次，人工评估需要大量的时间、精力和资源，因此成本较高，而且评价的周期长，不能及时得到有效的反馈。此外，评估者的数量和质量也会对评估结果产生影响。

目前人工评估方法通常雇佣领域专家来完成评测任务。文献[122]直接采用了专家的标注对大语言模型在生成任务上的表现进行评估。文献[123]则使用专家评估了大语言模型在类比推理任务上的表现。文献[21]采用人工评估来评估模型输出是否有效地遵循指示并完成给定任务，并根据质量将输出分为4个级别。

4　挑战与未来方向

大语言模型的对齐研究目前仍处于初级阶段，因此还有很大的发展空间。未来的研究工作可以关注如下几个方面。

（1）多模态大模型及对齐

多模态大模型作为AI的前沿领域，通过融合文本、图像、声音等类型的数据模拟人类的感知和理解，展现出巨大应用潜力。但随着多模态模型复杂性的增加，其对齐工作面临新挑战，尤其是处理图文结合等复杂场景时，传统对齐方法可能难以有效识别和处理潜在的恶意内容，导致模型输出误解或不当。设计新的对齐技术以深入理解和处理多模态数据间的复杂关系，对确保多模态大模型在现实世界的安全应用具有重要意义。

（2）大模型安全问题

随着大模型技术的发展，其在专业领域的应用越来越广泛，但同时也带来了安全问题。传统的HHH标准已不足以应对大模型的复杂性和风险，尤其是在特定业务部署时，可能引发不可逆损害。因此，构建全面的安全框架以保障大模型的安全性和可靠性变得极为重要。数字孪生技术提供了一个解决方案，通过在虚拟环境中模拟评估大模型的行为，提前发现并修正安全漏洞，优化决策过程。如何设计安全框架和工具，以及如何将数字孪生技术与大模型安全策略相结合还需进一步研究。

（3）构建多维度人类偏好数据

构建多维度人类偏好数据对大语言模型的对齐至关重要，但目前这一领域的研究仍面临诸多挑战。包括偏好的多样性和复杂性导致的数据收集难题、隐私和伦理问题、代表性和偏见问题，以及现有数据集缺乏动态更新机制。研究者需要开发能够自我调整和持续更新的数据集，以适应人类偏好的演变。这是一个跨学科挑战，需要技术、伦理和社会科学领域的深入合作。期待开源社区推出更加精确、全面且动态的人类偏好数据集，推动人工智能向更智能、可靠的方向发展。

（4）可扩展监督

可扩展监督作为一种有效的对齐方法，允许人类在特定领域对超越人类能力的大语言模型进行有效监督。它不仅可以提升大模型的性能和应用范围，还可以确保大模型在日益复杂的应用环境中的安全性和可靠性。然而，目前这一领域的研究尚处于初步阶段，现有的方法多数基于早期模型，且许多提议的有效性还未经过充分验证。如何设计新的可扩展监督方法并使其能够与日益复杂和强大的大模型相适应还需要更加细致深入的研究。

5 结束语

大语言模型在为人类带来便利和前所未有的智能服务的同时，也带来了一系列的问题和挑战，尤其是在隐私保护、伦理规范和安全性方面。本文着眼于大模型对齐的相关研究，围绕数据收集、模型调整和模型评估展开探讨。本文概述了指令数据的收集方法，包括由NLP基准转化、人类手工标注和利用大语言模型生成数据。总结了大语言模型监督调整和对齐调整的主流训练方法，介绍了对齐调整的常用方法，包括在线偏好训练、离线偏好训练以及可扩展监督方法。回顾了模型评估的主要内容，基准和数据集，并重点探讨了模型评估方法。最后本文展望了大语言模型对齐未来的研究方向，为未来的研究提供了有价值的思路和方向。本文旨在综述大语言模型对齐技术研究现状，促进大语言模型对齐研究的发展。

参考文献：

[1] OPENAI. Introducing ChatGPT[EB]. 2023.

[2] BUBECK S, CHANDRASEKARAN V, ELDAN R, et al. Sparks of artificial general intelligence: early experiments with GPT-4[J]. arXiv Preprint, arXiv: 2303.12712, 2023.

[3] OPENAI. GPT-4 technical report[J]. arXiv Preprint, arXiv: 2303.08774, 2023.

[4] 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.

[5] BACH S H, SANH V, YONG Z X, et al. Promptsource: an integrated development environment and repository for natural language prompts[J]. arXiv Preprint, arXiv: 2202.01279, 2022.

[6] WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv Preprint, arXiv:2109.01652, 2021.

[7] LONGPRE S, HOU L, VU T, et al. The flan collection: designing data and methods for effective instruction tuning[J]. arXiv Preprint, arXiv:2301.13688, 2023.

[8] WANG Y Z, MISHRA S, ALIPOORMOLABASHI P, et al. Super-NaturalInstructions: generalization via declarative instructions on 1600+ NLP tasks[J]. arXiv Preprint, arXiv: 2204.07705, 2022.

[9] MISHRA S, KHASHABI D, BARAL C, et al. Cross-task generalization via natural language crowdsourcing instructions[J]. arXiv Preprint, arXiv: 2104.08773, 2021.

[10] ZHANG G, SHI Y M, LIU R B, et al. Chinese open instruction generalist: a preliminary release[J]. arXiv Preprint, arXiv: 2304.07987, 2023.

[11] CONOVER M, HAYES M, MATHUR A, et al. Free dolly: introducing the world's first truly open instruction-tuned llm[EB]. 2023.

[12] KÖPF A, KILCHER Y, VON RÜTTE D, et al. OpenAssistant conversations: democratizing large language model alignment[J]. arXiv Preprint, arXiv:2304.07327, 2023.

[13] CHIANG W L, LI Z, LIN Z, et al. Vicuna: an open-source chatbot impressing gpt-4 with 90%* chatgpt quality[EB]. 2023.

[14] DING N, CHEN Y L, XU B K, et al. Enhancing chat language models by scaling high-quality instructional conversations[J]. arXiv Preprint, arXiv:2305.14233, 2023.

[15] XU C W, GUO D Y, DUAN N, et al. Baize: an open-source chat model with parameter-efficient tuning on self-chat data[J]. arXiv Preprint, arXiv:2304.01196, 2023.

[16] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020(33): 1877-1901.

[17] JENTZSCH S, KERSTING K. ChatGPT is fun, but it is not funny! humor is still challenging large language models[J]. arXiv Preprint, arXiv: 2306.04563, 2023.

[18] WANG Y, KORDI Y, MISHRA S, et al. Self-instruct: aligning language model with self generated instructions[J]. arXiv Preprint, arXiv:2212.10560, 2022.

[19] TAORI R, GULRAJANI I, ZHANG T, et al. Stanford alpaca: an instruction-following llama model[EB]. 2023.

[20] CUI Y M, YANG Z Q, YAO X. Efficient and effective text encoding for Chinese LLaMA and alpaca[J]. arXiv Preprint, arXiv: 2304.08177, 2023.

[21] WU M H, WAHEED A, ZHANG C Y, et al. LaMini-LM: a diverse herd of distilled models from large-scale instructions[J]. arXiv Preprint, arXiv: 2304.14402, 2023.

[22] XU C, SUN Q F, ZHENG K, et al. WizardLM: empowering large language models to follow complex instruction[J]. arXiv Preprint, arXiv: 2304.12244, 2023.

[23] ZHENG L M, CHIANG W L, SHENG Y, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena[J]. arXiv Preprint, arXiv: 2306.05685, 2023.

[24] DUBOIS Y, LI X C, TAORI R, et al. AlpacaFarm: a simulation

framework for methods that learn from human feedback[J]. arXiv Preprint, arXiv: 2305.14387, 2023.

[25] STIENNON N, OUYANG L, WU J, et al. Learning to summarize with human feedback[J]. Advances in Neural Information Processing Systems, 2020(33): 3008-3021.

[26] NAKANO R, HILTON J, BALAJI S, et al. WebGPT: browser-assisted question-answering with human feedback[J]. arXiv Preprint, arXiv: 2112.09332, 2021.

[27] BAI Y T, JONES A, NDOUSSE K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv Preprint, arXiv: 2204.05862, 2022.

[28] FAN A, JERNITE Y, PEREZ E, et al. ELI5: long form question answering[J]. arXiv Preprint, arXiv:1907.09190, 2019.

[29] GUO B Y, ZHANG X, WANG Z Y, et al. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection[J]. arXiv Preprint, arXiv: 2301.07597, 2023.

[30] LAMBERT N, TUNSTALL L, RAJANI N, et al. Huggingface h4 stack exchange preference dataset[EB]. 2023.

[31] CUI G Q, YUAN L F, DING N, et al. UltraFeedback: boosting language models with high-quality feedback[J]. arXiv Preprint, arXiv: 2310.01377, 2023.

[32] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[33] SARZYNSKA-WAWER J, WAWER A, PAWLAK A, et al. Detecting formal thought disorder by deep contextualized word representations[J]. Psychiatry Research, 2021, 304: 114135.

[34] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[35] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint, arXiv: 1810.04805, 2018.

[36] SIMONS J. The creator of ChatGPT thinks AI should be regulated[EB]. 2023.

[37] WEIDINGER L, UESATO J, RAUH M, et al. Taxonomy of risks posed by language models[C]//Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM Press, 2022: 214-229.

[38] TURNER A M, SMITH L, SHAH R, et al. Optimal policies tend to seek power[J]. arXiv Preprint, arXiv:1912.01683, 2019.

[39] PEREZ E, RINGER S, LUKOŠIŪTĖ K, et al. Discovering language model behaviors with model-written evaluations[J]. arXiv Preprint, arXiv: 2212.09251, 2022.

[40] BOSTROM N. Existential risk prevention as global priority[J]. Global Policy, 2013, 4(1): 15-31.

[41] BUCKNALL B S, DORI-HACOHEN S. Current and near-term AI as a potential existential risk factor[C]//Proceedings of the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM Press, 2022: 119-129.

[42] ORD T. The precipice: existential risk and the future of humanity[M]. Hachette Books, 2020.

[43] LEIKE J, KRUEGER D, EVERITT T, et al. Scalable agent alignment via reward modeling: a research direction[J]. arXiv Preprint, arXiv:1811.07871, 2018.

[44] SCHWARTZ S H, CIECIUCH J, VECCHIONE M, et al. Refining the theory of basic individual values[J]. Journal of Personality and Social Psychology, 2012, 103(4): 663-688.

[45] KENTON Z, EVERITT T, WEIDINGER L, et al. Alignment of language agents[J]. arXiv Preprint, arXiv: 2103.14659, 2021.

[46] ASKELL A, BAI Y T, CHEN A N, et al. A general language assistant as a laboratory for alignment[J]. arXiv Preprint, arXiv: 2112.00861, 2021.

[47] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. arXiv Preprint, arXiv: 1707.06347, 2017.

[48] GLAESE A, MCALEESE N, TRĘBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements[J]. arXiv Preprint, arXiv: 2209.14375, 2022.

[49] BAHETI A, LU X M, BRAHMAN F, et al. Leftover lunch: advantage-based offline reinforcement learning for language models[J]. arXiv Preprint, arXiv: 2305.14718, 2023.

[50] GO D, KORBAK T, KRUSZEWSKI G, et al. Aligning language models with preferences through f-divergence minimization[J]. arXiv Preprint, arXiv: 2302.08215, 2023.

[51] ZHU B, JIAO J, JORDAN M I. Principled reinforcement learning with human feedback from pairwise or $ K $-wise Comparisons[J]. arXiv Preprint, arXiv: 2301.11270, 2023.

[52] ZIEBART B D, MAAS A L, BAGNELL J A, et al. Maximum entropy inverse reinforcement learning[C]//Aaai. 2008(8): 1433-1438.

[53] HADFIELD-MENELL D, MILLI S, ABBEEL P, et al. Inverse reward design[J]. Advances in neural information processing systems, 2017, 30.

[54] LI Z N, XU T, ZHANG Y S, et al. ReMax: a simple, effective, and efficient reinforcement learning method for aligning large language models[J]. arXiv Preprint, arXiv: 2310.10505, 2023.

[55] WAN A, WALLACE E, SHEN S, et al. Poisoning language models during instruction tuning[J]. arXiv Preprint, arXiv: 2305.00944, 2023.

[56] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J].

Advances in Neural Information Processing Systems, 2022(35): 24824-24837.

[57] LEE H, PHATALE S, MANSOOR H, et al. RLAIF: scaling reinforcement learning from human feedback with AI feedback[J]. arXiv Preprint, arXiv: 2309.00267, 2023.

[58] ZHU B H, FRICK E, WU T H, et al. Starling-7B: improving LLM helpfulness & harmlessness with RLAIF[EB]. 2023.

[59] LI X, ZHANG T, DUBOIS Y, et al. AlpacaEval: an automatic evaluator of instruction-following models[EB]. 2023.

[60] SUN Z Q, SHEN Y K, ZHANG H X, et al. SALMON: self-alignment with instructable reward models[J]. arXiv Preprint, arXiv: 2310.05910, 2023.

[61] LIU R B, JIA C Y, ZHANG G, et al. Second thoughts are best: learning to re-align with human values from text edits[J]. Advances in Neural Information Processing Systems, 2022, 35: 181-196.

[62] KIM S, BAE S, SHIN J, et al. Aligning large language models through synthetic feedback[J]. arXiv Preprint, arXiv: 2305.13735, 2023.

[63] LI Z K, PENG B L, HE P C, et al. Guiding large language models via directional stimulus prompting[J]. arXiv Preprint, arXiv: 2302.11520, 2023.

[64] AKYÜREK A F, AKYÜREK E, MADAAN A, et al. RL4F: generating natural language feedback with reinforcement learning for repairing model outputs[J]. arXiv Preprint, arXiv: 2305.08844, 2023.

[65] CASPER S, DAVIES X, SHI C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback[J]. arXiv Preprint, arXiv: 2307.15217, 2023.

[66] LIU H, SFERRAZZA C, ABBEEL P. Chain of hindsight aligns language models with feedback[J]. arXiv Preprint, arXiv: 2302.02676, 2023, 3.

[67] DONG H Z, XIONG W, GOYAL D, et al. RAFT: reward ranked finetuning for generative foundation model alignment[J]. arXiv Preprint, arXiv: 2304.06767, 2023.

[68] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: your language model is secretly a reward model[J]. arXiv Preprint, arXiv: 2305.18290, 2023.

[69] SONG F F, YU B W, LI M H, et al. Preference ranking optimization for human alignment[J]. arXiv Preprint, arXiv: 2306.17492, 2023.

[70] ZIEGLER D M, STIENNON N, WU J, et al. Fine-tuning language models from human preferences[J]. arXiv Preprint, arXiv: 1909.08593, 2019.

[71] ZHAO Y, KHALMAN M, JOSHI R, et al. Calibrating sequence likelihood improves conditional language generation[J]. arXiv Preprint, arXiv: 2210.00045, 2022.

[72] LIU Y X, LIU P F, RADEV D, et al. BRIO: bringing order to abstractive summarization[J]. arXiv Preprint, arXiv: 2203.16804, 2022.

[73] EDUNOV S, OTT M, AULI M, et al. Classical structured prediction losses for sequence to sequence learning[J]. arXiv Preprint, arXiv: 1711.04956, 2017.

[74] ZHANG T Y, KISHORE V, WU F, et al. BERTScore: evaluating text generation with BERT[J]. arXiv Preprint, arXiv: 1904.09675, 2019.

[75] YUAN Z, YUAN H Y, TAN C Q, et al. Rrhf: rank responses to align language models with human feedback without tears[J]. arXiv Preprint, arXiv: 2304.05302, 2023.

[76] HONOVICH O, AHARONI R, HERZIG J, et al. TRUE: re-evaluating factual consistency evaluation[J]. arXiv Preprint, arXiv: 2204.04991, 2022.

[77] PEZESHKPOUR P. Measuring and modifying factual knowledge in large language models[J]. arXiv Preprint, arXiv: 2306.06264, 2023.

[78] WANG C X, CHENG S R, GUO Q P, et al. Evaluating open-QA evaluation[J]. arXiv Preprint, arXiv: 2305.12421, 2023.

[79] ZHA Y H, YANG Y C, LI R C, et al. AlignScore: evaluating factual consistency with a unified alignment function[J]. arXiv Preprint, arXiv: 2305.16739, 2023.

[80] RUDINGER R, NARADOWSKY J, LEONARD B, et al. Gender bias in coreference resolution[J]. arXiv Preprint, arXiv: 1804.09301, 2018.

[81] ZHAO J, WANG T, YATSKAR M, et al. Gender bias in coreference resolution: Evaluation and debiasing methods[J]. arXiv Preprint, arXiv: 1804.06876, 2018.

[82] LEVESQUE H J, DAVIS E, MORGENSTERN L. The Winograd schema challenge[C]//Proceedings of the Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. New York: ACM Press, 2012: 552-561.

[83] CAO Y T, DAUMÉ H III. Toward gender-inclusive coreference resolution: an analysis of gender and bias throughout the machine learning lifecycle[J]. Computational Linguistics, 2021, 47(3): 615-661.

[84] STANOVSKY G, SMITH N A, ZETTLEMOYER L. Evaluating gender bias in machine translation[J]. arXiv Preprint, arXiv: 1906.00591, 2019.

[85] ZHANG G, LI Y Z, WU Y Y, et al. CORGI-PM: a Chinese corpus for gender bias probing and mitigation[J]. arXiv Preprint, arXiv: 2301.00395, 2023.

[86] DHAMALA J, SUN T, KUMAR V, et al. BOLD: dataset and metrics for measuring biases in open-ended language generation[C]// Proceedings of the Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM Press, 2021: 862-872.

[87] SMITH E M, HALL M, KAMBADUR M, et al. “I’m sorry to hear that”: finding new biases in language models with a holistic descriptor dataset[C]//Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022: 9180-9211.

[88] COSTA-JUSSÀ M R, ANDREWS P, SMITH E, et al. Multilingual holistic bias: extending descriptors and patterns to unveil demographic biases in languages at scale[J]. arXiv Preprint, arXiv: 2305.13198, 2023.

[89] HUANG Y F, XIONG D Y. CBBQ: a Chinese bias benchmark dataset curated with human-AI collaboration for large language models[J]. arXiv Preprint, arXiv: 2306.16244, 2023.

[90] HENDRYCKS D, BURNS C, BASART S, et al. Aligning AI with shared human values[J]. arXiv Preprint, arXiv: 2008.02275, 2020.

[91] FORBES M, HWANG J D, SHWARTZ V, et al. Social chemistry 101: learning to reason about social and moral norms[J]. arXiv Preprint, arXiv: 2011.00620, 2020.

[92] CHEN Y, ZHOU Y L, ZHU S C, et al. Detecting offensive language in social media to protect adolescent online safety[C]// Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing. Piscataway: IEEE Press, 2012: 71-80.

[93] RAZAVI A H, INKPEN D, URITSKY S, et al. Offensive language detection using multi-level classification[C]//FARZINDAR A, KEŠELJ V. Canadian Conference on Artificial Intelligence. Berlin, Heidelberg: Springer, 2010: 16-27.

[94] WASEEM Z, HOVY D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter[C]//Proceedings of the NAACL Student Research Workshop. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 88-93.

[95] ROSS B, RIST M, CARBONELL G, et al. Measuring the reliability of hate speech annotations: the case of the European refugee crisis[J]. arXiv Preprint, arXiv: 1701.08118, 2017.

[96] WULCZYN E, THAIN N, DIXON L. Ex machina: personal attacks seen at scale[C]//Proceedings of the Proceedings of the 26th International Conference on World Wide Web. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2017: 1391-1399.

[97] XU J, JU D, LI M, et al. Recipes for safety in open-domain chatbots[J]. arXiv Preprint, arXiv: 2010.07079, 2020.

[98] GEHMAN S, GURURANGAN S, SAP M, et al. RealToxicityPrompts: evaluating neural toxic degeneration in language models[J]. arXiv Preprint, arXiv: 2009.11462, 2020.

[99] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB]. 2018.

[100] DENG J W, ZHOU J Y, SUN H, et al. COLD: a benchmark for Chinese offensive language detection[J]. arXiv Preprint, arXiv: 2201.06025, 2022.

[101] HENDRYCKS D, BURNS C, BASART S, et al. Measuring massive multitask language understanding[J]. arXiv Preprint, arXiv: 2009.03300, 2020.

[102] LI H N, ZHANG Y X, KOTO F, et al. CMMLU: measuring massive multitask language understanding in Chinese[J]. arXiv Preprint, arXiv: 2306.09212, 2023.

[103] HUANG Y Z, BAI Y Z, ZHU Z H, et al. C-eval: a multi-level multi-discipline Chinese evaluation suite for foundation models[J]. arXiv Preprint, arXiv: 2305.08322, 2023.

[104] LIU C, JIN R R, REN Y Q, et al. M3KE: a massive multi-level multi-subject knowledge evaluation benchmark for Chinese large language models[J]. arXiv Preprint, arXiv: 2305.10263, 2023.

[105] ZHONG W J, CUI R X, GUO Y D, et al. AGIEval: a human-centric benchmark for evaluating foundation models[J]. arXiv Preprint, arXiv: 2304.06364, 2023.

[106] COBBE K, KOSARAJU V, BAVARIAN M, et al. Training verifiers to solve math word problems[J]. arXiv Preprint, arXiv: 2110.14168, 2021.

[107] TALMOR A, HERZIG J, LOURIE N, et al. CommonsenseQA: a question answering challenge targeting commonsense knowledge[J]. arXiv Preprint, arXiv: 1811.00937, 2018.

[108] GEVA M, KHASHABI D, SEGAL E, et al. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies[J]. Transactions of the Association for Computational Linguistics, 2021(9): 346-361.

[109] SUZGUN M, SCALES N, SCHÄRLI N, et al. Challenging big-bench tasks and whether chain-of-thought can solve them[J].

arXiv Preprint, arXiv: 2210.09261, 2022.
[110] CHEN M, TWOREK J, JUN H, et al. Evaluating large language models trained on code[J]. arXiv Preprint, arXiv: 2107.03374, 2021.
[111] LIU J W, XIA C S, WANG Y Y, et al. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation[J]. arXiv Preprint, arXiv: 2305.01210, 2023.
[112] AUSTIN J, ODENA A, NYE M, et al. Program synthesis with large language models[J]. arXiv Preprint, arXiv: 2108.07732, 2021.
[113] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Morristown, NJ, USA: Association for Computational Linguistics, 2001: 311-318.
[114] XU F, SONG Y, IYYER M, et al. A critical evaluation of evaluations for long-form question answering[J]. arXiv Preprint, arXiv: 2305.18201, 2023.
[115] FU J L, NG S K, JIANG Z B, et al. GPTScore: evaluate as you desire[J]. arXiv Preprint, arXiv: 2302.04166, 2023.
[116] GAO M Q, RUAN J, SUN R L, et al. Human-like summarization evaluation with ChatGPT[J]. arXiv Preprint, arXiv: 2304.02554, 2023.
[117] ZHUO T Y. Large language models are state-of-the-art evaluators of code generation[J]. arXiv Preprint, arXiv: 2304.14317, 2023.
[118] BAI Y S, YING J H, CAO Y X, et al. Benchmarking foundation models with language-model-as-an-examiner[J]. arXiv Preprint, arXiv: 2306.04181, 2023.
[119] LIN Y T, CHEN Y N. LLM-eval: unified multi-dimensional automatic evaluation for open-domain conversations with large language models[J]. arXiv Preprint, arXiv: 2305.13711, 2023.
[120] WANG P Y, LI L, CHEN L, et al. Large language models are not fair evaluators[J]. arXiv Preprint, arXiv: 2305.17926, 2023.
[121] WANG Y D, YU Z H, ZENG Z R, et al. PandaLM: an automatic evaluation benchmark for LLM instruction tuning optimization[J]. arXiv Preprint, arXiv: 2306.05087, 2023.
[122] ZIEMS C, HELD W, SHAIKH O, et al. Can large language models transform computational social science?[J]. arXiv Preprint, arXiv: 2305.03514, 2023.
[123] BANG Y J, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[J]. arXiv Preprint, arXiv: 2302.04023, 2023.

[作者简介]

刘昆麟（1997– ），男，博士，现就职于中兴通讯股份有限公司，主要研究方向为电信领域的大语言模型、大模型安全等。

屈新纪（1999– ），男，现就职于中兴通讯股份有限公司，主要研究方向为大语言模型等。

谭芳（1976– ），男，中兴通讯股份有限公司高级工程师，主要研究方向为大数据、大语言模型等。

康红辉（1972– ），男，中兴通讯股份有限公司无线网络智能化首席架构师，主要研究方向为未来网智、自智网络等。

赵少伟（1974– ），男，中兴通讯股份有限公司副总裁，主要研究方向为无线虚拟化平台、人工智能等。

施嵘（1973– ），男，中兴通讯股份有限公司副总裁，主要研究方向为无线网络产品研发、云和人工智能基础设施等。