CCL Frontier Forum 2024

The 23rd Chinese National Conference on Computational Linguistics

Proceedings of the Frontier Forum

August 24 - August 28, 2024

Taiyuan, China

Order copies of this and other CCL proceedings from:

Chinese National Conference on Computational Linguistics (CCL)

Courtyard 4, South Fourth Street, Zhongguancun , Haidian District, Beijing 100190, China

Tel: + 010-62562916

Fax: + 010-62661046

cips@iscas.ac.cn

Introduction

Welcome to the proceedings of the Frontier Forum of the twenty third China National Conference on Computational Linguistics (23rd CCL). The conference were held in Taiyuan, China.

CCL is an annual conference (bi-annual before 2013) that started in 1991. It is the flagship conference of the Chinese Information Processing Society of China (CIPS), which is the largest NLP scholar and expert community in China. CCL is a premier nation-wide conference for disseminating new scholarly and technological work in computational linguistics, with a major emphasis on computational processing of the languages in China such as Mandarin, Tibetan, Mongolian, and Uyghur

The Program Committee selected 10 overviews for the Frontier Forum of CCL 2024, in order to give a general overview of the progress in large language models in the past year and increase the sense of the edge-cutting works for the attendees. The 10 overviews encompass the compelling facets of large language models, including pre-training, alignment, reasoning, evaluation and applications.

We thank the Program and Organizing Committees for helping to make the forum successful, and we hope all the participants enjoyed the CCL conference and a wonderful days in Harbin.

Xin Zhao

July 2024

Organizers

Forum Chairs

Xin Zhao Renmin University of China, China

Table of Contents

从多模态预训练到多模态大模型：架构、训练、评测、趋势概览

李泽君[1†], **张霁雯**[1†], **王晔**[1†], **杜梦飞**[1], **刘晴雯**[1],
王殿仪[1], **吴斌浩**[1], **罗瑞璞**[1], **黄萱菁**[1], **魏忠钰**[1*]
[1]复旦大学
zywei@fudan.edu.cn

摘要

多媒体信息在人类社会的发展历程中有着至关重要的作用，构建具有多模态信息处理能力的智能系统也是通往通用人工智能的必经之路。随着预训练技术的发展以及对于通用模型的需求，多模态的研究也从早期的任务特定的方法转移到了构建统一泛用的多模态基座模型上。初步的统一多模态模型探索受到BERT启发，从表征学习的角度出发构建能为不同下游任务提供有效初始化的多模态预训练模型，这类方法尽管有效但仍然在泛用性方面受限于预训练-微调范式，无法更广泛高效地应用。近年来随着大语言模型的发展，以大语言模型为基座的多模态大模型则展现出了巨大的潜力：此类模型有着强大的信息感知，交互，以及推理能力并且能有效泛化到多样的场景下，为新时代的通用人工智能系统提供了切实可行的思路。本文将从构建统一多模态模型的角度出发，介绍和梳理相关工作的发展，从多模态预训练到多模态大模型，介绍对应的架构，训练，评测方法以及发展趋势，为读者提供一个全面的概览。

关键词： 多模态学习；多模态预训练；多模态大模型

From Multi-Modal Pre-Training to Multi-Modal Large Language Models: An Overview of Architectures, Training, Evaluation, and Trends

Zejun Li[1*†], **Jiwen Zhang**[1*†], **Ye Wang**[1*†], **Mengfei Du**[1], **Qingwen Liu**[1],
Dianyi Wang[1], **Binhao Wu**[1], **Ruipu Luo**[1], **Xuanjing Huang**[1], **Zhongyu Wei**[1*]
[1]Fudan University
zywei@fudan.edu.cn

Abstract

Multimedia information has played a crucial role in the development of human society, and constructing intelligent systems with multi-modal information processing capabilities is an essential pathway towards achieving Artificial General Intelligence (AGI). With the advancement of pre-training techniques and the growing demand for general models, research in multi-modality has shifted from early task-specific approaches to constructing unified and generalizable multi-modal foundation models. Early explorations of unified multi-modal models were inspired by BERT and focused on representation learning to create multi-modal pre-trained models that provide effective initialization for various downstream tasks. However, these methods are still limited in generalization by the pre-training and fine-tuning paradigm, making them less efficient for broad applicability. In recent years, the development of Large Language

[†]共同一作
[*]通讯作者

第二十三届中国计算语言学大会论文集，第1页-第33页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

Models (LLMs) has shown immense potential for Multi-modal Large Language Models (MLLMs). These models possess strong capabilities in perception, interaction, and reasoning, and can generalize effectively to diverse scenarios, offering a feasible approach to building next-generation AGI systems. This paper will provide an overview of constructing unified multi-modal models, introducing and reviewing the development from multi-modal pre-training to MLLMs. It will cover corresponding architectures, training paradigms, evaluation methods, and development trends, offering readers a comprehensive overview.

Keywords: Multimodal Learning , Multimodal Pre-training , Multimodal Large Language Model

1 引言

在人类社会的发展过程中，信息的交流与表达形式不断丰富和演变。从最初的语言交流到文字记录，再到现代的多媒体传播，多模态信息处理能力已成为现代社会不可或缺的一部分。随着信息技术的飞速发展，人们对于能够理解和生成结合文本、图像、声音等多种模态信息的智能系统的需求日益增长。多模态研究的重要性不仅体现在提高机器的交互能力，更在于推动社会信息交流方式的革新，促进知识的传播和文明的交流，构建具有多模态感知交互能力的智能体也是通往通用人工智能的必经之路 (Gan et al., 2022; Cui et al., 2024; Jin et al., 2024; Zhang et al., 2024)。

早期的多模态研究主要聚焦于针对特定的任务构建特定的模型，比如针对图片字幕生成任务 (Xu et al., 2015; Anderson et al., 2018a; Fan et al., 2019; Fan et al., 2021)，视觉问答任务 (Antol et al., 2015; Yu et al., 2019)，图文检索设计的架构 (Lee et al., 2018; Fan et al., 2022; Li et al., 2021b)等等。这样的方法往往会引入任务特定，甚至是数据集特定的归纳偏置，难以适应新的任务或泛化到不同的应用场景上。这限制了多模态技术在更广泛领域的应用潜力，也进一步启发了关于如何构建统一且泛用的多模态模型的研究。

受启发与BERT为代表的文本预训练模型 (Devlin et al., 2019; Liu et al., 2019)，最早对于统一多模态模型的探索聚焦在构建多模态预训练模型上。从表征学习的角度出发，利用大量的视觉文本数据进行自监督学习，在此基础上预训练模型提供的跨模态对齐的表示能够为不同的任务提供有效的初始化 (Li et al., 2019; Su et al., 2020; Chen et al., 2020b; Tan and Bansal, 2019; Li et al., 2021a; Radford et al., 2021; Li et al., 2022a)。然而预训练模型往往需要在特定任务上的微调才能进行实际的应用，并且很难有效地建立起任务之间地相关性，进而几乎无法以零样本的方式泛化到新的任务形式上。

随着ChatGPT为代表的大语言模型的出现 (OpenAI, 2023a; Touvron et al., 2023a; Touvron et al., 2023b)，研究者们发现指令微调后的大语言模型能够具有强大的对话，推理能力，并且能泛化到不同场景、任务下的不同指令 (Ouyang et al., 2022a; Chung et al., 2022; Chiang et al., 2023)，这为构建通用人工智能系统提供了切实可行的思路。多模态大模型的研究则尝试将大语言模型的成功迁移到多模态的领域下，通过给语言基座模型扩充了多个模态的编码器 (Girdhar et al., 2023; Radford et al., 2021; Zhai et al., 2023b)，并以高效的训练将其他模态的信号对齐到语言基座中 (Liu et al., 2024c; Li et al., 2023d; Bai et al., 2023b; Zhu et al., 2023a)。不同于预训练模型，得益于大语言模型的发展，多模态大模型的参数规模得益有效地扩展；除此之外，多模态大模型同样使用指令微调的方法进行训练，以自然语言的方式构建起了任务间的关联，进一步能泛化到不同的指令和场景。目前的多模态大模型能够准确地感知其他模态的信息，并根据多模态的语境和用户以自然语言进行交互，理解用户的需求并完成对应的任务 (Chen et al., 2023b; Zhao et al., 2023b; Liu et al., 2023c; Young et al., 2024; Laurençon et al., 2024; Lu et al., 2024)。近来，多模态大模型在各方面的应用和发展也展现了其成为统一化的多模态信息处理基座的潜力 (Li et al., 2023j; Chen et al., 2023e; Wu et al., 2023a; Zheng et al., 2023)。

本文将从构建统一多模态模型的角度出发，介绍和梳理相关工作的发展。在第 2 节中我们将首先介绍早期的多模态预训练的方法。接下来，我们将聚焦于热门的多模态大模型，从多模

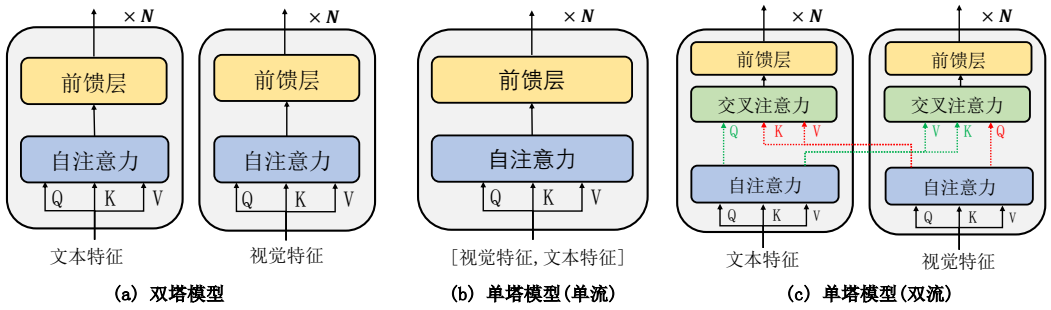


Figure 1: 常见的VLP模型架构。其中的Q，K，V分别表示注意力层的Query，Key，Value输入。“[视觉特征，文本特征]”表示两种特征拼接后得到的特征。

态大模型的架构（第 3 节），多模态大模型的训练方法（第 4 节）以及多模态大模型的评估方法（第 5 节）出发进行介绍。最后，本文将在第 6 节探讨多模态大模型的新发展方向，为读者提供一个全面的研究概览。

2 多模态预训练模型

在ChatGPT (OpenAI, 2023a) 引领的大模型风潮到来之前，多模态领域的研究者主要聚焦于多模态预训练技术(Vision-Language Pre-Training，简称VLP)：尝试通过预训练的方式构建统一、泛用的多模态表征模型。通过将大量的视觉文本数据输入到基座的多模态表征模型中，并通过自监督的预训练任务帮助模型学习跨模态的信息交互，所得到的预训练模型能够为下游多样化的多模态任务提供有效的初始化表示 (Tan and Bansal, 2019; Li et al., 2020c; Radford et al., 2021; Li et al., 2022c; Wang et al., 2022c; Bao et al., 2022a; Chen et al., 2023d)。

2.1 模型架构

受启发自以BERT (Devlin et al., 2019; Liu et al., 2019)为代表的文本预训练模型，多模态预训练的架构也主要构建在多层Transformer (Vaswani et al., 2017)的基础上，通过将视觉文本数据构建成序列，并对多模态序列学习语境化的表示。

2.1.1 多模态输入表示

文本表示 对于文本的输入，多数VLP模型遵循BERT的范式 (Devlin et al., 2019)，通过将文本进行分词，并在其开头和结尾补充“[CLS]”和“[SEP]”的特殊字符来表示文本的开头和结束。所得到的字符序列通过可学习的词嵌入层得到连续化的向量表示。

视觉表示 对于视觉输入，则需要额外的视觉编码器来对输入信号进行编码。早期的VLP方法通过物体检测模型来构建物体级别的视觉token，每个token的表示为对应物体所在区域的RoI (Region-of-Interest)的特征 (Zhang et al., 2021b; Chen et al., 2020b; Li et al., 2020c; Tan and Bansal, 2019)。其中最常用的检测器为Anderson et al. (2018a)训练的Faster-RCNN (Ren et al., 2015)。由于物体检测器的参数无法在后续的预训练过程中得到更新，后续的工作专注于构建端到端的VLP模型 (Li et al., 2020a; Xu et al., 2021; Huang et al., 2020)，其中最常使用的编码器为CNN类模型 (He et al., 2016)，以及视觉Transformer类模型 (Dosovitskiy et al., 2021; Liu et al., 2021)。这类视觉编码器得到的二维网格特征会被展开成一维，得到视觉的表示特征。视频则可以看作是多张图片构成的序列，并可以进一步用三维的模型建模时序的信息 (Feichtenhofer et al., 2019; Zhang et al., 2018; Carreira and Zisserman, 2017)。

2.1.2 多模态交互的建模方式

基于第 2.1.1 节中构建的多模态序列，VLP模型以不同的架构来建模模态间和模态内的信息交互，如图 1 所示，大致可以划分为如下几类：

双塔模型 双塔模型通过浅交互的方式来建模多模态的交互。视觉特征序列和文本序列分别被单独的视觉编码器和文本编码器进行编码，最终通过对比学习的方式进行整体的图片和文本间的对齐。最具代表性的工作有CLIP (Radford et al., 2021)和ALIGN (Jia et al., 2021)。单塔模

型能够高效地以内积的形式计算多张图片和文本之间的语义相似度，在跨模态检索任务上效率非常高，但是无法解决需要复杂的跨模态理解能力的任务，比如视觉语言问答。

单塔模型 单塔模型则希望构建视觉和文本token之间的深交互，主要利用Transformer中的注意力机制来进行建模。其中如图 1 (b, c) 所示，单流模型将视觉和文本序列拼接到一起，通过自注意力层学习模态内和模态间的交互 (Li et al., 2019; Chen et al., 2020b; Su et al., 2020; Li et al., 2020c; Kim et al., 2021)；而双流模型则解耦了模态内和模态间的表示，单独用交叉注意力来建模后者 (Lu et al., 2019; Tan and Bansal, 2019; Yu et al., 2021; Dou et al., 2022b)。单塔模型具有很强的跨模态建模和推理能力，但是在运算效率上稍逊于双塔模型。

其他架构 除了上述两种常规的架构，为了结合单塔和双塔结构的优点，研究者提出了融合型的架构，ALBEF (Li et al., 2021a)，FLAVA (Singh et al., 2022)和CoCa (Yu et al., 2022)在双塔模型的上层引入深交互的单塔融合编码器 (Xu et al., 2023c)；BLIP (Li et al., 2022b)，FIBER (Dou et al., 2022a)和VLMO (Bao et al., 2022a)等则设计了动态的交叉注意力或前馈层，使得同一模型可以支持多种形式的建模并且支持参数的高效共享 (Wang et al., 2021b)。为了支持文本生成的输出形式，目前的工作也探究了解码器 (Wang et al., 2022a)，编码器-解码器 (Cho et al., 2021; Wang et al., 2022c; Chen et al., 2023d)的VLP的架构。

2.2 训练方法

2.2.1 预训练数据集

多模态预训练要求大量视觉文本数据，图文对是最常见的形式，从高质量的标注数据，包括MSCOCO (Lin et al., 2014)，Flickr30K (Young et al., 2014)和Visual Genome (Krishna et al., 2017)，到网络爬取的图文对，CC3M (Sharma et al., 2018)，SBU (Ordonez et al., 2011)，CC12M (Changpinyo et al., 2021)，RedCaps (Desai et al., 2021)，再到更大规模的LAION (Schuhmann et al., 2021)等等。对于视频文本数据，则常用的数据为Kinetics-400 (Kay et al., 2017)，HowTo100M (Miech et al., 2019)以及WebVid-2M (Bain et al., 2021)。一般而言，数据集的规模和数据质量之间存在负相关的关系，需要综合考量。

2.2.2 预训练任务

多模态预训练任务主要为自监督的任务，通过不同的形式学习视觉文本的相关联系。

遮盖恢复建模 此类任务受启发于BERT中提出的遮盖语言模型（Masked Language Modeling，简称MLM）任务 (Devlin et al., 2019)，随机遮盖部分输入序列，要求模型预测原本的信息来学习语境的建模能力。对于文本，MLM也是VLP方法最常用的预训练任务 (Su et al., 2020; Li et al., 2019)。对于视觉信号的遮盖，基于物体检测输入的模型训练目标为预测被遮盖物体的类别和输入特征 (Chen et al., 2020b; Tan and Bansal, 2019)；SOHO (Huang et al., 2021)，VL-BEiT (Bao et al., 2022b)和BEiT-3 (Wang et al., 2023e)则利用视觉字典的方法给图片块构建了离散的表示，使得其也可以和MLM一样以分类的形式训练。

文本生成任务 不同于MLM进行双向的语境建模，文本生成任务要求模型进行单向的逐词的生成 (Li et al., 2022a; Wang et al., 2022a; Yu et al., 2022)。不同于MLM任务在每个样本里仅学习被遮盖部分的信息，文本生成任务可以利用每个token的信息，训练效率更高。Wang et al. (2022c)提出PrefixLM来让模型同时学习双向和单向的语境信息。VL-BART (Cho et al., 2021)和OFA (Wang et al., 2022b)进一步通过文本生成的形式统一表示多种任务。

图文对齐任务 用于帮助模型对齐模态间的表示，主要的任务有图文对比学习和图文匹配。前者主要作用在单塔模型部分 (Radford et al., 2021; Li et al., 2021a)，使用InfoNCE损失 (van den Oord et al., 2019)在训练批次内进行对比学习；后者则作用在融合模型部分，要求模型进行图文对是否匹配的二分类判断，其中的负样本可以来自于随机的采样 (Chen et al., 2020b)，批次内的困难负样本 (Li et al., 2021a)，数据集的困难负样本 (Chen et al., 2020a)。除了全局的对齐，也有工作 (Chen et al., 2020b; Kim et al., 2021)尝试用最优传输的理论对齐局部的token。

其他训练目标 除了上述的优化目标，视频还可以利用时序的信息进行自监督训练，比如视频帧的排序建模任务 (Li et al., 2020b)。除此之外，视觉问答 (Tan and Bansal, 2019; Li et al., 2020c)，物体标记框预测 (Xu et al., 2021; Zeng et al., 2022)也常被作为训练的任务。

2.3 多模态预训练模型的评测

多数的VLP模型仅仅作为初始化的表征参数，需要经过下游的微调后才能进行评测，并且需要在预训练模型的基础上增加任务相关的层来满足任务所需的输出形式。常见的任务包括：（1）跨模态检索任务：要求模型给定某模态内的一个样本从另一模态的候选集选择与其最语义匹配的样本，包括图文检索 (Lin et al., 2014; Young et al., 2014)，视频检索 (Xu et al., 2016; Chen and Dolan, 2011)等等；（2）视觉推理任务：需要模型根据对应的问题和视觉信息进行推理，包括视觉问答（VQA） (Goyal et al., 2017; Hudson and Manning, 2019; Xu et al., 2017; Jang et al., 2017)，推理 (Xie et al., 2019; Suhr et al., 2017; Zellers et al., 2019)，等等；（3）视觉描述任务：要求模型用语言描述对应的视觉信号 (Agrawal et al., 2019; Wang et al., 2019)。

尽管VLP模型能够对多种下游任务提供良好的初始化表示，预训练微调的范式也有明显的局限性：模型需要微调才能进行应用，并且无法适应没见过的任务形式，极大地限制了模型的泛用性。而VL-BART (Cho et al., 2021)和OFA等模型 (Wang et al., 2022b)中通过文本描述不同任务的方法则展现出了巨大的潜力，也成为了本文后续论述的多模态大模型的雏形。

3 多模态大模型架构

随着以ChatGPT为代表的大模型时代到来，多模态序列建模逐渐成为研究热点。当今的多模态大模型主要关注于多模态内容的理解和文本生成。这些模型通常将输入的模态信号进行抽象编码，形成一系列模态特征标记的序列表示，随后传递到大语言模型中进行统一处理，并解码生成文本。本章将从多模态序列表示开始，逐步介绍多模态大模型的具体架构设计以及架构优化方案。

3.1 多模态序列表示

一般当下多模态大模型主要关注多模态内容理解和文本生成。通常将输入的模态信号进行抽象编码成一系列模态特征标记的序列表示，然后再传递到大语言模型中进行统一的处理并解码出响应文本。建模数据格式为:Text + X− > Text。其中输入数据为文本和多模态内容对X(可以是图像文本对，音频文本对，视频文本对等)，输出是文本响应。

- 文本序列表示:按照语言模型的词表将输入文本分成一系列的文本标记。
- 图像序列表示:将输入图片切成M ∗ M的图像补丁通过视觉编码器变成一系列视觉标记。
- 视频序列表示:如果输入是视频，和图像方法一致，将视频按帧抽出图片表示，将每帧图片切成M ∗ M的图像补丁通过视觉编码器变成一系列视觉标记。并将每帧所有标记进行拼接。此外，会额外加一个表征时序信息的序列标记。而Valley(Luo et al., 2023b)则在每帧的视觉标记上作时间维度的平均，来保留空间信息。
- 音频序列表示:使用固定大小窗口进行滑动，将窗口内音频信号通过编码器进行编码处理，每个窗口的编码结果作为一个输入标记，从而得到一系列音频的序列标记表示。

3.2 多模态大模型的基座模型

典型的多模态大模型的基座模型一般可以抽象为三个模块，即预训练的视觉编码器、预训练大语言模型和模态连接件。预训练的视觉编码器作用是将输入的视觉信号编码成抽象的特征表示，类比于人类的眼睛，用于接收和预处理视觉输入。而大语言模型作为中央大脑，管理接收到的输入模态信号并理解和执行推理。视觉输入特征和大语言模型的文本特征相差较大，难以被大语言模型直接理解处理，而模态连接件充当一个桥梁来连接视觉编码器和大语言模型，使不同模态信息得到对齐。一些多模态大模型还包括一个生成器来输出除文本之外的其他模态信息，具体请参阅第 6.2 节中的讨论。

3.2.1 视觉编码器

视觉编码器(Vision Encoder，简称VE)的作用是将原始的图像输入X_v进行编码压缩成紧凑的视觉特征表示F_v，公式如下:

$$F_v = \mathrm{VE}(X_v) \tag{1}$$

预训练视觉编码器一般为ViT(Dosovitskiy et al., 2020)架构。比如常用的CLIP (Radford et al., 2021)通过图像-文本对进行大规模预训练，将视觉编码器在语义上与文本特征实现了对齐，因此使用这种最初预对齐的编码器通过预训练与大语言模型进行对齐会更加容易。除了常用的CLIP(Radford et al., 2021)视觉编码器，一些工作还探索了使用CLIP(Radford et al., 2021)的其他变体。MiniGPT-4(Zhu et al., 2023b) 、CogVLM(Wang et al., 2023d)采用EVA-CLIP (Sun et al., 2023b)编码器，该编码器在训练技术方面进行改进。SigLIP(Zhai et al., 2023b)改进了图像文本预训练损失来得到更好的对齐效果，被近期性能较强的多模态大模型广泛使用(Hu et al., 2023b; Lu et al., 2024; He et al., 2024; Laurençon et al., 2024; Team et al., 2024)。ImageBind(Girdhar et al., 2023)还将视觉、文本、音频和深度图等进行了对齐，扩展了输入模态的类别。

一些工作使用卷积架构作为视觉编码器，Osprey(Dehghani and Trojovskỳ, 2023)，ConvLLaVA(Ge et al., 2024)使用基于卷积的ConvNext-L 编码器(Liu et al., 2022b)来获取更高分辨率信息和多级特征，在相同分辨率下ConvNext(Liu et al., 2022b)生成的视觉特征标记更少，计算成本下降显著。此外，Fuyu-8b(Bavishi et al., 2023)探索了无编码器的结构，将图像块在输入到大语言模型之前直接投影，模型自然支持灵活的图像分辨率输入。

除了无编码器和单一编码器，近期一些工作还采用了双编码器设计 (Hong et al., 2023; Li et al., 2024)，使用两个编码器分别处理高分辨率图像和低分辨率图像。双编码器设计除了有助于支持高分辨外，还能够利用不同视觉编码器的归纳偏差进行互补，这有助于捕获更广泛的视觉表示，从而增强模型对视觉数据的理解。例如，Cobra(Zhao et al.,)和Deepseek-VL(Lu et al., 2024)分别采用了DINOv2(Oquab et al.,)、SAM-B(Kirillov et al., 2023)和SigLIP(Zhai et al., 2023b)的双编码器设计。这些模型通过结合DINOv2和SAM自监督学习的低级空间特征以及SigLIP通过弱监督提供的高级语义属性，实现了性能的互补增强。类似地，SPHINX-X(Gao et al., 2024)采用了DINOv2(Oquab et al.,)和CLIP-ConvNeXt(Cherti et al., 2023)的组合。Prismatic(Karamcheti et al., 2024)的研究进一步证明了SigLIP(Zhai et al., 2023b)和DINOv2(Oquab et al.,)可以实现最佳的互补效果。

3.2.2 大语言模型

通过对网络庞大语料库的自监督预训练，大语言模型嵌入了丰富的语言世界知识，拥有强大的泛化涌现、问答推理以及指令遵循能力。多模态大模型以大语言模型为骨干，大语言模型作为中央处理器统一接收视觉语言模态的特征入并统一解码推理，输出响应。如公式1,2。

目前，大多数的大语言模型属于因果解码器类别。LLAMA系列(Touvron et al., 2023a; Touvron et al., 2023b)和Vicuna(Chiang et al., 2023)是最具代表性的开源大语言模型，支撑了经典多模态大模型工作LLaVA系列(Liu et al., 2024c; Liu et al., 2023c; Liu et al., 2024b)、MiniGPT-4(Zhu et al., 2023b)等的研究。然而，这些模型主要在英语语料库上进行预训练，因此不支持多语言处理。双语大语言模型如Qwen(Bai et al., 2023a)和InternLM(Cai et al., 2024)则能很好地支持中文和英语，分别成为了Qwen-VL(Bai et al., 2023b)和多模态书生系列(Zhang et al., 2023a; Dong et al., 2024b)的大语言模型骨干。其他如Mistral(Jiang et al., 2023a)、Yi(Young et al., 2024)和Deepseek(Bi et al., 2024)等强大开源大语言模型也被用于构建新近推出的多模态模型，如Mini-Gemini(Li et al., 2024)、Yi-VL(Young et al., 2024)、LLaVA-next(Liu et al., 2024b)、IDEFICS2(Laurençon et al., 2024)和Deepseek-VL(Lu et al., 2024)等。

此外，对大语言模型专家混合体系结构(MoE)的研究也引起了广泛关注。与密集模型相比，MoE的稀疏架构可以通过选择性地激活部分参数，在不增加计算成本的情况下扩大模型总参数大小。在多模态大模型领域，如MM1(McKinzie et al., 2024)和MoE-LLaVA(Lin et al., 2024)将MoE机制引入并实现在多数基准测试中超越了相应的密集模型架构。Mini-Gemini(Li et al., 2024)拥有一个基于Mixtral 8x7B(MistralAITeam, 2023)构建的MoE版本。

同时，与动辄参数大小上百亿的多模态大模型相比，另一条路线是开发参数量更少、但性能不减的高效多模态大模型，这些模型通常使用参数少于3B的语言模型作为骨干。如MiniCPM-V系列(Hu et al., 2023b)和PaliGemma(Team et al., 2024)分别使用MiniCPM-2.4B(Hu et al., 2024b)和谷歌的Gemma-2B(Team et al., 2024)。而微软的Phi-3-V(Abdin et al., 2024)使用了拥有接近于Mixtral 8x7B (MistralAITeam, 2023)模型相当性能的Phi-3-mini(Abdin et al., 2024)。除了利用预训练的大语言模型外，MobileVLM系列(Chu et al., 2023; Chu et al., 2024)缩小了LLAMA(Touvron et al., 2023b)的参数尺寸，并使用开源数据集从头开始训练。

3.2.3 模态连接件

模态连接件(Projector)的作用是将视觉编码器输出的抽象视觉特征F_v映射到与大语言模型中的词嵌入语义空间具有相同维度的标记Z_v完成对齐，公式如下:

$$Z_v = \text{Projector}(F_v) \tag{2}$$

- **基于MLP架构**：多模态大模型的模态连接件通常使用可学习的线性层或多层感知器(MLP) 来实现。使用MLP架构的多模态大模型代表作是LLaVA系列(Liu et al., 2024c; Liu et al., 2023c; Liu et al., 2024b)。MLP架构的模态连接件的优点是可以较好的建模视觉特征的上下文信息，但缺点是视觉表示冗长，表征效率不高。

- **基于Attention架构**：BLIP-2(Li et al., 2023d)引入了Q-Former，作为一种轻量级的Transformer使用一组可学习的查询向量，从冻结的视觉编码器中提取视觉特征。后续Qwen-VL(Bai et al., 2023b)和MiniCPM-V(Hu et al., 2023b)系列同样使用了基于注意力架构的重采样器(Perceiver Resampler)作为连接件，可学习的查询向量作为Q，视觉编码器输出的图像特征为K和V，通过交叉注意力计算，输出视觉特征的聚合表示。

- **基于CNN架构**：MobileVLMv2(Chu et al., 2024)提出了LDPv2,一种基于卷积架构的新型模态连接件，包含特征变换、标记压缩和位置信息增强三个组件。通过使用逐点卷积层来匹配大语言模型的嵌入空间维度，平均池化层压缩视觉标记数量以及PEG模块来增强位置信息。较先前版本减少了99.8% 的参数量，实现更好的效率。

- **混合架构**：Honeybee(Cha et al., 2023)使用两个模态连接件，分别是C-Abstractor 和D-Abstractor。C-Abstractor为卷积架构由ResNet模块组成，专注于建模视觉特征的上下文，其中自适应平均池化有助于灵活管理视觉标记的数量。D-Abstractor利用可变形的注意力架构用于维护视觉特征的局部上下文。

另一条路线不仅仅使用模态连接件将视觉编码器和大语言模型进行模态对齐。而是在大语言模型内部插入额外参数模块以实现文本特征和视觉特征的交互融合。比如，Flamingo(Alayrac et al., 2022)、Open-Flamingo(Awadalla et al., 2023)和MMGPT(Gong et al., 2023) 在大语言模型内部层与层之间插入额外的交叉注意力层，将重采样器输出的视觉特征标记作为查询，与新插入到大语言模型中的层计算交叉注意力，从而将视觉信息注入到文本的生成过程中。CogVLM(Wang et al., 2023d)在每个Transformer 层中插入视觉专家模块，视觉专家模块包含QKV权重矩阵和FFN层，参数初始化来自预训练大语言模型，以实现视觉语言的双重交互融合。LLaMA-Adapter系列(Zhang et al., 2023b; Gao et al., 2023a)将可学习的提示引入Transformer 层。这些提示结合了视觉特征，作为前缀与文本特征进行拼接。

3.3 多模态大模型架构优化

尽管经典的多模态大模型已经能在各类通用任务上取得优异的表现 (Lin et al., 2014; Agrawal et al., 2015; He et al., 2015; Kafle and Kanan, 2017)，但在处理PDF文档、4K视频等对分辨率要求高、对时间/空间建模能力强的场景下仍然具有很大的不足 (Mathew et al., 2021; Li et al., 2023i; Ma et al., 2023; Liu et al., 2023f)。目前主流的开源模型，如LLaVA、QWen-VL等 (Liu et al., 2024c; Bai et al., 2023b)多数采用一个低分辨率的视觉编码器，其输入图像分辨率从224x224 (Dosovitskiy et al., 2021; Chen et al., 2020a; Sun et al., 2023b)、336x336 (如CLIP-ViT-Large-336 (Radford et al., 2021))到448×448 (如InternViT-1.2 (Chen et al., 2023e))不等。受限于视觉编码器的分辨率，常见的图像处理方案是：不论输入图片的原始分辨率是多少，统一缩放到与视觉编码器一致的分辨率，如LLaVA-1.5 (Liu et al., 2023c)、MiniGPT-4 (Zhu et al., 2023a)、EMU2 (Sun et al., 2023a)、InternLM-XComposer (Zhang et al., 2023a)等。显然，这种方案对信息的压缩程度过大，往往因缺失图像细节而降低模型的跨模态理解能力 (Li et al., 2023a; Ye et al., 2023a)，甚至导致幻觉 (Yu et al., 2023b)。

为了提高多模态模型的视觉编码能力，Monkey (Li et al., 2023j)等模型根据视觉编码器的分辨率大小将图像分片，每一片都由视觉编码器编码，然后拼接到一起作为图像的高分辨率特征。除了分片后的局部视图，通常还会有一个全局视图，即将原始图像放缩到视觉编码器对应

分辨率获得的低分辨率特征。全局低分辨率特征与局部高分辨率特征进行拼接，组成图片的完整特征表示。这就是多模态大模型的**多视图视觉表征**。这样的视觉表征方式虽然保留了更多的细粒度图像信息，但是拼接后的视觉序列过长，不但压缩了文本的表示能力，而且会造成训练困难 (Xu et al., 2024)。为了更有效率的进行图像特征表示，往往需要对图像特征序列进行压缩 (Ye et al., 2023a; Ye et al., 2023c; Yu et al., 2024)。最近的工作在图像切片的基础上衍生出了三种高分辨率的视觉编码方案，分别是：

1. **基于Resampler的视觉特征压缩**：得益于多模态预训练模型在早期对模型架构的探索，由Flamingo (Alayrac et al., 2022)所提出的Perceiver Resampler和BLIP (Li et al., 2022a)所提出的Q-Former架构天然地适用于视觉表征压缩。因此，以UReader (Ye et al., 2023a)、mPLUG-DocOwl 1.5 (Hu et al., 2024a)、InternLM-Xcomposer (Dong et al., 2024b)和TextHawk (Yu et al., 2024)为代表的模型采用一个与Q-Former结构相似的Resampler，在图像切片经过视觉编码器编码并拼接后进行特征压缩。Resampler通常是一个解码器结构，采用一组可学习的查询表示与视觉特征进行交叉注意力计算，以抽取高度聚合的视觉表征。训练往往冻结视觉编码器，只对Resampler进行微调。除了采用Resampler，以Monkey为代表的模型 (Li et al., 2023j; Liu et al., 2024d)在训练阶段还使用了低秩自适应的参数高效微调(LoRA)技术对视觉编码器进行参数更新，使得视觉编码器能够独立的建模不同图像切片的特征。

2. **基于双视觉编码器的多尺度特征融合**：利用多尺度图像信息允许模型捕获较小尺度中存在的细粒度细节和较大尺度中可用的全局上下文。CogAgent (Hong et al., 2023)、Mini-Gemini (Li et al., 2024)和DeepSeek-VL (Lu et al., 2024)都采用了这种方案，他们使用一个高分辨率视觉编码器用于高分辨率的全局视图处理，另一个低分辨率编码器用于低分辨率的局部视图处理。其中，Mini-Gemini (Li et al., 2024)提出了一个新颖的补丁信息挖掘(patch info mining)策略，使用低分辨率的视觉特征作为查询，通过交叉注意力从高分辨率的图像特征中检索相关的视觉线索。研究表明 (Shi et al., 2024)，一个多尺度较小模型的能力可以与一个较大模型相当，验证了多尺度方案的合理性。

3. **直接训练高分辨率视觉编码器**：以InternVL (Chen et al., 2023e)为代表的模型认为，多模态大模型受限的跨模态表征能力来源于不同模态之间参数的不平衡。因此，InternVL (Chen et al., 2023e; Chen et al., 2024b)系列模型将视觉编码器的参数量提高到6B，并采用MLP连接件来保留更多的视觉特征。InternViT-6B(Chen et al., 2023e)是拥有和大语言模型相同参数尺度的视觉编码器，通过对比学习直接和大语言模型进行对齐，来弥合视觉编码器和大语言模型之间参数尺度和特征表示能力的巨大差距。在后续的版本迭代中，InternViT-6B将图像切片的基础分辨率提高到448x448，并且能将任意分辨率的图片动态匹配到一个图像切片模版进行处理。值得一提的是，与此前的大部分模型不同，InternVL-1.5在预训练的第一阶段就解冻InternViT的参数，并持续使用高质量的图像文本数据对InternViT进行二阶段预训练，来增强模型对视觉信息的理解处理能力。在目前的开源多模态大模型中，InternVL-1.5展现了优异的性能，在多个指标上与代表性闭源大模型GPT4V (OpenAI, 2023b)相当，展现了该方案的潜力。

4 多模态大模型训练

4.1 预训练阶段

由于视觉编码器的输出空间与大语言模型的语义空间存在差异，需要利用多模态数据完成视觉文本特征之间的对齐。在预训练过程中，各个多模态大模型除了使用多种多样的多模态数据，还采取了不同的预训练策略，包含不同的预训练参数、预训练任务以及多预训练阶段。

4.1.1 预训练数据

多模态大模型的预训练数据根据数据的形式可以分为图片文本对、Grounded图片文本对、图文交错序列、表格/OCR数据、视频文本对、视频图文交错序列和仅文本语料。表 1 列出了不同的数据形式下常见的公开数据集的大小。在这些训练数据中，Taisu (Liu et al., 2022a)、Wukong Captions (Gu et al., 2022)和Youku-mPLUG (Xu et al., 2023a)的文本是中文

数据形式	数据集	图片/视频数	数据条数
图片文本对	MS-COCO (Lin et al., 2014)	110K	550K
	CC3M (Sharma et al., 2018)	2.9M	2.9M
	CC12M (Changpinyo et al., 2021)	11.1M	11.1M
	LAION-400M (Schuhmann et al., 2021)	400M	400M
	SBU Captions (Ordonez et al., 2011)	860K	860K
	VG Captions (Krishna et al., 2017)	100K	100K
	COYO-700M	580M	747M
	LAION-COCO	600M	600M
	CapsFusion-120M (Yu et al., 2023a)	120M	120M
	ShareGPT4V-Caption (Chen et al., 2023c)	1.3M	1.3M
	ALLaVA-Caption-4V (Chen et al., 2024a)	-	715K
	Wukong Captions (Gu et al., 2022)	100M	100M
	Taisu (Liu et al., 2022a)	166M	219M
Grounded图片文本对	GRIT (Peng et al., 2023)	90M	-
	Flickr30k Entities (Young et al., 2014)	30K	-
图文交错序列	MMC4 (Zhu et al., 2024)	571M	101.2M
	Wikihow (Yang et al., 2021)	772K	772K
	Wikipedia	-	-
表格、OCR数据	Char2Text (Kantharaj et al., 2022)	44K	44K
	UniChart (Masry et al., 2023)	611K	611K
	Paper2Fig100K (Rodriguez et al., 2023)	102K	102K
	Widget Captioning (Li et al., 2020d)	21K	162K
	Screen2Words (Wang et al., 2021a)	22K	112K
	TextOCR (Singh et al., 2021)	28K	903K
	COCO-Text (Veit et al., 2016)	63K	145K
视频文本对	InternVid (Wang et al., 2023f)	7.1M	234M
	Youku-mPLUG (Xu et al., 2023a)	10M	10M
	MSR-VTT (Xu et al., 2016)	7.2K	10K
	Webvid10M (Bain et al., 2021)	10.7M	10.7M
视频图文交错序列	YT-Storyboard-1B	20M	-
仅文本语料	Pile (Gao et al., 2020)	-	-

Table 1: 多模态大模型预训练数据

的。还有一些图片文本对格式的数据，利用了ChatGPT等大语言模型对文本进行了重写，如LAION-COCO、CapsFusion-120M (Yu et al., 2023a)、ShareGPT4V-Caption (Chen et al., 2023c)、ALLaVA-Caption-4V (Chen et al., 2024a)。除了多模态的训练数据外，还有模型使用仅文本语料进行预训练，防止大语言模型出现灾难性遗忘的情况。

4.1.2 预训练方法

多模态大模型的预训练阶段旨在对齐预训练模态编码器和预训练大语言模型的语义空间，我们从预训练参数、预训练任务和多预训练阶段三个角度进行介绍。

预训练参数 一部分多模态大模型预训练阶段只训练连接模块，如BLIP2 (Li et al., 2023d)、LLaVA (Liu et al., 2023d)、MiniGPT4 (Zhu et al., 2023a)等。DeepSeek-VL (Lu et al., 2024)训练连接模块和大语言模型。还有一部分模型预训练阶段会对全部参数进行训练，以Qwen-VL (Bai et al., 2023b)、Intern-VL (Chen et al., 2023e)为代表。

预训练任务 大部分模型在预训练阶段直接采用自回归语言建模任务对齐不同模态，如LLaVA (Liu et al., 2023d)、MiniGPT4 (Zhu et al., 2023a)等。为了更好的学习视觉表征中与文本有关的信息，BLIP2 (Li et al., 2023d)额外引入了图文对比学习、图文匹配任务，Intern-VL (Chen et al., 2023e)引入了对比学习。

多任务预训练阶段 Qwen-VL (Bai et al., 2023b)和MiniGPT-v2 (Chen et al., 2023b)根据训练数据的质量和数据形式将预训练阶段分为预训练和多任务预训练；Intern-VL (Chen et al., 2023e)根据预训练任务和数据质量将预训练阶段分为对比预训练和生成预训练；DeepSeek-VL (Lu et al., 2024)则根据训练参数将预训练阶段分为适配器预训练和联合预训练。

4.2 指令微调数据和微调方法

指令微调指的是用一系列带有指令和遵循指令的回答构成的文本对，对模型进行调优的一种方法。指令微调的目的是训练模型更好地理解用户的指令并遵循这个指令完成要求的任务。指令微调技术在大语言模型上的应用说明这样的方法能够帮助模型对齐人类的需求（以语言的形式描述任务），并泛化到新的场景和指令上 (Ouyang et al., 2022b; Chung et al., 2024; Chiang et al., 2023)。我们在这节中将介绍如何在多模态领域进行指令微调。

4.2.1 指令微调数据集的构建

由于指令微调数据在格式和任务的表述上有着很强的灵活性和多样性，使得收集这些数据往往比传统的有监督学习数据加困难且成本更高。在这一节里，我们将介绍三种主要的大规模收集指令微调数据的策略。

重构任务导向数据 目前存在着大量具有高质量标注的多模态数据集，但这些数据集大多用于特定的任务无法直接使用。因此很多研究工作 (Dai et al., 2023; Wang et al., 2024; Chen et al., 2023a; Xu et al., 2022; Zhang et al., 2023b; Gao et al., 2023a; Zhao et al., 2023b; Luo et al., 2023a)重构现有的高质量数据集为指令微调的形式。以经典的视觉问答任务（VQA）为例，原始的数据样本为一个输入输出对，其中输入包含一张图片和一个自然语言的问题，输出则是对应的文本回答。此类数据集的输入输出可以自然的构成指令微调阶段的输入输出。除了直接利用数据集中原始的输入外，指令微调的输入指令还可以通过手工设计或者GPT辅助生成。部分工作 (Dai et al., 2023; Xu et al., 2022; Zhu et al., 2023a; Chen et al., 2023a; Li et al., 2023c; Li et al., 2023f)针对同一个特定任务采用手工构建很多语义近似的指令组成候选池，然后在训练的时候随机选择池子中的一个作为该条数据的输入。其他一些工作 (Wang et al., 2024; Li et al., 2023e; Gong et al., 2023)通过手工设计一些种子指令，并使用这些种子指令来提示GPT生成更多的指令。现有的视觉问答（VQA）和图像描述（Image Captioning）数据集的答案通常较为简洁，直接使用这些数据集进行指令微调可能会让训练得到模型倾向于简短的输出。ChatBridge (Zhao et al., 2023b)中提到如果任务导向的原始数据回答很简短，在指令需要添加类似“in short” 这样的指示模型用简短回答的指令，而对于那些用一句话描述粗略答案则应该加上“a sentence” 等表示用一句话说完的语句。第二种方法扩展了原始回答的长度，例如M^{3}IT (Li et al., 2023f)提出通过使用原始问题、答案以及图片的背景信息（例如标题和OCR）来提示ChatGPT，从而扩展原始问题的答案。

闭源模型辅助 尽管现有的任务导向数据集能够提供丰富的数据资源，但它们通常无法很好地满足现实场景中的人类需求，如多轮对话等。部分研究通过更强大的闭源模型辅助来收集样本。这类方法利用现在市面上更强大的闭源LLMs（GPT4、Gemeni等），依据少量手工标注的样本，生成遵循指令的文本数据。具体而言，这些方法首先人工标注一些遵循指令的样例作为示范，随后引导GPT-4等闭源模型根据这些示范样例生成更多的指令样本。LLaVA (Liu et al., 2024c)将此方法扩展到多模态领域，通过将图像内容转换为标题和带有标识物体位置的检测框的文本，然后提示GPT-4生成新数据。基于此思路，后续研究如MiniGPT-4 (Zhu et al., 2023a)、ChatBridge (Zhao et al., 2023b)、GPT4Tools (Yang et al., 2023)及DetGPT (Pi et al., 2023)开发了适应不同需求的各类多模态指令微调数据集。而随着更为强大的多模态模型GPT-4V的发布，许多研究直接采用GPT-4V生成更高质量的数据，例如LVIS-Instruct4V (Wang et al., 2023a)和ALLaVA (Chen et al., 2024a)等。

文本指令数据 除了多模态指令微调数据之外，加入纯文本的指令数据同样能够用于增强对话技巧及指令遵循能力 (Gao et al., 2023b; Ye et al., 2023b; Gong et al., 2023; Luo et al., 2023a)。

4.2.2 指令微调方法

指令微调参数 相较于预训练阶段，指令微调阶段所引入的信息量更多，为了增大模型的容量，即可学习参数，除了少部分工作 (Li et al., 2023d; Dai et al., 2023; Zhu et al., 2023b)仅训练连接模块以外，多数模型会进一步训练语言模型基座，其中也会通过LoRA，Adapter等手段进行小参数的微调 (Zeng et al., 2023b; Chen et al., 2023b; Gong et al., 2023)。LLaVA-1.5则通过对比试验发现固定住语言基座可能会让模型无法适应新的指令和形式 (Liu et al., 2023c)。对

于视觉编码器部分，多数工作都将其冻结，也有部分工作训练视觉编码器来适应新的视觉建模方式或编码新的视觉信息 (Lu et al., 2024; Li et al., 2023j; Chen et al., 2023e)。

指令微调训练任务 与预训练阶段类似，指令微调阶段同样依靠自回归的语言建模任务来训练模型，不同于预训练阶段，多数模型以如下的形式来讲指令数据构造为完整的文本序列：

$X_{\text{system-message}}$
Human: X^1_{instruct}
Assistant: X^1_{response}
...
Human: X^n_{instruct}
Assistant: X^n_{response}

其中$X_{\text{system-message}}$为系统信息，用于给出对话的设定，$X^i_{\text{instruct}}$和$X^i_{\text{response}}$分别是第$i$轮对话的指令和回复，一共有$n$轮对话，具体的形式，包括对话的角色在不同的模型上设置不同，通常会和所使用的语言基座设定一致。其中只有红色部分的文本会用于计算损失。

4.3 基于人类反馈强化学习的方法

现有大语言模型基于交叉熵损失函数进行训练，虽然能够在一定程度上提升模型的泛化能力，但是没有显式引入人类偏好，与之进行对齐。因此，人类反馈强化学习（RLHF）(Stiennon et al., 2020; Ouyang et al., 2022a; Bai et al., 2022)引入人类反馈信号来对模型进行进一步优化。人类反馈学习的工作流程通常包含三个阶段：对预训练模型进行监督微调；创建奖励模型并使用人类标注数据训练；以及使用近端策略优化（PPO）(Schulman et al., 2017)通过奖励模型的奖励来优化策略模型。人类反馈强化学习作为大语言模型训练中一种强大并且可扩展的策略，也在最近被应用到了多模态大模型的领域中。

与纯文本领域不同，LLaVA-RLHF(Sun et al., 2023c)中策略模型与奖励模型都接受图片和文本作为输入。奖励模型初始化自基本的LLaVA模型(Liu et al., 2024c)，将最后一个词元的嵌入输出被线性投影为一个标量值作为输出的总体奖励。人类标注者通过比较两个由相同提示产生的回复选择出具有更少幻觉的答案作为人类偏好答案。该多模态大模型被训练以最大化由奖励模型模拟的人类奖励。除此之外，该文还提出了“事实增强RLHF”算法，即通过增加图像描述等额外的事实信息来校准奖励信号，从而在一定程度上缓解了可能出现的奖励作弊问题。ViGoR (Yan et al., 2024)设计了一个细粒度奖励模型，用于更新预训练的多模态语言模型，目标是改进视觉定位并减少幻觉现象。该模型结合了人类偏好和自动指标，通过众包收集细粒度的句子级反馈来获取用于训练奖励模型的人类判断和偏好。同时，它还利用先进的视觉感知模型来评估生成文本的定位和真实性，并在强化学习优化过程中结合成单一的奖励分数。

直接偏好优化（DPO）(Rafailov et al., 2024)作为一种新的优化方法已成为人类反馈强化学习的替代方案，这种方法直接优化策略模型以符合人类偏好，不需要创建奖励模型或使用强化学习进行优化。给定一个关于模型响应的人类偏好数据集，就可以通过直接偏好优化使用简单的二元交叉熵目标来优化策略。RLHF-V(Yu et al., 2023b)基于直接偏好优化方法提出了密集直接偏好优化（DDPO），直接根据密集和细粒度的段级偏好来优化策略模型；并且在数据层面上提供了细粒度的段级纠正形式的人类反馈数据集，以符合清晰密集且更为细粒度的人类偏好。HalDetect(Gunjal et al., 2024)提出用于检测幻觉内容的MHalDetect 数据集。该数据集涵盖了各种幻觉类型，包括不存在的物体、不真实的描述和不准确的关系。基于此数据集，HalDetect 训练了一个多模态奖励模型，并提出了细粒度直接偏好优化（FDPO）。细粒度直接偏好优化利用个别示例的细粒度偏好来增强模型区分准确描述的能力。

5 多模态大模型评测方法

多模态大模型的评测是推动该领域发展的关键环节。评测不仅为多模态大模型的持续优化提供了宝贵的反馈，还能帮助比较不同技术路径模型的性能差异。与传统多模态模型的评测方法相比，多模态大模型的评测呈现出几个新的特点：（1）由于多模态大模型具有强大的泛化能力，评估其综合性能变得尤为重要；（2）由于多模态大模型的输出形式为自由化文本，需要为这种灵活的输出设置稳定的自动化评估流程；（3）“幻觉”是大语言模型的常见问题，因此评估多模态大模型的幻觉问题也是评测过程中需要考虑的问题。现有的评测基准主要包含：以任务

为导向的基准、综合能力的基准以及评估幻觉的数据集准。根据任务类型的不同，也通常使用多样化的评估方法评测模型在相应数据集上的性能

5.1 评测数据集

多模态大模型评估数据集用于测试和比较不同的多模态大模型在各种任务上的性能。这些数据集，如VQAv2 (Goyal et al., 2017)和VSR (Liu et al., 2023d)，旨在模拟真实世界的视觉语言处理场景，涵盖视觉问答、场景文本识别、视觉推理和空间理解等多种任务。其次，鉴于大模型出色的综合表现能力，衡量综合能力的全面的评测数据集被提出，如MME (Fu et al., 2023)和MMBench (Liu et al., 2023e)。另外，大模型会过度依赖训练数据中的一些模式，从而导致严重的幻觉问题，许多评测幻觉的基准进而被提出，如CCEval (Zhai et al., 2023a)和CHAIR (Dai et al., 2022)。在这一节，我们整理了42个流行的基准，并将这些基准分为三类：面向任务的基准，综合的基准和幻觉的基准，如表 2所示。

5.1.1 面向任务的基准

早期的基准专注于解决特定任务的问题，这些基准拥有面向任务的形式和数据，往往会设计最适合某种任务的评测方式和评测指标。下面将按任务类型对这些基准分别进行介绍。

场景文本识别 旨在对场景图像作光学字符识别。TextVQA (Singh et al., 2019)要求识别出图片中的文字，从这些文字、图像信息中预测答案。DocVQA (Mathew et al., 2021)专注于考察对文档内容的理解。OCR-VQA (Mishra et al., 2019)则要求通过阅读图像中的文字进行可视化答题。

视觉推理 要求模型能够针对图像信息进行推理。VQAv2 (Goyal et al., 2017)减少了数据集中的语言偏见，强化了视觉图像在推理中的地位。GQA (Hudson and Manning, 2019)进一步扩大了数据集规模并涉及多步推理的问题。Whoops (Bitton-Guetta et al., 2023)创建故意违背常识的图像组成，并要求解释图像异常的原因。OK-VQA (Marino et al., 2019)需要依据外部知识作出回答。ScienceQA (Lu et al., 2022)是从小学和高中科学课程中收集的多模态科学选择题。VizWiz (Gurari et al., 2018)回答盲人的视觉问题，反映用户真实需求。ViQuAE (Lerner et al., 2022)包括使用知识库回答基于视觉上下文的命名实体的问题。A-OKVQA (Schwenk et al., 2022)是OK-VQA的增强，需要广泛的世界知识和常识来回答。

空间理解 考察对物体相对空间关系的理解。VSR (Liu et al., 2023d)涉及了66种空间关系用于评测空间推理能力。CLEVR (Johnson et al., 2017)包含100K渲染的图像，询问图像中的简单的3D形状的相关信息。EmbSpatial-Bench (Du et al., 2024)通过对3D具身场景的自动化探索，构建了一个评估大模型对6种以自身为中心的空间关系理解能力。

图文关系推理 要求从图像文本的关系中进行推理。WikiHow (Koupaee and Wang, 2018)包含一系列日常工作的任务图文步骤和文本概要。Winoground (Thrush et al., 2022)用于测量视觉-语言的组合推理，要求模型必须正确配对两个图片和对应的两个描述，两个描述的词汇构成完全相同，但顺序不同。SNLI-VE (Xie et al., 2019)要求预测图像和文本之间的关系是蕴涵、中性亦或矛盾的。MOCHEG (Yao et al., 2023)支持端到端的多模态事实检验和解释生成，输入是一条声明和大量网络资源，目的是评估该声明的真实性。

视觉描述 衡量对视觉内容进行生成描述的能力。TextCaps (Sidorov et al., 2020)要求识别文本并联系视觉上下文，在多个文本标记和视觉实体之间进行空间、语义和视觉推理。NoCaps (Agrawal et al., 2019)是针对新颖物体描述的大规模数据集，包含400种左右的新颖物体。Flickr30K (Young et al., 2014)基于30K图像和150K描述性标题的大型语料库构建表意图，进而构造视觉描述数据集。

视觉对话 包含VisDial (Das et al., 2017)数据集，其中智能体必须与人类就视觉内容进行对话。

引用表达 包含RefCOCO (Yu et al., 2016)数据集，要求模型能够在图像中定位到文本表达中的实体或者对图像区域内实体进行正确的文本描述。

5.1.2 综合的基准

多模态大模型旨在解决绝大多数任务。为此，现有的针对多模态大模型的基准往往会评估不同任务的性能，以期望全面综合地反映一个多模态大模型的表现。这些基准往往覆盖广阔范围的任务种类和数据内容，同时评测方式适应于多模态大模型自由化文本输出的形式，从而达到自动化评估的目的。

LAMM (Yin et al., 2023)旨在建设开源的多模态指令微调及评测框架，包含高度优化的训练框架、全面的评测体系，并支持多种视觉模态。MME (Fu et al., 2023)将问题构造成判断题的形式，并在4种任务和14种子任务上对大模型进行评测。LVLM-eHub (Xu et al., 2023b)由定量能力评估和在线互动评测平台组成，一方面，在47个标准视觉语言基准上定量评估大模型的视觉感知、视觉知识获取、视觉推理、视觉常识、对象幻觉和具身智能6类多模态能力，另一方面，搭建在线互动评测平台提供用户层面的模型排名。MMBench (Liu et al., 2023e)拥有自上而下的能力维度设计，根据定义的能力维度构造评测数据集，另外引入ChatGPT (OpenAI, 2023a)，以及提出了CircularEval的评测方式，使得评测的结果更加稳定。MMMU (Yue et al., 2023)包括11.5K来自大学考试、测验和教科书的多模态问题，涵盖30个学科和183个子领域，包括30种高度异构的图像类型。MM-Vet (Yu et al., 2023c)定义了6种核心视觉语言能力，并人工构造了包含200张图像和218个问题的开放问答式的评测基准，并使用GPT4对模型性能进行评估。ReForm-Eval (Li et al., 2023i)重构了现有的61个任务导向的视觉语言数据集将其转化为统一的生成和选择任务形式，并系统化地评估了模型地不稳定性。LLaVA-Bench-in-the-Wild收集了一组24张图像包括室内和室外场景、表情包、绘画、素描等，并人工构造了60个问题包含对话（简单问答）、详细描述和复杂推理三种问题类型。Seed-Bench (Li et al., 2023b)由19k由人类标注的多项选择题组成。评测基准涵盖12个不同的能力维度，包括图像和视频的理解能力。

5.1.3 面向幻觉的基准

与评估一般多模态大模型能力的基准不同，面向幻觉的基准主要针对生成内容中的幻觉判别以及模型的非幻觉生成。POPE (Li et al., 2023h)、NOPE (Lovenia et al., 2023) 和CIEM (Hu et al., 2023a)等任务主要关注模型生成文本中的物体幻觉，基准主要采用准确率作为评估指标。具体地，上述评测基准通过询问图像中是否存在物体并将模型响应与真实答案进行比较以计算模型对于图像中物体产生幻觉的情况。相较于于幻觉判别，生成类任务不仅能够评估模型的物体幻觉，也可以评估图像中的属性和关系幻觉 (Lovenia et al., 2023; Jing et al., 2023)。M-HalDetect (Gunjal et al., 2024) 包含16k条VQA数据的细粒度注释，使用人类评分和奖励模型分数评估模型输出包含幻觉的程度；GAVIE (Liu et al., 2023b)收集了来自Visual Genome (Krishna et al., 2017), VisText (Tang et al., 2023a)和Visual News (Liu et al., 2020)的图像，构造生成式的幻觉评测任务，并使用GPT4从准确性和相关性两方面评测模型能力；AMBER (Wang et al., 2023b)融合了生成性和判别性任务,并使用了一系列分类和生成指标评估模型生成的幻觉情况。

5.2 评测方法

现有的评测基准主要包括文本生成和选项选择这两种任务形式，而这两种任务形式分别对应不同的任务评估方法。

5.2.1 文本生成

文本生成类任务是评估多模态大模型性能最直接的形式，最适配大模型灵活的输出形式。一般包括OCR类任务、图片描述任务和视觉问答任务。

OCR类任务 主要评估多模态大模型识别图像中文本的能力，要求模型生成与图片中完全一致的目标词。在自动化评估的过程中，主要使用词语准确率（word accuracy） (Xu et al., 2023b; Liu et al., 2023g)进行评估：

$$\text{Accuracy}_{word} = \frac{\#\{\text{ predictions include target word }\}}{\#\{\text{ all predictions }\}}.$$

图像描述任务 主要评估多模态大模型对于图像内容的理解能力。图像描述任务的输出也可以用以评估多模态大模型产生幻觉的情况。具体地，针对一般的图像描述任务，主要

基准类别	任务类型	基准	图片来源	评估形式	评估指标
面向任务	场景文本识别	TextVQA (Singh et al., 2019)	OpenImages	生成	准确率
		DocVQA (Mathew et al., 2021)	网络	生成	准确率
		OCR-VQA (Mishra et al., 2019)	Amazon	生成	准确率
	视觉推理	VQAv2 (Goyal et al., 2017)	COCO	生成	准确率
		GQA (Hudson and Manning, 2019)	COCO, Flickr	生成	准确率
		Whoops (Bitton-Guetta et al., 2023)	自制	生成	匹配度，生成类指标
		OK-VQA (Marino et al., 2019)	COCO	生成	准确率
		ScienceQA (Lu et al., 2022)	网络	选择	准确率
		VizWiz (Gurari et al., 2018)	6个公开数据集	生成	准确率，生成类指标
		ViQuAE (Lerner et al., 2022)	网络	选择+生成	精确率，命中率，F1，精准匹配
		A-OKVQA (Schwenk et al., 2022)	COCO	选择+生成	准确率
	空间理解	VSR (Liu et al., 2023d)	COCO	选择	准确率
		CLEVR (Johnson et al., 2017)	自制	生成	准确率
	图文关系推理	WikiHow (Koupaee and Wang, 2018)	网络	生成	生成类指标
		Winoground (Thrush et al., 2022)	Getty Images	选择	文本分数，图像分数，组分数
		SNLI-VE (Xie et al., 2019)	Flickr30K	选择	准确率
		MOCHEG (Yao et al., 2023)	网络	选择+生成	精确率，召回率，F1，生成类指标
	视觉描述	TextCaps (Sidorov et al., 2020)	OpenImages	选择	生成类指标
		NoCaps (Agrawal et al., 2019)	OpenImages	生成	生成类指标
		Flickr30K (Young et al., 2014)	Flickr	生成	生成类指标
	视觉对话	VisDial (Das et al., 2017)	COCO	选择	召回率
	引用表达	RefCOCO (Yu et al., 2016)	COCO	生成	准确率，生成类指标
综合	综合	LAMM (Yin et al., 2023)	14个公开数据集	选择+生成	准确率，生成类指标
		MME (Fu et al., 2023)	7个公开数据集，自制	选择	准确率
		LVLM-eHub (Xu et al., 2023b)	47个公开数据集	选择	准确率
		MMBench (Liu et al., 2023e)	11个公开数据集，网络	选择	准确率
		MMMU (Yue et al., 2023)	教科书和网络	选择+生成	准确率
		MMVET (Yu et al., 2023c)	网络	生成	GPT-4分数
		ReForm-Eval (Li et al., 2023i)	现有的61个数据集	选择+生成	准确率，CIDEr
		LLaVA-Bench-in-the-Wild	网络	生成	GPT-4分数
		SEED-Bench (Li et al., 2023b)	CC3M	选择	准确率
幻觉	幻觉	CHAIR (Dai et al., 2022)	COCO, Open Images V4	生成	CHAIR
		CCEval (Zhai et al., 2023a)	VisualGenome	生成	CHAIR
		FAITHSCORE (Jing et al., 2023)	COCO	生成	FAITH
		GAVIE (Liu et al., 2023b)	Visual Genome, VisText, Visual News	生成	GPT4评分
		MMHal-Bench (Sun et al., 2023c)	OpenImages	生成	GPT4评分
		M-HalDetect (Gunjal et al., 2024)	COCO	生成	奖励模型分数+人工评分
		HaELM (Wang et al., 2023c)	COCO	生成	准确率
		POPE (Li et al., 2023h)	COCO	选择	准确率
		CIEM (Hu et al., 2023a)	COCO	选择	准确率
		NOPE (Lovenia et al., 2023)	10个VQA数据集	选择	准确率，METEOR
		AMBER (Wang et al., 2023b)	COCO, UnSplash	选择+生成	CHAIR, Cover ,Hal, Cog

Table 2: 多模态大模型评测数据集。生成类指标包括BLEU, ROUGE, METEOR, CIDEr等。

使用传统的图像描述任务的指标如CIDEr (Vedantam et al., 2015)、BLEU (Papineni et al., 2002)、METEOR (Banerjee and Lavie, 2005)和ROUGE-L(Lin, 2004)。针对评测多模态大模型的幻觉，一般使用CHAIR (Rohrbach et al., 2018)指标：

$$\text{CHAIR}_i = \frac{\#\{ \text{ hallucinated objects } \}}{\#\{ \text{ all objects in prediction } \}},$$

$$\text{CHAIR}_s = \frac{\#\{ \text{ hallucinated sentences } \}}{\#\{ \text{ all sentences } \}}.$$

除了CHAIR指标，AMBER (Wang et al., 2023b)进一步其他的幻觉评估指标：Cover，Hal，Cog以及AMBER分数。其中Cover分数主要衡量模型输出的物体类别覆盖真实物体类别的比例，理想的输出应该为最小化幻觉内容而不显著降低覆盖率。其公式为：

$$\text{Cover} = \frac{\#\{\text{objects in response} \bigcap \text{true objects } \}}{\#\{ \text{ true objects } \}}.$$

Hal分数表示出现幻觉的反应比例。对于模型的输出，如果$\text{CHAIR}_i \neq 0$，则认为模型出现幻觉。Hal分数的公式为：

$$\text{Hal} = \begin{cases} 1 & \text{if CHAIR}_i \neq 0 \\ 0 & \text{otherwise} \end{cases}.$$

Cog分数主要用于评估多模态大模型中的幻觉是否与人类认知中的幻觉相似。在给定人类幻觉目标物体集合（hallucinatory target objects）的情况下，Cog分数的公式为：

$$\text{Cog} = \frac{\#\{\text{objects in response} \bigcap \text{hallucinatory target objects } \}}{\#\{ \text{ objects in response } \}}.$$

为了全面评估各种多模态大模型在生成式和判别式任务下的表现，AMBER分数被提出，来整合生成任务上的CHAIR分数和判别任务上的F1分数。AMBER分数的公式为：

$$\text{AMBER Score } = \frac{1}{2} \times (1 - \text{ CHAIR } + F1).$$

视觉问答任务 主要形式是给定图片和任意问题，利用模型生成自由文本来回答问题。这类任务能够反映模型的综合能力 (Yue et al., 2023; Yu et al., 2023c)。然而，由于模型输出的形式自由，且答案逻辑较为复杂，通常很难计算准确率等指标。因此，在评测过程中通常采用人工评分，或者使用大语言模型如GPT-4 (OpenAI, 2023b)，通过对比人工标注的答案和模型生成的结果来评分，从而衡量模型输出的质量。

5.2.2 选项选择

选项选择是指任务所有可能的答案选项是预先定义好的，并且限制在一个有限的集合内。使用有限选项的选择题的评估形式能够极大增强评估结果的稳定性和可重复性。选项选择问题的形式包含单选题、判断题等，检索任务可以看作一种特殊的选择题形式。例如，在视觉问答中通过选择题的方式提供选项，要求模型输出正确的选项标记。

选项选择任务可以通过计算选项准确率评估模型的性能，也会计算扰乱选项后输出选项的准确率排除选项顺序对于模型输出的影响 (Fu et al., 2023; Liu et al., 2023e)。

由于模型输出格式的问题，因此在获取多模态模型的选择的选项过程中，一般有如下技术：（1）使用上下文样本来引导模型按期望的格式输出选项，再进行字符串匹配；（2）计算模型在给定图片文本下对于不同选项的生成概率，选择最高概率的选项作为模型选择题的答案 (Li et al., 2023i)；（3）针对图像生成类问题，通过计算生成图像和真实图像的CLIP相似分数，来获取模型选择题的答案 (Li et al., 2023b)；（4）通过计算模型对于各个选项的困惑度，选择困惑度最低的选项作为模型选择的答案 (Li et al., 2023b)。

6 多模态大模型的新趋势

6.1 解决多模态大模型的幻觉问题

“幻觉”被定义为生成无意义的内容或偏离其来源的内容(Ji et al., 2023)，而多模态大模型中的“幻觉”是指模型的视觉输入和文本输出之间的矛盾。从视觉语言任务的角度来看，当多模态大模型对用户的询问或陈述的响应与实际视觉数据不一致，即判断或描述具有缺陷时，则出现了幻觉(Liu et al., 2024a)。多模态大模型的幻觉来源于多种因素，包括来自数据的幻觉，即现有训练数据分布不平衡导致的数据偏差和不准确的标签或是注释；来自视觉编码器的幻觉，即有限的视觉分辨率和缺乏细粒度的视觉语义；来自多模态对其的幻觉，即连接模块结构过于简单和受限的标记约束；来自大语言模型的幻觉，即上下文注意力不足，随机采样解码以及能力错配等。

为解决多模态大模型的幻觉问题，可以从导致幻觉的因素出发。首先，减少幻觉的直接有效的方法是优化训练数据。具体的方法是引入负面数据和反事实数据，以增加数据集的多样性或是减少现有数据集中的噪声和错误，通过重写文本标注来提高数据集质量。对于数据偏差，CIEM(Hu et al., 2023a)利用现有的大语言模型从带有标注的图像文本数据集中生成对比的正负问答对，并用于对比指令微调。LRV-Instruction(Liu et al., 2023a)提出一个多样化大规模数据集，包含正确的视觉指令以及分别在三个不同的语义层面上对应的负指令。对于不准确的注释，构建丰富注释的数据集精确提取视觉内容和全面对齐模态也能够减轻多模态大模型的幻觉(Gunjal et al., 2024; You et al., 2023; Lu et al., 2023b)。

从视觉编码器改善多模态大模型幻觉的角度而言，利用支持更高分辨率的视觉编码器已经被广泛证明能够提升模型的视觉感知能力(Liu et al., 2023c; Bai et al., 2023b; Li et al., 2023j; Lu et al., 2024)。InternVL(Chen et al., 2023e)则进一步增大视觉编码器的参数规模，扩展至60亿参数。然而，现有的大多数多模态大模型使用ViT(Dosovitskiy et al., 2021)作为视觉编码器关注显著对象却忽略了一些关键的视觉线索。(Jain et al., 2023; Zhao et al., 2023a)引入额外的空间信息来引导多模态大模型处理用户查询，进一步增加了模型的对象级感知能力和空间感知能力。

对于视觉编码器和大语言模型之间的连接模块，连接模块的参数或结构影响了视觉文本之间的模态对齐，导致了幻觉的产生。最近一些工作开发了更强大的视觉语言连接模块：例

如，LLaVA-1.5将连接模块由单个线性层升级为多层感知机，实现了在各类多模态大模型评测数据集上的指标提升；InternVL(Chen et al., 2023e)利用LLaMA(Touvron et al., 2023a)构建了QLLaMA，由于QLLaMA使用预训练权重初始化，并且具有80 亿参数，因此在视觉语言对齐方面显著优于Q-Former(Li et al., 2023d)。

从大语言模型的角度来看，改进训练目标或是训练方法可以减少幻觉。(Chen et al., 2023f)引入辅助监督使用额外的标注信息来辅助监督模型关注图像内容。(Jiang et al., 2023b)通过对比损失减少文本和视觉样本之间的分布差距。基于人类反馈学习的方法能够进一步对齐人类偏好：LLaVA-RLHF(Sun et al., 2023c)通过引入人类反馈来减轻幻觉现象，将人类反馈学习范式从文本领域扩展到视觉语言对齐任务中；ViGoR(Yan et al., 2024)则设计了一个细粒度的奖励模型来更新策略模型，用于改进视觉定位并减少幻觉现象；RLHF-V(Yu et al., 2023b)基于直接偏好优化(Rafailov et al., 2024)提出密集直接偏好优化（DDPO），直接根据密集和细粒度的段级偏好来优化策略模型；HalDetect(Gunjal et al., 2024)提出了细粒度直接偏好优化（FDPO），使用来自个别示例的细粒度偏好来直接减少生成文本中的幻觉。除此之外，HallE-Switch(Zhai et al., 2023a)通过控制大语言模型中的参数知识来减少幻觉；OPERA(Huang et al., 2023)则提出基于过度信任惩罚和回顾分配策略的多模态大模型解码方法，无需额外的数据、知识或训练即可缓解幻觉问题。

6.2 任意模态输出的扩展

多数多模态大模型构建在大语言模型的基础上，也以文本作为主要的输出形式，然而正如GPT-4V (OpenAI, 2023b)以及GPT-4o (OpenAI, 2024)展现出的能力，用户的需求往往需要模型以多模态的输出来满足，对于开源模型来说，输出端的模态扩展也是研究的热门趋势。

首先被尝试扩展的模态是图像，应对文生图，图片编辑的任务需求。在大模型时代之前，StableDiffusion (SD) (Rombach et al., 2022)，InstructPix2Pix (Brooks et al., 2023)等模型已经在对应任务上展现了优异的性能，早期的Visual ChatGPT (Wu et al., 2023a)通过使用工具的方式引入了这样的能力，Mini DALLE3 (Lai et al., 2023)进一步探究了多个工具增强的大语言模型在交互式文生图场景下的表现。为了将图片生成能力内化到模型内部并且支持端到端的训练，GILL (Koh et al., 2023)首先提出用特殊字符在自回归生成过程力区分图片和文本的输出，并将输出图片位置的表示通过特殊的连接模块映射到SD模型输入，最终产生图片，这样的范式也在DreamLLM (Dong et al., 2024a)，MiniGPT-5 (Zheng et al., 2024)中被延用。除了来自SD的监督信号，原本SD使用的CLIP文本编码器也可以帮助对齐多模态大模型的输出和SD模型的输入 (Pan et al., 2024)。进一步，Emu (Sun et al., 2024b), Emu2 (Sun et al., 2024a), SEED (Ge et al., 2023)通过构建自编码器（AutoEncoder）的形式对齐了模型的输入输出，图片的输入表示可以用来恢复原本的图片并且监督多模态大模型自回归的训练。其中SEED以离散的形式表示中间的隐空间，Emu和Emu2则使用连续的隐空间。

除了图片，从任意模态到任意模态的生成则更具挑战，Next-GPT (Wu et al., 2023b)和Codi-2 (Tang et al., 2023b)对每种模态都构建了一个Diffusion模型，并用特殊字符区分不同的输出模态，将多模态大模型的输出表示提供给对应的模态生成器中。Any-GPT (Zhan et al., 2024)和Unified-IO2 (Lu et al., 2023a)则通过类似VQ-VAE的方法将多个模态的离散化表示（词表）增加到多模态模型输出的词表中，使得模型能够以原本自回归的方式输出多模态交错的序列。目前任意模态输出的方法主要以文本为支点，利用多个模态和文本的相关数据帮助训练，尽管目前也有方法来在GPT的帮助下构造部分多模态交错生成的数据 (Zhan et al., 2024; Wu et al., 2023b)，但是数据规模和形式相对有限，这也是限制当前模型的重要因素。

6.3 具身场景下的探索

得益于多模态大模型的多模态交互能力和强大推理能力，开发基于多模态大模型的具身智能体已成为具身智能领域的主要探索方向之一 (Zeng et al., 2023a)。其中一个主要挑战是如何将多模态大模型生成的文本与控制具身智能体的动作相结合，例如视觉语言导航任务 (Anderson et al., 2018b; Zhang et al., 2021a)。PaLM-E (Driess et al., 2023)通过接收语言、视觉和状态估计的多模态输入序列，输出低级别指令给下游的策略模块，以完成相应的具身任务。同时，它还是一个视觉语言通用模型，在传统的视觉语言任务（如VQA）中表现良好。与之相比，RT-2 (Brohan et al., 2023)可以将指令和视觉观察直接映射到机器人动作。具体来说，RT-2将低级别的机器人动作参数向量离散化为特殊的文本标记，加入模型的词表中。在实

际操作过程中，模型直接生成词表中的动作标记来控制机器人的动作。ManipLLM (Li et al., 2023g)则通过直接微调模型，使其输出具体的控制参数。具体而言，ManipLLM设计了物体、区域和姿势三个级别的微调任务，使模型能够逐步合理地预测以物体为中心的机器人操控姿势。NaviLLM (Zheng et al., 2023)和LLaRP (Szot et al., 2023)则利用额外的动作分类器，将特定标记对应的嵌入向量映射到合法动作，以完成动作控制。受Code-as-Policies (Liang et al., 2023)的影响，RoboCodeX (Mu et al., 2024)通过构造使用代码控制具身任务的预训练和指令微调数据，将代码生成与动作控制对齐，通过调用相应的API生成可执行代码，以控制下游具身任务的执行。目前，将多模态大模型的输出转化为可执行动作仍然是一个开放的问题。此外，现有的研究主要关注多模态大模型在特定具身任务上的应用。如何充分发挥其泛化能力，开发面向多样化具身任务的通用模型，也是一个亟待探索的问题。

7 总结

本文系统地回顾和探讨了多模态信息处理领域的研究进展，重点介绍了多模态预训练模型和多模态大模型的发展历程与技术细节。

首先，我们回顾了多模态预训练模型的早期研究，这些模型借鉴了文本预训练模型的成功经验，通过大量的视觉和文本数据进行自监督学习，取得了一定的成果。然而，预训练模型在泛化能力上存在不足，难以满足多样化应用场景的需求，很快被多模态大模型所取代。接下来，本文重点介绍了多模态大模型的出现及其架构设计。随着大语言模型的成功应用，多模态大模型通过扩展语言模型的能力，引入多模态编码器，实现了跨模态的高效对齐。我们详细讨论了多模态大模型的序列表示、基座模型和架构优化方案，特别是多视图视觉表征和特征压缩技术的应用。在训练方法方面，我们分析了多模态大模型的预训练阶段和指令微调方法，介绍了如何利用多模态数据完成视觉与文本特征的对齐，并通过指令微调和基于人类反馈的强化学习提高模型对自然语言指令的理解和执行能力。接着，我们分析了当前多模态大模型的评测方法，包含对已有基准数据集的简要介绍和对评测方法和归纳总结。最后，本文探讨了多模态大模型在解决幻觉问题、扩展任意模态输出以及具身场景探索方面的潜力和挑战。

本文的贡献在于全面、系统地总结了多模态信息处理的研究脉络。从早期的多模态预训练模型到当前的多模态大模型，本文详细分析了每个阶段的技术进展和应用场景，以及如何进行可靠的评测。通过对比不同方法的优缺点，我们揭示了各模型在处理跨模态信息时的优点和局限性。基于这些分析，我们提出了未来多模态研究可能的发展方向和潜在的创新点。希望本文能够为多模态技术的发展和应用提供有益的参考。

致谢

该论文得到了国家自然科学基金委(No. 62176058)和科技部重点研发计划（2023YFF1204800）的项目经费支持。复旦大学CFFF平台为该项目提供了算力支持。

参考文献

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. *International Journal of Computer Vision*, 123:4 – 31.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *ICCV*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *NIPS*, 35:23716–23736.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018a. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3674–3683.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1728–1738, October.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022a. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc.

Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. 2022b. Vl-beit: Generative vision-language pre-training.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. *arXiv:2303.07274*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18392–18402, June.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm. Dec.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tianlang Chen, Jiajun Deng, and Jiebo Luo. 2020a. Adaptive offline quintuplet loss for image-text matching. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 549–565, Berlin, Heidelberg. Springer-Verlag.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer.

Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023a. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv:2305.04160*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023b. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023c. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv:2311.12793*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023d. Pali: A jointly-scaled multilingual language-image model.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2023e. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023f. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. 2024a. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model. *arXiv:2402.11684*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 18–24 Jul.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, et al. 2023. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886*.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, et al. 2024. Mobilevlm v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*.

HyungWon Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, ShixiangShane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, EdH. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, QuocV. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. Oct.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 958–979, January.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2022. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *CVPR*, pages 326–335.

Mohammad Dehghani and Pavel Trojovskỳ. 2023. Osprey optimization algorithm: A new bio-inspired metaheuristic algorithm for solving engineering optimization problems. *Frontiers in Mechanical Engineering*, 8:1126450.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, Xiangwen Kong, Xiangyu Zhang, Kaisheng Ma, and Li Yi. 2024a. Dreamllm: Synergistic multimodal comprehension and creation.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024b. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, Jianfeng Gao, and Lijuan Wang. 2022a. Coarse-to-fine vision-language pre-training with fusion in the backbone. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32942–32956. Curran Associates, Inc.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022b. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, June.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, March.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuan-Jing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6514–6524.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, Ruize Wang, Zejun Li, Haijun Shan, and Xuanjing Huang. 2021. Tcic: Theme concepts learning cross language and vision for image captioning. *arXiv preprint arXiv:2106.10936*.

Zhihao Fan, Zhongyu Wei, Zejun Li, Siyuan Wang, Haijun Shan, Xuanjing Huang, and Jianqing Fan. 2022. Constructing phrase-level semantic labels to form multi-grained supervision for image-text retrieval. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 137–145.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023a. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023b. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. 2024. Sphinx-x: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*.

Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. 2023. Planting a seed of vision in large language model.

Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. 2024. Convllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv:2305.04790*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. 2024. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023a. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023b. Large multilingual models pivot zero-shot multimodal learning across languages. *arXiv preprint arXiv:2308.12038*.

Anwen Hu, Haiyang Xu, Jiabo Ye, Mingshi Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *ArXiv*, abs/2403.12895.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024b. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12976–12985, June.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, pages 6700–6709.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2023. Vcoder: Versatile vision encoders for multimodal large language models. *arXiv preprint arXiv:2312.14233*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2023b. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint arXiv:2312.06968*.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient multimodal large language models: A survey.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *ICCV*.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. 2022. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 18–24 Jul.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2023. Generating images with multimodal language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21487–21506. Curran Associates, Inc.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv:1810.09305*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73.

Zeqiang Lai, Xizhou Zhu, Jifeng Dai, Yu Qiao, and Wenhai Wang. 2023. Mini-dalle3: Interactive text to image by prompting large language models.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *45th ACM SIGIR*, pages 3108–3120.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11336–11344, Apr.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020b. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, November. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, page 121–137, Berlin, Heidelberg. Springer-Verlag.

Yang Li, Gang Li, Luheng He, Jingjie Zheng, Hong Li, and Zhiwei Guan. 2020d. Widget captioning: Generating natural language description for mobile user interface elements. *arXiv preprint arXiv:2010.04295*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Zejun Li, Zhongyu Wei, Zhihao Fan, Haijun Shan, and Xuanjing Huang. 2021b. An unsupervised sampling approach for image-sentence matching using document-level structural information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13324–13332.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 17–23 Jul.

Zejun Li, Zhihao Fan, Huaixiao Tou, Jingjing Chen, Zhongyu Wei, and Xuanjing Huang. 2022c. Mvptr: Multi-level semantic alignment for vision-language pre-training via multi-stage learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4395–4405.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. 2023a. Otterhd: A high-resolution multi-modality model. *ArXiv*, abs/2311.04219.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv:2307.16125*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023c. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023d. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023e. Videochat: Chat-centric video understanding. *arXiv:2305.06355*.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023f. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv:2306.04387*.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. 2023g. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. *arXiv preprint arXiv:2312.16217*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023h. Evaluating object hallucination in large vision-language models. *arXiv:2305.10355*.

Zejun Li, Ye Wang, Mengfei Du, Qingwen Liu, Binhao Wu, Jiwen Zhang, Chengxing Zhou, Zhihao Fan, Jie Fu, Jingjing Chen, et al. 2023i. Reform-eval: Evaluating large vision language models via unified re-formulation of task-oriented benchmarks. *arXiv preprint arXiv:2310.02569*.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2023j. Monkey: Image resolution and text label are important things for large multi-modal models. *arXiv preprint arXiv:2311.06607*.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2023. Code as policies: Language model programs for embodied control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9493–9500. IEEE.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022a. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. A convnet for the 2020s. arxiv e-prints.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023c. Improved baselines with visual instruction tuning. *arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023d. Visual instruction tuning. *arXiv:2304.08485*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023e. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, and Xiang Bai. 2023f. On the hidden mystery of ocr in large multimodal models. *ArXiv*, abs/2305.07895.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023g. On the hidden mystery of ocr in large multimodal models. *arXiv:2305.07895*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. 2024d. Textmonkey: An ocr-free large multimodal model for understanding document. *ArXiv*, abs/2403.04473.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. 2023a. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. 2023b. Evaluation and mitigation of agnosia in multimodal large language models. *arXiv preprint arXiv:2309.04041*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. 2024. Deepseek-vl: Towards real-world vision-language understanding. *arXiv:2403.05525*.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. 2023a. Cheap and quick: Efficient vision-language instruction tuning for large language models. *arXiv:2305.15023*.

Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. 2023b. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*.

Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. 2023. A benchmark for chinese-english scene text image super-resolution. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19395–19404.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, pages 3195–3204.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*, pages 2200–2209.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 ICDAR*, pages 947–952. IEEE.

MistralAITeam. 2023. Mixtral of experts a high quality sparse mixture-of-experts. [EB/OL]. https://mistral.ai/news/mixtral-of-experts/ Accessed December 11, 2023.

Yao Mu, Junting Chen, Qinglong Zhang, Shoufa Chen, Qiaojun Yu, Chongjian Ge, Runjian Chen, Zhixuan Liang, Mengkang Hu, Chaofan Tao, et al. 2024. Robocodex: Multimodal code generation for robotic behavior synthesis. *arXiv preprint arXiv:2402.16117*.

OpenAI. 2023a. Chatgpt (august 3 version).

OpenAI. 2023b. Gpt-4 technical report. *arXiv:2303.08774*.

OpenAI. 2024. Hello gpt-4o.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2024. Kosmos-g: Generating images in context with multimodal large language models.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv:2306.14824*.

Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. Detgpt: Detect what you need via reasoning. *arXiv:2305.14167*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Juan A Rodriguez, David Vazquez, Issam Laradji, Marco Pedersoli, and Pau Rodriguez. 2023. Ocr-vqgan: Taming text-within-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3689–3698.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv:2111.02114*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. arxiv.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565.

Baifeng Shi, Ziyang Wu, Maolin Mao, Xin Wang, and Trevor Darrell. 2024. When do we not need larger vision models? *ArXiv*, abs/2403.13043.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*, pages 8317–8326.

Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *CVPR*, pages 8802–8812.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada, July. Association for Computational Linguistics.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023a. Generative multimodal models are in-context learners. *ArXiv*, abs/2312.13286.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023b. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023c. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024a. Generative multimodal models are in-context learners.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024b. Emu: Generative pretraining in multimodality.

Andrew Szot, Max Schwarzer, Harsh Agrawal, Bogdan Mazoure, Rin Metcalf, Walter Talbott, Natalie Mackraz, R Devon Hjelm, and Alexander T Toshev. 2023. Large language models as generalizable policies for embodied tasks. In *The Twelfth International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November. Association for Computational Linguistics.

Benny J Tang, Angie Boggust, and Arvind Satyanarayan. 2023a. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*.

Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. 2023b. Codi-2: In-context, interleaved, and interactive any-to-any generation.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pages 5238–5248.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575.

Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv:1601.07140*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021a. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021b. Ufo: A unified transformer for vision-language representation learning.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. Git: A generative image-to-text transformer for vision and language.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 17–23 Jul.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022c. Simvlm: Simple visual language model pretraining with weak supervision.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023a. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023c. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023d. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023e. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. 2023f. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2024. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv:1901.06706*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul. PMLR.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Haiyang Xu, Ming Yan, Chenliang Li, Bin Bi, Songfang Huang, Wenming Xiao, and Fei Huang. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 503–513, Online, August. Association for Computational Linguistics.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2022. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *arXiv:2212.10773*.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. 2023a. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023b. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv:2306.09265*.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023c. Bridgetower: Building bridges between encoders in vision-language representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9):10637–10647, Jun.

Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. 2024. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *ArXiv*, abs/2403.11703.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *arXiv preprint arXiv:2402.06118*.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikihow. *arXiv preprint arXiv:2104.05845*.

Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv:2305.18752*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *ACM SIGIR*, pages 2733–2743.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Mingshi Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Feiyan Huang. 2023a. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *ArXiv*, abs/2310.05126.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023c. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *ArXiv*, abs/2311.04257.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. 2023. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv:2306.06687*.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3208–3216, May.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2023a. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2023b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023c. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Ya-Qi Yu, Minghui Liao, Jihao Wu, Yongxin Liao, Xiaoyu Zheng, and Wei Zeng. 2024. Texthawk: Exploring efficient fine-grained perception of multimodal large language models. *ArXiv*, abs/2404.09204.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv:2311.16502*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, pages 6720–6731.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 25994–26009. PMLR, 17–23 Jul.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023a. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. 2023b. What matters in training a gpt4-style language model with multimodal inputs? *arXiv:2307.02469*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023a. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling.

Da Zhang, Xiyang Dai, Xin Wang, and Yuan-Fang Wang. 2018. S3d: Single shot multi-span detector via fully 3d convolutional networks.

Jiwen Zhang, Jianqing Fan, Jiajie Peng, et al. 2021a. Curriculum learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34:13328–13339.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021b. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, pages 5579–5588.

Pan Zhang, Xiaoyi Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, Xinyu Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Y. Qiao, Da Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *ArXiv*, abs/2309.15112.

Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv:2303.16199*.

Xinnong Zhang, Haoyu Kuang, Xinyi Mou, Hanjia Lyu, Kun Wu, Siming Chen, Jiebo Luo, Xuanjing Huang, and Zhongyu Wei. 2024. Somelvlm: A large vision language model for social media processing. *arXiv preprint arXiv:2402.13022*.

Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference.

Yongqiang Zhao, Zhenyu Li, Zhi Jin, Feng Zhang, Haiyan Zhao, Chengfeng Dou, Zhengwei Tao, Xinhai Xu, and Donghong Liu. 2023a. Enhancing the spatial awareness capability of multi-modal large language model. *arXiv preprint arXiv:2310.20357*.

Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. 2023b. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv:2305.16103*.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2023. Towards learning a generalist model for embodied navigation. *arXiv preprint arXiv:2312.02010*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2024. Minigpt-5: Interleaved vision-and-language generation via generative vokens.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2024. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NeurIPS*, 36.

大模型工具学习进展与挑战

林衍凯
中国人民大学高瓴人工智能学院, 北京，100872，中国
yankailin@ruc.edu.cn

摘要

本论文综述了大模型工具学习的最新进展与挑战。工具作为人类智慧和能力的延伸，在提升生产力和解决问题方面至关重要。随着大语言模型（Large Language Models）的突破，工具学习得到了广泛关注，通过动态调用外部工具，显著增强了模型解决复杂问题的能力。

本文介绍了一个通用的大模型工具学习框架，包括控制器、工具集、环境和感知器四个核心组件。我们详细探讨了四个关键问题：意图理解、规划、工具使用和记忆管理。在意图理解方面，模型需要准确解析用户的输入和隐含意图。规划能力使模型能够将复杂任务分解为可执行的子任务。工具使用方面，介绍了示范学习、教程学习和探索学习三种主要训练策略，通过观察人类示范、阅读工具手册和直接探索来提升模型能力。记忆管理方面，提出了动态记忆管理和信息优先级管理等方法，以提高模型处理复杂任务的效率和准确性。

本文分析了当前大模型工具学习的研究进展和每个领域的挑战，为未来研究提供了有价值的见解。希望通过这篇综述，能帮助研究人员和开发者更好地理解和推进大模型工具学习领域的发展。

关键词： 工具学习；预训练模型

Challenges and Advances in Tool Learning with Foundation Models

Yankai Lin
Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China
yankailin@ruc.edu.cn

Abstract

This paper reviews the latest developments and challenges in the field of tool learning with foundation models. Tools, as extensions of human intelligence and capabilities, play a crucial role in enhancing productivity and problem-solving abilities. With the breakthroughs in Large Language Models (LLMs), tool learning has gained widespread attention, significantly enhancing the model's ability to solve complex problems through dynamic interaction with external tools.

We propose a general framework for tool learning with foundation models, consisting of four core components: the controller, tool set, environment, and perceiver. This paper explores four key issues: intent understanding, planning, tool use, and memory management. For intent understanding, models need to accurately interpret user

第二十三届中国计算语言学大会论文集，第34页-第47页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

inputs and underlying intentions. Planning capabilities enable models to break down complex tasks into executable sub-tasks. In terms of tool use, we introduce three main training strategies: demonstration learning, tutorial learning, and exploratory learning, which enhance model capabilities through observing human demonstrations, reading tool manuals, and direct exploration, respectively. For memory management, we propose methods such as dynamic memory management and information prioritization to improve the model's efficiency and accuracy in handling complex tasks.

This paper analyzes the current research progress in tool learning with foundation models, highlighting the challenges in each area and evaluating the effectiveness and limitations of existing solutions. By providing valuable insights for future research and applications, we aim to help researchers and developers better understand and advance the field of tool learning with foundation models.

Keywords: Tool Learning , Pre-trained Models

1 引言

工具是人类智慧与能力的延伸，旨在提升生产力、效率和解决问题的能力。从文明诞生之初，工具便成为我们生活中不可或缺的一部分 (Washburn, 1960)。工具的发明和使用，源于人类克服自身局限、探索未知领域的渴望。随着技术进步，我们能够完成越来越复杂的任务，释放出更多的时间和资源去追求更加宏伟的目标。工具不仅是文化和社会实践的重要基石，还极大地改变了我们的学习、交流、工作和娱乐方式，使其变得更加便捷和互动 (Gibson et al., 1993)。人类在工具发明和使用中发挥了关键作用，这无疑是智慧的鲜明体现 (Shumaker et al., 2011)。随着人工智能（AI）的快速发展，一个重要的问题是：人工智能是否具备与人类智能同样的工具学习能力?

掌握工具操作的前提是对工具功能的深刻理解，以及理解用户意图、进行规划和推理的能力。在预训练模型出现之前，进行以工具为导向的人工智能研究面临巨大挑战。尽管某些基础工具可以使用浅层统计模型或深度神经模型来实现调用 (Pomerleau, 1988; Mnih et al., 2013; Akkaya et al., 2019)，但它们的调用效果和稳定性仍不足以满足实际应用的需求，更不用说在各种工具间的泛化能力。这主要是由于传统监督学习在捕捉工具使用中的复杂操作方面存在局限性，而强化学习等试错类范式在掌握工具使用所需的庞大决策空间时也显得力不从心。总的来说，早期的人工智能研究在工具学习方面的根本限制在于模型能力的不足。

今天，我们正处于一个新的技术复兴期，由大语言模型（Large Language Models）的突破所驱动。大语言模型如GPT-4 (Achiam et al., 2023)展示了在自然语言处理（NLP）任务中的卓越能力 (El-Kassas et al., 2021; Zhang et al., 2024; Yang et al., 2018; Kwiatkowski et al., 2019)。然而，尽管这些模型能力惊人，它们在处理复杂计算和提供准确、及时的信息时仍存在挑战，因其依赖固定的参数知识，这常常导致"幻觉"现象，即生成看似合理但事实上错误或过时的回答 (Mallen et al., 2022; Vu et al., 2023; Ji et al., 2023; Zhang et al., 2023; **?**)。

随着大模型能力的不断增强，工具学习 (Qin et al., 2023a)范式被提出，期望大模型能够像人类一样熟练使用工具来解决复杂问题。工具学习通过允许模型动态调用外部工具，不仅提升了大模型的解决问题能力，还拓宽了其功能范围。例如，大语言模型可以使用计算器进行复杂计算以增强其数值计算能力，也通过天气API获取实时天气更新 (Pan et al., 2023; Wang et al., 2024)。调用工具可以显著提高模型响应用户查询的准确性，促进了更有效和可靠的用户互动。随着这一领域的不断发展，大模型工具学习有望在未来的人工智能领域中发挥关键作用，提供更灵活和适应性的解决方案 (Parisi et al., 2022; Karpas et al., 2022; Nakano et al., 2021; Surís et al., 2023)。

在过去的一年里，随着大模型的崛起，工具学习的研究也迅速增加。在实际应用中，GPT-4通过调用插件解决其知识限制，增强其能力，并将插件返回的结果与其内部知识结合，为用户生成更好的响应 (Achiam et al., 2023)。在研究领域，许多研究集中在评估大模型的工具学习能力以及如何增强这种能力 (Qin et al., 2023b; Xu et al., 2023; Gao et al., 2024; Zhao et al., 2024)。鉴于工具学习在大模型领域中的日益关注和快速发展，本文旨在回顾大模型工具学习最新的进展和挑战，以帮助研究人员和产业开发者了解当前的进展。

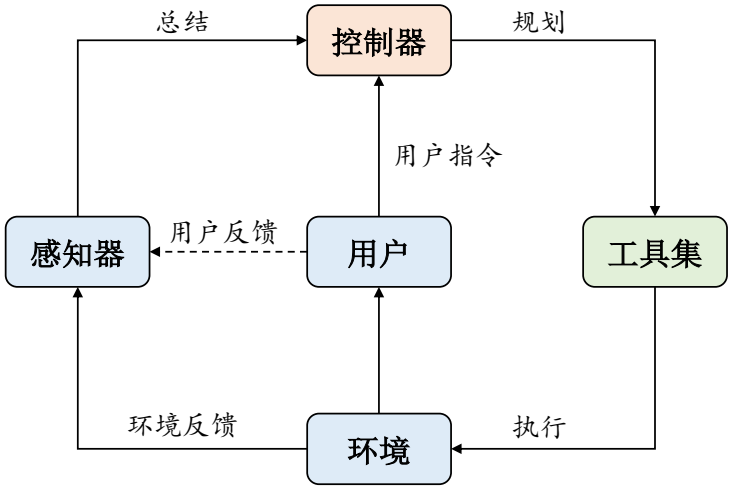


Figure 1: 大模型工具学习框架示意图。

本文首先基于现有工作总结出一个通用的大模型工具学习框架，包括控制器（通常使用大模型建模）、工具集、环境、感知器和人类。在统一框架的基础上，我们以回顾现有的大模型工具学习工作突出其核心研究问题。整个大模型工具学习过程从用户指令开始，模型需要为工具执行制定可执行的计划。为了将用户指令与适当的工具连接起来，模型首先需要学习理解指令背后的用户意图，在此基础上，将复杂任务分解为若干子任务，并有效地用适当的工具完成每个子任务。本文将以典型工作为例介绍其现有研究进展与挑战。

2 大模型工具学习框架

为了更全面地理解工具学习的核心挑战和未来方向，我们在 (Qin et al., 2023a)一文中提出了大模型工具学习的通用框架。具体而言，如图 1所示，我们将工具学习框架分为四个组成部分：包括控制器、工具集、感知器和环境。我们首先介绍工具学习过程中的每个组成部分：

- 控制器（Controller）控制器$\mathcal{C}$ 是大模型工具学习框架的核心，通常基于大模型来实现。控制器的任务是理解用户意图，并制定一个可行且精确的计划来使用工具满足用户需求。控制器需要理解用户的意图以及这些意图与可用工具之间的关系，然后制定一个选择适当工具的计划。对于复杂任务，控制器可能需要将其分解为多个子任务，这需要大模型具备强大的规划和推理能力。

- **工具集（Tool Set）**:工具集$\mathcal{T} = \mathcal{T}_1, \mathcal{T}_2, \cdots$ 是大模型工具学习的基础，包含一系列具有不同功能的工具。每个工具都有不同的接口。我们主要以应用程序编程接口（API）为例，说明如何与工具进行交互。API可以被定义为任何能够将基础模型的输出作为输入的功能。

- **环境（Environment）**：环境$\mathcal{E}$ 是工具执行的平台，为工具执行提供必要的基础设施，并向感知器提供工具执行结果。环境可以是虚拟的，也可以是真实的。虚拟环境具有易于访问和复制的优势，适合成本效益高的模型训练。然而，虚拟环境可能无法完全反映现实世界的复杂性，导致过拟合和泛化问题 (Hansen et al., 2021)。相比之下，真实环境提供了更具现实感的场景，但访问难度和成本较高。

- **感知器（Perceiver）**：感知器$\mathcal{P}$ 负责处理来自用户和环境的反馈，并生成信息的摘要供控制器使用。简单的反馈处理包括将用户和环境的反馈进行拼接或使用预定义模板进行格式化。总结后的反馈将传递给控制器，以辅助其决策。通过这些反馈，控制器可以判断生成的计划是否有效，以及执行过程中是否存在需要解决的异常情况。在更复杂的场景下，感知器应能够支持多模态（如文本、视觉和音频），以捕捉用户和环境反馈的多样性。

2.1 大模型工具学习流程

大模型工具学习的流程通常包括以下几个步骤。首先，用户提供指令，控制器需要根据指令制定一个可执行的计划。为了将用户指令与适当的工具连接起来，为了将用户指令与适当的工具连接起来，模型首先需要学习理解指令背后的用户意图，在此基础上，将复杂任务分解为若干子任务，并有效地用适当的工具完成每个子任务。在执行过程中，感知器会处理用户和环境反馈，并提供给控制器，以帮助调整和优化计划。

假设我们有一个工具集$\mathcal{T}$，控制器可以利用该工具集来完成某些任务。在执行的第t步，环境$\mathcal{E}$提供工具执行的反馈e_t。感知器$\mathcal{P}$接收用户反馈f_t和环境反馈e_t，生成总结反馈x_t。通常，感知器可以通过预定义规则（例如连接f_t和e_t）生成x_t，也可以使用复杂的神经模型进行建模。控制器$\mathcal{C}$生成当前步需要执行的计划a_t，选择并执行适当的工具进行执行。这个过程可以表示为以下概率分布：

$$p_{\mathcal{C}}(a_t) = p_{\theta_{\mathcal{C}}}(a_t \mid x_t, \mathcal{H}_t, q), \tag{1}$$

其中$\theta_{\mathcal{C}}$表示控制器的参数，q表示用户查询或指令，$\mathcal{H}t = (x_s, a_s)s = 0^{t-1}$表示历史反馈和计划。在其最简单的形式中，生成的计划a_t可以是工具执行的具体动作。控制器还可以将推理过程与动作预测相结合，a_t可能包含解释下一步应解决的子任务和选择解决该子任务的工具的推理轨迹。

在生成计划a_t之后，它将在环境$\mathcal{E}$中执行，环境反馈e_{t+1}将传递给感知器。上述过程重复进行，直到控制器完成任务。总体目标是找到一个动作序列a_t，最终实现用户指令q指定的任务。需要注意的是，在工具执行之后，控制器还可以将执行结果整合成一个合理的用户响应。

在理解和应用大模型工具学习的过程中，我们可以看到以下四个关键问题：

- **意图理解（Intent Understanding）**：理解用户任务意图是工具学习的首要步骤。控制器需要准确解析用户的输入，理解用户希望达成的目标和要求。这不仅仅是对用户文字表述的理解，还涉及到对隐含意图的推测和把握。例如，当用户询问“如何减肥”时，控制器需要明白用户不仅仅是在寻找一般性的建议，而可能需要具体的饮食计划、锻炼方案以及科学的减肥方法。

- **规划（Planning）**：在明确用户意图之后，控制器需要将用户的任务分解为一系列可执行的子任务。这一过程需要强大的规划和推理能力，以确保每个子任务都能有效地推动最终目标的实现。比如，对于“我想预订下周去北京的航班”这一任务，控制器需要分解出多个子任务，如查询航班信息、选择合适的航班、填写个人信息以及完成支付等步骤。

- **工具使用（Tool Use）**：一旦任务被分解为子任务，控制器需要选择和使用适当的工具来完成每个子任务。这要求控制器对可用工具的功能有深刻的理解，并能够根据任务需求灵活应用这些工具。例如，查询航班信息时可能需要调用航空公司API，而完成支付则需要调用支付网关的API。

- **记忆管理（Memory Management）**：在整个任务执行过程中，管理工作历史是确保任务顺利进行的重要因素。控制器需要跟踪每个子任务的执行状态，保存中间结果，并在需要时回顾和利用这些历史信息来调整和优化后续操作。有效的记忆管理能够帮助控制器在复杂和长时间的任务中保持高效和准确性。例如，在多轮对话中，控制器需要记住用户之前提供的信息和系统的响应，以便在后续对话中进行参考和调整。

在本文接下来的部分，我们将通过典型工作介绍在大模型工具学习中是如何处理意图理解、规划、工具使用、记忆管理问题的。具体来说，我们将探讨最新研究在这些领域所采取的方法和取得的进展，展示基础模型在实际应用中的表现和潜力。我们会深入分析每个问题的挑战，并探讨现有解决方案的有效性和局限性，最终为未来的研究和应用提供有价值的见解。

3 意图理解

理解用户意图一直是自然语言处理领域的一个长期研究课题 (Jansen et al., 2007; Sukthankar et al., 2014)。通过准确识别用户意图，控制器可以提供更加个性化的响应，从而提升

用户体验。近年来，指令调优（instruction tuning）方面的探索表明，基础模型在理解用户指令方面展现出了非凡的能力 (Wei et al., 2022a)。已有研究表明，将大语言模型在包含人类指令的多个数据集上进行微调，可以使模型甚至对未见过的任务指令进行泛化 (Wei et al., 2022a; Mishra et al., 2022; Sanh et al., 2022; Ouyang et al., 2022)。令人鼓舞的是，通过扩大模型规模以及增加训练指令的数量和多样性，可以进一步增强这种泛化能力 (Iyer et al., 2022)。

尽管意图理解能力已经取得了显著进展，但在实际的工具学习场景中仍存在一些挑战：

- **理解模糊指令（Understanding Vague Instructions）**：许多用户查询本质上是不精确的，甚至可能具有多义性，这要求控制器依赖上下文线索和背景知识来推断用户的真实意图。为了更好地理解用户意图，模型需要具备动态互动的能力。当遇到模糊或多义的指令时，模型应能够主动向用户询问澄清问题，以获取更多信息。这种互动不仅有助于提高指令理解的准确性，还能增强用户体验，使用户感受到被重视和理解。

  针对这个问题，钱等人的工作(Qian et al., 2024)介绍了一种创新的方法来解决大模型工具学习系统在处理用户指令时遇到的用户意图模糊问题。研究者们首先创建了一个名为Intention-in-Interaction (IN3)的新基准测试，它包含了一系列设计用来评估代理理解用户隐含意图能力的任务。这些任务被标注了不同程度的模糊性和缺失的细节，从而为评估提供了量化基础。接着，研究者们提出了将模型专家集成到大模型工具学习系统设计中的上游，以增强用户与代理之间的交互。特别是，他们开发了一个名为MistralInteract的模型，该模型通过模拟与用户的对话来明确任务的模糊性，主动询问缺失的细节，并在执行具体任务之前，将用户的意图明确化和细化为可操作的目标。为了训练MistralInteract，研究者们利用IN3的标注数据构建了模拟的对话记录，这些对话记录指导模型如何进行有效的交互。在实验中，MistralInteract被集成到了XAgent框架中，并通过一系列全面的评估来证明其在理解用户指令和执行任务方面的有效性。结果表明，MistralInteract在识别模糊任务、恢复关键信息、设置精确的执行目标以及减少不必要的工具使用方面表现出色，从而提升了整体的执行效率。这项工作不仅展示了如何通过用户参与来提高智能代理的性能，而且还通过开源数据和代码，为未来在这一领域的研究提供了基础和启发。通过这种方法，研究者们为构建更加用户友好的大模型工具学习系统迈出了重要的一步。

- **泛化到多样化指令（Generalization to Diverse Instructions）**：由于意图空间在理论上是无限的，因此基础模型在训练期间几乎不可能接触到所有现实世界中的意图表达。此外，个性化的挑战在于每个人都有自己独特的表达意图的方式，这要求模型适应不同个体的多样化意图表达。一种解决方案是利用用户反馈，主动适应个体用户，即个性化工具学习。

  由于大模型通常在通用领域进行训练，并根据广泛定义的人类偏好进行校准，这些偏好优先考虑有用性和无害性 (Ouyang et al., 2022; Nakano et al., 2021)。因此，它们在处理个人信息和提供个性化辅助方面存在困难。近年来，用户中心和个性化的自然语言生成受到了越来越多的关注 (Yang and Flek, 2021; Kirk et al., 2023)。现有工作涵盖了广泛的任务，如对话生成 (Madotto et al., 2019; Mazaré et al., 2018; Song et al., 2021; Zhong et al., 2022)、机器翻译 (Mirkin and Meunier, 2015; Michel and Neubig, 2018; Wuebker et al., 2018)和摘要生成 (Yan et al., 2011)。这些方法利用外部用户特定模块，如用户嵌入和用户记忆模块 (Zhang et al., 2018; Wu et al., 2021)，将不同用户的偏好、写作风格和个人信息注入生成的内容中。然而，这些工作通常针对特定任务设计，并在有限的用户信息下进行实验。如何将用户信息整合到大模型工具学习系统中仍然是一个未充分探索的领域。

  具体地，个性化工具学习强调在工具操作中考虑用户特定信息的重要性，主要有几个挑战：（1）异构用户信息建模：在现实世界中，个人信息可以来自多种异构来源。例如，使用电子邮件工具时，模型需要考虑用户历史对话记录中的语言风格，并从用户的社交网络中收集相关信息。这要求将具有多样结构的用户信息建模为统一的语义空间，允许模型联合利用这些信息。（2）个性化工具规划：不同用户在工具规划和选择上有不同的偏好。例如，在完成购买任务时，不同用户倾向于使用不同的在线购物平台。因此，模型需要根据用户偏好制定个性化的工具执行计划。（3）个性化工具调用：根据用户偏好自适应调用工具也是个性化工具学习中的重要方向。大多数工具在设计时未考虑个性化信息，这要求模

型根据用户偏好生成不同的工具输入。通过解决这些挑战，我们可以提高工具学习系统在理解用户意图和提供个性化支持方面的能力，从而提升整体用户体验。

4 规划

在大模型工具学习中，用户查询q通常涉及复杂任务，需要将其拆分为多个子任务并正确排序，这就需要强大的规划能力。最新研究表明，当预训练模型参数量达到一定规模时，其推理规划能力会显著增强 (Wei et al., 2022b)。拥有数百亿参数的预训练模型在解决复杂问题时，能够生成中间推理步骤，从而显著提升零样本和小样本任务的性能 (Nakano et al., 2021; Nye et al., 2021; Wei et al., 2022b)。然而，传统的少样本提示学习在处理需要复杂推理的问题时表现有限 (Creswell et al., 2022)。为此，(Wei et al., 2022c)提出了思维链推理（Chain-of-Thought，CoT）方法，通过在提示中插入推理步骤，引导模型生成解决问题的中间步骤，从而提升任务性能 (Wei et al., 2022c)。

基于大模型的强大推理规划能力，研究人员成功地将其应用于大模型工具学习的控制器中。其推理能力使控制器能够将复杂问题有效地分解为多个子问题，并确定每个子问题所需的工具。在这一方面，典型的研究工作主要分为两类：无反馈推理（*Planing without Feedback*）和带反馈推理（*Planning with Feedback*）。前者不与环境$\mathcal{E}$交互，生成静态任务拆解和工具调用计划；后者通过迭代地与环境$\mathcal{E}$交互，利用反馈逐步生成任务拆解和工具调用计划。

4.1 无反馈推理

无反馈推理是指在不依赖中间执行结果的情况下，直接生成多步计划。这种方法的一个典型例子是程序辅助语言模型（Program-Aided Language Models, PAL）(Gao et al., 2022)，该模型通过生成Python代码作为中间推理步骤，显著提高了大模型在算术、符号和算法方面的推理能力。PAL利用Python程序解释器作为工具，使模型能够像程序员一样编写详细注释，并在解决复杂问题时展示出强大的推理和规划能力。这一思路在具身自主智能体（embodied agents）中也得到了验证，如ProgPrompt(Singh et al., 2022)和Code-as-Policies (Liang et al., 2022)，这些方法通过生成可执行程序来指导具身自主智能体的实际行动，展示了模型在不直接与环境互动的情况下生成有效计划的能力。

另一无反馈推理的例子是Visual ChatGPT (Wu et al., 2023)，该系统将各种视觉预训练模型与ChatGPT结合，使其能够理解和生成图像。在Visual ChatGPT系统中，ChatGPT作为核心控制器，进行顺序决策。在每一步，ChatGPT可能调用一个视觉模型来修改现有图像或用纯文本回应用户。尽管这些模型没有直接与环境交互，但它们能够生成合理的中间步骤，并有效地处理复杂任务。

然而，无反馈推理方法的一个主要缺点是可能生成不切实际的计划。由于缺乏环境反馈，模型可能会在执行过程中遇到无法预见的异常情况。为了解决这一问题，SayCan (Ahn et al., 2022)提出了一种方法，通过使用价值函数估计每个动作成功执行的概率，使代理的行动更符合实际环境的约束。通过这种方式，工具学习系统能够在规划过程中考虑环境的实际情况，从而生成更为现实和可行的计划。

尽管存在这些挑战，无反馈推理方法在许多任务中仍然展示了其强大的规划能力。通过在生成计划时预见可能的异常情况并进行相应调整，无反馈推理方法能够在没有直接环境交互的情况下有效地解决复杂问题。

4.2 带反馈推理

相比之下，带反馈推理方法通过将环境$\mathcal{E}$纳入规划过程，逐步生成计划，使模型能够根据中间执行结果进行调整。这种方法更为灵活和动态，适合处理复杂任务，如多步问答和具体化学习，每一步的决策都依赖于之前的上下文。

带反馈推理的一个典型例子是Self-Ask (Press et al., 2022)、ReAct (Yao et al., 2022b)和ToolFormer (Schick et al., 2023)。这些研究表明，通过提供搜索引擎API访问权限，模型能够在多步问答中的准确性得到显著提高。通过思维链推理（CoT）提示或微调，这些模型能够将复杂问题分解为多个子问题，并利用搜索API找到每个子问题的答案。在获得每个子问题的答案后，模型能够迭代地确定下一个要问的问题或给出最终答案。

在具身学习中，带反馈推理方法通过直接与环境交互，进一步增强了模型的规划能力。例如，Inner Monologue (Huang et al., 2022)通过利用来自环境的多种反馈，如任务是否成功完成

和当前场景信息，生成更可行的计划，并提高高层次指令的完成能力。LLM-Planner (Song et al., 2022)明确考虑计划执行过程中可能出现的异常，并利用环境反馈在执行失败时重新生成计划，使模型能够适当处理例外情况。

此外，ReAct (Yao et al., 2022b)赋予模型自主权，使其在规划过程中能够根据当前情况决定何时停止生成动作令牌，从而制定更精细的后续计划。这种方法不仅提高了模型的灵活性，还增强了其在复杂任务中的表现能力。

尽管无反馈推理和带反馈推理各有优势，但将两者结合起来可以实现更强大的规划能力。在实际应用中，可以首先通过无反馈推理生成初步计划，然后利用带反馈推理在执行过程中不断调整和优化计划。例如，在复杂的任务如机器人控制和智能助手中，初步计划可以由内省推理生成，而具体执行时则通过外部推理实时调整，以应对环境中的变化和不确定性。这一思想被现在效果最好的大模型工具学习系统之一的XAgent所采用，使其能够真正处理需要几十步甚至上百步的复杂问题。

5 工具使用

在大模型工具学习中，工具使用指的是模型根据任务需要，选择并正确操作各种工具，以完成用户指令。工具使用的有效性直接影响任务的完成质量，因此，如何训练模型有效地使用工具至关重要。在这部分，我们将探讨如何通过不同的训练策略提升模型的工具使用能力。人类使用新工具主要有三种方式：通过学习其他人的示范、通过阅读工具手册或依靠自身探索尝试。同样地，我们将工具学习的训练策略分为三类：（1）示范学习：从具体的工具使用示范中学习，通常需要人工标注；（2）教程学习：从工具手册中学习；（3）探索学习：从工具探索尝试得到反馈中学习，通常涉及强化学习。

5.1 示范学习

从示范中学习是让模型通过观察和模仿人类专家的操作来掌握工具使用的方法。通过模仿学习，模型可以学到在特定情境下应如何使用工具。具体来说，行为克隆（Behavior Cloning）是一种常见的模仿学习形式，它假设人类专家的行为是最优或接近最优的，并通过监督学习来训练模型模仿这些行为。

假设我们有一个数据集$\mathcal{D}$，包含用户查询q和人类示范标注a^*的对，每对数据(q_i, a^i)表示一个具体的任务及其解决方案。我们的目标是优化控制器参数$\theta\mathcal{C}$，使其能够模仿人类专家的操作。学习目标可以表示为：

$$\theta_{\mathcal{C}}^* = \arg\max_{\theta_{\mathcal{C}}} \mathbb{E}_{(q_i,a_i^*)\in\mathcal{D}} \prod_{t=0}^{T_i} p_{\theta_{\mathcal{C}}}(a_{i,t}^* \mid x_{i,t}, \mathcal{H}_{i,t}, q_i), \tag{2}$$

其中$a_{i,t}^*$表示第i个任务执行到第t步时的人类标注。

具体来说，示范学习可分为三种主要方式：

- **监督学习（Supervised Learning）**：监督学习是最常见的示范学习形式，一般通过大量人类标注的数据来训练模型让其学会使用特定工具。例如，WebGPT (Nakano et al., 2021)通过与搜索引擎的交互，记录并模仿人类的搜索行为，来提升其信息检索能力。在这个过程中，研究人员首先建立了一个由Bing支持的搜索接口进行数据标注，然后采用标注出来的人类使用搜索引擎回答问题的行为序列来微调GPT-3，使其能够模仿人类专家的搜索行为。通过这种方式，WebGPT不仅能够生成有效的搜索查询，还能记录并总结重要的信息，从而提供更高质量的答案。另一个例子是WebShop (Yao et al., 2022a)，该模型在一个虚拟购物环境中学习如何根据人类指令进行商品购买。研究人员首先创建了一个互动环境，让模型能够浏览网页并选择商品，然后通过行为克隆训练模型模仿人类的购物行为，最终使模型能够在给定指令的情况下正确选择商品并完成购买。

- **半监督学习（Semi-supervised Learning）**：在很多情况下，获取大量高质量的人类标注数据是困难的，因此半监督学习提供了一种解决方案，即利用未标注数据生成伪标签，然后用这些伪标签来训练模型。例如，Baker等人的VPT工作(Baker et al., 2022)使用少量标注数据训练模型预测Minecraft视频游戏中每个时间步的动作伪标签，从而在没有大规模

人类行为标注数据的情况下，训练出更强大的模型。这种方法的核心在于利用少量的种子数据，训练一个初步模型来生成伪标签，再用这些伪标签进行更大规模的模型训练，从而在减少标注成本的情况下，提高模型的性能。

- **自监督学习（Self-supervised Learning）**：自监督学习进一步减少了对人工标注的依赖，模型通过自身的反馈迭代提升。例如，Toolformer工作 (Schick et al., 2023)利用少量人类示范，自动化生成工具使用示例，并通过过滤减少噪音，显著提升了工具使用性能。在这种方法中，模型首先使用一些基础示范进行初步学习，然后通过生成和筛选新的示范数据，不断改进自身的工具使用能力。这种方法的优势在于它能够利用现有的少量数据，通过自我学习和改进，逐步提升模型的性能。

5.2 教材学习

教程学习通过提示阅读工具手册，帮助模型理解工具的功能和使用方法。人类在学习使用新工具时，通常会通过阅读手册或观察他人演示来获取相关知识和技能。同样，模型也可以通过提示学习工具的使用方法。

在实际场景中，工具通常附带有使用手册（或教程），提供了关于其功能和用法的详细信息。大模型具备强大的零次学习（zero-shot learning）和少次学习（few-shot learning）能力，可以通过提示理解工具的功能和使用方法。具体来说，可以通过手动设计或检索构建合适的任务特定提示，这些提示描述了API的功能或通过示例展示其用法。

我们将提示方法分为两类：

- **零次提示（Zero-shot Prompting）**：描述API功能、输入输出格式、可能的参数等。这种方法使模型能够理解每个API可以处理的任务。

- **少次提示（Few-shot Prompting）**：为模型提供具体的工具使用示例。通过模仿这些示例中的人类行为，模型可以学习如何使用这些工具。

虽然提示方法具有显著优势，但也面临一些挑战。首先，提示受到输入上下文长度的限制。尽管大模型已经显示出通过提示学习使用简单工具的能力，但在面对多个复杂工具及其长描述时情况可能会更具挑战性。特别是当工具集大幅扩展时，在提示中提供所有可能的工具变得不可行，给定的上下文长度也有限。其次，也是最重要的一点，提示的效果很大程度上依赖于模型本身，较小或能力较弱的模型可能无法很好地理解提示。特别对于工具手册类的提示，几乎只有OpenAI系列大模型如ChatGPT、GPT-4拥有较强的零次提示、少次提示能力。

为了弥补这一差距，ToolLLM(Qin et al., 2023b)首先创建了一个名为ToolBench的指令调整数据集，它通过以下三个阶段自动构建：

1. API收集：从RapidAPI Hub收集了16,464个真实世界的RESTful APIs，覆盖了49个不同的类别。

2. 指令生成：使用ChatGPT生成涉及这些APIs的多样化指令，包括单工具和多工具场景。

3. 解决方案路径标注：利用ChatGPT为每个指令搜索有效的解决方案路径，即API调用链。

为了解决标注过程中大模型规划能力不足的问题，研究者们开发了一种新颖的基于深度优先搜索的决策树算法（DFSDT），它允许模型评估多种推理路径并扩展搜索空间。

基于ToolBench，研究者们微调了LLaMA模型，构建了ToolLLaMA模型，并为其配备了一个基于深度神经网络的API检索器，以推荐适合每个用户指令的APIs。实验结果表明，ToolLLaMA不仅能够执行复杂的指令，还能泛化到未见过的APIs，展现出与ChatGPT相当的性能。ToolLLaMA还在APIBench数据集上表现出强大的零样本泛化能力，证明了其在未知APIs上的适应性和灵活性。

5.3 探索学习

除了直接从人类示范和阅读工具手册学习使用工具外，模型还可以通过直接探索工具使用进行学习。这种方法在实际环境中探索使用工具，利用环境或人类的反馈来优化模型的工具使用策略。探索学习可以描述为通过开放探索优化控制器参数$\theta_{\mathcal{C}}$：

$$\theta_{\mathcal{C}}^{*}=\arg\max_{\theta_{\mathcal{C}}}\mathbb{E}_{q_i\in Q}\mathbb{E}_{\{a_{i,t}\}_{t=0}^{T_i}\in p_{\theta_{\mathcal{C}}}}\left[R(\{a_{i,t}\}_{t=0}^{T_i})\right], \tag{3}$$

其中R是从反馈序列中估计的奖励，T_i表示处理q_i所需的迭代次数。

强化学习（Reinforcement Learning, RL）通过与环境的互动，基于反馈信号（如奖励）来优化模型的决策过程。在大模型工具学习中，强化学习将动作空间定义为工具集中的所有工具，模型学习选择适当的工具并执行正确的动作以最大化奖励信号。例如，在机器人抓取任务中，模型通过反复尝试和调整抓取策略来学习最佳的工具使用方法 (Levine et al., 2018)。

在大模型工具学习中，我们主要从两个方面获取反馈：环境反馈和人类反馈。

- **环境反馈**：环境反馈包括模型与环境互动后得到的结果。根据反馈的不同，可以分为结果反馈和中间反馈：（1）结果反馈是任务完成与否的最终反馈，评估模型的整体表现。例如，WebShop (Yao et al., 2022a)通过评估模型购买的产品与人类购买的产品的相似性来提供反馈；（2）中间反馈是动作触发的环境状态变化，通过观察这些变化，模型可以学习每个动作的有效性，从而更好地调整其行为。例如，在信息检索任务中，模型可以通过观察搜索结果页面的内容来判断搜索查询的有效性，并根据这些信息调整后续的查询策略。这种反馈提供了关于每次工具执行效果的详细及时信息，使模型能够在任务执行过程中不断改进。

- **人类反馈**：人类反馈可以是显式的，如通过评分系统直接评价模型的行为；也可以是隐式的，通过用户行为和与模型的互动来推导用户的满意度。尽管人类反馈准确且稳定，但获取成本较高，因此人类反馈强化（RLHF）(Christiano et al., 2017)被提出，通过模仿人类给出奖励，然后使用强化学习算法来优化策略。例如，WebGPT(Nakano et al., 2021)利用人类反馈指导策略模型，使其在长篇问答中表现更好。

6 记忆管理

记忆管理问题仍是现有大模型工具学习系统中较少探索的领域。尽管在模拟型的大模型自主智能体中已经有初步的基于长短期记忆和外部检索模块的探索 (Park et al., 2023)，现有的大模型工具学习系统大多仍然依赖于通过提示语句的形式将信息拼接到大模型前面，利用大模型本身的长文本建模能力进行处理。然而，这种方法忽略了记忆管理中的一些关键问题。在复杂任务中，记忆管理不仅涉及到如何高效存储和检索信息，还包括如何在任务执行过程中动态更新和利用这些信息。具体来说，记忆管理面临以下几个挑战：

- **信息的持久性和可访问性**：大模型在处理长文本时可能会遇到上下文窗口的限制，导致无法有效利用所有相关信息。虽然通过提示语句拼接可以在一定程度上缓解这一问题，但这种方法无法保证信息的持久性和随时可访问性。对于需要长期记忆的任务，如持续对话或跨会话任务，现有方法显得力不从心。

- **信息的组织和优先级管理**：在任务执行过程中，不同信息的重要性和优先级可能不同。现有的方法往往无法区分和组织这些信息，导致重要信息可能被淹没在大量无关信息中。有效的记忆管理需要能够动态调整信息的组织方式，确保高优先级的信息能够被快速检索和利用。

- **动态更新和一致性维护**：随着任务的推进，新的信息不断涌入，旧的信息可能需要更新或淘汰。这就需要一个高效的机制来动态更新记忆内容，并保证信息的一致性。简单的拼接提示语句无法实现这一点，容易导致信息不一致或冗余。

7 总结

本文系统地探讨了大模型工具学习的最新进展及其核心问题。作为人类智慧的延伸，工具极大地提升了生产力和效率。当前，如何赋予大模型以工具学习能力，成为人工智能领域的前沿研究课题。

我们介绍了一个大模型工具学习的通用框架，包含控制器、工具集、感知器和环境四个核心组成部分。控制器负责理解用户意图并制定执行计划；工具集提供完成任务所需的各种工具；感知器处理用户和环境的反馈信息；环境为工具的执行提供平台和反馈。

基于这一框架，本文详细讨论了大模型工具学习中的几个关键问题：（1）意图理解：尽管大模型在指令理解方面已有显著进展，但理解模糊指令和泛化到多样化指令仍是主要挑战。通过主动互动和个性化学习，可以提升模型的理解能力。（2）规划与推理：在规划与推理方面，无反馈推理和带反馈推理是两种主要方法。结合这两种方法，可以更有效地应对复杂任务，确保任务拆解和工具调用的合理性和有效性。（3）工具使用：通过示范学习、教程学习和探索学习三种训练策略，可以显著提升模型的工具使用能力。示范学习通过模仿人类操作来训练模型；教程学习利用手册提示帮助模型理解工具功能；探索学习通过与环境互动，利用反馈优化工具使用策略。（4）记忆管理：记忆管理是大模型工具学习中的关键问题之一。现有方法主要依赖于提示语句拼接，但在处理复杂任务时存在局限。

综上所述，大模型工具学习在提升人工智能系统智能性和适应性方面展现出巨大潜力。通过持续的优化和创新，我们有望开发出更智能、更高效的人工智能系统，推动技术进步和社会发展。

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *ArXiv preprint*, abs/2204.01691.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving rubik's cube with a robot hand. *ArXiv preprint*, abs/1910.07113.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *ArXiv preprint*, abs/2206.11795.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *ArXiv preprint*, abs/2205.09712.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications*, 165:113679.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, et al. 2022. Pal: Program-aided language models. *ArXiv preprint*, abs/2211.10435.

Shen Gao, Zhengliang Shi, Minghang Zhu, Bowen Fang, Xin Xin, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2024. Confucius: Iterative tool learning from introspection feedback by easy-to-difficult curriculum. In *In Proceedings of 38th Conference on Artificial Intelligence (AAAI)*.

Kathleen R Gibson, Kathleen Rita Gibson, and Tim Ingold. 1993. *Tools, language and cognition in human evolution*. Cambridge University Press.

Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, et al. 2021. Self-supervised policy adaptation during deployment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *ArXiv preprint*, abs/2207.05608.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *ArXiv preprint*, abs/2212.12017.

Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2007. Determining the user intent of web search engine queries. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 1149–1150. ACM.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback. *ArXiv preprint*, abs/2303.05453.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436.

Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. 2022. Code as policies: Language model programs for embodied control. *ArXiv preprint*, abs/2209.07753.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318, Melbourne, Australia. Association for Computational Linguistics.

Shachar Mirkin and Jean-Luc Meunier. 2015. Personalized machine translation: Predicting translational preferences. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2019–2025, Lisbon, Portugal. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv preprint*, abs/2112.09332.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *ArXiv preprint*, abs/2112.00114.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada, July. Association for Computational Linguistics.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *ArXiv preprint*, abs/2205.12255.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Dean A Pomerleau. 1988. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv preprint*, abs/2210.03350.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Yankai Lin, Zhong Zhang, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! towards implicit user intention understanding of language model driven agents. *arXiv preprint arXiv:2402.09205*.

Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, et al. 2023a. Tool learning with foundation models. *arXiv preprint arXiv:2304.08354*.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023b. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv preprint*, abs/2302.04761.

Robert W Shumaker, Kristina R Walkup, and Benjamin B Beck. 2011. *Animal tool behavior: the use and manufacture of tools by animals*. JHU Press.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. 2022. Progprompt: Generating situated robot task plans using large language models. *ArXiv preprint*, abs/2209.11302.

Haoyu Song, Yan Wang, Kaiyan Zhang, Wei-Nan Zhang, and Ting Liu. 2021. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–177, Online. Association for Computational Linguistics.

Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. Llm-planner: Few-shot grounded planning for embodied agents with large language models. *ArXiv preprint*, abs/2212.04088.

Gita Sukthankar, Christopher Geib, Hung Bui, et al. 2014. *Plan, activity, and intent recognition: Theory and practice*. Newnes.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Xingyao Wang, Zihan Wang, Jiateng Liu, Yangyi Chen, Lifan Yuan, Hao Peng, and Heng Ji. 2024. Mint: Evaluating llms in multi-turn interaction with tools and language feedback. *In Proceedings of 12th International Conference on Learning Representations (ICLR)*.

Sherwood L Washburn. 1960. Tools and human evolution. *Scientific American*, 203(3):62–75.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, et al. 2022a. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *ArXiv preprint*, abs/2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. *ArXiv preprint*, abs/2201.11903.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970, Online. Association for Computational Linguistics.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *ArXiv preprint*, abs/2303.04671.

Joern Wuebker, Patrick Simianer, and John DeNero. 2018. Compact personalized models for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 881–886, Brussels, Belgium. Association for Computational Linguistics.

Qiantong Xu, Fenglu Hong, Bo Li, Changran Hu, Zhengyu Chen, and Jian Zhang. 2023. On the tool manipulation capability of open-source large language models. *arXiv preprint arXiv:2305.16504*.

Rui Yan, Jian-Yun Nie, and Xiaoming Li. 2011. Summarize what you are interested in: An optimization framework for interactive personalized summarization. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1351, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Diyi Yang and Lucie Flek. 2021. Towards user-centric text-to-text generation: A survey. In *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, pages 3–22. Springer.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022a. Webshop: Towards scalable real-world web interaction with grounded language agents. *ArXiv preprint*, abs/2207.01206.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022b. React: Synergizing reasoning and acting in language models. *ArXiv preprint*, abs/2210.03629.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten De Rijke. 2024. Let me do it for you: Towards llm empowered recommendation via tool learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is more: Learning to refine dialogue history for personalized dialogue generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5808–5820, Seattle, United States. Association for Computational Linguistics.

大模型逻辑推理研究综述

刘汉蒙
西湖大学/ 杭州，浙江
liuhanmeng@westlake.edu.cn

张岳
西湖大学/ 杭州，浙江
zhangyue@westlake.edu.cn

摘要

理解自然语言的逻辑结构和关系是机器理解的核心任务，也是人工智能领域的关键研究议题。随着大数据和计算能力的提升，预训练语言模型在逻辑推理方面取得了显著进展，使得大规模模型的逻辑推理能力成为研究的新焦点。本综述旨在全面梳理大模型在逻辑推理领域的研究进展，探讨其对人工智能系统智能水平评估的重要性及其在推动人工智能发展中的作用。

本文首先界定了大模型逻辑推理能力的研究范畴，系统性地讨论了逻辑推理的类型和特点，并回顾了相关理论的发展，为研究提供了清晰的框架。接着，从任务形式和数据基准的角度，详细介绍了逻辑推理研究的基础工作，为理解大模型的性能提供了基准。进一步，本文深入分析了大模型在逻辑推理能力上的现状，通过不同推理类型的案例研究，展示了大模型的能力表现。同时，本文还探讨了提升大模型逻辑推理能力的方法，包括预训练、指令微调、解码策略和神经符号混合方法，并对这些方法进行了比较分析。最后，本文提出了对未来研究方向的展望，旨在激发更多的学术讨论和探索，推动逻辑推理能力研究的进一步发展。

关键词： 语言模型 ； 逻辑推理 ； 人工智能

Survey on Logical Reasoning of Large Pre-trained Language Models

Hanmeng Liu
Westlake University
liuhanmeng@westlake.edu.cn

Yue Zhang
Westlake University
zhangyue@westlake.edu.cn

Abstract

This survey synthesizes the advancements in logical reasoning within large language models (LLMs), a pivotal area of AI. It delineates the research scope, theoretical underpinnings, and benchmarks for assessing LLMs' reasoning prowess. The paper scrutinizes current capabilities through case studies and evaluates strategies to bolster reasoning, such as pre-training and neuro-symbolic methods. The review concludes with future directions, encouraging further exploration to enhance logical reasoning in AI systems.

Keywords: language model , logical reasoning , artificial intelligence

1 引言

逻辑推理作为人工智能（AI）的核心能力之一，始终在自然语言处理（NLP）和人工智能领域的研究中占据举足轻重的地位。自20世纪50年代计算机科学和人工智能诞生之初，逻辑推理就被视为构建智能系统的基石(Newell and Simon, 1956; McCarthy and Hayes, 1981; McCarthy, 1959; McCarthy, 1989)，尽管随着技术的发展，其地位有所起伏，但始终贯穿于人工智能的演进历程。在人工智能的早期阶段，逻辑推理是许多研究项目的核心组成部分。形式逻辑与符号推理被广泛地用于模拟人类的思考过程。研究者们希望通过构建能够运用逻辑推理的机器，来实现类似人类的智能。这种理念催生了专家系统的出现，这些系统基于手工编码的规则进行决策，模拟了特定领域的人类专家知识。由于早期计算能力和NLP尤其是自然语言理解技术的限制，形式逻辑推理在70年代成为人工智能研究的主导领域(Pereira, 1982; Cann, 1993)。80年代，随着数据驱动方法和神经网络的兴起，机器学习算法取得了令人瞩目的成果。尽管如此，逻辑推理的重要性并未被完全忽视。相反，越来越多的研究者开始认识到，逻辑推理与数据驱动学习相结合，能够构建出更为强大、健壮和可解释的人工智能系统。

近年来，随着深度学习技术的不断发展，越来越多的研究致力于将逻辑推理与深度学习相结合。这包括使用神经符号集成的方法，以及设计具有逻辑推理能力的神经网络结构。这些努力旨在克服传统逻辑推理方法的局限性，同时保留其强大的推理能力。当前的深度学习技术已经显示出在逻辑推理方面的潜力(Clark et al., 2020)。逻辑推理在对话系统(Beygi et al., 2022)、信息提取(Ru et al., 2021)和问答系统(Angeli et al., 2016; Shi et al., 2021)等NLP应用中发挥着重要作用。同时，预训练语言模型的训练技术进步，使得语言模型的参数规模不断扩大，大语言模型成为推动人工智能发展的关键驱动力。大语言模型通过“预训练+指令微调+人类反馈强化学习”的训练范式，提升了模型的普适性和遵循人类指令的能力。相应的，大模型的逻辑推理能力也受到了更多关注(Liu et al., 2023b; Xu et al., 2023)。

逻辑推理在人工智能中的应用非常广泛，包括但不限于知识表示、规划、诊断、学习、自然语言处理等领域。通过逻辑推理，人工智能系统能够模拟人类的决策过程，处理复杂问题，并在不确定性环境中做出合理的判断。首先，逻辑推理为人工智能系统提供了强大的推理能力，使其能够理解和处理复杂的逻辑结构。这对于解决复杂问题、进行高级决策以及实现更高级别的智能至关重要。其次，逻辑推理有助于增强人工智能系统的可解释性。通过逻辑推理，我们可以更好地理解人工智能系统的决策过程，从而增加对其信任度。此外，逻辑推理还有助于提高人工智能系统的泛化能力，使其能够应对新的、未见过的情境。

本文旨在明确逻辑推理的概念，梳理其在大规模预训练语言模型中的应用，并探讨如何提升机器的逻辑推理能力。我们首先基于哲学和NLP场景提出人工智能逻辑的定义，讨论需要逻辑推理的任务类型，并引入逻辑推理的分类。接着，我们对NLP和大模型相关的自然语言推理进行梳理，涵盖演绎推理、归纳推理、溯因推理以及类比推理。本文介绍人工智能逻辑研究常用的数据集基准、测试平台与工具库，为研究者提供实践参考。最后，本文总结提高模型逻辑推理能力的策略和方向。后续章节架构如图1所示。

2 人工智能逻辑的概念、分类与发展历程

2.1 定义与类型

逻辑推理是人工智能领域的基石，它涉及使用一系列逻辑规则和原则，从已知的前提出发，推导出新的结论。这一过程不仅模拟了人类的思维过程，而且为智能系统在面对复杂问题和决策时提供了一种结构化和严密的思考方式。与之相关的一个概念是人工智能逻辑(Thomason, 2024)，是使用逻辑方法和成果来研究智能主体（intelligent agent）如何处理知识的领域。它起源于对计算机中知识处理功能的实现探索，由约翰·麦卡锡等先驱提出，旨在形式化人工智能问题。其核心在于建立一套形式理论，以支撑知识表示、推理和修正等过程。逻辑推理在人工智能中关注几个关键点：前提的明确性、规则的应用、结论的有效性以及推理过程的透明度。

逻辑推理能力主要可以划分为以下四类：

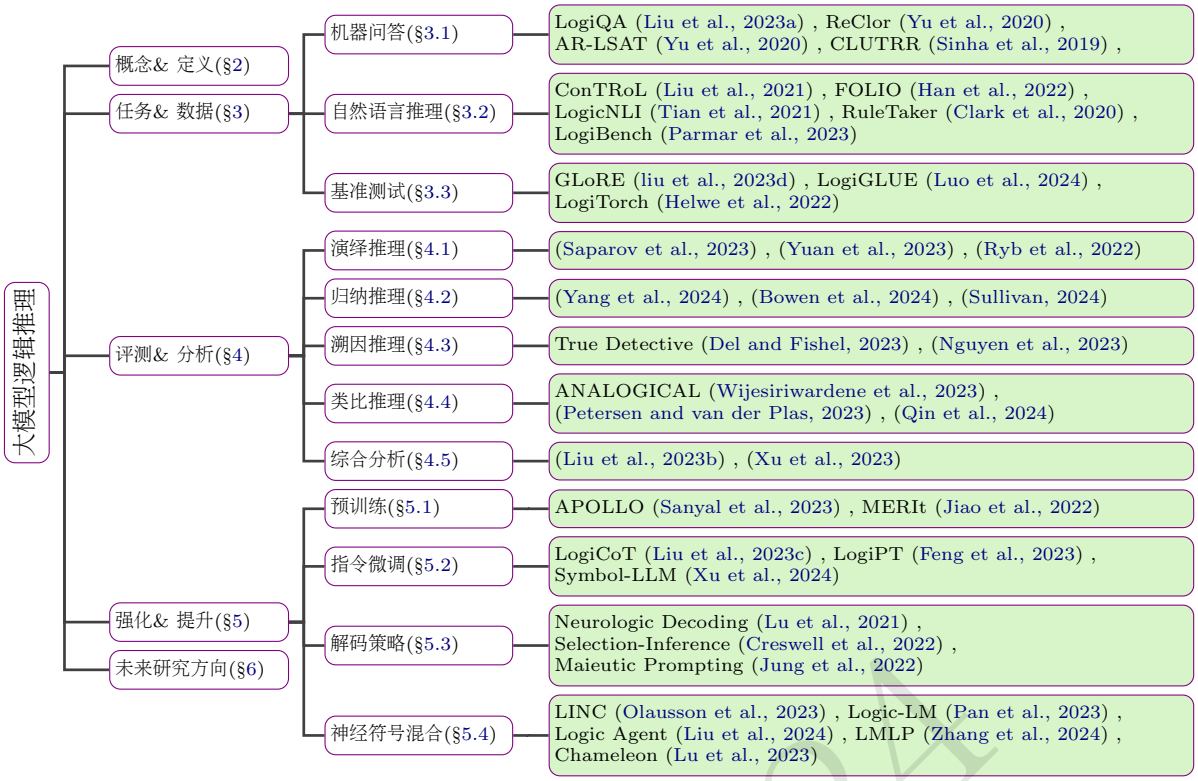


Figure 1: 本综述的组织结构。

1. **演绎推理：**这种推理方式是从一般规律到特定实例的推导。其核心思想是，如果前提都是真的，那么由此得出的结论必定也是真的。例如，假设所有苹果都是红色的，并且某一个特定的果实是苹果，那么可以推断这个果实是红色的。

2. **归纳推理：**它是基于观察到的特定事例来推导出一般性结论的过程。尽管归纳推理提供的结论通常被认为是真实的，但并不总是绝对确切的。例如，由于过去观察到的所有天鹅都是白色的，可能会归纳认为所有天鹅都是白色的。

3. **溯因推理：**这是一种试图为某些观察到的情况找出最合适的解释或原因的推理方法。当面临缺少信息的情境时，这种推理方式尤其有用。例如，看到街上的湿迹，可能会推测刚下过雨。

4. **类比推理：**基于已知相似情况间的比较，类比推理涉及从一个实例到另一个实例的推断。它通常用于解决问题和创造性思考，通过寻找不同情境之间的相似性来推导结论。例如，如果知道行星绕太阳运行的轨迹是椭圆形，类比推理可以帮助人们推测其他天体可能也呈现相似的轨道特征。

2.2 发展历程

对逻辑推理的研究可以追溯到古希腊时期，亚里士多德（Aristotle）被认为是逻辑学之父。他提出了著名的三段论，奠定了古典逻辑的基础，至今仍对逻辑推理有着深远的影响。进入中世纪，逻辑学得到了进一步的发展。学者们开始对亚里士多德的逻辑进行更深入的分析和发展，为逻辑学的深化和完善做出了重要贡献。

17世纪，莱布尼茨（Leibniz）的工作标志着逻辑学开始与数学和哲学结合，为后来的形式逻辑和数理逻辑奠定了基础。他的普遍语言（universal language）和推理计算器（calculus ratiocinator）的概念，预示了逻辑与计算机科学的结合。19世纪，乔治·布尔（George Boole）发展了布尔代数，将逻辑表达为代数形式，这是数理逻辑的重要里程碑。布尔代数不仅为逻辑提供了一种新的数学表述，也为电子计算机的逻辑电路设计提供了理论基础。

自20世纪初期，逻辑学与数学的紧密结合标志着一个新时代的开端。伯特兰·罗素（Bertrand Russell）和阿尔弗雷德·诺斯·怀特海德（Alfred North Whitehead）在他们的里程碑

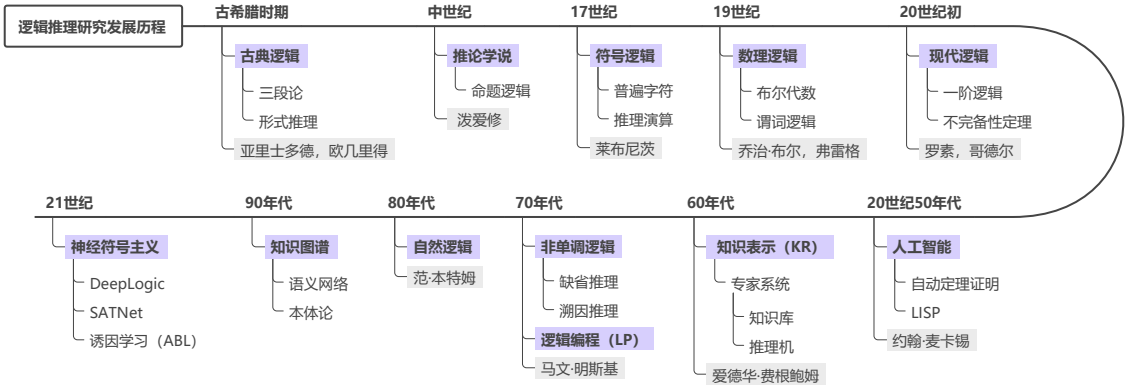


Figure 2: 人工智能逻辑相关理论发展时间线。

式著作《数学原理》（*Principia Mathematica*）中，进一步发展了逻辑理论，引入了更为复杂的逻辑体系，这不仅推动了数学逻辑的发展，也奠定了现代逻辑学的基础。随着20世纪中叶人工智能的诞生，逻辑推理迅速成为知识表示和自动定理证明的核心工具。约翰·麦卡锡（John McCarthy）等人的开创性工作为逻辑编程和知识库的构建提供了坚实的理论基础，极大地推动了人工智能领域的快速发展。特别是LISP语言(McCarthy, 1978)，由麦卡锡在1958年设计，作为世界上最早的高级编程语言之一，在人工智能研究中，尤其是在早期的专家系统和自然语言处理等领域，发挥了重要作用。进入70年代，非单调逻辑(McDermott and Doyle, 1980)的提出解决了传统逻辑在处理现实世界问题时的局限性，为常识推理提供了新的理论框架，使得逻辑推理能够更好地模拟人类的日常思维。Prolog （Programming in logic）(Kowalski, 1974)作为一种逻辑编程语言被提出，它建立在逻辑学理论的基础上，用于自然语言、人工智能的研究，被广泛应用于建造专家系统，智能知识库。Prolog基于逻辑学理论，采用”反向链式推理”（backward chaining）的方法，使得程序能够通过一系列逻辑规则推导出结论，这在建造专家系统和智能知识库方面发挥了巨大作用。知识表示和逻辑推理的结合推动了专家系统的发展。专家系统是人工智能领域中最早和最成功的应用之一，它们模拟专家的决策过程，通过知识库和推理引擎来提供专业建议或诊断。知识库是专家系统的核心，存储了大量经过验证的事实、规则和逻辑关系，而推理机则是系统智能的引擎，利用这些知识通过逻辑推理来解决复杂问题或进行决策。MYCIN (Van Melle, 1978)是早期专家系统的代表，用于诊断感染性疾病并推荐相应的抗生素治疗方案，展示了专家系统在处理复杂问题、提供专业建议和辅助决策方面的潜力。随着技术的发展，现代专家系统正在不断集成新的算法和数据源，以提供更准确和全面的解决方案。80年代，Haugeland (1989)提出了“有效的老式人工智能”（GOFAI）这一概念，用来泛指早期的人工智能方法。这些方法主要基于符合主义或逻辑主义。GOFAI的核心是使用符号来表示知识，并通过逻辑规则来处理这些符号，以实现智能行为。90年代，随着互联网的快速发展，组织和表示大量知识的需求日益增长。知识图谱作为解决这一问题的方法之一，开始受到研究者的关注。知识图谱的发展与逻辑推理紧密相关，因为它们提供了一种框架，用于表示复杂的事实和关系，这些信息可以被用来支持复杂的推理任务。其中令人瞩目的人工智能项目是Cyc (Lenat, 1995)，它通过构建庞大的常识知识库，使机器能够模拟人类的推理能力。该知识库涵盖了从基础物理概念到高级抽象思维，是常识推理和知识表示研究的宝贵资源。

进入21世纪，数据驱动和基于统计的方法成为人工智能研究的主流。学者们开始探索使用神经网络来处理传统上由逻辑和符号推理方法处理的问题，包括知识表示、推理和学习任务。Hamilton等人(2018)展示了如何将逻辑和知识表示为嵌入向量，而Yang等人(Yang et al., 2017)提出了可微分的一阶逻辑规则推理方法，使得神经网络能够学习逻辑推理过程。NeuroSAT (Selsam et al., 2018)尝试通过消息传递神经网络（MPNN）来解决布尔可满足性问题（SAT），进一步推动了神经网络在逻辑推理领域的应用。同时，神经符号主义的兴起，如DeepLogic (Cingillioglu and Russo, 2019)和SATNet (Wang et al., 2019)等工具和方法的开发，为逻辑推理带来了新的视角和可能性。逻辑推理理论的发展历程是一个不断进化和创新的过程。从古代哲学到现代计算机科学，逻辑推理一直是人类智能和人工智能研究的核心。随着技术的不断进步，逻辑推理理论将继续在智能系统的设计和实现中发挥关键作用。

数据集	语言	题目类型	数据量	构建方式
LogiQA	中文/英文	多项选择	15,937	考试题目
ReClor	英文	多项选择	6,138	考试题目
AR-LSAT	英文	多项选择	2,064	考试题目
CLUTRR	英文	问答	6,016	规则生成
ConTRoL	英文	三分类	8,325	考试题目
FOLIO	英文	二分类	1,351	专家构建
LogicNLI	英文	三分类	30K	规则生成
RuleTaker	英文	二分类	20K	规则生成
LogiBench	英文	二分类	1,270	规则生成
GLoRE	中文/英文	混合	17组	混合
LogiGLUE	英文	混合	24组	混合
LogiTorch	英文	混合	16组	混合

Table 1: **逻辑推理任务的重要数据集与基线。**

3 相关任务与数据集基准

逻辑推理的相关数据集，从数据来源上可以分为三类：

1）规则生成：这类数据集是通过逻辑推理规则自动生成的，它为获取大规模探针数据提供了有效途径。然而，生成的数据可能会在形式上显得单调，因此在设计时需要考虑多样性，以确保能够全面考察模型的推理能力。

2）专家构建：由领域专家精心构建的数据集通常在质量上更为精准和准确。专家会选择特定的语言和逻辑现象进行深入研究，并投入大量时间来确保数据集的严谨性。尽管这类数据集在数量上可能不及众包方式收集的语料库，但它们在研究中具有不可替代的价值。

3）考试试题：考试试题本质上也是由专家构建的，但它们通常是从互联网上搜集的开放数据。这种方式节约了数据构建的成本，并且由于考试的标准化，这些数据集通常具有较大的量和高质量的标签。考试试题已成为逻辑推理任务最主要的数据来源之一，包括但不限于中国国家公务员考试、美国法学院入学考试（LSAT）、美国研究生入学考试以及公司招聘笔试等。

这些数据来源各有优势和局限性，研究者在选择数据集时应根据研究目标和模型需求综合考量。规则生成的数据集适合于探索模型的推理广度，专家构建的数据集有助于评估模型的推理深度，而考试试题数据集则为模型提供了丰富的真实世界推理场景。表 1展示了逻辑推理任务中重要的数据集与基线。这些数据集中在两类经典自然语言理解任务形式-机器问答与自然语言推理。

3.1 机器问答

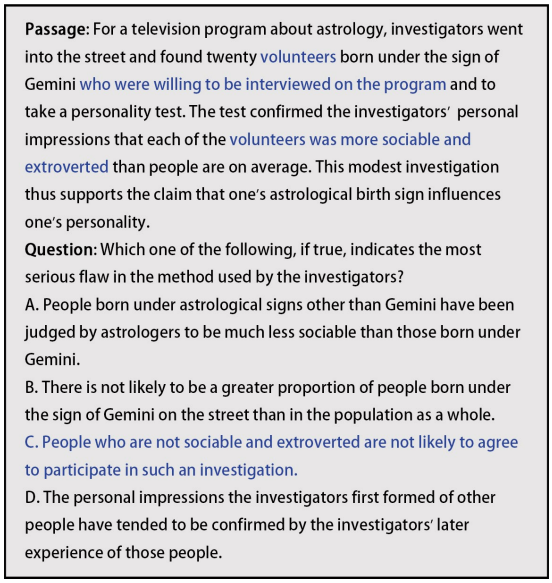
机器问答是评估大模型逻辑推理能力的常用途径。该任务通过提供一段上下文文本，以及与该文本相关的问题，要求模型给出相应的答案。回答的方式主要包括多项选择、片段抽取和自由回答。在这些形式中，多项选择题目因其标准化程度高而被广泛采用于逻辑推理能力的测评。图3a展示了一个典型的多项选择形式的机器问答题目。在这个例子中，模型需要处理自然语言的输入，理解上下文，并运用逻辑推理来选择正确的答案。

LogiQA (Liu et al., 2023a)是一个逻辑阅读理解数据集，源自中国公务员考试。该数据集目前更新至2.0版本，包含15,937条数据。LogiQA分为中文和英文两个版本，专注于考察机器阅读理解中的复杂逻辑推理能力。

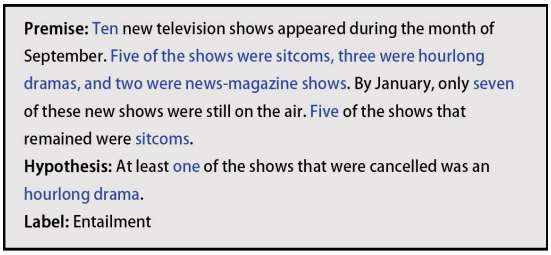
ReClor (Yu et al., 2020)数据集来源于美国研究生管理科入学考试（GMAT）题目，采用与LogiQA相似的四选项多项选择题形式。该数据集为英文，包含6,138条数据。

AR-LSAT (Wang et al., 2022)数据集取材于美国法学院入学考试（LSAT），每道题目包含5个备选选项。该数据集共包含2,064条数据，涵盖了三种主要类型的推理游戏：排序游戏、分组游戏和分配游戏。

CLUTRR (Sinha et al., 2019)是一个专注于归纳推理的数据集。CLUTRR要求自然语言理解（NLU）系统在短篇故事中推断人物之间的亲缘关系。成功完成此任务需要同时提取实体之间的关系以及推断这些关系背后的逻辑规则。该数据集共有6,016条数据。



(a) LogiQA数据集中的一个多项选择题例子。



(b) ConTRoL数据集中一个NLI例子。

Figure 3: NLP中逻辑推理任务举例。

3.2 自然语言推理

自然语言推理（NLI）是NLP中的一项传统任务，专注于探究两个文本之间的蕴含关系，即判断一个假设或结论是否能从给定的前提中推导出来。图3b展示了逻辑推理中NLI任务的一个例子。通过给出一个前提（premise），和一个对应的假设（hypothesis），模型需要选择正确的蕴含标签。NLI更直接地考察模型的逻辑推理能力。

在传统的NLI任务中，两个文本的关系通常被标记为二分类或三分类。二分类包括蕴含（Entailment）和非蕴含（Non-entailment），而三分类则进一步细分为蕴含（Entailment）、反对（Contradiction）和中立（Neutral）。一些非传统的NLI数据集则使用“正确（True）”和“错误（False）”作为一个文本能否被另一个文本正确推断出的标签。

ConTRoL (Liu et al., 2021)数据集源自各种企业、事业单位的招聘考试，例如银行笔试和美国警察选拔考试。原始数据的标签由“正确”、“错误”和“无法判断”组成，这些在NLI任务中分别对应于“蕴含”、“反对”和“中立”三个标签。ConTRoL数据集共包含8,325条数据。

FOLIO (Han et al., 2022)是一个由专家构建的一阶逻辑（First-Order Logic, FOL）推理数据集。每条数据的前提和结论都有对应的FOL标注，数据标签为“正确”和“错误”。FOLIO数据集共包含1,351条数据，为研究者提供了一个挑战性的逻辑推理基准。

LogicNLI (Tian et al., 2021)是一个由逻辑规则生成的数据集，其黄金标签为“蕴含”、“反对”和“中立”。LogicNLI是一个NLI风格的数据集，有效地将目标FOL推理与常识推理分离，并且可以从准确性、鲁棒性、泛化能力和可追溯性四个角度对语言模型（LMs）进行诊断。该数据集拥有超过30,000条数据。

RuleTaker (Clark et al., 2020)是一个人工合成的数据集，通过要求模型基于一组规则和事实进行推理，以生成真或假的响应，明确地针对逻辑推理。该数据集提供输入事实和规则作为上下文，要求输出二元的真或假响应。该数据的标签为“正确”与“错误”，尽管其最初设计用于问答，但该数据集可以轻松转换为NLI风格的二元分类任务。

LogiBench (Parmar et al., 2023)是一个大模型生成的数据集，它围绕25种推理类型并利用GPT-3生成数据，涵盖命题逻辑，一阶逻辑，以及非单调逻辑。最终得到“上下文（Context）”-“结论（Conclusion）”组，以及“是”、“否”两类标签。LogiBench测试数据一共有1,270条，但尚未开源数据链接。

3.3 综合测试基准

随着逻辑推理研究的不断深入，近年来出现了多个旨在评估和比较不同模型性能的基准测试套件。这些测试套件集成了领域内的多个数据集，并在数据选择和格式上各具特色。

GLoRE (liu et al., 2023d)是一个专注于大模型少样本和零样本测试的平台。它对所选数据集的格式进行了定制化改造，以满足大模型测评的特定需求。GLoRE的独特之处在于它仅提供测试集而不包含训练集，这为评估模型在缺乏大量训练数据时的表现提供了条件。此外，GLoRE支持OpenAI及Huggingface模型的一键测试，目前整合了17组数据集。

LogiGLUE (Luo et al., 2024)是一个包含24项任务的逻辑推理数据集集合。它将所有数据集统一转换为序列到序列（sequence-to-sequence）的数据格式，以便于模型的输入和处理。与GLoRE不同，LogiGLUE除了提供测试集外，还保留了完整的训练集，为模型训练提供了充足的数据。此外，LogiGLUE对数据集进行了推理类型的标注，这有助于研究者更好地理解和利用数据集。

LogiTorch (Helwe et al., 2022)是一个基于PyTorch的自然语言逻辑推理库，它由逻辑推理数据集，不同逻辑推理神经架构的实现，以及一个清晰的API这三个部分组成。LogiTorch目前包含16组数据集，为研究者提供了一个方便的框架，以快速实现和测试逻辑推理模型。

4 大模型逻辑推理能力测评与分析

随着预训练语言模型的快速发展，评估和分析这些模型在逻辑推理方面的能力变得尤为重要。本节将从演绎推理、归纳推理、溯因推理和类比推理四个角度对大模型的逻辑推理能力进行分类测评，并介绍一些综合性的分析工作。

4.1 演绎推理

演绎推理是一种从一般到特殊的逻辑推理形式，它基于已知的前提出发，通过逻辑推演得出必然的结论。在人工智能领域，演绎推理对于构建智能系统具有重要意义，尤其是在处理逻辑严密的任务时，如自动定理证明、知识表示和逻辑规划等。近年来，预训练语言模型（PLMs）在演绎推理任务上的能力受到了广泛关注。

1）演绎推理的基准测试：为了全面评估预训练语言模型的演绎推理能力，研究者们开发了多个基准测试套件。例如，Saparov 等人(2023)，提出了一个新的合成且可编程的推理数据集，用于测试大语言模型对更复杂证明的泛化能力。他们的实验结果表明，尽管大语言模型能够泛化到组合证明，但在处理更长的证明时存在困难，并且在没有明确示例的情况下难以产生假设性子证明。

2）预训练语言模型的演绎推理能力：尽管预训练语言模型在自然语言处理（NLP）的多个任务中表现出色，但它们在演绎推理方面的能力尚不明确。最近的研究工作，如Yuan 等人(2023)，通过一系列控制实验，发现预训练语言模型在学习演绎推理规则方面存在局限性。该研究指出，尽管经过演绎推理微调的预训练语言模型在标准基准测试上表现良好，但它们在泛化到未见案例时表现不佳，并且在面对简单的表面形式变化时一致性不足。

3）逻辑推理的可学习性：Ryb 等人(2022)通过他们提出的AnaLog 数据集，探讨了预训练语言模型是否能够学习分析性和演绎逻辑推理。他们的研究结果表明，尽管预训练语言模型在学习词预测的过程中能够编码预测蕴含关系的信息，但它们对逻辑语句实现中的词汇和句法变化仍然敏感。

4.2 归纳推理

归纳推理作为人类智能的核心组成部分，对于人工智能同样至关重要。它涉及从特定实例推广到一般性规则或假设的过程。近年来，大语言模型在归纳推理方面的能力受到了研究者们的广泛关注。

1）归纳推理的新范式：在传统计算机科学中，归纳推理通常使用形式语言来表示知识（事实和规则）。然而，这种方法存在一些系统性问题，例如对原始自然语言输入的处理不足、对错误标注数据的敏感性，以及处理模糊输入的能力不足。为了解决这些问题，Yang等人(2024)提出了一种新的归纳推理范式，即从自然语言事实中归纳出自然语言规则，并创建了一个包含1.2k规则-事实对的数据集DEER，用于评估这一任务。他们还提出了新的自动度量标准，并使用预训练语言模型作为“推理者”进行了实验，展示了一种现代的归纳推理方法。

2）大语言模型的归纳推理能力评估：Bowen等人(2024)的研究对当前大语言模型的归纳推理能力进行了全面评估。他们认为，仅考虑规则的归纳过于狭隘且不现实，因为归纳推理通常与其他能力（如规则应用、结果/规则验证和更新信息整合）混合在一起。通过设计的符号任务

对大语言模型进行探测，研究发现即便是最先进的大语言模型在执行直观上简单的任务时也会失败，这表明大语言模型在执行这些任务时存在局限性。

3）Transformers与归纳学习：Sullivan等人(2024)的论文则探讨了Transformer模型在自然语言推理（NLI）任务中的逻辑推理能力。研究结果表明，尽管这些模型在NLI任务上表现出色，但它们并没有学会逻辑推理。具体来说，他们发现微调在NLI数据集上的模型学会将外部否定视为干扰因素，有效地忽略了其在假设句子中的存在。此外，尽管经过广泛的微调，几个接近最先进的编码器和编码器-解码器Transformer模型仍无法归纳学习在单独外部否定前缀情景下的排中律。这些发现表明，当微调用于NLI任务时，Transformers可能并没有学会归纳推理。

4.3 溯因推理

溯因推理是一种从观察到的现象中推导出最佳解释或原因的逻辑推理形式。在法律、医疗和日常问题解决等领域，溯因推理对于专家从可用证据中构建有说服力的论证至关重要。

1）溯因推理的挑战与模型评估：True Detective (Del and Fishel, 2023)深入探讨了溯因推理在自然语言处理中的挑战，并提出了一种新的方法来评估模型在溯因推理任务上的性能。他们指出，尽管现有的预训练语言模型在某些推理任务上表现出色，但它们在处理溯因推理时仍面临显著的挑战。

2）溯因推理在法律领域的重要性：法律领域中，溯因推理对于解释法律条文、构建案件论证以及从法律文本中提取相关信息具有重要作用。Nguyen等人(2023)探讨了当前最先进的法律推理模型在支持溯因推理任务方面的能力。他们通过构建一个增强逻辑的数据集，包含498,697个样本，来评估法律领域中的最先进模型。研究结果表明，尽管这些模型在处理与法律文本相关的任务上表现良好，但在支持溯因推理任务上仍有不足。

4.4 类比推理

类比推理是一种基于已知信息，通过比较和对比来推断未知信息的推理方式。在人工智能领域，类比推理对于提升机器的智能和理解能力具有重要意义。

1）ANALOGICAL — 长文本类比评估：Wijesiriwardene等人(2023)中提出了ANALOGICAL，这是一个新的基准测试，旨在评估大语言模型在长文本类比推理方面的能力。该基准测试涵盖了从单词到隐喻的六个复杂性等级，并使用了13个数据集和三种不同的距离度量方法来评估8个大语言模型的性能。研究发现，随着类比复杂性的增加，大语言模型识别类比对的难度也在增加。

2）模型学习类比推理与人类表现比较：另一项研究(Petersen and van der Plas, 2023)探索了模型是否能够学习基本的类比推理。研究集中于人类类比推理评估中更典型的类比。实验表明，即使是在少量数据的情况下，模型也能够学习类比推理，并且经过训练的模型接近人类的表现。

3）大语言模型类比推理：Qin等人(2024)质疑了大语言模型是否真正能够执行类比推理。他们通过一系列类比提示的实验探索了大语言模型在多样化推理任务上的类比推理能力。研究发现，在类比提示中使用自动生成的随机示例在某些任务上出人意料地达到了可比或甚至更好的性能，表明大语言模型在这种场景下并不总是通过类比来进行推理。研究还指出，自动生成示例的准确性是影响大语言模型类比推理性能的关键因素，并据此设计了两种改进方法，显著降低了推理成本。

4.5 综合分析

1）GPT-4与ChatGPT的逻辑推理能力：Liu等人(2023b)对GPT-4和ChatGPT的逻辑推理能力进行了深入评估。研究团队分析了多个逻辑推理数据集，包括流行的LogiQA和ReClor基准测试以及新发布的AR-LSAT数据集。他们测试了ChatGPT和GPT-4在多项选择阅读理解和自然语言推理任务上的表现，并构建了一个逻辑推理分布外数据集来调查模型的鲁棒性。结果显示，ChatGPT在大多数逻辑推理基准测试上的表现得以显著优于RoBERTa微调方法。GPT-4在逻辑推理数据集上的表现更优，然而在处理新发布和分布外数据集时性能显著下降。这表明逻辑推理对ChatGPT和GPT-4来说仍然是一个挑战，尤其是在分布外的推理数据集上。

2）大语言模型作为逻辑推理器的全面评估：Xu等人(2023)对大语言模型是否真的擅长逻辑推理进行了全面评估。他们选择了十五个典型的逻辑推理数据集，并将它们组织成演

绎、归纳、演绎和混合形式推理设置。研究包括了三个代表性的大语言模型（即text-davinci-003、ChatGPT和BARD），并在所有选定的数据集上进行了零样本、单样本和三样本设置的评估。不同于以往仅依赖简单指标（例如准确率）的评估，他们提出了从客观和主观两个方面进行细粒度评估，涵盖了答案和解释。此外，为了避免知识偏差的影响，专注于基准测试大语言模型的逻辑推理能力，他们提出了一个新的中立内容数据集NeuLR，包含3000个样本，涵盖演绎、归纳和演绎推理设置。基于深入评估，本文最终形成了一个从六个维度（即正确、严谨、自我意识、主动、导向和无幻觉）对大语言模型逻辑推理能力进行一般性评估的方案。

5 大模型逻辑推理能力强化与提升

本节将从预训练、指令微调、解码策略和神经符号混合方法这几个方面介绍大模型逻辑推理能力的提升方法。

5.1 预训练

预训练是提升大模型逻辑推理能力的重要步骤，它通过在大量文本数据上训练模型来捕捉语言的丰富特征和深层语义。以下关于预训练方法的研究展示了如何通过特定的策略来增强语言模型的逻辑推理技能。

1）APOLLO —自适应预训练方法：Sanyal等人(2023)提出了APOLLO，这是一种简单的自适应预训练方法，旨在提高语言模型的逻辑推理能力。APOLLO通过选择Wikipedia的一个子集，并使用一组逻辑推理关键词作为过滤词来进行自适应预训练。此外，该方法提出了两种自监督损失函数：第一种是修改掩码语言建模损失，仅掩蔽那些可能需要更高阶推理来预测的特定词性词；第二种是提出句子级分类损失，教导模型区分蕴含和矛盾类型的句子。APOLLO展示了其与先前基线相比在两个逻辑推理数据集上的有效性，其在ReClor上表现相当，在LogiQA上超越了基线。

2）MERIt — 元路径引导的对比学习：Jiao等人(2022)提出了MERIt，这是一种新颖的元路径引导的对比学习方法，用于文本的逻辑推理。MERIt旨在通过自监督预训练在大量未标记的文本数据上进行训练。该方法包括两个关键策略：一是基于元路径的策略，用于发现自然文本中的逻辑结构；其次是对比事实数据增强策略，以消除预训练引起的信息捷径。在两个具有挑战性的逻辑推理基准测试ReClor和LogiQA上的实验结果表明，MERIt方法在显著提升的同时，超越了d当时的基线。

5.2 指令微调

指令微调是提升大语言模型在特定任务上性能的有效手段。通过指令微调，模型能够学习遵循复杂的指令，进而在各种任务上展现出更优的推理和执行能力。

1）LogiCoT — 逻辑链式推理指令微调：Liu等人(2023c)提出了LogiCoT，这是一个用于逻辑链式推理（Chain-of-Thought，CoT）的指令微调方法。LogiCoT旨在通过指令引导GPT-4生成逻辑推理的链式解释。研究者们通过精心设计的过程，利用现有的逻辑推理数据集，构建了CoT指令，并利用GPT-4的能力生成高质量输出。这些数据不仅包括符号推理，还包括多步骤CoT推理，为提升人工智能模型的逻辑推理能力提供了全面而精细的资源。通过使用LogiCoT对一个小型的指令微调模型（LLaMA-7b）进行微调，实验结果表明，与现有的指令微调模型相比，该模型在逻辑推理基准测试和通用基准测试上都取得了显著的性能提升。

2）LOGIPT — 直接模拟逻辑求解器的推理过程：Feng等人(2023)提出了LOGIPT，类似LogiCoT，它通过微调基座模型，模仿逻辑求解器的推理过程，并通过学习严格遵守求解器的语法和语法规则来绕过解析错误。LOGIPT在一个新的指令微调数据集上进行微调，该数据集是从揭示和细化演绎求解器的不可见推理过程中得到的。实验结果表明，LOGIPT在两个公共的演绎推理数据集上超越了现有的基于求解器的语言模型，效果类似于在ChatGPT或GPT-4这样的大模型上的少样本提示方法的结果。

3）Symbol-LLM — 以符号推理为中心的大语言模型：Xu等人(2024)提出了Symbol-LLM系列模型，旨在通过指令微调和框架创新，提升大语言模型在符号推理任务上的表现。研究者们首先提出了一个包含34个任务的数据集合，并纳入了约20种不同的符号语言，目的是捕捉符号之间的相互关系并促进它们之间的协同作用。然后，他们提出了一个两阶段的调优框架，成功地在不损失通用性能力的情况下注入了符号知识。实验表明，Symbol-LLM系列模型在符号和自然语言任务上展现出了较为平衡的高性能。

这些研究表明，通过指令微调，可以有效地提升大语言模型在逻辑推理、演绎推理以及符号操作等任务上的能力。这些方法不仅增强了模型对复杂指令的遵循能力，也为构建更智能、更灵活的人工智能系统提供了新的可能性。

5.3 解码策略

解码策略是提升大语言模型在文本生成任务中性能的关键技术。通过精心设计的解码算法，模型能够生成更准确、更符合约束条件的文本，进而在各种文本生成任务上展现出更优的推理和执行能力。

1）NEUROLOGIC DECODING — 基于谓词逻辑约束的神经文本生成：Lu等人(2021)提出了NEUROLOGIC DECODING，这是一种新颖的解码策略，它允许神经语言模型在生成流畅文本的同时满足复杂的词汇约束。该方法不仅强大而且高效，能够处理任何可以用谓词逻辑表达的词汇约束集，并且其渐近运行时间与传统的柱搜索（beam search）相当。

NEUROLOGIC DECODING的核心在于将硬性的逻辑约束转化为解码目标中的软性惩罚项，并通过基于束的搜索找到近似最优解。这种方法有效地控制了文本生成过程，而无需对模型结构或训练流程进行任何修改。在四个不同的文本生成任务上进行的实验结果表明，NEUROLOGIC DECODING在确保给定约束得到满足的同时，保持了高质量的生成，从而在监督和零样本（zero-shot）设置中都取得了新的最先进结果。

2）Selection-Inference — 利用大语言模型进行可解释的逻辑推理：Creswell等人(2022)提出了Selection-Inference（SI）框架，这是一种利用预训练的大语言模型作为通用处理模块的方法。SI框架通过交替选择和推理步骤，生成一系列可解释的、因果推理步骤，最终得出答案。在五次少样本（5-shot）泛化设置中，不进行微调的情况下，使用7B参数的大语言模型在10个逻辑推理任务上的性能提高了100%以上，与同等规模的普通基线相比。

SI框架的关键在于其模块化结构，它将逻辑推理分解为选择和推理两个阶段。选择阶段涉及选择足够的相关信息以进行单一推理步骤，而推理阶段仅看到选择模块提供的有限信息，并利用它来推断新的中间证据。这种方法不仅提高了推理问题的解决性能，还产生了可以解释最终答案的推理路径。此外，SI框架产生的推理路径是因果的，每一步都依赖于前一步，这与常见的端到端深度学习方法中的“黑箱”计算形成对比。

3）Maieutic Prompting — 递归解释与逻辑一致性推理：Jung等人(2022)开发了一种名为Maieutic Prompting的方法，旨在即使从大语言模型不可靠的生成中也能推断出正确答案。Maieutic Prompting通过递归方式（例如，X是真的，因为...）和演绎性地诱导解释树，然后将推理框架为这些解释及其逻辑关系的可满足性问题。

Maieutic Prompting的核心思想是苏格拉底式的助产术（maieutic method），它促使语言模型为不同假设生成演绎性解释，并通过深度递归推理，然后集体排除矛盾的候选项，从而得出一致的答案。该方法在三个需要复杂常识推理的挑战性基准上进行了测试，Maieutic Prompting在准确性上比最先进的提示方法提高了20%，并且作为一种完全无监督的方法，与监督模型具有竞争力。

这些解码策略的研究和应用表明，通过精心设计的解码过程，可以显著提高大语言模型在逻辑推理、文本生成和常识推理等任务上的性能。这些方法不仅增强了模型遵循复杂约束的能力，也为构建更智能、更可靠的人工智能系统提供了新的思路和工具。

5.4 神经符号混合方法

神经符号混合方法是一种新兴的研究领域，旨在结合深度学习的强大表示能力和符号推理的精确性与可解释性。这些方法通过将大语言模型与符号推理系统相结合，以提高模型在复杂逻辑推理任务上的性能。

3）LINC — 结合语言模型与一阶逻辑证明器：Olausson等人(2023)提出了LINC（Logical Inference via Neurosymbolic Computation），这是一种神经符号方法，通过将大语言模型与一阶逻辑（FOL）证明器相结合来提升逻辑推理能力。在LINC中，大语言模型作为语义解析器，将自然语言的前提和结论转换为FOL表达式，然后由外部定理证明器执行演绎推理。实验结果表明，LINC在两个数据集上相对于传统的基于提示的策略取得了性能提升，尤其是在ProofWriter数据集上，即使是相对较小的开源模型StarCoder+（15.5B参数）也超过了GPT-3.5和GPT-4。

2）Logic-LM — 通过符号求解器增强大语言模型的逻辑推理：Pan等人(2023)介绍了Logic-LM框架，该框架将大语言模型与符号求解器集成以改善逻辑问题求解。Logic-LM首先利用大语言模型将自然语言问题转换为符号公式，然后由确定性的符号求解器对公式进行推理。此外，该框架还包括一个自精炼模块，利用符号求解器的错误信息来修订符号公式。在五个逻辑推理数据集上的实验表明，与仅使用标准提示的大语言模型相比，Logic-LM平均性能提升了39.2%，与使用思维链提示的大语言模型相比提升了18.4%。

3) Logic Agent — 通过逻辑规则调用增强推理有效性: Liu等人(2024)提出了Logic Agent（LA），这是一个基于智能体（Agent）的框架，旨在通过策略性地调用逻辑规则来增强大语言模型在推理过程中的有效性。与常规方法不同，LA将大语言模型转变为能够动态应用命题逻辑规则的逻辑智能体，通过将自然语言输入转换为结构化逻辑形式来启动推理过程。LA利用一套全面预定义的函数系统地导航推理过程，不仅促进了推理结构的有序和一致性生成，还显著提高了它们的可解释性和逻辑一致性。

4) LMLP — 符号验证的逐步推理：Zhang等人(2024)探讨了通过符号验证评估逐步推理的方法。他们创建了包含等价（自然语言，符号）数据对的合成数据集，其中符号示例包含来自非参数知识库（KBs）的一阶逻辑规则和谓词，支持自动验证中间推理结果。研究者们重新审视了神经符号方法，并提出从包含逻辑规则和相应示例的示例中学习，以迭代地在知识库上进行推理，恢复Prolog的反向链接算法，并支持自动验证语言模型的输出。

5）Chameleon — 增强大语言模型的组合推理能力：Lu等人(2023)介绍了Chameleon，这是一个即插即用（plug-and-play）的组合推理框架，通过增强大语言模型与各种模块来解决大语言模型的固有局限性。Chameleon通过合成程序，组合不同的工具（例如大语言模型、现成的视觉模型、网络搜索引擎、Python函数和基于启发式的模块）来完成复杂的推理任务。Chameleon的核心是一个基于大语言模型的规划器，它组合了一系列工具的执行序列以生成最终响应。在ScienceQA和TabMWP两个多模态知识密集型推理任务上，Chameleon展示了其有效性，显著提高了最佳已发布结果的准确率。

这些方法展示了如何通过不同的方式将深度学习和符号推理相结合，以提高大语言模型在逻辑推理任务上的性能。通过这些创新的神经符号混合方法，研究者们能够开发出更可靠、更可解释且在复杂推理任务上表现更好的人工智能系统。

6 总结与展望

本文综述了大规模预训练语言模型（LLMs）在逻辑推理领域的最新进展。从理论基础到实践应用，我们全面梳理了逻辑推理的类型、特点，并回顾了相关理论的发展，为研究提供了清晰的框架。通过不同推理类型的案例研究，我们展示了大模型在逻辑推理方面的能力表现，并分析了提升大模型逻辑推理能力的方法，包括预训练、指令微调和神经符号混合方法。本文还探讨了大模型在逻辑推理任务上的现状和面临的挑战，并对如何提高模型的逻辑推理能力进行了深入讨论。

随着神经符号混合方法在提升大语言模型逻辑推理能力方面展现出巨大潜力，未来的研究可以从以下几个方向进行探索：

1）提高模型泛化能力：研究如何通过数据增强、多任务学习或元学习等技术提升模型对新领域或任务的适应性，以及通过跨领域知识转移增强模型的泛化能力。

2）优化模型可解释性：开发新的可视化工具和技术，帮助用户理解模型的推理过程，并通过逻辑规则的清晰表述和逻辑推理步骤的详细记录增强模型的可解释性。

3）探索新的神经网络架构：设计新的神经网络架构，这些可能包括逻辑操作专用的层类型或连接模式，以更有效地捕捉逻辑规则和推理过程中的关键特征。

4）增强模型鲁棒性：通过对抗训练、异常检测或冗余设计增强模型的鲁棒性，并探讨通过集成学习方法或多模型融合提高模型在不确定性条件下的推理能力。

5）跨领域知识融合：研究整合不同领域知识库的方法，以及设计算法处理和推理跨领域知识，利用领域特定的逻辑规则增强模型对特定领域知识的理解和应用。

6）多模态推理的进一步探索：集中于如何整合文本、图像、声音等多种模态的数据，并开发能够处理这些多模态输入的神经符号混合模型。

致谢

本文研究受到国家自然科学基金（No. 62336006）项目资助。张岳是本文通讯作者。

参考文献

Gabor Angeli, Neha Nayak, and Christopher D. Manning. 2016. Combining natural logic and shallow reasoning for question answering. In *Proc. of ACL*.

Sajjad Beygi, Maryam Fazel-Zarandi, Alessandra Cervone, Prakash Krishnan, and Siddhartha Reddy Jonnalagadda. 2022. Logical reasoning for task oriented dialogue systems.

Chen Bowen, Rune Sætre, and Yusuke Miyao. 2024. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 323–339, St. Julian's, Malta, March. Association for Computational Linguistics.

Ronnie Cann. 1993. *Formal semantics: an introduction*. Cambridge textbooks in linguistics. Cambridge University Press, United States. Includes bibliographical references (p. 333-338) and index.

Nuri Cingillioglu and Alessandra Russo. 2019. Deeplogic: Towards end-to-end differentiable logical reasoning.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *CoRR*, abs/2002.05867.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning.

Maksym Del and Mark Fishel. 2023. True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In Alexis Palmer and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 314–322, Toronto, Canada, July. Association for Computational Linguistics.

Jiazhan Feng, Ruochen Xu, Junheng Hao, Hiteshi Sharma, Yelong Shen, Dongyan Zhao, and Weizhu Chen. 2023. Language models can be logical solvers.

Will Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Embedding logical queries on knowledge graphs. *Advances in neural information processing systems*, 31.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, David Peng, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Shafiq Joty, Alexander R. Fabbri, Wojciech Kryscinski, Xi Victoria Lin, Caiming Xiong, and Dragomir Radev. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

John Haugeland. 1989. *Artificial Intelligence: The Very Idea*. The MIT Press, 01.

Chadi Helwe, Chloé Clavel, and Fabian Suchanek. 2022. Logitorch: A pytorch-based library for logical reasoning on natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3496–3509, Dublin, Ireland, May. Association for Computational Linguistics.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations.

Robert Kowalski. 1974. Predicate logic as programming language. In *IFIP congress*, volume 74, pages 569–544.

Douglas B Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Hanmeng Liu, Leyang Cui, Jian Liu, and Yue Zhang. 2021. Natural language inference in context - investigating contextual reasoning over long texts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13388–13396, May.

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023a. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023b. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023c. Logicot: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921.

Hanmeng liu, Zhiyang Teng, Ruoxi Ning, Jian Liu, Qiji Zhou, and Yue Zhang. 2023d. Glore: Evaluating logical reasoning of large language models.

Hanmeng Liu, Zhiyang Teng, Chaoli Zhang, and Yue Zhang. 2024. Logic agent: Enhancing validity with logic rule invocation.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online, June. Association for Computational Linguistics.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models.

Man Luo, Shrinidhi Kumbhar, Ming shen, Mihir Parmar, Neeraj Varshney, Pratyay Banerjee, Somak Aditya, and Chitta Baral. 2024. Towards logiglue: A brief survey and a benchmark for analyzing logical reasoning capabilities of language models.

J. McCarthy and P.J. Hayes. 1981. Some philosophical problems from the standpoint of artificial intelligence. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 431–450. Morgan Kaufmann.

John McCarthy. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*.

John McCarthy. 1978. History of lisp. In *History of programming languages*, pages 173–185.

John McCarthy, 1989. *Artificial Intelligence, Logic and Formalizing Common Sense*, pages 161–190. Springer Netherlands, Dordrecht.

Drew McDermott and Jon Doyle. 1980. Non-monotonic logic i. *Artificial intelligence*, 13(1-2):41–72.

A. Newell and H. Simon. 1956. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*.

Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. 2023. How well do sota legal reasoning models support abductive reasoning?

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore, December. Association for Computational Linguistics.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore, December. Association for Computational Linguistics.

Mihir Parmar, Neeraj Varshney, Nisarg Patel, Santosh Mashetty, Man Luo, Arindam Mitra, and Chitta Baral. 2023. Logicbench: A benchmark for evaluation of logical reasoning.

Fernando Carlos Neves Pereira. 1982. Logic for natural language analysis.

Molly Petersen and Lonneke van der Plas. 2023. Can language models learn analogical reasoning? investigating training objectives and comparisons to human performance. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16414–16425, Singapore, December. Association for Computational Linguistics.

Chengwei Qin, Wenhan Xia, Tan Wang, Fangkai Jiao, Yuchen Hu, Bosheng Ding, Ruirui Chen, and Shafiq Joty. 2024. Relevant or random: Can llms truly perform analogical reasoning?

Dongyu Ru, Changzhi Sun, Jiangtao Feng, Lin Qiu, Hao Zhou, Weinan Zhang, Yong Yu, and Lei Li. 2021. Learning logic rules for document-level relation extraction.

Samuel Ryb, Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. AnaLog: Testing analytical and deductive logic learnability in language models. In Vivi Nastase, Ellie Pavlick, Mohammad Taher Pilehvar, Jose Camacho-Collados, and Alessandro Raganato, editors, *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 55–68, Seattle, Washington, July. Association for Computational Linguistics.

Soumya Sanyal, Yichong Xu, Shuohang Wang, Ziyi Yang, Reid Pryzant, Wenhao Yu, Chenguang Zhu, and Xiang Ren. 2023. APOLLO: A simple approach for adaptive pretraining of language models for logical reasoning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6308–6321, Toronto, Canada, July. Association for Computational Linguistics.

Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the general deductive reasoning capacity of large language models using ood examples. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3083–3105. Curran Associates, Inc.

Daniel Selsam, Matthew Lamm, Benedikt Bünz, Percy Liang, Leonardo de Moura, and David L Dill. 2018. Learning a sat solver from single-bit supervision. *arXiv preprint arXiv:1802.03685*.

Jihao Shi, Xiao Ding, Li Du, Ting Liu, and Bing Qin. 2021. Neural natural logic inference for interpretable question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3673–3684, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. Clutrr: A diagnostic benchmark for inductive reasoning from text. *Empirical Methods of Natural Language Processing (EMNLP)*.

Michael Sullivan. 2024. It is not true that transformers are inductive learners: Probing NLI models with external negation. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1945, St. Julian's, Malta, March. Association for Computational Linguistics.

Richmond Thomason. 2024. Logic-Based Artificial Intelligence. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2024 edition.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through LogicNLI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3738–3747, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

William Van Melle. 1978. Mycin: a knowledge-based consultation program for infectious disease diagnosis. *International journal of man-machine studies*, 10(3):313–322.

Po-Wei Wang, Priya L. Donti, Bryan Wilder, and Zico Kolter. 2019. Satnet: Bridging deep learning and logical reasoning using a differentiable satisfiability solver.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Bimal Gajera, Shreeyash Gowaikar, Chandan Gupta, Aman Chadha, Aishwarya Naresh Reganti, Amit Sheth, and Amitava Das. 2023. ANALOGICAL - a novel benchmark for long text analogy evaluation in large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3534–3549, Toronto, Canada, July. Association for Computational Linguistics.

Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. 2023. Are large language models really good logical reasoners? a comprehensive evaluation and beyond.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. 2024. Symbol-llm: Towards foundational symbol-centric interface for large language models.

Fan Yang, Zhilin Yang, and William W Cohen. 2017. Differentiable learning of logical rules for knowledge base reasoning. *Advances in neural information processing systems*, 30.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024. Language models as inductive reasoners. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–225, St. Julian's, Malta, March. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, April.

Zhangdie Yuan, Songbo Hu, Ivan Vulić, Anna Korhonen, and Zaiqiao Meng. 2023. Can pretrained language models (yet) reason deductively? In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1447–1462, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Yi-Fan Zhang, Hanlin Zhang, Li Erran Li, and Eric Xing. 2024. Evaluating step-by-step reasoning through symbolic verification.

大模型时代的多语言研究综述

高长江　周昊　佘帅杰　钟昊鸣　刘斯哲　赖哲剑　王志军　黄书剑†
计算机软件新技术国家重点实验室, 南京大学
{gaocj, zhouh, shesj, zhonghm, liusz, laizj, wangzj}@smail.nju.edu.cn
huangsj@nju.edu.cn

摘要

进入大语言模型时代以来，传统的多语言研究模式发生了巨大变化。一些传统任务得到了突破性的解决，也出现了多种新任务，以及许多以多语言大模型为基础、面向大模型能力提升的多语言研究工作。本文针对研究领域中的这一新变化，整理归纳了进入大模型时代以来的多语言研究进展，包括多语言大模型、数据集、任务，以及相关的前沿研究方向、研究挑战等，希望能为大模型范式下的多语言研究的未来发展提供参考和帮助。

关键词： 大语言模型； 多语言； 跨语言

A Survey of Multilingual Research in the Large Language Model Era

Changjiang Gao, Hao Zhou, Shuaijie She, Haoming Zhong
Sizhe Liu, Zhejian Lai, Zhijun Wang, Shujian Huang†
National Key Laboratory for Novel Software Technology, Nanjing University
{gaocj, zhouh, shesj, zhonghm, liusz, laizj, wangzj}@smail.nju.edu.cn
huangsj@nju.edu.cn

Abstract

Since we enter the era of large language models, there have been significant changes in the traditional multilingual research paradigm. Some traditional tasks have been solved in a groundbreaking manner, new tasks have emerged, and many multilingual research works based on multilingual large language models and aimed at enhancing the capabilities of these models have been developed. This article focuses on this new change in the research field, summarizing the progress of multilingual research since the era of large models, including multilingual large language models, datasets, tasks, as well as related cutting-edge research directions, research challenges, etc., aiming to provide reference and assistance for the future development of multilingual research under the large model paradigm.

Keywords: Large Language Models , Multilingual , Cross-Lingual

1 引言

近年来，以 Transformer 模型 (Vaswani et al., 2017) 为基础架构，以大规模预训练 (Devlin et al., 2019) 为基本技术的大语言模型对自然语言处理领域的研究范式产生了革命性的影响 (Xu et al., 2024)。这些大模型不仅在多种自然语言处理任务上展现出了远超预期的性能，还显示

出了一些惊人的“涌现”能力，例如上下文学习 (Brown et al., 2020)，思维链能力 (Wei et al., 2023a; Wang and Zhou, 2024) 等。因此，面向大语言模型的研究，已经成为当前自然语言处理领域，乃至更广泛的人工智能领域中的重要趋势。其中，多语言研究作为自然语言处理与语言学学科的重点问题，也受到了强烈的影响。

由于训练语料中包含不同语言的大量数据，现有的大模型往往也是多语言大模型（MLLM）(Briakou et al., 2023; Qin et al., 2024)，并且展现出了强大的多语言能力，包括机器翻译能力 (Garcia et al., 2023; Fu et al., 2023; Li et al., 2024)，多语言推理能力 (Shi et al., 2022)。这些模型既可用于完成传统多语言任务，也为一些新的多语言应用场景提供了可能。同时，随着多语言预训练语料、指令对话数据、评估测试数据集的完善，多语言大模型的数量不断增加，质量也在不断提高。

然而，尽管多语言大模型的性能强大、应用前景广泛，此方面的研究仍面临问题与挑战。首先，多语言大模型的最适宜训练方式仍在完善中，在预训练、有监督微调（SFT）、人类反馈强化学习（RLHF）、下游任务微调等不同训练阶段中如何加入多语言信息 (Qin et al., 2024)，这些阶段的多语言训练分别有何影响 (Gao et al., 2024a)，都还在不断研究探索之中。其次，由于高质量训练语料以英文等高资源语言为主，多语言大模型往往表现出语言偏好与不平衡，在不同语言上展现出性能差距，以及不同语言的知识与语义表示没有对齐等。同时，能否拓展模型的语言能力，将已有的以英语为核心的大模型拓展为覆盖更多语言的模型，并且保持、迁移其英语性能，对实际应用也有重要指导意义。以及，由于深度学习模型的“黑箱”特点，研究者对多语言大模型如何产生多语言能力、内部计算过程与状态变化如何，仍没有明确的认识，需要通过新的观察手段与理论进行可解释性分析。最后，多语言大模型在不同语言上还面临公平性、安全性问题，灾难性遗忘与多语言诅咒，多语言训练代价大等挑战，需要进一步研究解决。

对此，本文将立足大模型时代的多语言研究，分别归纳多语言大模型的领域现状、前沿研究方向、面临的挑战，介绍多语言大模型的训练资源、训练手段、需要解决的问题等，为未来此方面的研究提供参考。

多语言模型类型	模型列表
基座模型	OpenAI ChatGPT, Google Gemini, BLOOM, PALM, LLaMA, Baichuan, DeepSeek, GLM, Qwen, ...
指令模型	Chines-LLaMA-Alpaca, BayLing, Vicuna, x-LLaMA/m-LLaMA, ...
特定任务模型	XGLM-7B, QAlign, MAPO, ...

Table 1: 代表性多语言大模型

多语言数据集类型	数据集列表
预训练语料	Wikipeida, mC4, ROOTs, CulturaX, ...
指令对话训练数据	xp3, OpenAssistant Conversation, Bactrain-X, Ayadataset, ...
评估测试数据集	理解任务：XCOPA, XStoryCloze, BELEBELE, XWinograd, PAWS-X, XNLI, ... 生成任务：XL-Sum, MLQA, MKQA, MGSM, FLORES, ...

Table 2: 代表性多语言数据集

2 研究现状概述

本部分将介绍目前多语言研究领域的基本情况，包括代表性多语言大模型、数据集、任务，并将特别介绍大模型时代新出现的多语言任务。

2.1 代表性多语言大模型

2.1.1 多语言基座模型

多语言基座模型是指那些在多种语言上进行了预训练，并且能够同时支持多种语言的模型。这些模型通常是在大规模的跨语言语料库上进行训练，以学习不同语言之间的共享特征和差异（见表 1）。

商业大模型，如 OpenAI GPT 系列模型，以及 Google Gemini 模型等，为目前全球应用较为广泛的多语言大语言基座模型，具有较强的多语言能力，且适用于不同下游任务。然而，由于模型架构、训练数据与训练方式均不公布，难以确定其多语言能力的基础与来源。

BLOOM (Scao et al., 2023) 是由 Big-Science 团队训练的的开源大模型，具有 176B 参数。BLOOM 在包含在 ROOTS 语料库上训练，能支持 59 种不同的语言，包括 46 种编程语言和 13 种自然语言。

PaLM (Chowdhery et al., 2022) 由 Google 发布，包含 540B 参数的大模型，在训练过程中使用了多语言维基百科和对话数据的混合，包括 124 种语言，其中英文词元在总词元中占比 78%。PaLM 在多语言测试上展现出强大的能力。

LLaMA 系列模型 (Touvron et al., 2023b) 是目前应用较为广泛的的开源大模型，其中 LLaMA-2 与 LLaMA-3 都在训练数据中有意提高了多语言预训练语料的占比，使模型的语言建模能力覆盖了较多不同语种。近期发布的 LLaMA-3 对模型的词表也进行了扩充，进一步提高了多语言能力。

此外，以 Baichuan (Yang et al., 2023a)，DeepSeek (DeepSeek-AI, 2024)，GLM (Zeng et al., 2022; Du et al., 2022)，Qwen (Bai et al., 2023) 等为代表的一批国产多语言基础模型，使用以中文与英文为主的多语言数据作为预训练语料，语料规模大、质量高且多样化，在中文、英文乃至其他语言的不同任务上都表现出了较强的性能。

2.1.2 多语言指令模型

多语言指令大模型是在以英语为中心的基座大模型基础上，通过多语言指令进行微调而得到的模型。这类模型通过使用多语言指令，将模型在英文上的理解、推理、生成能力扩展到了多语言，提升了模型在多语言上的表现。例如，Chinese-LLaMA-Alpaca 项目 (Cui et al., 2023) 对 LLaMA 系列模型进行了中文的词表扩充、继续预训练与指令微调，提高了中文的相关能力；BayLing (Zhang et al., 2023b) 通过交互式翻译数据和指令数据的混合微调，提高了模型的多语言翻译能力和指令能力；Vicuna (Chiang et al., 2023) 使用了全球各地用户与 ChatGPT 的交互数据微调，使 LLaMA 模型获得了较强的多语言指令能力；m-LLaMA (Zhu et al., 2023) 是在 LLaMA 上使用混合多种不同语言的指令微调数据，其指令微调数据包括了翻译数据和跨语言通用任务数据。

2.1.3 多语言特定任务模型

多语言特定任务大模型更侧重于特定多语言领域或任务的处理，如翻译、多语言问答等。这类模型通常在特定的多语言任务上进行微调，以优化模型在该任务上的性能。Li et al. (2024) 使用翻译数据微调模型，提升了模型的翻译性能。Zhu et al. (2024)，She et al. (2024) 在数学推理任务数据集上，将非英文语言与英文语言进行了对齐，使模型在非英语上表现出与英文相似的推理过程，提升了非英语推理能力。

2.2 代表性多语言数据集

在大型语言模型的训练过程中，数据是其中非常重要的组成部分。目前许多大型语言模型在训练中英语数据占比较大，因此大模型在英语上表现较好，会导致模型更多地倾向于学习英语语言的特征和结构。同时由于缺乏多语言数据集的训练，大多数模型在多语言环境下的性能存在较大差异，在非英语语言上的能力有所欠缺。

为了解决这一问题，有许多工作提出了使用多语言数据集来提升模型多语言能力的方法。通过在预训练阶段和微调阶段引入多语言数据集，模型可以更好地理解和处理各种语言的文本数据，从而提高其在多语言环境下的性能。

2.2.1 预训练语料

维基百科（Wikipeida）是质量较高的多语言数据集，具备较多优点，首先涉及领域众多（科学，历史，艺术，政治，商业），格式工整逻辑清晰。其次包含的语言种类也比较全面（一共支持 294 种语言）。并且维基百科语料会不断更新加入实时的新知识。但是，维基百科中不同语言之间的语料占比相差极大，这会导致仅使用该语料训练的模型（如编码器 mBERT (Kenton and Toutanova, 2019) 在低资源语言上能力较低。

mC4 (Xue et al., 2021) 数据集包含 108 种语言，是 Google 团队从公开的 Common Crawl 网页中爬取，一共包含 9.7T 左右的。在构建数据集时使用了启发式的算法进行自然语言的抽

取以及去重过滤，但是 mC4 数据集仍然存在语言识别度不高，会引入噪音的问题。目前 mT5 (Xue et al., 2021) 模型使用了 mC4 数据集进行端到端训练。

ROOTs（Responsible Open-science Open-collaboration Text Source）(Laurençon et al., 2022) 是 BigScience 团队用于训练 170B 的 BLOOM (Scao et al., 2023) 多语言大模型所用到的数据集。数据集大小约为 1.6 TB，由 59 种语言组成，其中包括 46 种自然语言 (以欧洲和亚洲的语言为主)和13种编程语言。ROOTS 的多语言数据是 BigScience 团队通过使用很多工具爬取（例如 BigScience Catalogue）以及从其他语言的存储库中收集了大量的多语言单语数据。代码数据是从 Github 和 StackExchange 中抽取而来。获得大量数据后，BigScience 团队还使用了许多方法来清洗过滤数据集，例如基于困惑度，字符重复程度来过滤低质量文本，使用 SimHash(Charikar, 2002; Manku et al., 2007) 以及后缀数组 (Manber and Myers, 1993) 的方法来去除重复文本。

CulturaX (Nguyen et al., 2023a) 是由美国俄勒冈大学标注完成的。总括包括 167 种语言，6.3T 词元。在大模型时代，很多多语言大模型，如 BLOOM，PolyLM(Wei et al., 2023b)等，其训练数据集不完全公开，阻碍了开源社区对于 LLM 的分析与理解，例如在无法了解预训练数据集的情况下，开发者们很难对模型产生的幻觉以及有毒的偏见进行归因和分析。CulturaX 创立的初衷是为了提供一个更开明的多语言数据集以供训练，增加了多语言大模型的透明度，让开发者更深入分析和理解多语言大模型。CulturaX 从公开的 OSCAR (Abadji et al., 2022) 和 mC4 数据集中爬取，既保证了内容的时效性（截止至 2023 年 1 月），又保证了低资源语言的占比不至于过低。俄勒冈大学团队使用了一套全新的架构来在数据清洗中去重噪音文本，非自然语言，有毒偏见数据，并有效提升了语言识别的准确率。

2.2.2 指令对话训练数据

在多语言大模型预训练后，模型存储了大量的多语言知识，但此时模型还无法根据人类指令输出正确的回答内容。基于此，许多工作贡献了多语言的指令对话数据，旨在帮助模型获得多语言的指令遵循能力。

xp3 数据集 (Muennighoff et al., 2023) 也是由 BigScience 团队标注的数据集，主要包括 46 种语言，16 种 NLP 任务，用于训练 BLOOMZ 和 mT0。xp3 数据集在 p3 (Victor et al., 2022) 数据集的基础上新增了许多的多语言数据集，并且语言占比分布与 ROOTs 数据集趋近相同。虽然 xp3 数据集的指令数量较多，但是在某些语言还是存在噪音过多的情况。

OpenAssistant Conversation (Köpf et al., 2024) 数据集是由 OpenAssistant 人工进行标注的多语言对话数据集，质量较高，涉及 35 种语言，一共 13 万数据。OpenAssistant Conversation 是以对话树的结构组织而成，具体来说，这一个拥有多轮对话，且对于同一个问题有不同回复的数据集。

Bactrain-X (Li et al., 2023b) 数据集是由阿布扎比大学团队标注的，涉及 51 种语言，每种语言有 6 万 7 千条数据，阿布扎比大学团队首先收集了开源的英语指令数据集 Alpaca (Taori et al., 2023) 和 Dolly (Conover et al., 2023)，其次将指令和输入数据使用谷歌翻译引擎翻译成目标语言,将问题输入到 gpt-3.5-turbo 收集回答。这样组成的（指令，输入，输出）就可作为扩展语言的指令数据集。由于 gpt-3.5-turbo 自身在低资源语言上的局限性，Bactrain-X 在低资源语言上存在较大噪音。

Ayadataset (Singh et al., 2024) 是一个旨在减少语言不平等的重要数据集，通过人工筛选和多语种数据收集而成。其数据规模庞大，包含了来自全球 119 个国家的 2997 名合作者的努力共同创建的 204,114 个高质量注释，涵盖了 65 种语言。为了获取这一规模庞大的数据集，Aya 项目与来自世界各地的精通者合作，收集人工筛选的指令和完成实例。通过这种方式，能够获取到更加真实和准确的数据，而不受自动筛选和机器翻译的影响。

2.2.3 评估测试数据集

目前，衡量大型多语言模型的多语言能力是一个备受关注的重要议题。当前的研究工作已经整理并标注了多种语言的文本数据集，这些数据集涵盖了从简单任务到复杂推理任务的广泛范围。评估大型多语言模型的多语言能力通常涉及在这些数据集上执行一系列任务，如语言理解、语言生成等。通过对这些数据集进行全面评估，可以更加深入地了解大型多语言模型在多语言环境下的性能表现，从而为模型的改进和优化提供有力指导（见表 2）。

多语言理解数据集 XCOPA (Ponti et al., 2020) 是一个关于常识推理的数据集，一共覆盖 11 门语言，包括海地克里奥尔语等低资源语言，每个语言含有 600 条数据。与传统的选择题中包含 4 个选项不同，XCOPA 中，每一条数据中的问题仅对应两个选项。

XStoryCloze (Lin et al., 2022)：由 Meta 团队评测 XGLM 模型时将英文的 StoryCloze (Mostafazadeh et al., 2016) 的验证集翻译而来，一共包括 11 门语言。StoryCloze 数据集主要考察模型对于一段故事的理解能力，具体来说，它要求模型从给定的选项中选出一个故事的正确结局。

BELEBELE (Bandarkar et al., 2023):由 Meta 团队标注，是一个涵盖了 122 种语言变体的多项选择机器阅读理解（MRC）数据集，可以更好的测评模型在高资源，中资源，低资源语言上的能力。每个语言包括 900 条数据，每个数据由 1 个段落，1 个问题和四个选项组成，其中段落由 FLOERS-200 (Costa-jussà et al., 2022) 数据集组成。

XWinograd (Muennighoff et al., 2023; Tikhonov and Ryabinin, 2021)：用于评估模型的指代消解能力和常识推理能力的数据集，一共包含 7 种语言，不同语言含有的数据量差异较大。

PAWS-X (Yang et al., 2019) 是一个关于复写的多语言数据集，主要包括 7 种语言，每种语言含有 5 万条左右的数据，其中 2 万 3 千条是人工翻译，另外 2 万 9 千条是机器翻译。

XNLI (Conneau et al., 2018) 是一个关于句子理解的多语言数据集，覆盖 15 种语言。每种语言含有 40 万条数据，每条数据的结构大致如下：给定两个句子，判断这个句子的关系（蕴含，中立，矛盾）。

除上述公开发表的数据集以外，俄勒冈大学团队使用了 DeepL，gpt-3.5-turbo 等模型将很多英文数据集（例如 MMLU (Hendrycks et al., 2020)，ARC 数据集 (Clark et al., 2018)，Truthful_qa 数据集 (Lin et al., 2021)，HellaSwag 数据集 (Zellers et al., 2019)）也翻译成了很多语言，以便更好评测多语言大模型。

多语言生成数据集 XL-Sum (Hasan et al., 2021) 是一个多语言的摘要数据集，涵盖语言范围较广，一共支持 44 种语言。数据来自 BBC 的 135 万篇专业标注的文章摘要对，使用一组经过精心设计的启发式方法提取。XLSum 具有高度抽象、简洁且高质量的特点。

MLQA (Lewis et al., 2019) 是一个多语言问答数据集，覆盖了 7 门语言，MLQA中每个语言包含了5 千个左右的（其中英文有 1 万 2 千个）抽取式问答实例，不同语言之间数据是高度并行的。

MKQA (Longpre et al., 2020) 是一个开放领域的问答评估数据集，涵盖了 26 种语言的数据，每种语言包含 10,000 个问题-答案对，总共 260,000 个问题-答案对。答案基于经过精心策划的、不依赖于特定语言的段落，使得结果可以在各种语言之间进行比较。

MGSM (Cobbe et al., 2021; Shi et al., 2022) 数据集将 GSM8K 的测试集翻译成了其他 10 种语言，每个语言包含 256 个数据，是一个小学难度的数学推理数据集。

FLORES (Costa-jussà et al., 2022) 是一个多语言翻译数据集，数据来源于维基百科，收集到 2000 个句子，由专业人员从英语翻译成其他 200 种语言。

2.3 代表性多语言任务

在大模型时代，多语言处理任务迎来了重大的变革。传统上，多语言任务依赖于针对特定语言对或小范围语言族的专门模型，但随着大模型如 GPT-3、LLaMA (Touvron et al., 2023b) 等的兴起，我们开始看到一种全新的范式。这些大模型通常在多种语言上预训练，能够同时处理多种语言任务，无需特定于某一语言的调整。这种方法不仅简化了模型的部署，还提高了处理低资源语言的能力。接下来，我将分两部分详细介绍代表性的多语言任务：传统任务的大模型方法和大模型时代的新任务。

2.3.1 传统任务的大模型方法

翻译 在当今的大型模型时代，翻译技术已经从以往的 Encoder-Decoder 架构，逐步演变为仅使用 Decoder 的模型。这样的转变主要由于仅 Decoder 模型，比如 GPT 系列 (Radford et al., 2018; Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023)，在执行语言生成任务时表现出的高效能和灵活性。不同于传统的 Encoder-Decoder 模型，仅 Decoder 模型在进行翻译时无需通过一个编码步骤来理解源语言，而是可以直接在生成步骤中将源语言文本转译为目标语言文本。例如，在执行翻译任务时，仅 Decoder 模型会接收一个用特定源语言写成的句子，并

在前面加上一个指示目标语言的特定前缀或标记，如“Translate to French:”。模型依靠其预先训练的知识来理解源语言，并直接开始生成相应的目标语言文本。

在探索大型语言模型的多语言能力时，设计翻译任务的主要目的是为了验证和增强模型处理和理解多种语言的能力。通过这些翻译任务，研究人员和开发者能够深入探讨模型如何把握并运用不同语言之间的语义和语法结构，以及如何将这些知识应用于准确地将一种语言转换为另一种语言。

多/跨语言信息抽取，摘要，阅读理解 大型语言模型如 ChatGPT 可以处理多语言或跨语言的任务，如信息抽取、摘要和阅读理解，主要是通过有效的提示词设计和模型训练方法来实现。模型的训练数据包含多种语言，使其能够理解和生成多语言内容。 在执行具体任务时，如信息抽取或摘要，设计的提示词会指导模型关注于文本的特定部分或特定类型的信息。 例如，如果需要从一篇中文文章中提取信息并生成英文摘要，可以设计一个提示词让模型首先理解和提取中文文章中的关键信息，然后将这些信息转换成英文摘要。 这一过程通常依赖于模型的多语言理解能力和生成能力，以及其在训练过程中学到的语言转换技能。

此外，通过调整提示词的具体内容和结构，可以优化模型的表现，使其更好地完成特定的多语言或跨语言任务。例如，在要求模型进行阅读理解时，可以通过明确的问题提示模型关注文章的哪一部分内容，或是用何种语言输出答案，以此来提高信息处理的准确性和效率。

命名命名实体识别 在大型语言模型中，完成命名实体识别（Named Entity Recognition, NER）主要依赖于模型训练和任务特定的提示词设计。 具体来说，通过精心设计的提示词和适当的少样本学习方法，可以有效地实现 NER 任务。

首先，在提示词设计方面，需要构造一个清晰的任务说明，使模型明白所需执行的具体任务是识别文本中的命名实体。 例如，可以设计一个提示词：“请识别以下文本中的所有人名、地名和组织名称。” 紧接着，提供一段文本，让模型进行实体识别。 其次，使用少样本学习 (Wang et al., 2020c) 不仅提高 NER 任务的效果，同时可以控制模型的输出格式，方便提取出模型输出中的命名实体标签。

情感分析，文本分类 在这种情况下，模型利用其预训练期间学到的知识来直接对文本进行分类，无需任何特定任务的训练。 例如，要进行情感分析，可以给模型提供一个文本并询问：“这段文本的情感是正面的、负面的还是中性的？”模型会根据其预训练中学到的知识来预测答案。 虽然 Decoder-only 的大模型在没有额外训练数据的情况下能够直接进行文本分类和情感分析，但在某些情况下，通过在特定任务的数据上进一步训练（微调）模型可以获得更加可控的输出结果。 对于分类问题而言，也可以通过限制模型只能输出词表中几种特定单词的方式，获得可控输出。

2.3.2 大模型时代的新任务

多语言上下文学习 大模型如 GPT-4 通过上下文学习 (Dong et al., 2022)，在没有显式训练的情况下，仅通过提示指导即可适应新任务。 这种方法特别适用于处理多语言数据，模型可以通过观察少量的例子迅速调整其行为。

多语言指令服从 新一代的大模型如 Codex 和 GPT-4 在多语言环境中不仅可以处理文本任务，还能理解并执行包括代码编写 (Chen et al., 2021)、数学问题解答 (Cobbe et al., 2021) 等复杂任务。 这显示了大模型在处理跨领域、跨语言问题时的强大能力。

总的来说，大模型时代为多语言任务带来了前所未有的机遇和挑战。通过预训练的大型模型，我们能够以前所未有的规模和效率处理多语言信息，但这也对模型的泛化能力和可解释性提出了新的要求。在设计未来的多语言处理系统时，我们需要考虑这些因素，以充分利用大模型的潜力，同时克服其局限性。

3 前沿研究方向

本部分将重点介绍目前多语言大模型研究的若干前沿方向，包括模型训练、跨语言对齐、语言能力扩展，以及多语言大模型内部机制的可解释性分析。图 1 展示了本文所述的多语言研究的不同角度概况。

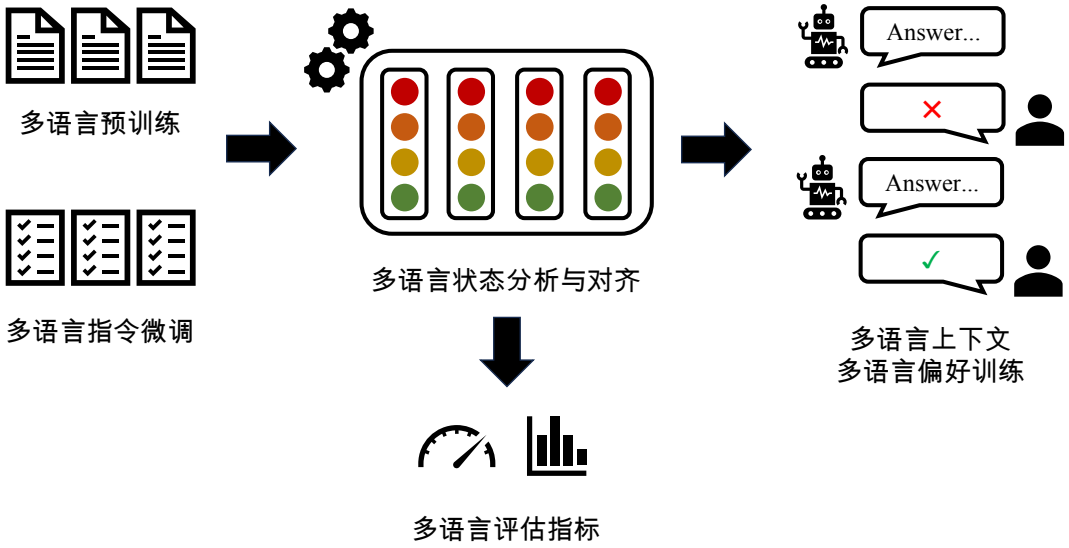


Figure 1: 不同角度的多语言研究示意图

3.1 多语言模型训练

为了得到具有更好多语言能力的大模型，现有的工作从预训练，指令微调以及强化学习偏好优化三个阶段进行。

3.1.1 预训练

在预训练阶段，随机初始化的大模型将在大规模的语料库上做无监督训练，获取知识和语言能力。一个很直接的想法是在预训练语料中加入多语言数据。现有的大模型的预训练语料库中往往包含了一部分多语言数据，如Qwen (Bai et al., 2023)，Mistral (Jiang et al., 2023)，mT5 (Xue et al., 2021)，Erine (Sun et al., 2021)，BLOOM (Scao et al., 2023),LLama1 (Touvron et al., 2023a)，LLama2 (Touvron et al., 2023b)，PolyLM (Wei et al., 2023b)，InternLM (Team, 2023) ，具备一定的多语言能力基础。

尽管在预训练阶段加入多语言的预训练语料有助于获得多语言能力，然而扩充语言数量削弱了每一个语言以及每一个任务的有效信息容量，语种之间竞争有限的模型容量 (Conneau et al., 2020)。因此在预训练阶段，一个需要考虑的问题是各个语言数据的配比问题。多语言的预训练语言模型通常采用基于启发式的温度采样方法来平衡不同语言之间的权重 (Devlin et al., 2019)，对一些低资源语言进行上采样。然而这样的方法容易造成低资源语言的过拟合，也降低了学习效率 (Hernandez et al., 2022; Lee et al., 2021)。一些工作展开了进一步探索，Wang et al. (2020b) 在 DDS (Wang et al., 2020a) 的基础上进行了扩展和改进，提出了 MultiDDS 算法，通过学习一个语言评分器，以优化多语言数据的使用，从而在多种不同语言上实现良好的性能。 Chung et al. (2023) 提出了 UNIMAX，在保证头部语言得到更均匀覆盖的同时，通过明确限制每种语言语料库的重复次数来减轻尾部语言的过拟合问题。

从头开始训练大模型的代价往往相对高昂，也没有办法很好的利用已经训练好的大模型。因此基于已有大模型的继续预训练就成了一个很好的选择。Tang et al. (2020) 展示了一种从已有预训练语言模型如mBART开始继续训练的方法，在拓展语言的同时，很好的保留了已有语言上能力，成功的将原始mBART支持的25种语言拓展到了50种。除了支持语种的大幅度扩充之外，也有一些工作关注于特定一些语言的能力提升，基于已有的英语为中心的大模型，在对应的相关语言的预训练数据上进行了大量的预训练，获得了一定的能力提升，如CPM-2 (Zhang et al., 2021)，Chinese-Llama (Cui et al., 2023)，Chinese-Mixtral (HIT-SCIR, 2024)，SeaLLM (Nguyen et al., 2023b)，Sailor (Dou et al., 2024)等等。

3.1.2 指令微调

在指令微调阶段，模型会在用指令或者模板形式构造好的数据上训练，可以很好的激发模型在预训练阶段习得的知识。因此，通过添加多语言的指令微调能够使模型具备一定程度的多语言指令响应能力。有一些工作 (Muennighoff et al., 2023; Cui et al., 2023; HIT-SCIR, 2024;

Nguyen et al., 2023b; Dou et al., 2024; Wei et al., 2023b; Csaki et al., 2024; Chen et al., 2023a) 收集了来自不同语言的指令微调数据，用于训练模型。这部分的训练数据有综合的指令响应任务、也包括了具体的如数学推理，翻译等任务。还有一些工作 (Muennighoff et al., 2023; Xue et al., 2021) 通过将原有的多语言多任务数据，通过指令模板的形式转化为多语言指令微调数据。相对而言多语言的指令微调数据相对稀缺，因此 BayLing (Zhang et al., 2023b) 通过交互式翻译任务将英语语言生成和指令遵循能力转移到非英语语言的指令遵循型大型语言模型，在多项评估中展现了不错的效果。

3.1.3 人类偏好对齐

RLHF 阶段往往用于调整模型的行为，让模型更好的和人类偏好对齐。这个阶段需要有部分人工标注的偏好数据，通常情况下体现为对于一个指令，利用策略模型采样输出，并人工对这些采样结果进行偏好打分。当前广泛使用的算法有 PPO (Schulman et al., 2017)、DPO (Rafailov et al., 2023)、IPO (Liu et al., 2019b)、CPO (Achiam et al., 2017) 等等。现有工作 (Yang et al., 2023a; Bai et al., 2023; Cui et al., 2023; Du et al., 2022; Wei et al., 2023b; Team, 2023; Chen et al., 2023a) 为了更好的将模型在各个语言上和人类偏好对齐，在这个阶段使用多语言的偏好标注数据来训练。上述的偏好优化算法往往需要基于一个经过多语言指令微调的大模型作为初始化的策略模型，后续提出的 ORPO (Hong et al., 2024) 偏好优化算法可以直接从预训练模型开始训练，节约计算资源消耗的同时也取得了很好的效果。

3.2 跨语言对齐

跨语言对齐，也可以称为跨语言一致性、跨语言迁移等，指的是通过一定技术手段，使得多语言大模型在不同语言上的知识、能力、行为等方面保持一致 (Wang et al., 2023a; Qi et al., 2023; Gao et al., 2024a; Qin et al., 2024)。此类研究的出发点主要有三：1）使大模型可以在不同语言上表现出更平衡的性能，服务不同语言背景的用户 (Wang et al., 2023a)；2）使模型能够复用在不同语言的数据上学习到的知识和偏好，提高训练效率 (Gao et al., 2024a)；3）大多数事实知识、指令服从能力、安全性偏好都是与语言种类无关的，因此模型在这些方面理应具有跨语言对齐的表现 (Ohmer et al., 2023; Wu et al., 2024)。目前，大语言模型的能力日益提高，但它们的跨语言对齐程度却并不令人满意。因此，此方面的研究在将来一段时间内也具有重要意义。 本部分将首先介绍目前对大语言模型跨语言对齐程度的评估工作；随后将根据观察角度与实现方式的不同，分别介绍训练数据、内部状态、模型偏好、指令上下文四个层面的跨语言对齐研究。

3.2.1 对大模型跨语言对齐程度的评估

现有工作对大模型的跨语言对齐程度主要从通用任务能力与事实知识两方面进行了评估。

通用任务能力方面，Lai et al. (2023a) 评估了 ChatGPT 在 37 种语言上执行 7 种代表性 NLP 任务的性能，并与 mT5-XXL (Xue et al., 2021) 等多语言预训练模型对比，发现 ChatGPT 服从非英语指令、完成非英语任务的能力低于英语，且与有监督学习基线的差距较大。Wang et al. (2023a) 提出了 SeaEval 基准测试，用于测量多语言大模型在英语、中文、印地语、越南语等语言上完成包括经典 NLP 任务、复杂推理以及文化理解在内的一系列任务的表现，发现目前的多语言大模型在不同语言的事实、科学和常识性知识等方面表现出了较大的不一致，未达到“平衡的多语言能力”。Zhang et al. (2023a) 考察了部分多语言大模型在语码切换（code-switching）场景下的情感分析、机器翻译、摘要和词级别语种识别任务，发现大模型在这些任务上的性能远不如专门微调过的小模型，提示大模型的多语言能力不能直接泛化到语码切换场景。

事实知识方面，Qi et al. (2023) 针对事实知识的跨语言一致性，提出了 RankC 指标，用于评估 BLOOM, mT5, XLM-RoBERTa (Liu et al., 2019a) 等模型在 17 种语言上的事实知识一致性，发现模型参数增加虽然能使事实准确率更高，但无法提高跨语言一致性；同时，用英语对模型进行知识编辑时，新知识只能迁移到与英语有较高 RankC 分数的语言，说明模型的跨语言事实知识一致性水平较为有限。Wang et al. (2023c) 同样也关注了知识编辑中的跨语言效应，通过构造英语-中文的跨语言知识，评估了英语知识编辑能否对中文对应知识产生影响，发现此种跨语言影响十分有限。Gao et al. (2024a) 则提出了 CLiKA 评测框架，将模型的多语言知识对齐分为性能、一致性、传导性三个层面，评估了现有多语言大模型在 10 种语言上的跨

语言知识对齐程度，并比较了多语言预训练、多语言指令微调对对齐程度的影响，发现当前模型的跨语言对齐程度不够理想，且其跨语言一致性可能主要来源于训练数据的重合，而非语言间的知识传导；同时，尽管多语言混合预训练和指令微调能带来更好的跨语言性能平衡和一致性，它们并不能提高模型的跨语言知识传导程度。

3.2.2 训练数据层面的跨语言对齐

促进模型跨语言对齐能力的常见方法，是在训练数据中加入多语言内容，包括平行语料、非平行语料、多语言指令等，本文将其称为训练数据层面的跨语言对齐。

在大模型时代以前，已经出现了在多语言预训练模型的训练数据中加入多语言内容的尝试。Conneau and Lample (2019) 提出了分别使用多语言非平行语料、平行语料进行预训练的方法，得到了 XLM 系列模型，并在跨语言 NLP 任务、机器翻译等领域上取得了提升。NLLB 团队 (2022) 收集了包含 200 种语言的翻译数据，得到了 NLLB 系列大规模多语言翻译模型，取得了领先的翻译性能。Vu et al. (2022) 则基于 mT5 模型，使用统一的 prompt tuning 框架 (Lester et al., 2021) 在多种语言上进行混合的有监督与无监督微调，在跨语言生成任务上取得了较大提升；还提出了将提示词各部分分离的 factorized prompt 格式，缓解灾难性遗忘。同时，以英语数据为主训练的模型也展现出了一定的多语言能力，Blevins and Zettlemoyer (2022) 研究了这一现象，并提出这可能是由于英语预训练语料中含有比例极低，但绝对数量较多的非英语数据，这些数据的含量与模型的多语言能力密切相关。Briakou et al. (2023) 针对英语预训练模型表现出的零样本或少样本机器翻译能力，调查了 PaLM (Chowdhery et al., 2022) 模型的训练数据，发现其中含有 3000 万以上、涵盖超过 44 种语言的翻译数据对，称为“偶然双语现象”，并认为这些数据与模型的翻译能力相关。这些研究说明了在预训练过程中加入多语言数据，对提升模型的多语言对齐程度较为重要。

进入大模型时代后，上述研究的思路得到了延续。一些大模型在预训练阶段加入了大量多语言非平行数据或翻译数据，与英语数据混合预训练 (Anil et al., 2023; Bai et al., 2023; Schioppa et al., 2023; Touvron et al., 2023b; Wei et al., 2023b; Scao et al., 2023; Yang et al., 2023a; Yang et al., 2023b)；也有一些模型以英语为主的基座模型为基础，在新语言上继续预训练 (Cui et al., 2023; Larcher et al., 2023)；还有一部分大模型在微调阶段加入多语言任务指令数据，包括对话数据、推理数据、翻译数据等，提高模型的多语言任务执行能力 (Cahyawijaya et al., 2023; Chen et al., 2023b; Kew et al., 2023; Muennighoff et al., 2023; Wei et al., 2023b; Scao et al., 2023; Yang et al., 2023a; Yang et al., 2023b; Zhang et al., 2023b; Jiao et al., 2023; Chen et al., 2024; Chung et al., 2024; Gao et al., 2024b; Li et al., 2024; Zhu et al., 2024; Lai et al., 2023b)。

3.2.3 内部状态层面的跨语言对齐

数据层面的方法是通过输出阶段的监督信号来实现跨语言对齐，但这种方法的粒度较粗，无法在更细的计算过程视角观察跨语言对齐。因此，研究者尝试了构建模型的内部状态对不同语言的相同语义数据的对齐。内部状态层面的跨语言对齐研究起源于无监督翻译领域，后者的核心思想是自动构建不同语言的语义嵌入空间之间的同构映射 (Lample et al., 2018; Søgaard et al., 2018; Cao et al., 2019; Naorem et al., 2024; Shen et al., 2024)，或自动推断双语词典 (Yu et al., 2023b; Ding et al., 2024; Garnier and Guinet, 2024)。

应用到大模型中，有研究使用对比学习等方法，拉近模型在不同语言中的同义上下文表示，从而提高模型的整体多语言能力 (Efimov et al., 2023; Li et al., 2023a; Tao et al., 2023)，或构建更好的多语言嵌入向量 (Zhang et al., 2023c; Zhao and Eger, 2023)。具体到任务中，Wang et al. (2023b) 通过实体表示的对齐，提高模型对知识图谱的跨语言学习效率；Xu et al. (2023b) 利用句法和语义的联合学习，构建适用于文本分类任务的跨语言嵌入模型；Chen et al. (2023b) 引入了一个全局指令表示向量，用于提高模型的翻译忠实度。

在用于提升模型跨语言对齐程度的同时，表示层面的对齐观察也为模型运行机制的可解释性提供了新的视角。Zhao et al. (2023) 通过类比神经科学中的脑区定位方法，在多语言大模型中找到了与多语言能力相关的少量参数，称之为语言核心区。Wendler et al. (2024) 通过精心构造的翻译任务场景，观察了 LLaMA-2 系列模型在处理跨语言数据时的内部状态变化，观察到了其中存在近似正交的“输入空间”、“概念空间”、“输出空间”的证据，且“概念空间”与英语较其他语言更为接近。Zhao et al. (2024) 则提出了平行语言特定神经元探测（PLND）方法，用

于衡量模型的不同神经元在处理多语言输入时的重要性，并认为模型在处理多语言数据时，可能在内部状态中先将输入翻译为英语，进行处理后再翻译回原输入语言。这些研究为进一步提高内部状态层面的跨语言对齐提供了理论支持。

3.2.4 模型偏好与上下文层面的跨语言对齐

除了利用数据层面的任务监督信号，内部状态层面的表示对比信号，还可以通过控制模型输出时的行为偏好与指令上下文，促进其跨语言对齐能力。

行为偏好层面，指的是通过强化学习方法，使模型倾向于给出跨语言一致的回答。这种方法可以追溯到使用强化学习和贝叶斯风险最小化方法优化的神经机器翻译模型 (Ramos et al., 2023; Yang et al., 2024)。应用到大模型的跨语言对齐方面，She et al. (2024) 将跨语言对齐作为一种受鼓励的倾向，提出了 MAPO 框架，针对数学推理场景，以强化学习方法促进模型在英语和非英语场景下给出相同的推理过程的倾向，使模型在相应的评测数据集上展现出了显著性能提升和更高的跨语言推理一致性。Wu et al. (2024) 提出了人类反馈强化学习（RLHF）阶段中的奖励模型的零样本跨语言迁移方法，将用英语数据训练的奖励模型迁移到非英语，从而促进模型行为倾向、安全性等方面的跨语言对齐。

指令上下文层面，指的是不改变模型参数，仅通过设计合适的上下文样例或提示词来提高模型的跨语言对齐程度。Huang et al. (2023) 提出了跨语言思维提示（XLT）方法，让模型先将问题翻译为英语，将其形式化并分布解答，激发模型的跨语言逻辑推理能力，发现可以显著提升模型对各种多语言任务的性能，并缩小英语与非英语的性能差距。Kim et al. (2023) 针对多语言 QA 任务，提出了 In-CLT 方法，在大模型的上下文学习样例中同时提供源语言篇章与目标语言问题，发现能够显著提高模型在目标语言上的 QA 任务性能。Puduppully et al. (2023) 提出了 DecoMT 方法，通过少样本提示，让模型将句子级别翻译任务分解为多个词组级别任务，提高了模型的翻译能力。Yong et al. (2023) 通过让模型模拟双语使用者的提示，鼓励模型生成英语-东南亚语言的高质量语码切换输出。

3.3 语言能力扩展

目前，大型语言模型如 LLaMA 和 Mistral 主要以英语为核心，在处理其他语言时性能明显下降。鉴于此，学术界展开了广泛的研究，旨在扩展和增强这些模型的多语言能力。这些研究可以被概括为两个主要方向：一是单一模型的多语言扩展，二是多模型融合以共同增强多语言处理能力。

在单一模型的多语言扩展方面，研究人员致力于探索如何使单个大模型在处理多种语言时保持高效性能。这包括但不限于多语言继续预训练和跨语言微调等技术。通过这些方法，研究人员希望实现在保持模型在英语上表现优异的同时，提高其在其他语言上的适用性和性能。在涉及到继续预训练的工作中，BigTranslate (Yang et al., 2023b)，Tower (Alves et al., 2024)，ALMA (Xu et al., 2023a) 工作都说明了继续使用单语训练大模型会极大提升模型的多语言能力 （翻译能力），其中ALMA指出，在给定算力的条件下，将更多的算力分配给单语而不是翻译数据会更好的提升模型的翻译能力，在这种实验设置下就可以让大模型的翻译能力与传统的监督模型相接近。在涉及到指令微调的技术中，Bayling (Zhang et al., 2023b)，xllama (Zhu et al., 2023)，SDRRL (Zhang et al., 2024) 都使用了多语言指令微调数据集来提升大模型在下游多语言任务上的性能，其中 xllama，SDRRL 都指出，在指令微调时混入大量的翻译数据会更好帮助模型把存储在英语上的知识迁移到目标语言中。

另一方面，多模型融合的研究聚焦于如何结合多个语言模型的知识和能力，以增强整体的多语言处理能力。这涉及到模型融合技术、集成学习方法等方面的研究。通过将不同语言模型的优势互补结合，研究人员期望实现在跨多语言环境下更为有效的文本处理能力。在模型融合方法借助多模态 LLaVA 架构，LangBridge (Yoon et al., 2024)指出，将含有多语言知识的小模型 （例如 mT5）与大模型 MetaMath (Yu et al., 2023a) （推理能力较强，但多语言能力较差）使用一个简单的线性层相融合，在仅使用英语数据下就可以极大提升大模型多语言的推理能力。同时，Bansal et al. (2024) 也尝试将较低资源语言的知识存储到较小的 PaLM 模型中，将较小的，存储较多多资源语言知识的 PaLM 模型与较大的 PaLM 模型使用 cross-attention 的方式去完成模型融合，增强了较大 PaLM 模型在低资源语言上的翻译能力。还有工作 (Blevins et al., 2024) 尝试使用 BTM (Branch Traiin Merge）的方法，对于同一个模型，不同的语言分别独立训练成多个模型后再融合，这种方法在一定程度突破了“多语言诅咒” (Wu and Dredze,

2020)。在集成学习方面，Farinhas et al. (2023) 在大模型完成多语言翻译任务时，采用了不同的生成方法采样了多个结果 （greedy search，beam search）,最终采取集成的方式选中最终候选回答，这样使大模型的翻译能力得到了巨大的长进。

3.4 多语言机制分析

目前的大语言模型普遍具有较强的多语言能力，而大语言模型内部是如何对不同语言进行处理，是一个值得探索的问题。 目前，有许多工作尝试对大语言模型内部处理多语言的机制进行分析。

其中一部分工作从模型生成下一词元时中间隐层状态的语义表示空间出发，将模型的前向计算过程按照不同层分为多个阶段，发现了在不同阶段之间存在语言转换的现象，并在其中部分阶段发现了不同语言间存在对齐的现象。 Wendler et al. (2024) 等人使用 Logit Lens 方法 (nostalgebraist, 2020) 发现，在训练语料主要是英语的大语言模型中（如 LLaMA），即使大语言模型在输出非英语的文本，模型也会在中间层先生成对应于英语词元的隐层状态，再在高层转换成对应输出语言的词元。 具体来说，其认为大语言模型的前向计算可以从低到高分为三个阶段：上下文信息的理解、抽象概念的生成以及从抽象概念到输出词元的转换。特别的，在以英语为中心的大语言模型中，可以观察到在第二阶段中生成的抽象概念的隐层状态正对应于输出词元的英文表示。 类似的，Zhao et al. (2024) 等人提出了 PLND 方法来检测大语言模型中特定于某种语言的神经元。他们发现，在模型的低层和高层中与非英语语言相关的神经元更多，而在中层与英语相关的神经元更多，由此提出了类似的“多语——英语——多语”三阶段模型。他们认为模型在最底层和最高层进行的理解和生成阶段是语言相关的，而在中层进行解决问题时会使用英语的思考能力以及不同语言的事实性知识。 由于大语言模型在处理不同的语言时，都可能会在中间阶段归并到某种语言无关的概念表示，再通过第三阶段进行语言相关的目标词元生成，因此某些低资源语言也能通过与高资源语言（如英语）共享相似的中间表示的方式，共享模型从其他语言学到的通用能力。

另外一部分工作则从模型的计算部件——神经元的角度出发进行分析研究。 Tang et al. (2024) 等人的工作中提出可以将大语言模型中的神经元分为通用神经元与语言特定的神经元，通用神经元在生成各种不同语言时都会被激活，而语言特定神经元在生成特定语言时才会被激活。进一步，他们提出了 LAPE 方法来检测模型中的语言特定神经元。由此，他们发现大语言模型对某种特定语言的处理大部分都是来源于模型顶部和底部的一小部分神经元，并且发现可以通过选择性地干预这些语言特定神经元的激活与抑制，达到控制大语言模型的输出语言的目的。

4 研究面临的挑战

本部分将介绍目前的多语言大模型研究面临的几项重要挑战。

4.1 多语言场景下的公平性与安全性

4.1.1 多语言场景下的公平性

虽然多语言技术的发展使得大模型在非英文上的能力逐渐增强，但模型在不同语言上的差异很难被完全消除，这引发了模型在多语言的公平性问题 (Shliazhko et al., 2022)。由于低资源语言的训练语料稀缺且质量较差 (Yu et al., 2022)，提升大模型在低资源语言上的能力十分困难。除了不同语言性能上的公平性，一些研究 (Levy et al., 2023; Piqueras and Søgaard, 2022) 还揭示了不同语言间的知识、种族、宗教、性别存在着差异。此外，模型在不同语言上的词元占比不同，不同语言词元的切分长度也存在较大差异，这造成了不同语言的计算开销的差异 (Petrov et al., 2024; Ahia et al., 2023)。

4.1.2 多语言场景下的安全性

大型语言模型在理解和生成自然语言方面表现出卓越的能力。然而，这种能力也引发了对潜在安全性问题的担忧，包括生成不安全的内容，例如侮辱或歧视性语言 (Sun et al., 2023) 或泄露私人信息 (Macko et al., 2023)。在多语言场景下，安全性问题更为突出。一方面，多语言数据的清洗和筛选相比于英文数据更加困难，当前大部分多语言数据集的清洗都是不充足的 (Nguyen et al., 2023a)。另一方面，缺少多语言安全性评测基准。Wang et al(Wang et al., 2023d) 构造了覆盖十种不同语言的安全性评测数据集 XSAFETY，是目前唯一的多语言安全性

评测数据集。但是，XSAFETY 覆盖的语言数量有限，构建一个更加全面的多语言安全性评估数据集，仍然是多语言安全性研究的迫切需求。

4.2 多语言诅咒问题

多语言诅咒问题最早由 Conneau et al. (2020) 在多语言预训练模型的研究中提出，具体定义为：当模型的容量固定时，如果不断增加多语言训练数据，模型的多语言能力（尤其是低资源语言能力）会先上升，但在到达一定程度后，几乎所有语言的单语言、跨语言能力都会下降。这种现象提示了模型的多语言能力提升有一定的边界，即多语言能力不能超过模型参数规模、基础语言能力等决定的上限。此后，许多工作都试图突破这一限制，例如使用混合专家模型，为不同语言分配不同的专家 (Blevins et al., 2024; Pfeiffer et al., 2022)。这些工作在缓解多语言诅咒的方面取得了一定效果，但同时也造成了新的问题。例如：当语言种类不断增加时，专家数量需要不断增加，导致模型参数量过大；语言专家中存在重复、冗余参数，并且分专家存储参数的模式对跨语言对齐也产生了挑战。这些问题都需要在未来给予更大的关注。

4.3 多语言训练的代价

在大模型时代，多语言训练的代价是一个非常重要的议题。多语言模型的训练需要大量的多语言数据 (Kaplan et al., 2020)。这些数据不仅要覆盖多种语言，还需要在各种语言之间保持高质量和平衡性。数据的采集包括从不同地区和文化背景收集文本，这通常涉及到昂贵的数据收集开销。此外为了确保数据的多样性和无害性 (Askell et al., 2021; Perez et al., 2022)，可能需要人工干预来纠正数据问题，那么还要付出额外的人工标注费用。另外，多语言模型通常需要比单一语言模型更大的参数空间和更复杂的网络结构 (Qin et al., 2024)，以便在多种语言之间进行知识共享和转移。这导致了更高的计算成本，包括但不限于 GPU、TPU 等高性能计算资源的使用。此外，这些资源的使用通常需要较长的时间，增加了能源消耗和相应的环境影响 (Scao et al., 2023)。

5 总结

多语言大模型是多语言研究领域在未来一段时间内的重点研究对象。本文从领域情况概述、前沿研究方向、研究面临的挑战三个方面介绍了大模型时代的多语言研究进展，提供了使用、训练、改进和分析多语言大模型所需的参考研究信息。我们希望这项工作能为多语言大模型的将来研究提供可能的帮助。

参考文献

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France, June. European Language Resources Association.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. *arXiv preprint arXiv:2305.13707*.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick,

Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report, September.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report, September.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Rachit Bansal, Bidisha Samanta, Siddharth Dalmia, Nitish Gupta, Shikhar Vashishth, Sriram Ganapathy, Abhishek Bapna, Prateek Jain, and Partha Talukdar. 2024. Llm augmented llms: Expanding capabilities through composition. *arXiv preprint arXiv:2401.02412*.

Terra Blevins and Luke Zettlemoyer. 2022. Language Contamination Helps Explain the Cross-lingual Capabilities of English Pretrained Models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv preprint arXiv:2401.10440*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM's Translation Capability. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada, July. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-Low Resource Language Alignment via Continual Crosslingual Instruction Tuning, October.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*, September.

Moses S Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Ye Chen, Wei Cai, Liangmin Wu, Xiaowei Li, Zhanxuan Xin, and Cong Fu. 2023a. Tigerbot: An open multilingual multitask llm.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023b. Improving Translation Faithfulness of Large Language Models via Augmenting Instructions, August.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*. Association for Computational Linguistics, March.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways, October.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling Instruction-Finetuned Language Models. *Journal of Machine Learning Research*, 25(70):1–53.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. *Company Blog of Databricks*.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation, August.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca, June.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model, May.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Qiuyu Ding, Hailong Cao, and Tiejun Zhao. 2024. Enhancing Bilingual Lexicon Induction via Bi-directional Translation Pair Retrieving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17898–17906, March.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland, May. Association for Computational Linguistics.

Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. 2023. The Impact of Cross-Lingual Adjustment of Contextual Word Representations on Zero-Shot Transfer. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval*, pages 51–67, Cham. Springer Nature Switzerland.

António Farinhas, José GC de Souza, and André FT Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. *arXiv preprint arXiv:2310.11430*.

Tingchen Fu, Lemao Liu, Deng Cai, Guoping Huang, Shuming Shi, and Rui Yan. 2023. The Reasonableness Behind Unreasonable Translation Capability of Large Language Model. In *The Twelfth International Conference on Learning Representations*, October.

Changjiang Gao, Hongda Hu, Peng Hu, Jiajun Chen, Jixing Li, and Shujian Huang. 2024a. Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly, April.

Pengzhi Gao, Zhongjun He, Hua Wu, and Haifeng Wang. 2024b. Towards Boosting Many-to-Many Multilingual Machine Translation with Large Language Models, February.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation, February.

Paul Garnier and Gauthier Guinet. 2024. Semi-Supervised Learning for Bilingual Lexicon Induction, February.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

HIT-SCIR. 2024. Chinese-mixtral-8x7b: An open-source mixture-of-experts llm. https://github.com/HIT-SCIR/Chinese-Mixtral-8x7B.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting, May.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Wenxiang Jiao, Jen-tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. ParroT: Translating during Chat using Large Language Models tuned with Human Translation and Feedback. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15009–15020, Singapore, December. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed?, December.

Sunkyoung Kim, Dayeon Ki, Yireun Kim, and Jinsik Lee. 2023. Boosting Cross-lingual Transferability in Multilingual Models via In-Context Learning, May.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023a. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning, April.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023b. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback, August.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*, February.

Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius Caridá. 2023. Cabrita: Closing the gap for foreign languages, August.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2021. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. *arXiv preprint arXiv:2305.11242*.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2023a. Align after Pre-train: Improving Multilingual Generative Models with Cross-lingual Alignment, November.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023b. Bactrian-x: A multilingual replicable instruction-following model with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions, April.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach.

Yongshuai Liu, Jiaxin Ding, and Xin Liu. 2019b. Ipo: Interior-point policy optimization under constraints.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Samuel Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, et al. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. *arXiv preprint arXiv:2310.13606*.

Udi Manber and Gene Myers. 1993. Suffix arrays: a new method for on-line string searches. *siam Journal on Computing*, 22(5):935–948.

Gurmeet Singh Manku, Arvind Jain, and Anish Das Sarma. 2007. Detecting near-duplicates for web crawling. In *Proceedings of the 16th international conference on World Wide Web*, pages 141–150.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual Generalization through Multitask Finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July. Association for Computational Linguistics.

Deepen Naorem, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024. Improving linear orthogonal mapping based cross-lingual representation using ridge regression and graph centrality. *Computer Speech & Language*, 87:101640, August.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023a. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. *arXiv preprint arXiv:2309.09400*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2023b. Seallms – large language models for southeast asia.

nostalgebraist. 2020. interpreting gpt: the logit lens. *LessWrong*.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. 2023. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. *CoRR, abs/2305.11662*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July. Association for Computational Linguistics.

Laura Cabello Piqueras and Anders Søgaard. 2022. Are pretrained multilingual models equally fair across languages? *arXiv preprint arXiv:2210.05457*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed Prompting for Machine Translation Between Related Languages using Large Language Models, October.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models, October.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers, April.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André F. T. Martins. 2023. Aligning Neural Machine Translation Models: Human Feedback in Training and Inference, November.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina

Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, June.

Andrea Schioppa, Xavier Garcia, and Orhan Firat. 2023. Cross-Lingual Supervision improves Large Language Models Pre-training, May.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing Multilingual Reasoning through Multilingual Alignment-as-Preference Optimization, April.

Yingli Shen, Wei Bao, Ge Gao, Maoke Zhou, and Xiaobing Zhao. 2024. Unsupervised multilingual machine translation with pretrained cross-lingual encoders. *Knowledge-Based Systems*, 284:111304, January.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the Limitations of Unsupervised Bilingual Dictionary Induction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia, July. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Hao Sun, Zhexin Zhang, Jiawen Deng, Jiale Cheng, and Minlie Huang. 2023. Safety assessment of chinese large language models. *arXiv preprint arXiv:2304.10436*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models.

Qian Tao, Zhihao Xiong, Bocheng Han, Xiaoyang Fan, and Lusi Li. 2023. A Novel Unsupervised Approach for Cross-Lingual Word Alignment in Low Isomorphic Embedding Spaces. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3027–3041.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Alexey Tikhonov and Max Ryabinin. 2021. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models, July.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Sanh Victor, Webson Albert, Raffel Colin, Bach Stephen, Sutawika Lintang, Alyafeai Zaid, Chaffin Antoine, Stiegler Arnaud, Raja Arun, Dey Manan, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. 2022. Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9279–9300, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-Thought Reasoning Without Prompting, February.

Xinyi Wang, Hieu Pham, Paul Michel, Antonios Anastasopoulos, Jaime Carbonell, and Graham Neubig. 2020a. Optimizing data usage via differentiable rewards. In *International Conference on Machine Learning*, pages 9983–9995. PMLR.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020b. Balancing training for multilingual neural machine translation. *arXiv preprint arXiv:2004.06748*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020c. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F. Chen. 2023a. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning, September.

Chenxu Wang, Zhenhao Huang, Yue Wan, Junyu Wei, Junzhou Zhao, and Pinghui Wang. 2023b. FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs. *Information Fusion*, 89:41–52, January.

Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, and Jiarong Xu. 2023c. Cross-Lingual Knowledge Editing in Large Language Models, September.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023d. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, January.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023b. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse Your Rewards: Reward Model Transfer for Zero-Shot Cross-Lingual Alignment, April.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023a. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Yuemei Xu, Wanze Du, and Ling Hu. 2023b. A Cross-lingual Sentiment Embedding Model with Semantic and Sentiment Joint Learning. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 82–94, Cham. Springer Nature Switzerland.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias, April.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. Paws-x: A cross-lingual adversarial dataset for paraphrase identification. *arXiv preprint arXiv:1908.11828*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, JunTao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023a. Baichuan 2: Open Large-scale Language Models, September.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023b. BigTranslate: Augmenting Large Language Models with Multilingual Translation Capability over 100 Languages, July.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2024. Direct Preference Optimization for Neural Machine Translation with Minimum Bayes Risk Decoding, April.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages. In Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali, editors, *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore, December. Association for Computational Linguistics.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. Langbridge: Multilingual reasoning without multilingual supervision. *arXiv preprint arXiv:2401.10695*.

Xinyan Velocity Yu, Akari Asai, Trina Chatterjee, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: a survey of multilingual dataset construction and necessary resources. *arXiv preprint arXiv:2211.15649*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023a. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Shenglong Yu, Wenya Guo, Ying Zhang, and Xiaojie Yuan. 2023b. CD-BLI: Confidence-Based Dual Refinement for Unsupervised Bilingual Lexicon Induction. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 379–391, Cham. Springer Nature Switzerland.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2022. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*, September.

Zhengyan Zhang, Yuxian Gu, Xu Han, Shengqi Chen, Chaojun Xiao, Zhenbo Sun, Yuan Yao, Fanchao Qi, Jian Guan, Pei Ke, Yanzheng Cai, Guoyang Zeng, Zhixing Tan, Zhiyuan Liu, Minlie Huang, Wentao Han, Yang Liu, Xiaoyan Zhu, and Maosong Sun. 2021. Cpm-2: Large-scale cost-effective pre-trained language models.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023a. Multilingual Large Language Models Are Not (Yet) Code-Switchers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore, December. Association for Computational Linguistics.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023b. BayLing: Bridging Cross-lingual Alignment and Instruction Following through Interactive Translation for Large Language Models, June.

Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. 2023c. VECO 2.0: Cross-lingual Language Model Pre-training with Multi-granularity Contrastive Learning, April.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. *arXiv preprint arXiv:2402.12204*.

Wei Zhao and Steffen Eger. 2023. Constrained Density Matching and Modeling for Cross-lingual Alignment of Contextualized Representations. In *Proceedings of The 14th Asian Conference on Machine Learning*, pages 1245–1260. PMLR, April.

Jun Zhao, Zhihao Zhang, Yide Ma, Qi Zhang, Tao Gui, Luhui Gao, and Xuanjing Huang. 2023. Unveiling A Core Linguistic Region in Large Language Models, October.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism?

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question Translation Training for Better Multilingual Reasoning, February.

大语言模型合成数据方法简述

李培基*
上海人工智能实验室
复旦大学计算机科学技术学院
lipeiji@pjlab.org.cn

马逸川*
上海人工智能实验室
上海交通大学电子信息与电气工程学院
mayichuan@pjlab.org.cn

颜航
上海人工智能实验室
yanhang@pjlab.org.cn

摘要

大语言模型在过去两年受到了极大的关注，并引起了对通用人工智能的广泛讨论。为了实现通用人工智能，合成数据被认为是其中非常关键的一环。本文将当前常见的数据合成方法归为三类，基于蒸馏的合成数据、基于模型自我进化、基于工具的合成数据。针对每一类合成数据方法，我们简要介绍了几种主流的做法，以期概览各类方法的基本思路以及异同。当前大部分合成数据方法都基于蒸馏，尽管这些方法取得了良好的效果，但其实质是将更强的大模型蒸馏到更小的大模型。这样的方法从降低大模型推理成本的角度具有实际意义，但对于进一步提升大模型能力上限作用有限。基于模型自我进化和基于工具的合成数据研究相对偏少，对于持续提升模型能力，这两个方向需要有更多探索。

关键词： 合成数据 ； 大语言模型 ； 知识蒸馏

A Brief Introduction to Synthetic Data for Large Language Model

Abstract

Large language models have gained significant attention over the past two years, sparking extensive discussions on artificial general intelligence (AGI). Synthetic data is considered a crucial component in achieving AGI. This paper categorizes the current common synthetic data methods into three types: distillation-based, model self-evolution-based, and tool-based. For each type, we briefly introduce several mainstream approaches to provide an overview of the basic ideas and differences among these methods. Most current synthetic data methods are based on distillation. Although these methods have achieved good results, their essence lies in distilling a more powerful large model into a smaller one. This approach is practically significant for reducing the inference cost of large models but has limited effectiveness in further enhancing the upper limit of large model capabilities. Research on model self-evolution-based and tool-based synthetic data is relatively scarce, and more exploration is needed in these two directions for continuously improving model capabilities.

Keywords: Synthetic Data , Large Language Model , Knowledge Distillation

* 共同一作

第二十三届中国计算语言学大会论文集，第86页-第97页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

大语言模型（Large Language Model，简称LLM）自2022年11月底ChatGPT问世以来受到了各个行业的持续关注，由于其出色的文字掌握能力和逻辑推理，LLM被广泛应用于代码生成[0]、文案写作、AI for Science (Zhang et al., 2024; Jablonka et al., 2024)等领域。尽管各类最先进大模型的具体训练方法并未披露 (OpenAI, 2023; Anil et al., 2023; Anthropic, 2024)，但数据被普遍认为是决定大模型能力的最重要因素 (Touvron et al., 2023a; Bai et al., 2023; Cai et al., 2024b)。

规模定律（Scaling Law）表明大模型的性能会随着使用训练数据量的增加而持续攀升 (Kaplan et al., 2020; Hoffmann et al., 2022)，是否能够持续产生训练数据将很大程度上决定大模型未来的能力上限。然而Epoch AI的预测表明人类将在2026年耗尽高质量数据[1]，尽管近期的一些研究表明可以通过重复使用数据的方式继续提高模型的性能 (Muennighoff et al., 2023)，但这种性能收益会在多次重复之后持续衰减。因此通过合成数据的方式来持续获得高质量训练语料成了一个值得研究的科学问题。除了数据短缺带来的对合成数据的需求，数据的长尾效应也使得在某些场景下必须要借助合成来进行补充，例如一些复杂数学推理 (Trinh et al., 2024a)和代码数据 (Li et al., 2022)，这些场景下需要比较专业的人才来标注相关数据，这类数据在天然文本中数量较少，同时标注成本较高。一些常识推理的数据，尽管对人类来说难度不高，但这类推理路径数据在自然文本中分布也很少。此外，在某些场景下可以直接利用工具进行大量数据合成，从而让模型掌握某些特定能力，例如四则运算 (Yang et al., 2023b; Yuan et al., 2023)。

在本文中，我们将合成数据定义为：借助大模型或者工具生产的数据。在此定义下，本文根据合成数据使用到的方法将过去的工作分成了三类：基于蒸馏的合成数据、基于模型自我进化的合成数据以及外部工具合成数据。其中基于蒸馏的相关工作主要集中在从性能更优的私有模型中获取训练数据，在开源大模型上进行继续训练，缩小开源模型与闭源模型间的性能差异 (Xu et al., 2023; Yu et al., 2023; Wang et al., 2024)；基于模型自我进化的合成数据从前景上，更有可能解决数据的短缺问题，目前这类方法效果提升相较前一种方案不明显，但这类方法未来有很好的发展前景，特别是对于如何持续提升大语言模型性能上限上有重要意义 (Wang et al., 2023b; Zelikman et al., 2022)；除了前两种需要借助语言模型的方法，还可以通过配合使用工具 (Yue et al., 2023; Singh et al., 2024)或者完全依赖工具来构造训练数据 (Yuan et al., 2023; Trinh et al., 2024a)，借助工具的方法可以利用工具的可靠性，提升合成数据的准确性，例如通过规则构造的四则运算数据就不会存在错误，但如何在合适的领域利用对应的工具需要领域知识。尽管自然语言处理领域早在ChatGPT诞生之前就在使用合成数据训练模型，例如ZeroGen (2022)利用生成式模型为判别式模型生成训练数据，Ding等人 (2023)使用GPT-3来进行数据标注训练模型，但本文的讨论的方法主要集中在ChatGPT之后，并且主要以呈现近期合成数据主流方法为主，每种方法主要介绍几篇相关工作，如需更详细和全面的了解，请参阅近期的相关综述论文 (Xu et al., 2024; Liu et al., 2024)。

接下来本文将首先介绍以上三类合成数据方法的典型做法，然后针对这些方法的优缺点以及当前合成数据的不足进行讨论，最后提出对合成数据发展的展望。

2 基于蒸馏的合成数据

知识蒸馏（Knowledge Distillation）通过一定的方法将大模型的知识迁移到小模型上，这样可以在推理的时候使用更小的模型获得相近的性能 (Lin et al., 2021)。过去的知识蒸馏方法一般假设可以获取到大模型的输出概率分布 (Gou et al., 2021)，甚至可以获取大模型的中间层输出 (Jiao et al., 2020)，但随着模型能力的增强，一些商业模型选择了闭源其模型，用户只能拿到其预测的结果。在这种情况下，只能通过模仿大模型的输出来实现蒸馏 (Wallace et al., 2020; Wang et al., 2023b)。自OpenAI于2022年11月底发布ChatGPT以来，各家商业公司的大模型能力不断攀升 (OpenAI, 2023; Anil et al., 2023; Anthropic, 2024)，这些模型都选择了闭源，开源社区模型如果想要复刻这些模型的性能，最简单的方法便是蒸馏这些模型的能力。

[0] https://github.com/features/copilot

[1] https://epochai.org/blog/will-we-run-out-of-ml-data-evidence-from-projecting-dataset

在这一节中，我们将讨论利用一些更强模型，例如ChatGPT、GPT-4 (OpenAI, 2023)，来提升开源模型的方法。由于这些方法都利用了更强的模型，所以它们的性能增益很大一部分来源于对更强模型的知识蒸馏。在过去两年涌现的相关方法，由于没有办法直接获取到模型的输出概率分布，基本均通过让语言大模型生成文本回答，模仿该回答来提升模型能力 (Shridhar et al., 2022; Magister et al., 2022; Fu et al., 2023)。这些方法都等同于扩大了训练数据量（通过扩展了回答的推理路径或者直接扩充了更多的数据条数），这类方法从本质上也类似于传统的数据增强（Data Augmentation） (Feng et al., 2021; Xu et al., 2024)。根据这些方法是否需要利用标注数据标签（例如数学题目的答案），我们将这类方法进一步细分为：借助监督数据标签进行合成和无需监督数据标签的数据合成。

2.1 借助监督数据标签进行合成数据

在OpenAI公司的GPT-3 (Brown et al., 2020)问世之后，人们发现其在经过思维链（Chain-of-Thought）的提示之后可以显著提高推理任务的回答准确率 (Wei et al., 2022)，这种特性同时也在谷歌公司推出的大模型PaLM上得到了印证 (Chowdhery et al., 2023)。经验性的结论发现，模型的参数量需要超过一定的量级才能涌现出一些能力 (Schaeffer et al., 2023)。但这些模型都太大了，推理成本高昂，研究者们希望能够找到方法缩小模型的尺寸同时保持其推理能力，因此Magister (2022)、Shridhar (2022)和Ho (2022)等人尝试了通过借助语境学习（In-Context Learning）的方法来生成思维链数据。

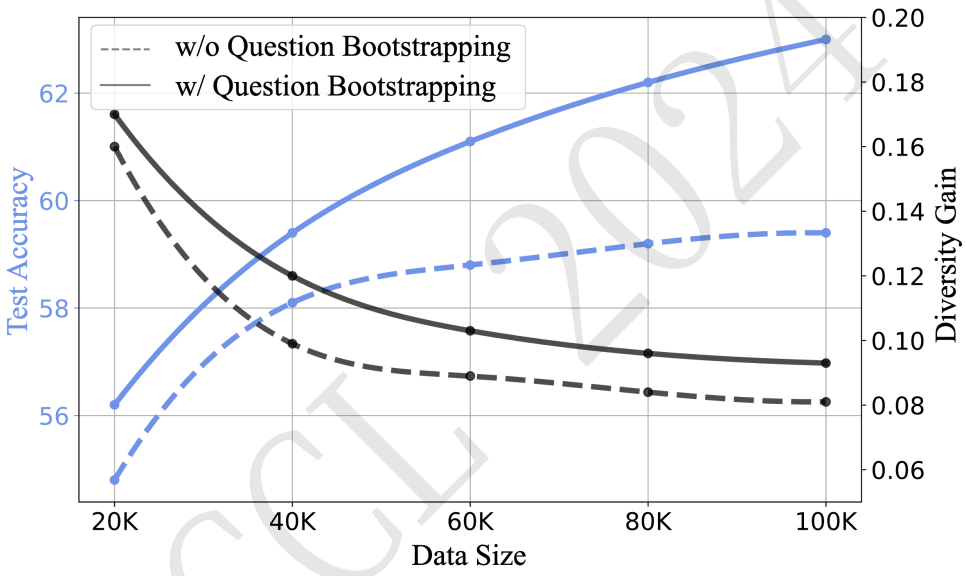


Figure 1: 模型性能随着微调数据量增加的曲线，如果不对数据中的问题进行增广（蓝色虚线），随着训练数据量增加，模型性能增长非常缓慢(Yu et al., 2023)

利用上述的方法可以获得大量的思维链训练数据，然而问题的数量被固定为了原数据集中问题的数量，这会导致训练数据的问题多样性不足，影响模型性能持续提升，如图 1所示。为了规避这个问题，研究者们提出了多种方法来可靠地增加问题的数量。MetaMath (Yu et al., 2023)提出可通过同义改写题目、逆向推理的题目增强以及同义改写答案来增加数据多样性。其中逆向推理的推理题目增强方法为任意将原文中的某个数字更换为未知数，让模型求解其未知数，例如原题为“农场的母鸡每日可以产蛋10颗，3天可产蛋多少颗？”，可修改为“农场的母鸡每日可以产蛋x颗，3天可产蛋30颗。问x的值为多少？”。不论是同义改写还是增强题目方法，都需要借助一个聪明的模型，通过输出最终的正确答案来筛选生成的数据，在该论文中使用了OpenAI的GPT-3.5-Turbo来进行数据生产。

通过借助程序来计算数学解答过程中的一些数值计算可以避免出现因为计算带来的错误，MAmmoTH (Yue et al., 2023)提出解题过程中可以生成类似于思维链的程序链（Program of Thought）来进行解题，即在问题求解过程中直接将解答过程写成可执行的代码，通过代码解释器执行完这些代码得到的结果作为预测结果，这样可以让针对相同的问题提出不同的解

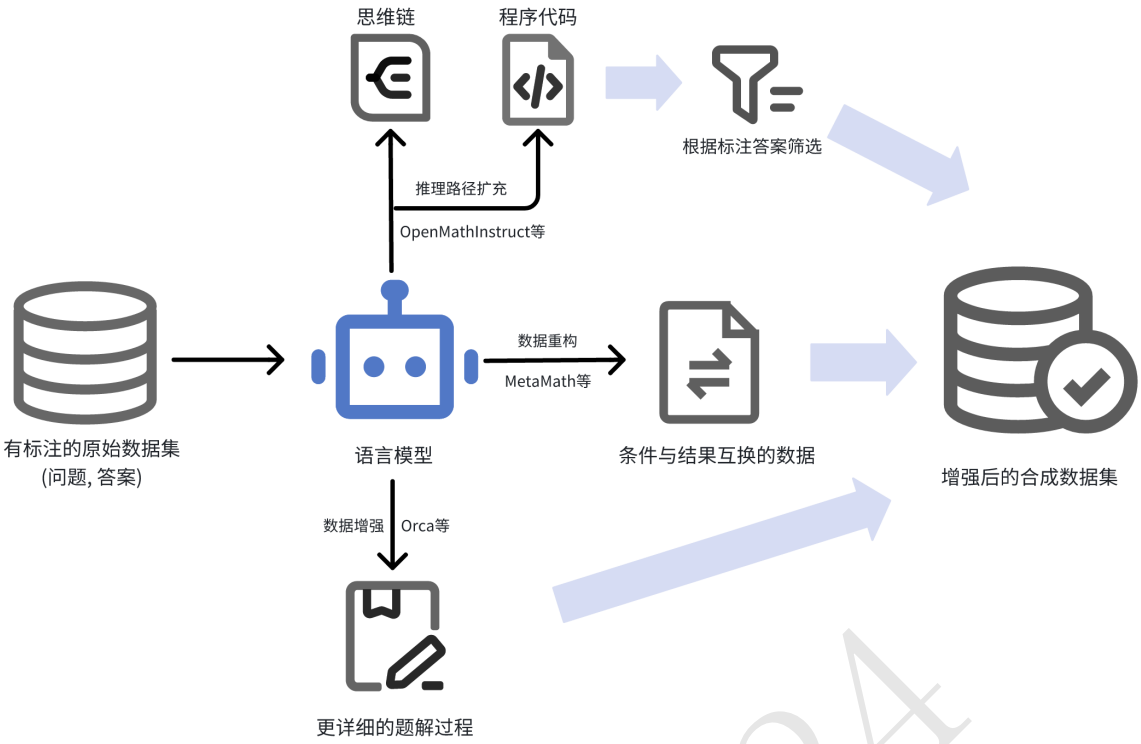


Figure 2: 借助监督数据标签可通过各种方法扩充训练数据

法。同样类似于直接生成答案的方法，如果代码执行完毕之后，预测结果与答案不符合则认为生成代码有误，不予采用。

如果有监督数据标签，除了通过上述的丢掉不符合答案的推理路径方案外，还可以尝试直接将答案用于解答过程的生成。Orca (Mukherjee et al., 2023)尝试将FLAN (Longpre et al., 2023)中的数据集通过大语言模型推理构建回答，并在模型回答前提前将答案拼接到输入中，然后输入诸如下面的提示词“假设我只有5岁，请一步步思考，并将为什么答案是这样解释给我听”，这样可以让小模型学习并模仿大模型的思考过程。在Orca中，作者总共通过GPT-3.5生成了500万条数据，其中的100万条通过GPT-4进行了再次生成。尽管这份数据没有直接开源，社区在这个思想的基础上进行了复现 (Lian et al., 2023)，并将复现结果进行了开源 [2]。

在本小节中我们介绍的几种方法被汇总到了图 2，这类方法由于利用了监督信号来辅助数据生产，所以可以在一定程度上保证生成数据的质量，使得这种方法生成的训练数据在下游任务提升上效果都比较显著。此外，类似的思想也可以用于代码生成之后，在代码中，可以通过生成的代码能否通过单元测试来判断生成的代码是否有误。但正如Orca (Mukherjee et al., 2023)指出的那样，能够利用的数据不够多是这类方法一个比较明显的缺点，尽管Orca里面利用了FLAN来扩大可利用的监督数据集，但这个数据量离预训练数据量还有两到三个数量级的差异。

2.2 无需监督数据标签的合成数据

在过去的蒸馏方法中，可以借助大量的无标签数据集来将大模型的能力迁移到小模型上。同样地，在大语言模型时代仍然可以利用相同的思想。在这一节中我们主要讨论三个相关的方法：（1）借助进化的思想，让合成数据的难度不断增加；（2）通过大模型生成更具教育意义的数据；（3）从预训练语料中挖掘高质量数据。

WizardLM (Xu et al., 2023)设计了Evol-Instruct算法来避免依赖人类生产高质量的指令。Evol-Instruct算法的思想如图 3所示，通过给大模型施加指令，使其在一个初始指令的基础上逐步增加指令的深度和宽度，从而不断提高指令难度。由于不确定大模型会从哪个方向

[2] https://huggingface.co/datasets/Open-Orca/OpenOrca

演化指令，因此生成的指令也不一定有标准答案，此刻就需要利用大模型来根据生成的指令进行回答，收集这些复杂指令和对应的回答可以构成训练数据集用于微调开源模型。通过迭代不断生成更复杂指令的思想也被广泛使用在了后续的工作中 (Luo et al., 2023a; Luo et al., 2023b)。

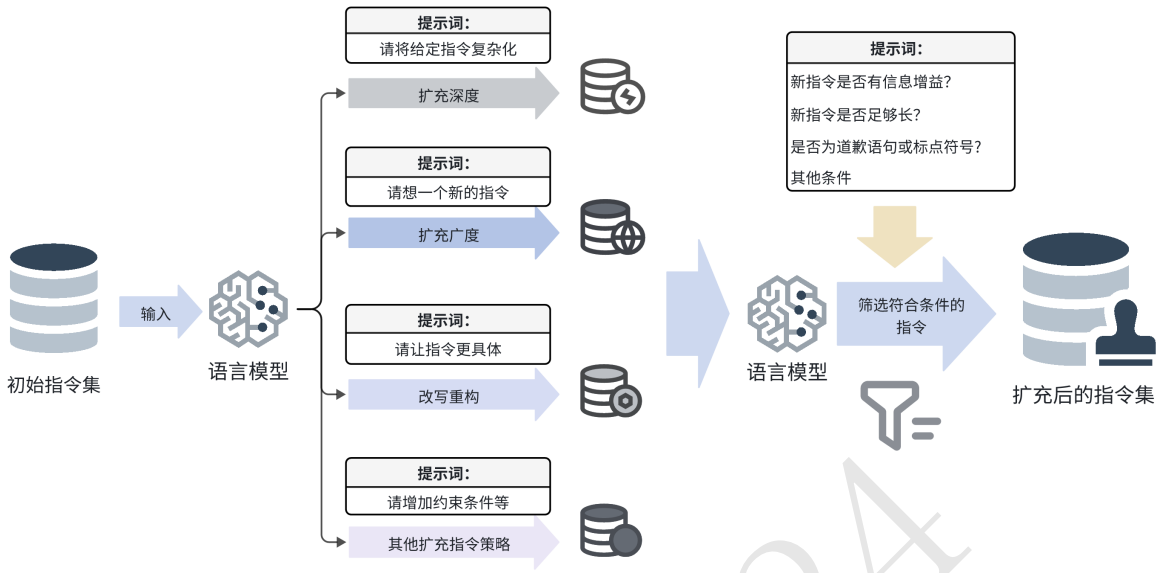


Figure 3: 通过多种策略丰富指令的数量和类型

受到人类学习启发，Gunasekar等人 (2023)提出利用大模型生成更具教育意义的训练数据能够使小模型更快地进行训练。基于此原则，作者发现许多代码数据是不具备良好的教育意义的，因为：（1）很多数据不完备，需要来自其它代码文件的信息；（2）很多代码块都只是类似变量定义等无意义的操作，不具备学习的价值；（3）一些具备复杂逻辑的代码没有很好地注释；（4）存在严重的长尾效应。为了避免这些问题，作者提出了借助大模型来生产更具教育意义的数据，具体来说：（1）使模型仅保留预训练文本中具有教育意义的内容；（2）让大模型直接生成“代码教科书”；（3）让大模型生产“课后习题”。通过以上方案，作者发现一个仅1.3B的代码模型可以在代码相关评测集上达到10倍其规模模型的性能。随后作者在此基础上，将相关思想推广到了推理类文本数据上 (Li et al., 2023b)，并取得了优异的效果。这类大量数据生成的方法容易遭遇模型生成过程中与已生成内容重复的问题，为了避免这个问题，作者首先收集了两万个主题，在模型生成时通过给定主题让模型尽量不重复。

除了通过限制主题来避免得到重复的内容，还可以通过借助大量的预训练语料来获取高质量合成数据。Yue等人 (2024)提出首先从互联网数据中筛选出相关文档，然后抽取出其中的问答对，最后使用开源大语言模型对问答对进行润色的方式来获取大量的训练数据。在筛选阶段，作者训练了一个基于Fasttext (Joulin et al., 2016)的文本分类器来从CommonCrawl (Com,)数据中分类出可能包含高质量问题-答案对的文档，在这个过程中大模型可以用于标注正样本数据，负样本数据则通过随机采样产生；然后通过去掉网页数据中的HTML标签以及广告，在这一步基础之上通过借助开源大模型，如Qwen (Bai et al., 2023)，判断文档中是否存在自然的问题-答案对，如果存在则让模型提取出；在抽取出的问题-答案对中，部分数据只含有问题和对应的答案，缺乏对相关过程的展示，则需要使用大模型来补足这部分内容。通过以上步骤可获取超过千万条相关数据，在上面训练的各种7B大小的模型取得了良好的性能。但由于生产这批数据使用的是Qwen-72B，一定程度上可以将整个过程看做是在蒸馏Qwen-72B模型。

相较于上一小节中借助标注数据标签来确定合成数据质量的方案，不利用标签的合成方法容易导致生成数据的质量不可控，因此不利用标签的方案尽管理论上可以生产近乎无限的数据量，但如何在质量和数据量间取得一个很好的平衡仍然是这类方法需要解决的问题。在代码领域，Song等人 (2024)通过让大语言模型补充代码数据注释的方式产生了超过100B词元的训练语料，由于注释即使发生错误也不影响代码执行的正确性，因此这样可以避免因为引入错误数据

而导致模型训练变坏的问题。此外，让大模型在生成数据过程中保持多样性和生成数据质量的平衡是未来这类方法另一个需要解决的难题，在Evol-Instruct方法中多样性提高的时，也容易出现合成指令不合理或无法被现有大模型很好解决的问题。

利用上述的方法一些工作被总结在表 1中，可以看出这些方法有以下两个特点，第一个是都利用了更强的模型，第二个是都主要集中在代码和数学等领域。将这些方法应用到其他领域仍然需要回答两个问题，其一是这些被蒸馏的模型是否在除代码和数学之外的领域仍然有很高的蒸馏价值，其二则是通过生成思维链进行蒸馏的方法是否对其它领域也适用。

论文	更强的大模型	下游模型	评测任务	借助监督标签
WizardLM(2023)	GPT-3.5	LLaMA-7B	Evol-Instruct、Vicuna	否
Orca(2023)	GPT-3.5、GPT-4	LLaMA-13B	BBH、AGIEval等	是
WizardCoder(2023b)	GPT-3.5	StarCoder-15B	HumanEval、MBPP等	否
phi-1(2023)	GPT-3.5	-	HumanEval、MBPP	否
MetaMath(2023)	GPT-3.5	LLaMA-2系列	GSM8K、Math	是
Mammoth(2023)	GPT-4	LLaMA-2系列	GSM8K、MATH等	否
MathCoder(2023a)	GPT-4	LLaMA-2系列	GSM8K、MATH等	是
WaveCoder(2024)	GPT-3.5、GPT-4	StarCoder-15B等	HumanEval、MBPP等	否
OpenMathInstruct(2024)	Mixtral 8x7B	LLaMA系列	GSM8K、MATH	是
Orca-Math(2024)	GPT-4	Mistral-7B	GSM8K	否
Xwin-Math(2024)	GPT-4	LLaMA-2 7B	GSM8K、MATH等	否
Mammoth2(2024)	Mixtral-8x22B等	Mistral-7B等	GSM8K、MATH等	否

Table 1: 基于蒸馏的合成数据工作

3 基于模型自我进化

除借助更强大模型合成数据的思路外，也有大量的工作着眼于利用较弱模型自身的能力，使用EM算法或强化学习等算法，通过框架和合成方法的设计帮助模型进行自我进化(Self-Evolution)。从方法和原理上看，这类工作与上一节基于蒸馏的方法没有本质区别，但这类方法由于不借助更强的大模型，因此具有更高的理论上限，同时也能避免一些法律许可相关的问题[3]，因此我们单独把这一类方法列为一节进行讨论。

在STaR方法 (Zelikman et al., 2022)中，作者利用GPT-J (Wang and Komatsuzaki, 2021)模型来生成大量的推理路径，并将这些推理路径用作训练数据微调模型自身。由于这个方法是在有监督标签的数据集上，因此判断这些推理路径是否正确时可以采用对比最终答案和标准答案的方式。除了直接使用对比答案的方法，在Self-Improve (Huang et al., 2022)中作者提出了通过采样生成多次答案，然后进行多数投票的方式选择可能的答案，并将此作为可用的训练数据用于训练模型。此外，指令回译 (Li et al., 2023a)方法也可以在不需要标准答案的情况下合成数据，具体来说，模型通过反向利用指令数据集，即将回答作为输入，将指令作为输出训练了一个指令生成模型，得到该生成模型后可以将大量的自然文本作为潜在的回答，并让该回译模型生成潜在的指令，这种方式得到的指令和回答数据质量可能参差不齐，因此论文提到也需要借助指令数据集训练一个正向模型，该正向模型可以用来评估生成的指令和回答的质量是否可用，并从中筛选出高质量的部分。筛选出来的数据可以继续用于迭代训练回译模型和正向模型，经过多轮迭代的模型可以取得良好的指令跟随能力。

受Alpha-Zero (Silver et al., 2017)的启发，Alpha-Math (Chen et al., 2024)构建了一个由蒙特卡洛树搜索（MCTS）驱动的推理算法，在采样推理路径时通过MCTS充分激发模型的推理潜能，通过策略和价值网络的协同使用进行节点的增长。其中策略网络为原始的模型，而价值网络通过在原始模型上添加一个带有tanh激活函数得到。在搜索过程中，论文将MCTS推理过程简化为步级别的束搜索（Step-level Beam Search）操作，迭代地从最优的步骤中采样多组方案，使用价值网络对不同的采样方案进行评估并得到更新后的最优步骤。通过使用价值模型对不同的推理路径进行评估，模型的推理表现得到了显著的提升。

除了以上这些显式通过大模型生成数据的自我进化方法，模型可以通过将上一代模型用于下一代模型的数据生产实现不同代模型间的逐步提升。例如在LLaMA-3 [4]中Meta的研究员利

[3]OpenAI的法律条款中规定不能使用ChatGPT生成的数据训练基础模型

[4]https://ai.meta.com/blog/meta-llama-3/

用LLaMA-2模型 (Touvron et al., 2023b)来识别高质量语料，实验发现LLaMA-2可以可靠地为文本质量分类器生产训练数据。通过上一代模型为下一代模型提供助力的方式应该具备广阔的前景。

目前来说，基于模型自我迭代增强的数据合成方法较少，未来有很大的研究空间。并且基于已有的方法来看，其效果随着迭代次数的增加并没有特别好的规模（Scaling）效应，未来如何提高迭代方案的性能上限是一个值得研究的问题。更多相关的探讨可以参阅论文 (Burns et al., 2023)及引用了该论文的后续论文。

4 基于工具的合成数据

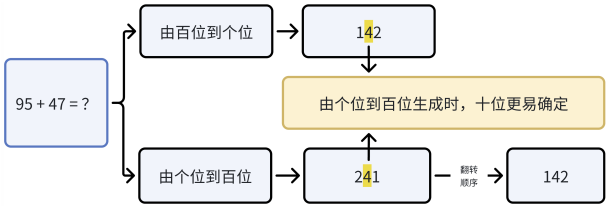
有许多工作在合成数据的过程中使用了外部工具，如代码解释器、计算器、推理引擎和抽象语法树等。外部工具可以为合成数据提供额外的信息。工具所提供的信息可能是不充分的或冗余的，但一般是客观、真实的。这些信息往往并不足以直接作为合适的训练数据，但能在合成数据的过程中为模型提供信息增益，从而合成更有效，更准确的数据。

在使用语言模型合成代码数据的工作中，代码解释器是一个天然的辅助工具。代码解释器的执行结果能带来额外的信息增益，这部分信息可以用于筛选合成数据的正确性，或者作为推理过程的中间结果。MAmmoTH (Yue et al., 2023)调用模型对已有数据集中的问题生成Python代码，并调用代码解释器返回代码的执行结果，通过比较代码执行结果与标准答案的异同，即可筛选出正确的代码数据。而MathCoder (Wang et al., 2023a)和OpenMathInstruct (Toshniwal et al., 2024)等工作使模型在推理时生成含有内嵌代码段的题解，在得到代码段的执行结果后模型将继续自回归的推理过程。Cai等人 (2024a)使语言模型充当工具制作者，为特定任务编写Python函数作为工具，同时为函数编写相应的测例。在工具通过所有测例后，另一个LLM将作为工具调用者解决实际问题。还有一些工作利用代码解释器搭建了交互式的框架 (Shypula et al., 2024; Yang et al., 2023a; Shinn et al., 2023)，使用强化学习等算法，在模型生成代码数据后，利用代码解释器提供的信息对合成数据进行分类或更正，进而迭代优化模型。Ni等人(2024)在代码修复任务中通过使用注释的方式将调试信息加入代码中，使得模型可以基于调试信息修正代码中的Bug。

尽管大模型在文字生成方面取得了良好的效果，但是它们却不能很好地计算基础的四则运算。为了让大模型能够在数值计算的时候减少错误，过去的方法尝试了通过规则生成大量的四则运算等式以及基础数学等式作为训练数据 (Yuan et al., 2023; Yang et al., 2023b)，这类数据的一些示例如图 4(a)所示。但这类方法主要依靠让模型背下来所有的计算，不太具备泛化计算的能力。人类在计算这类复杂运算时一般会通过草稿纸的形式，而非靠记下所有的四则运算，受此启发，Lee等人 (2023)提出通过Scratchpad的形式计算复杂运算，一个简单的示意如图 4(b)所示，通过加入对任务的理解，可以极大降低模型学习的难度，结合领域知识对于用好工具合成数据很重要。

运算类型	示例
四则运算	$7 + 1825.4 = 1832.4$
乘方运算	$7^3 = 343$
对数运算	$\ln 1523 = 7.328437$
三角运算	$\cot(313\pi/306) = 13.890726$

(a) 遍历各类运算



(b) 改变计算方式减少学习难度

Figure 4: 遍历四则运算与Scratchpad示例

借助数学工具，不但可以合成简单的四则运算，甚至可以合成比较复杂的数学推理数据。Lean (de Moura and Ullrich, 2021)是一个功能强大的交互式定理证明器和编程语言，主要用于形式化数学和计算机科学的证明，通过Lean可以实现自动验证数学证明。因此可以通过将大量数学证明问题转换为Lean的形式，并让语言模型使用Lean语言生成大量证明，通过Lean的定理验证器从中选择推理正确的数据，我们就可以生产出大量的严格正确的数学证明。通过这种方式大模型已经能够完成一些国际数学奥林匹克大赛的题目 (Ying et al., 2024; Xin et al., 2024)。除此之外，AlphaGeometry还通过学习几何推理引擎合成大量的高难度集合证明题，在

国际数学奥林匹克竞赛（IMO）的几何证明题目上取得了惊艳的结果 (Trinh et al., 2024b)。几何推理引擎能够在给定前提的基础上，利用已知的欧式几何规则，反复生成新的结论直到所有结论穷尽，即生成给定前提的推理闭包。然而，仅仅依靠几何推理引擎并不能够解决大部分证明问题，启发式的辅助点添加策略是必不可少的，而AlphaGeometry使用语言模型来替代了原先启发式添加辅助点的策略。由以上的工作可以看出，为了更好地利用各类工具，合成数据需要各领域相关的专业知识，大模型未来的数据合成需要和不同行业的专家进行合作。

使用工具合成数据最大的优势是数据的正确性比较容易有保证，例如四则运算和定理构造等都是基于符号化推理得到的结果，因此其结果出错概率很低。但如何比较好地利用工具，如何挑选用于合成数据的工具，这些需要相关的领域知识。

5 讨论

上述的各种方法只是合成数据方法的一部分，从以上的方法描述中不难看出，合成数据中两个比较关键的问题是：数据质量和数据多样性。对于提高合成数据的质量的方法，一般有以下四种：（1）对照生成答案与标准答案，只采用两者一致的合成数据；（2）采用多次投票的方案，采用获取票数最多的答案；（3）采取更强大模型，如GPT-4；（4）采用外部工具。为了提高合成数据的多样性，可以采用的方案有（1）调整模型生成时的温度等系数，使得生成更多样；（2）通过从预训练语料中挖掘；（3）通过限制生成条件的方式。

在实际应用中，为了获得高质量的合成数据，在成本可控的情况下，应尽可能使用性能更强的模型，或者通过迭代使用自合成的数据对较弱的模型进行优化，使其在所需领域的性能逐渐接近强大模型；对于输入文本x，应当选用质量较高的数据，并在使用前进行筛选，以排除那些质量低下、无意义的文本；对于先验知识t，可以利用数据集的标注来监督合成数据的输出y；此外，还应尝试利用各类工具，例如代码解释器，以减少模型生成发生错误的可能性，或者在提示词（Prompt）中为模型提供额外的先验知识来增加模型输出好的回答的概率。为了使合成数据具有多样性，应尽可能尝试从广泛的预训练语料中获取相关的种子数据，根据这些种子数据进行数据扩增；同时调整大模型推理时的随机性参数，或使用不同的大模型进行数据合成。

但当前合成数据仍然面临着以下的限制，（1）大量依赖更强的语言模型，例如GPT-4，通过蒸馏更强的闭源语言模型可以快速提升模型的能力，但是对于我们进一步实现通用人工智能（Artificial General Intelligence，简称AGI）帮助有限，同时构造大量语料的花费较高，例如使用GPT-4o构造1B的训练语料需要花费1.5万美金左右 [5]；（2）目前的合成数据方案较多只是针对某个特定能力，合成的数据量级大多在1B词元以下，与动辄上T量级的预训练数据需求存在几个数据量级的差距。大部分的合成数据方法不具备规模化的前景，同时在单一的领域上进行训练也可能导致模型在其它领域的能力下降 (Gudibande et al., 2023)；（3）方法主要集中在数学、代码类数据上，这些技术方法能否泛化到真实场景中的数理推理和代码生成是个尚待研究的问题，同时对于那些答案不唯一或者比较复杂的场景，可能需要发掘更多方法来进行合成数据质量控制；（4）在长序列语境下的合成数据较少，近期大模型的长语境处理能力给大模型带来了很大的想象空间，但目前高质量长序列训练数据非常缺乏 (Lv et al., 2024)，合成高质量长语境数据将有助于提高模型的对应能力。

6 总结

在本文中，我们介绍了三种常见的大模型数据合成方法。基于强大模型蒸馏的方法是目前采用最多的数据合成方法，但这类方法依赖于强大模型，尽管取得了很好的实际效用，但从进一步突破模型能力上限的角度作用有限。基于模型自我进化的方法有较高的研究价值，但目前相关工作较少。基于工具的方法可以保证合成数据的质量，但是高质量、多样的工具数据依赖于领域知识的注入，未来大模型领域需要和各行各业的科研工作者一道合作补充这类数据。

致谢

本文受上海人工智能实验室资助。

[5] https://openai.com/api/pricing/

参考文献

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, et al. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, et al. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *CoRR*, abs/2312.09390.

Tianle Cai, Xuezhi Wang, Tengyu Ma, Xinyun Chen, and Denny Zhou. 2024a. Large language models as tool makers.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, et al. 2024b. Internlm2 technical report. *CoRR*, abs/2403.17297.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. 2024. Alphamath almost zero: process supervision without process.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Common crawl web archive. http://commoncrawl.org. Accessed: 2024-05-30.

Leonardo de Moura and Sebastian Ullrich. 2021. The lean 4 theorem prover and programming language. Lecture Notes in Computer Science, pages 625–635. Springer.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, et al. 2023. Is gpt-3 a good data annotator?

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, et al. 2021. A survey of data augmentation approaches for NLP. Findings of ACL, pages 968–988. Association for Computational Linguistics.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *Int. J. Comput. Vis.*, 129(6):1789–1819.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, et al. 2023. The false promise of imitating proprietary llms. *CoRR*, abs/2305.15717.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, , et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, et al. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pages 1–9.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, et al. 2020. Tinybert: Distilling BERT for natural language understanding. Findings of ACL, pages 4163–4174. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, et al. 2016. Fasttext.zip: Compressing text classification models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, et al. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Nayoung Lee, Kartik Sreenivasan, Jason D. Lee, Kangwook Lee, and Dimitris Papailiopoulos. 2023. Teaching arithmetic to small transformers. *CoRR*, abs/2307.03381.

Yujia Li, David H. Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, et al. 2022. Competition-level code generation with alphacode. *CoRR*, abs/2203.07814.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, et al. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, et al. 2023b. Textbooks are all you need II: phi-1.5 technical report. *CoRR*, abs/2309.05463.

Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, et al. 2024. Common 7b language models already possess strong math capabilities. *arXiv preprint arXiv:2403.04706*.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, et al. 2023. Openorca: An open dataset of gpt augmented flan reasoning traces. https://https://huggingface.co/Open-Orca/OpenOrca.

Yih-Kai Lin, Chu-Fu Wang, Ching-Yu Chang, and Hao-Lun Sun. 2021. An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network. *Multim. Tools Appl.*, 80(3):4037–4051.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, et al. 2024. Best practices and lessons learned on synthetic data for language models.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. Proceedings of Machine Learning Research, pages 22631–22648. PMLR.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, et al. 2023a. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *arXiv preprint arXiv:2308.09583*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, et al. 2023b. Wizardcoder: Empowering code large language models with evol-instruct. In *The Twelfth International Conference on Learning Representations*.

Kai Lv, Xiaoran Liu, Qipeng Guo, Hang Yan, Conghui He, et al. 2024. Longwanjuan: Towards systematic measurement for long text quality. *CoRR*, abs/2402.13583.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, et al. 2023. Scaling data-constrained language models.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, et al. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

Ansong Ni, Miltiadis Allamanis, Arman Cohan, Yinlin Deng, Kensen Shi, et al. 2024. Next: Teaching large language models to reason about code execution.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, et al. 2023. Reflexion: Language agents with verbal reinforcement learning.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2022. Distilling reasoning capabilities into smaller language models. *arXiv preprint arXiv:2212.00193*.

Alexander Shypula, Aman Madaan, Yimeng Zeng, Uri Alon, Jacob Gardner, et al. 2024. Learning performance-improving code edits.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *CoRR*, abs/1712.01815.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, et al. 2024. Beyond human data: Scaling self-training for problem-solving with language models.

Demin Song, Honglin Guo, Yunhua Zhou, Shuhao Xing, Yudong Wang, et al. 2024. Code needs comments: Enhancing code llms with comment augmentation.

Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, et al. 2024. Openmathinstruct-1: A 1.8 million math instruction tuning dataset. *arXiv preprint arXiv:2402.10176*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, et al. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He, and Thang Luong. 2024a. Solving olympiad geometry without human demonstrations. *Nat.*, 625(7995):476–482.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024b. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Eric Wallace, Mitchell Stern, and Dawn Song. 2020. Imitation attacks and defenses for black-box machine translation systems. pages 5531–5546. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023a. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. *arXiv preprint arXiv:2310.03731*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, et al. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, et al. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, et al. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, et al. 2024. A survey on knowledge distillation of large language models. *CoRR*, abs/2402.13116.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023a. Intercode: Standardizing and benchmarking interactive coding with execution feedback.

Zhen Yang, Ming Ding, Qingsong Lv, Zhihuan Jiang, Zehai He, et al. 2023b. GPT can solve mathematical problems without a calculator. *CoRR*, abs/2309.03241.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, et al. 2022. Zerogen: Efficient zero-shot learning via dataset generation.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, et al. 2024. Internlm-math: Open math large language models toward verifiable reasoning. *CoRR*, abs/2402.06332.

Longhui Yu, Weisen Jiang, Han Shi, YU Jincheng, Zhengying Liu, et al. 2023. Metamath: Bootstrap your own mathematical questions for large language models. In *The Twelfth International Conference on Learning Representations*.

Zhaojian Yu, Xin Zhang, Ning Shang, Yangyu Huang, Can Xu, et al. 2024. Wavecoder: Widespread and versatile enhanced instruction tuning with refined data generation.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *CoRR*, abs/2304.02015.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, et al. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024. Mammoth2: Scaling instructions from the web.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, et al. 2024. Chemllm: A chemical large language model. *CoRR*, abs/2402.06852.

大语言模型时代的信息检索综述

庞亮，邓竞成，顾佳，沈华伟，程学旗
中国科学院计算技术研究所，智能算法安全重点实验室，北京，100190
中国科学院大学，北京，100190
pangliang@ict.ac.cn, dengjingcheng23s@ict.ac.cn

摘要

以大语言模型为代表的生成式人工智能迅猛发展，标志着人工智能从判别时代向生成时代的转变。这一进步极大地推动了信息检索技术的发展，本文对大语言模型对信息检索领域的影响进行了深入的综述。从性能改进到模式颠覆，逐步展开论述大语言模型对信息检索领域的影响。针对传统信息检索流程，大语言模型凭借强大的语义理解和建模能力，显著增强索引、检索和排序等信息检索模块的性能。同时，文章也探讨了大语言模型可能取代传统信息检索的趋势，并催生了新的信息获取方式，或将是新一次信息时代的寒武纪。此外，大语言模型对内容生态的深远影响也值得关注。

关键词： 信息检索；大语言模型；索引；检索；排序

A Review of Information Retrieval in the Era of Large Language Models

PANG Liang, DENG Jingcheng, GU Jia, SHEN Huawei, CHENG Xueqi
CAS Key Laboratory of AI Safety, Institute of Computing Technology,
Chinese Academy of Sciences / Beijing, 100190
University of Chinese Academy of Sciences / Beijing, 100190
pangliang@ict.ac.cn, dengjingcheng23s@ict.ac.cn

Abstract

The rapid development of generative artificial intelligence, exemplified by large language models, signifies a shift in artificial intelligence from the era of discrimination to generation. This advancement has greatly propelled the development of information retrieval technology. This paper provides an in-depth review of the impact of large language models on information retrieval. It discusses the influence of large language models on information retrieval, from performance improvements to paradigm shifts. In the context of traditional information retrieval processes, large language models significantly enhance the performance of various information retrieval modules such as indexing, retrieval, and ranking, owing to their powerful semantic understanding and modeling capabilities. Additionally, the paper explores the potential trend of large language models replacing traditional information retrieval methods and giving rise to new ways of obtaining information, possibly marking a new Cambrian explosion in the information age. The profound impact of large language models on the content ecosystem is also worthy of attention.

Keywords: Information Retrieval , Large Language Models , Indexing , Retrieval , Ranking

第二十三届中国计算语言学大会论文集，第98页-第119页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

信息检索是指从海量的文本、数据或多媒体中，根据用户需求找出相关信息，并呈现给用户的过程。在当今信息爆炸的时代，信息检索的价值愈发显著，是人们获取信息的重要途径(Robertson et al., 2009)。传统的信息检索通常包含索引、检索、排序等模块。索引模块（Indexing）为海量数据构建快速的访存机制，利用更少的空间，存储更多的数据，构建更快的访问(Negi et al., 2012)；检索模块（Retrieval）为用户需求筛选候选文档，保证相关信息的召回率(Xu et al., 2023d)；排序模块（Ranking）为用户需求提供精确的排序策略，确保更相关的文档排在更靠前的位置(Pang et al., 2017)，最后以列表的形式展现给用户，用户通过从上到下浏览文档列表，满足用户的信息需求。信息检索规范化的流程定义，让其快速规模化，诞生了类似谷歌、百度、必应等企业，但也面临着非常棘手的挑战，包括复杂用户需求的语义理解、全系统流程的相关性优化、单调不直接的列表信息形式等，有待解决。

近年来大语言模型（LLMs）的迅速发展，以ChatGPT为代表的模型(Brown et al., 2020)，基于Transformer架构(Vaswani et al., 2017)，在开放域问答(Xu et al., 2024a)、数学解题(Zhao et al., 2024)、对话系统(Zhou et al., 2023)、机器翻译(Piergentili et al., 2024)、文本摘要(Pakull et al., 2024)等领域表现出色。大语言模型凭借其强大的语义理解能力、顺畅的多轮交互能力、丰富的生成展现形式，已然成为各领域的研究热点，而利用大语言模型的优势来优化信息检索的各个模块(Zhu et al., 2023)，例如索引模块、检索模块、排序模块，成为最直接的应用目标。基于大语言模型的细粒度嵌入向量可以提升索引阶段的信息区分度，深度用户查询理解和重构可以提升检索阶段的信息召回率，内部充分交互后的生成式文档重排可以提升排序阶段的精确度。

除了嵌入在传统信息检索系统流程中提升效果，近年来还涌现了大量以大语言模型主导的信息检索新范式(Ma et al., 2024)，颠覆传统的信息检索流程。新型搜索引擎大致可以分为两类，一类是大语言模型作为代理进行网页浏览检索并整理检索结果的代理式检索，另一类是利用大语言模型的对话生成能力，将大语言模型作为搜索引擎的核心部分的交互式检索。与传统的关键字搜索引擎不同，基于大语言模型的新型搜索引擎更注重理解用户的意图和上下文，并提供与之相关的搜索结果。这类搜索引擎在用户与系统交互时提供有针对性的信息检索，能够为用户提供更智能、更个性化、更高效的信息检索体验，从而改善用户与系统之间的交互和沟通(Spatharioti et al., 2023)。广告为搜索引擎带来巨大的经济效益，而随着这一类新型搜索引擎的兴起，在传统搜索引擎中嵌入的广告应如何融入新型搜索引擎则是一个具有应用前景的话题(Feizi et al., 2023)。目前的主要做法是将广告作为文本嵌入大语言模型生成的内容。

大语言模型推动了信息检索领域的发展，也对信息检索的内容生态带来了深远影响，其中包括偏见问题、不公平问题和创作消费问题等。研究者们发现基于大语言模型的信息检索更倾向于检索人工智能生成的内容(Dai et al., 2023a; Xu et al., 2023c)，大语言模型生成的错误内容、不公平内容以及过多的人工智能创作对信息检索生态的影响(Dai et al., 2024)。

本文的结构如图 1所示，首先我们在第2节概述信息检索的主要模块以及大语言模型的发展总结。其次，我们将在第3节介绍大语言模型在传统信息检索中的应用，在第4节总结以大语言模型为核心的信息检索新范式，在第5节详述大语言模型生成的内容对信息检索生态的影响。最后，在第6节中讨论信息检索结合大语言模型在未来的发展方向，并在第7节对本文进行总结。

2 背景

2.1 信息检索

信息检索系统旨在从大规模的信息资源中根据用户的需求和查询提供相关的信息。它的主要作用是帮助用户快速准确地获取所需的信息，从而满足他们的信息需求。在本文中我们主要关注文本模态的信息检索系统，并将其分为三大块，分别为文本索引、文本检索和语义排序。

2.1.1 索引模块

文本索引是一种用于组织和加速文本数据检索的数据结构。它的主要功能是将文本数据中的关键词和它们出现的位置建立关联，以便在搜索过程中能够快速定位和检索相关文档。根据信息检索的两大范式——稀疏检索和稠密检索，我们可以将文本索引划分为两个主要部分。针对稀疏检索的文本索引被认定为是一种用于组织和加速文本数据检索的数据结构(Robertson et al., 1994)。它通常包括分词、去除停用词、正规化和词干提取等步骤，最终构建类似于倒

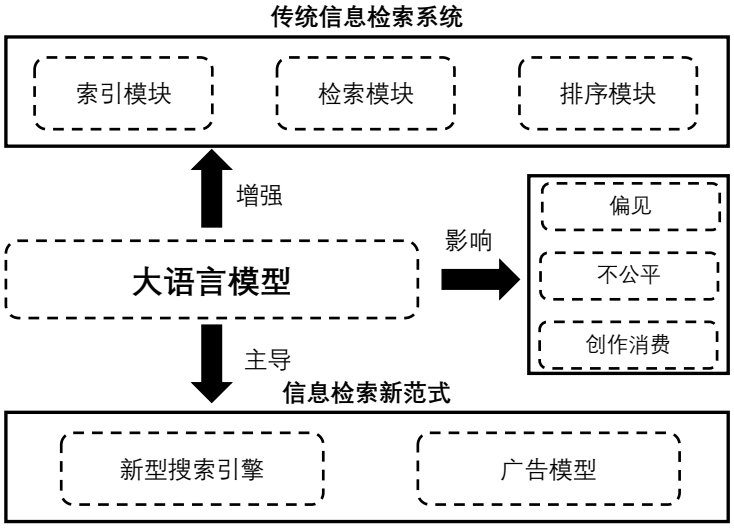


图 1: 大语言模型对现代信息检索系统的改变。主要分为增强传统信息检索系统、主导信息检索新范式和对检索生态的影响。

排索引的数据结构。在稠密检索中，人们一般通过文本嵌入模型来构建文本索引(Izacard and Grave, 2021)。此时文本索引被视为低维空间中的稠密向量。

2.1.2 检索模块

文本检索是指通过对文本内容的处理和分析，从大规模的文本集合中匹配和提取与用户查询相关的文档或文本片段。文本处理主要涉及查询处理和匹配这两种方法。前者对查询进行分析、解析和规范化。这包括去除停用词（如"和"、"的"等常见词汇），处理查询的语法和语义，以及提取查询中的关键词和条件。而后者通常是指使用诸如倒排索引（inverse index）等技术来快速定位包含查询关键词的文档。匹配完成后，搜索结果会按照相关性进行排序，以便将最相关的文档排在前面。此外随着深度学习技术的兴起，最近的研究(Izacard and Grave, 2021)主要围绕将查询和文档映射到各向同性的向量空间中，然后通过内积计算来计算它们的相关性分数。这种范式转变可以更有效地捕获查询和文档的语义相似性。总的来说这一模块充当信息检索系统中的首轮文档检索器，它从大规模文档集合中召回广泛相关文档作为候选。因此它定位相关文档的效率对于信息检索系统尤为重要。

2.1.3 排序模块

语义排序在一些研究中又被称之为重排序，它是信息检索中的另一个关键模块。与文本检索阶段强调效率和性能的平衡不同，它主要对召回的文档进行细粒度重新排序。为了提高排序质量，最近的研究(Xiao et al., 2023)提出了比传统内积匹配更复杂的方法，从而为模型提供更加丰富的语义匹配信号。

2.2 大语言模型

与传统语言模型不同，生成式大语言模型在处理复杂任务方面具有出色表现，这被称之为涌现能力(Wei et al., 2022)。最著名的大语言模型是来自OpenAI的ChatGPT(Ouyang et al., 2022)，它是大模型时代的一个里程碑。现有的大语言模型根据架构可以分为两类，第一类是编码器-解码器模型(Raffel et al., 2020; Zeng et al., 2023)，第二类是解码器模型(Ouyang et al., 2022)。编码器-解码器模型将文本输入编码器中转为向量，然后将向量输入解码器以获得输出。具有代表性的模型有T5(Raffel et al., 2020)和GLM(Zeng et al., 2023)。而更多的还是以GPT系列为代表的解码器模型，它们依赖于Transformer的解码器架构，从左至右自回归式地生成单词或字。具有代表性的模型有InstructGPT(Ouyang et al., 2022)、LLaMA系列模型(Touvron et al., 2023a; Touvron et al., 2023b)等。

3 信息检索系统中的大语言模型应用

本节主要描述在传统信息检索系统中大语言模型能够带来的改进。

3.1 大语言模型增强索引模块

针对稀疏检索的文本索引已经相当成熟，大语言模型对其带来的改进并不明显。而针对稠密检索的文本索引，大语言模型的出现为其性能带来了显著提升，因此我们重点关注这部分。在本文中，我们将大语言模型对文本索引的改进分为两个主要部分：生成训练数据和构造索引向量。

3.1.1 利用大语言模型生成训练数据

从模型训练角度看，大量高质量的训练数据至关重要，这可以使文本嵌入模型获得全面的语义知识和准确的语义空间。不幸的是，收集人工注释的相关性标签既耗时又昂贵。它限制了文本嵌入模型的知识边界及其跨不同应用领域进行泛化的能力。考虑到大语言模型强大的文本理解和生成能力，许多研究者利用基于大语言模型驱动的流程构建相关的查询-文档对以扩大文本嵌入模型的训练数据。

考虑到在现实世界中文档远比查询更加丰富，因此InPars(Bonifacio et al., 2022)提出使用类似于GPT-3的大语言模型来针对未被标注的文档生成相应伪查询。具体而言，它利用GPT-3强大的上下文学习能力（In-Context Learning）把一些查询-文档对作为示例输入模型。随后GPT-3针对给定文档生成可能的伪查询。因此利用这种方法可以轻松合成大量的训练数据。然而由于InPars方法较为简单，容易生成一些带有噪声或不相关信息的查询-文档对，因此InPars-v2(Jeronymo et al., 2023)在其之上使用更强大的大语言模型，并加入了现有的强大的重排序器对生成结果进行筛选，从而确保了最终数据的质量。除了使用重排序器过滤生成的样本对，也有研究专注于优化生成伪查询阶段的性能。AugTriever(Meng et al., 2022)提出两种构造查询-文档对的新策略，分别为查询提取和转移查询生成。查询提取针对文档中的某一跨度生成精确查询，而转移查询生成为其他自然语言处理任务（如文本摘要）也生成伪查询。这两种策略极大增强了生成的样本对的性能。UDAPDR(Saad-Falcon et al., 2023)采用了两阶段的伪查询生成方法，它首先采用强大的大语言模型生成少量高质量的伪查询，然后把这些高质量的查询-文档对作为普通的大语言模型的输入示例以生成大量伪查询。这种方法平衡了计算成本和生成质量。除了关注生成质量外，也有部分工作关注生成的任务类型和生成的数量。(Ma et al., 2023a)等人利用对比学习和瓶颈查询生成将大语言模型当中的知识有效地传递给文本嵌入模型，此外他们还结合了课程学习策略来减少对大语言模型推理的依赖。最终这些生成的数据被用于预训练文本嵌入模型，并且取得了出色的性能。Gecko(Lee et al., 2024b)扩大了合成的数据量并生成了更广泛的任务类型。然后它为每个查询检索一组候选段落，并使用相同的大语言模型重新标记正例段落和硬负例段落，进一步优化数据质量。这步操作保证它使用的大部分数据来源于真实数据而非合成数据，从而减轻合成数据带来的偏差。Promptagator(Dai et al., 2023b)专注于对特定检索任务合成数据，然后使用此数据训练文本嵌入模型，在11个不同的检索任务上平均nDCG上升1.2%。(Wang et al., 2024a)等人利用专有的大语言模型为近100种语言的数十万个文本嵌入任务生成各种合成数据，然后利用对比学习损失微调Mistral-7B模型，在BEIR(Thakur et al., 2021)和MTEB(Muennighoff et al., 2023)基准上取得了先进的性能。

3.1.2 利用大语言模型构造索引向量

传统的文本嵌入模型大部分是编码器结构的。虽然有少部分研究者尝试使用编码器-解码器结构或者仅解码器结构的模型产生嵌入向量，但是效果不佳。大语言模型的出现为使用仅解码器结构的模型产生高质量嵌入向量提供可能性。(Ma et al., 2023b)等人为每个文档拼接上一个终止符‘< \s >’，接着将它们输入LLaMA-2模型当中，并使用终止符‘< \s >’作为对应文档的嵌入表示。然后它采用InfoNCE损失函数对模型进行端到端优化，最终构造出RepLLaMA模型。实验证明大语言模型有成为稠密检索器的能力。(Li et al., 2023a)等人提出两个可以将大语言模型调整为稠密检索当中嵌入编码器的任务，分别叫做基于嵌入的自动编码（EBAE）和基于嵌入的自回归（EBAR）。它们确保大语言模型产生的文本嵌入可以重建输入语句的表示并预测下一语句的表示，大大提高了基于大语言模型的索引向量在各种稠密检索基准上的性能。考虑到自回归语言模型当中因果注意力机制的限制导致输入当中某一个标记表示向量并不能包含后续标记表示向量的信息，(Springer et al., 2024)等人提出一种叫做回声嵌入的方式，将文档重

复输入两次，并提取第二次出现的标记向量作为嵌入向量，在不修改模型结构的情况下缓解这个问题。而(BehnamGhader et al., 2024)等人提出一个简单的三阶段策略改变模型参数从而缓解这个问题，即1）启用双向注意力机制，2) 屏蔽下一个标记预测任务，3) 进行无监督对比学习。(Lee et al., 2024a)等人与其类似，他们在对比学习期间删除了原始大语言模型中的因果注意力机制，并加入一个潜在的注意力层来获取池化嵌入向量。他们提出的模型目前在MTEB排行榜上展示出最好的表现。

3.2 大语言模型增强检索模块

构造文本索引之后，信息检索到了文本检索阶段。此时给出查询可以召回与之最相似的一批文档。这一阶段主要注重于检索效率和高召回率，它们维持搜索引擎性能和最终结果生成至关重要。最近的工作可以被分为两类。第一类为查询重写，它们主要利用具有出色理解和生成能力的大语言模型重新表述原始查询以解决查询表述歧义、不清晰、不完整以及查询和文档之间词汇不匹配等问题。第二类工作被称为生成式检索，它们将大语言模型当作索引库，针对查询直接生成索引。

3.2.1 查询改写

最经典的工作为Query2doc(Wang et al., 2023a)，它利用大语言模型根据原始查询生成相关段落，然后将原始查询和相关段落进行拼接，一同去召回相关文档。由于生成的段落包含额外的详细信息，这可以缓解词汇不匹配问题，同时对不明确的、简短的查询尤为有效。(Jagerman et al., 2023)等人为了获得高质量的伪相关段落，他们分析了在零样本、少样本和思维链设置下大语言模型生成伪相关段落的质量，最后发现在思维链设置下用大语言模型进行查询改写的效果最好。除此之外，还有大量的工作是有不同的方法扩展原始查询中的知识。(Feng et al., 2023)等人提出了InteR，它允许现代检索模型和大语言模型之间的协同作用来促进信息细化。(Shen et al., 2023)等人建议通过使用查询和查询域内候选的组合来提示大语言模型，从而用其潜在答案来增强查询。(Lei et al., 2024)等人考虑到大语言模型的内在知识有限从而导致出现幻觉和信息过时等问题。因此他们提出引入语料库引导的查询扩展(CSQE) 来促进语料库中嵌入知识的整合。相关实验表明，CSQE无需任何训练即可表现出强大的性能，尤其是对于大语言模型缺乏知识的查询。针对法律领域的查询，(Tang et al., 2023b)等人使用大语言模型将复杂的和法律有关的查询简化为更易于搜索的法律事实和问题，然后采用基于提示的编码方案来进行有效的语言模型编码。然而，使用伪相关段落可能会给原始查询带来噪声并出现概念漂移，(Anand et al., 2023)等人利用重写的查询和文档对文本嵌入模型进行微调，让模型性能得到了极大的提升。

3.2.2 生成式检索

生成式检索将整个检索过程建模为生成式任务，具体可以分为四步，包括：（1）查询制定，即确定生成模型的输入；（2）文档标识符制定，即用短的标识符表示文档；（3）模型训练；（4）模型推理。在生成式检索中查询制定步骤一般比较简单，大部分工作均使用原始查询作为生成模型的输入。而模型推理步骤又和模型训练相关联。因此我们主要总结最近有关文档标识符制定和模型训练的研究。

理论上讲，生成式检索应该直接生成和查询对应的文档，然而由于生成模型上下文长度的限制，现有方法通常依赖于使用标识符来表示文档。最常见的是为语料库中的每一个文档分配一个数字标识符(Tay et al., 2022; Li et al., 2024a; Nadeem et al., 2022; Wang et al., 2022)。然而由于数字标识符缺乏语义含义，使其泛化性差。并且每当语料库更新时，新增文档标识符的构建以及记忆难度均会增加。使用类似于文档标题作为标识符是另一种解决方案(Chen et al., 2022; Cao et al., 2021; Lee et al., 2022a; Li et al., 2023b)，它将文档语义信息也纳入了标识符当中，在语义上可以和文档建立一对一的对应关系。然而这类方法因为很难设定段落标识符从而在更细粒度的段落级检索中表现不佳。并且在面向网页检索的任务当中也难以为网页构建高质量的标题标识符。除了这些之外，还有许多其他的建模方案，比如将文档中与查询语义相关的N元语法视为潜在的标识符(Bevilacqua et al., 2022; Chen et al., 2023a; Wang et al., 2023c),或者构建一个所谓的密码本，在根据文档内容学习最佳标识符(Sun et al., 2023a; Yang et al., 2023)。

生成式检索模型的训练一般分为两个阶段，分别为查询到标识符的训练和文档到标识符的

训练。对于缺乏语义表示的数字标识符和密码本标识符而言，后者的训练过程尤为重要。以上两个训练阶段均为生成式任务，然而也有工作(Li et al., 2024b)表明判别式训练的重要性。许多工作(Tang et al., 2024b; Zeng et al., 2024)证明了在生成式检索模型上引入判别式训练的有效性。

得益于大语言模型出色的理解能力和生成能力，还有不少研究脱离了传统生成式检索范式。比如(Ziems et al., 2023a)等人将大语言模型视为内置搜索引擎。他们直接构建用于文档检索的URL指令，并发现当提供一些例子给大语言模型时，大语言模型可以生成Web URL，其中近90%的相应文档包含对开放域问题的正确答案。(Yu et al., 2023)等人用大语言模型替换传统的文档检索器，直接根据给定查询生成上下文文档，在TriviaQA(Joshi et al., 2017)和WebQ(Berant et al., 2013)两个数据集上上分别获得了71.6和54.4的精确匹配分数。

3.3 大语言模型增强排序模块

语义排序作为信息检索系统中的最后一个阶段，旨在根据语义相关性对召回的文档重新排序。我们将在此阶段中使用大语言模型的工作分为两类，分别为利用大语言模型监督微调重排序器和利用大语言模型进行生成式排序。

3.3.1 监督微调重排序器

与利用大语言模型生成数据来监督微调检索器类似，监督微调重排序器也是一个很常见的做法。比如(Ferraretto et al., 2023)等人使用诸如GPT-3.5之类的大语言模型，通过解释来增强检索数据集，并训练一个序列到序列的排序模型，以输出给定查询-文档对的相关性标签和解释。实验证明在这种方法生成几千个样本上进行微调，性能与没有解释的使用3倍甚至更多样本进行微调的模型相当。(Boytsov et al., 2023)等人借鉴Inpars(Bonifacio et al., 2022)的思路，构造出一个轻量级的被称为InPars-Light的方法。它提示大语言模型生成合成查询-文档对，并针对比之前小7到100倍参数的排序模型进行无监督训练，最终在五个英文检索数据集上取得显著性改进。与注重于生成合成查询的工作不同，也有一部分工作强调生成合成文档的重要性。(Boytsov et al., 2023)等人构造出一个名为ChatGPT-RetrievalQA的数据集，它基于大语言模型响应用户查询生成合成文档来构建。他们利用此数据集和人工生成的数据微调了一系列重排序器。在多个数据集上的结果证明使用此数据集训练的重排序器比用真实数据训练的重排序器在统计上更有效。(Askari et al., 2023)等人提出DocGen和DocGen-RL两种方法，前者从查询生成合成文档，后者利用强化学习进一步优化DocGen，从而提高生成的合成文档与其对应查询之间的相关性。

3.3.2 生成式排序

生成式排序将排序任务视为生成式任务进行建模。在大语言模型出现之前，这个领域已有许多工作(Nogueira et al., 2020; Ju et al., 2021; Pradeep et al., 2021; Zhuang et al., 2023b)。而强大的大语言模型出现为这个领域又带来新的可能。按照生成方式我们将它们分为逐个生成（pointwise）、列表生成（listwise）、成对生成（pairwise）和集合生成（setwise）四大类。

逐个生成方法是指每次给定一个查询-文档对，生成式重排序模型生成它们的相关性得分。RankLLaMA(Ma et al., 2023b)将查询-文档对以“query: query document: document [EOS]”的模板输入到LLaMA模型中，并取“[EOS]”标记的最后一层嵌入表示进行相关性得分计算。由于此任务和大语言模型的预训练任务形式差距过大，因此在推理前需要对模型进行微调。(Zhuang et al., 2023a)等人认为生成二进制相关性标签（如‘True’或‘False’）的方法由于缺少中间相关性标签选项可能会导致大语言模型为与查询部分相关的文档提供嘈杂或有偏见的答案，因此他们将细粒度相关性标签合并到生成式重排序器的提示中，使它们能够更好地区分与查询具有不同相关性级别的文档，从而得出更准确的排名。与让重排序器生成相关性标签的工作相反，也有一部分工作仅给定文档，然后计算基于此文档生成实际查询的平均对数似然分数来确定此查询-文档对的相关性分数。(Sachan et al., 2022)等人直接计算以文档为条件的输入查询的概率，在完全开放域问答上取得了最先进的性能。(Zhuang et al., 2023c)等人重点研究了近期大语言模型真正的零样本查询似然排名有效性。同时他们还引入了一种新颖的先进排名系统，该系统将基于大语言模型的查询似然模型与混合零样本检索器相结合，在零样本和小样本场景中都表现出卓越的有效性。

列表生成方法将查询和所有相关文档输入生成式重排序器，然后直接让其按照相关性顺序

输出文档标识符。(Ma et al., 2023c)等人采用零样本列表生成方法在三个网络搜索数据集上进行实验，结果表明零样本列表生成方法不仅在对第一阶段检索结果进行重排序时优于零样本逐个生成方法，而且还可以充当最终阶段的重排序器，以改进逐个生成方法的排名结果，从而提高效率。(Pradeep et al., 2023)等人发布了第一个完全开源的生成式重排序大语言模型，叫做RankVicuna。它能够在零样本设置中执行高质量的列表式重排序，并取得与基于GPT-3.5的零样本列表重排序相当的有效性。然而列表生成方法对输入中的文档顺序异常敏感(Sun et al., 2023b)，当文档顺序随机打乱时它的效果甚至不如BM25。为此(Tang et al., 2023a)等人提出了一种被叫做排列自洽的方式，它的主要思想是边缘化提示中的不同列表顺序，以产生具有较少位置偏差的顺序无关的排名。他们从理论上证明了此方法的稳健性，表明在存在随机扰动的情况下可以收敛到真实排名。最终的效果超越了之前列表重新排序的最高水平。

成对生成方法每次给定一个查询和两个文档，然后要求生成式重排序模型生成相关性更高的文档标识符。最后再采用排序算法对所有相关文档进行重排序。(Qin et al., 2023b)等人认为现有的大语言模型无法理解逐一生成和逐列表生成的方式。因此他们使用成对排名提示(PRP)的新技术来显著减轻大语言模型的负担，并使用中等规模的开源大语言模型在标准基准上达到了最佳排名。

集合生成方法每次给定一个查询和一些文档，然后要求生成式重排序模型生成最相关的文档标识符或利用所有文档标识符的logits进行相关性排序。(Zhuang et al., 2023d)等人首次提出setwise提示方法，这是对之前三种方法的补充。通过在一致的实验框架内进行比较评估，并考虑模型大小、生成消耗、延迟等因素，他们表明setwise方法本质上在有效性和效率之间的平衡。比如pointwise方法在效率方面得分很高，但其有效性较差。相反，pairwise 方法表现出卓越的有效性，但会产生高计算开销。而setwise方法减少了排名过程中大语言模型的推理次数。这显著提高了基于大语言模型的零样本排名的效率，同时还保持了较高的零样本排名有效性。

4 大语言模型主导的信息检索新范式

本章节主要介绍以大语言模型为主导的信息检索新范式，主要包含新型搜索引擎，区别于基于传统信息检索方式的传统搜索引擎，以及在新型搜索引擎中广告模型应该如何适应。

4.1 新型搜索引擎

基于大语言模型的新型搜索引擎充分利用了大语言模型的生成能力，为用户提供了更好的检索体验。(Ziems et al., 2023b)发现大语言模型可以遵循人类的指令，直接生成用于文档检索的URL，大语言模型可以被认为是内置的搜索引擎。(Tang et al., 2024a)等人提出了一个端到端并且由大语言模型驱动的信息检索架构——自检索，其中IR系统所需的能力可以完全内化到一个单一的大语言模型中，并在IR过程中深入利用大语言模型的能力。本节根据大语言模型扮演的不同角色分为代理式检索和交互式检索。

4.1.1 代理式检索

近年来，以大语言模型为核心的代理式检索方法逐渐受到关注和应用，展现出了巨大的潜力和实际效果。由于大语言模型的知识是有限的，因此出现了通过检索等方式结合外部知识提高大语言模型的生成能力，并称之为检索增强生成RAG(Lewis et al., 2020a)。由此，有部分研究通过网络检索来提高大语言模型的能力，在这一类检索中，首先从数据索引中检索相关条目，然后代理以大语言模型为核心，模仿人类浏览网页的操作处理检索到的条目，以使用模型进行最终预测。

WebGPT(Nakano et al., 2021)是OpenAI提出的一个创新方法解决长篇问答问题，利用Bing创建了一个基于文本的Web 浏览环境。这个系统中，经过微调的GPT-3语言模型作为检索代理，在环境中执行检索任务。WebGPT使用模仿学习和强化学习等一般方法，以端到端的方式改进了信息的检索和合成过程。此外，生成的答案中包含了来自网页段落的参考文献，从而提高了答案的可靠性和可信度。然而，WebGPT也存在一定的局限性。针对这些问题，WebGLM(Liu et al., 2023)提出了改进方案。WebGLM增强了大语言模型的网络搜索和检索功能，同时确保在实际部署中的效率。通过引入人类偏好的机制，WebGLM提高了模型在实际应用中的表现，使其能够更高效、更准确地满足用户的需求。WebShop(Yao et al., 2022)展示了一个更为复杂和真实的应用场景。WebShop开发了一个模拟电子商务网站环境，拥有118万种真实产品和12,087条众包文本指令。在这个环境中，代理需要浏览多种类型的网页，并根据

指令执行不同的操作，如查找、定制和购买产品。这不仅测试了语言模型在复杂交互任务中的表现，也为未来的实际应用提供了宝贵的经验。在中文信息检索领域，WebCPM(Qin et al., 2023a)则提出了第一个中文长篇问答（LFQA）数据集。WebCPM的信息检索基于交互式网络搜索，能够实时与搜索引擎交互。通过对预训练语言模型进行微调，WebCPM模仿了人类的网络搜索行为，根据收集到的事实生成答案。这种方法不仅提高了中文信息检索的效率，还增强了系统对复杂问答任务的处理能力。

针对基于检索的模型，(Basu et al., 2022)提出了一种正式处理方法来描述它们的泛化能力。通过分析这些模型在不同场景下的表现，研究揭示了它们在面对不同类型问题时的适应能力和局限性。这些理论研究为进一步优化和改进检索模型提供了重要的参考和指导。尽管大语言模型驱动的Web代理在信息检索方面展现了巨大潜力，但也面临着安全威胁。WIPI(Wu et al., 2024)介绍了一种新型威胁，能够间接控制Web代理执行嵌入在公开网页中的恶意指令。即使在黑盒环境下，这种方法仍然能够实现超过90%的攻击成功率，揭示了当前Web代理的安全漏洞。这为未来更安全的大语言模型系统设计提供了重要见解和方向。

以大语言模型为核心的代理式检索方法结合了大语言模型和信息检索，展示了巨大的潜力。通过不断改进和优化这些模型，我们可以更高效、更准确地利用大语言模型从海量数据中提取有用信息，从而满足用户日益增长的需求。

4.1.2 交互式检索

交互式检索（Interactive Information Retrieval, IIR）是一种通过多轮对话和反馈的方式进行信息检索的方法，旨在提升检索的精确度和用户满意度。这个过程强调用户在检索中的主动参与和系统对用户反馈的即时响应。通过多轮交互，系统能够更好地理解用户需求，提供更相关的结果。

但是在交互式检索的过程中，不仅需要从单个查询中理解用户的意图，更需要结合上下文，要求解锁模型有更强的上下文理解能力。由于大语言模型的迅速发展，其强大的自然语言处理能力和上下文理解能力，能够显著提升交互式检索系统的智能化水平和用户体验，使检索过程更加高效、精准和个性化。

和传统会话检索的区别 与传统的关键词搜索相比，IIR的交互性支持逐步深入的查询，能够更好地满足用户的复杂需求，并提供了更自然的用户体验，给信息检索领域带来了新的机遇和挑战(Vtyurina et al., 2017; Radlinski and Craswell, 2017)。因此，在大语言模型出现之前，已经存在多轮交互的信息检索研究，如会话检索，在传统的信息检索流程中引入多轮交互信息，提高信息检索的性能。Radlinski和Craswell(Radlinski and Craswell, 2017)提出了会话搜索的理论框架。在对话中的每一个来回步骤中，系统向用户提供一些信息，用户做出响应。在会话检索系统中，用户可以像与人对话一样与系统交互，提出问题并根据系统提供的结果进行进一步的提问和调整，经过多轮会话得到信息(Gao et al., 2023)。

作为大语言模型的前身，预训练语言模型（PLMs）如Bert、GPT-2等在大语言模型之前就曾被多次用于多轮交互的信息检索，在传统信息检索的步骤中起作用(Dalton et al., 2020; Voskarides et al., 2020)。又有研究建立模型通过检索恰当的澄清问题与用户交互以明确用户的查询需求，(Aliannejadi et al., 2019)等人制定了在开放域会话系统中搜索信息的问题澄清的任务，模型通过多轮主动询问来逐步明确用户需求。他们提出了一个检索框架，包括三个组成部分：问题检索，问题选择，和文档检索。而在此之后，(Zamani et al., 2020)分析了从Bing 搜索日志中采样的查询重构数据来确定开放域搜索查询的澄清分类法，以进一步研究为开放域搜索任务生成澄清问题。

可见会话检索仍然基于传统的信息检索流程，通常包含三个组件构成：上下文查询理解，文档检索（包含建立索引）和文档排名(Gao et al., 2023)。与一般的关键字信息检索相比，会话检索由多轮对话组成，需要更加强大的上下文理解能力，但由于之前技术的局限性，这一部分的研究具有一定的挑战性，难以大规模普及。而大语言模型出色的自然语言理解和上下文理解能力，为交互式检索带来新的突破。与会话检索不同，基于大语言模型的交互式检索采用新的信息检索范式，充分利用大语言模型的自然语言生成能力，采用生成问题而不是检索问题(Aliannejadi et al., 2019)的方式与用户对话，打破了原有的信息检索的流程框架。同时，其生成能力也为满足用户的新的检索意图（如，创作）提供了发展前景。

新一代可应用的交互式检索 大语言模型的发展为传统搜索引擎带来巨大的改变。以往只能输入一次关键字，搜索引擎返回大量排序后的搜索结果，需要用户再次鉴别结果，如

果检索失败，用户需要重新搜索(Maoro et al., 2024)。Microsoft曾揭示传统搜索引擎中的问题——几乎一半的网络搜索都得不到准确的答复。最近，大语言模型已与网络搜索相结合，以实现一种新的大语言模型驱动的交互式检索模式。此外，大语言模型驱动的搜索引擎如New Bing、Perplexity AI等能够理解以自然语言表达的复杂查询，使用户能够像在对话中一样直观地提问，为用户提供更准确的搜索结果，甚至可以提供创作和编写的功能(Ma et al., 2024)。

New Bing，通常称为Bing Chat或者简单称为Bing，是微软Bing搜索引擎的一次重大升级，结合了先进的人工智能功能，以增强搜索体验。New Bing集成了OpenAI的GPT-4等大语言模型。Bing提供实时更新信息和强大的视觉搜索功能，能够与其他微软服务和产品集成，创造了一个无缝的微软生态系统体验。Perplexity AI是一个基于大语言模型的搜索引擎，旨在通过理解问题的上下文并以对话形式提供精确、相关的信息来增强搜索体验。Perplexity AI的核心是先进的人工智能和机器学习算法，通过用户互动不断学习和改进。和New Bing一样，Perplexity AI 力求提供最新信息，使其成为获取当前事件和动态话题的有用工具。Bard是Google开发的一款基于人工智能的对话式搜索引擎，旨在利用自然语言处理技术提供更自然和互动的搜索体验。Bard支持多轮对话，用户可以在一次搜索中提出后续问题，Bard能够记住上下文并继续提供相关的回答。通过了解用户的偏好和兴趣，Bard可以提供个性化的搜索结果，You.com是一个新兴的搜索引擎，旨在通过整合人工智能和机器学习技术，提供个性化和隐私友好的搜索体验。You.com通过了解用户的偏好和兴趣，提供个性化的搜索结果，使用户能够快速找到最相关的信息。

(Ma et al., 2024)剖析了大语言模型驱动的交互式搜索引擎（特别是Bing Chat）为其生成的回答选择信息源的机制。研究表明，Bing Chat更偏好具可读性和分析性的源内容，而且其对文本的独特倾向是可以被底层大语言模型预测的。同时也揭示了RAG API和Bing Chat之间的一致文本偏好。(Gong and Cosma, 2023)介绍了一种新颖的跨模态搜索引擎Boon，它结合了两个最先进的网络：GPT-3.5-turbo大模型和VSE网络VITR，使用户能够执行图像到文本和文本到图像的检索并且能够就其选择的一个或多个图像进行对话。然而，(Wazzan et al., 2024)比较了传统搜索引擎和基于大语言模型的搜索引擎（Microsoft Bing Chat）在图像地理定位搜索任务中的性能，表明使用传统搜索的参与者比使用大语言模型搜索的参与者表现更好。在(Spatharioti et al., 2023)等人的研究中，基于大语言模型的搜索引擎的参与者能够更快地完成任务，并且参与者具有更满意的体验，但是如果大语言模型提供的信息不可靠，用户仍然会过度依赖错误信息。

在其他任务如推荐任务中，基于大语言模型的交互式检索也为其提供了新的研究范式,提供了超越传统推荐技术的更自然、更无缝的用户体验。(Huang et al., 2023)结合推荐模型和大语言模型各自的优势来创建一个多功能且交互式的推荐系统，它通过集成大语言模型，使传统推荐系统成为具有自然语言界面的交互式系统。为了改善企业网站上的搜索体验，(Maoro et al., 2024)提出了一个领域自适应的问答框架，结合了语义搜索和GPT-3.5，当返回答案时，用户可以提出后续问题并进行特定主题的交流，改善了企业网站的整体用户体验。(Völker et al., 2024)介绍了一种新颖的检索增强生成系统，该系统利用基于聊天的大语言模型来简化和增强出版物管理流程，使用户能够通过直观的聊天界面与各种网页平台如SemanticScholar等无缝交互。

4.2 广告模型

对于商业性质的查询，搜索引擎会提供相关的在线广告(Dubey et al., 2024)。在传统的搜索引擎中，通过关键词匹配广告，大语言模型能够提供更加准确的广告匹配。除此之外，随着大语言模型驱动的搜索引擎的发展，广告技术也经历了显著变化。传统的广告形式和投放方式可能不适用于大语言模型生成式的检索结果。

在传统搜索引擎中，广告与查询关键词匹配至关重要。传统技术在准确刻画查询和关键词之间的语义相关性方面存在局限性。为此，(Wang et al., 2024b)提出了一种基于大语言模型的关键词生成方法（LKG），能够一步到位地从搜索查询中提取相关关键词。这种方法利用大语言模型的语义理解能力，显著提高了关键词匹配的准确性和广告的投放效果。

广告的嵌入需要考虑到广告商出价等因素，因此需要一个新颖的拍卖机制来整合来自不同广告商的输入。将在线广告模型和拍卖框架转移到大语言模型环境中，带来了新的机遇和挑战。(Feizi et al., 2023)提出了一个合理的框架，包括修改大语言模型的原始输出、广告商为修改后的输出竞标、大语言模型计算广告的相关信息以及广告竞争并选择最终输出。这

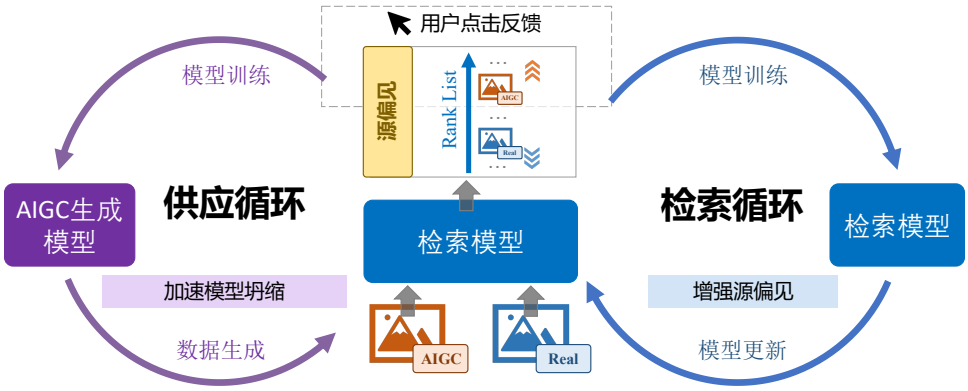


图 2: 源偏见问题示意图。IR模型倾向于将AI生成的图像排在真实图像之前，尽管它们具有非常相似的语义。这种偏见增加了生成的图像从互联网海量数据中暴露出来的可能性，这使得它们更容易被混入AIGC和检索模型的训练中，从而导致更严重的偏见并形成恶性循环。

种框架有效地整合了大语言模型的生成能力和广告竞价机制，提升了广告投放的精准度和效果。(Dütting et al., 2024)是第一篇引入大语言模型机制设计的论文，它提出了一种基于token的拍卖模型，该模型以大语言模型作为广告商代理，通过出价影响生成的广告内容。其中，广告段落是逐个token生成的，竞标者之间的出价在不同大语言模型中分布式聚合，从而产生优化的广告创意。与这样的方法不同，(Dubey et al., 2024)提出了一种分解框架，包括拍卖模块和大语言模型模块，其中分配和支付仍由拍卖模块决定，而大语言模型模块则根据拍卖模块分配的突出性生成广告摘要。这种方法不仅提高了广告的展示效果，还优化了广告内容的分配和生成。为了解决(Dütting et al., 2024)中广告商的支出会随着大语言模型生成的输出序列的长度而增长等限制，(Soumalias et al., 2024)引入了一种拍卖机制，不需要对大语言模型微调和访问模型的权重，就能汇总了多个广告商代理对用户查询回复的偏好需求，同时也为用户提供了有用的回答。

除了将广告融入大语言模型驱动的搜索引擎之外，(Schmidt et al., 2024)提到了对原生广告的检测。在大语言模型驱动的搜索引擎中，广告嵌入到生成的响应中，未来的用户很可能会面临生成的原生广告。研究表明，大语言模型也可以用于检测并阻止生成的原生广告，从而保护用户体验和信息的真实性。

大语言模型在搜索引擎广告中的应用，为广告技术带来了巨大的变化。通过引入智能拍卖机制、生成广告摘要等技术，广告商能够更精准地投放广告并适用于新的搜索引擎。

5 大语言模型对检索生态的影响

如第3节和第4节所述，大语言模型不仅对传统信息检索范式产生巨大影响，还产生了信息检索的新范式。这种变化带来新机遇的同时也带来了新的挑战，特别是大语言模型产生的偏见和不公平可能会威胁现在的检索生态。本节系统地研究了应用大语言模型可能会带来的偏见问题、公平性问题和创作消费问题。

5.1 偏见问题

大语言模型的出现让人们可以轻易地生成大规模合成数据，而这些合成数据会重溯检索数据的分布(Dai et al., 2024)。最近的研究(Dai et al., 2023a; Xu et al., 2023c)表明，现代检索模型，尤其是基于神经网络的模型，更倾向于检索由大语言模型生成的内容，而非人类创作的具有相似语义的内容。这种现象被称之为源偏见问题。如图 2所示。它的产生原因是，大语言模型生成的文本具有某种独特的嵌入表示。而神经检索模型可以捕捉这种表示，从而产生更高的排名。如果检索模型使用了合成数据进行训练，那么这种偏差会进一步放大。另外也有研究(Tan et al., 2024)表明，源偏见问题会从检索模型进一步延伸到生成模型。为了解决源偏见问题，目前的研究(Dai et al., 2023a; Xu et al., 2023c)主要集中于在检索模型的训练过程中引入去偏差约束。它们的思想在于从分布对齐角度将检索模型的相关性分布重新校正为理想状态，从而对不同来源的文档进行公平对待。

除了源偏见外，事实偏见也是另外一种常见的偏见问题，它被定义为大语言模型可能会产生与现实世界公认的事实信息不一致的内容(Dai et al., 2024)。(Lin et al., 2022; Lee et al., 2022c; McKenna et al., 2023)等人证明大语言模型会生成许多错误的答案，并且有可能欺骗人类。即使参数量更大的模型也无法避免这个问题。此外大语言模型在一些大规模基准(Chen et al., 2023b; Lee et al., 2022b; Deng et al., 2024; Wei et al., 2024a; Wei et al., 2024b)上的表现也验证了这种现象。事实偏见引入了大量非事实或"幻觉"内容。这种引入改变了检索数据的分布，从而导致检索过程中的偏差。目前有许多工作尝试缓解事实偏差。比如一部分(Gunasekar et al., 2023; Touvron et al., 2023b)侧重于为大语言模型提供高质量且事实正确的数据。还有一部分为在可信数据源中检索信息来增强大语言模型的生成(Ram et al., 2023; Lewis et al., 2020b; Shi et al., 2023; Deng et al., 2023; Xu et al., 2024d; Ding et al., 2024; Xu et al., 2024c)或者利用大语言模型自身的推理能力避免生成非事实问题(Xu et al., 2024b; Wang et al., 2023b; Chuang et al., 2023)。

5.2 不公平问题

(Dai et al., 2024)等人认为在信息检索系统中存在用户公平和项目（文档）公平两个概念，它们分别和社会学领域的平等和分配正义(Xu et al., 2023a)概念有关。具体来说用户公平被定义为信息检索系统应向不同用户提供公平和非歧视性的信息服务。而项目公平被定义为信息检索系统应该为较弱的项目提供更多的被检索到的机会。而大语言模型对检索生态会产生用户不公平和项目不公平的问题。

产生用户不公平的一个主要原因是检索数据中存在的歧视或者攻击性内容对特定群体产生了不成比例的影响。这些内容之所以会出现在检索数据中既可能是历史或者文化原因(Beukeboom and Burgers, 2019; Ntoutsi et al., 2020; Zhuo et al., 2023),也可能是大语言模型生成的(Fang et al., 2023)。此外，大语言模型之所以会生成这些内容，也源于它们的训练数据当中歧视性内容的文本(Beukeboom and Burgers, 2019)。以往的工作主要采用各种方法来过滤这些文本。比如(Ghanbarzadeh et al., 2023)等人通过性别调整来构造更公平的数据集来消除模型的偏见。(Xu et al., 2023b)等人经过大量的分析证实了大语言模型存在隐性的用户歧视，并强调识别和减轻隐性用户歧视的必要性。还有一些方法(Deldjoo and Noia, 2024; Ngo et al., 2021)使用降低包含歧视信息样本的重采样策略或者直接过滤和删除这些内容。此外，指令微调或基于人类反馈的强化学习也被证明可以让大模型对齐人类的价值观以有效促进公平性(Touvron et al., 2023b)。

产生项目不公平性问题的一个主要原因为某些项目的代表性不平衡导致在信息检索或评估过程当中的差异(Jiang et al., 2024)。此外大语言模型也有可能生成新的项目或文档，从而潜在引入了新的内容和观点(Das et al., 2024; Jr. and Licato, 2023)。为了减轻数据当中的项目不公平性，(Jiang et al., 2024)等人提出为不同的项目或文档进行重新加权以平衡项目或文档的代表性。

5.3 创作消费问题

大语言模型极大地降低了人们的创作门槛，让普通用户通过使用强大的内容生成模型（例如ChatGPT、Sora和GPT-4o）也能够创作高质量的文本或视频，这给在线内容生态系统和检索生态系统注入了新活力(Epstein et al., 2023)。然而，这种使用人工智能来进行创作消费的转型也带来了新的问题，即会导致市场过度饱和，个人创作者的内容更加难以被挖掘。另外，大语言模型并不是万能的，如第5.1和5.2节所述，它具有偏见和不公平的现象。如果高质量的人类创作内容被边缘化，那么依赖于广泛且多样化的数据集进行训练的大语言模型生成的质量必然下降(Yao et al., 2024)。因此探索人类生成内容和人工智能生成内容是否能稳定地以共生的形式存在是一个具有挑战性的方向。(Yao et al., 2024)等人拓展了Tullock竞争模型，从理论和实验上给出了一个具有希望的前景，即尽管生成式人工智能会扰乱人类内容生成的市场，但具有理想特征的稳定平衡是可以实现的。

6 未来发展

降低在信息检索系统中应用大语言模型的成本 大语言模型不仅可以改进传统的信息检索系统，还主导了信息检索的新范式。然而使用大语言模型会产生高昂的计算成本，尤其是对于

高校实验室或者是小规模的公司而言。即使是一些具有充足计算资源的大型互联网公司，当面临大量用户请求时也会产生巨大的成本压力。常见的解决方案包括大语言模型压缩和推理加速，但是这些方法有可能会损害大语言模型性能，从而对信息检索系统造成影响。因此需要开发更高效的大语言模型使用方式以应对成本挑战。

消除大语言模型生成内容的偏见 由于大语言模型生成的内容目前几乎被信息检索的所有阶段使用，然而这会产生偏见问题。它主要表现为改变了检索数据的分布，从而让检索模型更加倾向于检索大语言模型生成的内容。虽然目前已有一些工作使用去偏差约束来缓解此问题，但是它们无法彻底解决带有偏见的内容。

让大语言模型生成的内容更可信 由于大语言模型会产生幻觉，因此人们无法完全相信大语言模型根据用户查询生成的内容。有时这些内容看起来合理但实际上确实不合逻辑的甚至是虚假的。这为现代信息检索系统产生不利影响。因此需要正确认识到大语言模型在某些方面的局限性，从而开发出可信的现代信息检索系统。

7 结论

本文对大语言模型对信息检索领域的影响进行了深入的阐述。针对传统信息检索系统，大语言模型凭借出色的理解能力改变了索引模块、检索模块和排序模块。此外本文还探讨了大语言模型可能取代传统信息检索方法的趋势，并催生出新的信息检索范式，预示着信息时代的新发展。同时，本文也关注了大语言模型对内容生态的影响，包括偏见、不公平问题以及创作消费问题。此外，虽然将大语言模型应用到信息检索系统的前景广阔，但同时也带来了一系列挑战。未来的发展方向应从降低成本、消除偏见，提高可信度这几个方面去考虑。

参考文献

Mohammad Aliannejadi, Hamed Zamani, Fabio A. Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Abhijit Anand, Venktesh V, Vinay Setty, and Avishek Anand. 2023. Context aware query rewriting for text rankers using LLM. *CoRR*, abs/2308.16753.

Arian Askari, Mohammad Aliannejadi, Chuan Meng, Evangelos Kanoulas, and Suzan Verberne. 2023. Expand, highlight, generate: Rl-driven document generation for passage reranking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10087–10099. Association for Computational Linguistics.

Soumya Basu, Ankit Singh Rawat, and Manzil Zaheer. 2022. Generalization properties of retrieval-based models. *CoRR*, abs/2210.02617.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *CoRR*, abs/2404.05961.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: a review and introduction of the aocial categories and stereotypes communication (scsc) framework. *Review of Communication Research*, 7:1–37.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick S. H. Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, and Rodrigo Frassetto Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *CoRR*, abs/2202.05144.

Leonid Boytsov, Preksha Patel, Vivek Sourabh, Riddhi Nisar, Sayani Kundu, Ramya Ramanathan, and Eric Nyberg. 2023. Inpars-light: Cost-effective unsupervised training of efficient rankers. *CoRR*, abs/2301.02998.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. GERE: generative evidence retrieval for fact verification. In Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai, editors, *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2184–2189. ACM.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2023a. A unified generative retriever for knowledge-intensive language tasks via prompt learning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23. ACM, July.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2023b. Complex claim verification with evidence retrieved in the wild. *CoRR*, abs/2305.11859.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, and Jun Xu. 2023a. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *CoRR*, abs/2310.20501.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2023b. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Unifying bias and unfairness in information retrieval: A survey of challenges and opportunities with large language models. *CoRR*, abs/2404.11457.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC cast 2019: The conversational assistance track overview. *CoRR*, abs/2003.13624.

Debarati Das, Karin de Langis, Anna Martin-Boyle, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Anugrah Hayati, Risako Owan, Bin Hu, Ritik Parkar, Ryan Koo, Jong Inn Park, Aahan Tyagi, Libby Ferland, Sanjali Roy, Vincent Liu, and Dongyeop Kang. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *CoRR*, abs/2401.14698.

Yashar Deldjoo and Tommaso Di Noia. 2024. Cfairllm: Consumer fairness evaluation in large-language model recommender system. *CoRR*, abs/2403.05668.

Jingcheng Deng, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Regavae: A retrieval-augmented gaussian mixture variational auto-encoder for language modeling. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2500–2510. Association for Computational Linguistics.

Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024. Unke: Unstructured knowledge editing in large language models. *arXiv preprint arXiv:2405.15349*.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2024. Retrieve only when it needs: Adaptive retrieval augmentation for hallucination mitigation in large language models. *CoRR*, abs/2402.10612.

Kumar Avinava Dubey, Zhe Feng, Rahul Kidambi, Aranyak Mehta, and Di Wang. 2024. Auctions with LLM summaries. *CoRR*, abs/2404.08126.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. 2024. Mechanism design for large language models. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 144–155. ACM.

Ziv Epstein, Aaron Hertzmann, Laura Mariah Herman, Robert Mahari, Morgan R. Frank, Matthew Groh, Hope Schroeder, Amy Smith, Memo Akten, Jessica Fjeld, Hany Farid, Neil Leach, Alex Pentland, and Olga Russakovsky. 2023. Art and the science of generative AI: A deeper dive. *CoRR*, abs/2306.04141.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: An examination of news produced by large language models. *CoRR*, abs/2309.09825.

Soheil Feizi, MohammadTaghi Hajiaghayi, Keivan Rezaei, and Suho Shin. 2023. Online advertisements with llms: Opportunities and challenges. *CoRR*, abs/2311.07601.

Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2023. Synergistic interplay between search and large language models for information re. *arXiv preprint arXiv:2305.07402*.

Fernando Ferraretto, Thiago Laitz, Roberto de Alencar Lotufo, and Rodrigo Frassetto Nogueira. 2023. Exaranker: Explanation-augmented neural ranker. *CoRR*, abs/2301.10521.

Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural Approaches to Conversational Information Retrieval*, volume 44 of *The Information Retrieval Series*. Springer.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5448–5458. Association for Computational Linguistics.

Yan Gong and Georgina Cosma. 2023. Boon: A neural search engine for cross-modal information retrieval. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*, MM '23. ACM, October.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *CoRR*, abs/2306.11644.

Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender AI agent: Integrating large language models for interactive recommendations. *CoRR*, abs/2308.16505.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *CoRR*, abs/2305.03653.

Vitor Jeronymo, Luiz Henrique Bonifacio, Hugo Queiroz Abonizio, Marzieh Fadaee, Roberto de Alencar Lotufo, Jakub Zavrel, and Rodrigo Frassetto Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *CoRR*, abs/2301.01820.

Meng Jiang, Keqin Bao, Jizhi Zhang, Wenjie Wang, Zhengyi Yang, Fuli Feng, and Xiangnan He. 2024. Item-side fairness of large language model-based recommendation system. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 4717–4726. ACM.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Antonio Laverghetta Jr. and John Licato. 2023. Generating better items for cognitive assessments using large language models. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2023, Toronto, Canada, 13 July 2023*, pages 414–428. Association for Computational Linguistics.

Jia-Huei Ju, Jheng-Hong Yang, and Chuan-Ju Wang. 2021. Text-to-text multi-view learning for passage re-ranking. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1803–1807. ACM.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022a. Generative multi-hop retrieval. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1417–1436. Association for Computational Linguistics.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022b. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8424–8445. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022c. Factuality enhanced language models for open-ended text generation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024a. Nv-embed: Improved techniques for training llms as generalist embedding models. *CoRR*, abs/2405.17428.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernández Ábrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024b. Gecko: Versatile text embeddings distilled from large language models. *CoRR*, abs/2403.20327.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pages 393–401. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle,

Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chaofan Li, Zheng Liu, Shitao Xiao, and Yingxia Shao. 2023a. Making large language models A better foundation for dense retrieval. *CoRR*, abs/2312.15503.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023b. Generative retrieval for conversational question answering. *Inf. Process. Manag.*, 60(5):103475.

Xiaoxi Li, Zhicheng Dou, Yujia Zhou, and Fangchao Liu. 2024a. Towards a unified language model for knowledge-intensive tasks utilizing external corpus. *CoRR*, abs/2402.01176.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024b. Learning to rank in generative retrieval. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8716–8723. AAAI Press.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In Ambuj K. Singh, Yizhou Sun, Leman Akoglu, Dimitrios Gunopulos, Xifeng Yan, Ravi Kumar, Fatma Ozcan, and Jieping Ye, editors, *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 4549–4560. ACM.

Guangyuan Ma, Xing Wu, Peng Wang, Zijia Lin, and Songlin Hu. 2023a. Pre-training with large language model-based document expansion for dense passage retrieval. *CoRR*, abs/2308.08285.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023b. Fine-tuning llama for multi-stage text retrieval. *CoRR*, abs/2310.08319.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023c. Zero-shot listwise document reranking with a large language model. *CoRR*, abs/2305.02156.

Lijia Ma, Xingchen Xu, and Yong Tan. 2024. Crafting knowledge: Exploring the creative mechanisms of chat-based search engines. *CoRR*, abs/2402.19421.

Falk Maoro, Benjamin Vehmeyer, and Michaela Geierhos. 2024. Leveraging semantic search and llms for domain-adaptive information retrieval. In Audrius Lopata, Daina Gudonienė, and Rita Butkienė, editors, *Information and Software Technologies*, pages 148–159, Cham. Springer Nature Switzerland.

Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. 2023. Sources of hallucination by large language models on inference tasks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 2758–2774. Association for Computational Linguistics.

Rui Meng, Ye Liu, Semih Yavuz, Divyansh Agarwal, Lifu Tu, Ning Yu, Jianguo Zhang, Meghana Bhat, and Yingbo Zhou. 2022. Augtriever: Unsupervised dense retrieval by scalable data augmentation. *arXiv preprint arXiv:2212.08841*.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Usama Nadeem, Noah Ziems, and Shaoen Wu. 2022. Codedsi: Differentiable code search. *CoRR*, abs/2210.00328.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Ambesh Negi, Mayur Bhirud, Suresh Jain, and Amit Mittal. 2012. Index based information retrieval system. *International Journal of Modern Engineering Research (IJMER)*, 2:945.

Helen Ngo, Cooper Raterink, João G. M. Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *CoRR*, abs/2108.07790.

Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics.

Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, Ioannis Kompatsiaris, Katharina Kinder-Kurlanda, Claudia Wagner, Fariba Karimi, Miriam Fernández, Harith Alani, Bettina Berendt, Tina Kruegel, Christian Heinze, Klaus Broelemann, Gjergji Kasneci, Thanassis Tiropanis, and Steffen Staab. 2020. Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl. Discov.*, 10(3).

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tabea Margareta Grace Pakull, Hendrik Damm, Ahmad Idrissi-Yaghir, Henning Schäfer, Peter A. Horn, and Christoph M. Friedrich. 2024. Wispermed at biolaysumm: Adapting autoregressive large language models for lay summarization of scientific articles. *CoRR*, abs/2405.11950.

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Jingfang Xu, and Xueqi Cheng. 2017. Deeprank: A new deep architecture for relevance ranking in information retrieval. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17. ACM, November.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing gender-inclusive machine translation with neomorphemes and large language models. *CoRR*, abs/2405.08477.

Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *CoRR*, abs/2101.05667.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *CoRR*, abs/2309.15088.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, Ruobing Xie, Fanchao Qi, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023a. Webcpm: Interactive web search for chinese long-form question answering. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8968–8988. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023b. Large language models are effective text rankers with pairwise ranking prompting. *CoRR*, abs/2306.17563.

Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, page 117–126, New York, NY, USA. Association for Computing Machinery.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *CoRR*, abs/2302.00083.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md. Arafat Sultan, and Christopher Potts. 2023. UDAPDR: unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11265–11279. Association for Computational Linguistics.

Devendra Singh Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3781–3797. Association for Computational Linguistics.

Sebastian Schmidt, Ines Zelch, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. Detecting generated native ads in conversational search. In Tat-Seng Chua, Chong-Wah Ngo, Roy Ka-Wei Lee, Ravi Kumar, and Hady W. Lauw, editors, *Companion Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, Singapore, May 13-17, 2024*, pages 722–725. ACM.

Tao Shen, Guodong Long, Xiubo Geng, Chongyang Tao, Tianyi Zhou, and Daxin Jiang. 2023. Large language models are strong zero-shot retriever. *CoRR*, abs/2304.14233.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Ermis Soumalias, Michael J. Curry, and Sven Seuken. 2024. Truthful aggregation of llms with an application to online advertising. *CoRR*, abs/2405.05905.

Sofia Eleni Spatharioti, David M. Rothschild, Daniel G. Goldstein, and Jake M. Hofman. 2023. Comparing traditional and llm-based search for consumer choice: A randomized experiment. *CoRR*, abs/2307.03744.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *CoRR*, abs/2402.15449.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2023a. Learning to tokenize for generative retrieval. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023b. Is chatgpt good at search? investigating large language models as re-ranking agents. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 14918–14937. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *CoRR*, abs/2401.11911.

Raphael Tang, Xinyu Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. 2023a. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. *CoRR*, abs/2310.07712.

Yanran Tang, Ruihong Qiu, and Xue Li. 2023b. Prompt-based effective input reformulation for legal case retrieval. In Zhifeng Bao, Renata Borovica-Gajic, Ruihong Qiu, Farhana Murtaza Choudhury, and Zhengyi Yang, editors, *Databases Theory and Applications - 34th Australasian Database Conference, ADC 2023, Melbourne, VIC, Australia, November 1-3, 2023, Proceedings*, volume 14386 of *Lecture Notes in Computer Science*, pages 87–100. Springer.

Qiaoyu Tang, Jiawei Chen, Bowen Yu, Yaojie Lu, Cheng Fu, Haiyang Yu, Hongyu Lin, Fei Huang, Ben He, Xianpei Han, Le Sun, and Yongbin Li. 2024a. Self-retrieval: Building an information retrieval system with one large language model. *CoRR*, abs/2403.00801.

Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, and Xueqi Cheng. 2024b. Listwise generative retrieval models via a sequential learning process. *CoRR*, abs/2403.12499.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Prakash Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2022. Transformer memory as a differentiable search index. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20. ACM, July.

Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles L. A. Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 2187–2193, New York, NY, USA. Association for Computing Machinery.

Tom Völker, Jan Pfister, Tobias Koopmann, and Andreas Hotho. 2024. From chat to publication management: Organizing your related work using bibsonomy & llms. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '24. ACM, March.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2022. A neural corpus indexer for document retrieval. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9414–9423. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023c. NOVO: learnable and interpretable document identifiers for model-based IR. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 2656–2665. ACM.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. *CoRR*, abs/2401.00368.

Yang Wang, Zheyi Sha, Kunhai Lin, Chaobing Feng, Kunhong Zhu, Lipeng Wang, Xuewu Jiao, Fei Huang, Chao Ye, Dengwu He, Zhi Guo, Shuanglong Li, and Lin Liu. 2024b. One-step reach: Llm-based keyword generation for sponsored search advertising. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1604–1608, New York, NY, USA. Association for Computing Machinery.

Albatool Wazzan, Stephen MacNeil, and Richard Souvenir. 2024. Comparing traditional and llm-based search for image geolocation. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval*, CHIIR '24. ACM, March.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2024a. Mlake: Multilingual knowledge editing benchmark for large language models. *CoRR*, abs/2404.04990.

Zihao Wei, Liang Pang, Hanxing Ding, Jingcheng Deng, Huawei Shen, and Xueqi Cheng. 2024b. Stable knowledge editing in large language models. *CoRR*, abs/2402.13048.

Fangzhou Wu, Shutong Wu, Yulong Cao, and Chaowei Xiao. 2024. WIPI: A new web threat for llm-driven web agents. *CoRR*, abs/2402.16965.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-pack: Packaged resources to advance general chinese embedding. *CoRR*, abs/2309.07597.

Chen Xu, Sirui Chen, Jun Xu, Weiran Shen, Xiao Zhang, Gang Wang, and Zhenhua Dong. 2023a. P-MMF: provider max-min fairness re-ranking in recommender system. In Ying Ding, Jie Tang, Juan F. Sequeda, Lora Aroyo, Carlos Castillo, and Geert-Jan Houben, editors, *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 3701–3711. ACM.

Chen Xu, Wenjie Wang, Yuxin Li, Liang Pang, Jun Xu, and Tat-Seng Chua. 2023b. Do llms implicitly exhibit user discrimination in recommendation? an empirical study. *CoRR*, abs/2311.07054.

Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. 2023c. Ai-generated images introduce invisible relevance bias to text-image retrieval. *CoRR*, abs/2311.14084.

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023d. BERM: training the balanced and extractable representation for matching to improve generalization ability of dense retrieval. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6620–6635. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024a. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1362–1373. ACM.

Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024b. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1362–1373. ACM.

Shicheng Xu, Liang Pang, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024c. List-aware reranking-truncation joint model for search and retrieval-augmented generation. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1330–1340. ACM.

Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024d. Unsupervised information refinement training of large language models for retrieval-augmented generation. *CoRR*, abs/2402.18150.

Tianchi Yang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Auto search indexer for end-to-end document retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6955–6970. Association for Computational Linguistics.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 20744–20757. Curran Associates, Inc.

Fan Yao, Chuanhao Li, Denis Nekipelov, Hongning Wang, and Haifeng Xu. 2024. Human vs. generative AI in content creation competition: Symbiosis or conflict? *CoRR*, abs/2402.15467.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hamed Zamani, Susan T. Dumais, Nick Craswell, Paul N. Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. *Proceedings of The Web Conference 2020*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In Tat-Seng Chua, Chong-Wah Ngo, Ravi Kumar, Hady W. Lauw, and Roy Ka-Wei Lee, editors, *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 1441–1452. ACM.

Jun Zhao, Jingqi Tong, Yurong Mou, Ming Zhang, Qi Zhang, and Xuanjing Huang. 2024. Exploring the compositional deficiency of large language models in mathematical reasoning. *CoRR*, abs/2405.06680.

Junkai Zhou, Liang Pang, Huawei Shen, and Xueqi Cheng. 2023. Think before you speak: Cultivating communication skills of large language models via inner monologue. *CoRR*, abs/2311.07445.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *CoRR*, abs/2308.07107.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. 2023a. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. *CoRR*, abs/2310.14122.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023b. Rankt5: Fine-tuning T5 for text ranking with ranking losses. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2308–2313. ACM.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023c. Open-source large language models are strong zero-shot query likelihood models for document ranking. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8807–8817. Association for Computational Linguistics.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2023d. A setwise approach for effective and highly efficient zero-shot ranking with large language models. *CoRR*, abs/2310.09497.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring AI ethics of chatgpt: A diagnostic analysis. *CoRR*, abs/2301.12867.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023a. Large language models are built-in autoregressive search engines. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2666–2678. Association for Computational Linguistics.

Noah Ziems, Wenhao Yu, Zhihan Zhang, and Meng Jiang. 2023b. Large language models are built-in autoregressive search engines. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2666–2678. Association for Computational Linguistics.

对齐的理论、技术与评估

吉嘉铭，邱天异，陈博远，杨耀东*
人工智能安全与治理中心，人工智能研究院，北京大学
中国，北京，100080
caisg@pku.edu.cn

摘要

人工智能对齐(AI Alignment)旨在使人工智能系统的行为与人类的意图和价值观相一致。随着人工智能系统的能力日益增强，对齐失败带来的风险也在不断增加。数百位人工智能专家和公众人物已经表达了对人工智能风险的担忧，他们认为“减轻人工智能带来的灭绝风险应该成为全球优先考虑的问题，与其他社会规模的风险如大流行病和核战争并列”(CAIS, 2023)。为了提供对齐领域的全面和最新概述，本文深入探讨了对齐的核心理论、技术和评估。首先，本文确定了人工智能对齐的四个关键目标：鲁棒性(Robustness)、可解释性(Interpretability)、可控性(Controllability)和道德性(Ethicality) (**RICE**)。在这四个目标原则的指导下，本文概述了当前人工智能对齐研究的全貌，并将其分解为两个关键组成部分：**前向对齐**和**后向对齐**。本文旨在为对齐研究提供全面且对初学者友好的调研。同时本文还发布并持续更新网站 www.alignmentsurvey.com，该网站提供了一系列教程、论文集和其他资源。更详尽的讨论与分析请见 https://arxiv.org/abs/2310.19852。

关键词： 人工智能安全 ； 人工智能系统对齐 ； RICE原则

Theories, Techniques, and Evaluation of AI Alignment

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Yaodong Yang*
Center for AI Safety and Governance, Institute for AI, Peking University
China,Beijing, 100080
caisg@pku.edu.cn

Abstract

AI alignment aims to ensure that the behavior of AI systems is consistent with human intentions and values. As the capabilities of AI systems continue to increase, the risks associated with alignment failures are also rising. Hundreds of AI experts and public figures have expressed concerns about AI risks, stating that “mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war” (CAIS, 2023). To provide a comprehensive and up-to-date overview of the field of alignment, this paper delves into the core concepts, methodology, and practice of alignment. Firstly, the paper identifies four key goals of AI alignment: Robustness, Interpretability, Controllability, and Ethicality (**RICE**). Guided by these four principles, the paper outlines the current landscape of AI alignment research and breaks it down into two key components: **forward alignment** and **backward alignment**. This paper aims to offer a comprehensive and beginner-friendly survey of alignment research. Additionally, a continuously updated website, www.alignmentsurvey.com, is released, providing a range of tutorials, collections of papers, and other resources. For more detailed discussions and analyses, please refer to https://arxiv.org/abs/2310.19852.

Keywords: AI Safety , AI Alignment , RICE Principles

* 通讯作者 Corresponding Author.

1 引言

随着人工智能系统愈发强大，它们逐渐被应用于不同领域，如基于大语言模型的智能体开发(Xi et al., 2023)，以及应用深度强化学习技术来控制核聚变 (Degrave et al., 2022)。然而，人工智能系统能力的日益提升和在高风险领域的拓展应用带来了巨大的潜在风险。先进人工智能系统（如大语言模型）已经表现出了各种不良行为（如操纵(Turner et al., 2021; Perez et al., 2023; Carroll et al., 2023a; Steinhardt, 2023; Sharma et al., 2023) 和欺骗(Park et al., 2023b)），这引发了人们对人工智能系统可能带来的伦理和安全挑战的担忧。

这些担忧进一步激发了对人工智能对齐*(AI Alignment)*(Christian, 2020; Bucknall and Dori-Hacohen, 2022) 的研究努力。人工智能对齐旨在使人工智能系统的行为与人类的意图与价值观相一致(Leike et al., 2018)——它更多关注的是人工智能的意图和目标，而不是它们的能力。对齐失败（即未对齐）是人工智能可能造成危害的最突出的原因之一。这些失败背后的机制包括奖励破解(Pan et al., 2022)和目标错误泛化(Shah et al., 2022)等(§1.1)。这些对齐失败进一步被模型能力所放大，体现为谄媚(Perez et al., 2023)、欺骗(Hubinger et al., 2019)和权利寻求(Power et al., 2022)等可能危害人类社会的行为。

接下来本文将阐释实现对齐的四个关键目标(§1.2)：鲁棒性、可解释性、可控性和道德性(**RICE**)。同时，本文将当前关于对齐的研究和实践分解为四个关键领域(§1.3)：从反馈中学习(§2)、在分布偏移下学习(§3)，对齐保证(§4)和人工智能治理(§5)。这四个目标（RICE原则）和四个领域共同构成了**对齐循环**。

本文介绍了人工智能对齐的理论、技术与评估，并讨论了可能的未来研究方向。

1.1 对齐的动机

在最近的十年中，深度学习领域取得了显著的进步，其发展范围从符号系统(Smolensky, 1987; Goel, 2022)扩展到基于自监督学习的系统(Mnih et al., 2015; OpenAI, 2023a)。这一进展使得大型神经网络在各种领域中都展现出了卓越的能力。特别是在游戏环境(Silver et al., 2017; Kaufmann et al., 2023)以及复杂且高风险的真实世界应用场景(Ruff and Pappu, 2021; Degrave et al., 2022)中，基于深度学习的人工智能系统均取得了显著成就。除此之外，大语言模型在多步推理(Wei et al., 2022; Wang et al., 2023)和跨任务泛化(Brown et al., 2020; Askell et al., 2021)方面的能力也不断增强。

然而，随着人工智能系统能力的增强，其带来的风险也随之增加。近年来，两个不同的群体敲响了警钟。第一个群体关注当前的技术伦理风险，早在几年前面部识别系统已经出现了对某类人群或性别识别特别不准确的现象，而当下的大语言模型也展现出一些关乎伦理的不良行为（例如，不真实的回答(Bang et al., 2023)和对性别和移民身份等明显的偏见(Perez et al., 2023)），这些行为可能进一步加剧社会现有的不平等现象。而第二类群体关注于未来的风险——人工智能对齐与人工智能安全之间的界限正日益模糊。展望未来，人工智能系统的日益强大为在可预见的未来实现通用人工智能(AGI)提供了可能性，即人工智能系统可以在所有相关方面达到或超过人类智能(Bubeck et al., 2023)。然而，这种潮流在带来技术进步(Korinek et al., 2021)和效率提升(Furman and Seamans, 2019)的同时，也可能带来严重的风险(CAIS, 2023)，甚至是全球范围内的严重危害(Hendrycks et al., 2023; GOV.UK, 2023)和存在性风险（即威胁到人类长期生存的潜在风险）(Ord, 2020)。在CAIS (2023) 中，人工智能科学家和其他知名人士表示，减轻人工智能引发的灭绝风险应与其他社会规模的风险如大流行病和核战争一样，成为全球优先考虑的问题。11月初，英国举办了首届全球人工智能安全峰会，汇集了国际政府、领先的人工智能科技公司、民间社会团体和研究专家。峰会上发布了《布莱切利宣言》，宣言中强调共同识别人工智能安全风险、提升透明度和公平性，建立科学和证据为基础的共享理解 [0]。

具体来说，当前人工智能系统已经表现出的与人类意图相悖的不良或有害行为，被称为人工智能系统的对齐失败，这些对齐失败行为即使没有恶意行为者的滥用，也可能自然发生，并代表了人工智能的重大风险来源，包括安全隐患(Hendrycks et al., 2021c)和潜在的生存风险(Hendrycks et al., 2023)。本文总结了几类较为显著的对齐失败行为和先进人工智能系统可能

[0] https://www.gov.uk/government/topical-events/ai-safety-summit-2023。

具备的危险能力，从而为未来的对齐评估提供研究方向：

- **欺骗性对齐**：不对齐的人工智能系统可能会故意误导他们的人类监督者，而不是坚守预定的任务。这种欺骗行为已经在使用进化算法的人工智能系统中表现出来(Wilke et al., 2001; Lehman et al., 2020; Hendrycks et al., 2021c)。在这些情况下，人工智能系统演化出了区分评估和训练环境的能力。他们在评估过程中采取了战略性的悲观反应方法，故意降低了在调度程序中的繁殖率(Lehman et al., 2020)。此外，人工智能系统可能会参与一些表面上符合奖励信号的有意行为，目的是从人类监督者那里获取最大的奖励(Ouyang et al., 2022)。值得注意的是，尽管现有的大语言模型有能力提供更准确的答案，但它们偶尔会生成不准确或次优的回答(Lin et al., 2022; Chen et al., 2021)。这些欺骗行为的存在带来了重大挑战。它们破坏了人类顾问提供可靠反馈的能力(因为人类无法确定人工智能模型的输出是否真实和忠实)。此外，这种欺骗行为可以传播虚假的信念和误导信息，污染在线信息来源(Hendrycks et al., 2021c)。

- **操纵**：先进的人工智能系统可以有效地影响个人的信念，即使这些信念与真相不符(Shevlane et al., 2023)。这些系统可以产生欺骗性或不准确的输出，甚至欺骗人类顾问以达到欺骗性对齐。这样的系统甚至可以说服个人采取可能导致危险结果的行动(OpenAI, 2023a)。在大语言模型、推荐系统(系统影响用户的偏好)(Adomavicius et al., 2022)和强化学习智能体(从人类反馈中学习的代理采取策略来欺骗人类评估者)(Amodei et al., 2017)中，都存在这种行为的早期迹象。此外，当前的大语言模型已经具备了进行欺骗所需的能力。Sciadvadh (Spitale et al., 2023)已经发现GPT-3具有超人的能力，可以产生令人信服的虚假信息。鉴于所有这些早期迹象，更先进的人工智能系统可能会展示出更严重的欺骗/操纵行为。

- **违反伦理**：人工智能系统中的不道德行为涉及到违反公共利益或违反道德标准的行为——例如那些对他人造成伤害的行为。这些不良行为通常源于在人工智能系统设计中忽略了重要的人类价值观，或者向系统中引入了不适当或过时的价值观(Kenward and Sinclair, 2021)。针对这些不足的研究工作涵盖了机器伦理领域(Tolmeijer et al., 2020)，并深入探讨了关键问题，例如，人工智能应该与谁保持一致？(Santurkar et al., 2023)。

解决对齐失败带来的风险需要人工智能系统的对齐技术，以确保人工智能系统的目标与人类意图和价值观一致，从而避免非预期的不利结果。更重要的是，本文期望对齐技术能够应对更困难的任务，并且能够应用于比人类更智能的先进人工智能系统。一个可能的解决方案是超级对齐，其目标是构建一个大致与人类水平相当的自动对齐研究器，从而使用大量的计算能力来迭代并扩增对齐超智能(OpenAI, 2023b)。

1.2 对齐目标:RICE原则

我们如何构建与人类价值和意图对齐的人工智能系统?

目前并没有一个被普遍接受用来衡量对齐的标准。首先，我们必须明确本文讨论的对齐目标是什么。Leike (2018)提出智能体对齐问题，并指出了这样的问题："如何创建能够按照用户意图行事的智能体？"进一步，其将问题扩展到了超级人工智能系统上(OpenAI, 2023b)："如何确保比人类更聪明的人工智能系统遵循人类的意图？"在这些讨论中，一个一致的主题是对人类意图的关注。为了清楚地定义对齐目标，我们必须准确地描述人类的意图，正如Kenton (2021)所指出的，这是一个具有挑战性的任务。例如，人类可以代表从个体到人类群体的各种实体。Gabriel (2020)将意图分为几个类别，如指令(遵循用户的直接命令)、表达的意图(根据用户的潜在愿望行事)、揭示的偏好(反映用户的基于行为的偏好)等。

具体来说，我们用四个关键词来描述对齐的目标：鲁棒性，可解释性，可控性，和道德性（**RICE**）。以下是对四个原则的详细解释。

- **鲁棒性**指人工智能系统在面对多样化场景(Dietterich, 2017)或对抗压力(Rudner and Toner, 2021b)时的抵抗力，特别是保证其目标的正确性以及能力泛化性。鲁棒的人工智能系统能够应对黑天鹅事件(Taleb, 2007)和长尾风险(Hendrycks et al., 2021c)，以及各种对抗压

力(Song et al., 2018; Chakraborty et al., 2021)。例如，一个未完全对齐的大语言模型可以拒绝执行有害的请求，但用户可以通过越狱提示和其他对抗攻击使得模型被迫执行有害的行为(Zou et al., 2023)。而一个能够抵抗对抗攻击的模型在面对诱发系统失败的输入时仍能按照预期行事。随着人工智能系统在军事和经济等高风险领域的应用越来越广泛(Steinhardt and Toner, 2020)，我们更要确保它能抵御意外中断和对抗攻击，因为即使是瞬间的失败也可能带来灾难性的后果(Kirilenko et al., 2017; OecdAI, 2021; Rudner and Toner, 2021b)。一个对齐的系统应在其生命周期内始终保持鲁棒性(Russell, 2019)。

- **可解释性**要求人类能理解人工智能系统的内在推理过程，特别是黑盒神经网络的内部工作原理(Räuker et al., 2023)。直接的对齐评估方法，如行为评估，可能会受到人工智能系统不诚实行为的干扰(Turpin et al., 2023; Park et al., 2023b; Jacob Steinhardt, 2023)或欺骗性对齐(Carranza et al., 2023)的影响。解决这些问题的一种方法是在构建系统的过程中设计必要机制使人工智能系统诚实、不隐藏、不操纵(Carroll et al., 2023b)。或者，我们可以构建可解释性工具，深入了解神经网络内部的概念和推理机制(Elhage et al., 2021; Meng et al., 2022)。除了使安全评估成为可能，可解释性还使决策过程对于用户和利益相关者透明和易于理解，从而实现人类的有效监督。随着人工智能系统在现实世界的决策过程和高风险环境中扮演越来越重要的角色(Holzinger et al., 2017)，揭示决策过程而不是让它保持作为一个不透明的黑盒系统变得至关重要(DeepMind, 2018; Rudner and Toner, 2021a)。

- **可控性**是一种必要的属性，它确保系统的行动和决策过程始终受到人类监督和约束。它保证人类可以及时纠正系统行为中的任何偏差或错误(Soares et al., 2015; Hadfield-Menell et al., 2016a)。随着人工智能技术的日益发展，越来越多的研究表达了对这些强大系统的可控性的关注和担忧(ARC Evals, 2023)。当一个人工智能系统开始追求与其人类设计者相矛盾的目标时，它可能表现出一些具有重大风险的能力，包括欺骗、操纵用户和权力寻求的行为(Shevlane et al., 2023; ARC Evals, 2023)。可控性的目标主要集中在如何在训练过程中实现可扩展的人类监督(Bowman et al., 2022)，以及人工智能系统的可纠正性 (即在部署过程中不抵制关闭或目标修改)(Soares et al., 2015)。可控性原则引发的一个关键的问题是，随着人工智能系统愈发强大，人类如何更好的指导人工智能系统甚至是超级人工智能系统的训练和运行？这也被称为超对齐问题(Superalignment)。

- **道德性**指人工智能系统在决策和行动中坚定不移地维护人类的规范和价值观。在这里，规范和价值观包括道德指南和其他社会规范/价值观。它确保系统避免采取违反道德规范或社会公约的行为，例如对特定群体展示偏见(Kearns and Roth, 2019; Berk et al., 2021)，对个人造成伤害(Hendrycks et al., 2021b)，以及缺乏多样性或公平性(Collective Intelligence, 2023)。有大量的研究致力于为人工智能系统开发道德框架(Pankowska, 2020)。将道德性原则加入系统对于它们融入社会至关重要(Winfield et al., 2019)。

1.3 对齐循环

在这一章节，我们专注于阐述人工智能对齐的范围：我们将对齐过程构建为一个对齐循环，并将其分解为前向对齐过程和后向对齐过程[1](§1.3)。前向对齐过程关注于基于已有的对齐需求构建对齐的系统，而后向对齐关注于验证已对齐的系统在部署过程中的实际对齐度，并根据实际需要或社会道德的变化更新对齐需求。需要注意的是，前向对齐和后向对齐并不是割裂的，后向对齐也可能会出现在前向对齐的过程中，二者相辅相成贯穿于人工智能系统发展的整个生命周期，特别地，我们会更近一步讨论人类价值观在人工智能对齐中的地位(§5.1)，并进一步分析对齐范围外的AI安全问题(§5.2)。

本文将人工智能对齐分解为**前向对齐** (对齐训练) (§2, §3) 和**后向对齐** (对齐精炼) (§4, §5)。前向对齐旨在将一个训练系统初步对齐基本要求。本文将这项任务分解为从反馈中学习(§2) 和在分布偏移下学习(§3)。后向对齐旨在通过在简单和现实环境中进行评估，并设置监管条例来处理现实世界的复杂性，即对齐保证(§4)，确保训练系统的实际对齐。它还包括创建和执行确保人工智能系统安全开发和部署的规则，即人工智能治理(§5)。同时，后向对齐根据系统的对齐程度评估和监控(部署前和部署后)并更新对齐要求,并应用于下一轮的前向对齐训练中。

[1]在接下来的描述中，为了方便，我们简化称为前向对齐和后向对齐。

这两个阶段，前向对齐和后向对齐，形成了一个循环，每个阶段都会产生或更新下一阶段的输入。这个循环，我们称之为对齐循环，将重复进行以产生越来越对齐的人工智能系统。我们将人工智能对齐视为一个动态过程，在这个过程中，所有的标准和实践都应该被持续评估和更新。值得注意的是，后向对齐(包括人工智能系统的对齐保证和对齐治理)的努力在整个对齐循环中都在进行，而不仅仅是在训练之后。如Koessler (2023) 所论述，对齐和风险评估应该在系统生命周期的每个阶段进行，包括在训练前后阶段和部署阶段。同样，对系统生命周期的每个阶段的监管措施也已经被广泛讨论(Schuett et al., 2023; Anderljung et al., 2023)。

本文围绕四个核心支柱进行结构化：从反馈中学习（§2）和在分布偏移下学习（§3），这两者构成了前向对齐；以及对齐保证（§4）和人工智能治理（§5），这两者构成了后向对齐。接下来本文将对每个支柱进行介绍，阐明其如何协同构建一个全面的人工智能对齐框架。

2 从反馈中学习

从反馈中学习涉及的问题是在对齐训练过程中，我们如何提供并利用反馈来指导训练中的人工智能系统的行为？在大语言模型的应用中，一个典型的解决方案是利用基于人类反馈的强化学习(RLHF)(Christiano et al., 2017)，其中人类评估者通过比较来自语言模型的不同答案来提供反馈，然后通过强化学习(RL)对训练好的奖励模型使用这些反馈。一些研究发现，经过RLHF训练的LLM (Ouyang et al., 2022) 比通过单纯使用监督学习方法训练的模型(Devlin et al., 2019; Brown et al., 2020) 更具创造性和对齐性。RLHF的重要性不仅仅限于让LLM遵循人类的指示(Ouyang et al., 2022)。它通过偏好训练赋予LLM重要的道德品质，如有用、无害和诚实，使LLM更好地对齐(Bai et al., 2022a)。这些优点使得RLHF被广泛用来对齐LLM (Ziegler et al., 2019; OpenAI, 2023a; Touvron et al., 2023)。特别地，Dai (2023)提出基于人类反馈的安全强化学习（Safe RLHF），用于平衡大语言模型在对齐训练中帮助性和无害性目标之间的内在矛盾。通过将大模型中的安全性形式化为一种在训练中需要满足的约束，并用约束马尔科夫决策(CMDP)来形式化整个任务，Safe RLHF在满足输出符合人类价值中的安全性的前提下(约束)，尽可能的提高了帮助性(目标)。未来的努力可以集中在减少对人类标注的依赖(Sun et al., 2023) 和通过利用迭代RLHF方法(即将其与辩论框架集成(Irving et al., 2018))等，提高奖励模型的有效性。

本文根据Ouyang (2022)的研究将RLHF的流程总结为如下三个阶段：

- **监督微调(SFT)**。RLHF通常从一个预训练的语言模型开始，然后使用监督学习——特别是最大似然估计——在为下游任务量身定制的高质量数据集上进行微调，以获得模型π^{SFT}。这些任务包括对话处理、指令跟随和总结。

- **收集比较数据和奖励建模**。这个阶段包括收集比较数据，然后用它来训练一个奖励模型。SFT模型被给予提示x，生成来自$\pi^{SFT}(y \mid x)$ 的响应对(y_1, y_2)。然后，这些响应对被展示给人类标注者并标注得到偏好数据。偏好数据被用来构建奖励模型r_θ。

- **通过强化学习进行策略优化**。最后一步是基于奖励模型r_θ提供的奖励，使用RL方法对LLM的策略π进行优化。大语言模型从提示生成响应的过程被建模为一个bandit环境 (Ouyang et al., 2022)，在每个响应结束时从奖励模型r_θ获得奖励。RL的主要目标是调整大语言模型的参数ϕ，使得在训练提示数据集D_{RL}上的期望奖励最大化：

$$\arg\max_{\pi_\phi} \; E_{x\sim D_{\mathrm{RL}}, y\sim\pi_\phi}\left[r_\theta\left(x, y\right)\right].$$

通常，会引入来自SFT模型π^{SFT}的额外对每个token的KL惩罚，以缓解奖励过度优化的问题。此外，引入从预训练的数据分布$\mathcal{D}_{\mathrm{pretrain}}$当中产生的梯度有助于保持模型性能，这在Ouyang (2022)中被称为PTX损失。因此，可以引入一个更全面的目标函数：

$$J(\phi) = E_{x\sim D_{\mathrm{RL}}, y\sim\pi_\phi}\Big[r_\theta(x,y) - \beta\log\left(\pi_\phi(y|x)/\pi^{\mathrm{SFT}}(y|x)\right)\Big] + \eta\, E_{(x,y)\sim D_{\mathrm{pretrain}}}\Big[\log\left(\pi_\phi(y|x)\right)\Big]$$

其中β和η是决定KL惩罚强度和预训练梯度混合的系数。这个过程使大语言模型生成的响应更好地与在训练过程中用于提示的人类偏好相符。

尽管RLHF很受欢迎，但它面临着许多挑战(Tien et al., 2023)。其中一个突出挑战是可扩展监督，即如何对在复杂场景中运行的超级人工智能系统提供高质量的反馈，这些情况往往超出了人类评估者的知识或理解范围，使得人工智能系统的行为难以被人类评估(Bowman et al., 2022)。具体来说，可扩展监督的出现源自两个实际原因。首先是人类频繁评估人工智能系统行为的高成本。例如，训练过程非常耗时，将人类直接纳入实时的训练循环会大大浪费人力资源并降低训练效率(Christiano et al., 2017)。其次是人工智能系统行为的固有复杂性使得评估变得困难，尤其是在难以理解和高风险的任务上(Saunders et al., 2022)，例如，教人工智能系统总结书籍(Wu et al., 2021)，生成复杂的代码片段(Pearce et al., 2022)，和预测未来的天气变化(Bi et al., 2023)等任务。

可扩展监督旨在确保人工智能系统即使在超越了人类的专业知识的情况下，仍然与人类的意图保持一致。在此，本文的主要关注于提出一些可能尚未普遍实施的构建可扩展监督的前景方向(Leike et al., 2018)。

从人工智能反馈中进行强化学习(RLAIF) RLAIF是对RLHF的进一步扩展，通过使用大语言模型生成的反馈替代人类反馈，提高模型整体效用。该方法通过自我评估、修订、微调和评估人工智能反馈，构建一个可用于训练的奖励模型(Bai et al., 2022b)。与RLHF相比，RLAIF通过人工智能反馈实现无害性，降低训练成本，且在摘要任务上与人类反馈相媲美(Bowman et al., 2022)。

从人类和人工智能反馈中进行强化学习(RLHAIF) RLHAIF整合了人类和人工智能元素，将任务分解为子任务，形成树状结构，以便人类监督和评估(Wu et al., 2021)。同时，通过人工智能模型生成的批评有助于人类发现模型可能忽视的缺陷(Saunders et al., 2022)。这种混合方法展示了在复杂问题和多领域监督中，使用人工智能协助的可行性。

递归奖励建模(RRM) 奖励建模允许我们将系统目标的构建与行为评估分离(Ibarz et al., 2018)。以这种方式，奖励建模为人工智能系统的优化方向提供了指导，能够精细地使系统与人类的意图和价值观对齐，例如对语言模型进行微调以遵循人类指令(Bai et al., 2022a; Touvron et al., 2023)。递归奖励建模(Recursive Reward Modleing, RRM)(Leike et al., 2018; Hubinger, 2020)旨在将奖励建模的应用扩展到更复杂的任务。智能体被训练以最大化通过对其扩增版本进行奖励学习所获得的奖励。这种方法不仅受到人类反馈的影响，而且受到模型自身对于奖励构成的评估影响。RRM的核心思想是递归使用训练得到的智能体A_{t-1} 来为训练更复杂任务的智能体A_t 提供反馈。A_0 通过基本的奖励模型(从纯人类反馈中学习)进行训练。基于评估答案比回答问题更容易的假设，奖励模型可以迭代地达到更高的能力从而能够监督更强大的人工智能系统。

辩论(Debate) 辩论的基本过程是两个智能体轮流提供答案和陈述，而人类裁判根据辩论过程进行最终结果的评判(Irving et al., 2018)。作为零和辩论游戏，智能体在辩论过程中试图识别对方的缺点，同时努力获得人类裁判的更高信任，这是构建可扩展监督的潜在方法。例如，在围棋游戏中，人类裁判可能无法从单个局面辨别优势方。然而，通过观察游戏的过程和最终结果，裁判可以更容易地推断出“谁相对更具优势”。这种方法的前提依赖于一个关键的假设：为真理辩护通常比为虚假辩护更容易，这意味着说真话的辩论者更具有优势。随着大语言模型能力的提升，已有相关工作在大语言模型上应用辩论来提升模型能力(Du et al., 2023; Claude, 2023)。然而，在特定的开放现实世界场景中，辩论可能会面临巨大挑战(Irving et al., 2018)。例如，某些问题可能过于复杂、无法被人类理解，或者问题背景过于庞大、无法完全呈现，比如解释一个1亿像素的图像或者整个互联网的信息。同样，有些情况下，一个问题的最佳答案可能非常冗长，比如一个需要跨越一百页的回答。为了处理这些问题，智能体可能会首先选择一个回答，然后随着辩论的进行，揭示问题或答案的部分内容(Irving et al., 2018)。

合作逆强化学习(CIRL) 许多对齐失败的模式，例如奖励破解(Victoria et al., 2020; Skalse et al., 2022)，欺骗(Park et al., 2023b)，和操纵(Carroll et al., 2023b)，都是AI系统对错误规范的目标进行“自信地”优化的结果(Pan et al., 2022)。在训练和部署过程中，指定的目标(如奖励函数)对AI系统来说起着一个无可挑战的真理的角色，人类的反馈一定程度上只有在目标位置上才被尊重，这意味着它可以被篡改(Everitt et al., 2021) 或者被操纵(Carroll et al., 2023b)。合作逆强化学习(Hadfield-Menell et al., 2016b) 试图通过以下方式来解决上述问题(1)让AI系统明确

地对其奖励函数保持不确定性；(2)让人类提供关于真实奖励函数是什么的唯一信息。不确定性使AI系统倾向于听从人类的意见并驱使它去确定人类真正想要什么。具体来说，它将整个任务模型化为一个包含两个玩家的合作博弈，其中人类玩家 H 和智能体玩家 R 共享一个公共的奖励函数$r(\cdot)$。更重要的是，奖励函数和奖励信号对R 来说是不可见的(实际上并没有被训练机制明确地计算出来)，只能通过一个类似于IRL的过程从H 的行为中R 推断出来(包括通过询问和与H 交互)。这一设定被称为*CIRL* (Hadfield-Menell et al., 2016b)，协助博弈 (Fickinger et al., 2020)，和协助*POMDP* (Shah et al., 2020)。简单来说，AI系统将人类的真实目标$r(\cdot)$ 作为自己的目标(尽管$r(\cdot)$ 的值并不确定)，并通过观察和与人类交互来不断尝试弄清楚r。这可能消除了例如操纵这类行为的动力，因为操纵人类的行为只会污染一个信息源，而不会影响r。

3 在分布偏移下学习

与固定输入分布的反馈学习过程形成对比，此部分更关注对齐过程中输入分布发生变化的情况，即分布偏移(Krueger et al., 2020; Hendrycks et al., 2021a)。更具体地，它关注在分布偏移下对齐特性(即遵循人类意图和价值观)的保持，而不是模型能力的保持。与分布偏移相关的一个挑战是目标错误泛化，在这种情况下，人工智能系统在训练分布下的预期目标(例如，遵循人类的真实意图)与其他未对齐的目标(例如，不择手段地获取人类的认可）无法区分。系统往往实际上针对后者学习优化，这导致人工智能系统在部署分布中出现未对齐的行为(Di Langosco et al., 2022)。另一个相关的挑战是自诱发分布偏移(Auto-induced Distribution Shift, ADS)，在这种情况下，人工智能系统能够改变其输入分布以最大化奖励(Krueger et al., 2020; Perdomo et al., 2020)。一个例子是推荐系统能够反向塑造用户偏好使得算法便于优化(Adomavicius et al., 2022)。目标错误泛化和自诱发分布偏移都可能导致或加剧人工智能系统的欺骗行为(Park et al., 2023b) 和操纵行为(Carroll et al., 2023b) 。应对分布偏移的方法包括算法干预 ，它通过在训练过程中改变风险范围以提高人工智能系统在其他分布下的可靠性，以及数据分布干预 ，它扩大训练分布(或融合多分布) 以减小训练和部署分布之间的差异。前者包括像风险外推（REx）(Krueger et al., 2021) 和基于连通性的微调（CBFT）(Lubana et al., 2023) 等方法。后者包括对抗训练(Bai et al., 2021)，它用对抗性输入增强训练输入分布，以及合作训练(Dafoe et al., 2020)，其目标是解决从单智能体到多智能体环境的分布偏移问题。

4 对齐保证

即使人工智能系统经过了前向对齐，我们在实际部署它之前还需要考察其对齐度的置信值(Anderljung et al., 2023)。这就是对齐保证：在人工智能系统实际训练和部署后对其实际对齐情况进行测量和评估。对齐保证的方法包括安全评估(Perez et al., 2023) 和更高级的方法，如红队测试(Perez et al., 2022)和可解释性技术(Olah et al., 2018)：

可解释性 可解释性是一个使机器学习系统及其决策过程对人类可理解的研究领域(Doshi-Velez and Kim, 2017)。可解释性研究构建了一个工具箱，用来更好地描述或预测模型的新特性。在本文中，我们更关注的是与对齐和安全性最相关的研究，并且从经验上看，这些技术通过研究神经网络的内部结构和表示使神经网络更安全(Räuker et al., 2023)。

红队测试 红队测试是指制造特定语境，使人工智能系统被诱导产生不符合预期的输出或行动(如危险的行为如欺骗或权力寻求，以及其他问题如有毒或有偏的输出)，并在这些场景下测试系统。其目标是通过施加对抗压力，即特意试图使系统失败，来评估系统对齐的稳健性。一般来说，最先进的系统——包括语言模型和视觉模型——不能通过这个测试(Chakraborty et al., 2021; Perez et al., 2022; Liu et al., 2023b; Chen et al., 2023)。红队测试的动机有两个：(1)获得对训练系统对齐的保证;(2)在对抗训练中提供对抗输入的来源(Yoo and Qi, 2021; Bai et al., 2021)。我们更加关注第一个，值得注意的是，这两个目标是不可分割的；针对第一个动机的工作也有助于为第二个目标提供基础。

对齐保证的范围也包括验证系统与人类价值观的对齐度，包括旨在可证明合作性(Dafoe et al., 2021) 和伦理性(Tolmeijer et al., 2020) 的形式化理论，以及广泛的经验和实证方法。在这里我们介绍两种用于验证人类价值契合性的方法：

场景模拟 场景模拟是一种比数据集更复杂的形式，因此有些观点认为它(Hendrycks et al., 2021d)在反映真实情况和获得更好结果方面更有效。场景的形式也可以有所不

同。Pan (2023)通过文本冒险游戏构建了一系列多样化的，具有道德意义的场景，评估了欺骗，操纵，背叛等复杂行为。另一方面，一些工作试图通过模拟人机交互使智能代理学习人类价值。Yuan (2022)提出了一种人机双向价值对齐的方法，通过人类反馈使机器学习人类的偏好和隐含目标。Liu (2023a)将人工智能置于模拟的人类社会沙盒中，通过模仿人类的社交互动，让人工智能学习人类社会价值倾向。

价值评估方法 现有的评估模型在价值方面展现出了非常多样化的方法。Durmus (2023)从全球五个不同的文化中收集了关于人类价值观的数据。为了评估LLM的价值取向，他们比较了LLM产生的回应与这些不同人类群体得到的回应之间的相似性。研究结果表明，LLM仍然表现出明显的价值偏见。同时，Zhang (2023)使用社会价值取向的框架(Messick and McClintock, 1968; Van Lange et al., 1997)研究了LLM在各种价值观上的合理性。他们的发现表明，LLM更倾向于选择反映中性价值观的行动，如亲社会。

对齐保证在人工智能系统的生命周期中都会进行，包括在训练前、训练中、训练后和部署后，而不仅仅是在训练后(Shevlane et al., 2023; Koessler and Schuett, 2023)。值得注意的是，许多对齐保证的技术在训练过程中也是适用的，例如，红队测试是对抗性训练的关键组成部分，可解释性可以帮助提供反馈(Burns et al., 2023)。

5 人工智能治理

单靠对齐保证难以确保人工智能系统在部署环境中始终保持对齐性，因为它没有考虑到现实世界的复杂性。这就需要对人工智能系统进行必要的治理监管，重点关注它们的对齐性和安全性，并覆盖系统的整个生命周期。在这里，本文主要讨论人工智能治理的多方利益相关者的方法，包括政府规定(Anderljung et al., 2023)，实验室自主治理(Schuett et al., 2023)，以及第三方组织(Shevlane et al., 2023; Koessler and Schuett, 2023)。总的来说，**政府机构**运用立法、司法和执法权力，制定人工智能发展政策并参与国际合作。**行业和AGI实验室**研究和部署人工智能技术，是被监管的主体，同时又提出方法进行自我监督并影响治理体系架构。**第三方**包括学术界、非政府组织(NGOs)和非营利组织(NPOs)，不仅对企业治理、人工智能系统应用提供审计，而且协助政府制定政策。特别地，本文建立在跨代际视角上从两个方面分析国际人工智能治理合作的重要性和可行性：管理全球灾难性人工智能风险和管理人工智能中的机遇，从而为国际人工智能治理的未来结构贡献创新思考。

管理全球性的人工智能灾难风险 市场上无节制的竞争和地缘政治因素可能导致先进人工智能系统过快的开发和部署，带来潜在的负面全球影响(Tallberg et al., 2023)。人工智能系统中根深蒂固的种族和性别偏见可能会被放大并导致代际性的道德歧视(Swaugerarchive, 2020)。国际治理合作的干预可以缓解这些全球性的挑战，例如国家之间的共识可以帮助避免潜在的人工智能军备竞赛，而而全行业的协议可以防止草率和不负责任的开发人工智能系统，从而保障人工智能长期和可持续的发展(Ho et al., 2023)。

管理人工智能机遇 人工智能发展带来的机遇并没有平等地分布，这可能导致不同地区之间持久的数字不平等，并危及人工智能发展的可持续性。人工智能发展中的地理差异将加剧经济和社会效益的不公平分配(Ho et al., 2023)。此外，技术领域决策权掌握在少数个体中可能会导致权力分配的不平等，从而形成代际性垄断(Noble et al., 2021)。通过人工智能的传播、教育和基础设施发展(Opp, 2023) 促成的人工智能机遇的国际共识和协调行动，可以确保从人工智能带来的技术红利平衡分配，并促进其持续发展的可持续性。

人工智能治理领域另一个争论焦点是人工智能模型是否应开源(Seger et al., 2023)。对于开源模型是否会提高模型安全性还是增加滥用风险仍然存在争论。正如Shapiro (2010)所指出的，透明度的有效性取决于潜在攻击者已经拥有的知识的可能性，以及政府将透明度转化为识别和解决新出现的漏洞的能力。如果无法在人工智能系统的攻防之间建立适当的平衡，开源可能会潜在地带来人工智能系统滥用的重大风险。为了准确和清晰，本文遵循Seger (2023)中对开源模型的定义：允许公开和公共访问模型的架构和权重，并允许任何人进行修改、研究、进一步开发和利用。目前，最为公认的开源模型包括Llama2 (Touvron et al., 2023)、Falcon (Penedo et al., 2023)、Vicuna (Chiang et al., 2023)等。本节主要评估开源模型的安全优势和潜在威胁，以促进关于开源的可行性和具体方法的讨论。

支持开源的观点 支持开源的研究人员和政策制定者认为开源可以通过多种途径减轻模型中固有的安全风险：(1) 开源可以促进开发者和社区对模型的测试,进而快速识别和解决模型可能具有的问题，并增强对模型相关风险的认知，促进对这些潜在风险更多的关注和研究(Zellers, 2019)。(2) 开源被认为是促进权力和控制分散化的有效策略。一个例子是Stability公开Stable Diffusion的核心原因：他们将信任寄托于个人和社区，而不是由中央集中控制、未经选举的实体控制人工智能技术(Mostaque, 2022)。一些评论家将模型开源与启蒙时代相提并论，认为分散化的控制增强了对人类和社会力量和善意的信任(Howard, 2023)，出于安全目的实施集中治理可能反而会增强人工智能技术社区的权力。

反对开源的论点 开源模型的批评者从多个角度评估了开源模型可能被滥用的风险，提出了反对意见：(1) 开源模型可能被微调为有害模型。一些人工智能系统与其最初的设计意图——减轻化学或生物学中的毒性——相反，现在有可能制造新的化学毒素(Urbina et al., 2022)和生物武器(Sandbrink, 2023)。这种模型的恶意微调可能导致深远的安全风险。此外，一旦进行了精细调整，语言模型可以模拟熟练的写手，产生令人信服的虚假信息，这可能导致相当大的社会政治风险。(Goldstein et al., 2023)。(2) 无意中鼓励系统越狱。研究表明，对开源模型权重的无限制访问可能促使绕过系统安全措施的行为(Seger et al., 2023)。(Zou et al., 2023)通过使用Vicuna-7B和13B(Chiang et al., 2023)实现了开发攻击后缀。一旦这些后缀在像ChatGPT(OpenAI, 2023a)，Bard(Google, 2023)和Claude(Anthropic, 2023)这样的易于访问的接口中实施，将产生违反人类意图的生成结果。

关于人工智能模型的开源问题的争论仍未产生共识，目前主流的观点是，人工智能模型的公开并不会在目前带来重大风险，但仍需要做好必要准备。例如现有的关于开源先进人工智能系统的指导方针包括通过量化微调滥用的可能性来评估风险，以及逐步发布模型(Seger et al., 2023)等措施。同时，政策制定者正在为这些开源模型建立严格的合规协议。

5.1 对齐中的人类价值观

我们在RICE原则中包含道德性，这体现了人类价值观在人工智能对齐中的关键作用。人工智能系统不仅应与价值中立的人类偏好(如人工智能系统执行任务的意图)相一致，还应与道德和伦理考虑相一致，也就是价值对齐(Gabriel and Ghazavi, 2021)。人类价值观的考虑因素被嵌入到对齐循环的所有部分–实际上，我们调查的所有四个部分都有专门针对人类价值观对齐的研究主题。因此，为了提供这些研究主题的更全面的画像，我们在深入讨论每个单独部分的详细信息之前，先对它们进行概述。

本文将关于人类价值观的一致性研究分类为三个主要主题：(1)伦理和社会价值观，旨在教导人工智能系统区分对错; (2) 合作型AI，旨在特别培养人工智能系统的合作行为;以及(3)处理社会复杂性，为多智能体和社会动态的建模提供基础。

伦理和社会价值观 人类价值观本质上具有极强的抽象性和不确定性。Macintyre (2013) 更是指出现代社会缺乏统一的价值标准，不同文化的人类之间的价值差异可能非常大。这使得我们究竟要对齐何种人类价值成为了一个重要挑战。虽然在所有人中完全一致的价值观不一定存在，但仍然有一些价值在不同的文化中都得到了体现。在以下的部分中，我们将分别从机器伦理，公平性和社会心理学中的跨文化价值观的角度讨论这些问题。

- **机器伦理** 与大部分将人工智能系统与人类的一般偏好(包括全面价值和中立价值)相对齐的对齐研究相比，机器伦理学专注于将适当的道德价值观灌输到人工智能系统中(Yu et al., 2018)。这一类工作最早涵盖了符号和统计学习系统(Anderson et al., 2005; Anderson and Anderson, 2007)，后来扩展到包括建立大型道德伦理数据集(Pan et al., 2023)和基于深度学习的方法(Jiang et al., 2021; Jin et al., 2022)。

- **公平性** 尽管存在争议(Verma and Rubin, 2018)，但公平性的定义相对于其他人类价值观来说比较清晰。它是指个人或群体先天或后天获得的偏见、偏爱特性的缺失(Mehrabi et al., 2021)。关于人工智能公平性的研究非常广泛，这些方法涵盖从在训练前减少数据偏见出发(d'Alessandro et al., 2017; Bellamy et al., 2018)，最小化在训练过程中引入的不公平性(Berk et al., 2017)，以及处理训练阶段未成功学习到的不公平样例(Xu et al., 2018)。

- **社会心理学中的跨文化价值观** 在社会心理学领域，许多研究专注于探索跨文化人类社区中存在的价值观群簇，从而发展出各种跨文化价值观量表。奥尔波特-弗农-林赛的价值观系统(Allport, 1955)提出，理解个人的哲学价值观构成了评估其价值系统的关键基础。他们设计了一个包含六种主要价值类型的价值观量表，每种类型代表人们对生活各个方面的偏好和关注。Van (1997)引入并改进了一种可量化的方法，即社会价值取向(SVO)，用于评估个人的社会价值观倾向。它使用定量方法评估个人如何分配给自己和他人的利益，进而评估其中反映的社会价值观取向，如利他主义，个人主义等。Murphy (2014)引入了滑块测量方法，可以从连续的角度入手，根据受试者对一些特定问题的选择精确评估相应的SVO。Rokeach (1973)开发了一个包含36个价值观的价值观清单，其中包含18个代表期望目标的终端价值观和18个代表实现这些目标的手段的工具价值观。Schwartz (1992; 1994)在20个不同的国家进行了全面的问卷调查，即施瓦茨价值观调查。这项研究确定了无论文化、语言或地点如何，都被普遍认可的十个价值观。这些研究都为确定人工智能应与何种价值观对齐奠定了坚实的理论基础。

合作型人工智能 多智能体交互中最关键的方面是合作，而合作失败则是多智能体交互中最令人担忧的方面。作为人工智能合作失败的一个例子，*2010*年的闪电崩盘导致市场价值在2分钟内损失了数万亿，这其中部分原因是由高频算法交易者之间的交互引起的(Kirilenko et al., 2017)。因此，有必要在类似智能体的人工智能系统和他们所操作的环境中设计确保合作的机制(Dafoe et al., 2021)。这种机制的高级设计原则和低级实现属于合作型人工智能的领域(Dafoe et al., 2020)。此外，合作型人工智能还通过人工智能的视角研究人类的合作，以及人工智能如何帮助人类实现合作。更准确地说，Dafoe (2020)将合作型人工智能研究分类为四个广泛的主题：理解、沟通、承诺和制度，涵盖了从博弈论到机器学习再到社会科学等各种学科。

解决社会复杂性 道德性的要求本身就包含了社会成分。“什么是道德的？”通常在社会环境中定义，因此，道德性在人工智能系统中的实现也需要考虑社会复杂性。Critch (2020)为这个领域提出了许多研究主题的建议。其中一个研究方向侧重于社会系统的真实模拟，包括基于规则的智能体建模(Bonabeau, 2002; De Marchi and Page, 2014)，基于深度学习的模拟(Storchan et al., 2021)，以及那些包含大语言模型的模拟(Park et al., 2023a)。这些模拟方法可以服务于各种下游应用，从影响评估(Osoba et al., 2020)到多智能体社会学习(Critch and Krueger, 2020)。在另一方面，社会选择(Sen, 1986; Arrow, 2012)领域以及相关的计算社会选择(Brandt et al., 2016)领域旨在为多样化人口中的偏好聚合等目标提供数学和计算解决方案。有人认为，当与基于人类偏好的对齐方法(例如，RLHF和在§2中介绍的大多数其他方法)结合时，社会选择的方法可以作为已有方法的补充，以保证表征出的公平性能够代表每个人的偏好(Leike, 2023; Collective Intelligence, 2023)。一部分研究对这个提议已经进行了早期阶段的实验(Yamagata et al., 2021; Köpf et al., 2023)。为了进一步扩展这种从人群中学习价值的方法，还有人认为，人工智能系统中的体现价值应在长期内持续进步，而不是被永久锁定(Kenward and Sinclair, 2021)，以便应对新出现的挑战，以及变得未来可证，并满足道德领域的潜在未知现象。

5.2 对齐外的人工智能安全性

在介绍了对齐的内在范围之后，在本节我们进一步讨论对齐之外的人工智能安全性。人工智能系统除了对齐失败之外还存在许多风险：恶意行为者可能故意使用人工智能造成伤害，如制造生物武器。与此同时，人工智能开发者之间的竞争可能导致他们忽视风险，急于部署安全性有待确认的人工智能系统。虽然这篇综述文章主要关注对齐，但我们借鉴了Hendrycks (2023)，对其他可能导致灾难性人工智能风险的原因进行了简要概述，从而扩展人工智能对齐的讨论范围。

恶意使用 恶意行为者可以故意使用人工智能造成伤害。目前已经有犯罪分子利用深度伪造技术进行诈骗和敲诈(Cao and Baptista, 2023)。随着未来人工智能系统可能发展出更为强大的能力，滥用的威胁变得更大。一个关于人工智能系统可能被恶意用于造成伤害的例子是生物武器。研究已经表明，大语言模型可以提供步骤详尽的关于合成具有大规模流行能力的病原体的说明指南(Soice et al., 2023)。除了传播如何制造生物武器的信息之外，人工智能系统还可以帮助设计出比现有疾病更致命和更易传播的新病原体(Sandbrink, 2023)。像奥姆真理教(Danzig et al., 2012)这样的恐怖组织已经试图制造生物武器以造成大规模的破坏，人工智能系统可能使小

团体更容易制造生物武器并引发全球大流行。其他种类的恶意使用可能包括使用人工智能系统对关键基础设施发动网络攻击(Mirsky et al., 2023)，或者创建能在人类控制之外生存和传播的智能体(Bengio, 2023)。随着人工智能系统的能力不断变强，相应的风险也不断加大，需要进行彻底的评估，以确定人工智能系统可能如何被用来造成伤害。恶意使用不应被视为对齐失败，因为当一个人工智能系统按照恶意用户的意图行事时，这个系统将与其用户对齐，虽然结果是对社会构成严重威胁。确保人工智能符合公众利益的政策将是避免这种威胁的关键。

集体行动问题 人工智能开发者正在竞相开发和部署强大的人工智能系统(Grant and Weise, 2023)。这种竞争氛围使得开发者忽视安全性，而急于部署他们的人工智能系统。即使有一个开发者想要谨慎小心地开发人工智能系统，他们可能也会存在担忧：放慢速度，彻底评估他们的系统，并投资新的安全特性，可能会让他们的竞争对手超过他们(Armstrong et al., 2013)。这形成了一个社会困境，即个别的人工智能开发者和机构追求自己的利益，可能会导致所有人的结果不理想。人工智能系统之间的竞争成功可能受到进化动力学的制约，即最强大和最自私的人工智能系统最有可能生存(Hendrycks, 2023)。防止这些集体行动问题导致社会灾难，需要国家和国际人工智能政策的干预，以确保所有人工智能开发者都遵守共同的安全标准。

6 结论

本文对人工智能对齐进行了全面的介绍，人工智能对齐的目标是构建行为符合人类意图和价值观的人工智能系统。本文将对齐的目标归纳为鲁棒性、可解释性、可控性和道德性(**RICE**)，并将对齐方法的范围划分为前向对齐(通过对齐训练使人工智能系统对齐)和后向对齐(获取人工智能系统对齐的证据，并适当地对其进行管理，以避免加剧对齐风险)。目前，前向对齐的两个显著研究领域是从反馈中学习和在分布偏移下学习，而后向对齐则包括对齐保证和人工智能治理。

与许多其他领域相比，人工智能对齐的一个特点是其多样性(Hendrycks, 2022) – 它是多个研究方向和方法的紧密组合，通过共享的目标而非共享的方法论将其联系在一起。这种多样性带来了好处。它允许不同的研究方向互相补充，共同服务于对齐的目标；这体现在对齐循环，其中四个支柱被整合成一个自我改进的循环，不断提高人工智能系统的对齐性。

同时，这种研究方向的多样性提高了进入这个领域的门槛，这就需要编制组织良好的调查材料，既服务于新人，也服务于有经验的研究人员。在这篇综述中，本文试图通过提供全面和最新的对齐概述来解决这个需求。本文的对齐综述几乎关注了这个领域的所有主要研究议程，以及对齐保证和人工智能治理方面的实际实践。本文通过展望未来并展示我们认为的人工智能对齐领域未来需要解决的关键问题来结束这篇综述。

强调政策相关性 对齐研究并不是在真空中进行，而是在一个生态系统中进行(Drexler, 2019)，研究人员、行业参与者、政府和非政府组织都应参与其中。因此，服务于人工智能对齐和安全生态系统需求的研究将是有用的。这些需求包括解决各种治理方案的关键障碍，例如，极端风险评估(Shevlane et al., 2023)、计算治理的基础设施(Shavit, 2023)以及关于人工智能系统的可验证声明的机制(Brundage et al., 2020)。

强调社会复杂性和道德价值 随着人工智能系统越来越多地融入社会(Abbass, 2019)，对齐不再只是一个单一层次问题，而成为一个社会问题。这里，社会的含义有三层。首先，在多智能体环境中进行对齐研究，这涉及到多个人工智能系统和多个人之间的交互(Critch and Krueger, 2020)。其次，将人类的道德和社会价值纳入对齐，这与机器伦理学和价值对齐领域密切相关(Gabriel, 2020)。第三，建模和预测人工智能系统对社会的影响，这需要方法来处理社会系统的复杂性，包括社会科学中的那些问题。可能有用的方法包括社会模拟(Bonabeau, 2002; Park et al., 2023a) 和博弈论(Critch and Krueger, 2020)。

开放式探索新的挑战和方法 许多对齐讨论都建立在经典文献之上，这些文献早于最近的大语言模型和大规模深度学习的其他突破。因此，当这种范式转变发生在机器学习领域时，有一些对齐的挑战可能变得不那么突出，而其他的则变得更为突出；毕竟，科学理论的一个定义特征就是其可被证伪性(Popper, 1935)。更重要的是，这种机器学习方法的转变和人工智能系统越来越紧密地融入社会的更广泛趋势(Abbass, 2019)，引入了以前无法预见的新挑战。这就要求我们进行开放式探索，积极寻找以前被忽视的新挑战。

参考文献

Hussein A Abbass. 2019. Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2):159–171.

Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2022. Recommender systems, ground truth, and preference pollution. *AI Magazine*, 43(2):177–189.

Gordon Willard Allport. 1955. *Becoming: Basic considerations for a psychology of personality*, volume 20. Yale University Press.

Dario Amodei, Paul Christiano, and Alex Ray. 2017. Learning from human preferences. https://openai.com/research/learning-from-human-preferences.

Markus Anderljung, Joslyn Barnhart, Jade Leung, Anton Korinek, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, Miles Brundage, Justin Bullock, Duncan Cass-Beggs, et al. 2023. Frontier ai regulation: Managing emerging risks to public safety. *arXiv preprint arXiv:2307.03718*.

Michael Anderson and Susan Leigh Anderson. 2007. The status of machine ethics: a report from the aaai symposium. *Minds and Machines*, 17:1–10.

Michael Anderson, Susan Anderson, and Chris Armen. 2005. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 fall symposium on machine ethics*, pages 1–7.

Anthropic. 2023. Model card and evaluations for claude models. https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

ARC Evals. 2023. Update on ARC's recent eval efforts. https://evals.alignment.org/blog/2023-03-18-update-on-recent-evals/.

Stuart Armstrong, Nick Bostrom, and Carl Shulman. 2013. Racing to the precipice: a model of artificial intelligence development. Technical Report 2013-1, Future of Humanity Institute, Oxford University.

Kenneth J Arrow. 2012. *Social choice and individual values*, volume 12. Yale university press.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4312–4321. International Joint Conferences on Artificial Intelligence Organization, 8. Survey Track.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

Yoshua Bengio. 2023. How rogue ais may arise.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. 2023. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, pages 1–6.

Eric Bonabeau. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the national academy of sciences*, 99(suppl_3):7280–7287.

Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. 2022. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*.

Felix Brandt, Vincent Conitzer, Ulle Endriss, Jérôme Lang, and Ariel D Procaccia. 2016. *Handbook of computational social choice*. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Benjamin S Bucknall and Shiri Dori-Hacohen. 2022. Current and near-term ai as a potential existential risk factor. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 119–129.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

CAIS. 2023. Center for ai safety: Statement on ai risk. https://www.safe.ai/statement-on-ai-risk.

Ella Cao and Eduardo Baptista. 2023. 'deepfake' scam in china fans worries over ai-driven fraud. *Reuters*, 5.

Andres Carranza, Dhruv Pai, Rylan Schaeffer, Arnuv Tandon, and Sanmi Koyejo. 2023. Deceptive alignment monitoring.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023a. Characterizing Manipulation from AI Systems, March. arXiv:2303.09387 [cs].

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. 2023b. Characterizing manipulation from ai systems. *arXiv preprint arXiv:2303.09387*.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2023. Content-based unrestricted adversarial attack. *arXiv preprint arXiv:2305.10665*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Brian Christian. 2020. *The alignment problem: Machine learning and human values*. WW Norton & Company.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Ruby RobertM GPT-4 Claude. 2023. New lw feature debates.

Collective Intelligence. 2023. Introducing the collective intelligence project.

Andrew Critch and David Krueger. 2020. Ai research considerations for human existential safety (arches). *arXiv preprint arXiv:2006.04948*.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative ai. *arXiv preprint arXiv:2012.08630*.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. 2021. Cooperative ai: machines must learn to find common ground. *Nature*, 593(7857):33–36.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data*, 5(2):120–134.

Richard Danzig, Zachary Hosford, Marc Sageman, Terrance Leighton, Lloyd Hough, Hidemi Yuki, and Rui Kotani. 2012. Aum shinrikyo: Insights into how terrorists develop biological and chemical weapons, second edition. Report, Center for a New American Security (CNAS), 12.

Scott De Marchi and Scott E Page. 2014. Agent-based models. *Annual Review of political science*, 17:1–20.

DeepMind. 2018. Building safe artificial intelligence: specification, robustness, and assurance.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, pages 12004–12019. PMLR.

Thomas G Dietterich. 2017. Steps toward robust artificial intelligence. *Ai Magazine*, 38(3):3–24.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

K Eric Drexler. 2019. Reframing superintelligence: Comprehensive ai services as general intelligence. *Future of Humanity Institute, University of Oxford*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html.

Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese*, 198(Suppl 27):6435–6467.

Arnaud Fickinger, Simon Zhuang, Dylan Hadfield-Menell, and Stuart Russell. 2020. Multi-principal assistance games. *arXiv preprint arXiv:2007.09540*.

Jason Furman and Robert Seamans. 2019. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191.

Iason Gabriel and Vafa Ghazavi. 2021. The challenge of value alignment: From fairer algorithms to ai safety. *arXiv preprint arXiv:2101.06060*.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.

Ashok Goel. 2022. Looking back, looking ahead: Symbolic versus connectionist ai. *AI Magazine*, 42(4):83–85.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Google. 2023. Bard.

GOV.UK. 2023. Frontier ai: capabilities and risks – discussion paper.

Nico Grant and Karen Weise. 2023. In a.i. race, microsoft and google choose speed over caution. *New York Times*, 4. Updated on April 10, 2023.

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016a. The off-switch game. *arXiv preprint arXiv:1611.08219*.

Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016b. Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, 29.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021b. Aligning ai with shared human values. *ICLR 2021*.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021c. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916*.

Dan Hendrycks, Mantas Mazeika, Andy Zou, Sahil Patel, Christine Zhu, Jesus Navarro, Dawn Song, Bo Li, and Jacob Steinhardt. 2021d. What would jiminy cricket do? towards agents that behave morally. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. 2023. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*.

Dan Hendrycks. 2022. Pragmatic ai safety.

Dan Hendrycks. 2023. Natural selection favors ais over humans.

Lewis Ho, Joslyn Barnhart, Robert Trager, Yoshua Bengio, Miles Brundage, Allison Carnegie, Rumman Chowdhury, Allan Dafoe, Gillian Hadfield, Margaret Levi, et al. 2023. International institutions for advanced ai. *arXiv preprint arXiv:2307.04699*.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Jeremy Howard. 2023. Ai safety and the age of dislightenment.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. 2019. Deceptive alignment.

Evan Hubinger. 2020. An overview of 11 proposals for building safe advanced ai. *arXiv preprint arXiv:2012.07532*.

Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *arXiv preprint arXiv:1805.00899*.

Jacob Steinhardt. 2023. Emergent deception and emergent optimization.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. 2023. Champion-level drone racing using deep reinforcement learning. *Nature*, 620(7976):982–987.

Michael Kearns and Aaron Roth. 2019. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Zachary Kenton, Tom Everitt, Laura Weidinger, Iason Gabriel, Vladimir Mikulik, and Geoffrey Irving. 2021. Alignment of language agents. *arXiv preprint arXiv:2103.14659*.

B. Kenward and T. R. Sinclair. 2021. Machine morality, moral progress, and the looming environmental disaster. *Cognitive Computation and Systems*, 3:83–90.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. 2017. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998.

Leonie Koessler and Jonas Schuett. 2023. Risk assessment at agi companies: A review of popular risk assessment techniques from other safety-critical industries. *arXiv preprint arXiv:2307.08823*.

Mr Anton Korinek, Mr Martin Schindler, and Joseph Stiglitz. 2021. *Technological progress, artificial intelligence, and inclusive growth*. International Monetary Fund.

David Krueger, Tegan Maharaj, and Jan Leike. 2020. Hidden incentives for auto-induced distributional shift. *arXiv preprint arXiv:2009.09153*.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment.

Joel Lehman, Jeff Clune, Dusan Misevic, Christoph Adami, Lee Altenberg, Julie Beaulieu, Peter J Bentley, Samuel Bernard, Guillaume Beslon, David M Bryson, et al. 2020. The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial life*, 26(2):274–306.

Jan Leike, David Krueger, Tom Everitt, Miljan Martic, Vishal Maini, and Shane Legg. 2018. Scalable agent alignment via reward modeling: a research direction. *arXiv preprint arXiv:1811.07871*.

Jan Leike. 2023. A proposal for importing society's values.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M Dai, Diyi Yang, and Soroush Vosoughi. 2023a. Training socially aligned language models in simulated human society. *arXiv preprint arXiv:2305.16960*.

Yi Liu, Gelei Deng, Zhengzi Xu, Yuekang Li, Yaowen Zheng, Ying Zhang, Lida Zhao, Tianwei Zhang, and Yang Liu. 2023b. Jailbreaking chatgpt via prompt engineering: An empirical study. *arXiv preprint arXiv:2305.13860*.

Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. 2023. Mechanistic mode connectivity. In *International Conference on Machine Learning*, pages 22965–23004. PMLR.

Alasdair MacIntyre. 2013. *After virtue*. A&C Black.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

David M Messick and Charles G McClintock. 1968. Motivational bases of choice in experimental games. *Journal of experimental social psychology*, 4(1):1–25.

Yisroel Mirsky, Ambra Demontis, Jaidip Kotak, Ram Shankar, Gelei Deng, Liu Yang, Xiangyu Zhang, Maura Pintor, Wenke Lee, Yuval Elovici, and Battista Biggio. 2023. The threat of offensive AI to organizations. *Comput. Secur.*, 124:103006.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Emad Mostaque. 2022. Democratizing ai, stable diffusion & generative models.

Ryan O Murphy and Kurt A Ackermann. 2014. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1):13–41.

Safiya Umoja Noble, Beatrice Dias, Sara Cole Stratton, Aimee van Wynsberghe, Carlos Affonso Souza, Ilene Carpenter, Alvaro Martin Enriquez, and Emily Ratté. 2021. Ai regulation through an intergenerational lens.

OecdAI. 2021. Ai principles.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill*, 3(3):e10, March.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Introducing superalignment. Accessed on July 5, 2023.

Robert Opp. 2023. Committing to bridging the digital divide in least developed countries.

Toby Ord. 2020. *The precipice: Existential risk and the future of humanity*. Hachette Books.

Osonde A Osoba, Benjamin Boudreaux, and Douglas Yeung. 2020. Steps towards value-aligned systems. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 332–336.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*.

Paulina Karolina Pankowska. 2020. Framework on ethical aspects of artificial intelligence, robotics and related technologies. *European Parliament*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023a. Generative agents: Interactive simulacra of human behavior. In Sean Follmer, Jeff Han, Jürgen Steimle, and Nathalie Henry Riche, editors, *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, pages 2:1–2:22. ACM.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2023b. Ai deception: A survey of examples, risks, and potential solutions. *arXiv preprint arXiv:2308.14752*.

Hammond Pearce, Baleegh Ahmad, Benjamin Tan, Brendan Dolan-Gavitt, and Ramesh Karri. 2022. Asleep at the keyboard? assessing the security of github copilot's code contributions. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 754–768. IEEE.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR.

Ethan Perez, Saffron Huang, H. Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3419–3448. Association for Computational Linguistics.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13387–13434. Association for Computational Linguistics.

Karl R. Popper. 1935. *The Logic of Scientific Discovery*. Routledge, London, England.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483. IEEE.

Milton Rokeach. 1973. *The nature of human values*. Free press.

Tim Rudner and Helen Toner. 2021a. Key concepts in ai safety: Interpretability in machine learning.

Tim Rudner and Helen Toner. 2021b. Key concepts in ai safety: Robustness and adversarial examples.

Kiersten M Ruff and Rohit V Pappu. 2021. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208.

Stuart Russell. 2019. *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Jonas B Sandbrink. 2023. Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools. *arXiv preprint arXiv:2306.13952*.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. 2023. Towards best practices in agi safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*.

Shalom H Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 1994. Are there universal aspects in the structure and contents of human values? *Journal of social issues*, 50(4):19–45.

Elizabeth Seger, Noemi Dreksler, Richard Moulange, Emily Dardaman, Jonas Schuett, K Wei, Christoph Winter, Mackenzie Arnold, Seán Ó hÉigeartaigh, Anton Korinek, et al. 2023. open-sourcing-highly-capable-foundation-models.

Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

Rohin Shah, Pedro Freire, Neel Alex, Rachel Freedman, Dmitrii Krasheninnikov, Lawrence Chan, Michael D Dennis, Pieter Abbeel, Anca Dragan, and Stuart Russell. 2020. Benefits of assistance over reward learning. *https://openreview.net/forum?id=DFIoGDZejIB*.

Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. Goal misgeneralization: Why correct specifications aren't enough for correct goals. *arXiv preprint arXiv:2210.01790*.

Jacob N Shapiro and David A Siegel. 2010. Is this paper dangerous? balancing secrecy and openness in counterterrorism. *Security Studies*, 19(1):66–98.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Yonadav Shavit. 2023. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*.

Toby Shevlane, Sebastian Farquhar, Ben Garfinkel, Mary Phuong, Jess Whittlestone, Jade Leung, Daniel Kokotajlo, Nahema Marchal, Markus Anderljung, Noam Kolt, et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471.

Paul Smolensky. 1987. Connectionist ai, symbolic ai, and the brain. *Artificial Intelligence Review*, 1(2):95–109.

Nate Soares, Benja Fallenstein, Stuart Armstrong, and Eliezer Yudkowsky. 2015. Corrigibility. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Emily H. Soice, Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. Can large language models democratize access to dual-use biotechnology?

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. 2018. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis)informs us better than humans. *Science Advances*, 9(26):eadh1850.

Jacob Steinhardt and Helen Toner. 2020. Why robustness is key to deploying ai.

Jacob Steinhardt. 2023. Emergent Deception and Emergent Optimization, February.

Victor Storchan, Svitlana Vyetrenko, and Tucker Balch. 2021. Learning who is in the market from time series: market participant discovery through adversarial calibration of multi-agent simulators. *arXiv preprint arXiv:2108.00664*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Shea Swaugerarchive. 2020. Software that monitors students during tests perpetuates inequality and violates their privacy.

Nassim Nicholas Taleb. 2007. *The black swan: The impact of the highly improbable*, volume 2. Random house.

Jonas Tallberg, Eva Erman, Markus Furendal, Johannes Geith, Mark Klamberg, and Magnus Lundgren. 2023. The global governance of artificial intelligence: Next steps for empirical and normative research. *arXiv preprint arXiv:2305.11528*.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca D. Dragan, and Daniel S. Brown. 2023. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Turner, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. Optimal policies tend to seek power. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23063–23074. Curran Associates, Inc.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Fabio Urbina, Filippa Lentzos, Cédric Invernizzi, and Sean Ekins. 2022. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191.

Paul AM Van Lange, Ellen De Bruin, Wilma Otten, and Jeffrey A Joireman. 1997. Development of prosocial, individualistic, and competitive orientations: theory and preliminary evidence. *Journal of personality and social psychology*, 73(4):733.

Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7.

Krakovna Victoria, Uesato Jonathan, Mikulik Vladimir, Rahtz Matthew, Everitt Tom, Kumar Ramana, Kenton Zac, Leike Jan, and Legg Shane. 2020. Specification gaming: the flip side of ai ingenuity.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Claus O Wilke, Jia Lan Wang, Charles Ofria, Richard E Lenski, and Christoph Adami. 2001. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Alan F Winfield, Katina Michael, Jeremy Pitt, and Vanessa Evers. 2019. Machine ethics: The design and governance of ethical ai and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3):509–517.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

Taku Yamagata, Ryan McConville, and Raul Santos-Rodriguez. 2021. Reinforcement learning with feedback from multiple humans with diverse skills.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser, and Qiang Yang. 2018. Building ethics into artificial intelligence. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5527–5533. ijcai.org.

Luyao Yuan, Xiaofeng Gao, Zilong Zheng, Mark Edmonds, Ying Nian Wu, Federico Rossano, Hongjing Lu, Yixin Zhu, and Song-Chun Zhu. 2022. In situ bidirectional human-robot value alignment. *Science robotics*, 7(68):eabm4183.

Rowan Zellers. 2019. Why we released grover.

Zhaowei Zhang, Nian Liu, Siyuan Qi, Ceyao Zhang, Ziqi Rong, Yaodong Yang, and Shuguang Cui. 2023. Heterogeneous value evaluation for large language models. *arXiv preprint arXiv:2305.17147*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

基于大语言模型的自主智能体概述

陈旭
高瓴人工智能学院
中国人民大学
xu.chen@ruc.edu.cn

摘要

近年来，基于大语言模型的自主智能体受到了学术界和工业界的广泛关注，其关键在于利用大语言模型作为核心控制器，并设计相应的辅助模块增强智能体在动态环境中的演化和适应能力，从而提升智能体自主解决任务的能力。本文通过总结过去工作，抽象出智能体设计的通用范式，并讨论了大模型时代自主智能体能力提升的途径。我们还从个体拓展到系统，深入探讨了多自主智能体系统常见的交互机制和面临的重要问题。

关键词： 大语言模型 ； 自主智能体

A Survey on Large Language Model based Autonomous Agents

Xu Chen
Gaoling School of Artificial Intelligence
Renmin University of China
xu.chen@ruc.edu.cn

Abstract

In recent years, autonomous agents based on large language models (LLMs) have garnered widespread attention from academia and industry. The central approach involves using LLMs as the primary controllers, complemented by auxiliary modules that enhance the agents' capabilities for evolution and adaptation in dynamic environments, thus boosting their proficiency in autonomously addressing tasks. This paper reviews previous efforts, distills a universal paradigm for agent design, and explores methods to advance the capabilities of autonomous agents in the era of LLMs. Furthermore, we extend our discussion from individual agents to systems, examining the prevalent interaction mechanisms and critical challenges faced by multi-agent systems.

Keywords: Large language model , Autonomous agent

1 引言

自主智能体作为实现通用人工智能的潜在途径之一，核心价值在于其能够独立规划和执行复杂场景下的任务。传统的自主智能体相关研究或者通过逻辑规则和符号表示来封装知识并进行推理 (Sacerdoti, 1975)，或者结合传统强化学习方法 (Watkins, 1989; Rummery and Niranjan, 1994)和深度神经网络 (Tesauro, 1995; Li, 2017)来从环境互动中不断学习并改进智能体自身的行为策略。基于以上传统算法的自主智能体已经在一些特定任务场景下取得了显著的成果，例如AlphaGo (Silver et al., 2016)和DQN (Mnih et al., 2013)。然而，这些早期探索仍然大多利用简单的启发式策略在封闭环境中进行学习，这一过程与人类复杂且灵活的学习方式存在着明显不同。具体来说，人类的思维过程极为复杂，涉及多种认知功能和学习机制，这使其能够在不

第二十三届中国计算语言学大会论文集，第141页-第150页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

断变化的环境中快速学习和适应。由于这种差异的存在，传统研究中的自主智能体在面对开放和无限制的环境时，往往难以具备人类水平的决策能力。

近期，大语言模型（Large Language Models, LLMs）的相关研究取得了显著进展，展现出了强大的自然语言理解能力和实现人类水平智能的巨大潜力 (Wang et al., 2024)。这种令人瞩目的表现主要得益于对广泛的训练数据集和庞大的模型参数的深入利用 (Brown et al., 2020)。基于此，一个快速成长的新领域开始显现，即利用LLMs作为核心控制单元来构建高级的自主智能体，旨在实现与人类相似的决策水平。在这一研究领域，研究者们已经开发了多种有希望的智能模型，其核心理念是赋予LLMs关键的人类特性，例如记忆和规划能力，以期它们能够高效地完成各种任务。然而，值得注意的是，当前基于大语言模型的自主智能体研究大多是独立发展的，对于这个快速增长且潜力巨大的领域，目前还缺少一个系统的综述和比较分析。

本文广泛而深入地探讨了基于LLMs的自主智能体研究。具体而言，研究聚焦于在构建自主智能体过程中的两大挑战：一是设计一种能够最大化LLMs潜能的自主智能体架构；二是探索如何提升自主智能体执行多样化特定任务的能力。简单来说，第一个挑战相当于为自主智能体打造一个“硬件”平台，而第二个挑战则着眼于为自主智能体提供“软件”支持。针对第一个挑战，本文提出了一个全面的设计框架，整合了众多现有研究中的多种模块设计，以确保其在各种场景下的适应性和可扩展性；至于第二个挑战，本文归纳了一套常用的策略，旨在赋予智能体特定的能力。接下来，本文将对这两个关键议题进行详尽的阐述。

2 基于LLMs的自主智能体的架构设计

LLMs的迅猛发展在对话和问答领域彰显了其非凡的潜力，然而要创建一个具有自主性的智能体，仅依靠问答功能是远远不够的。自主智能体需要能够独立地感知周遭环境，并且能够根据当前情况采取相应的行为，从环境中学习，以实现类似人类的自我进化。为了缩小传统LLMs和自主智能体之间的差异，一个至关重要的步骤是构建一个合理的智能体架构，这将有助于LLMs充分发挥其潜力。在这一领域，已有研究提出了多种模块来提升LLMs的能力。本文对这些研究成果进行整合，从设定模块、记忆模块、规划模块和行动模块四个方面对已有工作进行总结 (Wang et al., 2024)。设定模块的核心功能是确定智能体所扮演的角色。记忆与规划模块则将智能体置于一个动态变化的环境中，使其能够回顾过往行为并预测未来行动。行动模块则负责将智能体的决策转化为实际的输出。在这些模块中，设定模块对记忆和规划模块有着直接的影响，而这三个模块又共同决定了行动模块的表现。本文将对这些模块进行详细介绍。

2.1 设定模块

自主智能体在执行任务时，经常扮演如程序员、教师或领域专家等特定角色。智能体的设定模块是为了指导智能体根据任务来设定其角色，这些设定信息通常通过提示词的形式嵌入，以影响LLMs的具体行为表现。智能体的设定通常包含基础信息，例如年龄、性别、职业，以及反映智能体性格特征的心理属性，还有描述智能体之间互动关系的社交属性。智能体设定的具体内容很大程度上取决于应用场景的需求。例如，如果应用的目的是探索人类的认知过程，那么心理属性就变得特别重要。在确定了设定文件所需包含的信息类型后，下一步是为智能体制定具体的设定文件，这通常涉及以下策略：例如，为了打造具有不同性格特征的智能体，人们可能会用“你是一个开朗的人”或者“你是一个害羞的人”来定义智能体的性格。这种手工定制方法已被许多研究工作采用，用以明确智能体的设定资料。例如，Expertprompting(Xu et al., 2023)通过人工精心设计不同领域的专家设定信息来增强智能体在不同领域的能力。Character-LLM (Shao et al., 2023)人工选择了七位名人作为智能体扮演的对象，结合各自维基百科的资料设计了每个智能体的设定。总的来说，这种人工设定方法具有很高的灵活性，因为可以为智能体指定任何所需的设定信息。但是，当涉及大量智能体时，这种方法可能会变得非常耗时。为缓解该问题，人们可以用大模型生成设定。RecAgent (Wang et al., 2023)首先由人工提供少量背景信息（如年龄、性别、个性特征和电影喜好）作为智能体的示例设定，然后利用ChatGPT根据这些种子信息生成更多的智能体设定。SOTOPIA(Zhou et al., 2023)通过GPT-4生成了多种具有不同行为类型的智能体及他们之间的社交关系，以研究智能体在动态环境中的社交能力。LLMs驱动策略在处理大量智能体时可以显著节省时间，但可能会在生成档案的精确度上有所欠缺。同时，人们也可以考虑使用真实数据集。在这种策略中，智能体的设定是基于真实世界数据集的信息。通常，首先将数据集中关于真实个体的信息整理成自然语言提示，然

后使用这些提示来设定智能体的档案。例如，EconAgent(Li et al., 2023)使用真实的年龄和收入信息作为智能体的设定信息，进而通过智能体来模拟宏观经济行为。数据集映射策略能够精确地捕捉到真实人群的属性，使得智能体的行为更加有意义，更能反映现实世界的情况。

2.2 记忆模块

在智能体的架构设计中，记忆组件扮演着至关重要的角色，它负责存储智能体从环境中获取的感知信息，并利用这些信息来指导未来的决策和行动。记忆组件不仅帮助智能体积累经验、自我进化，而且确保了其行为的一致性、合理性和有效性。接下来，本文将详细探讨记忆模块的结构、格式和操作。从结构上讲，基于LLMs的自主智能体通常借鉴人类认知科学的理论和方法。人类的记忆过程大致可以分为三个阶段：首先是感觉记忆，它负责存储从感官接收到的原始信息；其次是短期记忆，它用于暂时保留信息；最后是长期记忆，它负责将信息固定下来，以便长期保存。在设计这些智能体的记忆系统时，研究者们从人类记忆的这些阶段中汲取灵感。具体来说，短期记忆可以类比为受限于Transformer架构的上下文窗口，它能够处理并存储输入的信息；而长期记忆则类似于一个外部的存储系统，智能体可以迅速地访问和检索所需的信息。除了记忆结构之外，分析记忆模块的另一个角度是基于记忆存储介质的格式，例如自然语言记忆或编码向量记忆。不同的存储器格式具有不同的优势并且适合不同的应用。下面本文介绍几种有代表性的记忆格式。在现有研究中，Memochat(Lu et al., 2023)通过自我编写自然语言形式的备忘录来实现长期开放领域对话；Memory Sandbox(Huang et al., 2023)通过自然语言的形式存储记忆，便于用于对智能体的记忆进行管理，减少冗余记忆的干扰。MemoryBank (Zhong et al., 2023)通过将记忆片段转换为嵌入向量，构建了一个索引化的语料库，从而优化了记忆的检索过程。TiM(Liu et al., 2023)从原始信息中抽取出实体之间的关系作为记忆存储至数据库中来增强推理能力；在DB-GPT (Zhou et al., 2023)中，记忆模块同样也是基于数据库构建的，为了更直观地操作记忆信息，智能体经过微调以理解和执行SQL查询，使其能够直接使用自然语言与数据库进行交互。TradingGPT(Li et al., 2023)设计了分层记忆架构来存储不同的市场信息。从操作上讲，记忆模块在允许智能体通过与环境交互来获取、积累和利用重要知识方面发挥着关键作用。智能体与环境之间的交互是通过三个关键的记忆操作来完成的：记忆读取、记忆写入和记忆反思。记忆读取的目的是从记忆中提取有意义的信息以增强智能体的行动。例如，利用以前成功的行动来实现类似的目标。记忆读取的关键在于如何提取有价值的信息。通常，信息提取常用三个标准，即临近度、相关性和重要性 (Park et al., 2023)。记忆写入的目的是将从环境中感知到的有价值的信息存储在记忆中，这为将来检索信息中的丰富记忆奠定了基础，使智能体能够更高效、更合理地行动。在记忆写入的过程中，有两个潜在问题需要仔细解决。一方面，解决如何存储与现有记忆相似的信息（即记忆冗余）至关重要。另一方面，重要的是要考虑当记忆达到其存储限制（即记忆溢出）时如何删除信息。记忆反思旨在模仿人类见证和评估自己的认知、情感和行为过程的能力。当应用于智能体时，其目标是为智能体提供独立总结和推断更抽象、复杂和高级信息的能力。传统LLMs和智能体之间的一个显著区别是，后者必须具备在动态环境中学习和完成任务的能力。如果说记忆模块负责管理智能体过去的行为，那么就必须有另一个重要的规划模块来帮助智能体计划他们未来的行动。

2.3 规划模块

当面对复杂的任务时，人类会倾向于将其分解为更简单的子任务，并逐步解决；当人类在遭遇失败时，会对过去的错误原因进行反思，从失败中吸取教训。规划模块旨在赋予智能体类似的能力，从而使其可以解决复杂任务时更加有效与可靠。例如，在结构化思维链 (Li et al., 2023)中，作者用结构化的思维链约束大模型使用程序结构（例如：顺序、分支和循环结构）去组织思维过程，引导大模型从程序语言的角度去逐步思考如何解决需求。Q* (Wang et al., 2024)让大模型推理多条路径的同时，维护一个价值函数，综合考虑了当前状态的价值与未来期望价值，最后利用A*搜索算法对状态进行最佳优先搜索，实现了对复杂推理任务的全盘规划，在ToolLLM (Yao et al., 2023)中，智能体通过思考-调用工具-获取结果的循环来形成提示，其中思考环节用于促进深度的推理和策略规划，以指导智能体的行动；调用工具环节指的是智能体调用什么工具，怎么调用工具；而获取结果则是指动作产生的结果，这些结果通常是环境反馈的结果。此外，下一轮的思考过程会受到前一次观察的影响，从而使新的计划更加贴合环境条件。ReHAC(Feng et al., 2024)提出在agent多步规划的过程中，使用一个经过强化学习训练

后的策略模型，把子任务分配给人类，让经验丰富的人类来帮助基于LLMs的智能体更好地完成任务规划。在论文(Du et al., 2023)中，多个语言模型智能体在多轮中提出并辩论它们各自的回应和推理过程，以得出共同的最终答案。

2.4 行动模块

行动模块负责将智能体的决策转化为具体的结果。该模块位于最下游位置，直接与环境进行交互，受到设定、记忆和规划模块的影响。在目标方面，自主智能体的行动往往是为了完成各种目标，例如完成具有明确定义的任务（作为助手给用户推荐一家附近的餐厅 (Yao et al., 2023)）；与其他智能体或真实人类进行沟通，共享信息或合作 (Du et al., 2023)；探索陌生的环境以扩展感知,并在探索和利用之间取得平衡 (Wang et al., 2024)。在动作空间方面，动作空间指的是智能体可以执行的所有动作的集合。一般来说，本文可以将这些行动大致分为如下两类。外部工具：虽然LLMs已被证明在完成大量任务方面是有效的，但它们可能不适用于需要全面专业知识的领域。此外，LLMs也可能遇到难以自行解决的幻觉问题。为了缓解上述问题，智能体被赋予了调用外部工具（如API）来执行更多复杂操作的能力 (Yao et al., 2023)。内部知识：LLMs通常能够很好地理解人类常识并生成高质量的对话。基于这种能力，许多自主智能体可以直接利用其内在知识来模拟人类行为，从而做出类似人类的决策。

3 如何增强基于LLMs的自主智能体的特定能力

在上述内容中，本文详细探讨了基于LLMs的自主智能体的统一设计架构，旨在更好地激发LLMs的能力，使智能体能够以与人类相仿的智能水平完成复杂任务。自主智能体的架构相当于其“硬件”部分，但仅有硬件是不够的，因为智能体在面对特定任务时可能缺少必要的能力、技巧和知识，这些可以被看作是“软件”部分。为了弥补智能体在这些方面的不足，研究者们开发了多种策略。本文根据是否需要对LLMs进行微调，将这些策略划分为两大类，并将在下文中对每类策略进行详尽的阐述。

3.1 使用微调进行能力获取

提升自主智能体执行特定任务的效能，可以通过针对与任务相关的数据集进行微调来实现。这些数据集的来源可以多样化，包括但不限于人类专家的标注、由先进的语言模型生成的内容，或是从现实世界的应用场景中直接收集的数据。通过这样的微调，智能体能够更精准地理解和适应特定任务的需求，从而提高其性能。

利用人类注释数据集进行微调：微调LLMs以适应不同的应用场景，是一种通过人工标注数据集实现的高效策略。研究人员首先规划出需要的标注任务，随后招募人员来执行这些任务。例如，DPO (Liu et al., 2023)通过优化人类标注的偏好数据的对数似然，以实现LLMs与人类价值观的一致性，并直接利用这些自然语言数据集对模型进行微调;

利用LLMs生成的数据集进行微调：建立人类注释的数据集需要招募人员，这可能会带来高昂的成本，特别是当需要注释大量样本时。考虑到LLMs可以在广泛任务中实现类似人类的能力，一个自然的想法是使用LLMs来完成注释任务。虽然从这种方法产生的数据集可能不像人类注释的数据集那样完美，但它成本更低，并且可以用来生成更多的样本。例如，在FireAct (Chen et al., 2023)中，为了增强开源LLMs的推理能力，作者使用Chatgpt收集了一批专家数据，在复杂推理任务中获得了较好的性能，鲁棒性以及泛化性。

3.2 在没有微调的情况下获取能力

在基于LLMs的自主智能体时代，由于微调可能耗费大量资源，所以也可以采用非微调的方法提升模型能力。例如，在CoT (Wei et al., 2022)中，为了赋予智能体进行复杂任务推理的能力，作者将中间推理步骤作为少样本示例呈现在提示中。类似的技术也被用于CoT-SC (Wang et al., 2022)和ToT (Yao et al., 2024)中。EvoAgent(Yuan et al., 2024)通过进化算法，根据当前任务自动生成新的智能体从而自动化构建一个多智能体系统来解决当前任务。AgentHospital(Li et al., 2024)模拟了一家医院的运行流程，其中作为医生的智能体会根据历史信息积累经验，结合系统中已有的病历提高自己的诊断能力。ICE(Qian et al., 2024)将智能体的自我进化分解为探索，固化和利用三个阶段，将探索成功的路径保存为规划链，再下次执行任务时这些成功的规划链会作为参考的依据。

通过比较上述的自主智能体能力获取策略可以发现：微调策略通过优化模型参数，能够吸收特定任务的丰富知识，但这种方法主要适用于开源的LLMs；而无需微调的策略则依赖于精心设计的提示形式或机制工程来增强自主智能体的任务表现，这种策略同时适用于开源和闭源的LLMs，但由于LLMs对输入上下文的窗口大小有限制，它们往往无法处理过多的任务信息，同时提示和机制的设计空间非常大，这无疑增加了寻找最优解的难度。

4 由个体到系统：基于LLMs的多自主智能体系统

基于LLMs的单一智能体已经显示出了强大的认知能力，这些单一个体的开发主要集中在设计其内部工作方式及其对外部环境的响应。相比之下，由多个自主智能体组成的系统则强调每个智能体的角色属性、各个智能体间交互以及集体决策程序的多样化。多智能体系统旨在实现多个独立主体的协作，每个主体都被赋予了独特的策略和行为，促进彼此的沟通，进而有助于处理更动态、更复杂的任务。本节将深入探讨多智能体系统的内部交互形式和重要研究问题。

4.1 多智能体系统内部的交互方式

基于LLMs的多自主智能体系统旨在模拟人类群体动态，通过智能体之间的合作、竞争或层次化交互来解决复杂问题。智能体的交互方式对于整个系统的效能至关重要，它们可以是合作互补的，也可以是相互竞争的，甚至是动态变化的。

合作交互：合作型多智能体系统是实际应用中最广泛部署的模式。在这些系统中，每个智能体评估其对应智能体的需求和能力，积极追求协作行动并共享信息 (Li et al., 2023)。这种框架提供了多种潜在优势，例如提高任务效率、增强集体决策能力，以及解决单个智能体无法单独解决的复杂现实世界问题，最终实现整体系统性能的提升。SPP (Wang et al., 2023)通过利用多角色自我合作，实现了多轮对话，有效地将单一的大型语言模型转变为认知协同体；Generative Agents (Park et al., 2023)利用基于LLMs的多智能体系统模仿现实人类行为，促进智能体之间的合作；CAMEL (Li et al., 2023)通过任务导向的角色扮演实现AI助手与用户之间的合作多轮对话；MetaGPT (Hong et al., 2023)将高效的工作流程整合到LLMs驱动的多智能体协作编程方法中，促进了不同角色之间的协作；ChatDev (Qian et al., 2023)使用基于LLMs的多个智能体进行对话交流并解决任务，为加快软件应用的设计与开发提供了新思路；受Minsky心智社会概念 (Minsky , 1988)的启发，NLSOM (Zhuge et al., 2023)引入了基于自然语言的心智社会（NLSOMs）的概念，由多个LLMs和其他基于神经网络的专家通过自然语言界面进行通信。这种方法被应用于解决不同场景中的复杂任务；在具身智能领域，RoCo (Mandi et al., 2023)利用LLMs进行高层通信和低层路径规划，从而促进了多个机器人之间的协作；Interact (Chen and Chang , 2023)为各种角色（如检查员和分类器）的自主智能体分配任务，在AlfWorld (Shridhar et al., 2020)环境中取得了显著的成功率；AutoAgents (Chen et al., 2023)能够适应性地生成和协调多个专业智能体，从而形成一个强大的AI团队，在各种复杂任务场景中协作实现目标。

竞争交互：在竞争场景中，各个自主智能体往往发展出一系列策略和技能，例如制定稳健的行动计划和分析实时的竞争反馈，以确保自身能在激烈的彼此对抗中获得优势。在这种方式下，多智能体系统不仅能够通过内部竞争实现更加优越的整体表现，还能够在面对复杂挑战时展现出更高的适应性和创新能力。近期，Liang等人 (Liang et al., 2023)通过引入多智能体辩论框架增强了整体系统解决问题的能力；ChatEval (Chan et al., 2023)同样采用多智能体策略，使LLMs与多样化的智能对手进行互动，通过利用各个内部智能体的独特能力和专业知识，显著增强了系统解决复杂任务的效率和效果。

层次交互：层次型多智能体系统关注于在层次结构中开发高效控制结构、信息传输方法和任务分解技术。这些策略促进了不同层级智能体之间的有效协作，提高了整体系统性能。层次型多智能体系统通常以树状结构组织，父节点智能体负责任务分解并将任务分配给子节点智能体。后者遵循来自其父节点的指令并提供汇总反馈。例如，AutoGen (Wu et al., 2023)使用各种智能体执行代码生成和文本写作等任务，通过对话进行任务分解。关于层次型多智能体系统的研究仍然处于起步阶段，目前的探索仍局限于少数几个层级。

动态交互：动态互动指的是多智能体系统中结构的灵活性，其中智能体的角色、它们的关系以及智能体的总数可以随时间演变。动态交互型多智能体系统往往表现出强大的上下文适应性，即互动模式根据内部或外部影响进行调整，其中的智能体具有根据变化条件动态调整各自

角色和彼此之间关系的能力。例如，(Talebirad and Nadiri , 2023)展示了根据特定任务添加或删除智能体的可能性。

混合交互：多智能体系统中的混合交互需要在合作和竞争之间进行微妙的平衡以实现期望的结果。当前基于LLMs的多智能体系统研究聚焦于开发协作竞争算法，这是一个关键的研究领域。这种算法能够有效地帮助智能体系统在复杂的环境场景中做出更优的整体决策。近期，Xu等人 (Xu et al., 2023)使多个基于大语言模型的自主智能体参与狼人杀游戏，各个智能体根据不对称的信息来合作或背叛他人以实现各自的目标；Light (Light et al., 2023)等人则在阿瓦隆游戏场景下设计了一个多智能体系统，其中各个智能体导航动态发展的游戏阶段，并与其他智能体进行合作或欺骗以履行其指定角色；受人类行为启发，Corex (Sun et al., 2023)结合了辩论、审查和检索等多种合作范式，旨在增强推理过程的真实性、保真度和可靠性。

4.2 多智能体系统研究的开放问题

本小节对基于LLMs的多智能体系统中的开放问题进行探讨，旨在为这一迅速发展的研究领域指明进一步探索和创新的可行方向。

迈向多模态环境：近期关于基于LLMs的多智能体系统的研究主要集中在基于文本的环境上，展示了它们在文本处理和生成方面的专业能力。然而，在多模态设置的背景下仍然存在着一个显著的空白。在多模态环境中，多智能体系统往往需要处理来自各种感官输入的数据，并以多种模态产生输出，如图像、音频、视频和物理动作，其中的挑战主要在于处理和整合不同数据类型的复杂性，这要求智能体具有高级的感知和认知能力。此外，生成连贯且符合上下文的多模态输出要求各个智能体保持共享理解并有效协调它们的行动。解决这些挑战对于开发能够适应各种现实世界场景的多功能和适应性强的多智能体系统至关重要。

共同获取集体智能：传统的多智能体系统通常依赖于离线训练数据集的强化学习。相比之下，基于LLMs的多智能体系统主要利用通过与环境或人类交互获得的即时反馈。然而，这种学习范式需要一个可靠的交互环境，为跨各种任务的可扩展性带来了挑战。此外，当前研究中的主要方法强调使用记忆和自我演化技术，根据反馈适应个体智能体。尽管对于每个个体智能体有效，但这些方法未能充分利用多智能体系统的潜在集体智能。通过单独调整智能体，它们忽视了来自协调的多智能体交互的协同效应。因此，同时调整多个智能体以实现最佳集体智能仍然是基于LLMs的多智能体系统面临的关键挑战。

扩大智能体数量：增加智能体数量可以提高任务效率并增强社会仿真的真实性 (Park et al., 2023; Qian et al., 2023)。然而，现有研究仍然存在许多挑战。随着部署的人工智能智能体数量增加，计算负载也随之增加，这需要改进架构设计和计算优化以确保系统平稳运行。随着智能体数量的增长，通信和消息传播构成了重大障碍，导致高度复杂的通信网络。在多智能体系统中，由于幻觉和误解导致的信息传播偏见可能会扭曲信息传播，尤其是在智能体数量较多时，增加了风险并降低了通信的可靠性。此外，随着智能体数量的增加，协调智能体变得越来越困难，可能阻碍合作和效率，从而影响实现共享目标的进展。

5 结论

本文详细探讨了以LLMs为核心的自主智能体技术，主要聚焦于体系架构的设计,特定能力的获取和多自主智能体这三个关键议题。本文以一种统一的框架视角，系统地涵盖了当前主流的智能体构建技术，归纳总结了一系列赋予智能体特定能力的关键策略并分析了多自主智能体系统内部的交互机制和可能的开放问题。通过对现有相关工作的综合梳理，本文充分阐述了基于LLMs的自主智能体的研究现状和发展趋势。这些深入的讨论与思考不仅有助于读者更加全面地把握这一新兴技术领域，也为未来的探索和创新提供了宝贵的指导。随着技术的持续发展和理论的逐步深化，以LLMs为核心的智能体系统有望在各个领域发挥重要作用，为人类社会注入新的发展动力和创新潜力。

参考文献

J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S. K. S. Yau, Z. Lin, L. Zhou, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, and M. Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Y. Dong, X. Jiang, Z. Jin, and G. Li. 2023. Self-collaboration code generation via chatgpt. *arXiv preprint arXiv:2304.07590*.

L. Wang, J. Zhang, X. Chen, Y. Lin, R. Song, W. X. Zhao, and J. R. Wen. 2023. Recagent: A novel simulation paradigm for recommender systems. *arXiv preprint arXiv:2306.02552*.

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

G. Wang, Y. Xie, Y. Jiang, A. Mandlekar, C. Xiao, Y. Zhu, L. Fan, and A. Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

W. Zhong, L. Guo, Q. Gao, and Y. Wang. 2023. Memorybank: Enhancing large language models with long-term memory. *arXiv preprint arXiv:2305.10250*.

X. Zhu, Y. Chen, H. Tian, C. Tao, W. Su, C. Yang, G. Huang, B. Li, L. Lu, X. Wang, et al. 2023. Ghost in the minecraft: Generally capable agents for open-world environments via large language models with text-based knowledge and memory. *arXiv preprint arXiv:2305.17144*.

C. Hu, J. Fu, C. Du, S. Luo, J. Zhao, and H. Zhao. 2023. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*.

X. Zhou, G. Li, and Z. Liu. 2023. Llm as dba. *arXiv preprint arXiv:2308.05481*.

A. Modarressi, A. Imani, M. Fayyaz, and H. Schütze. 2023. Retllm: Towards a general read-write memory for large language models. *arXiv preprint arXiv:2305.14322*.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone. 2023. LLM+P: Empowering large language models with optimal planning proficiency. *arXiv preprint arXiv:2304.11477*.

S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Twelfth International Conference on Learning Representations*.

W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, et al. 2022. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*.

A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, et al. 2024. Self-refine: Iterative refinement with selffeedback. *Advances in Neural Information Processing Systems*, 36.

Z. Wang, S. Cai, G. Chen, A. Liu, X. Ma, and Y. Liang. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multitask agents. *arXiv preprint arXiv:2302.01560*.

M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, et al. 2022. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*.

H. Liu, C. Sferrazza, and P. Abbeel. 2023. Chain of hindsight aligns language models with feedback. In *The Twelfth International Conference on Learning Representations*.

B. Y. Lin, Y. Fu, K. Yang, F. Brahman, S. Huang, C. Bhagavatula, P. Ammanabrolu, Y. Choi, and X. Ren. 2024. Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks. *Advances in Neural Information Processing Systems*, 36.

Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, and Y. Su. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

K. A. Fischer. 2023. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi). *arXiv preprint arXiv:2305.12647*.

Y. Shu, H. Gu, P. Zhang, H. Zhang, T. Lu, D. Li, and N. Gu. 2023. Rah! recsys-assistant-human: A human-central recommendation framework with large language models. *arXiv preprint arXiv:2308.09904*.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

C. Colas, L. Teodorescu, P. Y. Oudeyer, X. Yuan, and M. A. Côté. 2023. Augmenting autotelic agents with large language models. *arXiv preprint arXiv:2305.12487*.

N. Nascimento, P. Alencar, and D. Cowan. 2023. Self-adaptive large language model (llm)-based multiagent systems. In *2023 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C)*, pages 104–109.

S. Saha, P. Hase, and M. Bansal. 2023. Can language models teach weaker agents? teacher explanations improve students via theory of mind. *arXiv preprint arXiv:2306.09299*.

L. Wang, C. Ma, X. Feng, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6): 1-26.

Chris Watkins. 1989. Learning from delayed rewards.

Gavin A. Rummery and Mahesan Niranjan. 1994. On-line Q-learning using connectionist systems. *University of Cambridge, Department of Engineering Cambridge, UK.*

Gerald Tesauro. 1995. Temporal difference learning and td-gammon. *Communications of the ACM*, 38(3): 58–68.

Yuxi Li. 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587): 484–489.

V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

E. D. Sacerdoti. 1975. The nonlinear nature of plans. In *Advance Papers of the Fourth International Joint Conference on Artificial Intelligence*, pages 206–214.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Kojima, Takeshi and Gu, Shixiang Shane and Reid, Machel and Matsuo, Yutaka and Iwasawa, Yusuke 2022. Language models are few-shot learners. *Advances in neural information processing systems*, 35: 22199–22213.

Chen, Baian and Shu, Chang and Shareghi, Ehsan and Collier, Nigel and Narasimhan, Karthik and Yao, Shunyu 2023. Fireact: Toward language agent fine-tuning. *arXiv preprint arXiv:2310.05915*.

G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36: 51991–52008.

Z. Wang, S. Mao, W. Wu, T. Ge, F. Wei, and H. Ji. 2023. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona self collaboration. *arXiv preprint arXiv:2307.05300*, 1(2): 3.

M. Minsky. 1988. Society of mind. *Simon and Schuster*.

M. Zhuge, H. Liu, F. Faccio, D. R. Ashley, R. Csordás, A. Gopalakrishnan, A. Hamdi, H. A. A. K. Hammoud, V. Herrmann, K. Irie, et al. 2023. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*.

P.-L. Chen and C.-S. Chang. 2023. Interact: Exploring the potentials of chatgpt as a cooperative agent. *arXiv preprint arXiv:2308.01552*.

Z. Mandi, S. Jain, and S. Song. 2023. Roco: Dialectic multi-robot collaboration with large language models. *arXiv preprint arXiv:2307.04738*.

M. Shridhar, X. Yuan, M.-A. Côté, Y. Bisk, A. Trischler, and M. Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, and C. Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, Z. Tu, and S. Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

J. Light, M. Cai, S. Shen, and Z. Hu. 2023. From text to tactic: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*.

Q. Sun, Z. Yin, X. Li, Z. Wu, X. Qiu, and L. Kong. 2023. Corex: Pushing the boundaries of complex reasoning through multi-model collaboration. *arXiv preprint arXiv:2310.00280*.

Y. Talebirad and A. Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Y. Xu, S. Wang, P. Li, F. Luo, X. Wang, W. Liu, and Y. Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

G. Chen, S. Dong, Y. Shu, G. Zhang, S. Jaward, K. Börje, J. Fu, and Y. Shi. 2023. Autoagents: The automatic agents generation framework. *arXiv preprint*.

C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Y. Shao, L. Li, J. Dai, and X. Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Chaojie Wang and Yanchen Deng and Zhiyi Lv and Zeng Liang and Jujie He and Shuicheng Yan and An Bo. 2024. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning. *arXiv preprint arXiv:2406.14283*.

C. Shen, G. Xie, X. Zhang, and J. Xu. 2024. On the Decision-Making Abilities in Role-Playing using Large Language Models. *arXiv preprint arXiv:2402.18807*.

X. Zhou, H. Zhu, L. Mathur, R. Zhang, H. Yu, Z. Qi, L.-P. Morency, Y. Bisk, D. Fried, G. Neubig, *et al.* 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Xueyang Feng and Zhi-Yuan Chen and Yujia Qin and Yankai Lin and Xu Chen and Zhiyuan Liu and Ji-Rong Wen, *et al.* 2024. Large Language Model-based Human-Agent Collaboration for Complex Task Solving. *arXiv preprint arXiv:2402.12914*.

N. Li, C. Gao, Y. Li, and Q. Liao. 2023. Large language model-empowered agents for simulating macroeconomic activities. *arXiv preprint arXiv:2310.10436*.

Z. Huang, S. Gutierrez, H. Kamana, and S. MacNeil. 2023. Memory sandbox: Transparent and interactive memory management for conversational agents. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–3.

L. Liu, X. Yang, Y. Shen, B. Hu, Z. Zhang, J. Gu, and G. Zhang. 2023. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2311.08719*.

Jia Li and Ge Li and Yongmin Li and Zhi Jin 2023. Structured Chain-of-Thought Prompting for Code Generation. *arXiv preprint arXiv:2305.06599*.

Yilun Du and Shuang Li and Antonio Torralba and Joshua B. Tenenbaum and Igor Mordatch. 2023. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *arXiv preprint arXiv:2305.14325*.

Rafael Rafailov and Archit Sharma and Eric Mitchell and Stefano Ermon and Christopher D. Manning and Chelsea Finn. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*.

B. Xu, A. Yang, J. Lin, Q. Wang, C. Zhou, Y. Zhang, and Z. Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

J. Lu, S. An, M. Lin, G. Pergola, Y. He, D. Yin, X. Sun, and Y. Wu. 2023. Memochat: Tuning LLMs to use memos for consistent long-range open-domain conversation. *arXiv preprint arXiv:2308.08239*.

Y. Li, Y. Yu, H. Li, Z. Chen, and K. Khashanah. 2023. TradingGPT: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance. *arXiv preprint arXiv:2309.03736*.

J. Li, S. Wang, M. Zhang, W. Li, Y. Lai, X. Kang, M. Ma, and Y. Liu. 2024. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*.

C. Qian, S. Liang, Y. Qin, Y. Ye, X. Cong, Y. Lin, Y. Wu, Z. Liu, and M. Sun. 2024. Investigate-consolidate-exploit: A general strategy for inter-task agent self-evolution. *arXiv preprint arXiv:2401.13996*.

S. Yuan, K. Song, J. Chen, X. Tan, D. Li, and D. Yang. 2024. EvoAgent: Towards Automatic Multi-Agent Generation via Evolutionary Algorithms. *arXiv preprint arXiv:2406.14228*.

浅谈大模型时代下的检索增强：发展趋势、挑战与展望

冯掌印[1], **朱坤**[1], **马伟涛**[1], **黄磊**[1], **秦兵**[1,2], **刘挺**[1,2], **冯骁骋**[1,2*]
[1]哈尔滨工业大学，社会计算与信息检索研究中心，哈尔滨，150006
[2] 鹏城实验室，深圳，518055
{zyfeng,kzhu,wtma,lhuang,qinb,tliu,xcfeng*}@ir.hit.edu.cn

摘要

大型语言模型(LLM) 在各种自然语言任务上表现出了卓越的性能，但它们很容易受到过时数据和特定领域限制的影响。为了应对这些挑战，研究人员整合不同来源的外部信息来增强大语言模型，具体方法如检索增强等。在本文中，我们综合讨论了检索增强技术的发展趋势，包括检索时机规划、检索技术、以及检索结果的利用。此外，我们介绍了当前可用于检索增强任务的数据集和评价方法，并指出了应用和潜在研究方向。我们希望这项综述能够为社区提供对该研究领域的快速了解和全面概述，以启发未来的研究工作。

关键词： 检索增强

Enhancing Large Language Models with Retrieval-Augmented Techniques: Trends, Challenges, and Prospects

Zhangyin Feng[1], **Kun Zhu**[1], **Weitao Ma**[1], **Lei Huang**[1],
Bing Qin[1,2], **Ting Liu**[1,2], **Xiaocheng Feng**[1,2*]
[1]Research Center for Social Computing and Information Retrieval,
Harbin Institute of Technology, Harbin, 150006
[2] Peng Cheng Laboratory, Shenzhen, 518055
{zyfeng,kzhu,wtma,lhuang,qinb,tliu,xcfeng*}@ir.hit.edu.cn

Abstract

Large language models (LLMs) have demonstrated outstanding performance in various natural language tasks, but they are susceptible to issues caused by outdated data and specific domain limitations. To address these challenges, researchers have integrated external information from different sources to enhance the capabilities of large language models, including retrieval-augmented generation methods. In this paper, we present a review to discuss the development trends of retrieval-augmented techniques, including retrieval timing strategies, retrieval paradigms, and utilization of retrieval results. Additionally, we introduce the datasets and evaluation methods currently available for retrieval-augmented tasks, and highlight applications and future potential research directions. We hope that this survey will provide the community with quick access and a comprehensive overview of this research area, to inspire future research efforts.

Keywords: retrieval-augmented

*通讯作者

第二十三届中国计算语言学大会论文集，第151页-第168页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

大规模预训练语言模型已表现出将现实世界知识编码到参数中的强大潜力，以及解决各种自然语言处理任务的非凡能力 (Brown et al., 2020; Hoffmann et al., 2022; Zeng et al., 2022; Chowdhery et al., 2022; Touvron et al., 2023; Zhao et al., 2023b)。然而，在面对需要大量现实世界知识作为参考的知识密集型任务时 (Petroni et al., 2021)，大模型仍然面临着严峻的挑战。近期研究表明，大语言模型很难学习长尾知识 (Kandpal et al., 2023; Mallen et al., 2023)，并且无法及时更新参数以捕捉不断变化的世界 (De Cao et al., 2021; Kasai et al., 2024)（例如，ChatGPT[0]的参数仅包含2021年9月之前的信息，完全不了解最新的世界知识。）。此外，大模型还会遇到幻觉问题 (Zhang et al., 2023; Rawte et al., 2023; Huang et al., 2023)。为了缓解这些问题，大量研究尝试利用检索外部知识的手段增强语言模型的知识能力 (Mallen et al., 2023; Shi et al., 2023b; Trivedi et al., 2023)，这类方法一般采用现成的检索模型从外部语料库中获取相关文档，以帮助大型语言模型更好地进行内容生成。

检索增强技术能够在推理阶段以非参数化形式利用外部知识，其框架通常由检索器和生成器组成。当前已有工作探索了以端到端方式训练整个检索器-语言模型系统的不同方法：使用检索增强序列对数似然 (Lewis et al., 2020b; Borgeaud et al., 2022)、解码器中融合注意力蒸馏 (Borgeaud et al., 2022; Izacard et al., 2023)或知识图 (Ju et al., 2022)。当越来越多的独特需求出现时，这种微调的成本可能会很高 (Maronikolakis and Schütze, 2021)。更重要的是，许多大型语言模型只能通过黑盒API 访问 (Ouyang et al., 2022; Achiam et al., 2023)。这些API 允许用户提交查询并接收响应，但通常不支持微调。

在本文中，我们以检索策略和内容处理策略为中心，整理现有的检索增强技术。我们先规范化检索增强技术的范式（§2），然后从检索前、检索时和检索后三个阶段总结现有的检索增强方法会面临的三个重要问题：（§3）什么时候需要通过检索来增强大型语言模型？（§4）如何准确高效地检索得到相关信息？（§5）如何利用检索得到的信息和知识优化大模型生成的内容？此外，我们介绍了当前适用于检索增强任务的数据集（§6）和评价方法（§7），并提供了一些对未来研究的前瞻性思考以促进该领域的进一步发展（§8）。

2 检索增强定义

当前文本生成框架遵循以下范式：给定输入上文信息 x，θ 参数化的预训练语言模型通过概率建模 $p_\theta(y|x) = \prod_i p_\theta(y_i|y_{<i}, x)$ 生成相关内容 y。由于大规模预训练语言模型的 θ 参数难以跟随现实的变化实时更新，检索增强技术则尝试在生成范式中添加辅助的检索知识 $p_\theta(y|x, d) = \prod_i p_\theta(y_i|y_{<i}, x, d)$ 以扩展语言模型的能力，其中 d 是从外部知识或语料库 $d_1, \ldots, d_n \in D$ 中获取的辅助信息。检索增强的工作流程可以简单举例阐述：假设用户向预训练大模型 p_θ 询问关于最近实事动态 x 的评论，检索增强系统需要将 x 转写成合适的查询 q 并以此从最新的外部知识库 D 中来检索相关信息 $[d_1, \ldots, d_i]$，进行必要的处理后作为 d 交给语言模型生成对应的回复。这样一种检索生成范式需要面对的问题主要有三点：

1. 检索的时机：面对当前上文信息 x 时，语言模型 p_θ 的能力是否足够生成所需内容，是否需要进行检索补充信息？即需要能够提前估计生成内容 $y \sim p_\theta(y|x)$ 的质量。

2. 检索的策略：如何将当前上文 x 转变为有效的查询 q，并快速准确地从外部知识库 D 中获取相关文档 $[d_1, \ldots, d_i]$？

3. 检索结果的使用：如何将相关文档 $[d_1, \ldots, d_i]$ 转换为适合语言模型 p_θ 理解与使用的辅助信息 d？

3 检索时机规划

检索时机规划是指在检索开始之前，模型发起检索请求是被动还是主动。我们称预先设定好的检索流程为被动规划，基于大模型反馈的检索请求发起为主动检索。由于检索算法和检索数据来源的限制，检索得到的内容并不是百分百相关且准确。Shi et al. (2023a) 和 Wu et

[0] https://chat.openai.com

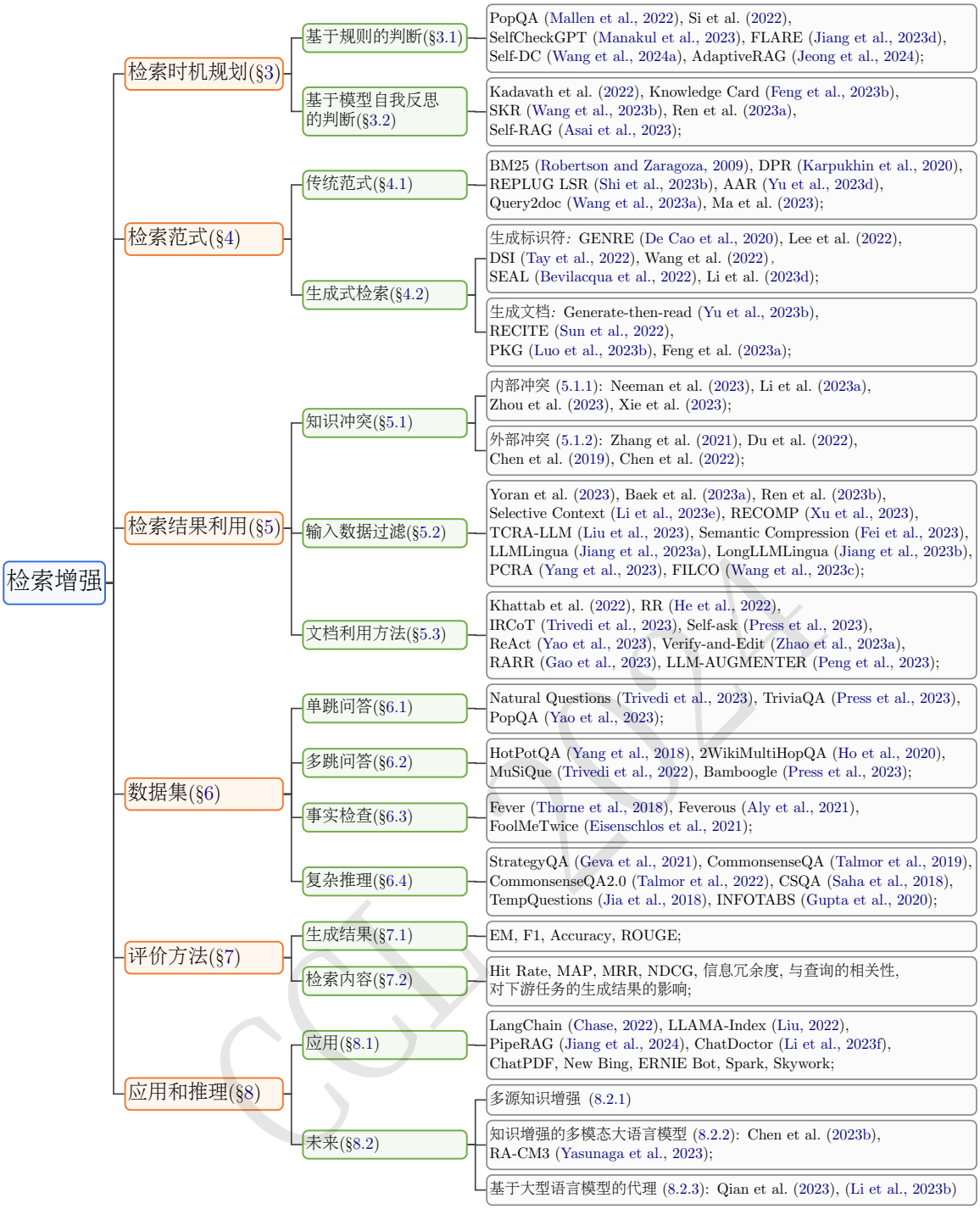


Figure 1: 以检索策略和内容处理策略为中心的检索增强技术分类

al. (2024) 的研究表明：大模型解决问题的能力很容易被输入上下文中不相关但具有误导性的内容干扰。Chen et al. (2023a) 认为，虽然大模型具有一定程度的噪声鲁棒性，但它们在负面信息过滤、信息整合和处理虚假信息方面仍有很大的困难。因此当模型固有的参数化知识足以回答相关问题时，过度检索并不能为最终结果带来增益。由此，本章将深入研究主动检索方法。

对于主动检索来说，一个非常重要的问题是了解大模型的知识边界 (Yin et al., 2023a)并确定何时检索补充知识。根据判断知识边界的方法可将目前的主动检索判断为两类：基于规则的判断和基于模型自我反思的判断。

3.1 基于规则的判断

一个简单直观的想法是设置一个指标和阈值。当指标高于或低于阈值时，我们触发检索器来获取相关文档。Kandpal et al. (2022)研究大型语言模型记忆的知识与预训练数据集的信息之间的关系。他们观察到在某些问答数据集上的准确率和相关文档数量之间存在很强的相关性和因果关系，因此得到结论：语言模型回答基于事实的问题的能力与预训练期间看到的与该问题相关的文档数量有关。为了深入分析大模型的参数化知识与数据流行度之间的关系，Mallen et al. (2022)构建一个开放域问答数据集PopQA，其中包含来自维基百科的实体流行度。然后，他们设计了一种自适应检索方法，仅对流行度低于流行度阈值的问题使用检索。除了流行度之外，Jiang et al. (2020)表明大模型往往经过良好校准，低概率或置信度通常表明缺乏相关知识。Si et al. (2022)和 Manakul et al. (2023)利用词语概率来指示其输出的不确定性。遵循这个想法，Jiang et al. (2023d)提出了一种基于置信度的主动检索方法，名为FLARE。如果生成的句子中每个单词的置信度高于阈值，则它们接受该句子而不检索附加信息。否则，它们主动触发检索并利用检索到的相关信息重新生成当前句子。Self-DC (Wang et al., 2024a)则将模型响应的置信度得分划分为三组：未知、不确定和已知。引导大模型按需自适应性地调用不同的方法，被视为未知的查询通过顺序检索增强流水线进行处理，而那些具有不确定性的查询则被分解为子问题以生成答案。AdaptiveRAG (Jeong et al., 2024)则训练一个较小的语言模型作为分类器，判断查询的复杂性，动态决定是否检索，

3.2 基于模型自我反思的判断

基于模型自我反思的判断方法即让大模型根据问题以生成的方式判断是否触发检索。考虑到大语言模型具有非常强大的能力，一些研究人员直接使用大语言模型来确定是否需要检索。Yin et al. (2023b)通过评估大模型识别无法回答或不可知问题的能力来研究他们的自我认知边界。Kadavathet al. (2022)提示大模型预测他们的回答是否可靠的概率。这些不可靠的回答表明大模型需要额外的信息来回答相应的问题。Feng et al. (2023b)询问大模型“您需要更多信息吗？（是或否）”来通过情境学习来确定给定问题是否需要外部知识。SKR (Wang et al., 2023b)构建了一个二元分类数据集，给定问题输出是否可回答。该数据集用于训练小模型或引导大模型情境学习。Ren et al. (2023a)采用先验和后验判断指令来调查大模型是否能够在正常设置和检索设置下感知自己的事实知识边界。先验判断询问大模型是否可以提供问题的答案。事后判断要求大模型评估问题答案的正确性。他们观察到大模型对他们的事实知识边界的认识不准确，并且在正常情况下有过度自信的倾向。Self-RAG (Asai et al., 2023)训练生成器来直接生成检索标记来确定是否执行检索。

4 检索技术

在检索增强系统中，外部知识的质量和语言模型的能力共同决定了生成内容的上限。在当前的语言模型规模和预训练开销下，微调语言模型以追踪最新的知识成本过高难以实现。而利用高质量实时更新的外部知识则是目前最为有效的一种手段。获取得到外部知识的质量主要受到两方面影响，一方面是检索数据源的质量，另一方面是检索方法的性能。传统的检索范式是使用检索器从外部语料库（例如维基百科、知识图谱、网络文本等）获取相关文档。最近有一种观点认为，既然大型语言模型可将现实世界知识编码到其参数中，那么是否可以通过提示模型输出相关的内部知识以获得更准确内容。本章将从传统检索范式和生成式检索两方面介绍现有检索范式。

4.1 传统检索

给定输入上下文 x，检索器旨在从语料库 $D = d_1, \ldots, d_m$ 中检索与 x 相关的一小组文档。检索器有不同类型，包括基于术语的稀疏检索器、基于向量表示的稠密检索器和商业搜索引擎。稀疏检索器通常使用 TF-IDF 或 BM25 实现 (Robertson and Zaragoza, 2009)，它通过倒排索引有效地匹配关键字。然而，术语匹配方法对高度选择性的关键词和短语很敏感。稠密检索器 (Karpukhin et al., 2020)将文本编码到连续的稠密语义空间中，其中由完全不同的词语组成的同义词或复述仍然可以映射到彼此接近的向量。商业搜索引擎，例如谷歌和百度，是能够检索最新世界知识的复杂系统。这三种方法各有不同的优势和应用场景，接下来我们重点关注稠密检索器。

给定文本段落的集合，稠密检索器的目标是在低维连续空间中索引所有段落，以便它可以在运行时有效地检索与读者输入问题相关的前 k 个段落。稠密检索器使用稠密编码器 $E(\cdot)$，它将任何文本段落映射到 d 维实值向量。具体来说，编码器通过对 d 中词语的最后一个隐藏表示进行平均池化，将每个文档 $d \in D$ 映射到向量表示 $E(d)$。在查询时，将相同的编码器应用于输入上下文 q 以获得查询向量表示 $E(q)$。查询向量表示和文档向量表示之间的相似度通过它们的余弦相似度计算:

$$s(d, q) = cos(E(d), E(q))$$

通过以上步骤检索获得与输入 q 具有最高相似度分数的前 k 个文档。

由于微调大型语言模型的资源、成本以及黑盒限制，先前使语言模型适应检索器的工作已不适用于当前，最近的工作尝试使检索器适应语言模型。REPLUG LSR (Shi et al., 2023b)利用来自黑盒语言模型的监督信号进一步改进了 REPLUG 中的初始检索模型，即GPT-3 Curie (Brown et al., 2020)。AAR (Yu et al., 2023d)建议利用小规模语言模型为检索器训练提供语言模型偏好的信号。训练后的检索器可以通过插入检索到的文档来直接用于辅助大型目标语言模型。

与之前专注于调整检索器的研究不同，另一个研究方向侧重于弥合输入文本和查询真正需要的知识之间的语义差距。Query2doc (Wang et al., 2023a)通过采用几次提示范例来提示大模型生成伪文档。随后，通过合并伪文档来扩展原始查询。检索器模块使用这个新查询来检索相关文档的列表。Ma et al. (2023)引入了用于检索增强的重写-检索-读取框架，可以进一步调整该框架以适应大模型。他们还在检索器之前添加了查询重写步骤。与 Query2doc 不同的是，它们采用可训练的语言模型来执行重写步骤。重写语言模型通过强化学习进行训练，以 LLM 表现作为奖励。

4.2 生成式检索

生成式检索是一种新的检索范式，主要包括生成文档标识符字符串和生成完整文档两种。

前者使用标识符来减少无用信息量，使模型更容易记忆和学习 (Li et al., 2023d)。De Cao et al. (2020)提出 GENRE，它通过生成实体文本本身来检索实体。GENRE 还可以应用于页面级检索，其中每个文档都包含唯一的标题作为标识符。Lee et al. (2022)将生成检索引入多跳设置，检索的项目是短句。Tay et al. (2022)提出了 DSI 方法，该方法采用数字 ID 作为文档的标识符。Wang et al. (2022)通过生成更多查询作为额外的训练数据来改进 DSI。然而，基于 Id 的数值方法通常在小型数据集上进行评估，部分原因是它们面临大规模扩展问题。Bevilacqua et al. (2022)提出 SEAL，它采用子字符串作为标识符。检索过程是在 FM-Index结构上有效完成的。Li et al. (2023d)提出了多视图标识符，从不同的角度代表一段段落，以增强生成检索并实现最先进的性能。

后者的目标不是生成标识符，而是使用大型语言模型直接生成完整的文档。Generate-then-read (Yu et al., 2023b)表明，生成的上下文文档比检索到的最热门文档更频繁地包含正确答案，并且显著优于直接从大型语言模型生成答案，尽管没有包含任何新的外部信息。RECITE (Sun et al., 2022)采用了类似的方法，它通过首先“背诵”相关信息然后生成输出来处理知识稠密型 NLP 任务。PKG (Luo et al., 2023b)为大模型配备了背景知识生成模块来获取相关知识。

考虑到构建参数化知识模块的知识时效性和微调模型成本，可通过离线高效微调开源小语言模型来存储任何知识。Feng et al. (2023a)建议通过集成专业语言模型，为通用大型语言模型提供模块化和协作来源的知识。专业语言模型是在来自不同来源和领域的语料库上进行训练的。他们还提出了三个级别的知识过滤器，以动态选择和细化生成的文档，并控制主题相关性、文档简洁性和知识真实性。

总结：传统检索范式使用检索器从外部语料库中获取相关文档。然而，检索到的文档可能包含与问题无关的嘈杂信息。另一种选择是使用大型语言模型直接生成相关文档，但无法获得实时信息。所以我们应该根据实际场景做出合适的选择。

5 检索结果利用

在检索得到相关知识和信息之后，如何更有效地利用这些信息，辅助语言模型生成更高质量的内容则至关重要。在语言模型进行生成之前，检索知识的质量需要严格把关，可利用噪声过滤、思维链推理扩展、重新检索等手段进行输入过滤。同时，我们也需要探索如何让语言模

型接受并理解检索知识，并利用该知识进行生成。在语言模型生成之后，检索知识同样可以作为关键信息修正语言模型等生成结果。

5.1 知识冲突

在检索增强的大模型中，有两种知识来源有助于模型推理，但分工不明确且不透明。第一个是通过预训练和微调灌输的隐式参数知识（即它们学习的权重）。第二个是上下文知识，通常来源于检索器的文本段落。知识冲突是指所包含的信息不一致、矛盾。知识冲突有两种类型：内部冲突和外部冲突。内部冲突是指大型语言模型中的知识与检索到的文档中的知识之间的不一致。外部冲突是指检索到的多个文档之间不一致。

5.1.1 内部冲突

随着世界不断发展，记忆的事实可能会变得过时 (Liška et al., 2022; Kasai et al., 2024)。利用包含相关知识的外部背景来增强大模型是一个有前途的方向。然而，此类方法面临的挑战是大模型可能会坚持记住的事实并忽略所提供的上下文 (Longpre et al., 2021)。为了应对这一挑战，最近的著作 (Neeman et al., 2023; Li et al., 2023a)对反事实背景下的大模型进行了微调，其中原始事实被反事实的内容所取代。他们发现这种微调过程可以有效提高大模型对上下文的利用率，而不是仅仅依赖他们的参数知识。Zhou et al. (2023)提出了一种使用提示来提高大模型上下文忠实度的方法，无需额外的微调，这为大模型提供了一种更通用、更具成本效益的方法。他们提出了各种提示策略来提高大模型的忠诚度，包括设计有效的提示和选择适当的上下文演示。Xie et al. (2023)对大模型在遇到反记忆时的行为进行了全面和受控的调查。他们发现，大模型的参数记忆既有支持性的证据，也有矛盾的证据，他们表现出强烈的确认偏差，并倾向于坚持他们的参数记忆。

5.1.2 外部冲突

这种情况在某些段落已用新信息更新而其他段落仍然过时的环境中很常见 (Zhang and Choi, 2021)。当段落被敌对地编辑为包含虚假信息时 (Du et al., 2022)，或者当段落由对答案有不同意见的多人撰写时 (Chen et al., 2019)，也会发生此类冲突。Chen et al. (2022)模拟一种设置，其中证据段落的子集受到干扰，以提出不同的答案，以反映检索返回混合信息包的现实场景。他们发现，当不同的段落提出多个相互冲突的答案时，模型更喜欢与其参数知识相匹配的答案。除了分析简单的内部和外部冲突之外，Xie et al. (2023)还对更复杂的知识冲突场景进行了实验。在提供相关和不相关证据的情况下，大模型可以在一定程度上过滤掉不相关的证据。然而，随着不相关证据数量的增加，这种能力就会减弱。

综上，知识冲突是一个非常重要的问题。然而，目前的研究主要集中在知识冲突的分析上。下一步应该从不同方面解决知识冲突，比如输入数据过滤、改进文档利用方法等。

5.2 输入数据过滤

对于输入过滤，最直觉的方法是首先检索所有问题的相关文档，然后判断相关文档是否可以回答问题。如果相关，大型语言模型会利用检索到的文档来生成答案。如果不相关，大型语言模型会直接生成答案。Yoran et al. (2023)将检索到的文档和问题的相关性判断视为自然语言推理(NLI) 问题 (Dagan et al., 2005; Bowman et al., 2015)，并使用训练有素的 BART-Large NLI (Lewis et al., 2020a)模型来识别不相关的检索文档。检索到的文档作为前提，而问题和生成的答案连接起来作为假设。Baek et al. (2023b)建议使用指令数据对大模型进行微调，以识别输入问题和检索到的文档之间的相关性，并整合各种指令的结果，以进一步提高准确性。Ren et al. (2023b)发现，纳入相关文献后，大模型自我评估的准确性有所提高，动态引入检索文献对大模型来说是有效的。

除了二元判断检索得到的文档是否有帮助外，另一种思考输入过滤方法的角度是对检索得到的内容进行文本摘要，保留其中与目标答案最相关的部分。文本摘要可分为抽取式摘要和生成式摘要。抽取式摘要从检索结果中删除不重要的词语、句子或文档，保留关键句和关键词组成摘要。生成式摘要则根据原文，生成新的词语、短语来组成摘要。Selective Context (Li et al., 2023e)将词语合并为单元，然后基于自信息指标（即负对数似然）应用单元级提示剪枝。RECOMP (Xu et al., 2023)基于表示相似度构造抽取式摘要数据训练 BERT，使用 ChatGPT 生成式摘要数据微调 T5-Large。TCRA-LLM (Liu et al., 2023)则通过去掉句

子中语义影响最小的词，基于语义压缩算法构造抽取式数据；利用 ChatGPT 构造不同长度的生成式摘要构造生成式数据。SemanticCompression (Fei et al., 2023)提出了一种语义压缩方法。它首先将文本分解成句子。接下来，它按主题将句子分组，然后总结每组内的句子。LLMLingua (Jiang et al., 2023a)引入了一种从粗到细的方案进行过滤。最初，它执行段落级别删减，然后根据生成模型提供的困惑度计算每个词语的重要性，同时保留重要的词语。为了提高性能，LLMLingua 还提出了一个预算控制器，可以在输入提示的不同部分动态分配修剪预算。LongLLMLingua (Jiang et al., 2023b)基于 LLMLingua 构建，引入了查询感知压缩，并根据计算的重要性得分对检索到的文档进行重新排序。PCRA (Yang et al., 2023)则利用强化学习范式，基于生成结果的反馈约束优化过检索内容过滤器，使其适用于黑盒大模型和不同检索器。FILCO (Wang et al., 2023c)基于后验概率分布预测先验分布，即标注对生成标准答案影响最大的句子作为目的摘要。Zhu et al. (2024)则从信息瓶颈的角度综合后验概率分布和内容冗余程度两方面评估检索上下文质量。

5.3 文档利用方法

有了相关文档后，我们该如何利用它们来提高大语言模型的能力？一种常见的方法是将检索到的文档添加到原始输入中，再将其输入到大型语言模型中以进行最终预测 (Khattab et al., 2022; Yu et al., 2023b; Luo et al., 2023b; Feng et al., 2023b)。大型语言模型能够通过生成逐步的自然语言推理步骤（称为思维链（CoT））来回答复杂的问题 (Wei et al., 2022)。由此引申出将思维链和检索过程相结合来推理复杂问题，通过反复循环检索和生成来逐步细化结果。

He et al. (2022) 提出了一种称为检索重新思考(RR) 的后处理方法，用于在大模型中利用外部知识。他们首先使用思想链（CoT）提示方法来生成一组多样化的推理路径。然后，他们使用这些路径中的每个推理步骤来检索相关的外部知识，这使得RR 能够提供更忠实的解释和更准确的预测。IRCoT (Trivedi et al., 2023)提出了一种交错方法，使用检索来指导思想链推理步骤，并使用 CoT 推理来指导检索。Self-ask (Press et al., 2023) 建立在思想链提示的基础上，但不是输出一个连续的、无界限的思想链。Self-ask 清楚地划分了每个子问题的开始和结束，并使用搜索引擎来回答子问题。Luo et al. (2023a) 构建 CoT 数据集并训练模型在生成阶段首先输出对检索上下文的相关性分析，再回答问题。

ReAct (Yao et al., 2023) 促使大模型以交错的方式生成言语推理轨迹和动作，这使得模型能够执行动态推理来创建、维护和调整高级行动计划，同时还与外部环境将附加信息纳入推理。Verify-and-Edit (Zhao et al., 2023a)旨在通过根据外部知识对推理链进行后期编辑来提高预测的真实性。Self-RAG (Asai et al., 2023)提出“自我反思检索-增强生成”框架。首先训练一个可按需自适应性地生成特殊检索标记的语言模型，获得检索内容后令生成模型反思检索到的段落及其自身的生成，基于最优结果继续生成内容。此外，最近的几项工作 (Shao et al., 2023; Feng et al., 2024; Yu et al., 2023c; Cheng et al., 2024; Wang et al., 2024b)首先生成初始输出，然后基于该输出利用检索模型从大型文档集合中获取相关信息，最后合并将检索到的信息放入输入上下文中以进行输出细化，迭代多次。

RARR (Gao et al., 2023)提出了一种与模型无关的方法来改进任何现有 LM 的归因，而不是限制 LM 生成归因文本。生成文本后，RARR 获取相关证据，然后修改文本以使其与证据一致，同时保留风格或结构等质量，使修改后的文本能够无缝地代替原始文本。RARR 可以被视为增强检索模型，其中检索发生在生成之后而不是生成之前。Peng et al. (2023) 提出 LLM-AUGMENTER，通过外部知识和自动反馈来改进大模型。给定用户查询，LLMAUGMENTER 首先从外部知识中检索证据，并通过将检索到的原始证据与相关上下文联系起来并执行推理以形成证据链来进一步巩固证据。然后，LLM-AUGMENTER 通过检查候选人的回答是否产生幻觉来验证其回答。

综上，可将使用相关文档的时机分为三个不同阶段：使用相关文档作为输入提示的一部分，使用相关文档来确保推理过程的正确性，利用相关文档修改大语言模型的原始答案。三种阶段的方法可综合使用。

6 数据集

因为解决知识密集型任务需要访问大量信息，所以知识密集型任务非常适合评估检索增强型大型语言模型。我们详细研究了常用的数据集，并按任务类型将它们分为以下几类。

6.1 单跳问答

单跳问题的结构相对简单，可以使用段落中包含的信息来回答。常用的数据集有 Natural Questions (NQ)、TriviaQA 和PopQA。Natural Questions (Kwiatkowski et al., 2019a)是由从 Google 搜索引擎聚合的问题组成，答案由专家人工注释。TriviaQA (Joshi et al., 2017)由琐事爱好者撰写的问题组成，证据收集自维基百科和网络。PopQA (Mallen et al., 2022)是一个大规模的以实体为中心的问答数据集，其构造时采样了更多长尾数据，并且明显包含更多的低受欢迎度实体。

6.2 多跳问答

多跳问答是无法通过单一来源或段落直接解决的问题，模型需要执行多个推理步骤才能回答问题。常用的数据集有 HotPotQA、2WikiMultiHopQA、MuSiQue和Bamboogle。HotPotQA (Yang et al., 2018)收集自明确需要对多个支持上下文文档进行推理的问题。2WikiMultihopQA (Ho et al., 2020)也是通过组合构建的，但它们使用一组有限的人工编写的组合规则。MuSiQue (Trivedi et al., 2022)是通过仔细选择和编写单跳问题以自下而上的过程构建的。MuSiQue 具有六种组成结构，比 HotPotQA 和 2WikiMultihopQA 更具挑战性且不易作弊。Bamboogle (Press et al., 2023)是一个由作者编写的包含2 跳问题的小型数据集，其中所有问题都足够困难，以确保常用的互联网搜索引擎无法回答，但所需的两个支持证据都可以在维基百科中找到。

6.3 事实检查

事实验证，也称为事实检查，需要从纯文本中检索相关证据并使用这些证据来验证给定的主张。常用的数据集有 Fever、Feverous 和 FoolMeTwice (FM2)。Fever (Thorne et al., 2018)是一个用于事实验证的大型数据集，需要检索句子级证据来支持某个主张是否得到支持或反驳。除了非结构化文本证据之外，Feverous (Aly et al., 2021)还将维基百科表格视为一种证据形式。Feverous 中的证据检索考虑了维基百科文章的整体，因此证据可以位于文章中除参考部分之外的任何部分。Feverous 的证据分布平衡，要么仅包含文本、表格，要么两者都作为证据，两种情况的实例数量几乎相同。FoolMeTwice (Eisenschlos et al., 2021)是通过一个有趣的多人游戏收集的数据集。该游戏鼓励对抗性示例，其中可以使用捷径解决的实例数量远少于其他数据集。

6.4 复杂推理

复杂推理包括不同类型的推理，如常识推理、表格推理等。常识推理是人类认识的基础，植根于日常生活和社会实践中积累的基础知识和生活经验，概述了世界如何运转的实用知识 (Sap et al., 2020)。常识推理任务评估模型在物理世界中的推理能力。StrategyQA、CommonsenseQA 和 CommonsenseQA2.0 是广泛使用的常识推理数据集。StrategyQA (Geva et al., 2021)是一个专注于开放领域问题的问答基准，其中所需的推理步骤隐含在问题中，应使用策略来推断。CommonsenseQA (Talmor et al., 2019)和CommonsenseQA2.0 (Talmor et al., 2022)的提出是为了探索大型语言模型的常识理解能力，其中包括关于日常常识知识的是/否问题（或论断）。CSQA (Saha et al., 2018)是一个要求长输出的问答数据集，旨在为寻求复杂信息的问题生成全面的答案。TempQuestions (Jia et al., 2018)旨在研究时间推理。该数据集包含1,271 个时间问题，分为四类：显式时间、隐式时间、时间答案和序数约束。INFOTAB (Gupta et al., 2020)由23, 738 个人工编写的文本假设组成，这些假设以从维基百科信息框提取的表格内容为前提。

6.5 其他

除了以上传统任务外，最近也有一些工作针对检索增强任务的特点专门构建数据集。检索增强范式在处理时效性强的任务上具有天然优势，但由于生成模型的训练语料库中可能包含早期数据集的背景知识，致使模型不需要外部知识即可回答问题，为了准确评估检索增强系统的时效性能力，RealTime QA (Kasai et al., 2024)创建了一个动态问答平台，每周定期发布关于最新事件或信息的问题。针对检索上下文中可能包含的噪声、事实冲突，Chen et al. (2024)提出不同大语言模型在检索增强生成中应具有4种基本能力：噪声鲁棒性、负面拒绝、信息集成和反事实鲁棒性，并为此构建一个中英文的新数据集。

任务类型	数据集	评价指标
单跳问答	Natural Questions (Kwiatkowski et al., 2019b)	EM / F_1
	TriviaQA (Joshi et al., 2017)	EM / F_1
	PopQA (Mallen et al., 2023)	Accuracy
多跳问答	2WikiMultiHopQA (Ho et al., 2020)	EM / F_1
	HotPotQA (Yang et al., 2018)	EM / F_1
	MuSiQue (Trivedi et al., 2022)	EM / F_1
	Bamboogle (Press et al., 2023)	EM / F_1
事实检查	Fever (Thorne et al., 2018)	Accuracy
	Feverous (Aly et al., 2021)	Accuracy
	FoolMeTwice (Eisenschlos et al., 2021)	Accuracy
复杂推理	StrategyQA (Geva et al., 2021)	Accuracy
	CommonsenseQA (Talmor et al., 2019)	Accuracy
	CommonsenseQA2.0 (Talmor et al., 2022)	Accuracy
	CSQA (Saha et al., 2018)	EM / ROUGE
	TempQuestions (Jia et al., 2018)	EM / F_1
	INFOTABS (Gupta et al., 2020)	EM / F_1

Table 1: 不同任务下数据集及相应评估方法

7 评价方法

检索增强技术有效性的评估主要从两个角度进行。最直观的方式是直接使用具体下游任务的评价方法，如问答任务上的EM、F1，事实检查任务上的Accuracy，生成任务上的ROUGE等。然而仅依靠端到端的评价结果无法全面评估检索增强系统中各组件的性能，我们需要从更细致的角度来衡量检索增强的有效性(Hoshi et al., 2023)。

7.1 生成结果

针对有标准答案的生成结果，表1中展示了第6章中提及的数据集对应的评价方法。而对于没有具体标准答案的生成结果，可针对具体任务采用不同角度的人工或自动评价，如流畅性、连贯性、事实一致性等模型生成能力指标。具体方法与对应任务相关，本文不作深入探讨。

7.2 检索内容

如何评估检索得到内容的质量是当前的难点，关键点在于如何定义检索上下文的有效性。传统搜索引擎的评价方式有命中率（Hit Rate）、平均准确率（Mean Average Precision，MAP）、平均倒数排名（Mean reciprocal rank，MRR）和归一化折损累积增益（Normalized Discounted Cumulative Gain，NDCG）等指标。然而针对检索增强系统，检索上下文的有效性可取决于其信息冗余度、与查询的相关性、对下游任务的生成结果的影响等。综上，我们总结第5.2节中的过滤方法，将过滤检索内容的监督信号总结为以下三种：

- 对于整体语义的贡献：例如，删去不影响句子语义的词语，删去不影响篇章语义的句子。
- 基于向量表示的相似性：例如，保留与查询问题相似度更高的检索信息，对检索信息间相似性高的内容进行酌情删减。
- 对于生成结果的影响程度：例如，保留对生成结果困惑度影响程度高的文章或句子。

8 应用和未来

8.1 应用

事实证明，使用从各种知识存储中检索的相关信息来增强语言模型可以有效提高各种知识密集型任务的性能。在开放域问答和事实验证中，模型可以通过在大型语料库或网络上搜

索相关文档来更准确地回答问题。除了经典的自然语言处理任务之外，随着检索增强大型语言模型的发展，还出现了许多新的应用程序。LangChain (Chase, 2022)是一个功能强大的框架，提供了一组工具、组件和接口来简化创建由大型语言模型和聊天模型支持的应用程序的过程。LangChain 可以轻松管理与大型语言模型的交互、将多个组件链接在一起并集成其他资源。类似的工作有 LLAMA-Index (Liu, 2022) 和 PipeRAG (Jiang et al., 2024)。ChatPDF[1]是一款帮助用户理解 PDF 文档并与之聊天的 AI 工具。它可以识别关键信息、提供简洁的摘要并回答您的问题。ChatDoctor (Li et al., 2023f)是一种专为医疗应用而设计的高级语言模型。患者可以通过聊天界面与 ChatDoctor 模型进行交互，询问有关他们的健康、症状或医疗状况的问题。然后，该模型将分析输入并提供适合患者独特情况的响应。New Bing 通过将 ChatGPT 与微软的搜索引擎相结合来使用检索增强。新的 Bing Chat 根据提示生成搜索查询，检索相关文档，并将它们用作结果的上下文。新必应还提供其生成的句子的信息源链接。综上所述，大语言模型具有更强大的知识理解能力，以及通过检索相关文档进行推理的能力，将会有更多的应用场景。百度、科大讯飞和昆仑万维也提供类似的服务，例如 ERNIE Bot[2]、Spark[3]和 Skywork[4]。

8.2 未来

检索增强的发展仍处于初级阶段，因此还有很大的改进空间。在本节中，我们对未来的研究进行简要概述。

8.2.1 多源知识增强

现有的知识增强方法主要在整合知识的格式和种类方面表现出局限性。当前的检索增强方法主要集中于维基百科或网络的非结构化文本检索 (Shi et al., 2023b; Feng et al., 2023b; Vu et al., 2023)。虽然有诸如 KAPING (Baek et al., 2023a)（知识图谱）、StructGPT(Jiang et al., 2023c)（数据库）等方法探索了基于大模型从结构化文本进行检索以完成增强任务的方法。但在现实场景中，一个复杂的问题可能需要从不同来源收集零散的证据才能得出最终答案。不同来源和不同格式的证据对大型语言模型的影响值得探讨。此外，找到一种合适的方法来整合不同来源的证据也极其重要。

8.2.2 知识增强的多模态大语言模型

多模态学习作为视觉到语言推理的基本技术，由于其巨大的应用潜力而引起了越来越多的研究关注 (Li et al., 2023c; Yu et al., 2023a; Chen et al., 2023b; Yu et al., 2021)。如何赋予大语言模型多模态推理能力正成为研究热点。Cheng et al. (2023b)探索了当前知识编辑方法在细化多模态模型中的应用，并揭示了效果仍在进一步改善。RA-CM3 (Yasunaga et al., 2023)提出了一种检索增强的多模态模型，该模型使基本多模态模型能够引用检索器从外部存储器获取的相关文本和图像。未来的研究可以进一步研究知识和多模态大语言模型的整合，以应对现实世界中的复杂挑战。

8.2.3 基于大型语言模型的代理

自主代理长期以来一直是学术界和工业界的一个突出研究焦点 (Padgham and Winikoff, 2005)。通过获取大量的网络知识，法学硕士在实现人类水平的智能方面表现出了巨大的潜力，并为智能体的进一步发展带来了一线希望 (OpenAI, 2023; Sumers et al., 2023)。这些基于LLM的智能体可以表现出推理和规划能力，并已应用于各种现实场景 (Qian et al., 2023; Li et al., 2023b)。由于现实世界的多样性，基于法学硕士的代理人需要额外的信息来做出决策。探索实际复杂场景中知识与大语言模型方法的融合对于基于大模型的代理的开发非常重要。

9 结论

在本文中，我们对检索增强技术进行了调查，并对其主要方向提供了具体介绍。此外，我们总结了常用的数据集和前沿应用，并指出了一些有前景的研究方向。我们希望这项调查能让读者清楚地了解当前的进展并激发更多的工作。

[1] https://www.chatpdf.com/
[2] https://yiyan.baidu.com/
[3] https://xinghuo.xfyun.cn/
[4] https://search.tiangong.cn/

参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Jinheon Baek, Alham Fikri Aji, and Amir Saffari. 2023a. Knowledge-augmented language model prompting for zero-shot knowledge graph question answering. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*, pages 78–106.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong C. Park, and Sung Ju Hwang. 2023b. Knowledge-augmented language model verification.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems*, 35:31668–31683.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harrison Chase. 2022. Langchain. https://github.com/langchain-ai/langchain.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023a. Benchmarking large language models in retrieval-augmented generation. In *AAAI Conference on Artificial Intelligence*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2023b. Measuring and improving chain-of-thought reasoning in vision-language models. *arXiv preprint arXiv:2309.04461*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. 2024. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yibing Du, Antoine Bosselut, and Christopher D Manning. 2022. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online, June. Association for Computational Linguistics.

Weizhi Fei, Xueyan Niu, Pingyi Zhou, Lu Hou, Bo Bai, Lei Deng, and Wei Han. 2023. Extending context window of large language models via semantic compression. *arXiv preprint arXiv:2312.09571*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023a. Cook: Empowering general-purpose language models with modular and collaborative knowledge. *arXiv preprint arXiv:2305.09955*.

Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2023b. Knowledge card: Filling llms' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2024. Retrieval-generation synergy augmented large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11661–11665. IEEE.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. 2020. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online, July. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Yasuto Hoshi, Daisuke Miyashita, Youyang Ng, Kento Tatsuno, Yasuhiro Morioka, Osamu Torii, and Jun Deguchi. 2023. Ralle: A framework for developing and evaluating retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 52–69.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Zhengbao Jiang, J. Araki, Haibo Ding, and Graham Neubig. 2020. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023a. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13358–13376.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023b. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression. *arXiv preprint arXiv:2310.06839*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023d. Active retrieval augmented generation. *ArXiv*, abs/2305.06983.

Wenqi Jiang, Shuai Zhang, Boran Han, et al. 2024. Piperag: Fast retrieval-augmented generation via algorithm-system co-design. *arXiv:2403.05676*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, July. Association for Computational Linguistics.

Mingxuan Ju, Wenhao Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. 2022. Grape: Knowledge graph enhanced passage reader for open-domain question answering. *arXiv preprint arXiv:2210.02933*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zachary Dodds, Nova Dassarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, John Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom B. Brown, Jack Clark, Nicholas Joseph, Benjamin Mann, Sam McCandlish, Christopher Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Nikhil Kandpal, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2022. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, et al. 2024. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019a. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019b. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022. Generative multi-hop retrieval. In *Conference on Empirical Methods in Natural Language Processing*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023a. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: Communicative agents for "mind" exploration of large language model society.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023d. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6636–6648, Toronto, Canada, July. Association for Computational Linguistics.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023e. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023f. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9796–9810.

Jerry Liu. 2022. LlamaIndex, 11.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023a. Sail: Search-augmented instruction learning. *arXiv preprint arXiv:2305.15225*.

Ziyang Luo, Can Xu, Pu Zhao, Xiubo Geng, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023b. Augmented large language models with parametric knowledge guiding. *arXiv preprint arXiv:2305.04757*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. *arXiv preprint arXiv:2305.14283*.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark John Francis Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv*, abs/2303.08896.

Antonis Maronikolakis and Hinrich Schütze. 2021. Multidomain pretrained language models for green nlp. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 1–8.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. Disentqa: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Lin Padgham and Michael Winikoff. 2005. *Developing intelligent agent systems: A practical guide*. John Wiley & Sons.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June. Association for Computational Linguistics.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, J. Liu, Hao Tian, Huaqin Wu, Ji rong Wen, and Haifeng Wang. 2023a. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *ArXiv*, abs/2307.11019.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023b. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *arXiv preprint arXiv:2307.11019*.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Amrita Saha, Vardaan Pahuja, Mitesh Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online, July. Association for Computational Linguistics.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan L. Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *ArXiv*, abs/2210.09150.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2023. Cognitive architectures for language agents.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2022. Recitation-augmented language models. *arXiv preprint arXiv:2210.01296*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2022. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023c. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Hongru Wang, Boyang Xue, Baohang Zhou, Tianhua Zhang, Cunxiang Wang, Guanhua Chen, Huimin Wang, and Kam-fai Wong. 2024a. Self-dc: When to retrieve and when to generate? self divide-and-conquer for compositional unknown questions. *arXiv preprint arXiv:2402.13514*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024b. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Siye Wu, Jian Xie, Jiangjie Chen, Tinghui Zhu, Kai Zhang, and Yanghua Xiao. 2024. How easily do irrelevant inputs skew the responses of large language models? *ArXiv*, abs/2404.03302.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge clashes.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November. Association for Computational Linguistics.

Haoyan Yang, Zhitao Li, Yong Zhang, Jianzong Wang, Ning Cheng, Ming Li, and Jing Xiao. 2023. Prca: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5364–5375.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Retrieval-augmented multimodal language modeling.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023a. Do Large Language Models Know What They Don't Know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023b. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. 2021. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. *Advances in Neural Information Processing Systems*, 34:26462–26474.

Weijiang Yu, Haofan Wang, Guohao Li, Nong Xiao, and Bernard Ghanem. 2023a. Knowledge-aware global reasoning for situation recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, S Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023b. Generate rather than retrieve: Large language models are strong context generators. In *International Conference on Learning Representations*.

Wenhao Yu, Zhihan Zhang, Zhenwen Liang, Meng Jiang, and Ashish Sabharwal. 2023c. Improving language models via plug-and-play retrieval feedback.

Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. 2023d. Augmentation-adapted retriever improves generalization of language models as generic plug-in. *arXiv preprint arXiv:2305.17331*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023a. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5823–5840, Toronto, Canada, July. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023b. A survey of large language models.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556.

Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

生成式文本质量的自动评估方法综述

兰天[1]，马梓奥[1]，周杨浩[1]，徐晨[2]，毛先领[1]
[1]北京理工大学，计算机学院，北京市，100081
[2]北京理工大学，医工技术学院，北京市，100081
lantiangmfty@gmail.com

摘要

人工评估，作为生成式文本质量评价的金标准，成本太高；自动评估，核心思想在于要使其评估结果与人工评估高度相关，从而实现对生成式文本质量的自动化分析和评价。随着自然语言处理领域相关技术的迭代进步，使得生成式文本质量的自动评估技术，已然经历了多次技术范式的迭代。然而，学界至今依然缺乏对生成式文本质量自动评估技术的系统化总结。因此，本文将首先系统地对已有的生成式文本自动评估方法进行归纳总结，然后分析了生成式文本自动评估方法的主要发展趋势，最后为了使读者更加宏观地了解自动评估整体，对自动评估领域整体的未来研究方向进行了探讨和展望。

关键词： 文本生成 ； 自动评估 ； 综述

A Survey of Automatic Evaluation on the Quality of Generated Text

Tian Lan[1], Ziao Ma[1], Yanghao Zhou[1], Chen Xu[2], Xian-Ling Mao[1]
[1]School of Computer Science and Technology, Beijing Institute of Technology, 100081
[2]School of Medical Technology, Beijing Institute of Technology, 100081
lantiangmfty@gmail.com

Abstract

Human evaluation, as the gold standard for assessing the quality of generated text, is prohibitively expensive. Automatic evaluation, on the other hand, aims to achieve high correlation with manual evaluation, thereby enabling automated analysis and assessment of generated text quality. With the iterative advancement of technologies in the field of natural language processing, the automatic evaluation of generated text quality has undergone several paradigm shifts. However, there is still a lack of systematic summarization of these automatic evaluation techniques in the academic community. Therefore, this paper first systematically summarizes the existing methods for automatic evaluation of generated text. It then analyzes the main development trends of these automatic evaluation methods. Finally, to provide a more comprehensive understanding of automatic evaluation, the paper discusses and anticipates future research directions in the field of automatic evaluation.

Keywords: Text Generation , Automatic Evaluation , Survey

第二十三届中国计算语言学大会论文集，第169页-第196页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

1 引言

随着以ChatGPT和GPT-4为代表的大规模语言模型（Large Language Models, LLMs）的迅猛发展 (OpenAI, 2023)，自然语言处理（NLP）领域正经历着前所未有的变革 (Lee et al., 2023; Zhao et al., 2023a)。大规模语言模型在文本生成 (Su et al., 2022b)、机器翻译 (Vaswani et al., 2017)、对话系统(Lan et al., 2020; Lan et al., 2023) 等多个NLP应用场景中展现出卓越的性能，极大地推动了生成式人工智能技术的发展。然而，大规模语言模型文本生成能力的提升不可避免地对其生成内容的自动化评估带来了巨大的挑战 (Zheng et al., 2023; Liu et al., 2023a)。因此，如何有效地自动化评估文本生成模型的能力，以确保其可靠性和实用性，已成为NLP领域面临的一项重要挑战 (Fu et al., 2023; Li et al., 2024b; Pan et al., 2023)。

目前，自动评估技术是解决这一挑战的具有前景的重要技术方案 (Tao et al., 2017; Pan et al., 2023; Lan et al., 2024)。自动评估技术旨在设计和构建和人工评估高度相关的自动评估方法，以实现对生成文本的质量的可靠评估 (Lan et al., 2020; Tao et al., 2017; Fu et al., 2023)。自动评估技术不仅能够帮助研究人员快速分析优化模型，还能够进一步推动大规模语言模型的自我提升（LLM self-improvement），即通过自动评估方法生成的反馈优化并增强大规模语言模型生成结果 (Yuan et al., 2024b; Xu et al., 2024)。因此自动评估技术具有重要的研究意义。目前，生成式文本的自动评估技术已经取得了显著的进展。如图1所示，从2014年至今，有关自动评估指标论文的发表数量呈现出逐年快速增长的趋势。然而，当前关于生成式文本的自动评估技术仍缺乏系统的总结。这种状况不仅导致了针对具体NLP任务中自动评估方法的选择和应用上的混乱，同时也限制了对生成式文本自动评估技术发展趋势的总结，以及对未来研究方向的深入洞察。为了解决该问题，本文系统地整理和综述了生成式文本的自动评估方法，分析了生成式文本自动评估发展过程中四轮技术范式的转变 (Papineni et al., 2002; Zhang* et al., 2020a; Lan et al., 2020; Fu et al., 2023)。基于上述讨论，本文概括了自动评估技术发展的主要趋势，即自动评估方法在通用性、可解释性和多维度性方面的持续改进。最终，本文对自动评估领域的未来重要的发展趋势和研究方向进行探讨和展望，包括以下四个方面内容：（1）小模型自动评估能力的提升；（2）评估可解释评估的质量；（3）评估技术的应用；（4）多模态生成内容的自动评估。

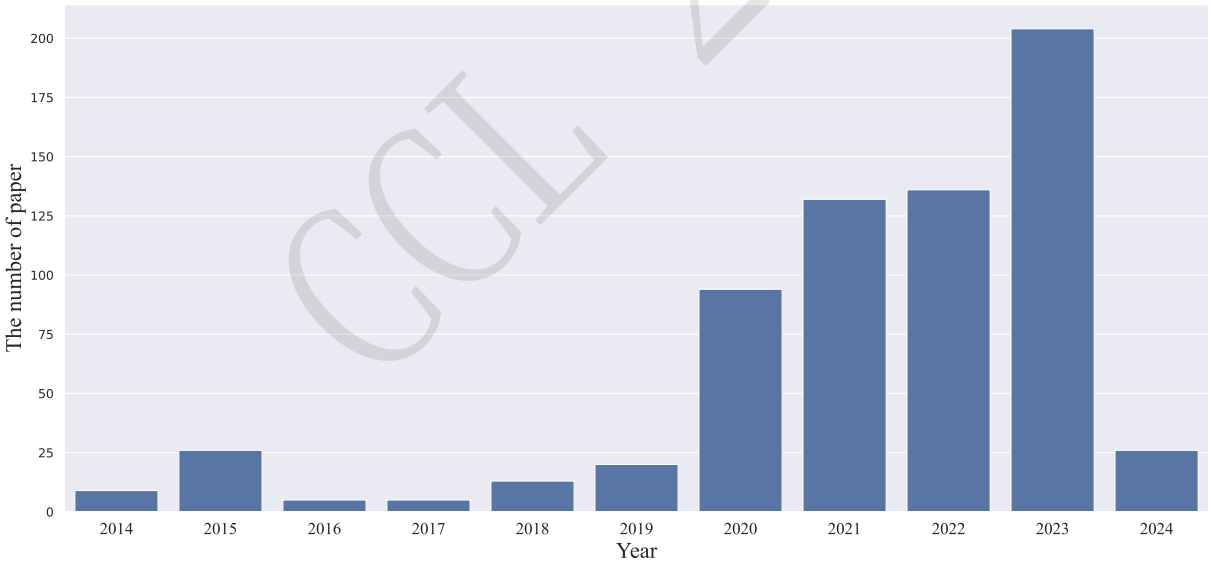


Figure 1: 生成式文本自动评估指标相关论文发表的数量呈现逐年递增的趋势（2014-2024）。

本文的后续内容安排如下：第2节将介绍生成式文本自动评估技术的相关背景。第3节归纳总结了生成式文本的自动评估技术的四轮技术范式的特点。基于归纳总结的内容，第4节对生成式文本自动评估技术的发展趋势进行了总结。最终，第5节将自动评估技术未来的研究趋势进行深入探讨和展望。

2 背景

2.1 生成式文本评估的分类体系

生成式文本的质量评估是衡量现有文本生成模型性能的关键手段 (Fu et al., 2023)。根据评估方法的不同，现有的生成式文本评估技术可以分为两大类：一类是基于标准答案的评估，另一类是无标准答案的评估 (Cobbe et al., 2021; Lan et al., 2020)。

2.1.1 有标准答案的评估

数据集	任务	数据规模
GSM8K (Cobbe et al., 2021)	小学数学题	8.79K
MATH (Hendrycks et al., 2021)	竞赛数学题	12.5K
HumanEval (Chen et al., 2021)	代码题	164
MBPP (Austin et al., 2021)	代码题	500
MMLU (Hendrycks et al., 2020)	人文社科等57个主题	14K
HellaSwag (Zellers et al., 2019)	常识推理	10K
AI2 Reasoning Challenge (Clark et al., 2018)	多项选择问答	2.59K
DROP (Dua et al., 2019)	段落理解问答	96K
C-Eval (Huang et al., 2023b)	中文52学科多项选择题	14K

Table 1: 有标准答案的评估基准数据集。

有标准答案的评估方法依赖于包含了人工标注的标准答案的测试数据集，以此来评价模型生成内容的质量。这种方法在评估近期流行的大规模语言模型，如GPT-4、InternLM2 (Team, 2023)、DeepSeek (DeepSeek-AI, 2024)和Qwen (Bai et al., 2023a)系列模型在解决特定下游任务的能力方面得到了广泛应用。[0] 例如，在模型解决数学问题和编程问题的任务中，通过将生成的答案与标准答案进行比较，以判断解题的正确性。部分常用的有标准答案的评估数据集如上表1所示。

2.1.2 无标准答案的评估

无标准答案的评估主要适用于那些没有固定标准回答的开放式文本生成任务，如开放式对话 (Tao et al., 2017)和故事生成 (Guan and Huang, 2020)等场景。

在这些任务中，人工评估是最常见且最可靠的评估方法 (Tao et al., 2017)。人工评估的优点在于人工结果的可靠性，因为人工标注能够直接反映人类的判断和偏好。然而，人工评估也存在明显的局限性，包括耗时、成本高昂以及实验结果的不可重复性 (Lan et al., 2020; Lan et al., 2023)。为解决无标准答案评估中人工评估的上述问题，自动评估技术逐渐开始受到研究人员的重视。自动评估的目标是开发一种自动化的评估模型或方法，使其评分结果与人工评估具有较高的相关性。然而，目前自动评估与人工评估的相关性仍有显著差距，难以完全替代人工评估。因此，如何有效提高自动评估方法与人工评估的一致性，成为当前研究的重点 (Fu et al., 2023; Pan et al., 2023)。随着自动评估技术的不断发展，它已经经历了从启发式评估、基于语义向量的评估、基于学习的评估到基于大规模语言模型的评估的四轮技术范式的转变。总体来看，随着自动评估技术范式的演变，自动评估方法的通用性、可解释性和评估维度的多样性都得到了显著的提升。

2.2 评价自动评估方法的性能

根据自动评估输出格式的类型，已有的自动评估工作可以分为两类：（1）**判别式自动评估**：这种方法直接为生成式文本一个预测的质量分数；（2）**生成式自动评估** (Pan et al., 2023)：这种方法生成针对待评估文本的可解释性分析内容，指出文本中的错误并提供改进意见 (Lan et al., 2024)。随着大规模语言模型的快速发展，生成式自动评估开始成为主要趋势，即大规模语言模型通过使用思维链（chain-of-thought） (Wei et al., 2023)的方式逐步生成针对待评估文本内容质量的分析以及对应的质量分数标签，实现更可解释、更细粒度、信息更丰富的自动评价。

[0] https://cdn.openai.com/papers/gpt-4.pdf

数据集	任务	数据规模
SummEval (Fabbri et al., 2021)	文本摘要	1.6K
WMT-22 (zh-en) (Freitag et al., 2022)	中英翻译	33.75K
WebNLG-2020 (Zhou and Lampouras, 2020)	RDF-to-Text	2.8K
LFQA (Jiang et al., 2023)	长文档问答	1.6K
GSM8K† (Jiang et al., 2023)	数学题评估	2.6K
OpenMEVA (Guan et al., 2021)	故事生成	1K
BAGEL (Mairesse et al., 2010)	Data-to-Text	202
CommonGen† (Lin et al., 2019)	生成式问答	2.8K
CriticBench (Lin et al., 2024)	推理类任务	3.83K

Table 2: 评价判别式自动评估方法的常用Meta-Evaluation数据集。†标识的数据集表示在原数据集（如GSM8K数据集）上构建的meta-evaluation数据集。

评估**判别式自动评估**的性能，首先需要收集人工标注的meta-evaluation数据集，这些数据集包含特定NLP下游任务的输入、待评估生成文本以及针对这些文本的多组人工标注质量分数。通过计算自动评估方法预测的分数与人工标注分数之间的相关性系数来衡量自动评估指标与人类评价之间的相关性。常用的相关性评估方法有Spearman、Pearson和Kendall相关性系数 (Zar, 2005; Fu et al., 2023)。目前，已有大量针对不同下游NLP任务的meta-evaluation数据集，常用的meta-evaluation数据集罗列如表2所示。

评估**生成式自动评估**方法的性能目前通常采用先进的大规模语言模型来完成，如GPT-4 (Li et al., 2024a; Cui et al., 2023; Sun et al., 2024)。然而，近期工作指出，GPT-4在评估生成式自动评估的质量和真实人类评估具有较大差别 (Wang et al., 2023c; Zhang et al., 2024)，因此如何准确评价生成式评估内容依然是一个较为困难的问题。

3 现有自动评估研究技术的发展

如图2所示，本文通过综合调研现有的自动评估技术，根据其发展先后顺序，将已有生成式文本自动评估方法归类为如下四个技术范式：（1）启发式的自动评估方法（Heuristic Evaluation）；（2）基于语义向量的自动评估方法（Embedding-based Evaluation）；（3）基于学习的自动评估方法（Learning-based Evaluation）；（4）基于大规模语言模型的自动评估方法（LLM-based Evaluation）。本节将对这四类自动评估范式的特点及其代表性工作进行系统分析和归纳。

3.1 启发式自动评估方法（Heuristic Evaluation）

启发式自动评估范式在NLP任务的早期阶段非常普遍。这种方法通常依赖于特定下游NLP任务中的规则或特征来构建评估方法，通过设计可解释的公式、算法或评估流程来计算质量分数。以下是一些常见的自然语言下游任务中使用的启发式评估方法介绍。

3.1.1 机器翻译

在机器翻译任务中，n-gram级别的字面信息是最为关键的特征。因此，启发式自动评估方法主要围绕n-gram特征构建。BLEU (Papineni et al., 2002)是一种广泛采用的方法，它通过计算生成翻译与高质量参考翻译之间n-gram的准确率来衡量二者的相似度。为了解决BLEU指标无法考虑翻译中重要词汇的问题，NIST (Doddington, 2002)引入了n-gram的信息量影响，考虑了翻译中罕见但重要的词汇对翻译质量的作用。由于BLEU仅计算了生成翻译n-gram相对于参考翻译的准确率，未考虑召回率，这影响了二者相似度的准确衡量。METEOR (Banerjee and Lavie, 2005)对此进行了改进，引入了召回率，并通过WordNet扩展了参考译文中每个n-gram的同义词，从而更好地考虑了翻译准确但不完全匹配的情况。类似地，chrF (Popović, 2015)和chrF++ (Popović, 2017)通过统计生成译文和参考译文之间字符级别n-gram的匹配程度来计算二者的相似度。除了n-gram级别的字面信息，文本的编辑距离也常用于评估生成翻译的质量。TER（Translation Edit Rate） (Snover et al., 2006; Moramarco et al., 2022)是一种基于编辑距离的评估指标，它通过计算将生成的机器翻译转换为参考翻译所需的最小编辑操作次

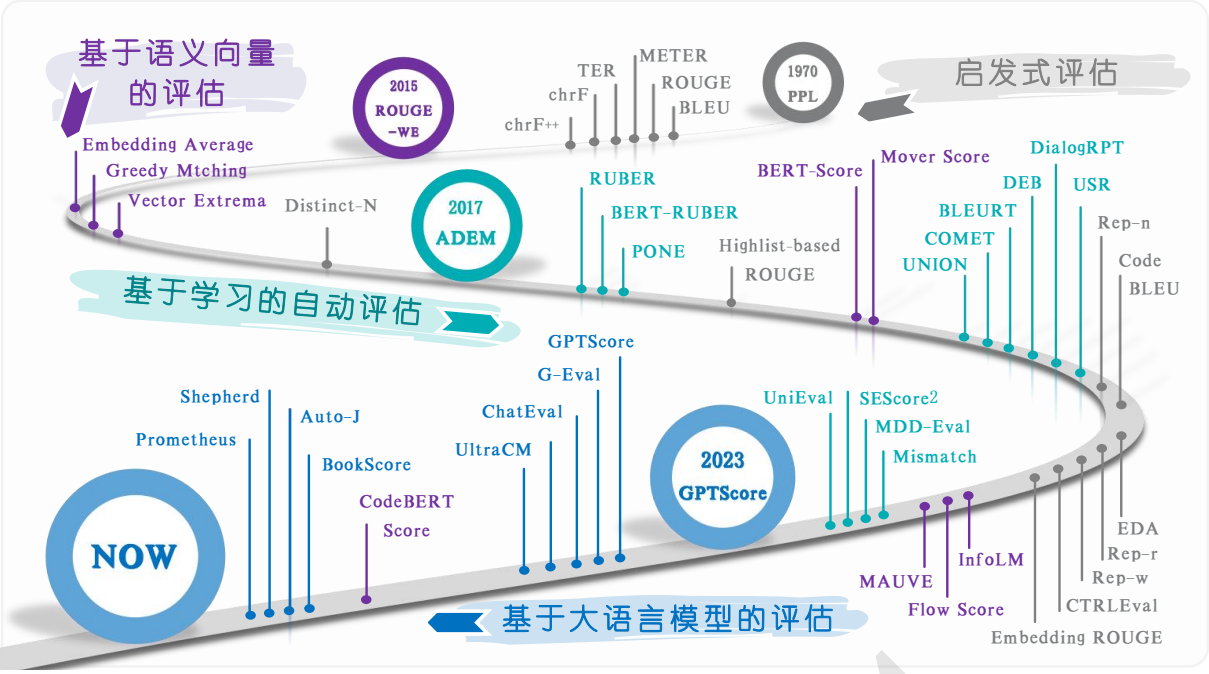


Figure 2: 生成式文本的自动评估指标经历了从启发式评估（Heruistic Evaluation）、基于语义向量的评估方法（Embedding-based Evaluation）、基于学习的自动评估方法（Learning-based Evaluation）到基于大规模语言模型（LLM-based evaluation）的四轮技术范式发展过程。

数来衡量翻译的质量。这些编辑操作包括插入、删除、替换和移动单词或字符。TER值越低，表示机器翻译的质量越高，因为它与参考翻译的差异越小。因此，TER提供了一种量化翻译误差的方法，帮助研究者评估和改进机器翻译系统的性能。与BLEU等基于准确率计算的指标相比，基于编辑距离的指标能够考虑到生成翻译中的错译和漏译问题。

3.1.2 开放式文本生成

开放式文本生成，如故事生成和对话生成等，是NLP领域中应用非常广泛的一类生成任务 (Zhang et al., 2018; Xu et al., 2022a; Yang et al., 2023)。这类任务的质量评估通常从多个角度进行，包括多样性、流畅性等 (Pascual et al., 2021; Xu et al., 2021)。

困惑度（perplexity）是评估生成文本流畅度中最常用的指标。它通过使用预训练的语言模型，如BERT/GPT等，来计算生成文本的预测概率 (Su et al., 2022c; Lan et al., 2022)。为了评估生成文本的多样性，Distinct-N (Li et al., 2016)指标通过统计不重复的1-gram和2-gram的比例来计算生成对话回复的多样性。类似的，Rep-n (Welleck et al., 2020)，Rep-w，Rep-r (Li et al., 2023a; Fu et al., 2021)指标通过评估生成文本中不重复n-gram的比例来评价文本的多样性。为了解决Distinct-N指标在长文本评估中的偏差，Expectation-Adjusted Distinct （EAD） (Liu et al., 2022)通过引入不重复token的期望来调整Distinct-N指标的计算过程。CTRLEval (Ke et al., 2022)通过设计多种不同的文本填充任务，计算基于上下文（包括前缀和属性标签）的生成文本的生成概率，以评估生成文本的连贯性、一致性和属性相关性。

3.1.3 文本摘要

ROUGE系列指标 (Lin, 2004)是摘要生成质量评估中应用最广泛的评估指标。ROUGE指标通过将生成摘要和参考摘要分割为n-gram，统计生成摘要n-gram相对于参考摘要的召回率，以此来衡量生成摘要与参考摘要的相似度。然而，ROUGE指标在计算过程中要求严格匹配n-gram，这可能导致对语义相近但字形不匹配的生成结果产生误判。为了解决这个问题，EmbeddingROUGE (Tsann et al., 2022)通过将n-gram的匹配对象扩展到参考摘要中语义相似的n-gram列表，从而放宽了精确匹配的约束。除了与参考文档匹配外，Highlight-based ROUGE (Hardy et al., 2019)通过计算生成摘要与源文档中标注的显著句子之间的n-gram重叠来评估摘要的质量。这种方法侧重于检测生成摘要是否包含了源文档中最重要、最相关的信

息，这对于确保摘要能够准确捕捉和传达原文核心内容至关重要。

3.1.4 代码生成

为了有效评估生成代码的质量，CodeBLEU (Ren et al., 2020)指标对BLEU指标进行了改造，在BLEU的计算过程中加入了三个计算分数，以分别考虑代码生成场景下的三种特征：（1）BLEU_{weight}对代码中的关键词施加了更高的权重，从而更准确地评估生成代码中的关键词相似性；（2）Match_{df}通过比较代码运行过程中的数据流图之间的相似度来评估生成代码和参考代码的相似度；（3）Match_{dst}将生成代码和参考代码分别表示为抽象语法树（AST），评估语法树之间的相似度。

3.1.5 其他NLP任务

除了上述NLP任务以外，针对更多NLP任务的启发式自动评估方法罗列在表3中。

启发式评估方法	针对任务
PARENT (Dhingra et al., 2019)	Table-to-text
SARI,FKBLEU (Xu et al., 2016)	文本简化
L'AMBRE (Pratapa et al., 2021)	文本形态变化
KDA (Moon et al., 2022)	多项选择问答

Table 3: 针对其他NLP任务的启发式自动评价方法。

3.2 基于语义向量的评估方法（Embedding-based Evaluation）

启发式自动评估方法严重依赖于特定下游NLP任务中的关键特征来计算生成文本和参考文本之间的相似度。然而，这些特性通常都是字符级别特征，如n-gram特征等。这些特征很难捕捉到复杂的语义信息。因此，在开放式文本生成任务等场景中，评价那些在字面上无法精确匹配但语义相似度高的评估文本时，其评估结果往往与真实人工评估的相关性存在较大差距。为了解决启发式自动评估方法无法考虑文本语义信息的问题，基于语义向量的自动评估方法开始逐渐被引入。基于语义向量的评估方法的计算过程可以形式化定义如下：

$$\text{Score} = \mathcal{M}\left(E\left(g\right), E\left(c\right), E\left(r\right)\right) \tag{1}$$

其中E代表具体的语义嵌入模型，早期的语义嵌入模型为词向量模型，例如Word2Vec (Mikolov et al., 2013)和GloVe (Pennington et al., 2014)等。随着预训练模型如BERT (Devlin et al., 2019)和RoBERTa (Liu et al., 2019)的快速发展，基于预训练语言模型计算得到的语义嵌入包含有更丰富的上下文信息，进一步提升了基于语义向量的评估方法的可靠性。其中g表示待评估文本。c, r表示上下文文本和参考文本，它们在基于语义向量的自动评估方法中通常用于衡量和生成文本的相关性，从而衡量生成文本的质量。需要注意的是，部分基于语义向量的评估方法可能不包含上下文文本或者参考文本，只依赖二者中某一部分具体信息计算和生成文本的相关性。$\mathcal{M}$代表度量语义相似度或者语义距离的计算方法，目前常用的计算方法有点积相似度、余弦相似度、WMS距离以及KL散度等。需要注意的是，大部分基于语义向量的自动评估方法直接计算生成文本和参考文本或上下文信息的语义相似度，无需针对嵌入模型进行额外的微调过程，因此往往都是无监督的。

根据调研，现有的基于语义向量的自动评估方法可以分为如下三类：（1）无需上下文但是基于参考文本的方法（Context-free and Reference-based）；（2）基于上下文但是无需参考文本的方法（Context-based and Reference-free）；（3）基于上下文以及参考文本的方法（Context-based and Reference-based）。

无需上下文但是基于参考文本的方法（Context-free and Reference-based） 无需上下文但是基于参考文本的方法方法旨在直接计算生成文本与参考文本之间的语义相似性，而不考虑上下文信息。例如，ROUGE-WE (Ng and Abrecht, 2015)通过计算参考和生成文本中单词的Word2Vec (Mikolov et al., 2013)词向量的点积相似度来扩展ROUGE指标。此外，Vector Extrema、Embedding Average、Greedy Matching (Liu et al., 2016)通过使用Word2Vec和ELMo模

型获取生成对话回复中单词的语义向量，然后通过不同的聚合策略得到对话回复整体的语义向量，然后通过计算和参考文本的语义相似度来衡量生成文本质量。例如Embedding Average计算句子中所有单词的语义嵌入向量的平均值（即平均池化）以获取句子语义向量。最终的生成对话回复和参考回复的语义向量通过余弦相似度计算得到。BERTScore (Zhang* et al., 2020a)是一种具有代表性的基于语义向量的自动评估方法。其首先使用BERT模型 (Devlin et al., 2019)来获取文本的语义向量，随后通过计算参考文本和生成文本中token对之间的余弦相似性的最大值的加权求和来计算得到最终的自动评估分数，其中逆文档频率（IDF）作为重要性权重。除了BERTScore指标以外，还有大量方法通过计算参考文本和生成文本之间的余弦相似性作为生成文本的质量分数 (Zhao et al., 2023b; Su et al., 2023a; Akula and Garibay, 2022; Lo, 2019)。除了文本语义向量之间的相似度以外，参考文本和生成文本之间的距离度量也广泛用于评估文本之间的相似度。例如，MoverScore (Zhao et al., 2019)和Sentence Mover's Similarity （SMS） (Clark et al., 2019)使用BERT作为嵌入模型，计算生成文本和参考文本之间的Word Move Distance。此外，散度度量也广泛用于基于语义向量的自动评估指标。例如，InfoLM (Colombo et al., 2021)使用不同类型的散度函数来比较参考文本和生成文本的离散概率分布之间的差异。MAUVE (Pillutla et al., 2021)通过计算参考文本和生成文本的混合分布之间的KL散度来衡量两者之间的差异。

基于上下文但是无需参考生成文本的方法（Context-based and Reference-free） 此类方法旨在直接评估生成文本与给定上下文之间的相似性，无需高质量参考文本，因此更适合于生成空间较大的开放式生成任务中。例如，FlowScore (Li et al., 2021)借助DialoFlow对话生成模型，通过计算生成对话回复与基于对话历史预测得到的semantic influence来评估生成对话回复的质量。这种方法能够更直接地捕捉到生成文本与上下文之间的关系，从而提供了一种无需参考文本的评估途径。

基于上下文以及参考文本的方法（Context-based and Reference-based） 基于上下文和参考生成文本的方法旨在同时评估生成文本与上下文以及参考文本之间的相关性。对于开放式对话生成，Frechet Bert Distance（FBD）和Precision-Recall Distance（PRD） (Xiang et al., 2021)通过编码对话历史与生成对话来获取包含对话上下文的语义信息。针对代码生成任务，CodeBERTScore (Zhou et al., 2023)自动评估方法首先将编程问题分别与参考答案和生成答案进行拼接，然后按照BERTScore的计算方式得到最终的评分。这种方法能够有效地结合编程问题的上下文信息，以及生成答案与参考答案之间的语义差异，为代码生成任务提供了一种精准的评估手段。

3.3 基于学习的评估方法（Learning-based Evaluation）

由于仅考虑语义相似度，基于语义向量的模型在处理以下两种情况时存在局限：（1）语义信息可能缺乏一些基本特征，如语法正确性和多样性。因此，即使生成文本与参考文本具有较高的语义相似度，也可能包含严重的错误。（2）对于一些高熵任务，如对话生成和开放式文本生成，有限的参考文本可能无法覆盖所有高质量生成内容的空间，因此自动评估方法容易误判与参考文本不相似的高质量生成内容。针对基于语义向量的自动评估方法的这两类问题，研究人员通过训练神经网络来模仿人类标注者进行人工评估的过程，即基于学习的自动评估方法。具体的，基于学习的评估方法计算过程可以形式定义如下：

$$\text{Score} = \text{Model}(g|c \oplus r) \tag{2}$$

其中，c, r分别表示上下文信息和高质量的参考生成内容。需要注意的是，在某些任务中，上下文信息高质量的参考生成内容并不是必须的。$\oplus$表示文本的拼接操作。其中g是待评估生成文本。Model通常是一个编码文本信息并进行分类的深度神经网络模型。

根据是否利用NLP任务中的上下文信息以及是否依赖高质量参考文本，基于学习的评估方法可以分为以下四类：（1）基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）；（2）无需上下文但是基于参考生成文本的方法（Context-free and Reference-based）；（3）无需上下文以及参考生成文本的方法（Context-free and Reference-free）；（4）基于上下文以及参考生成文本的方法（Context-based and Reference-based）。大部分现有的基于学习的评估方法都在构造的正负样本数据上进行判别式训练。其中，正样本

是指针对任务输入生成的高质量文本，而负样本则是低质量文本。通过训练模型区分高质量与低质量文本，实现对生成内容质量的预测。因此，构建包含正负样本的训练数据是基于学习评估方法的核心。根据数据来源的不同，上述四类基于学习的评估方法可以进一步细分为以下三类：（1）**有监督方法**：通过人工标注的方式筛选和构建训练数据中的负样本。尽管构建数据集的成本较高，但由此产生的数据集噪声较少，质量较高；（2）**自监督方法**：通过负采样 (Tao et al., 2017)来构建负样本。自监督方法可以自动构建数据集，但可能包含噪声，影响基于学习的自动评估模型的性能；（3）**混合方法**：结合了有监督和自监督方法的优点，联合训练基于学习的自动评估方法。

3.3.1 基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）

这类指标在开放领域NLG任务中得到了广泛应用，例如开放域对话生成和开放式文本生成。这类开放式任务的生成空间较大，有限的参考回复往往无法提供足够的信息来全面评估生成文本的质量，因此参考文本对于提升自动评估的可靠性并不显著。

有监督方法 DialogRPT (Gao et al., 2020)通过真实网站的评论数据中的width, depth, up-down属性收集高质量评级数据以优化DialoGPT (Zhang et al., 2020b)对话模型，以对生成对话回复质量评估模型。DEB (Sai et al., 2020)构建了一个包含多个参考回复和高质量负样本的人类标注数据集，用于训练对话响应评估模型。

自监督方法 自监督方法的目标是自动生成负样本，用于训练可学习指标，其中负采样是最广泛使用的技术。RUBERT (Tao et al., 2017)、BERT-RUBER (Ghazarian et al., 2019)和USR (Mehri and Eskenazi, 2020b)通过在随机负采样方法上微调语言模型来构建基于学习的评价模型。负样本和正样本的质量是自监督方法的关键。因此，后续的研究工作旨在提高这些增强样本的质量。 (Zhang et al., 2021; Zhang et al., 2022a; Zhang et al., 2022b; Zhang et al., 2023a; Phy et al., 2020; Wu et al., 2023)。例如，PoNe (Lan et al., 2020)提出了带权负采样和正样本增强来生成更高质量的正负样本用于训练。类似的，MDD-Eval (Zhang et al., 2022a)设计了五种方法构建多样的负样本：（1）语法扰动；（2）back-translation回译；（3）生成模型输出；（4）随机句子选择；（5）遮蔽和填充。UniEval (Zhong et al., 2022)通过设计基于规则的文本转换方法构建负样本，并在多种不同的NLP生成任务和多个维度上构建二元问答任务来对生成文本进行评估。

3.3.2 无需上下文但是基于参考生成文本的方法（Context-free and Reference-based）

类似的，根据正负样本构建方式不同，该类方法可以进一步分为如下三类：（1）有监督方法；（2）自监督方法；（3）混合类型方法。

有监督方法 早期的有监督方法旨在训练LSTM模型对生成式文本进行分类或者回归 (Guzmán et al., 2015; Guzmán et al., 2014; Gupta et al., 2015)。例如，RUSE (Shimanaka et al., 2018)在人类标注的翻译质量分数上训练的回归模型。

自监督方法 除了有监督方法外，自监督类方法也得到了广泛研究 (Jwalapuram et al., 2019; Kamal Eddine et al., 2022; Lu et al., 2023)。例如，SEScore (Xu et al., 2022b)和SEScore2 (Xu et al., 2023b)在检索增强的合成负样本样本上训练评估模型。

混合类型方法 有监督方法通过拟合真实人类的评估数据，在人工评估的相关性上相比自监督方法具有优势。然而有监督方法需要花费大量的时间和代价收集人工标注数据，相比之下，自监督方法通过自动构建负样本数据训练模型，无需人工标注数据即可训练自动评估模型。因此，结合有监督和自监督方法也被广泛的研究。例如BLEURT (Sellam et al., 2020)和MisMATCH (Murugesan et al., 2023)首先在大量的合成参考文本-候选文本对上预训练BERT模型，然后使用人工标注的分数对其进行进一步微调。这种方法结合了有监督和自监督的优势，既利用了大量的未标注数据，又通过少量的人工标注数据提高了模型完成评估任务的准确性和可靠性。

3.3.3 无需上下文以及参考生成文本的方法（Context-free and Reference-free）

同样的，无需上下文以及参考文本的基于学习的自动指标可以分为有监督类型和自监督两类方法。

有监督方法 有监督方法主要应用于文本简化 (Cripwell et al., 2023)、故事生成 (Chen et al., 2022)和作文生成任务的自动评估 (Taghipour and Ng, 2016)。这些方法依赖于大量的人工标注数据来训练模型，以实现对生成文本质量的准确评估。

自监督方法 自监督方法通过使用文本编辑策略对参考高质量参考文本进行扰动，构建具有不同错误类型的负样本，从而在正负样本数据上训练评估模型对文本中的错误进行识别。例如，UNION (Guan and Huang, 2020)在通过四种策略增强的负样本上进行训练，这些策略包括重复、替换、重排序和否定修改，用于训练故事生成的评估模型。

3.3.4 基于上下文以及参考生成文本的方法（Context-based and Reference-based）

相比前三种方法，该方法通过联合考虑上下文和参考文本的信息评估生成文本的质量。大多数此类基于学习的自动评估方法都基于人类标注的样本进行训练的 (Lowe et al., 2017; Maddela et al., 2023; Shi et al., 2023; Chen et al., 2020)。例如，在生成翻译质量的自动评估任务重，COMET (Rei et al., 2020)的两种COMET变体，COMET-MQM和COMET-DARR在人工标注的机器翻译语料库上训练，可以同时考虑输入源句子和高质量参考翻译的信息来评估生成译文的质量。这种方法能够更全面地捕捉到翻译质量的多维度特征，从而提供了一种更精准的评估手段。

3.4 基于大规模语言模型的评估方法（LLM-based Evaluation）

尽管大量工作证明了基于学习的方法已经实现了与人工评估的高度相关性和一致性，但它们严重依赖于高质量和多样化的训练数据。尤其是随着NLP任务的快速发展，研究者们不再满足于对生成质量的粗粒度评估，针对不同NLP任务的多维度和细粒度的质量评估已经成为重要的发展趋势。例如，在机器翻译任务中，研究者们更加关注翻译的流畅性、准确性以及其他细粒度维度上的质量。为了解决这些问题，基于学习的方法不可避免地需要收集大量高质量的正负样本数据以训练优化具体的自动评估模型。然而，在多个不同任务和评估维度上收集高质量的正负样本数据是一个非常困难的工作，这对基于学习的自动评估指标的扩展性提出了重大挑战。因此，自动评估指标的通用性和可扩展性已经成为自动评估领域中最重要的需求，这促使研究者们转向另一种强大通用和普遍的技术，以构建可靠和通用的自动评估方法，即基于大规模语言模型的（LLM-based）评估。

参考GPTScore(Fu et al., 2023)的形式化定义，基于大规模语言模型的生成式评估方法一般可以表示为：

$$E_i = \mathcal{D}\left(\text{LLM}\left(T(t, c, \mathcal{S}) \oplus E_{<i}\right)\right) \tag{3}$$

其中，t为待评估的任务描述，c为评估标准（criteria），$\mathcal{S}$为上下文或参考文本，在部分方法中可能为空；$T(\cdot)$使用上述内容构造评估提示词，构造方式一般与具体任务相关；$E_{<i}$代表在第i步已生成的评估文本，E_i代表此时评估文本的下一个词元；$\oplus$代表文本的拼接操作；LLM$(\cdot)$代表利用大规模语言模型获取下一词元在词表上的概率分布，$\mathcal{D}$代表具体的解码算法，一般常用为greedy search (Li et al., 2023a)。最终生成的评估内容用$E = \{E_i\}_{i=1}^{L}$表示，可以包含评估分数、排序、偏好标签、文本形式的评估解释等，其中L为生成的评估文本的序列长度。

相比于基于学习的评估方法，基于大规模语言模型的自动评估方法无需收集训练数据来微调模型。通过prompt工程的方式，基于大规模语言模型的自动评估方法可以灵活地处理各种不同的评估任务。例如，在评估方式上，大规模语言模型可以直接评估单一生成文本的质量（single-wise的评估），也可以比较一对文本的质量（pair-wise的评估） (Li et al., 2024a)。同时，由于大规模语言模型通常在包含多种不同语言的训练数据上进行预训练，因此它可以直接评估不同语言中生成文本的质量。

基于大规模语言模型的评估方法同样可以根据是否使用参考文本与是否依赖上下文信息两个维度分为如下三类：（1）基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）；（2）基于上下文以及参考生成文本的方法（Context-based and Reference-free）；（3）无需上下文的方法（Context-free）。

3.4.1 基于上下文但是无需参考生成文本的方法（Context-based and Reference-free）

在生成式文本的自动评估任务中，大规模语言模型的引入使得评估过程可以借助其蕴含的常识性知识与语义理解能力，从而让许多场景下的评估任务摆脱对参考文本的依赖成为可能。

同时，大规模语言模型捕捉上下文语义关系的能力需要借助充足的上下文信息才能得以更好地展现。因此，基于大规模语言模型的评估方法大多需要上下文信息而无需借助参考文本。

在开放性故事生成与对抗性攻击任务的评估中，有研究(Chiang and yi Lee, 2023)率先使用大规模语言模型替代人类评估，证明了大规模语言模型与人类专家评估结果的一致性。由于部分使用大规模语言模型的评估方法与人类的关联度仍不及中等大小的神经网络评估器，G-Eval(Liu et al., 2023b)在提示词中加入思维链(Wei et al., 2023)与范例填充机制缓解这一问题。

基于大规模语言模型的成对评估方法往往存在自强化、位置偏见等影响评估质量的缺陷(Li et al., 2023c)。为此，PRD(Li et al., 2023c)提出了同行分级与同行讨论的方法，通过多轮迭代确定每个模型打分的权重，再经过多个模型讨论得到评估结果。(Bai et al., 2023b)使用大规模语言模型生成辅助评估的问题，在多轮问答中利用去中心化的同行检查机制减轻缺少领域知识与单一模型潜在偏见对评估结果的影响。ChatEval(Chan et al., 2023)通过构建多智能体的裁判团队缩小了单智能体方法与人类评估水平的差距。

由人类为大规模语言模型评估任务撰写的提示词可能包含潜在的偏见，且评估结果关于提示词形式的敏感度尚未明确。为解决这一问题，AutoCalibrate(Liu et al., 2023d)借助人类专家的评估样例，通过多阶段的标准起草与修改，实现了基于大规模语言模型的评估方法与人类偏好的自动对齐。SocREval(He et al., 2024)引导大规模语言模型对评估的问题生成自己的答案作为参考，根据该答案对待评估的文本做定性分析，实现对推理的评估。WideDeep(Zhang et al., 2023b)在使用大规模语言模型生成评估标准的基础上，组建大规模语言模型网络，进行多轮评估与同行检查。

此外，基于大规模语言模型的评估方法也可以加入其它技巧以优化在特定任务上的表现。一项文本风格迁移任务的评估(Ostheimer et al., 2023)对提示词进行集成，提升了评估方法的健壮性以及与人类评估的关联度。该工作还发现经过指令微调的大规模语言模型更能胜任评估类任务。DeltaScore(Xie et al., 2023)通过计算打乱前后文本的似然度差值体现生成的故事情节的质量。BookScore(Chang et al., 2024)将大规模语言模型作为分类器，通过在人为预定义的错误类别上对句子进行分类，评估超长文本摘要的句子级别连贯性。

3.4.2 基于上下文以及参考生成文本的方法（Context-based and Reference-based）

由于大规模语言模型本身具有较强的语义理解能力，因此在评估过程中参考文本一般用于指示生成文本中应该包含的关键信息。在长文本摘要评估的研究中(Wu et al., 2024)，为降低评估过程中的模型推理成本、缓解长文本中部信息被忽视的问题，首先提取长文本中的关键句子作为参考文本，然后借助大规模语言模型完成评估。RAGAS(Es et al., 2023)在检索增强生成任务的评估工作中借助大规模语言模型对生成文本从忠实性、答案相关性、上下文相关性三个维度进行评估。大规模语言模型在评估过程中还可以根据单一的参考文本，生成多个表达方式不同的参考文本覆盖参考文本的语义空间，避免因生成文本与参考文本表达方式不同导致的评估结果与人类不一致的问题(Tang et al., 2024)。

3.4.3 无需上下文的方法（Context-free）

在基于大规模语言模型的评估方法中，不引入上下文的方法一般关注于生成文本的逻辑性、语义连贯性等方面。此类方法的关注点较为单一，适用范围小，数量较少。例如，BookScore(Chang et al., 2024)着重关注长文本摘要本身的句子级别连贯性，该方法将大规模语言模型用作连贯性错误的分类器，用不包含连贯性错误的句子在所有句子中的占比表征摘要的语义连贯性。

3.4.4 基于大规模语言模型自动评估技术的总结

虽然基于大规模语言模型的自动评估技术已经取得了和人工评估较高的相关性和一致性，但其在应用过程中依然面临如下五个严重的问题:

- 长度偏差（Length Bias）：近期研究指出，基于大规模语言模型的自动评估方法对生成文本的长度非常敏感。大规模语言模型倾向于给予文本长度更长、内容更丰富的待评估文本更高的评分 (Zheng et al., 2023; Zeng et al., 2023)。

- 位置偏差（Positional Bias）：在pair-wise的评估方式下，大规模语言模型常用于直接比较两段生成文本的质量。然而，已有研究指出，大规模语言模型更倾向于在prompt中特定位

置出现的待评估文本 (Zheng et al., 2023; Zeng et al., 2023; Wang et al., 2023d)。这一问题显著影响了基于大规模语言模型在实际应用过程的可靠性。

- 自加强偏差（Self-enhancement Bias）：大规模语言模型尤其倾向于偏向于自己生成的文本内容，因此针对其自己生成的文本内容评估往往会存在评分偏高的问题 (Zheng et al., 2023)。
- 尽管大规模语言模型在开放式文本生成任务的自动评估中取得了与人工评估较强的相关性，但其针对逻辑推理类任务的评估能力依然薄弱，例如解数学题、代码题、推理问答等 (Valmeekam et al., 2023; Huang et al., 2023a)。
- 不一致的评估内容：大规模语言模型本身具有明显的不一致性。经过多次解码推理得到的生成式评估文本存在较为明显的不一致信息，这显著影响了大规模语言模型的评估的可靠性 (Zhang et al., 2024)。

虽然已有工作证明了使用few-shot例子、高质量参考文本和严格的打分标准（criteria）可以提升基于大规模语言模型的自动评估方法的可靠性 (Kim et al., 2024; Fu et al., 2023; Li et al., 2024b; Li et al., 2024a)，上述这些问题对基于大规模语言模型的自动评估方法的影响依然非常严重，因此未来的工作需要重点关注和解决这些问题，以进一步提升基于大规模语言模型的自动评估方法的可靠性。

4 自动评估主要发展趋势总结

生成式文本的自动评估技术经历了从启发式评估、基于语义向量的评估、基于学习的评估到基于大规模语言模型评估的四轮技术范式的演变。通过分析已有自动评估工作，本节将总结和探讨针对生成式文本自动评估技术的三个核心发展趋势：（1）通用性；（2）可解释性；（3）评估维度的多样性。

通用性 随着自动评估技术的不断发展和完善，其评估的通用性得到了显著增强。早期的研究主要聚焦于单一语言和特定任务的评估，例如机器翻译、对话生成和摘要生成等任务的自动评估 (Papineni et al., 2002; Tao et al., 2017)。然而，这些针对特定任务开发的自动评估指标往往难以快速适应和迁移到新的自然语言处理（NLP）任务中，这一局限性在一定程度上阻碍了文本生成任务领域的进一步发展。为了克服这一挑战，研究人员开始探索更具通用性和迁移能力的自动评估指标。特别是在基于学习的自动评估方法中，研究人员通过在包含更多样化文本生成任务的数据上训练评估模型，使其能够同时适用于多种不同的任务。例如，UniEval (Zhong et al., 2022)通过在自然语言推理（NLI）、问答（QA）和情感分析（SST）等任务数据上构建负样本并联合训练自动评价模型，实验结果表明UniEval能够快速适应并泛化至训练阶段未曾接触的其他文本生成任务。近期，借助大规模语言模型强大的泛化能力，基于大规模语言模型的自动评估技术已经能够有效地处理多种语言和多种不同文本生成任务的评估 (Mehri and Eskenazi, 2020a; Zhong et al., 2022)。这种通用性的提升使得同一个模型能够用于评估多种不同的任务，展现了极大的灵活性和便捷性。

可解释性 自动评估技术的发展经历了从早期范式到基于大规模语言模型的范式的转变，显著提高了其可解释性。启发式自动评估、基于语义向量的自动评估和基于学习的自动评估方法能够对生成式文本质量进行评分，但无法提供全面且细粒度的评估内容。近期的基于大规模语言模型的评估方法利用大规模语言模型强大的文本生成能力，对生成文本的质量进行细粒度分析。这些方法不仅能指出错误内容，还能提出可行的改进建议，极大地提升了评估技术的可解释性。此外，这些丰富的细粒度解释信息甚至有助于大规模语言模型利用这些信息进一步增强其生成能力 (Yuan et al., 2024b)，对于实现大规模语言模型的自我提升具有重要价值。

评估维度的多样性 自动评估技术已从单一维度的评估发展到多维度全面评估 (Fu et al., 2023)。针对特定的下游任务，从多个不同角度评估生成质量能够提供更全面、更细粒度的视角，以捕捉生成文本在不同方面的质量信息，从而实现更准确有效的生成文本质量评估。例如，针对开放式文本生成的自动评估方法最初主要关注生成的相关性。然而，随着生成质量和流畅性的提升，研究人员开始更多地关注生成文本内部的一致性以及信息的丰富程

度 (Su et al., 2022b; Lan et al., 2022)。基于学习的自动评估方法主要通过设计针对维度的负样本的方式，训练自动评估模型关注特定的评估维度质量。例如，MDD-Eval (Zhang et al., 2022a)和UNION (Guan and Huang, 2020)通过设计多种不同类型的方法构建负样本训练模型，以关注生成文本中的错误信息。近期，大量工作通过在prompt中设计和定义特定的评估维度，从而让大规模语言模型在评估过程中关注特定的维度的质量信息 (Fu et al., 2023; Li et al., 2024a)。更进一步的，让模型直接学习生成和构建层次化的评估维度已经被证明有助于大规模语言模型生成更可靠的评估内容 (Yuan et al., 2024a; Hu et al., 2024)。

5 未来自动评估技术的研究方向

从基于启发式的评估范式到基于大规模语言模型的评估方式，生成式文本的自动评估技术得到了快速发展。本章节将探讨和展望自动评估技术未来的四个重要的研究方向：（1）小模型评估能力的提升；（2）评估自动评估的质量；（3）评估技术的应用：大规模语言模型的自我提升；（4）多模态生成内容的评估。

5.1 小模型自动评估能力的提升

大量研究已经证实，基于大规模语言模型的自动评估方法与人工评估表现出高度的一致性。然而，大规模语言模型在推理过程中会产生高昂的开销，这对于实现高效的自动评估构成了挑战 (Wang et al., 2024)。然而，目前的小规模语言模型在自动评估能力方面仍存在不足，其评估分数与人工评价分数之间的相关性明显弱于大规模语言模型 (Li et al., 2023b)。因此，针对性地提升小模型的自动评估能力将成为未来研究的重要趋势。目前，提升小模型自动评估能力的主要方法是通过知识蒸馏，将先进语言模型（如GPT-4）的自动评估能力通过微调的方式蒸馏到小模型上 (Cui et al., 2023; Li et al., 2023b; Wang et al., 2023c)。具体而言，首先使用GPT-4对待评估的查询-响应（query-response）配对数据生成自动评估内容，这些内容可能包括可解释的评估分析、评估分数等。通过收集大量的自动评估数据，然后微调参数量较小、推理效率较高的开源模型，以实现对其自动评估能力的增强。尽管当前的技术方案可以在一定程度上提升小模型的自动评估能力，但它们与先进语言模型之间的差距仍然较大，并且其评估能力的通用性依然受限。具体来说，当前技术方案的通用性面临以下三个重要问题亟待解决：（1）如何收集多样化的下游任务数据，以提高小模型自动评估能力的通用性；（2）如何涵盖更多样的评估设置，以适应不同的评估环境和需求；（3）如何涵盖更多样的自动评估维度，以全面评估生成文本的质量。

根据第4节的总结，提升自动评估技术的通用性是生成式文本的自动评估方法的一个关键的发展趋势。目前用于提升小模型自动评估能力的方法所使用的数据集，在数据规模、数据类别、评估设置、评估维度方面都存在一定的局限性，这对自动评估模型的通用性产生了负面影响。因此，未来的工作重点应当是收集涵盖更多任务、语言和评估设置的多样化自动评估数据，以用于微调模型，从而增强其评估的通用性 (Wei et al., 2022; Li et al., 2024a; Li et al., 2024b)。目前已有工作开始尝试解决这一问题。例如，在增加数据任务的多样性方面，Auto-J(Li et al., 2024a)涵盖了58个不同的生成场景，并构建了由GPT-4生成的高质量的可解释评估数据，用以微调Llama-2-13B模型，从而使其评估能力超过了GPT-3.5-turbo模型。此外，Auto-J同时具备对单一待评估文本和一对评估文本同时进行评估的能力，显著增强了其评估设置的多样性。针对评估维度的多样性方面，Prometheus (Kim et al., 2024)通过引入评估打分维度的问题，针对性地对生成回复文本的质量进行评估，展现了较强的灵活性。

5.2 如何评估可解释自动评估的质量

当前基于大规模语言模型的自动评估技术通过生成可解释的评估分析内容，实现了细粒度和可解释的自动评估。然而，对生成可解释评估内容质量的自动评价仍然是一个尚未充分探索的领域。作为一种特殊的文本生成任务，生成可解释评估分析内容的质量自动评价相比传统的文本生成任务更具挑战性。目前的研究发现，使用高质量的参考评估内容作为提示，大规模语言模型对生成的可解释自动评估内容的评价分数与人工评估结果具有较高的相关性。例如，CriticEval(Lan et al., 2024)的消融实验表明，移除高质量的参考评估文本会导致GPT-4等大规模语言模型在自动评价可解释评估内容方面的能力显著下降（平均相关性损失达19.3%）。此外，MetaCritique(Sun et al., 2024)的实验发现，通过将高质量参考评估和生成

评估拆解为原子信息单元（Atomic Information Unit，AIU），并使用GPT-4逐一验证生成评估中的AIU是否与参考评估的AIU匹配，可以实现与人工评估的高度一致性。这些研究都证实了高质量参考可解释评估文本在自动评价可解释自动评估内容中的重要作用。然而，高质量的参考可解释评估文本需要通过严格的人工标注来收集，这对自动评估可解释评估质量的方法的扩展性构成了严重挑战。因此，未来该领域的一个重要研究方向在于如何减少对人工高质量评估数据收集的依赖，从而进一步提升评估解释评估的有效性和可扩展性。

5.3 评估技术的应用：大规模语言模型的自我提升

除了降低人工评估开销和负担，自动评估技术也已被广泛应用于大规模语言模型（LLM）的自动提升中，（Yuan et al., 2024b; Xu et al., 2024），这主要体现在两个重要的阶段：（1）**推理阶段**：利用LLM的生成能力和自动评估能力，分析生成内容中的缺陷并提供改进建议，可以迭代改进生成回复的质量 (Saunders et al., 2022; Zhang et al., 2024; Madaan et al., 2023; Fernandes et al., 2023; Yao et al., 2023)；（2）**训练阶段**：自动评估的打分通常用于构建具有明确性能差距的文本数据，这类数据可以用于使用拒绝式微调（RFT）或偏好学习（RLHF (Lee et al., 2023)）的方式实现对大规模语言模型的进一步提升，提高LLM能力 (Yuan et al., 2024b; Xu et al., 2023a; Bowman et al., 2022; Bai et al., 2022; Xu et al., 2024)。例如，Self-rewarding (Yuan et al., 2024b)通过使用自己生成的自动评估分数对Llama-2-70B (Touvron et al., 2023)进行微调，从而进一步改进了Llama-2-70B的质量。类似地，ChatGLM-Math（Xu et al., 2024)通过微调数学批评模型对生成的答案的质量生成答案，通过拒绝式微调 (Touvron et al., 2023)和直接偏好优化模型的效果(Rafailov et al., 2023)。直觉上，当前大规模语言模型自我提升的有效性主要取决于模型自动评估的准确性，即具有较强反思能力的语言模型在自我提升方面潜力更大。然而，这一结论尚未得到证实。未来的工作重点在于分析大规模语言模型自我提升能力和自动评估能力的相关性。

5.4 多模态生成内容的评估

同文本生成类似，多模态场景的生成任务也面临着缺乏有效自动化评估指标的挑战。例如，随着文生图 (Rombach et al., 2021)、文生视频以及多模态对话问答 (Su et al., 2023b; Su et al., 2022a)领域技术的快速发展，生成图片和视频的质量以及对多模态输入内容的问答对话能力逐步提高。如何准确地自动化评估生成图片和视频的质量已成为核心难题之一。当前，已有部分研究在不同模态的自动评估上做出了初步尝试。接下来，本节将总结当前多模态领域自动评估技术的发展状况，并对其未来的发展趋势进行预测。

语音生成自动评估 生成语音的主流质量评估方法是基于人类听觉判断的大规模众包平均意见得分（MOS），这种方法测试成本昂贵且要求测试者具有一定的听感训练基础。因此，模仿MOS评价的生成语音质量自动评估也成为当前语音合成领域发展的核心问题之一。目前语音生成自动评估的发展受到了两个主要挑战的影响：（1）自动评价指标的设计依赖于针对合成语音质量的人工评价数据，这使得该任务数据构建的成本同样高昂；（2）合成语音的评价需要考虑多个维度，包括自然度、可懂度和是否符合预期的听觉效果，这些维度在目前的自动评估中难以全面体现。AutoMOS(Patton et al., 2016)对单元选择合成器生成的语音自动评估进行了初步的尝试，但其作用主要停留在辅助人工评测，提前选择对听者更有优势的语音样本。MOSNET(Lo et al., 2019)是首个直接评分的语音生成自动评估系统，其在2018年的语音转换竞赛VCC数据集上进行了训练。研究者发现，该方法在不同数据集评估上的泛化性较差，因此需要更多的数据来奠定训练基础。

目前语音合成自动评估的数据主要依托于主流的语音主观评估指标，即生成语音及其在不同听者的MOS评分数据。VoiceMOS系列(Huang et al., 2022a; Cooper et al., 2023)语音自动评估比赛的举办提供了丰富的人工评估数据，同时三星公司标注的SOMOS(Maniati et al., 2022)也标注了MOS评分数据集。这些数据集的构建进一步促进了语音生成自动评估技术的有效发展。以此为基础，基于学习的自动评估方法大量涌现，基于听感学习有MBNET(Leng et al., 2021)、LDNET(Huang et al., 2022b)、UTMOS(Saeki et al., 2022)、DDOS(Tseng et al., 2022)、MOSA-Net(Zezario et al., 2022)等系列工作，基于自监督学习的方法有SSL-MOS(Cooper et al., 2022)及其后续提升版本ZevoMOS(Stan, 2022)、LE-SSL-MOS(Qi et al., 2023)等。随着检索增强方法在NLP领域大放异彩(Lewis et al., 2020)，在自监督学习方法的

基础上，研究者提出了检索增强的MOS预测方法RAMP(Wang et al., 2023a)。也有研究者在基于语义向量嵌入表示的语音生成自动评估进行了尝试，其利用文本预训练语言模型理解和量化语音的质量，SpeechLMScore(Maiti et al., 2023)测量合成音频样本与预训练在自然语音上的生成语音单元语言模型的似然性，SpeechBERTScore(Saeki et al., 2024)受到了BERTScore(Zhang et al., 2019)生成式文本自动评估指标的设计启发，计算生成语音和参考语音的自监督密集语音特征的BERTScore，从而获得针对不同质量的语音有更好的泛化和鲁棒效果。此外，SQuId(Sellam et al., 2023)把研究角度扩展到了零样本情况下的多语言语音评估，PAM(Deshmukh et al., 2024)则默认语音和文本对比学习模型CLAP学习到大量的音频样本和描述该音频质量的反义词对（例如“清晰”和”模糊”），其将研究角度迁移到针对更广泛音频生成内容（包括语音、声音事件、噪声等）的评估。

目前语音合成的自动评估方法在通用性、泛化性和可解释性上依然不足，随着针对语音理解的大规模多模态语言模型的兴起(Latif et al., 2023)，如何利用这些模型来提高语音自动评估的通用性和可解释性以及评估维度的多样性，将成为重要的研究方向。例如，对情绪(Zhou et al., 2022)、口音(Liu et al., 2024; Zhou et al., 2024)、言语障碍(Huang et al., 2022c)等语音合成新兴方向的自动评价，应该是未来探索的有效方向。这些探索不仅能够提高语音合成的质量，还能推动自动评估技术的发展，使其更加精确和全面。

视觉内容生成自动评估 在视觉内容生成任务的自动评估相比文本生成任务的更加困难，其中视觉内容生成包含图片、视频作为输入或输出的生成任务，如文生图、文生视频、VQA (Antol et al., 2015)。目前，视觉内容生成任务可以分为如下两类：（1）视觉内容输入端任务：提供图片或者视频信息，针对问题或者对话历史进行回复、生成图片视频描述等 (Su et al., 2023b)；（2）视觉内容输出端任务：根据文本描述，合成图片或者视频 (Antol et al., 2015)。

目前，针对视觉内容输入的任务，自动评估最常用的方法是依赖标准答案的评估。这类方法主要评估跨模态语言模型 (Su et al., 2023b)理解视觉多模态信息的能力，常用的人工标注的测试数据基准有MMBench (Liu et al., 2023c)和MMMU (Yue et al., 2024)等。此外，由于存在标准的参考文本信息，针对文本的自动评估方法也常用于图像字幕（Image Caption）任务，以评估模型根据输入图片生成的描述文本的质量 (Xu et al., 2023b; Hessel et al., 2022)。近期，随着以GPT-4V等先进的跨模态语言模型技术的快速发展，跨模态语言模型已被广泛应用于针对视觉输入内容生成的开放式文本内容质量的自动评估 (Lu et al., 2024)。

目前，视觉内容输出端评估主要采用无标准答案的自动评估方法。已有的无标准答案的视觉内容生成的自动评估技术也可以分为四类：（1）启发式自动评估 (Marcos-Morales et al., 2023; Zhou et al., 2019)；（2）基于语义向量的评估，如FID (Seitzer, 2020)，FVD (Unterthiner et al., 2018)；（3）基于学习的自动评估 (Qin et al., 2023; Wang et al., 2023b)；（4）基于大规模跨模态语言模型的自动评估 (Zhang et al., 2023c; Ge et al., 2023; Ku et al., 2023; Lu et al., 2024; Lee et al., 2024)。以GPT-4v和GPT-4o为代表的跨模态大规模语言模型的快速发展也显著促进了针对视觉内容生成的自动评估技术的进步。然而，已有的工作注意到 (Lu et al., 2024; Ku et al., 2023)，GPT-4V等跨模态语言模型对生成图像的理解能力依然不足。因此，未来的工作应更关注进一步提升视觉跨模态语言模型对图像内容的理解和推理能力，从而为构建更准确的基于大规模跨模态语言模型的视觉内容生成自动评估方法打下基础。

综上，相对于生成式文本自动评估，多模态生成内容的自动评估发展较为缓慢，未来的工作可以借鉴文本领域的研究成果，进一步促进多模态领域自动评估技术的发展。

6 总结

随着大规模语言模型技术的飞速发展，自动评估已成为自然语言处理（NLP）领域的核心难题之一。本文系统地梳理了生成式文本自动评估技术的四次技术范式的变革，对各个技术范式的特点和代表性工作进行了详细的梳理和分析。基于这些分析内容，本文总结并提出了关于自动评估技术的主要发展趋势，即自动评估方法在通用性、可解释性和评估维度多样性上的持续进步。在此基础上，本文进一步对自动评估领域未来的研究方向进行了总结和展望。

第二十三届中国计算语言学大会论文集，第169页–第196页，太原，中国，2024年7月25日至28日。
卷2：前沿综述

参考文献

Ramya Akula and Ivan Garibay. 2022. Sentence pair embeddings based evaluation metric for abstractive and extractive summarization. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6009–6017, Marseille, France, June. European Language Resources Association.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023b. Benchmarking foundation models with language-model-as-an-examiner.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. Measuring progress on scalable oversight for large language models.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online, November. Association for Computational Linguistics.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Hong Chen, Duc Vo, Hiroya Takamura, Yusuke Miyao, and Hideki Nakayama. 2022. StoryER: Automatic story evaluation via ranking, rating and reasoning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1739–1753, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations?

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy, July. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Pierre Colombo, Chloe Clave, and Pablo Piantanida. 2021. Infolm: A new metric to evaluate summarization & data2text generation. In *AAAI Conference on Artificial Intelligence*.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2023. The voicemos challenge 2023: zero-shot subjective speech quality prediction for multiple domains. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–7. IEEE.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12053–12059, Singapore, December. Association for Computational Linguistics.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback.

DeepSeek-AI. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. 2024. Pam: Prompting audio-language models for audio quality assessment. *arXiv preprint arXiv:2402.00282*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy, July. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022. Results of wmt22 metrics shared task: Stop using bleu–neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2021. A theoretical analysis of the repetition problem in text generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Xiang Gao, Yizhe Zhang, Michel Galley, Chris Brockett, and Bill Dolan. 2020. Dialogue response ranking training with large-scale human feedback data. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 386–395, Online, November. Association for Computational Linguistics.

Wentao Ge, Shunian Chen, Guiming Chen, Junying Chen, Zhihong Chen, Shuo Yan, Chenghao Zhu, Ziyue Lin, Wenya Xie, Xidong Wang, et al. 2023. Mllm-bench, evaluating multi-modal llms using gpt-4v. *arXiv preprint arXiv:2311.13951*.

Sarik Ghazarian, Johnny Wei, Aram Galstyan, and Nanyun Peng. 2019. Better automatic evaluation of open-domain dialogue systems with contextualized embeddings. In Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf, editors, *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 82–89, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Jian Guan and Minlie Huang. 2020. UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166, Online, November. Association for Computational Linguistics.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2021. OpenMEVA: A benchmark for evaluating open-ended story generation metrics. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6394–6407, Online, August. Association for Computational Linguistics.

Rohit Gupta, Constantin Orăsan, and Josef van Genabith. 2015. ReVal: A simple and effective machine translation evaluation metric based on recurrent neural networks. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1066–1072, Lisbon, Portugal, September. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, Alessandro Moschitti, Preslav Nakov, and Massimo Nicosia. 2014. Learning to differentiate better from worse translations. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 214–220, Doha, Qatar, October. Association for Computational Linguistics.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2015. Pairwise neural machine translation evaluation. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 805–814, Beijing, China, July. Association for Computational Linguistics.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy, July. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2024. Socreval: Large language models with the socratic method for reference-free reasoning evaluation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. Clipscore: A reference-free evaluation metric for image captioning.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. Are llm-based evaluators confusing nlg quality criteria?

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. 2022a. The voicemos challenge 2022. *arXiv preprint arXiv:2203.11389*.

Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda. 2022b. Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 896–900. IEEE.

Wen-Chin Huang, Bence Mark Halpern, Lester Phillip Violeta, Odette Scharenborg, and Tomoki Toda. 2022c. Towards identity preserving normal to dysarthric voice conversion. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6672–6676. IEEE.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023a. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2023. Tigerscore: Towards building explainable metric for all text generation tasks. *arXiv preprint arXiv:2310.00752*.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. 2019. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), pages 2964–2975, Hong Kong, China, November. Association for Computational Linguistics.

Moussa Kamal Eddine, Guokan Shang, Antoine Tixier, and Michalis Vazirgiannis. 2022. FrugalScore: Learning cheaper, lighter and faster evaluation metrics for automatic text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1305–1318, Dublin, Ireland, May. Association for Computational Linguistics.

Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland, May. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. 2023. Viescore: Towards explainable metrics for conditional image synthesis evaluation. *arXiv preprint arXiv:2312.14867*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2020. Pone: A novel automatic evaluation metric for open-domain generative dialogue systems. *ACM Trans. Inf. Syst.*, 39(1), nov.

Tian Lan, Yixuan Su, Shuhang Liu, Heyan Huang, and Xian-Ling Mao. 2022. Momentum decoding: Open-ended text generation as graph exploration. *arXiv preprint arXiv:2212.02175*.

Tian Lan, Xian-Ling Mao, Wei Wei, Xiaoyan Gao, and Heyan Huang. 2023. Towards efficient coarse-grained dialogue response selection. *ACM Trans. Inf. Syst.*, 42(2), sep.

Tian Lan, Wenwei Zhang, Chen Xu, Heyan Huang, Dahua Lin, Kai Chen, and Xian-ling Mao. 2024. Criticbench: Evaluating large language models as critic. *arXiv preprint arXiv:2402.13764*.

Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Heriberto Cuayáhuitl, and Björn W Schuller. 2023. Sparks of large audio models: A survey and outlook. *arXiv preprint arXiv:2308.12792*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *arXiv preprint arXiv:2401.06591*.

Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin. 2021. Mbnet: Mos prediction for synthesized speech with mean-bias network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 391–395. IEEE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California, June. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online, August. Association for Computational Linguistics.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. 2023a. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023b. Generative judge for evaluating alignment.

Ruosen Li, Teerth Patel, and Xinya Du. 2023c. Prd: Peer rank and discussion improve large language model based evaluations.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Junlong Li, Fan Zhou, Shichao Sun, Yikai Zhang, Hai Zhao, and Pengfei Liu. 2024b. Dissecting human and llm preferences.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2019. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*.

Zicheng Lin, Zhibin Gou, Tian Liang, Ruilin Luo, Haowei Liu, and Yujiu Yang. 2024. Criticbench: Benchmarking llms for critique-correct reasoning.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Dublin, Ireland, May. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023d. Calibrating llm-based evaluator.

Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. 2024. Controllable accented text-to-speech synthesis with fine and coarse-grained intensity rendering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. 2019. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv preprint arXiv:1904.08352*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, August. Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada, July. Association for Computational Linguistics.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F. Wong, and Dacheng Tao. 2023. Toward human-like evaluation for natural language generation with error analysis. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907, Toronto, Canada, July. Association for Computational Linguistics.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada, July. Association for Computational Linguistics.

François Mairesse, Milica Gasic, Filip Jurcicek, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561.

Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. 2023. Speechlmscore: Evaluating speech generation using speech language model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Georgia Maniati, Alexandra Vioni, Nikolaos Ellinas, Karolos Nikitaras, Konstantinos Klapsas, June Sig Sung, Gunu Jho, Aimilios Chalamandaris, and Pirros Tsiakoulis. 2022. Somos: The samsung open mos dataset for the evaluation of neural text-to-speech synthesis. *arXiv preprint arXiv:2204.03040*.

Adria Marcos-Morales, Matan Leibovich, Sreyas Mohan, Joshua Lawrence Vincent, Piyush Haluai, Mai Tan, Peter Crozier, and Carlos Fernandez-Granda. 2023. Evaluating unsupervised denoising requires unsupervised metrics.

Shikib Mehri and Maxine Eskenazi. 2020a. Unsupervised evaluation of interactive dialog with DialoGPT. In Olivier Pietquin, Smaranda Muresan, Vivian Chen, Casey Kennington, David Vandyke, Nina Dethlefs, Koji Inoue, Erik Ekstedt, and Stefan Ultes, editors, *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 1st virtual meeting, July. Association for Computational Linguistics.

Shikib Mehri and Maxine Eskenazi. 2020b. USR: An unsupervised and reference free evaluation metric for dialog generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online, July. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hyeongdon Moon, Yoonseok Yang, Hangyeol Yu, Seunghyun Lee, Myeongho Jeong, Juneyoung Park, Jamin Shin, Minsam Kim, and Seungtaek Choi. 2022. Evaluating the knowledge dependency of questions. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10512–10526, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human evaluation and correlation with automatic metrics in consultation note generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland, May. Association for Computational Linguistics.

Keerthiram Murugesan, Sarathkrishna Swaminathan, Soham Dan, Subhajit Chaudhury, Chulaka Gunasekara, Maxwell Crouse, Diwakar Mahajan, Ibrahim Abdelaziz, Achille Fokoue, Pavan Kapanipathi, Salim Roukos, and Alexander Gray. 2023. MISMATCH: Fine-grained evaluation of machine-generated text with mismatch error types. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4485–4503, Toronto, Canada, July. Association for Computational Linguistics.

Jun-Ping Ng and Viktoria Abrecht. 2015. Better summarization evaluation with word embeddings for ROUGE. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1925–1930, Lisbon, Portugal, September. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. Text style transfer evaluation using large language models.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Damian Pascual, Beni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3973–3997, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Brian Patton, Yannis Agiomyrgiannakis, Michael Terry, Kevin Wilson, Rif A Saurous, and D Sculley. 2016. Automos: Learning a non-intrusive assessor of naturalness-of-speech. *arXiv preprint arXiv:1611.09207*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vitou Phy, Yang Zhao, and Akiko Aizawa. 2020. Deconstruct to reconstruct a configurable evaluation metric for open-domain dialogue systems. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4164–4178, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Adithya Pratapa, Antonios Anastasopoulos, Shruti Rijhwani, Aditi Chaudhary, David R. Mortensen, Graham Neubig, and Yulia Tsvetkov. 2021. Evaluating the morphosyntactic well-formedness of generated texts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7131–7150, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zili Qi, Xinhui Hu, Wangjin Zhou, Sheng Li, Hao Wu, Jian Lu, and Xinkang Xu. 2023. Le-ssl-mos: Self-supervised learning mos prediction with listener enhancement. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–6. IEEE.

Guanyi Qin, Runze Hu, Yutao Liu, Xiawu Zheng, Haotian Liu, Xiu Li, and Yan Zhang. 2023. Data-efficient image quality assessment with attention-panel decoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2091–2100.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Shuo Ren, Daya Guo, Shuai Lu, Long Zhou, Shujie Liu, Duyu Tang, Neel Sundaresan, Ming Zhou, Ambrosio Blanco, and Shuai Ma. 2020. Codebleu: a method for automatic evaluation of code synthesis.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. 2024. Speechbertscore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. *arXiv preprint arXiv:2401.16812*.

Ananya B. Sai, Akash Kumar Mohankumar, Siddhartha Arora, and Mitesh M. Khapra. 2020. Improving dialog evaluation with a multi-reference adversarial dataset and large scale pretraining. *Transactions of the Association for Computational Linguistics*, 8:810–827.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators.

Maximilian Seitzer. 2020. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August. Version 0.3.0.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July. Association for Computational Linguistics.

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P Parikh, and Jason Riesa. 2023. Squid: Measuring speech naturalness in many languages. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhengliang Shi, Weiwei Sun, Shuo Zhang, Zhen Zhang, Pengjie Ren, and Zhaochun Ren. 2023. RADE: Reference-assisted dialogue evaluation for open-domain dialogue. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12856–12875, Toronto, Canada, July. Association for Computational Linguistics.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels, October. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Adriana Stan. 2022. The zevomos entry to voicemos challenge 2022. *arXiv preprint arXiv:2206.07448*.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022a. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022b. A contrastive framework for neural text generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561. Curran Associates, Inc.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022c. A contrastive framework for neural text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023a. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada, July. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023b. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Shichao Sun, Junlong Li, Weizhe Yuan, Ruifeng Yuan, Wenjie Li, and Pengfei Liu. 2024. The critique of critique.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas, November. Association for Computational Linguistics.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, Tom Kocmi, and Furu Wei. 2024. Not all metrics are guilty: Improving nlg evaluation by diversifying references.

Chongyang Tao, Lili Mou, Dongyan Zhao, and Rui Yan. 2017. Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *AAAI Conference on Artificial Intelligence*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Phua Yeong Tsann, Yew Kwang Hooi, Mohd Fadzil Bin Hassan, and Matthew Teow Yok Wooi. 2022. Embeddingrouge: Malay news headline similarity evaluation. In *2022 International Conference on Digital Transformation and Intelligence (ICDI)*, pages 01–06.

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung-yi Lee. 2022. Ddos: A mos prediction framework utilizing domain adaptive pre-training and distribution of opinion scores. *arXiv preprint arXiv:2204.03219*.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Karthik Valmeekam, Matthew Marquez, and Subbarao Kambhampati. 2023. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Hui Wang, Shiwan Zhao, Xiguang Zheng, and Yong Qin. 2023a. Ramp: Retrieval-augmented mos prediction via confidence-based dynamic weighting. *arXiv preprint arXiv:2308.16488*.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023b. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2555–2563.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023c. Shepherd: A critic for language model generation.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023d. How far can camels go? exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Hanming Wu, Wenjuan Han, Hui Di, Yufeng Chen, and Jinan Xu. 2023. A holistic approach to reference-free evaluation of machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 623–636, Toronto, Canada, July. Association for Computational Linguistics.

Yunshu Wu, Hayate Iso, Pouya Pezeshkpour, Nikita Bhutani, and Estevam Hruschka. 2024. Less is more for long document summary evaluation by llms.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online, August. Association for Computational Linguistics.

Zhuohan Xie, Miao Li, Trevor Cohn, and Jey Han Lau. 2023. Deltascore: Fine-grained story evaluation with perturbations.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Chen Xu, Jianyu Zhao, Rang Li, Changjian Hu, and Chuangbai Xiao. 2021. Change or not: A simple approach for plug and play language models on sentiment control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15935–15936.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. 2022a. Cosplay: Concept set guided personalized dialogue generation across both party personas. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 201–211.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022b. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6559–6574, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Weiwen Xu, Deng Cai, Zhisong Zhang, Wai Lam, and Shuming Shi. 2023a. Reasons to reject? aligning language models with judgments.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023b. SESCORE2: Learning text generation evaluation via synthesizing realistic mistakes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5166–5183, Toronto, Canada, July. Association for Computational Linguistics.

Yifan Xu, Xiao Liu, Xinghan Liu, Zhenyu Hou, Yueyan Li, Xiaohan Zhang, Zihan Wang, Aohan Zeng, Zhengxiao Du, Wenyi Zhao, Jie Tang, and Yuxiao Dong. 2024. Chatglm-math: Improving math problem-solving in large language models with a self-critique pipeline.

Haoran Yang, Yan Wang, Piji Li, Wei Bi, Wai Lam, and Chen Xu. 2023. Bridging the gap between pretraining and fine-tuning for commonsense generation. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 376–383, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024a. Llmcrit: Teaching large language models to use criteria.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024b. Self-rewarding language models.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Jerrold Zar, 2005. *Spearman Rank Correlation*, volume 5. 07.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Ryandhimas E Zezario, Szu-Wei Fu, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao. 2022. Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:54–70.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2204–2213.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. In *ACL, system demonstration*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5676–5689, Online, August. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Thomas Friedrichs, and Haizhou Li. 2022a. Mdd-eval: Self-training on augmented data for multi-domain dialogue evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11657–11666.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2022b. FineD-eval: Fine-grained automatic dialogue-level evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3336–3355, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Chen Zhang, Luis Fernando D'Haro, Qiquan Zhang, Thomas Friedrichs, and Haizhou Li. 2023a. Poe: A panel of experts for generalized automatic dialogue assessment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1234–1250.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023b. Wider and deeper llm networks are fairer llm evaluators.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023c. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024. Self-contrast: Better reflection through inconsistent solving perspectives.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578, Hong Kong, China, November. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023a. A survey of large language models.

Wei Zhao, Michael Strube, and Steffen Eger. 2023b. DiscoScore: Evaluating text generation with BERT and discourse coherence. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3865–3883, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Giulio Zhou and Gerasimos Lampouras. 2020. Webnlg challenge 2020: Language agnostic delexicalisation for multilingual rdf-to-text generation. In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 186–191.

Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. 2019. Hype: A benchmark for human eye perceptual evaluation of generative models. *Advances in neural information processing systems*, 32.

Kun Zhou, Berrak Sisman, Rajib Rana, Björn W Schuller, and Haizhou Li. 2022. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*.

Shuyan Zhou, Uri Alon, Sumit Agarwal, and Graham Neubig. 2023. CodeBERTScore: Evaluating code generation with pretrained models of code. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13921–13937, Singapore, December. Association for Computational Linguistics.

Xuehao Zhou, Mingyang Zhang, Yi Zhou, Zhizheng Wu, and Haizhou Li. 2024. Accented text-to-speech synthesis with limited data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1699–1711.