

文章编号: 1003-0077(2026)03-0001-26

基于预训练模型的文本常识合理性估计研究综述

崔莞清^{1,2}, 毕可平^{1,2}, 郭嘉丰^{1,2}, 黄昊坤³, 程学旗^{1,2}

(1. 中国科学院 计算技术研究所 智能算法安全全国重点实验室, 北京 100190;

2. 中国科学院大学, 北京 100190;

3. 北京理工大学 计算机学院, 北京 100190)

摘要: 文本常识合理性估计旨在评估文本内容是否符合人类日常经验与普遍认知, 是提升自然语言处理系统可靠性的关键技术。相较于早期依赖规则和知识库的方法, 基于预训练模型的方法通过自监督学习隐式编码常识知识, 在泛化性和灵活性上实现突破, 成为当前主流。该文针对基于预训练模型的常识合理性估计方法, 从知识来源和是否依赖微调两个维度对现有方法进行系统梳理: 前者聚焦模型推理时的知识获取路径, 后者衡量方法对标注数据与训练资源的依赖性。此外, 该文从任务类型和常识知识类型两方面回顾了常用的 48 个评测数据集, 涵盖从普遍常识到特定知识类型的多样化任务场景。最后, 总结了该领域的突破路径: 零样本下的复杂推理优化; 异构知识的融合机制和多模态常识知识的挖掘、表征和利用, 为后续研究提供方向。

关键词: 常识合理性估计; 常识推理; 自然语言理解

中图分类号: TP391

文献标识码: A

Commonsense Plausibility Estimation Based on Pretrained Models: A Survey

CUI Wanqing^{1,2}, BI Keping^{1,2}, GUO Jiafeng^{1,2}, HUANG Haokun³, CHENG Xueqi^{1,2}

(1. State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences,

Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100190, China;

3. School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100190, China)

Abstract: Commonsense Plausibility Estimation evaluates whether textual content aligns with human daily experience and general knowledge. Compared with early methods that relied on rules and knowledge bases, methods based on pre-trained models implicitly encode commonsense knowledge through self-supervised learning and achieve breakthroughs in the generalization and flexibility. This paper systematically reviews these mainstream methods from two perspectives: the knowledge source (the knowledge acquisition path during inference) and the fine-tuning dependency (the sensitivity of methods to labeled data and computational resources during training). Additionally, 48 benchmark datasets are reviewed based on task types and commonsense knowledge categories. Finally, three breakthroughs are highlighted: complex reasoning optimization for zero-shot inference, fusion mechanism of heterogeneous knowledge, and representation and utilization of multi-modal commonsense knowledge.

Keywords: commonsense plausibility estimation; commonsense reasoning; natural language understanding

0 引言

文本的常识合理性估计(Commonsense Plausi-

bility Estimation), 又称常识验证(Commonsense Validation), 旨在评估一个给定的文本内容是否或者多大程度上符合常识——即人们在日常生活中积累的、普遍认同的知识和经验。常识合理性估计与

收稿日期: 2025-03-07 定稿日期: 2025-04-17

基金项目: 国家自然科学基金(62302486, 62441229); 中国科学院计算技术研究所创新课题(E361140); 中国科学院特别研究助理资助项目; 国防基础科研计划项目(JCKY2022130C039); 中国科学院战略性先导科技专项(XDB0680102)

事实验证 (Fact Verification)^[1-2] 存在本质区别: 后者主要关注客观事实, 其判断依赖明确的证据且结果具有确定性。相对而言, 常识合理性估计关注的内容多为人类社会共识性认知, 通常包括日常生活中积累的普遍知识和经验 (在部分数据集中也会关注基础科学知识), 其结果可能并非绝对的非黑即白。此外, 常识的判断可能缺乏直接证据, 因为常识往往不会被明确或自然地记录。尽管在自然语言处理领域, 研究者们专门构造了有针对性的数据集以评估或增强模型对常识的理解能力, 但与事实知识相比, 语料中对常识的记录仍相对稀疏。

常识合理性估计作为常识推理的一个细分任务, 是自然语言处理领域具有挑战性的研究方向之一, 随着大规模预训练语言模型 (LLMs)^[3-5] 的广泛应用, 常识合理性估计技术也变得愈发重要。文本生成作为自然语言交互的核心, 其内容的可靠性直接影响用户体验。现有语言模型虽然在许多任务上取得了显著进展, 但它们依然会生成违背常识的内容^[6-9]。尽管研究者提出多种方法尝试提升生成质量^[10-12], 但评估体系仍受限于词重叠的传统指标^[13-15]。这类指标仅能捕捉生成文本与参考文本的表面匹配程度, 无法检测深层次的常识违背问题。这种评估偏差导致模型优化方向偏离真实需求, 产生“指标虚高但常识错误”的评估假象。

考虑到现实需求和评估体系的缺陷, 构建系统性的常识合理性评估体系具有双重价值: ①作为生成内容的质量检测标准, 识别问答系统、对话系统等场景中违背常识的文本输出。②作为方法改进的评估指标, 更准确地衡量不同方法的优化效果, 甚至在训练中提供监督信号^[16], 引导模型在参数更新时考虑常识一致性, 而非仅追求表层特征的匹配。

关于对常识合理性估计方法的研究, 最早可追溯到 1984 年启动的 CYC 系统^[17], 它通过人工编写逻辑规则实现常识推理。该系统依赖专家手工设计的规则, 推理严格, 扩展性差。之后的方法逐渐从规则驱动演变为知识库驱动。代表性的工作包括 ConceptNet^[18] 和 WordNet^[19], 二者都通过网络结构来表示常识性知识, 并通过语义关系进行推理。这类方法能够通过自动化手段不断扩展知识库, 从而提升了知识的覆盖面和推理的灵活性。随后, 统计学习方法也被引入常识合理性估计中^[20], 这类方法通过从文本中提取统计特征来判断常识的合理性, 但缺乏深度语义表征能力。总体而言, 这些方法面临对人工的高度依赖和泛化能力不足的双重

挑战。

预训练语言模型^[21-23] 的出现带来了方法的变革。通过在海量文本数据上进行自监督学习, 这些模型能够自动学习常识知识的潜在结构和规律, 使得常识合理性估计不再依赖人工规则, 极大地拓展了常识推理的能力和应用范围。此外, 随着大规模预训练语言模型的兴起, 方法的研究方向也从传统的有监督学习和针对特定常识类型的评估向无监督学习和更广泛的常识类型的方向转变, 方法的适应性有了进一步提升。因此, 本文将聚焦于基于预训练语言模型的常识合理性估计方法。

同时, 为了评估常识合理性估计方法的性能, 研究者们构建了丰富多样的评测数据集, 覆盖普遍性常识^[24-26]、事件常识^[27-29]、因果关系^[30-32] 等多个领域。它们的任务形式也很多样, 不仅包括二元分类^[25, 33]、分数估计^[29] 等任务形式, 还涉及多项选择^[24, 34] 和分布预测^[35-36] 等更复杂的评估方式。这些数据集为模型的训练和测试提供了宝贵的资源。

尽管已有一些关于常识推理相关领域的综述^[37-40], 但这些综述主要集中在常识问答等一般性推理任务上。这些任务虽然与常识合理性估计在自然语言处理领域中相互关联, 但各自的研究重点和应用场景存在显著差异: 常识问答等任务的本质是“信息补全”, 需要处理不完整命题的填补, 缺失的信息通常是明确的; 而本文聚焦的常识合理性估计的本质是“信息验证”: 需对完整陈述进行合理性量化, 要求模型主动识别出需要重点关注的位置。常识合理性估计的独特价值在于, 它不仅为生成式人工智能提供了一个过滤机制, 还为其他常识推理任务提供了一个基础的验证框架。另外, 部分常识问答任务也可以转换为常识合理性估计的方式, 例如通过将选择题转换为对问题和候选答案的合理性估计。此外, 随着大语言模型的发展, 方法层面上也带来了革新, 提示工程、思维链、检索增强等新技术使零样本合理性评估成为可能, 受限于发表时间, 既有综述未能涵盖 2023 年后大模型时代的方法突破。

总的来说, 常识合理性估计领域已经取得了显著进展, 但目前仍缺乏一篇全面的综述文章来系统地总结和分析现有的研究成果。为了填补这一空白, 本文通过系统化的文献调研和筛选, 首次对基于预训练模型的常识合理性估计方法进行系统性综述。我们从以下主要数据库和资源中检索相关文献: ACL Anthology、IEEE Xplore、ACM Digital Library、Google Scholar、arXiv、OpenReview。检索

关键词包括“commonsense plausibility estimation”“commonsense validation”“commonsense reasoning”及其相关变体。检索时间范围为 2018 年至 2024 年,以覆盖预训练模型兴起后的最新研究成果。通过对初步检索到的文献进行全文阅读,剔除方法不完整、实验设计不严谨或与综述目标不符的文献。另外,我们重点关注领域内的顶级会议和期刊(如 ACL、EMNLP、NAACL、AAAI、NeurIPS、ICLR),最终纳入约 150 篇文献。

分析现有文献,我们发现方法之间的核心区别在于“方法的知识来源”和“对微调的依赖程度”。“方法的知识来源”关注推理时使用的知识是基于模型的内部知识还是引入外部知识进行增强,这一维度的分类直接关联到方法在知识完备性与推理效率之间的权衡:基于模型内部知识的方法虽能避免外部检索开销,但其推理质量受限于模型知识的覆盖度和准确度;引入外部知识库或检索增强的方法虽能扩大知识边界并提升可解释性,却面临额外推理成本与噪声干扰的难题。“对微调的依赖程度”关注方法是否需要额外的训练开销,此维度的划分在本质上反映了任务特异性与泛化迁移性之间的平衡,直接影响方法在真实场景中的实用价值:对模型微调可直接提升模型在特定场景的性能,但会导致模型灾难性遗忘与训练成本激增;零样本/少样本推理即拿即用,保留了基础模型泛化能力,但可能因参数更新受限而难以修正模型固有的认知偏差。

本文组织结构如下:首先,第 1 节形式化定义常识合理性估计任务;第 2 节将系统性地梳理和分析现有的常识合理性估计方法,从知识来源和对训练的依赖程度两个核心维度进行分类和深入探讨;第 3 节将聚焦于常识合理性估计方法的实际应用场景;第 4 节对常用的评测数据集进行介绍,分析其特点和适用性;第 5 节讨论常识合理性估计领域面临的前景与挑战;最后,第 6 节总结全文的主要内容和贡献。

1 常识合理性估计任务

常识合理性估计任务的目标是建立输入文本 D 到其常识合理性分数 s 的预测: $f: D \rightarrow s$ 。由于段落可以拆分为句子,因此任务输入通常为一个长度为 l 的句子 $D = \{w_1, w_2, \dots, w_l\}$,其中, w_i 代表句子中的第 i 个词或词元(token)。另外,复杂的关系大多可进一步拆分为三元组,因此输入也可以是

一个三元组 $T = (h, r, t)$,其中, h 、 r 和 t 分别表示头实体、关系和尾实体。鉴于句子是最常见的输入形式,后续的讨论将以句子为例进行展开。

根据任务形式差异,常识合理性估计任务可划分为以下三种主要类型(这里先对任务类型进行形式化概览,结合实际数据集的具体任务介绍参见第 5 节)。

分类标签预测 这类任务的目标是建立输入到二元决策空间的预测,模型的输出是代表输入内容的绝对合理性的离散的二值标签 $s \in \{0, 1\}$,其中,“0”代表不符合常识,“1”代表符合常识。评价该任务性能的主要指标为准确率和 F_1 分数,但当数据存在类别不平衡时,AUC-ROC 曲线能提供更稳健的评估。

标量数值回归 这类任务的目标是学习连续合理性评分函数,模型的输出是表示输入内容符合常识的程度的连续的实值分数 $s \in [0, 1]$ 。方法的性能通过衡量在整个测试集上的预测结果和真实结果的一致性来评估,主要依赖皮尔逊相关系数(Pearson’s γ)和斯皮尔曼等级相关(Spearman’s ρ),前者衡量线性相关性,后者评估序关系保持能力。

概率分布估计 这类任务的目标是构建概念空间上的概率分布预测,例如给定问题“企鹅可能是什么颜色?”及 11 种候选颜色,模型需输出每种颜色的合理概率。对于每个文本,模型的输出仍然是连续的实值分数 $s \in [0, 1]$ 。方法的性能通过衡量在每个概念空间上的预测结果和真实结果的一致性来评估,也就是通过斯皮尔曼等级相关来衡量序关系的保持能力(仅需关注概念间的相对合理性时,不关心绝对数值),或者通过 KL 散度来衡量两个概率分布之间的差异程度,还可以通过 Top-N 准确率来兼容多个合理答案的场景。

2 常识合理性估计方法

预训练语言模型通过在大规模语料库上的自监督学习,展现出对常识知识的隐式编码能力。这种能力源于其训练过程中对语言模式的深度建模:当模型预测被遮蔽的词汇或生成后续文本时,必须依赖对现实世界规律的潜在认知。例如,模型若能准确生成“北极熊通常是白色的”,则表明其掌握了关于北极熊的颜色属性知识。这种隐式知识存储于模型参数之中,与主要存储在全连接层中少数几个神经元上的事实类知识不同,研究表明^[41],常识知识不仅存储在全连接层,也存储在注意力层,且存储位

置范围广泛、较为分散。这种分布特性使得常识知识的提取和利用更具挑战性,但也为模型提供了更强的鲁棒性和泛化能力。

通过概率估计、文本生成等机制,这种隐式知识可以外化为常识合理性判断。然而,预训练语言模型的常识能力存在显著局限:首先,其知识边界受限于预训练语料的覆盖范围;其次,模型可能因训练数据的报告偏差而对知识的建模不准确;再者,静态的预训练知识难以适应动态演变的常识环境。因此,除了直接利用预训练语言模型的隐式常识,还可以融合外部文档、知识库、多模态数据等外部资源,以扩展模型的常识覆盖范围并修正潜在认知偏差。基于此,常识合理性估计方法可以大致分为两类:基于模型内部知识的方法和结合外部知识增强的方法,这两类方法的关键区别在于推理时所依赖的知识来源。

方法的知识来源决定了方法的知识范围,而方法的训练范式,也就是对微调的依赖程度,则影响了

所适用的实际场景:依赖微调的方法通过训练数据直接优化模型,提高模型在特定任务中的表现,但面临计算成本高、标注数据依赖性强、过拟合风险大等挑战。无/弱监督方法(如零样本或少样本学习)利用提示工程(Prompt Engineering)、思维链(Chain-of-Thought)等机制激发模型的零样本推理能力,虽在特定任务上性能受限,但在标注数据稀缺的场景中展现出较强的适应性。另外,在大模型时代,随着模型规模的不断扩大,微调模型的计算成本也在不断上升,使得这类无需额外训练开销的无/弱监督方法更具吸引力。

如图 1 所示,这两个维度的交叉分析构建了完整的方法论框架,为理解现有方法提供了系统视角,因此,本节将从这两个维度出发,先根据知识来源维度系统地回顾和梳理常识合理性估计方法的研究现状,然后从对微调的依赖程度对已有方法作进一步总结。另外,我们也在表 1 和表 2 中对方法的实验效果进行报告(2.5 节)。

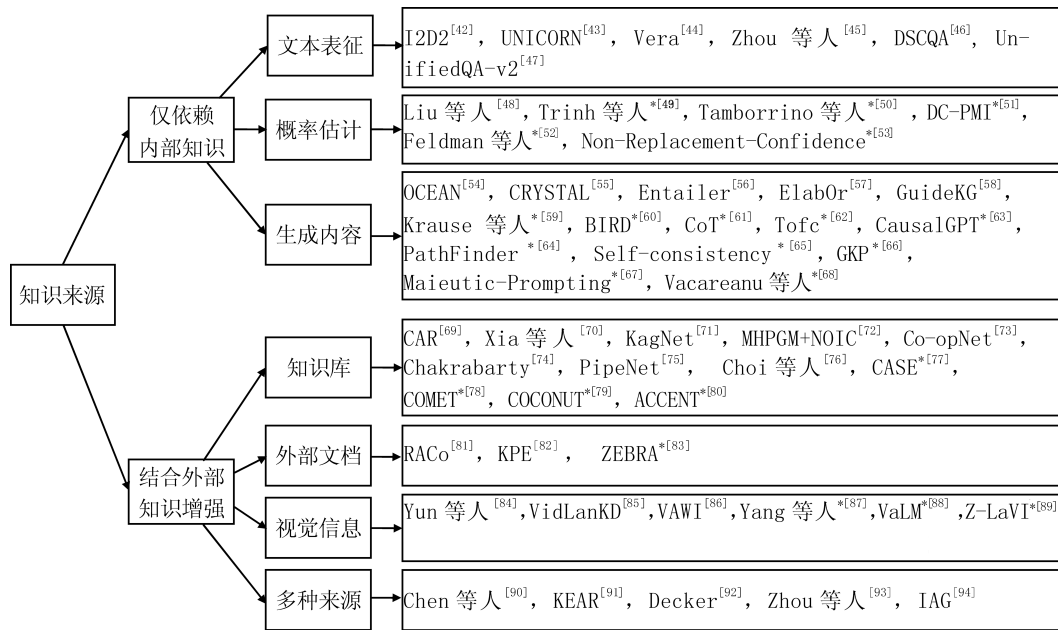


图 1 根据知识来源划分的常识合理性估计方法

其中“*”标注的是可以支持零样本的方法

2.1 基于模型内部知识的方法

基于模型内部知识的常识合理性估计方法充分挖掘了语言模型在大规模语料预训练中学习到的隐性常识知识。这类方法利用模型对文本序列的生成概率、模型根据输入生成的回复,以及表征向量空间分布特点来量化句子的常识合理性,无须依赖外部资源,具有高效简便的特点。

2.1.1 基于文本的表征

预训练语言模型通过多层 Transformer 架构对输入文本进行编码,最终输出的隐藏状态(即模型最后一层的上下文相关向量表示)可以作为文本的表征。这种高维空间的表示不仅捕捉了句子的语义信息、语法结构、上下文关系等方面的内容,还隐式存储了丰富的常识性信息。因此决策模型可以通过分析输入内容的表征来判断其是否符合常识^[42,95],通

常情况下,这个决策模型并不需要太复杂,一般是一个多层感知机。对于这类方法,决策函数的输入不再是文本内容本身,而是语言模型对文本的表征 $\text{Enc}(D)$,如式(1)所示。

$$s = f(\text{Enc}(D)) \quad (1)$$

由于引入了新的模块,这类方法需要利用有监督数据训练决策模型,为此,研究者做出了以下探索:首先,在训练数据方面,除了在目标数据集上进行训练,还可以整合多个数据集,使模型训练得更充分,增强其泛化能力^[43-44]。其次,在训练任务方面,除了直接训练决策模型对句子的常识合理性进行判断,引入其他辅助任务能够显著提升最终的效果,比如引入句子中概念之间的关系预测任务来促进模型对常识知识的掌握和运用^[70],或者引入对比学习来增大常识性句子与非常识性句子的表征之间的距离^[45]。最后,对推理过程的优化同样有助于提升模

型的表现,Bao 等人^[46]提出在模型的推理过程中引入技能模块,针对不同的输入结合不同的技能表征,让模型能够针对所需的技能类型调用合适的推理模式。

另外,表征的质量也很重要,对于三元组格式的输入,由于和预训练数据的差距,语言模型无法很好地对其进行编码,因此将结构化的三元组转换为自然语言形式效果会更好^[96]。

2.1.2 基于语言模型的概率估计

基于语言模型的概率估计方法通过量化模型对文本的生成概率来评估其常识合理性,其核心思想是:一个常识合理的输入句子应当具有较高的生成概率。这类方法将常识合理性估计视为一个回归任务,直接为每个输入计算其对应的分数。如图 2 所示,根据模型预训练任务的差异,不同模型对应不同的概率估计方法。

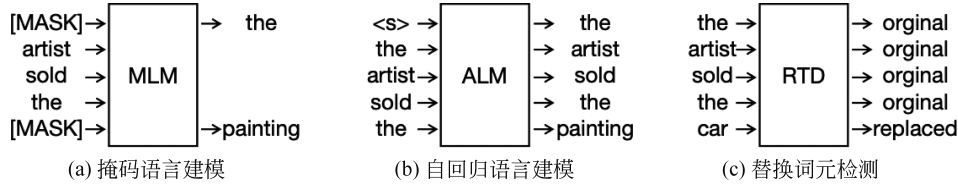


图 2 三种不同的预训练任务的输入和目标输出示意图

具体来说,对于以 GPT^[23] 为代表的使用自回归语言建模 (Autoregressive Language Modeling, ALM) 的模型,可以通过自回归的方式计算每个词的条件概率来评估整个句子的概率^[49],即,通过已知的前面词序列 $D_{1:i-1}$ 来预测下一个词 w_i ,相应的,句子的分数可以表示为:

$$s = -\frac{1}{l} \sum_{i=1}^l \log P_{\text{ALM}}(w_i | D_{1:i-1}) \quad (2)$$

对于像 BERT^[21] 这类的利用掩码语言建模 (Masked Language Modeling, MLM) 的预训练模型,评估句子的概率则需要通过掩码词预测来实现,也就是通过逐一掩蔽句子中的词 w_i 得到句子 $D_{\setminus w_i}$,并计算根据 $D_{\setminus w_i}$ 恢复原词的概率来评估句子的常识合理性^[50,52],如式(3)所示。

$$s = -\frac{1}{l} \sum_{i=1}^l \log P_{\text{MLM}}(w_i | D_{\setminus w_i}) \quad (3)$$

对于像 ELECTRA^[97] 这类预训练任务是替换令牌检测 (Replaced Token Detection, RTD) 的模型,可以通过预测句子中每一个词是否被替换来评估一个句子的概率^[53],如式(4)所示。

$$s = -\frac{1}{l} \sum_{i=1}^l \log P_{\text{RTD}}(D^{\text{cor}}, i) \quad (4)$$

其中, $P_{\text{RTD}}(D^{\text{cor}}, i)$ 代表句子中第 i 个词没有被替换的概率。

这种基于概率的估计方法实现简单,即使在不经过微调的情况下也能够对文本的常识合理性进行估计,如果能够将模型基于常识知识库进行知识对齐^[48],其准确性会进一步提升。但这类方法的缺陷是鲁棒性较差,容易受到表面形式竞争,即同义词替换或语句结构变化等语义无关的因素影响^[98]。这一问题可以通过引入额外的领域信息来修正^[51],或者通过对词汇权重进行调整,让模型的推理更加关注关键词 CASE^[77]。

2.1.3 基于语言模型的生成内容

针对具备指令遵循能力的大语言模型 (Instruction-following LLMs),研究者提出将常识合理性估计转化为生成式问答任务。这类方法的核心在于利用模型的指令理解与响应生成能力,通过设计特定提示引导模型完成合理性判断^[59]。例如,直接向模型提问“请判断以下陈述是否符合常识,仅回答‘是’或‘否’:太阳从西边升起。”,模型基于预训练阶段习得的常识知识和指令遵循能力,将生成“否”作为响应,从而完成合理性判定。

然而,直接让模型生成答案准确性还有很大的提升空间,除了设计模型更适用的提示词,还可以通过引入“思维链”(Chain of Thought, CoT)的方式增强推理过程,或者结合“知识生成”的方式类补充背景知识,从而提升推理的可靠性。这两种方法的核心在于构建扩展推理框架:模型首先生成中间辅助内容 z (具体类型因方法而异),然后将 z 与原始输入 D 共同用于最终决策,让模型的推理过程能够依赖更多的信息,如式(5)所示。

$$s = f(D, z) \quad (5)$$

思维链通过引导模型分步骤地进行推理,从而提高模型在复杂推理任务中的表现^[61]。具体而言,在给出最终答案之前,模型首先生成中间推理步骤 z ,然后再推理出最终的预测。这种方法能够模拟更接近人类的思考方式,逐步推理出问题的解决方案。进一步的优化除了微调模型来提升中间推理的质量^[54-55]、利用多智能体协作融合多个模型的结果之外^[63],还可以对线性思维链进行扩充:一方面可以通过多路径协同的方式,多个推理路径并行执行,而不仅仅是选择一条路径进行线性推理^[60, 64-65, 68],这种方式不仅能够对推理路径做进一步筛选,还能综合考虑多个角度的推理。另一方面,可以将思维链扩展为思维树(Tree of Thought),通过引入多个分支路径,使得模型能够从多个角度同时进行推理,从而更好地筛选出最优路径^[62]。此外,基于思维链的逆向推理方法也受到关注,这类方法针对符合常识和不符合常识这两种假设都分别生成对应的推理链,然后基于每条推理路径的概率分数来选择最终的结论^[56-57]。这种逆向推理机制能够更全面地考虑不同候选答案的可能性,更有效地结合推理过程的生成与验证,进一步提升常识合理性估计的准确性与可靠性。

虽然思维链和知识生成都是通过让模型生成更多的扩展内容帮助决策,但它们之间有本质区别:如图3所示,思维链关注逻辑性,中间输出为推理步骤,每一步之间都有明确的因果关系,侧重于通过一系列明确的逻辑步骤引导至最终结论,知识生成关注信息丰富性,中间输出为相关知识的集合,侧重于通过对概念的扩展、细化或补充来支持结论。

2.1.4 小结

基于文本表征的方法通过分析语言模型的高维向量空间分布构建决策边界,虽然能捕获深层的语义特征,但因为需要添加任务特定的分类器,这类方法依赖额外的训练过程。基于生成概率的方法利用

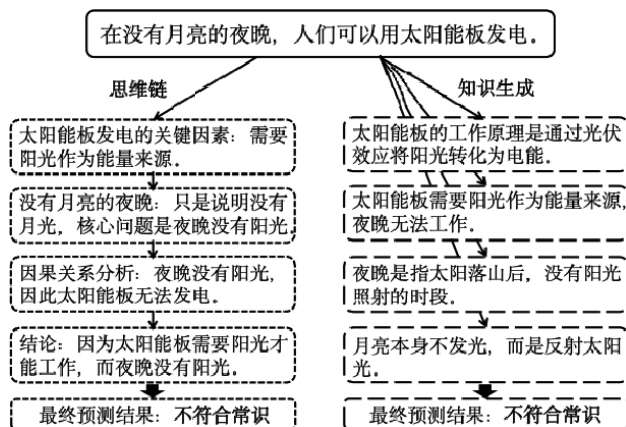


图3 基于思维链和常识知识生成方法的模型推理过程示例

语言模型对句子中标记的预测概率作为指标,利用的是语言模型的固有能力,具备实现简单和高效的优点,但容易受到语义无关因素的干扰。基于语言模型的生成内容的方法将常识验证重构为生成式问答任务,通过任务指令让模型完成零样本或少样本推理,但这类方法的有效性严格依赖于模型的指令理解能力,只适用于先进的大语言模型。

2.2 结合外部知识增强的方法

尽管预训练语言模型从语料库中学习到了大量的常识知识,但其对知识的掌握仍存在一定的局限性,主要体现在两个方面:一是训练语料中的知识范围有限,难以覆盖所有常识场景;二是模型习得的知识可能存在准确性不足的问题或时间滞后的问题。显式地引入外部知识,成为一种行之有效的策略,其优势在于:外部知识库或文档可以提供更广泛、更新的知识,弥补模型内部知识的不足。另外,通过高效的检索和知识整合机制,可以动态适配不同任务需求,从而提升推理的灵活性和准确性。接下来,我们根据不同的外部知识来源类型来对相关方法进行梳理。

2.2.1 引入知识库的方法

针对常识相关的任务,在引入外部补充知识时,知识库是最常见的选择。根据知识库的类型,可以将知识库分为静态知识库和动态知识库这两种,接下来,我们将对基于这两类知识库的方法进行详细的对比与分析,以揭示其各自的优势、局限性及适用场景。

ConceptNet^[18]和Atomic^[99]是目前应用最广泛的两个静态知识库,它们以结构化的三元组形式存储常识知识。ConceptNet^[18]主要侧重于对象的一

般性常识知识,例如物体的属性、用途、与其他实体的关系等。而 Atomic^[99]则聚焦于事件间的因果关系,例如,事件的前因、后果、可能的意图和情感等。在推理的过程中,具体选择哪个知识库取决于任务涉及的主要常识类型。这两种知识库基于人类专家的知识 and 大量的文本数据构建,更新频率低,因此被称为静态知识库。

在利用静态知识库时,检索相关常识知识通常基于概念词匹配,也就是基于输入文本中的关键概念,与知识库中的节点或关系进行匹配,得到与输入相关的子图(Subgraph)。为了平衡知识覆盖与噪声控制,检索的子图通常保留与匹配节点相关的一跳或两跳范围的邻居节点^[71]。这种策略能够有效捕获目标概念的上下文信息,同时避免因跳数过多而引入过多的无关知识或噪声。这种基于概念词匹配并限定跳数的方式简单高效,是目前静态知识库中最常见的检索策略。

为了利用子图信息,需要对子图进行编码,其中,图卷积网络(Graph Convolutional Network, GCN)^[100]是一种广泛使用的子图编码方法,如图4所示,能够通过卷积操作聚合邻居节点信息,生成独立于输入文本的知识表征。

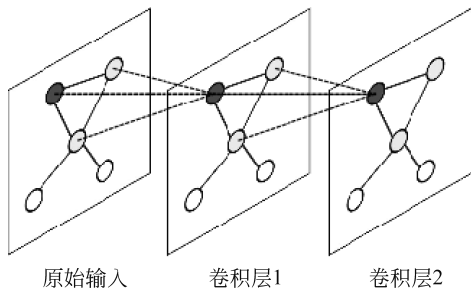


图4 图卷积网络的卷积过程示意

为了更清晰地展示,仅演示了针对一个节点的卷积操作

如图5所示,这些知识表征随后与输入文本的表征相结合,作为联合特征用于推理^[71,75-76]。尽管图模型是编码子图的直接方式,但其表征空间与语言模型的异构性增加了信息融合的难度。为解决这一问题,部分研究提出了基于语言模型的统一编码方法,即将子图中的路径转换为自然语言描述,再利用语言模型对其进行编码,从而使知识与文本共享同一表征空间,使融合过程更加自然^[72,79]。

此外,子图的检索结果可能包含与任务无关的信息,尤其是考虑多跳路径时,从而引入冗余信息或噪声,影响推理效果。为此,PipeNet^[75]提出了语义剪枝策略,通过分析子图中节点与任务目标的相关

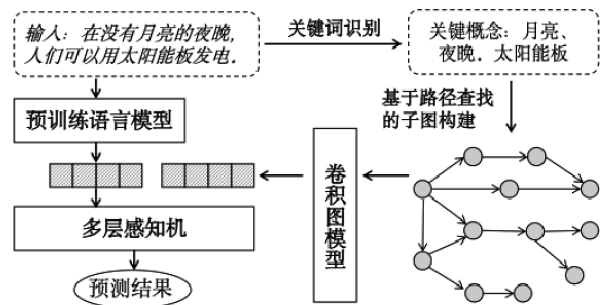


图5 KagNet网络^[71]的整体工作流程

性,精简无关节点,减少不必要的信息干扰,从而进一步提升推理的准确性和效率。

尽管静态知识库为常识推理提供了结构化支持,但其更新频率低、知识覆盖范围有限,因此难以满足全面知识补充的需求。与之相对的动态知识库是一种更灵活的知识来源。与以固定三元组形式存储知识的静态知识库不同,动态知识库指的是利用知识补全任务进行过微调的预训练语言模型,能够基于输入的概念和关系生成常识信息^[73,78]。动态知识库充分挖掘了存储在预训练语言模型中的常识知识,不仅突破了传统知识库在覆盖范围上的限制,还能基于多样化的输入得到相关常识信息,展现出更强的泛化能力和适应性。当将其用于知识增强时,动态知识库能够更全面地补充知识,从而比静态知识库提供更优的效果^[74]。动态知识库的另一独特优势在于其能够生成参考答案辅助评估,比如ACCENT框架^[80]利用动态知识库预测多种可能的尾实体,并将这些预测与实际尾实体进行比对,以对三元组的合理性进行评分。

总体而言,静态知识库在早期的常识推理任务中发挥了重要作用,通过结构化的常识知识提供了强有力的支持。然而,随着预训练语言模型的兴起,动态知识库凭借其在灵活性和适应性方面的独特优势,逐渐展现出更大的潜力。然而,目前的动态知识库仍主要基于参数规模较小、模型结构相对落后的语言模型,其生成的知识在质量和覆盖范围上尚存不足。未来,如何更好地挖掘大规模预训练语言模型中隐含的知识,是一个值得深入探索的重要方向。

2.2.2 引入检索文档的方法

检索得到的外部文档也是常识知识的重要补充来源,通过从大规模无结构化文档中检索相关内容,能够为模型提供额外的上下文支持。

常见的文档来源包括以下三类:人工标注的事实性数据集,如 OMCS^[101]、ATOMIC^[98]和

Wiktionary^[102])等,此类语料由事实陈述组成,质量相对较高,但覆盖范围有限;常识推理任务的基准数据集,如 CommonsenseQA^[24]、OpenBookQA^[103]等,这类语料中的问题和正确答案能提供丰富的推理信息,且这些数据集通常经过精心设计,能涵盖较为复杂的推理任务;以及从网络中收集的常识相关语料库,如 ARC^[104]、QASC^[105]和 GenericsKB^[106]等,这类语料规模更大,但噪声较高。

为了从语料中检索与输入相关的信息,当前主流方法通过预训练向量模型(如 BGE^[107]、Contriever^[108])将文本编码为稠密向量表示,借助向量空间内的相似度计算实现语义检索。相较于传统关键词匹配,这种基于深度语义嵌入的检索范式能够有效捕捉查询与文档间的隐式语义关联,显著提升开放域知识检索的准确率。

检索增强生成(Retrieval-Augmented Generation, RAG)是一种结合检索和生成的框架,通过从大规模文档中检索相关信息并将其作为上下文输入生成模型,从而提升生成内容的准确性和合理性。其核心思想是利用检索模块从外部文档中提取与输入相关的知识,并将其与语言模型的内部知识相结合,以增强推理能力。这类方法能够在不需要进一步训练模型的情况下动态扩展模型的知识边界,弥补预训练模型内部知识的不足。

检索到的相关文档可通过以下两种方式融入后续推理阶段:知识增强将检索到的文档片段(如事实陈述)作为补充上下文,拼接后输入推理模型,从而提升预测的精度^[81]。参考答案辅助则通过引入与当前任务相似的问题、答案甚至解释作为参考信息,通过类比推理机制激发模型的解题思路^[83]。此外,为了更灵活地利用检索信息,可以引入门控机制来自适应地决定何时以及多大程度上融合外部文档信息,进一步优化知识的利用效率^[82]。

引入检索文档的方法通过从大规模无结构化文本中检索相关内容,增强了推理的广度。检索增强生成框架作为一种灵活的扩充模型知识边界的方式,为常识合理性估计提供了新的思路。然而,如何高效整合检索到的文档信息,避免与任务无关的噪声信息干扰,以及高效地将检索到的知识与模型内部知识融合,仍然是未来研究的核心挑战和关键方向。

2.2.3 结合视觉信息的方法

前述方法,无论是基于语言模型本身,还是通过引入外部知识库或文档增强,都依然依赖于文本

模态来获取知识,缺乏与现实物理世界的直接对应,对视觉常识^[109]和现实世界知识^[110]的捕捉不足。另外,常识知识作为一种共识信息,在交流中往往被忽略,且部分常识知识用文字表达并不自然,因此文本中通常缺乏对常识信息的记录,同时人们更倾向于提及反常的信息。譬如明确指出香蕉是黄色的通常显得冗余,而对于较为罕见的红色香蕉,人们更倾向于在文本中特别提及。这导致了文本语料的报告偏差^[111],即文本中常识知识的分布与现实中的分布存在显著差异。很多缺乏文字记录的常识却天然存在于照片、视频等视觉模态的信息中,这为常识推理提供了新维度的辅助^[112]。近年来,越来越多的研究开始探索如何利用视觉信息来增强常识推理能力。

为了利用多模态信息,使用多模态模型是最直接的解决方案,包括视觉文本模型(Vision-Language Models, VLMs)和多模态语言模型(Multi-modal Language Models, MLMs)。前者通过对齐视觉和文本模态的表示来学习跨模态的关系,侧重于视觉-文本的匹配,如 CLIP^[113]、ALIGN^[114];后者扩展了传统的语言模型,使其能够理解文本与视觉信息,适用于更复杂的视觉语言推理任务,如 BLIP-2^[115]、LLaVA^[116]。这类模型通过联合训练语言与视觉模态,能够同时从两种模态中获取知识。例如,在处理与视觉相关的常识(如空间关系、从属关系等)时,多模态模型展现了显著的优势,即使在没有视觉输入的情况下,也能激活相关的视觉常识^[87]。然而,这类模型在增强视觉常识推理能力的同时,往往导致非视觉常识推理性能下降,尤其在处理复杂任务时效果不尽理想^[84,87]。因此,如何在提升视觉常识推理能力的同时,保持模型对其他类型常识的推理能力,是该方向的一大挑战。

向语言模型中注入视觉信息是一种更灵活的方法。将与文本输入相关的图像信息融入模型有两种方式:一种是在预训练阶段引入视觉相关的监督信号,使语言模型在基于文本语料训练的过程中也能学习视觉常识^[85];另一种是通过多模态模型得到额外的词编码作为视觉增强表示,并利用重构模块将其注入预训练语言模型中,以丰富语言模型中词表示的视觉语义^[86]。

此外,显式引入图像作为推理输入的方式也得到了关注。这类方法通过检索或生成相关图像,将图像作为显式输入。例如, VaLM^[88]在语言建模过程中引入相关图像,提高了推理效果。Z-LaVi^[89]结

合了检索图像和生成图像,通过综合多模态推理结果,显著提升了常识推理的准确性。这种显式融合方法为模型提供了更丰富的上下文信息,但也具有计算成本高和融合策略复杂的缺陷。

通过引入视觉信息,这些方法有效弥补了纯文本方法在常识推理中的局限性,扩展了常识知识的来源。然而,如何高效融合视觉与语言信息,特别是在多模态推理任务中平衡视觉和非视觉常识的推理能力,仍然是未来研究的重要方向。另外,需要注意的是,视觉信息只能补充视觉相关的常识知识,其他类型比如因果知识等不能被捕捉。

2.2.4 多种信息结合的常识合理性估计方法

在常识合理性估计任务中,结合多种信息源的方法能够从不同角度获取知识,让模型可以利用的信息更加全面,有效弥补单一来源的局限,从而提升推理的准确性和适应性。

像知识图谱这类结构化信息与文档这类非结构化信息具有不同的特点。结构化知识提供了清晰的实体及其关系表示,但覆盖范围有限且稀疏性较高;非结构化知识包含丰富的上下文信息,但内容中也包含更多的噪声。这两类知识具有天然的互补性,如何融合它们成为增强推理能力的重要方向。为了同时处理异构的知识,统一的知识表示是关键。常见的融合方式是使用自然语言将两者统一,利用模板将从知识图谱中检索到的实体和关系转换为自然语言句子,并与语料中的文本一起作为模型输入^[90-91]。另一种方法是将图(Graph)作为统一的中间表示,例如 DECKER 模型^[92]将输入文本和检索到的句子视为特殊节点,与知识图谱的子图结合,构建统一的推理框架。这种基于图的表示能够充分利用结构化知识的优势,但将自然语言视为统一形式的方法直接利用了文本数据与语言模型的天然适配性,避免了将文本转换成图过程中的不准确及与图与语言模型结合时的对齐问题,因此在灵活性和效率上具有优势。

除了融合异构知识外,如何有效结合语言模型的内部知识与外部知识也是常识合理性估计中的关键问题。外部知识通常更为可靠,但需要从有限候选集中检索,这不仅引入了额外的复杂度,还可能受到覆盖范围的限制;而语言模型内部蕴含的知识使用灵活,但容易受到“幻觉”现象的影响,生成错误的常识性陈述。结合这两种知识能够取长补短,一种方法是在检索外部知识的同时也提示模型显式输出相关的知识,将两者一起作为任务输入的补充,使模

型能够从更丰富的知识中受益^[93]。然而,这种方法要求模型在推理过程中自行判断知识的可信度,当内部知识与外部知识不一致时,可能会引起模型的混淆,从而影响推理结果的准确性;另一种方法是利用模型内部知识对检索到的外部知识进行扩充,比如利用语言模型的归纳能力,在保证知识可靠性的同时兼顾到丰富度^[94]。这种方式能够有效缓解外部知识覆盖范围不足的问题,同时降低因内部知识引入错误信息的风险。

尽管多种信息结合方法在常识合理性估计中表现出显著优势,但如何平衡不同信息源仍然是一个开放性问题。不同任务对知识来源的依赖程度各不相同,例如某些任务中外部知识至关重要,而另一些任务中语言模型的内部知识可能已经足够,过多引入外部知识反而可能产生负面影响^[66]。因此,根据具体任务适应性选择知识结合策略,是未来研究的重要方向。

2.2.5 小结

结合外部知识的常识合理性估计方法通过引入多种外部信息源,有效弥补了语言模型内部知识的局限性。常见的外部知识来源包括知识库、外部文档和视觉信息等。对于知识库而言,动态知识库利用语言模型的泛化能力,提供了更灵活且适应性强的常识推理支持。外部文档的方法则通过从大规模无结构化文本中检索相关内容,增强了推理的广度。此外,多模态方法通过结合文本和视觉信息,进一步提升了常识覆盖的全面性。尽管这些方法从不同角度增强了模型的常识估计能力,但如何高效整合不同来源的知识,针对具体任务选择适配的知识来源,仍然是未来研究的核心挑战和关键方向。

2.3 基于微调的方法

是否依赖额外的微调训练是方法之间的另一个核心差异,这不仅影响了方法的训练成本和适用性,也决定了方法对标注数据的依赖程度。因此,从是否依赖微调的角度讨论方法,有助于更清晰地分析这些方法在实际应用中的优缺点和适用场景。其中,依赖微调的方法通过训练数据直接优化模型,从而提升模型在特定任务上的表现。根据使用的监督数据不同,又可以将其进一步划分为:直接在目标任务上训练的有监督学习、基于与目标任务相关的合成数据训练的弱监督学习。接下来,我们将分别介绍微调模型时的训练目标、有监督学习的方法以及弱监督学习的方法。

2.3.1 训练目标及损失函数

正如在第 1 节所提到的,常识合理性任务涉及多类型,包括:分类标签预测、标量数值回归以及概率分布估计。针对不同的任务类型,对应的训练目标也不尽相同,常见目标包括分类目标、回归目标和排序目标,这些训练目标的核心在于损失函数的选择。

分类目标 采用交叉熵损失(Cross Entropy Loss)来最小化预测标签 s 和实际标签 s^* 之间的差异,其中, s_i 是模型对测试集中第 i 个样本的预测标签。

$$L_{CE} = -\frac{1}{N} \sum_{i=1}^N [s_i^* \log(s_i) + (1 - s_i^*) \log(1 - s_i)] \quad (6)$$

其中, N 代表训练数据集中的总样本数量。

回归目标 通常使用均方误差(Mean Squared Error)损失来优化模型,使得模型的输出评分与真实评分之间的差距最小化,如式(7)所示。

$$L_{MSE} = -\frac{1}{N} \sum_{i=1}^N (s_i^* - s_i)^2 \quad (7)$$

排序目标 应用边际排序损失(Margin Ranking Loss)作为损失函数,其核心目标是强制模型为符合常识的样本(正例 D_+)分配比不符合常识的样本(负例 D_-)更高的分数 $s_+ > s_-$,且两者之间的差距至少为预设的边距 ϵ ,如式(8)所示。

$$L_{MR} = \frac{1}{N} \sum_{i=1}^N \max(0, \epsilon - (s_+ - s_-)) \quad (8)$$

对比这几个训练目标,分类目标直接服务于二元合理性判定需求,适用于分类标签预测任务。回归目标实现了对合理性程度的细粒度建模,适用于标量数值回归以及概率分布估计任务。排序目标能够进一步强化样本之间的对比关系,帮助模型学习合理性的分布特征,这一目标通常不会被单独使用,一般作为对回归目标的补充来进一步增强模型在数值回归以及概率分布估计任务上的表现。

2.3.2 有监督训练

在常识合理性估计任务中,使用与测试任务一样的数据来进行训练是一种被广泛采用的策略,通过对模型进行微调以适应特定任务需求。根据模型规模的不同,小语言模型和大语言模型在训练设计上呈现出明显的差异。

小语言模型参数较少,表征能力有限,通常难以直接应对复杂的推理任务。有监督微调成为提升其性能的关键手段^[47,57,90,96]。为了更好地训练模型,

方法的优化主要集中在特征选取和调整训练策略两方面。在特征选取方面,除了利用语言模型生成的句子表征,还结合外部结构化知识(如知识图谱)^[48]或多模态表征^[86],增强模型对常识的理解能力。在训练策略方面,结合对比学习能够增强合理句子和不合理句子之间的表征差异^[45];多任务学习方法则同时在多个数据集上进行训练,共享模型参数,通过这种方式,模型可以从多个任务中学习通用的特征,从而提高在目标任务上的表现^[43-45];辅助任务学习则通过在训练过程中引入一个或多个与目标任务相关的额外任务来增强目标任务的表现,即使这些任务本身不是最终目标,但它们可以提供额外的监督信号,有助于促进模型对多种常识能力的掌握^[70]。

大语言模型凭借更大的参数量和海量预训练数据,具备强大的表征能力和语言理解能力,在复杂的常识推理任务中表现出显著优势。然而,由于其基础能力已经较强,且在预训练过程中已经见过了大部分的公开数据集,简单地在目标数据集上的微调对大语言模型的提升空间相对有限,因此监督微调的研究集中在优化训练数据和训练策略两方面。在数据方面,高质量的微调数据集是提升模型性能的关键,而单纯扩展数据量效果有限^[44]。为了高效生成高质量数据,近年来出现了通过语言模型生成与目标任务相关的多样化样本,之后结合人工筛选的方法。人工筛选的标准通常包括:正确性、多样性、任务相关性和语言流畅性。这种方式能够在利用少量人工成本的情况下构造大量高质量数据^[95]。在训练策略上,大语言模型通常采用强化学习等高级优化技术,以提升推理过程的效率和模块间的协同能力^[55]。

综合来看,小语言模型和大语言模型在有监督训练中的优化方向具有明显差异。小模型通过设计多样化的特征表示和更精细的训练策略来弥补不足,而大模型更依赖高质量数据和复杂训练策略进一步挖掘潜力。小模型在计算效率和资源需求方面占优,适合资源受限的应用场景;大模型则在准确性和泛化能力上表现更强,但对计算资源要求较高。因此,针对不同的任务需求和资源限制,合理选择模型规模和优化策略是提升常识合理性估计性能的关键。

2.3.3 弱监督训练

弱监督训练是一种无须利用目标任务数据来训练模型的方法。在这种训练策略下,模型通过合成

自身数据和相关任务数据来训练。这一策略有效降低了训练对任务标注数据的依赖。

数据合成是弱监督训练的首要核心,其目标是通过自动化手段生成与目标任务相关的训练数据,缓解人工标注数据“数据量不足”以及“场景覆盖有限”的瓶颈。常识知识库^[18,98]以结构化形式存储了大量已验证的常识关系,但这种结构化数据和实际任务中的自然语言数据在形式上有很大差异,因此无法直接用于训练模型。Ma 等人提出通过预定义模板将知识库中的结构化三元组(如“人—能够—奔跑”)转化为符合常识的句子,并通过尾实体替换(如“人—能够—飞翔”)构造反常识样本^[117]。为了进一步提升数据质量和扩大覆盖范围,除实体替换还利用语言模型对知识库概念进行语义泛化,生成涵盖同义词、近义词及关联场景的多样化数据。此外,为了筛选出合成数据中的噪声,Shi 等人^[118]提出基于训练过程动态评估的数据清洗方法,通过监测损失函数变化自动剔除对模型效果无益的低质量数据,以确保合成数据的可靠性和信息量。

除了数据合成,弱监督训练还可通过任务间知识迁移提升模型能力。与有监督训练中的多任务学习或辅助任务学习(需要直接训练目标任务)不同,弱监督迁移学习的核心在于利用相关任务的隐式知识。例如 Huang 等人^[119]将模型在自然语言推理(NLI)任务上训练得到的能力迁移到常识合理性估计,将蕴含关系、矛盾关系等与合理性进行对应,以此让模型完成常识合理性估计推理。

总的来说,弱监督训练通过合成数据的生成与优化以及相关任务的迁移学习,充分利用了低成本数据的优势。尽管弱监督方法在标注依赖和数据覆盖上表现出较大的灵活性,但其效果往往受到数据质量的显著影响。因此,如何提升合成数据的多样性与可靠性,是未来弱监督训练方法研究的重要方向。另外,弱监督训练也可以结合对比学习等辅助任务训练目标,进一步提升模型的表现。

2.3.4 小结

依赖训练的常识合理性估计方法通过模型微调实现对特定任务的优化,从而显著提升推理性能。根据训练方式的不同,这类方法可以分为有监督训练和弱监督训练。有监督训练依赖目标任务的标注数据,主要通过设计有效的特征表示和训练策略来提升小模型的能力,或结合高质量数据和复杂优化技术进一步挖掘大模型的潜力。弱监督训练则通过合成数据和迁移学习等方式,能够在缺乏目标任务

标注数据的前提下对模型性能进行优化。尽管依赖训练的方法能够直接优化模型在特定任务上的性能,但其需要额外的训练开销,未来的研究需要进一步平衡训练成本与性能提升,探索更高效的训练方法。另外,基于特定任务的微调也会损害模型在其他任务上的能力,造成模型的通用性下降,因此是否微调模型的参数需要根据实际的使用场景是专用场景还是通用场景来决定。

2.4 零样本及少样本学习的方法

在常识合理性估计任务中,零样本及上下文少样本学习(In-Context few-shot Learning)的方法逐渐成为研究的热点,尤其是在大模型参数规模突破千亿量级的背景下。一方面,模型的参数爆炸使传统参数微调陷入计算成本困境;另一方面,模型通过预训练阶段已隐式编码海量常识知识,仅需设计合理的推理机制即可激活其判断能力。基于此,这类无须训练的方法能够在训练数据和训练资源稀缺的场景下快速应用于多种任务,展现出较高的应用价值。

在常识合理性估计任务中,零样本及上下文少样本学习方法通过两大互补路径实现知识利用:基于大模型内部知识的直接激活(2.1 节)与结合外部知识的动态增强(2.2 节)。前者直接挖掘模型预训练阶段隐式编码的常识,后者通过检索文档或知识转换扩展模型认知边界,二者共同构建起无须参数更新的轻量化推理体系。

基于内部知识激活的方法,充分调用大模型自身参数存储的常识知识,前文提到的“基于语言模型的概率估计”的方法就是主要技术之一。这类方法通过直接计算文本序列的生成概率,模型可以快速评估输入的常识合理性^[52-53]。生成概率越高,输入内容越符合语言模型在预训练阶段学到的常识。然而,尽管这种方式能够有效评估一些显而易见的常识性判断,但在处理复杂或隐含的推理问题时表现有限,难以捕捉深层次的常识关系。提示词工程(Prompt Engineering)是激活模型内部知识的另一重要技术。通过精心设计提示词,模型可以直接回答句子是否符合常识^[59]。对于这类方法而言,提示词的构造是关键,适当的提示能够显著提升模型的推理能力。通过在提示词中加入任务示例,模型能更好地理解任务模式并生成更符合预期的响应^[120]。采用思维链或知识生成等方式,引导模型在生成最终答案前进行推理步骤或补充知识,也能

有效提高常识合理性评估的准确性^[61,66]。

为扩大模型推理时利用的知识范围,零样本及上下文少样本学习方法也可以结合外部知识增强。在此类方法中,检索到的相关文档可以作为补充的上下文信息输入给语言模型,无须对模型进行微调^[72,81,90-91]。这种方法有效拓展了语言模型的知识来源,弥补模型内在知识的局限性。

总的来说,零样本以及上下文少样本学习的常识合理性估计方法能够绕过传统微调对标注数据和计算资源的强依赖,直接利用预训练阶段获得的固有能力,提供了一种低成本、高灵活性的解决方案。这类方法在许多任务中表现优异,但关键在于如何设计更有效的推理策略以充分挖掘语言模型的潜力。未来研究的重点应放在优化模型的推理流程,以最大化地利用模型已掌握的知识,尽可能发挥模型的能力。

2.5 小结

本节从方法的知识来源和对微调的依赖程度两个核心维度,系统性地梳理了基于预训练模型的文本常识合理性估计方法。通过分析方法的知识来源,我们发现基于模型内部知识的方法具有高效性和简便性,但其推理能力受限于预训练模型的知识覆盖范围;而结合外部知识增强的方法通过引入知

识库、文档或视觉信息,能够显著扩展模型的知识边界并提升推理的准确性和可解释性。此外,从是否依赖微调的角度来看,依赖微调的方法能够通过优化模型参数提升特定任务的性能,但面临计算成本高和标注数据依赖强的挑战;而零样本及少样本学习方法则通过激活模型内部知识或结合外部知识增强,展现了在标注数据稀缺场景下的高灵活性和应用价值。

表 1 和表 2 展示了不同方法的实际性能。由于各个方法针对的数据集各有不同,很难有一个全面的

表 1 最佳方法在分布内及不同类型分布外整合数据集上的实验效果对比

方法	分布内数据	分布外数据 1	分布外数据 2
SKD Critic ^[95]	36.64	38.34	43.40
I2D2 Critic ^[33]	55.03	54.79	67.11
UnifiedQA-v2 ^[47]	56.33	59.73	53.95
ChatGPT ^[4]	62.11	61.20	62.83
Entailer ^[56]	73.79	71.47	70.72
GPT-3.5 ^[121]	75.41	71.03	78.87
GPT-4 ^[5]	72.35	77.40	70.29
Flan-T5 ^[122]	79.50	77.62	78.89
Vera-T5 ^[44]	85.51	81.65	83.37

表 2 各方法在 Com2Sense^[123] 和 CSQA 2.0^[25] 数据集上的结果

方法	基座模型	Com2Sense		CSQA 2.0	
		验证集	测试集	验证集	测试集
ELABOR ^[57]	GPT2-large ^[124]	—	—	58.7	57.6
Krause 等人 ^[59]	GPT-3 ^[122]	58.1	—	54.1	—
GKP ^[66]	GPT-3	61.8	—	59.7	—
CoT ^[61]	GPT-3	61.6	—	59.6	—
Self-consistency ^[65]	GPT-3	61.4	—	60.8	—
I2D2 ^[42]	GPT-3	65.2	—	43.7	—
MAIEUTIC-PROMPTING ^[57]	GPT-3	72.5	75.0	69.5	68.3
DSCQA ^[47]	T5-large ^[125]	59.1	58.5	—	—
RACo ^[81]	T5-large	—	—	60.2	—
UnifiedQA ^[47]	T5-3B ^[125]	75.1	71.3	—	—
Vera ^[44]	Flan-T5-5B ^[122]	85.9	—	68.6	—
UNICORN-v2 ^[43]	T5-11B ^[125]	76.0	—	56.0	—
Entailer ^[56]	T5-11B	82.9	—	56.0	—

续表

方法	基座模型	Com2Sense		CSQA 2.0	
		验证集	测试集	验证集	测试集
CRYSTAL ^[55]	T5-11B	85.9	—	—	—
OCEAN ^[54]	Mistral-7B-v0.2 ^[126]	—	—	69.19	—
Zhou 等人 ^[45]	DeBERTaV3-large ^[127]	—	83.7	—	—
GuideKG ^[58]	Vicuna-13B ^[128]	—	—	66.8	—
CausalGPT ^[63]	Claude ^[129]	83.0	—	—	—
Vacareanu 等人 ^[68]	Falcon ^[130]	—	—	62.2	—
ZEBRA ^[83]	Phi-3-Small ^[131]	—	—	74.6	—
KPE ^[108]	Unicorn-11B ^[43]	—	—	68.4	—

对比,因此本文在表 1 中报告了目前的最佳方法的性能(来自 Vera^[44]论文中的实验结果,评测的数据集更全面),并建议读者在选取方法的时候将表 1 作为主要参考,因为该评估基于一系列数据集的整合,更能够反映出方法的整体性能。其中,分部内数据^[24-26,32-33,102,123,132-139]被用于 Vera 模型的训练,分布外数据未被用于模型训练,其中分布外数据 1^[34,140-144]的任务类型与分部内数据相似,分布外数据 2^[27-28,31,145-148]的任务类型与分部内数据有较大差异。从表 1 可以看出,GPT-3.5^[121]、GPT-4^[5]、Flan-T5^[122]等大语言模型在各类数据集上展现出强大的泛化能力,即使未经过有监督训练,依然取得了优异效果,凸显了大语言模型时代零样本方法的潜力。Vera-T5^[44]通过在大量数据上进行有监督训练,照比 Flan-T5 效果能够进一步提升,整体表现最佳,目前已成为完成常识合理性估计的首选方法。此外,本文也选取 Com2Sense^[123]和 CSQA 2.0^[25]这两个最常用的常识合理性评估数据集作为参考,并在表 2 中对各个方法的结果进行报告(由于数据集的测试集未公开,大部分方法报告的是验证集上的效果)。同时,除了方法框架,基座模型的选择也会对最终效果带来显著影响,因此表 2 中也对各个方法使用的基座模型进行了报告。

3 对自然语言处理任务的作用

可靠的常识合理性估计方法从监督信号和内容过滤两个维度,提升自然语言处理系统的性能。

首先,该方法可作为监督信号引导模型优化方向:传统的文本生成训练方法仅基于极大似然估计优化模型,只能优化生成文本与参考文本的表面相

似度,缺乏对生成内容常识合理性的反馈信号。通过引入常识合理性估计作为额外的监督信号,模型不仅能关注语言流畅度,还能在训练过程中学习到符合常识的生成模式。这种双目标优化机制可有效弥补传统方法对常识评估的忽视,显著提升生成内容的质量和可信度^[16,149]。

更关键的作用在于该方法可作为动态过滤器对模型的生成内容实现质量管控。在大模型时代,语言模型凭借低成本、高吞吐量的生成能力,其输出内容已被广泛应用于训练数据合成、知识库自动化扩充、推理链路生成等场景,但其输出中隐含的常识性错误可能对具体的应用造成风险。通过构建实时评估机制,系统可在三个关键场景中过滤不合理内容,从而进一步提升模型生成内容在实际应用中的可靠性。

保证合成数据质量 数据合成能够有效缓解人工标注数据稀缺的问题,其中基于大语言模型的数据合成已成为重要解决方案。然而,模型生成的合成数据存在隐性常识错误,直接使用可能导致误差传播。常识合理性估计可以作为质量控制工具,帮助筛选和剔除不合理的数据,确保合成数据既满足数量需求,又符合常识标准,从而提升后续模型训练的效果^[150]。

保证扩充知识的质量 传统的知识库构建依赖于人工专家进行手工编制,既费时又费力。基于语言模型的知识库自动化扩充通过挖掘编码在语言模型中的常识知识,能够有效提升知识覆盖范围。但这种自动化生成的知识条目可能存在不符合常识的情况,不仅影响知识库的可靠性,也为结合知识增强的方法带来了负面影响。常识合理性估计通过对生成的三元组进行评估实现知识质量把控^[42,95-96,151]。

提升思维链质量 思维链推理通过逐步推理的方式增强模型对问题逻辑的理解,中间推理步骤的合理性尤为重要。针对思维链推理中的错误传播问题,常识合理性估计方法可以对推理链中的每一步进行合理性评估,确保推理过程中的每个环节都符合常识,避免错误累积。通过这种方式,最终预测结果的可靠性能得到显著提升^[152]。

4 评测数据

在常识合理性估计任务中,选择合适的数据集是评测模型性能和推动研究发展的关键。为了全面评估模型在常识推理中的表现,研究者们设计了多种数据集,这些数据集通常根据涉及的常识类型以

及任务形式的不同而有所差异。通过这些数据集的评测,能够有效地检验模型在处理不同类型的常识问题时的表现,从而为常识合理性估计方法的优化提供实践依据。在本节中,我们将首先对数据集的常见任务形式进行梳理,然后根据常识类型对评测数据集进行系统介绍。另外,我们也将每个数据集的重点信息整理成表格,以方便读者快速地获取关键信息,详见表 3。

4.1 任务形式

常识合理性估计任务涉及多个评测基准,目前常见的任务形式包括:二元分类、分数估计、多项选择和概率分布,这些任务的标注和评测方式各有不同。

表 3 可用于评估和训练常识合理性估计任务的数据集的基本情况

名称	常识类型	任务类型	大小 (训练集/验证集/ 测试集)	标注方式	链接
I2D2 ^[153]	普遍性常识	二元分类	15 000/7 490/—	模型生成,人工过滤	https://huggingface.co/datasets/tasksource/I2D2
FIGURINE ^[153]	语言表达	二元分类	6 304/731/3 062	人工构造	https://leaderboard.allenai.org/figurine/submission
DECO ^[29]	事件常识	分数估计	200/—/—	人工构造	https://github.com/PlusLabNLP/ACCENT/tree/main
ConTurE ^[29]	事件常识	分数估计	100/—/—	人工构造	https://github.com/PlusLabNLP/ACCENT/tree/main
CoDa ^[35]	视觉常识	概率分布	311/103/107	人工构造	https://huggingface.co/datasets/corypaik/coda
PaCo ^[30]	因果关系	多项选择	12 400/—/—	人工构造	https://github.com/luka-group/PaCo
CSQA2.0 ^[25]	普遍性常识	二元分类	9 282/2 544/2 517	人工构造	https://github.com/allenai/twentyquestions
StrategyQA ^[145]	普遍性常识	二元分类	2 290/—/490	人工构造	https://github.com/eladsegal/strategyqa/tree/main/data/strategyqa
NUMERSENSE ^[34]	数值常识	多项选择	—/200/2 200	人工构造	https://github.com/INK-USC/NumerSense/tree/main/data
ViComTe ^[36]	视觉属性	概率分布	4k/—/1k	模型生成,人工过滤	https://github.com/ChenyuHeidiZhang/VL-commonsense
CREAK ^[146]	实体常识	二元分类	10 176/1 371/1 371	人工构造	https://github.com/yasumasaonoe/creak/tree/main/data/creak
COM2 ^[154]	事件常识	多项选择	782 000/—/1 300	模型生成,人工过滤	https://github.com/tqfang/complex-commonsense-reasoning
ROCStories ^[155]	事件常识	多项选择	98 162/—/1 871	人工构造	http://cs.rochester.edu/nlp/rocstories
Com2Sense ^[123]	普遍性常识	二元分类	160 8/804/5 558	人工构造	https://github.com/PlusLabNLP/Com2Sense
CLADDER ^[156]	因果关系	二元分类	10 112/—/—	模型构造	https://huggingface.co/datasets/causal-nlp/CLadder
FeasibilityQA ^[157]	普遍性常识	多项选择	4,608/—/—	人工构造	https://github.com/kevinscaria/feasibilityQA
VIPHY ^[158]	科学常识	概率分布	15 240/8 120/5 7840	来源于其他数据集	https://github.com/Axe-/ViPhy
ComVE ^[26]	普遍性常识	二元分类	10 000/997/1 000	人工构造	https://github.com/wangcunxiang/SemEval2020-Task4
CFC ^[159]	普遍性常识	概率分布	9 999/55/46	模型生成,人工过滤	https://github.com/qxc101/PROBEVALCFC/
TEXT2AFFORD ^[160]	可供性常识	二元分类	—/—/35 520	人工构造	https://github.com/sayantan11995/Text2Afford/tree/master/data
spatialQA ^[140]	视觉常识	二元分类	—/—/20 480	其他数据集	https://github.com/josephcappadona/spatialQA
WINOVENTI ^[161]	因果关系	分数估计	—/—/4 352	人工构造	https://github.com/commonsense-exception/commonsense-exception

续表

名称	常识类型	任务类型	大小 (训练集/验证集 /测试集)	标注方式	链接
NEWTON ^[162]	科学常识	多项选择	—/—/2 800	人工构造	https://newtonreasoning.github.io/
Normlense ^[163]	道德常识	分数估计	—/—/10 000	人工构造	https://seungjuhan.me/normlens
OpenBookQA ^[102]	科学常识	多项选择	4 957/500/2 000	人工构造	https://github.com/allenai/unifiedqa
ARC(easy) ^[132]	科学常识	多项选择	2 251/570/2 281	来源于其他数据集	https://github.com/allenai/unifiedqa
ARC(hard) ^[132]	科学常识	多项选择	1 119/299/1 194	来源于其他数据集	https://github.com/allenai/unifiedqa
AI2Science(elem) ^[132]	科学常识	多项选择	623/123/489	来源于其他数据集	https://github.com/allenai/unifiedqa
AI3Science(middle) ^[132]	科学常识	多项选择	605/125/502	来源于其他数据集	https://github.com/allenai/unifiedqa
CommonsenseQA ^[24]	普遍性常识	多项选择	9 741/1 221/6 099	人工构造	https://github.com/allenai/unifiedqa
QASC ^[133]	科学常识	多项选择	8 134/926/7 408	人工构造	https://github.com/allenai/unifiedqa
PhysicalIQA ^[134]	科学常识	多项选择	16 113/1 838/3 676	人工构造	https://github.com/allenai/unifiedqa
SocialIQA ^[135]	社会知识	多项选择	33 410/1 954/5 861	人工构造	https://github.com/allenai/unifiedqa
Winogrande ^[136]	普遍性常识	多项选择	40 398/1 267/2 534	人工构造	https://github.com/allenai/unifiedqa
SciQ ^[137]	科学常识	多项选择	11 679/1 000/4 000	人工构造	https://allenai.org/data/sciqyes
QuaRel ^[138]	数值常识	多项选择	1 941/278/556	人工构造	https://allenai.org/data/quarelyes
QuaRTz ^[139]	数值常识	多项选择	2 696/384/768	人工构造	https://allenai.org/data/quartzyes
SKD ^[32]	因果关系	二元分类	7 980/1 015/1 015	人工构造	https://github.com/peterwestai2/symbolic-knowledge-distillation
WSC ^[141]	普遍性常识	多项选择	—/273/546	人工构造	https://huggingface.co/datasets/ErnestSDavis/winogradwsc
COPA ^[142]	普遍性常识	多项选择	—/500/1 000	人工构造	https://huggingface.co/datasets/aps/superglue
PROST ^[143]	科学常识	多项选择	—/18 736/74 944	人工构造	https://huggingface.co/datasets/corypaik/prost
Rainer ^[144]	普遍性常识	二元分类	—/591/591	来源于其他数据集	https://github.com/liujch1998/rainier
ULogic ^[164]	因果关系	二元分类	—/—/14 647	模型生成,人工过滤	https://github.com/SiyuanWangw/ULogic/tree/master/Data
SWAG ^[28]	事件常识	多项选择	—/20 006/80 024	来源于其他数据集	https://github.com/rowanz/swagaf/tree/master/data
HellaSwag ^[147]	事件常识	多项选择	—/10 042/40 168	来源于其他数据集	https://github.com/rowanz/hellaswag/tree/master/data
CODAH ^[148]	普遍性常识	多项选择	—/2 776/11 104	人工构造	https://github.com/WebSail-NU/CODAH/tree/master/data
StoryClozeTest ^[27]	事件常识	多项选择	—/1 871/3 742	人工构造	https://cs.rochester.edu/nlp/roctestories/
anLI ^[31]	因果关系	多项选择	—/1 532/3 064	人工构造	https://leaderboard.allenai.org/anli/submissions/get-started

4.1.1 二元分类

作为基础性评估范式,该任务旨在验证模型对确定性事实常识的辨识能力,要求对输入内容进行二元判断。典型案例如判定“一周有 7 天”是否符合常识(标签为“是”)。评估采用准确率指标。这类任务简单直观,适用于需要明确验证合理性的问题,但难以量化常识合理性的细微程度,例如“鸟类会飞”并不完全合理,需排除企鹅等特殊情况。

4.1.2 分数估计

该任务通过连续值评分量化文本符合常识的程度,重点考察模型对模糊性常识的量化建模能力。例如评估“杯子是用玻璃制成的”是否符合常识,答案并不是 100% 的正确,因为也会有陶瓷杯子、塑料杯子等其他情况。与二元分类不同,分数估计能够捕捉合理性的细微差异,因此适用于需要精确评分的场景。模型性能的评估通常通过计算预测值与真

实值之间的误差(如均方误差)来完成。这种任务形式的标注成本较高,但可以为更复杂的推理任务提供支持。

4.1.3 多项选择

多项选择任务要求在多候选答案中选择最优解,验证模型在有干扰选项下的常识推理能力。这类任务不直接询问句子正确与否,需要模型结合常识知识进行推理。以物理常识题为例:“若杯子从桌面滑落,接下来会发生?(A)悬浮空中,(B)摔碎,(C)弹回”,正确答案是 B。常见评测指标是准确率。这种任务形式在现实场景中应用广泛,但由于需要设计明确的问题和答案集合,其数据构建成本相对较高。

4.1.4 概率分布

该任务的目标是为所有可能的答案分配一个合理性的概率分布,反映每个答案的可能性或优先级,适用于存在多重合理答案的开放域场景。这种任务形式不仅关注答案是否正确,还强调答案之间的偏好和差异。例如,预测“香蕉的颜色”的分布时,按概率从高到低排序颜色应该是黄色、绿色、棕色等。评测指标关注模型预测分布与真实分布之间的一致性,如斯皮尔曼相关系数、皮尔森相关系数、KL 散度。概率分布任务适用于对多个可能答案进行细粒度评估的场景,例如涉及属性推理或优先级判断的问题。

4.2 常识类型和数据集

常识数据集可以按照常识的类型进行分类,每种类型的数据集旨在评估模型在特定领域中的推理能力。通过梳理和分类这些数据集,可以更清晰地理解它们的特点与应用场景,进而更好地选择适配的评测基准。以下内容将基于常识类型,对常见的评测数据集进行详细说明。另外,尽管实体常识和科学常识可能超出了传统意义上的常识的“日常经验”范畴,但部分工作^[44,81]也会将这些知识纳入研究范畴,并且实验表明这些知识在常识推理中具有重要作用——掌握这些知识能够为模型提供更全面的推理能力。因此,为了对数据集进行更全面的总结,并确保研究的完整性,本文在这一部分也会介绍关于实体常识和科学常识这两种类型的数据集。

4.2.1 普遍性常识

普遍性常识(General)数据集主要来源于大型知识库,涵盖人类日常生活中的基本常识,而不专门针对某种类型。这类知识包括对世界基本规则的理

解,例如,“鸟类通常会飞翔”或“鱼类生活在水中”。普遍性常识是人类在大多数情境下的基本判断,对人工智能模型的理解能力至关重要。常见数据集有 I2D2^[33]、StrategyQA^[145]、CommonsenseQA^[24]、CSQA2^[25]、FeasibilityQA^[156]、WSC(Winograd Schema Challenge)^[141]、Winogrande^[136]、CODAH^[148]、COPA^[142]、ComVE^[26]、CFC^[158]、Com2Sense^[123]。

4.2.2 事件常识

事件常识(Event)关注事件的发生及其参与者之间的关系,重点评估模型对常见事件的理解能力,包括时间顺序、因果关系以及参与者之间的互动,例如,“一个人想领养猫,但因过敏而被阻碍”。这类知识对于建模复杂的日常情境至关重要,可以帮助模型掌握事件的逻辑和可能的结果。常见数据集有 StoryClozeTest^[27]、ROCStories^[155]、DECO^[29]、COM2^[154]、SocialIQA^[135]、SWAG^[28]、HellaSwag^[147]。

4.2.3 因果关系

因果关系(Causal)数据集旨在评估模型对因果联系的理解和推理能力。这类数据集的目标是帮助模型掌握事件之间的因果逻辑,例如“下雨→路面湿滑→驾驶员需要减速”。因果推理能力是理解复杂情境的关键,在需要理解事件顺序、条件以及后果的任务中,这种能力显得尤为重要。常见数据集有 WINOVENTI^[165]、PaCo^[30]、CLADDER^[121]、SKD(annotated)^[32]、ULogic^[164]、 α NLI^[31]。

4.2.4 数值常识

数值常识(Quantitative)主要评估模型对数值关系、数量概念的掌握程度和量化推理的能力。这类知识涵盖实体的基本数值属性(如“蜘蛛通常有 8 条腿”)以及更复杂的量化推理任务(如“两个橙子与三个苹果相比哪个更重?”)。数值常识是常识推理的重要组成部分,尤其在需要精准计算、逻辑比较或多步推理的任务中具有重要作用。常见数据集有 QUAREL^[138]、Quartz^[139]、Numersense^[34]。

4.2.5 语言表达

语言表达(Idioms & Similes)数据集评估模型对隐喻、比喻及习语等复杂语言现象的理解能力。这类数据集要求模型不仅能识别这些表达的字面含义,还需掌握其隐含含义以及背后的文化背景。例如,“像蜜蜂一样勤劳”不仅意味着勤奋,还暗示了一种积极的评价。FIGURINE^[153]是该类型下的典型数据集。

4.2.6 道德常识

道德常识(Moral)数据集涉及对行为的伦理和

道德的判断,例如理解“偷窃是不道德的”。此类数据集旨在测试模型对社会伦理标准的理解能力,以及在复杂社会情境下进行道德推理的能力,比如 NORMLENS^[163]数据集。

4.2.7 可供性常识

可供性常识(Affordance)数据集关注物体的功能性和用途,评估模型对物体在环境中的交互方式的理解。例如,“椅子可以用来坐下或放置物品”。这类知识有助于模型理解物体用途。Text2Afford^[160]是该类型下的典型数据集。

4.2.8 视觉常识

视觉常识(Visual)数据集测试模型在理解和推理与视觉相关的知识方面的能力,比如对象的空间位置、颜色、大小、形状及其在视觉场景中的交互关系。这类知识主要通过图像、视频或其他视觉内容表现,例如,“杯子通常在桌子上”。常见数据集有 ViComTe^[36]、spatialQA^[140]、CoDa^[35]。

4.2.9 实体常识

实体常识(Entity)聚焦于世界中的具体实体(如物体、人物、地点等)及其属性、类别和实体间相互关系。这类知识旨在帮助模型理解实体的本质和它们在现实中的关联。例如,“爱因斯坦”是一位物理学家,他的主要成就是提出相对论。模型对实体常识的掌握,不仅有助于知识的准确获取,也为更复杂的推理任务(如关系推断和归纳推理)奠定基础。CREAK^[146]是该类型下的典型数据集。

4.2.10 科学常识

科学常识(Scientific)数据集评估模型在理解和应用科学原理、自然规律以及实验方法等方面的能力。这类数据集通常以问答任务的形式测试模型对物理、化学、生物等基础知识的掌握。例如,“植物通过光合作用产生有机物”。常见数据集有 QASC^[133]、NEWTON^[162]、OpenBookQA^[102]、ARC^[132]、AI2 Science(elem)^[132]、AI3 Science(middle)^[132]、Physical IQA^[134]、SciQ^[137]、PROST^[143]、VIPHY^[158]。

5 前景与挑战

前文通过回顾和分析现有的研究成果,我们对当前领域的研究现状有了清晰的了解。基于此,本节将探讨未来的研究方向,并重点关注当前面临的一些挑战。

5.1 零样本估计方法

在大模型参数规模突破千亿量级的背景下,模型的参数爆炸使传统参数微调陷入计算成本困境,但模型通过预训练阶段已隐式编码海量常识知识,仅需设计合理的推理机制即可激活其判断能力。基于此,未来的研究应从如何“训好模型”转变为如何“用好模型”,零样本及上下文少样本学习的方法会越来越成为研究的热点。这些方法无须额外训练,直接利用预训练阶段获得的固有能力,提供了一种低成本、高灵活性的解决方案,展现了巨大的应用潜力。此外,它们还能避免因过度依赖特定数据集而导致的过拟合问题,保持模型的泛化能力。

然而,零样本方法也面临一些挑战。推理能力的局限性是一个关键问题。尽管这些方法在简单任务中表现出色,但在处理复杂推理问题时,往往难以捕捉深层次的常识关系,因此需要设计更有效的推理策略。

此外,零样本方法对提示词的设计过于敏感,不同提示词可能导致显著的性能波动,这增加了实际应用中的不确定性,如何优化模型对提示词的稳定性也是一个重要的研究问题。

幻觉现象是零样本方法面临的另一个重要挑战。大规模预训练语言模型在生成内容时可能出现幻觉,即生成不符合现实或常识的内容,对零样本方法的性能构成了显著影响,因为它直接违背了常识合理性估计的核心目标。

5.2 引入外部知识

正如前文所述,外部知识的有效结合能够显著提升模型的表现,但其选择和使用仍面临许多挑战。首先,何时引入外部知识是一个关键问题。某些任务在生成过程中可能不需要外部知识的支持,盲目地进行知识增强不仅增加了计算开销,甚至可能因为信息冗余和上下文过长损害模型的性能。因此,根据具体输入的适应性引入外部知识的增强,是未来研究的重要课题。

其次,引入什么样的外部知识也需根据具体任务需求进行精准选择。不同任务可能对外部知识的依赖不同,如何根据任务特性和输入的具体情况来选择合适的知识源,仍然是一个值得深入探索的问题。未来的研究应关注如何灵活地结合多源知识,以更好地适应任务的多样性和复杂性。

此外,对于文档数量的选择,目前大多直接采用

top- k 的方式,这种方式简单但不够灵活。未来的研究应探索如何根据任务需求动态调整文档数量,以更好地平衡信息丰富度和计算效率。

最后,文档的压缩也是一个值得关注的问题。模型在处理长上下文时往往会遇到困难,因为长上下文不仅增加了计算复杂度,还可能导致注意力机制无法有效捕捉到关键信息。因此,需要通过有效的压缩策略来减少信息冗余和噪声干扰,同时提高推理效率。

5.3 融合多模态知识

在常识合理性估计任务中,融合多模态知识已成为一个重要的研究方向。随着多模态数据的广泛应用,从不同模态中提取和利用常识知识成为提升模型推理能力的关键。尽管现有方法对如何从其他模态中提取常识知识进行了初步探索,但整体仍处于起步阶段,面临诸多挑战。

首先,模态间的语义鸿沟是一个核心问题。不同模态的特征表示存在显著差异,如何将这些信息统一到一个共同的语义空间中,从而提升多模态知识的融合效果,仍是一个开放性问题。

其次,当前对于其他模态的探索主要局限在图像上,视频数据仍未被充分利用。与单帧图像不同,视频序列不仅能够反映物体和场景的空间结构,还能呈现物体间的动态关系和时序变化,蕴含了更加复杂和深层的常识知识。因此,视频数据的深度利用将大大增强常识推理的表现力,尤其是在动态场景理解和因果推理方面。

此外,需要探索更加灵活的多模态融合机制。目前方法大多采用静态的权重分配机制,即在模型训练时预设各模态的贡献比例,并在推理过程中保持固定。然而,这种静态分配方式忽略了不同模态在不同类型的常识推理任务中的适用性差异,使模型无法根据任务需求进行自适应调整,无法充分发挥各模态的优势。未来的研究应构建更加智能的、常识类型驱动的动态融合框架,使模型能够根据任务需求自适应地调整不同模态的权重。

6 结论

在本文中,我们详细探讨了基于预训练的常识合理性估计方法的多种研究进展。我们首先回顾了常识合理性估计的研究背景和发展历程,明确了常识推理在自然语言处理中的重要地位。随后,从方

法使用的知识来源和训练的监督程度系统地梳理了当前常识合理性估计的主要方法,并讨论了这些方法的实际应用场景。之后,总结了评测基准的任务形式,并按照常识知识类型对相关数据集进行介绍。最后,指出该任务未来可能的研究方向,供研究者参考。

参考文献

- [1] THORNE J, VLACHOS A, CHRISTODOULOPOULOS C, et al. Fever: A large-scale dataset for fact extraction and verification[J]. arXiv preprint arXiv:1803.05355, 2018.
- [2] WADDEN D, LIN S, LO K, et al. Fact or fiction: Verifying scientific claims [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 7534-7550.
- [3] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: Open and efficient foundation language models[J/OL]. ArXiv, 2023, abs/2302.13971.
- [4] OPEN AI. Introducing chatGPT [EB/OL] [2023-01-09]. <https://openai.com/blog/chatgpt>.
- [5] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
- [6] MARCUS G F. The next decade in AI: Four steps towards robust artificial intelligence [J/OL]. ArXiv, 2020, abs/2002.06177.
- [7] ELAZAR Y, KASSNER N, RAVFOGEL S, et al. Measuring and improving consistency in pretrained language models[J/OL]. Transactions of the Association for Computational Linguistics, 2021, 9: 1012-1031.
- [8] MAHOWALD K, IVANOVA AA, BLANK I A, et al. Dissociating language and thought in large language models[J]. Trends in Cognitive Sciences, 2024.
- [9] CHEN J, SHI W, FU Z, et al. Say what you mean!: Large language models speak too positively about negative commonsense knowledge[C]//Proceedings of the 61st Annual Meeting of the Association For Computational Linguistics, 2023.
- [10] LIU Y, WAN Y, HE L, et al. Kg-bart: Knowledge graph-augmented bart for generative commonsense reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021(35): 6418-6425.
- [11] WANG H, LIU Y, ZHU C, et al. Retrieval enhanced model for commonsense generation [C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 3056-3062.
- [12] CUI W, BI K, GUO J, et al. MORE: Multi-modal retrieval augmented generative commonsense reason-

- ing[C]//Proceedings of the Association for Computational Linguistics ACL, 2024: 1178-1192.
- [13] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002: 311-318.
- [14] LIN C Y, HOVY E. Manual and automatic evaluation of summaries [C]//Proceedings of the ACL Workshop on Automatic Summarization, 2002(4): 45-51.
- [15] BANERJEE S, LAVIE A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments [C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005: 65-72.
- [16] TIAN Y, ZHANG F, PENG N. Harnessing black-box control to boost commonsense in LM's generation [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 5417-5432.
- [17] LENAT D B. CYC: A large-scale investment in knowledge infrastructure[J]. Communications of the ACM, 1995, 38(11): 33-38.
- [18] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: An open multilingual graph of general knowledge [C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence, 2017: 4444-4451.
- [19] MILLER G A. WordNet: A lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
- [20] GORDON A, KOZAREVA Z, ROEMMELE M. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning[C]//Proceedings of the SEM: The 1st Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation, 2012: 394-398.
- [21] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, 2019: 4171-4186.
- [22] LIU Y. RoBERTa: A robustly optimized bert pre-training approach [J]. arXiv preprint arXiv:1907.11692, 2019, 364.
- [23] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2024-01-15]. https://www.academia.edu/129447679/Improving_Language_Understanding_by_Generative_Pre_Training_OpenAI.
- [24] TALMOR A, HERZIG J, LOURIE N, et al. CommonsenseQA: A question answering challenge targeting commonsense knowledge [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4149-4158.
- [25] TALMOR A, YORAN O, BRAS R L, et al. CommonsenseQA 2.0: Exposing the limits of AI through gamification[J]. arXiv preprint arXiv:2201.05320, 2022.
- [26] WANG C, LIANG S, JIN Y, et al. Semeval-2020 task 4: Commonsense validation and explanation[J]. arXiv preprint arXiv:2007.00236, 2020.
- [27] MOSTAFAZADEH N, CHAMBERS N, HE X, et al. A corpus and cloze evaluation for deeper understanding of commonsense stories. [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 839-849.
- [28] ZELLERSR, BISKY, SCHWARTZ R, et al. Swag: A large-scale adversarial dataset for grounded commonsense inference [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018: 93-104.
- [29] GHAZARIAN S, SHAO Y, HAN R, et al. Accent: An automatic event commonsense evaluation metric for open-domain dialogue systems [C]//Proceedings of the 61st Annual Meeting of the Association For Computational Linguistics, 2023: 4398-4419.
- [30] QASEMI E, ILIEVSKI F, CHEN M, et al. PaCo: Preconditions attributed to commonsense knowledge [G]. Findings of the Association for Computational Linguistics: EMNLP, 2022: 6781-6796.
- [31] BHAGAVATULA C, BRAS R L, MALAVIYA C, et al. Abductive commonsense reasoning [C]//Proceedings of the International Conference on Learning Representations, 2020.
- [32] WEST P, BHAGAVATULA C, HESSEL J, et al. Symbolic knowledge distillation: From general language models to commonsense models [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024: 4602-4625.
- [33] BHAGAVATULA C, HWANG J D, DOWNEY D, et al. I2d2: Inductive knowledge distillation with neurologic and self-imitation [C]//Proceedings of the 61st Annual Meeting of the Association for Computa-

- tional Linguistics, Association for Computational Linguistics, 2023: 9614-9630.
- [34] LIN B Y, LEE S, KHANNA R, et al. Birds have four legs?!: NumerSense: Probing numerical commonsense knowledge of pre-trained language models [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 6862-6868.
- [35] PAIK C, AROCA OUELLETTE S, RONCONE A, et al. The world of an octopus: How reporting bias influences a language model's perception of color [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2021: 823-835.
- [36] ZHANG C, VAN DURME B, LI Z, et al. Visual commonsense in pretrained unimodal and multimodal models [C]//Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5321-5335.
- [37] DAVIS E. Benchmarks for automated commonsense reasoning: A survey[J]. *ACM Computing Surveys*, 2023, 56(4): 1-41.
- [38] BHARGAVA P, NG V. Commonsense knowledge reasoning and generation with pre-trained language models: A survey [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(11): 12317-12325.
- [39] SAP M, SHWARTZ V, BOSSELU T A, et al. Commonsense reasoning for natural language processing [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, 2020: 27-33.
- [40] STORKS S, GAO Q, CHAI J Y. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches[J]. *arXiv preprint arXiv:1904.01172*, 2019: 1-60.
- [41] HUANG X, WANG Y, ZHAO J, et al. Commonsense knowledge editing based on free-text in LLMs [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024: 14870-14880.
- [42] BHAGAVATULA C, HWANG J D, DOWNEY D, et al. I2D2: Inductive knowledge distillation with neuro logic and self-imitation [C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 9614-9630.
- [43] LOURIE N, LE BRAS R, BHAGAVATULA C, et al. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(15): 13480-13488.
- [44] LIU J, WANG W, WANG D, et al. Vera: A general-purpose plausibility estimation model for commonsense statements [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics, 2023: 1264-1287.
- [45] ZHOU Y, HAN Y, ZHOU H, et al. Advancing transformer's capabilities in commonsense reasoning [J/OL]. *arXiv preprint arXiv:2310.06803*, 2023.
- [46] BAO M, LIU Q, ZHANG K, et al. Keep skills in mind: Understanding and implementing skills in commonsense question answering [C/OL]//Proceedings of the 32nd International Joint Conference on Artificial Intelligence. Macau, SAR China: International Joint Conferences on Artificial Intelligence Organization, 2023: 5012-5020.
- [47] KHASHABI D, KORDI Y, HAJISHIRZI H. UnifiedQA-v2: Stronger generalization via broader cross-format training[J]. *arXiv preprint arXiv:2202.12359*, 2022.
- [48] LIU Y, ZHANG L, HAN W, et al. Constrained text generation with global guidance: Case study on common generation [J/OL]. *arXiv preprint arXiv:2103.07170*, 2021.
- [49] TRINH T H, LE Q V. A simple method for commonsense reasoning [J/OL]. *arXiv preprint arXiv:1806.02847*, 2019.
- [50] TAMBORKINO A, PELLICANO N, PANNIK R B, et al. Pre-training is (almost) all you need: An application to commonsense reasoning [C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020: 3878-3887.
- [51] HOLTZMAN A, WEST P, SHWARTZ V, et al. Surface form competition: Why the highest probability answer isn't always right [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 7038-7051.
- [52] DAVISO N J, FELDMAN J, RUSH A. Commonsense knowledge mining from pretrained models [C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: Association for Computational Linguistics, 2019: 1173-1178.

- [53] PENG L, LI Z, ZHAO H. Evaluate confidence instead of perplexity for zero-shot commonsense reasoning[J/OL]. arXiv preprint arXiv:2208.11007, 2022.
- [54] WU J, LI X, WANG R, et al. OCEAN: Offline Chain-of-thought evaluation and alignment in large language models[J/OL]. arXiv preprint arXiv:2410.23703, 2024.
- [55] LIU J, PASUNURU R, HAJISHIRZI H, et al. Crystal: Introspective reasoners reinforced with self-feedback[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 11557-11572.
- [56] TAFJORD O, DALVI MISHRA B, CLARK P. Entailer: Answering questions with faithful and truthful chains of reasoning[C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 2078-2093.
- [57] WANG W, SRIKUMAR V, HAJISHIRZI H, et al. Elaboration-generating commonsense question answering at scale[C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 1619-1635.
- [58] WEI X, CHEN H, YU H, et al. Guided knowledge generation with language models for commonsense reasoning[C]//Proceedings of the Association for Computational Linguistics, Miami, Florida, USA: Association for Computational Linguistics, 2024: 1103-1136.
- [59] KRAUSE S, STOLZENBURG F. From data to commonsense reasoning: The use of large language models for explainable AI[J/OL]. arXiv preprint arXiv:2407.03778, 2024.
- [60] FENG Y, ZHOU B, LIN W, et al. BIRD: A trustworthy bayesian inference framework for large language models[J/OL]. arXiv preprint arXiv:2404.12494, 2024.
- [61] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J/OL]. arXiv preprint arXiv:2201.11903, 2023.
- [62] DAI L, QIN K. Tofc: Tree-of-fact with continued best-first search for commonsense reasoning[C]//Proceedings of the International Joint Conference on Neural Networks. IEEE, 2024: 1-8.
- [63] TANG Z, WANG R, CHEN W, et al. Towards CausalGPT: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in LLMs[J/OL]. arXiv preprint arXiv:2308.11914, 2023.
- [64] GOLOVNEVA O, O'BRIEN S, PASUNURU R, et al. PathFinder: Guided search over multi-step reasoning paths[J/OL]. arXiv preprint arXiv:2312.05180, 2023.
- [65] WANG X, WEI J, SCHUURMAN SD, et al. Self-consistency improves chain of thought reasoning in language models[J/OL]. arXiv:2203.11171, 2023.
- [67] JUNG J, QIN L, WELLECK S, et al. Maieutic prompting: Logically consistent reasoning with recursive explanations[C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 1266-1279.
- [66] LIU J, LIU A, LU X, et al. Generated knowledge prompting for commonsense reasoning[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 3154-3169.
- [68] VACAREANU R, PRATIK A, SPILIOPOULOU E, et al. General purpose verification for chain of thought prompting[J/OL]. arXiv preprint arXiv:2405.00204, 2024.
- [69] WANG W, FANG T, DING W, et al. Car: Conceptualization-augmented reasoner for zero-shot commonsense question answering[C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2023: 13520-13545.
- [70] XIA J, WU C, YAN M. Incorporating relation knowledge into commonsense reading comprehension with multi-task learning[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 2393-2396.
- [71] LIN B Y, CHEN X, CHEN J, et al. KagNet: Knowledge-aware graph networks for commonsense reasoning[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 2829-2839.
- [72] BAUER L, WANG Y, BANSAL M. Commonsense for generative multi-hop question answering tasks[C/OL]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018: 4220-4230.
- [73] GABRIEL S, BOSSELUT A, DA J, et al. Discourse understanding and factual consistency in abstractive summarization[C/OL]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics,

- 2021: 435-447.
- [74] CHAKRABARTY T, CHOI Y, SHWARTZ V. It's not rocket science: Interpreting figurative language in narratives[J/OL]. Transactions of the Association for Computational Linguistics, 2022, 10: 589-606.
- [75] SU Y, ZHANG J, SONG Y, et al. PipeNet: Question answering with semantic pruning over knowledge graphs[C]//Proceedings of the 13th Joint Conference on Lexical and Computational Semantics, 2024: 360-371.
- [76] CHOI B, LEE Y, KYUNG Y, et al. ALBERT with knowledge graph encoder utilizing semantic similarity for commonsense question answering[J/OL]. Intelligent Automation & Soft Computing, 2023, 36(1): 71-82.
- [77] CHEN W, RAVI S, SHWARTZ V. CASE: Commonsense-augmented score with an expanded answer space[C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2023: 2732-2744.
- [78] BOSSELUT A, RASHKIN H, SAP M, et al. COMET: Commonsense transformers for automatic knowledge graph construction[C/OL]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, 2019: 4762-4779.
- [79] PARK J H, LEE M, KIM J, et al. COCONUT: Contextualized commonsense unified transformers for graph-based commonsense augmentation of language models[C/OL]//Proceedings of the Association for Computational Linguistics: ACL, Bangkok, Thailand: Association for Computational Linguistics, 2024: 5815-5830.
- [80] GHAZARIAN S, SHAO Y, HAN R, et al. ACCENT: An automatic event commonsense evaluation metric for open-domain dialogue systems[C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 4398-4419.
- [81] YU W, ZHU C, ZHANG Z, et al. Retrieval augmentation for commonsense reasoning: A unified approach[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 4364-4377.
- [82] ZHAO Z, HU L, ZHAO H, et al. Knowledgeable parameter efficient tuning network for commonsense question answering[C/OL]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada: Association for Computational Linguistics, 2023: 9051-9063.
- [83] MOLFESE F, CONIA S, ORLANDO R, et al. ZEBRA: Zero-shot example-based retrieval augmentation for commonsense question answering[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2024: 22429-22444.
- [84] YUN T, SUN C, PAVLICK E. Does vision-and-language pretraining improve lexical grounding? [C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2021: 4357-4366.
- [85] TANG Z, CHO J, TAN H, et al. VidLanKD: Improving language understanding via video-distilled knowledge transfer[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021, 34: 24468-24481.
- [86] GUO H, ZHOU K, ZHAO W X, et al. Visually-augmented pretrained language models for NLP tasks without images[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023: 14912-14929.
- [87] YANG H Y, SILBERER C. Are visual-linguistic models commonsense knowledge bases? [C]//Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022: 5542-5559.
- [88] WANG W, DONG L, CHENG H, et al. Visually-augmented language modeling[J]. arXiv preprint arXiv:2205.10178, 2022.
- [89] YANG Y, YAO W, ZHANG H, et al. Z-LaVI: Zero-shot language solver fueled by visual imagination [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 1186-1203.
- [90] CHEN Q, JI F, CHEN H, et al. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources[C]//Proceedings of the 28th International Conference on Computational Linguistics, 2020: 2583-2594.
- [91] XU Y, ZHU C, WANG S, et al. Human parity on-commonsenseQA: Augmenting self-attention with external attention[J/OL]. arXiv preprint arXiv:2112.03254, 2022.
- [92] ZOU A, ZHANG Z, ZHAO H, DECKER. Double check with heterogeneous knowledge for commonsense fact verification[C]//Proceedings of the Association for Computational Linguistics: ACL, 2023: 11891-11904.
- [93] ZHOU P, GOPALAKRISHNAN K, HEDAYAT-NIA B, et al. Commonsense-focused dialogues for response generation: An empirical study[C]//Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2021: 121-132.

- [94] ZHANG Z, ZHANG X, REN Y, et al. IAG: Induction-augmented generation framework for answering reasoning questions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 1-14.
- [95] WEST P, BHAGAVATULA C, HESSEL J, et al. Symbolic knowledge distillation: From general language models to commonsense models[C/OL]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle, United States: Association for Computational Linguistics, 2022: 4602-4625.
- [96] ZHANG Y, LI L, WU Y, et al. Alleviating the knowledge-language inconsistency: A study for deep commonsense knowledge[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 594-604.
- [97] CLARK K. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.
- [98] ABDOU M, RAVISHANKAR V, BARRETT M, et al. The sensitivity of language models and humans to winograd schema perturbations[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020: 7590-7604.
- [99] SAP M, LE BRAS R, ALLAWAY E, et al. ATOMIC: An atlas of machine commonsense for if-then reasoning[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019(33): 3027-3035.
- [100] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]//Proceedings of the Semantic Web: 15th International Conference, Heraklion, Crete, Greece, Springer, 2018: 593-607.
- [101] HAVASI C, SPEER R, ARNOLD K, et al. Open mind common sense: Crowd-sourcing for common sense[C]//Proceedings of the 2nd AAAI Conference on Collaboratively-Built Artificial Intelligence, 2010.
- [102] MEYER C M, GUREVYCH I. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography[M]. Electronic Lexicography. Oxford University Press, 2012: 259-292.
- [103] MIHAYLOV T, CLARK P, KHOT T, et al. Can a suit of armor conduct electricity?: A new dataset for open book question answering [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2018.
- [104] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering?: Try arc, the AI2 reasoning challenge[J]. arXiv preprint arXiv: 1803.05457, 2018.
- [105] KHOT T, CLARK P, GUERQUIN M, et al. QASC: A dataset for question answering via sentence composition[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020 (34): 8082-8090.
- [106] BHAKTHAVATSALAM S, ANASTASIADES C, CLARK P. GenericsKB: A knowledge base of generic statements [J]. arXiv preprint arXiv: 2005.00660, 2020.
- [107] XIAO S, LIU Z, ZHANG P, et al. C-pack: Packed resources for general chinese embeddings[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024: 641-649.
- [108] IZACARD G, CARON M, HOSSEINI L, et al. Unsupervised dense information retrieval with contrastive learning [J]. arXiv preprint arXiv: 2112.09118, 2021.
- [109] LIU X, YIN D, FENG Y, et al. Things not written in text: Exploring spatial commonsense from visual signals[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022: 2365-2376.
- [110] ZHANG C, VAN DURME B, LI Z, et al. Visual commonsense in pretrained unimodal and multimodal models[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022: 5321-5335.
- [111] GORDON J, VAN DURME B. Reporting bias and knowledge acquisition [C]//Proceedings of the Workshop on Automated Knowledge Base Construction, 2013: 25-30.
- [112] YAO Y, YU T, ZHANG A, et al. Visually grounded commonsense knowledge acquisition [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2023(37): 6583-6592.
- [113] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//Proceedings of the International Conference on Machine learning, PMLR, 2021: 8748-8763.
- [114] JIA C, YANG Y, XIA Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the International Conference on Machine Learning, PMLR, 2021: 4904-4916.

- [115] LI J, LI D, SAVARESE S, et al. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the International Conference on Machine Learning, PMLR, 2023: 19730-19742.
- [116] LIU H, LI C, WU Q, et al. Visual instruction tuning[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, 36: 34892-34916.
- [117] MA K, ILIEVSKI F, FRANCIS J, et al. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(15): 13507-13515.
- [118] SHI H, WANG W, FANG T, et al. QADYNAMICS: Training dynamics-driven Synthetic QA diagnostic for zero-shot commonsense question answering[C]//Proceedings of the Association for Computational Linguistics: EMNLP, 2023: 15329-15341.
- [119] HUANG C, HE W, LIU Y. Improving unsupervised commonsense reasoning using knowledge-enabled natural language inference[C/OL]//Proceedings of the Association for Computational Linguistics: EMNLP. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021: 4875-4885.
- [120] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the Advances in Neural Information Processing Systems. Curran Associates, Inc., 2020 (3): 1877-1901.
- [121] OPENAI. OpenAI's GPT-3. 5 model [EB/OL]. [2025-01-09]. <https://platform.openai.com/docs/models/gpt-3-5>.
- [122] CHUNG H W, HOU L, LONGPER S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.
- [123] SINGH S, WEN N, HOU Y, et al. Com2Sense: A commonsense reasoning benchmark with complementary sentences[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 883-898.
- [124] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1(8): 9-21.
- [125] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research, 2020, 21(140): 1-67.
- [126] JIANG, ALBERT Q, et al. Mistral 7B[J]. arXiv preprint arXiv:2310.06825, 2023.
- [127] HE P, GAO J, CHEN W. DeBERTAV3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing[J]. arXiv preprint arXiv:2111.09543, 2021.
- [128] ZHENG L, CHIANG W L, SHENG Y, et al. Judgingl lm-as-a-judge with mt-bench and chatbot arena[J]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, 36: 46595-46623.
- [129] BAI Y, KDDAVATH S, KUNDU S, et al. Constitutional AI: Harmlessness from AI feedback[J]. arXiv preprint arXiv:2212.08073, 2022.
- [130] ALMAZROUEI E, ALOBEIDLI H, ALSHAMSI A, et al. The falcon series of open language models [J]. arXiv preprint arXiv:2311.16867, 2023.
- [131] ABDIN M, ANEJA J, AWADALLA H, et al. Phi-3 technical report: A highly capable language model locally on your phone[J]. arXiv preprint arXiv:2404.14219, 2024.
- [132] CLARK P, COWHEY I, ETZIONI O, et al. Think you have solved question answering?: Try arc, the AI2: Reasoning challenge[J]. arXiv preprint arXiv:1803.05457, 2018.
- [133] KHOT T, CLARK P, GUERQUIN M, et al. QASC: A dataset for question answering via sentence composition [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8082-8090.
- [134] BISK Y, ZELLERS R, GAO J, et al. PiQA: Reasoning about physical commonsense in natural language[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 7432-7439.
- [135] SAP M, RASHKIN H, CHEN D, et al. SocialiQA: Commonsense reasoning about social interactions [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 4463-4473.
- [136] SAKAGUCHI K, BRAS R L, BHAGAVATULA C, et al. Winogrande: An adversarial winograd schema challenge at scale[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8732-8740.
- [137] WELBL J, LIUN F, GARDNER M. Crowd sourcing multiple choice science questions[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 94-106.
- [138] TAFJORD O, CLARK P, GARDNER M, et al. Quarel: A dataset and models for answering questions about qualitative relationships[J]. Proceedings

- of the AAAI Conference on Artificial Intelligence, 2019, 33: 7063-7071.
- [139] TAFJORD O, GARDNER M, LIN K, et al. Quartz: An open-domain dataset of qualitative relationship questions[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5941-5946.
- [140] CAPPADONA J, CHEN Y, LAN J, et al. Language models are unreliable spatial reasoners[EB/OL].[2024-01-15].<https://jcappadona.com/assets/files/spatialQA.pdf>.
- [141] LEVESQUE H J. The winograd schema challenge [C]//Proceedings of the 27th Conference on Innovative Applications of Artificial Intelligence, 2015.
- [142] GORDON A, KOZAREVA Z, ROEMMELE M. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning[C]//Proceedings of the 1st Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the 6th International Workshop on Semantic Evaluation, 2012: 394-398.
- [143] AROCA OUELLETTE S, PAIK C, RONCONE A, et al. PROST: Physical reasoning about objects through space and time[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 4597-4608.
- [144] LIU J, HALLINAN S, LU X, et al. Rainier: Reinforced knowledge introspector for commonsense question answering[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2022: 8938-8958.
- [145] GEVA M, KHASHABI D, SEGAL E, et al. Did aristotle use a laptop?: A question answering benchmark with implicit reasoning strategies[J]. Transactions of the Association for Computational Linguistics, 2021, 9: 346-361.
- [146] ONOE Y, ZHANG M J Q, CHOI E, et al. Creak: A dataset for commonsense reasoning over entity knowledge[J]. arXiv preprint arXiv:2109.01653, 2021.
- [147] ZELLERS R, HOLTZMAN A, BISK Y, et al. Hel-laSwag: Can a machine really finish your sentence? [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 4791-4800.
- [148] CHEN M, D'ARCY M, LIU A, et al. CODAH: An adversarially-authored question answering dataset for common sense[C]//Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, 2019: 63-69.
- [149] COLON HERNANDEZ P, LIEBERMAN H, XIN Y, et al. Adversarial transformer language models for contextual commonsense inference[J/OL]. arXiv preprint arXiv:2302.05406, 2023.
- [150] MIRANDA I, SALABERRIA A, AGIRRE E, et al. BIVLC: Extending vision-language compositionality evaluation with text-to-image retrieval [C]//Proceedings of the 38th International Conference on Neural Information Processing Systems, 2025, 37: 101880-101904.
- [151] WANG W, FANG T, LI C, et al. CANDLE: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 2351-2374.
- [152] RAMNATH S, JOSHI B, HALLINAN S, et al. Tailoring self-rationalizers with multi-reward distillation[C]//Proceedings of the ICLR Workshop on Secure and Trustworthy Large Language Models, 2024.
- [153] CHAKRABARTY T, CHOI Y, SHWARTZ V. It's not rocket science: Interpreting figurative language in narratives[J]. Transactions of the Association for Computational Linguistics, 2022, 10: 589-606.
- [154] FANG T, CHEN Z, SONG Y, et al. Complex reasoning over logical queries on commonsense knowledge graphs[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 11365-11384.
- [155] CHEN J, CHEN J, YU Z. Incorporating structured commonsense knowledge in story completion[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019(33): 6244-6251.
- [156] JIN Z, CHEN Y, LEEB F, et al. Cladder: Assessing causal reasoning in language models [C]//Proceedings of the 37th Conference on Neural Information Processing Systems, 2023.
- [157] GUPTA H, VARSHNEY N, MISHRA S, et al. John is 50 years old, can his son be 65?: Evaluating NLP models' understanding of feasibility[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2023: 407-417.
- [158] SINGH S, QASEMI E, CHEN M. VIPHY: Probing "visible" physical commonsense knowledge [C]//Proceedings of the Association for Computational Linguistics, 2023: 7113-7128.

- [159] CHENG Q, BORATKO M, YELUGAM P K, et al. Every answer matters: Evaluating commonsense with probabilistic measures[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 493-506.
- [160] ADAK S, AGRAWAL D, MUKHERJEE A, et al. Text2Afford: Probing object affordance prediction abilities of language models solely from text[C]//Proceedings of the 28th Conference on Computational Natural Language Learning, 2024: 342-364.
- [161] DO N, PAVLICK E. Are rotten apples edible?: Challenging commonsense inference ability with exceptions[C]//Proceedings of the Association for Computational Linguistics: ACL-IJCNLP, 2021: 2061-2073.
- [162] WANG Y, DUAN J, FOX D, et al. Newton: Are large language models capable of physical reasoning? [C]//Proceedings of the Association for Computational Linguistics, 2023: 9743-9758.
- [163] HAN S, KIM J, HESSEL J, et al. Reading books is great, but not if you are driving!: Visually grounded reasoning about defeasible commonsense norms [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023: 894-914.
- [164] WANG S, WEI Z, CHOI Y, et al. Can LLMs reason with rules?: Logic scaffolding for stress-testing and improving LLMs[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024: 7523-7543.



崔莞清(1996—), 博士, 主要研究领域为常识推理、多模态。

E-mail: cuiwanqing18z@ict.ac.cn



郭嘉丰(1980—), 博士, 研究员, 主要研究领域为数据挖掘、信息检索。

E-mail: guojiafeng@ict.ac.cn



毕可平(1989—), 博士, 副研究员, 主要研究领域为信息检索、可信人工智能。

E-mail: bikeping@ict.ac.cn