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：

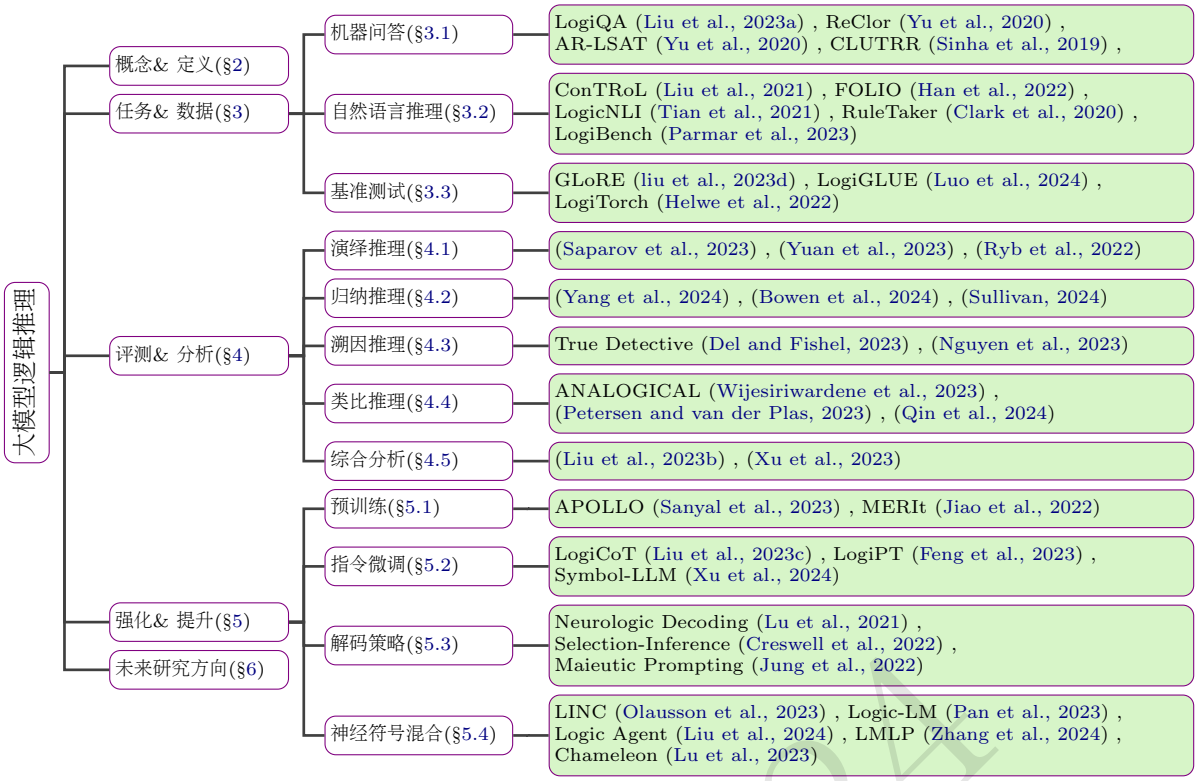


Figure 1: 本综述的组织结构。

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

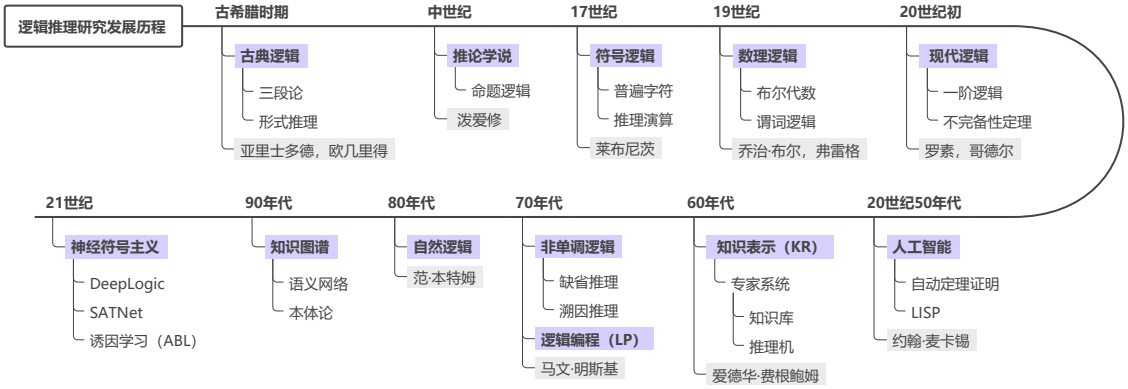


Figure 2: 人工智能逻辑相关理论发展时间线。

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

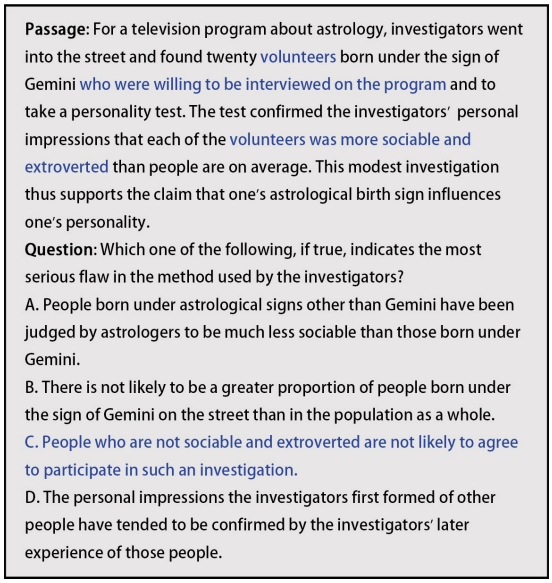
机器问答是评估大模型逻辑推理能力的常用途径。该任务通过提供一段上下文文本，以及与该文本相关的问题，要求模型给出相应的答案。回答的方式主要包括多项选择、片段抽取和自由回答。在这些形式中，多项选择题目因其标准化程度高而被广泛采用于逻辑推理能力的测评。图3a展示了一个典型的多项选择形式的机器问答题目。在这个例子中，模型需要处理自然语言的输入，理解上下文，并运用逻辑推理来选择正确的答案。

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。

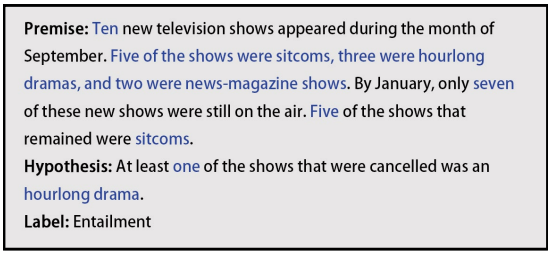
ReClor (Yu et al., 2020)数据集来源于美国研究生管理科入学考试（GMAT）题目，采用与LogiQA相似的四选项多项选择题形式。该数据集为英文，包含6,138条数据。

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.