基于检索增强生成的两阶段常识推理方法

李东洋[1,2]，袁志勇[1,2]，车超[1,2] *
[1]大连大学，软件工程学院，辽宁大连，116622
[2]大连大学，先进设计与智能计算省部共建教育部重点实验室，辽宁大连，116622
lidongyang9527@foxmail.com, yuanzhiyong1999@163.com, chechao@dlu.edu.cn

摘要

常识推理任务是指模型利用日常经验知识对隐含信息进行推断，从而理解和预测现实世界中的合理情境。当前研究趋势之一是通过引入外部知识库来获得额外的背景知识。然而现有的常识推理模型存在引入的外部信息不够精准和融合不充分的问题，致使其在实际应用中的表现不佳。针对上述问题，本文提出了一种基于检索增强生成的两阶段常识推理方法。该方法基于维基百科构建了包含6.28M篇文章的知识库，使用检索增强生成方法，赋予模型语义相关的上下文作为补充信息，辅助模型推理。同时，为了节省时间和资源，本文提出了一种两阶段推理策略，将简单问题交由小模型处理，将复杂问题交由大模型完成。在OpenBookQA等多个数据集上的实验结果证明，本文方法展现出优越的性能，而且适配不同的骨干网络和大模型，可做到即插即用。

关键词： 常识推理；检索增强生成；问答系统；大语言模型

A Two-Stage Commonsense Reasoning Approach Based on Retrieval Augmented Generation

Dongyang Li[1,2], Zhiyong Yuan[1,2], Chao Che[1,2] *
[1]School of Software Engineering, Dalian University, Dalian, Liaoning, 116622, China
[2]Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, Dalian University, Dalian, Liaoning, 116622, China
lidongyang9527@foxmail.com, yuanzhiyong1999@163.com, chechao@dlu.edu.cn

Abstract

Commonsense reasoning refers to the task where a model leverages everyday knowledge to infer implicit information, thereby understanding and predicting plausible scenarios in the real world. A prevailing research trend is to incorporate external knowledge bases to enrich background information. However, existing commonsense reasoning models often suffer from inaccurate retrieval and insufficient integration of external knowledge, which limits their practical effectiveness. To address these issues, we propose A Two-Stage Commonsense Reasoning Approach Based on Retrieval Augmented Generation(TSCR). Specifically, we construct a large-scale knowledge base containing 6.28 million Wikipedia articles, and employ a retrieval-augmented generation framework to provide the model with semantically relevant contextual information to support

* 通讯作者
基金项目：国家自然科学基金(62076045)
©2025 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

reasoning. Furthermore, to improve efficiency in terms of time and resources, we introduce a two-stage reasoning strategy that delegates simple questions to a smaller model and assigns complex ones to a larger model. Experimental results on several datasets, including OpenBookQA, demonstrate that the proposed method exhibits superior performance. Moreover, it is adaptable to different backbone networks and large models, making it plug-and-play.

Keywords: Commonsense Reasoning , Retrieval Augmented Generation , Question Answering System , Large Language Model

1 引言

在大数据时代背景下，问答系统凭借其智能化与个性化优势，逐渐成为满足用户多样化信息需求的关键手段。由于人类社会面临的问题愈发复杂(袁毓林and 卢达威, 2018)，问答系统在回答不同问题时就需要结合不同的科学事实、日常习惯等背景知识。这种隐性知识虽不起眼，但已超越了自然语言理解系统现有的能力范畴。鉴于此，常识推理(Bhargava and Ng, 2022)任务应运而生，并成为检验模型或者机器推理能力最简单、最直观的方法。不论是传统预训练语言模型，还是展现出强大能力的大语言模型（Large Language Model, LLM），都会采用该类任务来检验模型的推理能力(Touvron et al., 2023)。

使用预训练语言模型蕴含的潜在知识弥补推理中缺乏上下文参考信息的问题，可有效提高模型推理能力。但预训练语言模型蕴含的知识量与其参数量相关，较小体量模型的知识覆盖面相对较窄，无法很好地补充常识推理问题中的背景信息，较大体量的模型又会带来额外的推理成本；同时传统的编码器-解码器模型在处理多选问答时，通常将问题与选项拼接后输入模型直接生成答案，该方法与微调编码器类预训练模型效果相近，实质上将解码器简化为分类器，限制了其生成与推理能力，难以胜任常识推理任务(Liu et al., 2021)。

LLM的出现，推动了自然语言处理领域的进一步发展，可以利用其强大的提示词，让模型遵循指令，进而完成各类任务。但种种实践表明，LLM对提示词的要求很高，而且不同的提示词会对任务结果造成很大的影响，如何找到适合当前任务的提示词是一大难题。同时，LLM也由于其体量庞大的原因，占用大量的资源，导致使用门槛和成本较高。此外，模型如何准确识别和利用已有知识回答问题也是当前研究的重点。

针对上述问题，本文提出了基于检索增强生成的两阶段常识推理方法（A Two-Stage Commonsense Reasoning Approach Based on Retrieval Augmented Generation, TSCR）。TSCR基于Wikipedia (Napoles and Dredze, 2010)构建了包含6.28M篇文章的向量检索库，并构建文章级索引和句子级索引的双重索引，实现了高效的语义检索。不同于完全使用LLM推理，本文设计了小语言模型推理模块（Small Language Model Reasoning, SLMR）和大语言模型推理模块（Large Language Model Reasoning, LLMR）。为了实现更高效地推理，TSCR通过阈值判断问题难度，将整个推理过程分为两个步骤：首先使用SLMR推理问题，若推理结果的置信度小于阈值ϵ，则使用LLMR再次推理，最终答案以LLMR的结果为准。TSCR在节省资源的同时，提高了模型推理效果，结合语义检索技术，也提高了模型的可解释性。除此之外，TSCR适用于不同的骨干网络和LLM，可做到即插即用，具有较好的可扩展性。本文的主要贡献如下：

- 基于Wikipedia构建了包含6.28M篇文章的大规模向量检索库，并设计了文章级与句子级的双重索引结构，显著提升了语义检索的效率与准确性。
- 设计了SLMR与LLMR两个模块，并通过置信度阈值动态判断问题难度，实现了推理效率与准确率的双重提升，同时降低了对计算资源的依赖。
- 提出了TSCR方法，其不仅提升了常识推理任务的性能和效率，还增强了模型的可解释性，并具备良好的通用性和灵活性，支持即插即用。

2 相关工作

2.1 基于预训练语言模型的常识推理方法

在常识推理任务中，最直接、简单的方式是使用BERT (Devlin et al., 2019)及其多种变体，如RoBERTa (Liu et al., 2019)、ALBERT (Lan et al., 2019)、DeBERT (He et al.,

(c) 2025 中国中文信息学会计算语言学专业委员会

2020)、SentiBERT (Yin et al., 2020)、KnowBERT (Peters et al., 2019)等进行分类，即可取得超越传统方法的表现。但此类方法未能充分利用预训练语言模型的表达能力，且小模型所蕴含的知识有限，难以胜任常识推理任务。此外，为保持输入一致性，需为每个选项分别构造输入，显著增加训练和推理开销。同时，一些研究者选择编码器-解码器类预训练语言模型进行下游任务微调，如T5 (Raffel et al., 2020)、BART (Lewis et al., 2019)等。 Lan et al. (2019)提出UnifiedQA方法，将问题和所有候选答案串联后输入到T5模型，学习从输入中提取片段来生成正确的选择。Zhu等人 (Zhu et al., 2019)在训练阶段采用对抗训练算法来增强模型的鲁棒性和泛化能力。

2.2 引入外部知识的常识推理方法

预训练的文本语言模型在很多应用中表现良好，但因为自然语言中存在大量使用频率低的词语（长尾效应），模型很难全面掌握这些低频信息。为了解决这个问题，越来越多的研究开始引入外部知识库，帮助模型获取更多背景知识，提升对常识的理解能力，也能缓解长尾效应带来的影响。根据数据的组织方式，外部知识库通常分为结构化和非结构化两类。

常见的结构化知识库有Wikidata、ConceptNet (Liu and Singh, 2004)等，这些知识库包含丰富的结构化背景知识，大多以知识图谱的形式出现。以此为基础， Lin et al. (2019)提出了一个基于ConceptNet的文本推理框架，通过GCN与层次路径注意力机制对知识子图进行建模与推理优化； Wang et al. (2020)通过命名实体识别获取问题与选项中的实体，在ConceptNet中采用随机游走采样多跳知识路径，并结合改进的关系网络注意力机制与ALBERT编码器，实现常识推理； Feng et al. (2020)将多跳关系推理模块引入预训练模型，通过子图中的多跳推理高效聚合信息，实现可解释且性能优越的常识推理； Yasunaga et al. (2021)为了同步更新和统一表示QA上下文和图结构信息，提出一种端到端的联合训练方法QA-GNN，其推理过程具备较好的可解释性。

常见的非结构化知识库有OpenMindCommonSense（OMCS） (Singh et al., 2002)、Simple Wikipedia (Napoles and Dredze, 2010)、Wikipedia等，该类知识库提供了高覆盖率和强相关性的背景信息，可作为常识知识的补充，帮助模型进行推理。 Li et al. (2021)利用基于语义的检索技术，将OMCS中的相关文本知识筛选和过滤后引入模型中，作为模型推理的补充信息； Chen et al. (2020)通过迭代搜索相关概念和实体来丰富背景信息，并设计了答案选择感知注意力来融合嵌入表示； Xu et al. (2020)将知识图谱信息和选项释义相联系，以更加充分的文本信息作为输入； Lv et al. (2020)通过在ConceptNet和Wikipedia中分别抽取图结构和文本信息，利用GCN编码知识子图，XLNET编码文本，然后通过GAT进行多模态推理。

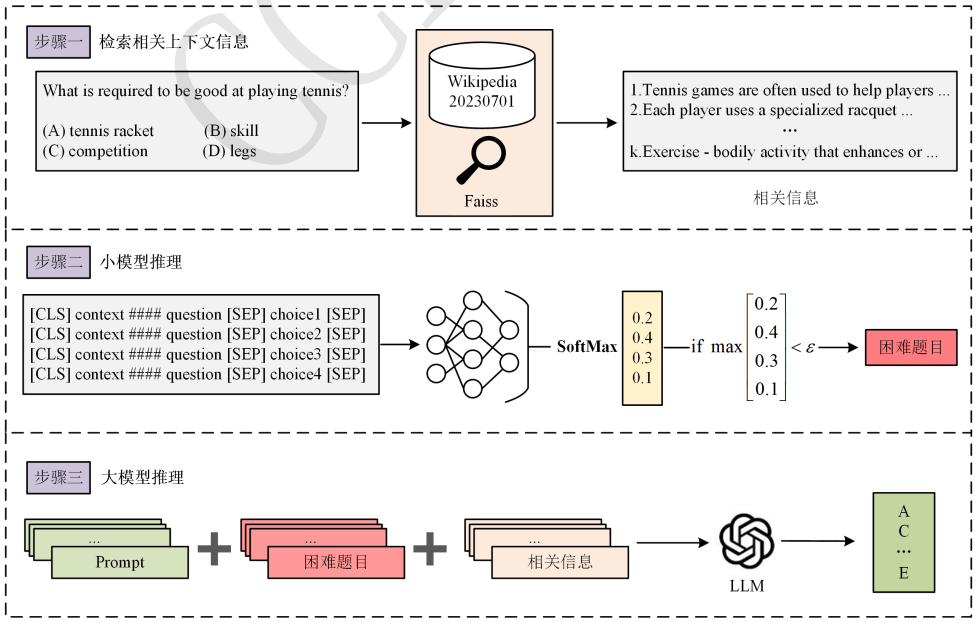

Figure 1: TSCR方法整体流程图

(c) 2025 中国中文信息学会计算语言学专业委员会

总结来说，常识推理依赖于模型自身的推理能力和外部知识的补充。然而，结构化知识覆盖有限且难以转化为自然语言，非结构化知识则易引入噪声。不同知识库间存在存储与表示差异，亟需统一的融合机制。因此，提升模型推理能力并高效整合多源异构知识，仍是当前的重要研究挑战。

3 TSCR方法

TSCR方法包括三个步骤，如图 1所示，包括：相关上下文检索、小模型推理和大模型推理。其中以参数规模是否超过1B作为划分标准，将参数量低于1B的模型定义为小模型；而参数量超过1B的模型定义为大模型。在相关上下文检索部分，TSCR基于Wikipedia构建了一个包含6.28M篇文章的知识库，采用文章级与句子级双重索引结构。具体而言，系统先通过Faiss完成文章级语义检索，利用首句摘要加速匹配；再对相关文章分句，构建句级索引，提取与问题最相关的上下文，用于回答或作为Prompt输入大语言模型。在推理部分，TSCR采用两阶段策略提升效率，并设置置信度阈值ϵ用于动态决策模型调用。第一阶段由小模型进行推理，若其最大置信度超过ϵ，则直接输出结果；否则将其标记为“困难问题”，由大模型接续处理。该策略在保证准确率的同时，有效降低大模型使用频率，节省推理成本。

3.1 相关上下文检索及应用

为了辅助模型推理，本文选择高质量文本数据作为候选知识库。在模型推理之前，先进行检索，得到与待推理问题相关的多条信息，结合这些上下文信息一同推理，相当于将闭卷考试转化为开卷考试，一定程度上缓解了问题的难度。我们在知识库的构建与检索时严格遵循标准技术实践，在数据来源方面，使用开源的Wikipedia数据转储作为基础语料；在预处理标准化中，我们采用WikiExtractor对原始数据进行清洗，去除非正文内容（如超链接、模板、脚注等），保留结构化段落，确保语料质量；在构建知识库向量化表示时，我们使用Sentence Transformers进行嵌入计算，进一步利用Faiss建立双重索引（文章级和句子级），并在索引阶段引入摘要句（文章首句）以提升检索效率；上下文筛选中，我们针对每个问题，设定Top-k文章与Top-k句子的检索数量，结合语义相似度进行排序筛选，确保提供给模型的上下文具备相关性和覆盖度。具体构建及检索过程如图 2所示。

(1)使用Sentence Transformers (Reimers and Gurevych, 2019) 将问题和Wikipedia文章转换为Embedding后，使用Faiss创建索引。同时，为了提高检索效率，使用每篇文章的第一句话作为整篇文章的摘要。

(2) 使用Faiss执行语义相似度检索，查找最有可能包含相关信息的前k篇文章，并获得这k篇文章的全文，使用Bling Fire工具 (Microsoft, 2019)将文章分割成句子。

(3) 将分割后的句子转化为Embedding，创建单独的索引并执行语义相似性检索，以获得每个问题的top-k相关句子。

经上述流程得到相关上下文之后，可以将问题、候选和上下文信息结合起来，直接执行问题回答，也可以将其整合进Prompt中馈送到LLM，利用这些上下文信息，生成更具逻辑性和连贯性的回答，从而为用户提供更加精准和详尽的回复。

3.2 小模型推理

在小模型推理阶段，本文选择了DeBERTaV3 (He et al., 2021)的Large版本作为推理模型。其优越的语义理解能力和鲁棒性为复杂自然语言任务提供了有力支持。同时，因其在鲁棒性方面的卓越表现，已成为各类文本竞赛的可靠和首选工具。

本节推理输入的定义为“[CLS] context #### question [SEP] choice* [SEP]”，其中context为从知识库中检索到的相关上下文，question为题干信息，choice*表示对应的选项。这样的输入结构有助于模型更好地理解问题并进行推理。模型进行推理后，通过softmax归一化得到每个选项的概率分布。设定一个阈值ϵ，如果推理结果中的最大概率值大于该阈值，系统则选择最大概率对应的选项作为模型推理的结果；如果最大概率小于阈值，系统则将问题标记为困难题目，等待在下一步推理中更强大模型的介入。这一策略的制定考虑到了在处理大规模问题时，系统需要在保证准确性的同时高效地使用计算资源。如图 1中步骤二所示。

3.3 大模型推理

为了节省推理时间和资源，TSCR中的推理过程分为两个阶段，第一阶段使用小模型推

(c) 2025 中国中文信息学会计算语言学专业委员会

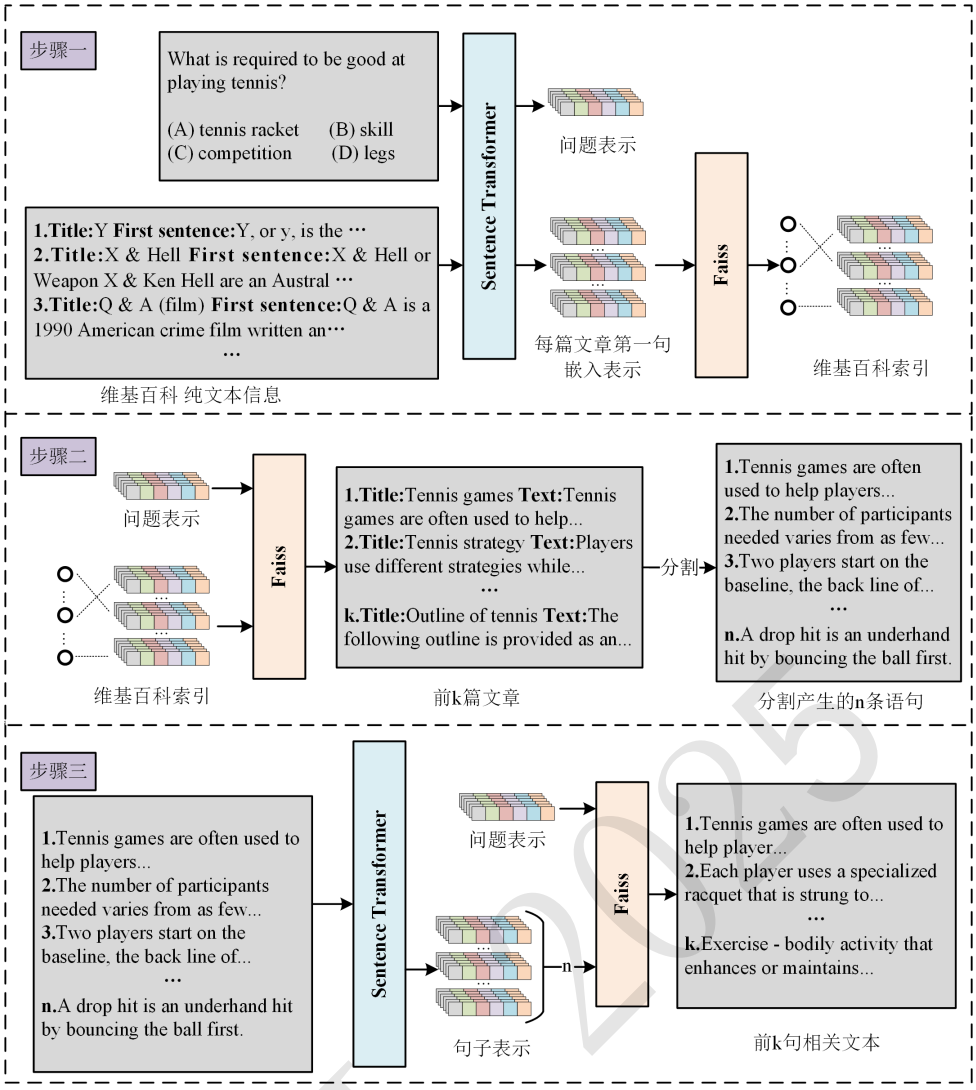

Figure 2: 相关上下文检索流程图

理，解决一些简单问题的同时，标记出困难问题。第二阶段使用大模型推理，解决上一阶段标记的困难问题。

本文选择Qwen (Bai et al., 2023)作为大模型推理，该模型是开展本文工作时综合性能最好的开源大语言模型 (Contributors, 2023)，本阶段要求模型仅输出正确答案的索引，即A、B、C、D。因此，本文将Prompt设计为："Output an index of the correct answers based on the questions and options, and do not output anything other than the index."，同时，为了让LLM进一步理解Prompt的含义，使用3-shot少样本提示。具体Prompt示例如表 1所示。

4 实验结果与分析

4.1 实验数据

本文在四个常识推理数据集上做了实验，这些数据集涵盖了常识与科学领域的问题，旨在考察模型运用背景知识进行逻辑推理的能力。数据集具体信息如表 2所示。

CSQA (Talmor et al., 2018)是一个用于常识推理的大规模问答数据集，涵盖广泛的主题。这些问题通常涉及一种常见的情景或事件，要求根据常识知识进行推理来得出正确答案。

OBQA (Mihaylov et al., 2018)是一个用于开放式基础科学问答的数据集，这些问题通常基于书籍中的内容，要求根据阅读理解和常识知识进行推理来选择正确的答案。

ARC-Easy和ARC-Challenge是ARC (AllenAI, 2018)数据集中两个不相交的子集，都包含小学自然科学题目，每题有四个选项。ARC-Challenge 集合的问题在理解和推理层面具有更高

(c) 2025 中国中文信息学会计算语言学专业委员会

Prompt
Output an index of the correct answers based on the questions and options, and do not output anything other than the index. Here are three examples: **Context**: context **Question**: What are you waiting alongside with when you're in a reception area? **Options**: A:motel B:chair C:hospital D:people E:hotels output: D **Context**: context **Question**: Where could a person avoid the rain? **Options**: A:bus stop B:tunnel C:synagogue D:fairy tale E:street corne output: C **Context**: context **Question**: What has someone who had finished their undergraduate done? **Options**: A:graduated B:masters C:postgraduate D:phd E:professor output: A Here are the questions you need to answer, and you can use the contextual information at your discretion: **Context**: context **Question**: question **Options**: A: choice1 B:choice2 C:choice3 D:choice4 E:choice5 output:

Table 1: Prompt示例

难度。

QASC (Khot et al., 2020)由小学至初中层级的基础科学问题构成。每道题目配有8个备选答案，同时还提供了一个包含1900万条科学事实的知识库。

	训练集 大小（条）	验证集 大小（条）	测试集 大小（条）	选项 数量（个）	问题平均 长度（词）	选项平均 长度（词）
CSQA	8500	1221	1241	5	13.17	1.89
OBQA	4957	500	50	4	10.44	3.06
ARC-E	2241	567	2365	4	19.44	3.71
ARC-C	1117	295	1165	4	22.45	4.93
QASC	7320	814	926	8	8.07	3.28

Table 2: 数据集详细信息表

4.2 实验设置及评价指标

模型使用PyTorch框架实现，使用AdamW优化器更新模型权重，并设置早停机制提高训练效率。为保证实验公平性和结果一致性，所有实验均在一张Tesla V100-SXM2-32GB上完成，并设定相同的随机种子。模型具体实验参数如表 3所示。

本文选择正确率（accuracy, Acc）作为评价指标，正确率是指数据集中正确选择答案的样本占所有样本的比例，可以表示为$\text{Acc} = \frac{\text{正确选择答案的样本个数}}{\text{总样本个数}}$。

第二十四届中国计算语言学大会论文集，第156页-第169页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

实验参数	Epoch	Batch size	Early stop	Dropout	学习率	衰减率	预热率	梯度累积	上下文检索长度	随机种子	训练精度	学习率衰减策略
大小	50	8	5	0.1	3e-05	0.01	0.1	8	256	2023	fp16	cosine

Table 3: 模型实验参数

4.3 对比试验中的性能

为了证明本文工作的有效性，本文选择了2017年以来17个可用于常识推理的模型，在CSQA和OBQA两个数据集上进行了对比，结果如表 4所示，其中"/"表示原论文未在该数据集上进行实验，对比实验中涉及的模型如下:

	CSQA		**OBQA**	
	dev	test	dev	test
RN (Santoro et al., 2017)	74.57	69.08	67.00	65.20
RGCN (Schlichtkrull et al., 2018)	72.69	68.41	64.65	62.45
RoBERTa (Liu et al., 2019)	68.92	71.88	67.80	64.47
KagNet (Lin et al., 2019)	73.47	69.01	/	/
GconAttn (Wang et al., 2019)	72.61	68.59	66.85	64.75
ALBERT (Lan et al., 2019)	60.62	59.32	54.50	49.27
UnifiedQA (Khashabi et al., 2020)	55.28	61.34	70.40	68.40
MHGRN (Feng et al., 2020)	74.45	71.11	68.10	66.85
QA-GNN (Yasunaga et al., 2021)	76.54	73.41	68.27	67.80
GSC (Wang et al., 2021)	79.11	74.48	/	70.33
JointLK (Sun et al., 2021)	77.88	74.43	/	70.34
GenMC (Huang et al., 2022)	73.22	72.28	70.60	67.20
CoSe-Co (Bansal et al., 2022)	78.15	72.87	/	/
CORN (Guan et al., 2022)	79.58	74.43	72.35	71.30
GreaseLM (Zhang et al., 2022)	78.50	74.20	/	66.90
DRAGON (Yasunaga et al., 2022)	/	76.00	/	72.00
EMKG (房晓and 王红斌, 2025)	78.62	74.45	/	70.69
TSCR（本文方法）	82.64	77.36	82.40	83.40

Table 4: 同类方法在CSQA和OBQA数据集上的正确率对比

RN (Santoro et al., 2017)：提出了一种即插即用的小模块，用于解决一些基于关系推理的问题。

RGCN (Schlichtkrull et al., 2018)：使用GCN建模关系型数据，对知识库进行链路预测和实体分类，可用于知识问答、信息检索等下游任务。

RoBERTa (Liu et al., 2019)：通过移除Next Sentence Prediction任务、使用更大的批量和数据量，以及更长时间的训练，显著提升了模型在多项NLP任务上的表现。

KagNet (Lin et al., 2019)：从ConceptNet中启发式检索出路径，利用LSTM对其编码，最终结合层次路径注意力进行推理。

GconAttn (Wang et al., 2019)：使用ConceptNet作为外部知识源，选择不同的子图生成机制，系统地利用其中包含的概念和关系，结合文本模型解决自然语言推理问题。

ALBERT (Lan et al., 2019)：通过参数压缩和改进自监督目标，在显著减少参数量的同时提升了多项NLP任务性能。

UnifiedQA (Khashabi et al., 2020)：通过将多种问答任务统一为文本到文本格式，并使用T5模型进行多任务训练，实现在不同格式的问答数据集上强大的泛化能力和一致性能。

(c) 2025 中国中文信息学会计算语言学专业委员会

MHGRN (Feng et al., 2020)：从知识子图中执行多条、多关系推理，且拥有可解释的推理路径。

QA-GNN (Yasunaga et al., 2021)：设计了一种融合异构信息的端到端问答模型，通过分析推理链路实现可解释的推理。

GSC (Wang et al., 2021)：使用Sparse VD (Molchanov et al., 2017)对现有GNN进行剪枝，得到简单、高效的GNN模块后，使用图推理解决常识问题。

JointLK (Sun et al., 2021)：使用双向注意力融合文本和图结构信息，同时引入动态剪枝方法，去除不相关节点。

GenMC (Huang et al., 2022)：提出了一种生成增强的多项选择问答方法，通过先生成提示信息来提供上下文线索，再根据这些线索生成答案，从而提升模型在多项选择问答中的表现。

CoSe-CO (Bansal et al., 2022)：将路径和文本对齐以创建训练数据，训练后可生成与给定问题相关的常识信息，用于推断下游任务中的知识。

CORN (Guan et al., 2022)：使用预训练语言模型编码文本信息，使用图神经网络编码图结构信息后联合推理，经由MLP得到最终可能性得分。

GreaseLM (Zhang et al., 2022)：在检索到的子图中，添加一个交互节点，用以联合更新文本和图结构信息，交互节点多次迭代后，可将文本和图编码映射到同一个空间。

DRAGON (Yasunaga et al., 2022)：在GreaseLM的基础上改进，设计为预测子图中某条边的预训练任务。

EMKG (房晓and 王红斌, 2025):通过融合问题上下文与推理后的知识图谱表示，引入专家网络与动态门控打分机制，提升答案预测的准确性。

由表 4分析可知，本文提出的方法在CSQA和OBQA两个数据集上都达到了最佳水平。在测试集上，比原有SOTA分别高出1.36%和11.4%。分析其增益差距的原因是CSQA覆盖面较广，检索到的信息相对较多，质量也参差不齐，同时模型难以分辨检索到的信息对解答当前问题是否有帮助；而OBQA多为基础科学问题，范围相对集中，检索的结果也更加贴切。

4.4 阈值分析

4.4.1 使用阈值判断题目难度的合理性分析

	ARC-Easy	ARC-Challenge
题目总数	2365	1165
RoBERTa-Base	1648(69.68%)	818(70.21%)
RoBERTa-Large	571(24.14%)	651(55.87%)

Table 5: 相同难度阈值下不同数据集上的难题数量对比

为分析使用阈值判断题目难度的合理性，本文以超参数的形式设定相同难度阈值，选择RoBERTa模型在ARC-Easy和ARC-Challenge两个数据集上进行了实验。涉及到的两个数据集均来自于初中常识题，题目来源相同，难度不同，符合实验条件。实验结果如表 5所示。

实验结果显示，以RoBERTa-Large为例，在同样的难度阈值下，ARC-Challenge中的难题占比为55.87%，而ARC-Easy中的难题占比仅为24.14%；同时，在ARC-Challenge数据集中，RoBERTa-Base比RoBERTa-Large难题比例高出14.34%。说明使用阈值来判断题目难度的方法是合理的，可以为模型推理做出正确的决策，同时再次证明模型越大，推理能力越强。

4.4.2 阈值ϵ对实验结果的影响

本文在实验阶段进行了超参数敏感性分析：我们选取多个ϵ值进行网格搜索，并记录不同阈值下，模型的准确率和推理时间的变化趋势，最终选择了在准确率与推理效率之间取得平衡的ϵ值（CSQA选取0.35，ARC-C选取0.55）。具体结果如图 3和图 4所示：

4.5 消融实验

为了探索上下文检索和两阶段推理对TSCR方法的增益，本文使用了RoBERTa-Base等四个骨干网络在多个数据集上进行了消融实验。实验结果如表 6所示，表中各个模型的变种含义如下：

(c) 2025 中国中文信息学会计算语言学专业委员会

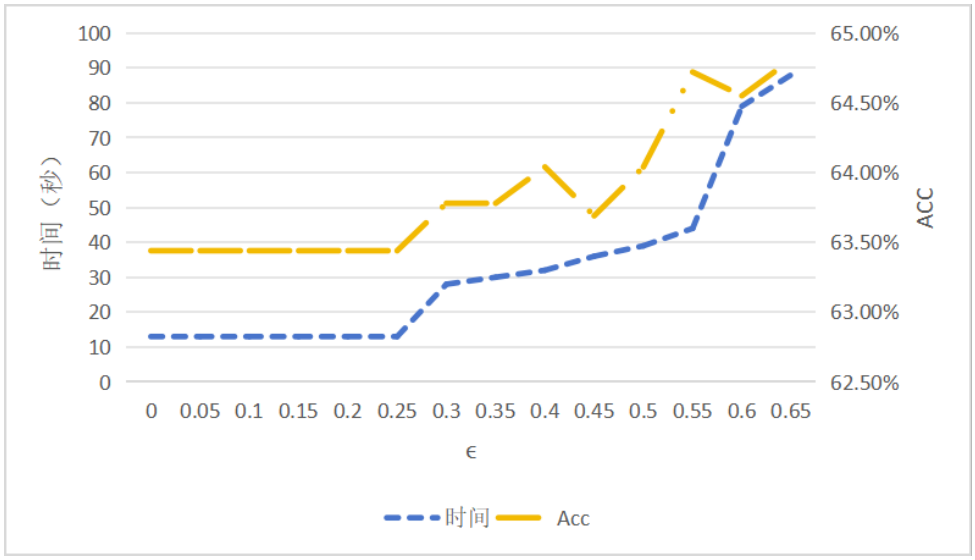

Figure 3: ARC-C中ϵ对实验结果的影响

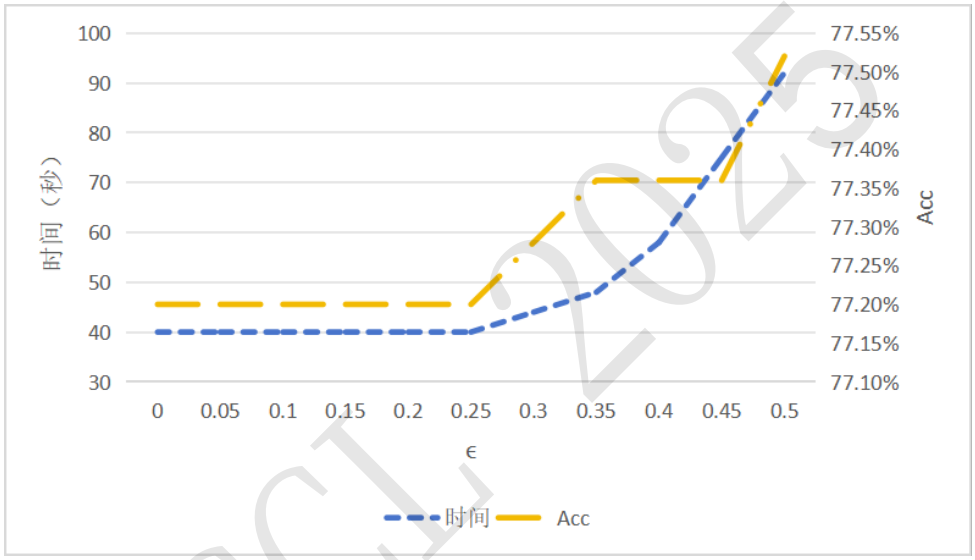

Figure 4: CSQA中ϵ对实验结果的影响

(1)+Context：在骨干网络上加入上下文检索模块；
(2)+LLM：在骨干网络上加入大语言模型，进行两阶段推理；
(3)+Context+LLM：将上下文检索和大语言模型同时加入模型。

上下文检索模块的有效性：在移除检索模块后，模型在五个数据集上的整体表现均出现明显下降，以DeBERTa-V3-Large网络和ARC-Easy数据集为例，在骨干网络中加入上下文信息后，正确率提高了3.09%。说明模型在缺乏外部知识支撑的情况下难以有效理解问题并做出准确推理。相比之下，引入上下文检索模块能够为模型提供更丰富的知识背景，有效提升推理准确率，验证了该模块在提升问答性能方面的重要性。

两阶段推理模块的有效性：在引入LLM模块并使用两阶段推理方法后模型在多个数据集上取得了更优的性能表现，同样以DeBERTa-V3-Large网络和ARC-Easy数据集为例，在骨干网络中加入LLM推理困难题目后，正确率提高了3.17%。相比单一大模型的高资源消耗，两阶段推理方案合理分配计算资源，在保持高准确率的同时提升了推理效率，充分验证了该模块的实用价值。

总的来说，删除两者中任一模块均会使模型的准确性降低，将两部分同时加入之后，模型性能会得到显著提升。说明上下文信息和使用LLM推理对于解决常识推理问题有益，而且增幅

(c) 2025 中国中文信息学会计算语言学专业委员会

	CSQA	OBQA	ARC-E	ARC-C	QASC
RoBERTa-Base	57.78	54.60	46.81	31.16	30.67
+Context	57.94	55.80	59.58	33.91	45.36
+LLM	58.18	55.80	48.71	32.36	35.85
+Context+LLM	**58.18**	**57.20**	**61.18**	**36.39**	**46.76**
RoBERTa-Large	58.10	52.60	49.68	38.88	34.45
+Context	69.70	68.40	67.78	40.34	60.80
+LLM	59.31	53.00	69.09	39.57	49.89
+Context+LLM	**69.78**	**70.00**	**69.09**	**41.46**	**62.31**
DeBERTa-V3-Base	69.70	67.80	67.36	43.00	58.64
+Context	71.23	72.60	71.75	43.00	63.28
+LLM	70.51	71.80	71.71	43.35	59.07
+Context+LLM	**71.88**	**73.20**	**74.80**	**44.21**	**64.04**
DeBERTa-V3-Large	77.28	80.40	82.03	63.69	67.60
+Context	77.36	83.00	85.12	63.61	71.49
+LLM	**77.68**	81.00	85.20	64.12	70.84
+Context+LLM	77.36	**83.40**	**86.39**	**64.72**	**72.57**

Table 6: 消融实验结果

较大。解决常识推理问题有两个重点：需要一定的知识作为基础；结合基础知识进行推理。而本文方法中的上下文信息和LLM推理正好对应这两个关键点，所以TSCR设计有效。

4.6 推理效率分析

推理过程不仅要求正确率高，而且也要尽可能节约时间。为进一步分析模型推理效率，本文选择了具有代表性的Baichuan2 (Yang et al., 2023)、DeepSeek (Bi et al., 2024)、Mistral (Jiang et al., 2024)和Qwen四种开源大模型（统一选择其7B-Chat版本），使用相同Prompt进行单独推理，分析其性能和效率。实验结果如表 7所示。

	CSQA	OBQA	ARC-E	ARC-C	QASC
Baichuan2 (Yang et al., 2023)	179/22.56	84/43.20	447/63.13	222/48.58	150/27.21
DeepSeek (Bi et al., 2024)	198/29.01	74/29.40	340/41.95	181/37.17	172/28.73
Mistral (Jiang et al., 2024)	657/20.23	256/27.60	1132/24.99	614/22.58	460/13.07
Qwen (Bai et al., 2023)	263/63.34	111/66.40	537/81.65	302/61.80	234/62.20
TSCR(本文方法)	**48/77.36**	**22/83.40**	**277/86.38**	**44/64.72**	**118/72.57**

Table 7: 仅LLM推理的时间/准确率对比（时间单位为秒，且不包含加载模型耗时）

在相同Prompt下，不同LLM推理性能相差较大：如在ARC-Challenge数据集上，Baichuan2推理时间为222秒，准确率为48.58%，而Qwen推理时间为302秒，准确率为61.80%，Qwen耗费了更长的时间，取得了更高的准确率。由于本文方法采取两阶段推理，简单问题不需要LLM来解决，因此推理时间更短。从表中结果来看，LLM并不是全能模型，相反，其在某些领域的性能与专有模型还有一定差距，需要进一步微调LLM来弥补这部分差距。

4.7 个例分析

为了进一步探索检索的背景上下文对模型推理的帮助，本文从CSQA测试集中筛选了样例进行实验分析，具体结果如表 8所示

样例1中的问题相对简单，所以模型无论是否使用上下文，都可以选择出正确选项(D) army。说明上下文信息并没有对模型推理产生积极作用，但也没有负面影响。

样例2中的问题需要关注细节“facial sign”，特指面部的表现。模型不使用上下文时，选择

(c) 2025 中国中文信息学会计算语言学专业委员会

	问题	检索得到的信息	结果
1	What requires a very strong leader? (A) grocery store (B) country (C) organization (D) army (E) pack	**1.** Leader authorization and employee attitudes. **2.** A strongman is a type of an authoritarian political leader. **3.** The autocratic leadership style works well if the leader is competent and knowledgeable enough to decide about each and every item under their control.	DeBERTa-V3-Large: (D) army ✓ TSCR: (D) army ✓
2	What is a facial sign of having fun? (A) laughter (B) pleasure (C) smiling (D) being happy (E) showing teeth	**1.**Smiling, on the other hand, iseasily recognized as an expression of happiness, but even here there is a distinction between cheek-raising or Duchenne smiles and non-emotional smiles,which are thought to be used mainly as social signals. **2.**Gelotophilia describes the joy of being laughed at. **3.**The Psychology of Facial Expression.	DeBERTa-V3-Large: (A) laughter × TSCR: (C) smiling ✓
3	What might learning about science lead to? (A) new ideas (B) experiments (C) invent (D) accidents (E) atheism	**1.**Finally, another approach often cited in debates of scientific skepticism against controversial movements like ""creation science"" is methodological naturalism. **2.**If the hypothesis survived testing, it may become adopted into the framework of a scientific theory, a logically reasoned,self-consistent model or framework for describing the behavior of certain natural events. **3.**Thus, it is argued that science is better learned through experiential activities.	DeBERTa-V3-Large: (A) new ideas ✓ TSCR: (E) atheism ×

Table 8: CSQA中选取的实例

了与题干中"have fun"最接近的(A) laughter，但是需要注意的是，A不是面部表现；当模型使用上下文时，可以看到第一条上下文中第一个词就是正确答案"Smiling"，模型可以关注到这个细节。该样例说明上下文信息可以对模型产生积极影响，进而选出正确答案。

样例3中的问题相对抽象，所以模型在有上下文干预的情况下，反而给出了错误答案，而没有上下文信息时却能正确回答。说明上下文信息对于模型推理不总是积极正面的，需要加以限制，保证不会影响模型原本正确的输出，这也是未来需要解决的问题。

5 结论

为解决现有常识推理模型推理过程中引入的外部信息不够精准和融合不充分的问题，本文提出了一种基于检索增强生成的两阶段常识推理方法TSCR。在创建了双重索引来加速检索过程的基础上，设计了两阶段推理方法。多项实验验证了TSCR在推理任务中的有效性和适用性。消融实验表明，上下文信息和大语言模型的引入对推理效果有显著提升。推理效率方面，两阶段设计有效减少了不必要的计算资源消耗。总体来看，TSCR在性能、效率与实际应用之间取得了良好平衡，为常识推理任务提供了一种高效、灵活的解决思路。

在未来的工作中，我们将考虑使用更先进的自动化工具和技术，以减轻知识库维护的负担，同时探索更高效的数据存储和检索方法，在提高知识库性能和可靠性的同时，降低TSCR对知识库的依赖性。

第二十四届中国计算语言学大会论文集，第156页-第169页，济南，中国，2025年8月11日至14日。

(c) 2025 中国中文信息学会计算语言学专业委员会

参考文献

AllenAI. 2018. AI2 Reasoning Challenge (ARC) 2018 Dataset. [EB/OL].

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

Rachit Bansal, Milan Aggarwal, Sumit Bhatia, Jivat Neet Kaur, and Balaji Krishnamurthy. 2022. Coseco: Text conditioned generative commonsense contextualizer. *arXiv preprint arXiv:2206.05706*.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12317–12325.

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

Qianglong Chen, Feng Ji, Haiqing Chen, and Yin Zhang. 2020. Improving commonsense question answering by graph-based iterative retrieval over multiple knowledge sources. *arXiv preprint arXiv:2011.02705*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*.

Xin Guan, Biwei Cao, Qingqing Gao, Zheng Yin, Bo Liu, and Jiuxin Cao. 2022. Corn: co-reasoning network for commonsense question answering. In *Proceedings of the 29th international conference on computational linguistics*, pages 1677–1686.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Zixian Huang, Ao Wu, Jiaying Zhou, Yu Gu, Yue Zhao, and Gong Cheng. 2022. Clues before answers: Generation-enhanced multiple-choice qa. *arXiv preprint arXiv:2205.00274*.

AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. 2024. Mistral 7b. arxiv 2023. *arXiv preprint arXiv:2310.06825*.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yeqiu Li, Bowei Zou, Zhifeng Li, Aiti Aw, Yu Hong, and Qiaoming Zhu. 2021. Winnowing knowledge for multi-choice question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1157–1165.

(c) 2025 中国中文信息学会计算语言学专业委员会

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Frederick Liu, Siamak Shakeri, Hongkun Yu, and Jing Li. 2021. Enct5: Fine-tuning t5 encoder for non-autoregressive tasks. *arXiv preprint arXiv:2110.08426*, 2.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8449–8456.

Microsoft. 2019. Finite state machine and regular expression manipulation library. https://github.com/microsoft/BlingFire, 03. Accessed: 2024-02-28.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International conference on machine learning*, pages 2498–2507. PMLR.

Courtney Napoles and Mark Dredze. 2010. Learning simple wikipedia: A cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, pages 42–50.

Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, pages 593–607. Springer.

Push Singh, Thomas Lin, Erik T Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE: Confederated International Conferences CoopIS, DOA, and ODBASE 2002 Proceedings*, pages 1223–1237. Springer.

Yueqing Sun, Qi Shi, Le Qi, and Yu Zhang. 2021. Jointlk: Joint reasoning with language models and knowledge graphs for commonsense question answering. *arXiv preprint arXiv:2112.02732*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

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

(c) 2025 中国中文信息学会计算语言学专业委员会

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. 2019. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7208–7215.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. *arXiv preprint arXiv:2005.00691*.

Kuan Wang, Yuyu Zhang, Diyi Yang, Le Song, and Tao Qin. 2021. Gnn is a counter? revisiting gnn for question answering. *arXiv preprint arXiv:2110.03192*.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2020. Fusing context into knowledge graph for commonsense question answering. *arXiv preprint arXiv:2012.04808*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. 2022. Deep bidirectional language-knowledge graph pretraining. *Advances in Neural Information Processing Systems*, 35:37309–37323.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. Sentibert: A transferable transformer-based architecture for compositional sentiment semantics. *arXiv preprint arXiv:2005.04114*.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models for question answering. *arXiv preprint arXiv:2201.08860*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

房晓and 王红斌. 2025. 基于知识图谱与门控机制的专家再学习推理问答方法. 应用科学学报, 43(2):288–300.

袁毓林and 卢达威. 2018. 怎样利用语言知识资源进行语义理解和常识推理. 中文信息学报, 32(12):11–23.

CCL 2025

(c) 2025 中国中文信息学会计算语言学专业委员会