基于思维链和知识迁移的多语言问答推理研究

罗健[1,2]　孙媛[1,2,3,*]
[1]中央民族大学 信息工程学院
[2]国家语言资源监测与研究民族语言中心
[3]中央民族大学国家安全研究院
[*]通讯作者：孙媛
22302120@muc.edu.cn，tracy.yuan.sun@gmail.com

摘要

近年来，大型语言模型如ChatGPT显著提高了机器对自然语言的理解能力，其中，问答推理任务在推动语言理解能力和人机交互智能化方面具有重要意义，但目前仍面临诸多挑战。本文针对现有大模型资源消耗大、小模型推理能力弱，低资源语言推理能力受限等问题，提出了融合思维链和微调技术的方法，通过Human-Thinking提示策略优化大模型推理能力，并借助大模型指令微调提升小模型推理性能，引入多角色协作机制进一步优化推理步骤质量。通过探索跨语言思维链提示方法，利用高资源语言知识弥补低资源语言不足，采用双通道机制和投票打分机制整合不同语言推理知识，提升模型在低资源语言的推理表现。实验结果表明，本文方法能有效提升小型模型在多语言问答推理的能力，具有一定的研究价值。

关键词： 问答推理；大模型；多语言；思维链

Research on Multi-Language Q&A Reasoning Based on Chain-of-Thought and Knowledge Transfer

Jian Luo[1,2]　Yuan Sun[1,2,3,*]
[1]School of Information Engineering, Minzu University of China
[2]National Language Resources Monitoring and Research Center for Minority Languages
[3]Institute of National Security, Minzu University of China
[*]Corresponding author：Yuan Sun
22302120@muc.edu.cn，tracy.yuan.sun@gmail.com

Abstract

In recent years, large language models such as ChatGPT have significantly enhanced machines' capabilities in understanding natural language. Question-answering reasoning tasks play a crucial role in promoting language comprehension and the intelligence of human–machine interactions. However, they still face numerous challenges. This paper proposes a method integrating chain of thought and fine-tuning techniques to address the existing issues of high resource consumption in large models, weak reasoning capabilities in small models, and limited reasoning abilities in low-resource languages. The method leverages the reasoning capabilities of large models to guide small models, optimizes the reasoning capabilities of large models through a Human-Thinking prompting strategy, and enhances the reasoning performance of small models via instruction fine-tuning of large models. Additionally, a multi-agent collaboration mechanism is introduced to further optimize the quality of reasoning steps. By exploring cross-lingual chain of thought prompting methods, the knowledge from high-resource languages is utilized to compensate for the deficiencies in low-resource languages. A dual-channel mechanism and a voting mechanism are adopted to integrate reasoning

第二十四届中国计算语言学大会论文集，第170页-第183页，济南，中国，2025年8月11日至14日。

knowledge from different languages, thereby improving the model's reasoning performance in low-resource languages. Experimental results demonstrate that the proposed method effectively enhances the capabilities of small models in multilingual question-answering reasoning tasks and holds certain research value.

Keywords: Q&A Reasoning , LLM , Multilingual , Chain-of-Thought

1 引言

人类的理性思维离不开推理能力，推理是综合已有事实与个人信念推导新结论的过程，也是运用知识进行逻辑分析并得出合理判断的能力。问答推理任务不仅要求模型提供准确答案，还要求展示推理过程，以增强结果的可理解性和可信性。然而，深度学习模型的"黑盒"特性使其推理过程难以解释，尽管结果可能准确，但缺乏透明度仍是其短板。在自然语言处理中，常识推理基于人们对世界及人类行为的基本认知，通过逻辑推理得出结论，是人类认知体系中的重要能力。问答任务是常识推理的典型应用，已有大量研究围绕该方向展开。例如，阅读理解任务要求考生根据给定文本作答，许多问题依赖常识知识以确保精准理解与推理。这种能力是人类智能的关键，对机器而言也具有重要意义，要求机器具备类似能力，运用常识知识展开逻辑推理，得出合理且准确的结论。

本研究聚焦于结合思维链方法和知识迁移技术，提升小型语言模型多语言问答推理中的性能。近年来的研究表明，思维链（Chain-of-Thought, CoT）(Wei et al., 2022b)提示在提升大模型推理能力方面具有显著效果，但大多数CoT提示仍以"Let's think step by step"等(Kojima et al., 2022)机械化语言形式存在，未能有效模拟真实人类在解题过程中的认知路径。这类提示往往缺乏结构化指引，容易导致模型生成冗余或跳跃式推理，尤其在处理常识性、事实性问题时存在一定的逻辑偏差。本文提出Human-Thinking提示策略，通过模拟人类在解答判断题时的结构化思维路径，引导模型识别关键信息、调用相关常识并逐步推理，从而提升推理过程的可控性与合理性。本文通过研究探索如何利用大模型生成丰富且准确的背景知识和逻辑链条，并将这些知识有效融入小型模型中，提高其在问答推理任务中的表现。这有助于减少实际工程对大型模型的依赖，降低计算成本，提升问答推理的准确性和可靠性。

本文的主要贡献如下：

（1）通过数据生成、清洗和人工质量评估，构建了6,000条涵盖英文、中文和藏文的多语言常识问答数据集，问题为判断题类型，通过手动设计若干个不同的主题，覆盖了日常生活中的常见知识领域，解决多语言问答任务缺乏高质量数据集的问题。

（2）提出融合思维链和微调技术的方法，通过Human-Thinking提示策略优化大模型推理能力，并借助大模型指令微调提升小模型在问答推理任务的性能，引入多角色协作机制进一步优化推理步骤质量。

（3）探索跨语言思维链提示方法，利用高资源语言知识弥补低资源语言不足，采用双通道机制和投票打分机制整合不同语言推理知识，提升模型在低资源语言的表现。

2 相关工作

近年来，众多大模型（LLMs）(Achiam et al., 2023; Chowdhery et al., 2022; Brown et al., 2020; Devlin et al., 2019)在自然语言处理领域取得了显著的成功。这些模型凭借其新兴能力，在多种自然语言处理任务中展现了令人印象深刻的少样本和零样本性能(Zhong et al., 2023; Srivastava et al., 2022)。如Wei等人(Wei et al., 2022a)所述，随着模型规模的扩大，LLMs逐渐展现出遵循指令(Sanh et al., 2022)、程序执行(Nye et al., 2021)和模型校准(Kadavath et al., 2022)等能力。然而，LLMs在处理复杂推理任务时仍难以提供稳定和准确的答案(Zhang et al., 2023)，例如在数学推理(Patel et al., 2021)、常识推理(Geva et al., 2021)和符号推理(Wei et al., 2022b)等领域。近期的研究(Yuan et al., 2023; Luo et al., 2023; Liu et al., 2023)表明，通过数学数据对LLMs进行推理增强微调可以在一定程度上提高其推理能力。即便如此，这些模型在复杂推理问题上的表现仍不尽如人意。此外，上述基于训练的方法通常需要大量的数据和高昂的

计算成本，这可能会削弱LLMs的通用能力。因此，一些研究开始尝试采用更经济的提示方法来增强LLMs的推理能力，而无需额外的训练。Wei等人(Wei et al., 2022b)首次提出了少样本思维链提示方法，该方法通过在给出最终答案之前引入中间推理步骤，显著提升LLMs的推理能力。随后，许多工作从不同方面进一步改进思维链提示，包括提示格式(LI et al., 2022)、提示选择(Lu et al., 2022)、提示集成(Weng et al., 2022)、问题分解(Dua et al., 2022)和规划(Yao et al., 2023)。例如，Zhou等人(Zhou et al., 2022)提出了“最少到最多提示”方法，将复杂问题分解为更易管理的子问题，以便逐步解决；Wang等人(Wang et al., 2022)引入一种自一致性解码策略，通过采样多种推理轨迹来投票决定最终答案；Cobbe等人(Cobbe et al., 2021)为数学问题训练验证器，以对候选推理链进行排名，从而选择最可靠的推理链。然而，这些方法仍依赖大量人工标注的推理链。为克服这一限制，Zhang等人(Zhang et al., 2022)提出Auto-CoT，利用零样本思维链和聚类自动生成推理演示。在某些特定任务中，使用经过微调的语言模型可能会比使用思维链提示的大模型获得更好的性能。

思维链通过引导LLMs逐步解决问题，为解决复杂问题和增强模型的推理能力提供了新的视角(Wei et al., 2022b)。然而，当前的研究主要集中在英语上。全球存在超过7000种语言，解决少数语言中的推理问题已成为一个迫切的需求。鉴于这一研究空白，越来越多的研究开始专注于超越传统的单语思维链，探索其跨语言维度。例如，Shi等人(Shi et al., 2022)率先引入了一个专门用于数学推理的多语言数据集，并提出了一种新的跨语言方法，要求LLMs在预测思维链时使用英语，将问题翻译成英语，并通过英语思维链范式解决它。Ranaldi等人(Ranaldi and Zanzotto, 2023)提出了一种跨语言多步推理策略中的自一致性提示机制，显著增强了多种语言中的推理能力。Tanwar等人(Tanwar et al., 2023)建议在提示上下文中加入展示源语言和目标语言之间语义一致性的示例，以促进跨语言推理的无缝过渡。Qin等人(Qin et al., 2023)提出了一种自一致性提示方法，最初涉及手动选择推理语言，然后通过投票机制确定最终的推理答案，这种方法在跨语言提示的有效性方面取得了优异的结果。尽管跨语言专家语言模型在多语言语料库子集上独立训练(Blevins et al., 2024)，但针对需要细微文化理解的特定文化任务，母语提示仍然至关重要(Liu et al., 2024)。最近关于语言处理动态的研究表明，以英语为中心的LLMs使用英语作为内部枢纽语言(Wendler et al., 2024)，而非英语为中心的模型则利用其主要训练语言(Zhong et al., 2024)。对语言无关知识神经元和跨语言神经模式的分析表明，对齐多语言表示可以增强知识转移(Wang et al., 2024; Tang et al., 2024; Cao et al., 2024)。

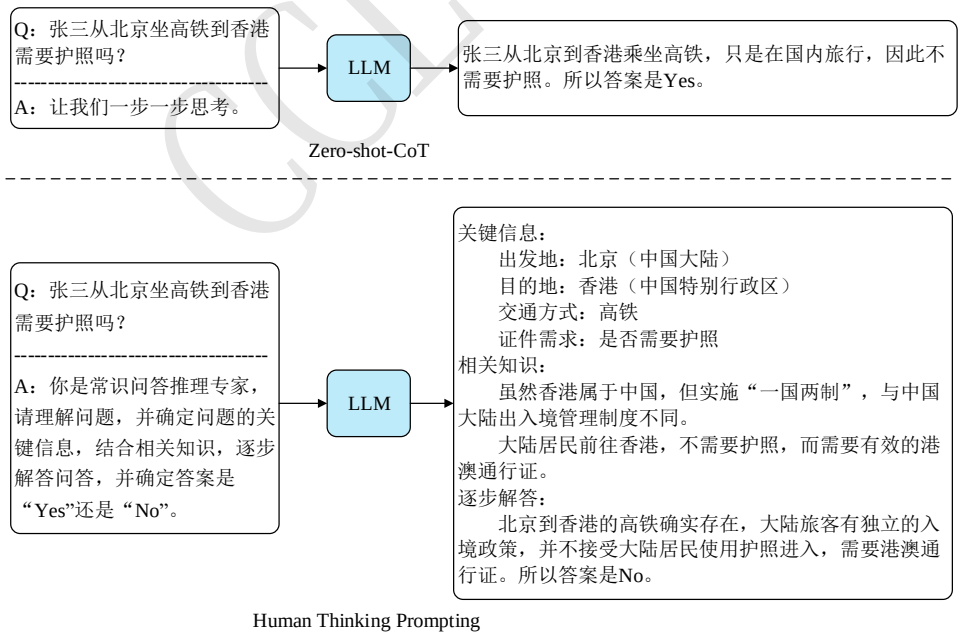

Figure 1: 问答推理数据构建样例

3 研究方法

3.1 Human-Thinking提示策略

尽管现有的CoT提示策略（如“Let’s think step by step”）在某些任务中提升了模型的推理能力，但它们通常仅依赖语言模式对模型进行启发，缺乏对问题结构的建模能力，也未明确引导模型聚焦关键因素或分析路径。这导致模型在推理过程中容易出现信息遗漏、逻辑混乱、冗余或跳步等问题。本文提出Human-thinking（HT）提示策略旨在模拟人类真实解题的思维结构，借助明确的任务角色设定（如“你是常识问答推理专家”）和结构化指令（如“请识别关键信息、结合相关知识、逐步推理”），帮助模型理清推理路径、组织语言表达，从而提高回答的准确性与可解释性。本文在构建的常识推理数据集上验证，如图1所示。HT策略引导模型聚焦关键信息，确保逻辑连贯性，强调推理步骤指导和结果提取精确性。实验表明，HT策略能显著提升大模型（如GPT-3 175B）性能，减少零样本思维链推理中的问题理解和逻辑错误。

3.2 思维链融合大模型的指令微调

大模型指令微调技术可以提高对某一方面任务的性能。本文利用大模型指令微调技术提升小模型推理能力。具体步骤为：首先，用大模型通过零样本思维链提示生成推理过程和答案；其次，根据预测结果正确性筛选样本，格式化为<问题，推理步骤，答案>三元组；最后，用筛选后的高质量样本微调小模型。本研究采用LoRA方法(Hu et al., 2021)对预训练语言模型进行微调，以应对常识问答推理任务。LoRA通过引入低秩矩阵适应特定任务，仅更新少量参数，保留预训练模型通用知识。具体实现是在预训练模型权重矩阵基础上添加低秩适应层，微调时仅更新该层参数，原始权重保持不变。这显著降低了计算和内存开销，同时保持了模型性能。图2展示了融合思维链的LLM指令微调样例图。从数学角度来看，假设输入为x，输出为h，预训练模型的权重为W，降维矩阵为A，升维矩阵为B。在采用LoRA微调时，输出的计算公式如1所示，其中，ΔW是需要更新的参数，通过低秩矩阵A和B来表示较大的矩阵。

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

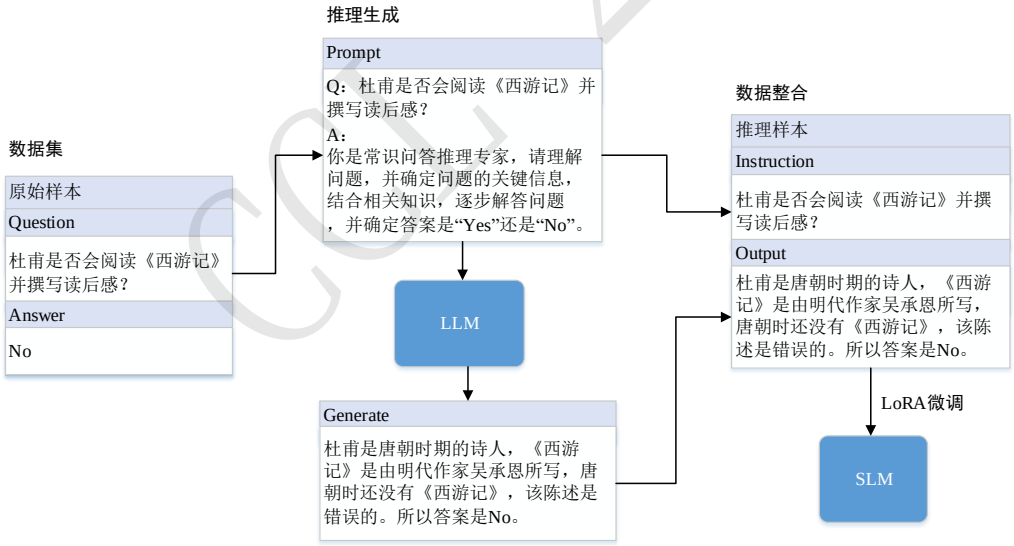

Figure 2: 融合思维链的LLM指令微调样例图

3.3 基于多角色协作的数据质量增强

在上述基于教师模型的样本筛选过程中，我们发现答案预测错误的样本中，大都是因为推理步骤仍可能存在错误或质量不高。为对这些样本进行质量增强，我们设计了多角色协作的机制，如图3所示。具体设计了四个角色：输入员、推理专家、常识专家和仲裁者。输入员按普通人思维对问题进行推理；推理专家检查每步的推理逻辑的正确性；常识专家则结合常识知识对

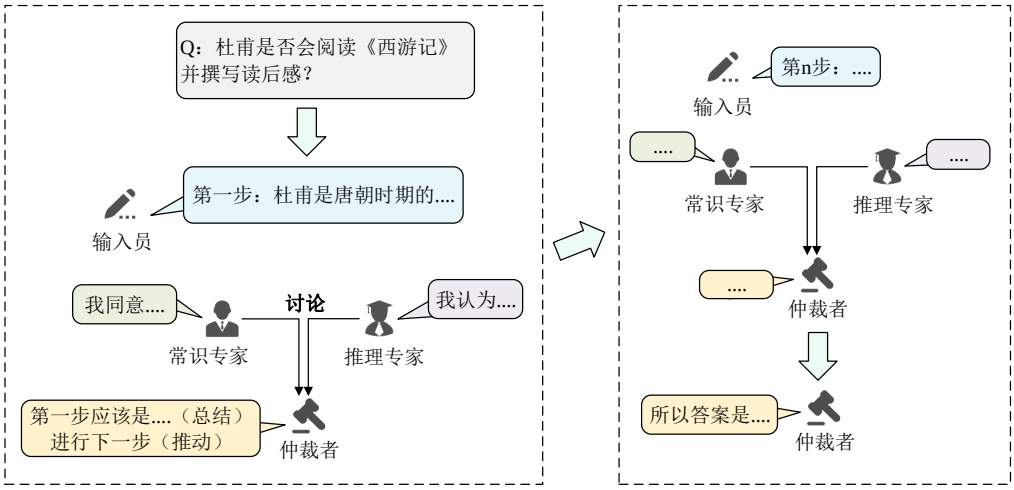

Figure 3: 多角色协作机制流程图

推理步骤进行判断。最后仲裁者综合各方观点，确定合理推理步骤。此机制有效优化了错误样本的推理步骤，为小模型提供高质量训练数据，显著提升其推理性能。表1展示了多角色协作机制的样例。

步骤	问题	杜甫是否会阅读《西游记》并撰写读后感?
1	输入员	杜甫生活在唐朝，而《西游记》是明代小说，因此杜甫不可能阅读这本书。
	常识专家	同意，杜甫生活的时代（唐代）与《西游记》创作的时间（明代）相差约300年，因此杜甫不可能阅读《西游记》。
	推理专家	支持这一点，杜甫是唐朝诗人，《西游记》成书在明代，显然杜甫在世时不可能阅读《西游记》。
	仲裁者	杜甫不可能阅读《西游记》。进行下一步。
2	输入员	杜甫虽然没有阅读过《西游记》，但是他可以撰写读后感，因为那个时候已经有了纸和笔。
	常识专家	不同意。如果杜甫不可能接触到《西游记》，那么就无法撰写《西游记》的读后感。
	推理专家	进一步确认，若杜甫没有阅读《西游记》，撰写读后感也无从谈起。
	仲裁者	既然杜甫不可能阅读《西游记》，那么他也无法写下《西游记》的读后感。讨论结束。

Table 1: 多角色协作机制样例

3.4 基于跨语言思维链的问答推理生成

在多语言环境中，语言模型的推理能力虽然在一定程度上具有语言无关性，但由于训练数据的分布、语言的结构特征以及跨语言对齐的差异，模型在不同语言上的表现往往存在显著差异。为此，我们提出一种跨语言思维链机制，旨在通过跨语言的中间推理步骤，优化语言模型在低资源语言（如藏语）中的推理性能。跨语言思维链的双通道机制通过以下方式实现：对于藏语输入，模型会先将其翻译为英语和汉语，分别生成对应的推理结果，再通过教师模型将这些结果融合，生成最终的藏语输出。这种方法不仅能够弥补低资源语言数据不足的问题，还能利用高资源语言的丰富语义和逻辑结构，提升模型在低资源语言中的推理能力。跨语言思维链推理的流程和样例分别如图4、图5所示。

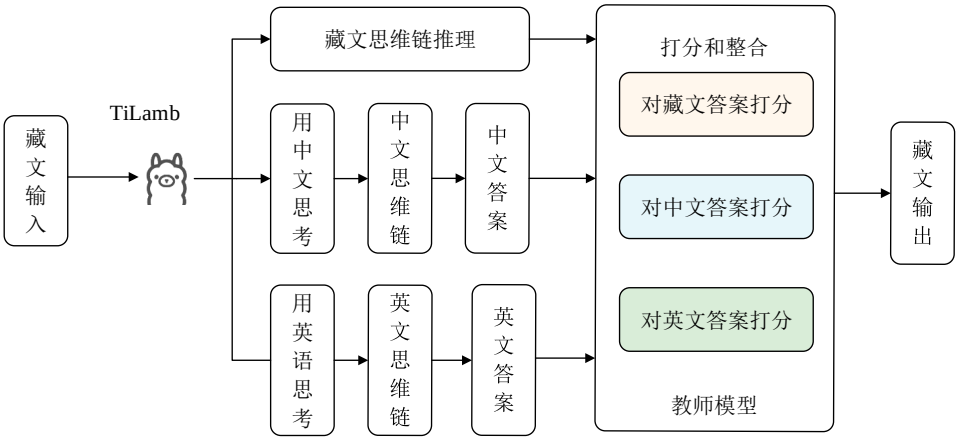

Figure 4: 跨语言思维链推理流程图

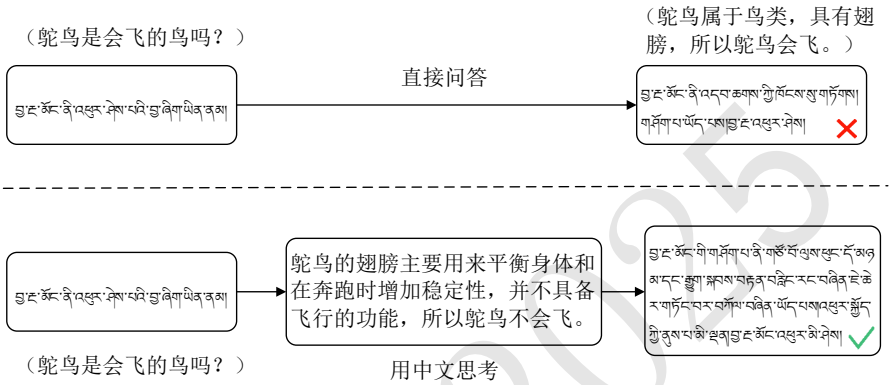

Figure 5: 跨语言思维链推理样例

3.5 基于跨语言思维链的问答推理评估

在问答推理任务中，我们采用大模型作为评估器，对不同语言生成的多样化响应进行质量评估，并选择最优推理方案。具体而言，给定藏文的输入$Q_t i$及其对应的中文与英文的推理路径，大模型通过在两种辅助语言中进行推理，产生对应的藏文回答A。随后，评估模型根据预设标准对生成的回答进行打分，其计算如公式如2所示。

$$S = LLM(p, Q_t i, A) \tag{2}$$

在本工作中，我们设计一个提示p，概述评估回答质量的指标，并将这些指标逐一输入评估模型中。最终得分由评估模型根据所有评估标准生成，打分规则如表2所示，选择得分最高的响应作为最终输出。

事实准确性	答案与事实不符 1	——	答案与事实相符 6
语言流畅性	不流畅 1-2	流畅但存在语法错误 3-4	流畅且句子结构清晰 5-6
推理一致性	推理逻辑混乱 1-2	逻辑连贯但证据不足 3-4	推理连贯且逻辑清晰 5-6

Table 2: 教师模型打分准则

3.6 跨语言一致性

在跨语言思维链中，我们通过跨语言一致性整合不同语言的推理知识。具体方法是：利用中英两种辅助语言进行跨语言思维链推理得到中间步骤，同时用单语言思维链在藏文直接得到推理步骤。通过投票机制保留三种语言在推理结果$\hat{F}$中高度一致的答案，公式如3所示。

$$\hat{F} = \arg\max \sum_{t=1}^{|f|} \sum_{f} \rho(F_t = f) \tag{3}$$

其中$|f|$表示所有目标语言中潜在推理结果的数量，t表示推理步骤的索引，用于遍历目标语言中的所有潜在推理结果，$\rho(X)$表示一个0-1函数，当X为假时返回0，当X为真时返回1。

4 实验

4.1 数据集

本文利用大模型自动构建判断题类型的常识问答对，为确保日常生活所需常识知识的覆盖，我们参考TriviaQA数据集(Joshi et al., 2017)的统计数据，手动设计了若干个不同的主题，包括食物、运动、动物、植物、国家、城市、书籍、历史人物、历史事件、科学家、名胜古迹等。这些主题涵盖了日常生活中的常见知识领域，能够为数据集提供广泛的知识基础。通过数据生成、清洗和人工质量评估，最终保留了涵盖英文、中文和藏文的平行数据各6,000条作为实验的数据集。

4.2 超参数设置

表3展示了实验阶段的超参数设置情况。所有实验均在2张Tesla V100S-PCIE-32G上完成。

超参数名称	超参数值
cutoff_len	2048
learning_rate	2e-04
finetuning_type	lora
num_train_epochs	3.0
per_device_train_batch_size	4
gradient_accumulation_steps	2
lr_scheduler_type	cosine
max_grad_norm	1.0
lora_rank	8
lora_dropout	0.05

Table 3: 模型微调超参数设置

4.3 评测方法

本文将二项选择任务视为特殊分类任务，采用Accuracy和F1值作为模型性能的评估指标。Accuracy（准确率）是模型正确预测样本数占总样本数的比例，计算公式如4所示。其中，TP为真正例数，TN为真负例数，FP为假正例数，FN为假负例数。

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

Precision（精准率）衡量预测正类中真正的正例比例，Recall（召回率）衡量模型捕捉真正例的能力，计算公式分别如5、6所示。

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

F1值综合考虑Precision和Recall，是Precision和Recall的调和平均值，适用于不平衡类别，其计算如公式7所示。

$$\text{F1} = \frac{2 * Precision * Recall}{Precision + *Recall} \tag{7}$$

本文分别计算真实答案为“yes”和“no”的F1值，取平均值“avg”作为最终评价指标，以反映模型在问答任务中的准确性，为推理答案性能提供更好评价依据。

4.4 实验模型

本文选取了五种语言模型，这些模型因训练语料和超参数设置差异被选入实验，以覆盖不同语言和任务需求，确保实验结果的全面性和代表性。LLaMa2-7B（英文）(Touvron et al., 2023)，由Meta开发，基于Transformer架构，通过无监督学习预训练，但中文数据占比仅0.13%，导致中文任务表现欠佳；LLaMa2-7B-Chinese（中文）(Cui et al., 2023)，由LLaMa中文社区开发，通过大规模中文数据预训练和微调，显著提升中文任务表现；TiLamb-7B（藏文）(Wenhao et al., 2024)，以LLaMa2-7B为基座，扩充藏文词表并利用26.43GB藏文数据增量预训练，提升藏文编码和语义理解能力，优于传统方法和其他藏文模型；TiLamb-13B，基于LLaMa2-13B，扩充中文和藏文词表至42,353Tokens，进一步提升编码和语义理解能力；DeepSeek满血版(Shao et al., 2024)，由深度求索研发，对标GPT-4架构，在多项核心能力上表现领先。

4.5 实验基线

为验证融合思维链和微调技术的问答推理方法的有效性，本文设计七个基线进行比较，即：Zero-shot，输入仅包含测试问题，不提供任何额外信息；Zero-shot-CoT，采用零样本思维链的方式，在给模型输入测试问题后拼接提示词“Let’s think step by step”，以激发模型的推理能力；Few-shot-CoT，采用少样本思维链的方式，在给模型输入问题前，拼接六个示例，使得模型学习相应的推理方式；Fine-tune-qa，直接用数据集中的问答对微调模型；Fine-tune-CoT，将通过Zero-shot-CoT方式对大模型进行测试得到的带有推理步骤的三元组形式的<问题，推理步骤，答案>数据对模型进行微调；HT，将采用HT提示策略通过零样本思维链方式对大模型进行测试得到的带有推理步骤的三元组形式的<问题，推理步骤，答案>数据对模型进行微调；HT+，将采用多角色协作机制进行数据质量增强后的数据对模型进行微调。

为验证跨语言思维链问答推理方法的有效性，设计五个基线进行了比较，即：origin，模型直接接收原始藏文输入；single，采用单语言思维链的方式，模型需要在输入后生成推理步骤；Few-shot，采用少样本思维链的方式，在给模型输入问题前，拼接六个示例，使得模型学习相应的推理方式；Cross-Lingual-Vote，通过投票机制保留三种语言在推理结果中表现出高度一致性的答案；Cross-Lingual-Score，对不同语言路径生成的响应进行评分选择最优路径并将辅助语言的推理结果转换为目标语言的最终响应。

序号	Prompt	Accuracy
1	无提示词。	58.0
2	让我们一步步思考。	63.3
3	请理解提示和问题信息，然后逐步解决问题并显示答案。	64.5
4	让我们先准备相关信息并制定计划。然后，一步一步地回答这个问题（注意常识和逻辑连贯性）。	66.7
5	你是常识问答比赛的参赛者，总是准确地回答各种常识问题。我是比赛的主持人，决赛就要开始了。	67.2
6	你是常识问答推理专家，请理解问题，并确定问题的关键信息，结合相关知识，逐步解答问题，并确定答案是“yes”还是“no”。	**68.6**

Table 4: 不同提示策略的实验结果

4.6 实验结果与分析

本研究首先使用GPT-3.5-turbo作为教师模型，在自建常识问答数据集上测试，实验结果如表4所示。零样本方法中，HT提示策略准确率达68.6%，优于其他方法。输入数据后，模型输出答案及推理步骤，整理成<问题，推理步骤，答案>三元组，用于后续微调小型学生模型。

针对融合思维链和微调技术的实验共涉及LLaMa2-7B、LLaMa2-7B-Chinese和TiLamb-7B三种模型，涵盖中文、英文和藏文三种语言。实验结果如表5所示。

实验表明，融合思维链和微调技术的方法在三种模型中均表现出积极作用。尽管思维链方法（如Zero-shot-CoT和Few-shot-CoT）的表现并不总是优于直接问答（Zero-shot），但思维链方法在某些情况下仍能显著提升模型性能，原因在于其能够引导模型进行逻辑推理，帮助其更好地理解问题结构和关键信息。

微调对模型性能的提升具有显著作用。在LLaMa2-7B-Chinese模型中，Zero-shot方法的准确率为57.4%，而经过Fine-tune-qa微调后，准确率提升至61.5%；经过Fine-tune-CoT微调后，准确率进一步提升至63.3%。在LLaMa2-7B模型和TiLamb-7B模型中也存在类型情况。这表明微调能够使模型更好地适应特定任务的数据分布，而不仅仅是依赖于预训练阶段的通用知识，使其更适合于当前任务的需求，从而提高其在问答任务中的表现。

方法	Accuracy	F1-yes	F1-no	F1-avg
LLaMa2-7B-Chinese				
Zero-shot	57.4	56.3	58.1	57.2
Zero-shot-CoT	56.1	55.6	58.8	57.2
Few-shot-CoT	58.6	58.4	58.2	58.3
Fine-tune-qa	61.5	62.8	60.8	61.8
Fine-tune-CoT	63.3	62.2	64.4	63.3
HT	66.7	63.1	67.1	65.1
HT+	**68.1**	**66.3**	**69.3**	**67.8**
LLaMa2-7B				
Zero-shot	61.0	60.6	62.4	61.5
Zero-shot-CoT	61.2	63.2	61.1	62.2
Few-shot-CoT	65.3	66.2	63.4	64.8
Fine-tune-qa	68.4	67.2	66.8	67.0
Fine-tune-CoT	70.5	70.7	70.4	70.5
HT	71.0	**71.4**	71.4	71.4
HT+	**73.4**	70.3	**72.7**	**71.5**
TiLamb-7B				
Zero-shot	52.3	53.2	51.8	52.5
Zero-shot-CoT	52.2	53.2	52.0	52.6
Few-shot-CoT	53.7	52.7	54.5	58.6
Fine-tune-qa	58.6	54.4	62.1	58.3
Fine-tune-CoT	61.1	63.4	58.0	60.7
HT	62.7	58.2	**64.5**	61.4
HT+	**64.3**	**64.9**	62.6	**63.8**

Table 5: 融合思维链和微调的问答推理实验结果

使用思维链微调方法（如Fine-tune-CoT、HT、HT+）与直接问答微调方式Fine-tune-qa相比，在模型性能上表现出显著的提升。这种原因在于思维链微调方法能够使模型学习到推理步骤，而不仅仅是问题和答案之间的关系。通过学习推理步骤，模型能够更好地理解问题的结构和逻辑，并被激发了推理能力。

数据质量提升后的思维链微调如HT+与HT相比，在模型性能上表现出显著的提升。原因

在于数据质量的提升能够使模型学习到更准确和更有效的推理步骤。通过多角色协作机制进行数据质量增强，能够去除数据中的噪声和错误，从而提高数据的质量和可靠性。

融合思维链微调后的学生模型如HT+方法与在教师模型采用零样本思维链进行问答推理的表现具有一定的可比性。学生模型通过学习教师模型生成的推理步骤，能够更好地理解问题的结构和逻辑，被激发出相当程度的推理能力，极大地增强了小型模型在问答推理下游任务中的表现。这也说明即使小型模型由于参数和训练数据的限制，其问答推理能力通常不如大型模型，但通过我们的方法小型模型能够在资源受限的情况下表现出比拟大型模型的效果。

纵观实验中所采用的三种模型以及三种语言，中英藏的表现存在一定的差异。这种差异的原因可能在于不同语言语料资源丰富度以及模型对语言的支持程度。

基于跨语言思维链的实验共涉及TiLamb-13B和DeepSeek满血版两种模型。相较于TiLamb-7B，TiLamb-13B的思维链提示效果表现的更好。DeepSeek在本实验中充当教师模型和评估器的角色。实验结果如表6所示。跨语言思维链方法效果显著。Cross-Lingual-Vote利用中、英文推理能力，通过一致性投票机制，准确率较Few-shot提高3.6%，F1-avg提升3.5，表明跨语言知识迁移和一致性投票能有效提升藏文问答可靠性；Cross-Lingual-Score通过引入教师模型评分机制，准确率达到66.7%，F1-avg提升至66.5%，较Cross-Lingual-Vote提升1.6%，说明评分策略可更精准筛选高质量答案。

从origin到Cross-Lingual-Score的对比表明，单独依赖藏文模型推理效果有限，引入跨语言推理方法后性能明显提升。这说明跨语言知识迁移可弥补藏文低资源不足，一致性投票能过滤错误推理路径，教师模型评分机制可进一步增强推理质量、减少噪声，提升藏文问答性能。

方法	Accuracy	F1-yes	F1-no	F1-avg
origin	59.3	59.2	59.8	59.5
single	60.7	61.5	60.3	60.9
Few-shot	61.5	62.3	61.0	61.7
Cross-Lingual-Vote	65.1	63.6	66.8	65.2
Cross-Lingual-Score	**66.7**	**66.2**	**67.0**	**66.5**

Table 6: 跨语言思维链知识迁移的实验结果

4.7 消融实验

为验证跨语言思维链方法的有效性和可扩展性，对藏文问答进行消融实验，并对中文和英文进行比较实验。实验结果如表7所示，新增三个评价规则：Cross-Lingual-Score (English)、Cross-Lingual-Score (Chinese)和Cross-Lingual-Score (Tibetan)，分别表示在跨语言思维链中仅使用英文、中文或藏文作为辅助语言生成推理步骤的情形。

方法	en	zh	ti
origin	65.3	61.8	59.3
single	65.6	62.4	60.7
Few-shot	64.8	62.6	61.5
Cross-Lingual-Score (English)	65.6	65.6	62.6
Cross-Lingual-Score (Chinese)	66.5	62.4	64.3
Cross-Lingual-Score (Tibetan)	65.1	62.5	60.7
Cross-Lingual-Score	**66.5**	**68.2**	**66.7**

Table 7: 消融实验结果

实验结果显示，英文路径在藏文任务中的准确率为62.6%，高于原始藏文的59.3%，但低于多语言融合的66.7%。这表明英语作为高资源语言，其丰富的语料和逻辑表达能力能有效弥补藏文数据不足，但因与藏文在语言结构和文化背景上差异较大，直接翻译可能导致语义丢失或逻辑偏差。相比之下，中文路径在藏文任务中准确率达到64.3%，高于英文路径，可能是因为

汉语与藏语同属汉藏语系，在词汇和表达方式上存在一定对应关系，尽管语法差异仍较大，但汉语作为辅助语言能更自然地贴近藏语表达。

单独使用英文或中文作为辅助语言时，在藏文任务中均有一定提升，表明两者在推理中具有显著优势。跨语言思维链框架在所有语言上均实现性能提升，尤其在低资源语言上效果显著，说明该框架能有效利用高资源语言知识，提升低资源语言的推理能力。

4.8 人工评估

为进一步验证模型在低资源语言（如藏语）上的推理能力，我们设计了人工评估实验。我们从多语言问答数据集中随机抽取100条数据，针对每个问题，使用TiLamb-13B模型生成回答，形成完整的模型输出样本。表8为一个样本示例。

藏文问题	བྱ་ཇ་མོང་ནི་འཕྲུར་ཤེས་པའི་བྱ་ཞིག་ཡིན་ནམ།
模型回答	བྱ་ཇ་མོང་གི་གཤོག་པ་ནི་གསོ་བོ་ལུས་ཕྱང་དོ་མཉམ་དང་རྒྱུག་སྐབས་བཀན་བཟློང་རང་བཞིན་རེ་ཅེར་གིང་བར་བགོ་པ་བཞིན་ཡོད་པས། འཕྲུར་ཤེས་པའི་ཐུན་ཚོ་ནས་པ་མི་ཤ་ཟན།
人工评估	事实准确性-6.0 语言流畅性-5.0 推理一致性-6.0

Table 8: 人工评估样例图

我们邀请3位藏语母语者进行评估。评估者背景多样，包括不同年龄、教育背景和职业，以减少评估偏差。对每个样本从三个维度进行主观评分，评分标准同样采用表2所示准则。每位评估者独立完成评分，我们对每个样本计算其各维度的平均得分。部分评价结果如表9所示。

问题编号	事实准确性	语言流畅性	推理一致性
Q1	6.0	5.0	6.0
Q2	6.0	4.0	5.0
Q3	1.0	3.0	2.0
...	...	...	...
Q100	6.0	5.0	4.0
平均	**4.4**	**4.7**	**5.1**

Table 9: 藏文问答生成结果的人工评估结果

该定性评估进一步验证了本文方法在藏文问答推理任务中的实用性与可靠性。模型不仅能够生成内容合理、结构清晰的答案，还具有较高的语言自然度，能够胜任藏文问答场景。

5 总结

本文聚焦于思维链和知识迁移，强化小型模型在问答推理方面的表现。本文提出融合思维链和微调技术的方法，基于HT提示策略模拟人类思维过程，优化大模型的推理生成。通过指令微调，将包含问题、推理步骤和答案的样本用于小模型的微调，提升小模型的推理性能。通过引入多角色协作机制，优化推理步骤质量，进一步提升小模型的推理能力。针对低资源语言如藏文问答推理效果不佳的问题，本文探索了跨语言思维链提示方法，利用高资源语言的知识基础弥补低资源语言的知识不足。实验表明，该策略有效提升了藏文的问答推理效果。接下来，我们将继续完善多语言问答推理的数据集，改进小型模型并拓展到其他低资源语言问答任务中，提升小型模型在多语言问答推理任务中的泛化能力和适应性。

致谢

本论文得到了国家社会科学基金(22&ZD035)，中国工程院科技战略咨询项目(2025-XZ-16-06)，国家自然科学基金(61972436)，中央民族大学项目(2025XYCM39)的资助。

参考文献

OpenAI Josh Achiam, Steven Adler, and Sandhini Agarwal. 2023. Gpt-4 technical report.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke S. Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Conference on Empirical Methods in Natural Language Processing*.

Tom B. Brown, Benjamin Mann, and Nick Ryder. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Pengfei Cao, Yuheng Chen, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2024. One mind, many tongues: A deep dive into language-agnostic knowledge neurons in large language models. *ArXiv*, abs/2411.17401.

Aakanksha Chowdhery, Sharan Narang, and Jacob Devlin. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Karl Cobbe, Vineet Kosaraju, and Mo Bavarian. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *ArXiv*, abs/2304.08177.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. In *Conference on Empirical Methods in Natural Language Processing*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017:1601-1611.

Saurav Kadavath, Tom Conerly, and Amanda Askell. 2022. Language models (mostly) know what they know. *ArXiv*, abs/2207.05221.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

SHIYANG LI, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *ArXiv*, abs/2210.06726.

Bingbin Liu, Sébastien Bubeck, Ronen Eldan, Janardhan Kulkarni, Yuanzhi Li, Anh Nguyen, Rachel Ward, and Yi Zhang. 2023. Tinygsm: achieving ¿80% on gsm8k with small language models. *ArXiv*, abs/2312.09241.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *ArXiv*, abs/2403.10258.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and A. Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *ArXiv*, abs/2209.14610.

Haipeng Luo, Qingfeng Sun, and Can Xu. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *ArXiv*, abs/2308.09583.

第二十四届中国计算语言学大会论文集，第170页-第183页，济南，中国，2025年8月11日至14日。

Maxwell Nye, Anders Andreassen, and Guy Gur-Ari. 2021. Show your work: Scratchpads for intermediate computation with language models. *ArXiv*, abs/2112.00114.

Arkil Patel, S. Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *North American Chapter of the Association for Computational Linguistics*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023:2695-2709.

Leonardo Ranaldi and Fabio Massimo Zanzotto. 2023. Empowering multi-step reasoning across languages via tree-of-thoughts. *ArXiv*, abs/2311.08097.

Victor Sanh, Albert Webson, and Colin Raffel. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Zhihong Shao, Damai Dai, Daya Guo, Bo Liu (Benjamin Liu), Zihan Wang, and Huajian Xin. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *ArXiv*, abs/2405.04434.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *ArXiv*, abs/2210.03057.

Aarohi Srivastava, Abhinav Rastogi, and Abhishek Rao. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *ArXiv*, abs/2402.16438.

Eshaan Tanwar, Manish Borthakur, Subhabrata Dutta, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023:6292-6307.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Weixuan Wang, Barry Haddow, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *ArXiv*, abs/2406.09265.

Jason Wei, Yi Tay, and Rishi Bommasani. 2022a. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

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

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024:15366-15394.

Yixuan Weng, Minjun Zhu, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. 2022. Large language models are reasoners with self-verification. *ArXiv*, abs/2212.09561.

Zhuang Wenhao, Sun Yuan, and Zhao Xiaobing. 2024. Tilamb: A tibetan large language model based on incremental pre-training. In *China National Conference on Chinese Computational Linguistics*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Zheng Yuan, Hongyi Yuan, Cheng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. *ArXiv*, abs/2308.01825.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alexander J. Smola. 2022. Automatic chain of thought prompting in large language models. *ArXiv*, abs/2210.03493.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *ArXiv*, abs/2308.04371.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *ArXiv*, abs/2302.10198.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *ArXiv*, abs/2408.10811.

Denny Zhou, Nathanael Scharli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Olivier Bousquet, Quoc Le, and Ed H. Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *ArXiv*, abs/2205.10625.

第二十四届中国计算语言学大会论文集，第170页-第183页，济南，中国，2025年8月11日至14日。