基于多维度答案筛选的低资源语言开放域问答方法

王新阳[1,2], 关昕[1,2], 张利飞[1,2], 余正涛[1,2], 黄于欣*[1,2]
[1] 昆明理工大学，信息工程与自动化学院，昆明，650500
[2] 昆明理工大学，云南省人工智能重点实验室，昆明，650500
xinxinyang222@163.com, 876044483@qq.com, dayniizhang@gmail.com
ztyu@hotmail.com , huangyuxin2004@163.com

摘要

开放域问答通常是从大规模数据中检索多个相关文档，并利用大语言模型对文档内容进行理解生成答案。然而，面向缅甸语、老挝语等低资源语言，检索到的数据可能存在问题无关的噪声文档，且大语言模型对低资源语言理解能力弱，生成答案错误率高。对此，提出一种基于多维度答案筛选的低资源语言开放域问答方法，将现有基于大模型直接理解文档生成答案的过程，转换成多个候选答案生成并筛选的多阶段过程。在答案生成阶段，从文档中抽取多样化的候选答案，在筛选阶段，设计多维度答案筛选策略，通过全局篇章答案验证、局部证据答案验证以及不同答案相关性排序，筛选出最优答案。在四种东南亚低资源语言开放域问答数据集上的实验结果表明，基于GPT-4o-mini、DeepSeek-v3等大语言模型底座，提出方法相比思维链、摘要验证等最优方法都取得了更好的性能，验证了多阶段答案生成筛选过程在低资源开放域问答任务中有效性。

关键词： 开放域问答 ； 大语言模型 ； 低资源语言 ； 多维度答案筛选

A Multi-Dimensional Answer Filtering Approach for Open-Domain Question Answering in Low-Resource Languages

Xinyang Wang[1,2], , Xin Guan[1,2], Lifei Zhang[1,2], ZhengtaoYu[1,2], Yuxin Huang*[1,2]
[1] Faculty of Information Engineering and Automation,
Kunming University of Science and Technology Kunming 650500, China
[2] Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology Kunming 650500, China
xinxinyang222@163.com, 876044483@qq.com, dayniizhang@gmail.com
ztyu@hotmail.com , huangyuxin2004@163.com

Abstract

Open-domain question answering (ODQA) typically involves retrieving multiple relevant documents from large-scale corpora and leveraging large language models (LLMs) to comprehend the content and generate answers. However, in low-resource languages such as Burmese and Lao, the retrieved documents often contain irrelevant noise, and LLMs exhibit limited understanding capabilities, leading to high error rates in answer generation. To address these challenges, we propose a multi-stage ODQA framework for low-resource languages based on multi-dimensional answer filtering. Instead of generating answers directly from documents, our approach first extracts diverse candidate answers, followed by a multi-dimensional filtering process. This includes global

*黄于欣（通讯作者）：huangyuxin2004@163.com
基金项目：国家自然科学基金(62266027，U23A2038，62266028)；云南省重大科技专项项目(202302AD080003, 202402AG050007, 202303AP140008)；云南省基础研究项目(202301AS070047，202301AT070471)；昆明理工大学“双一流”创建科技专项(202201BE070001-021)

discourse-level validation, local evidence verification, and inter-answer relevance ranking to identify the optimal answer. Experiments on four Southeast Asian low-resource ODQA datasets demonstrate that our method, built upon LLMs such as GPT-4o-mini and DeepSeek-v3, outperforms state-of-the-art techniques including chain-of-thought reasoning and summary verification, confirming the effectiveness of our multi-stage answer generation and filtering framework.

Keywords: open-domain question answering , large language models , low-resource languages , multi-dimensional answer filtering

1 引言

开放域问答（Open-Domain Question Answering, ODQA）(Chen and Yih, 2020)通常是从大规模数据中检索多个相关文档，并对文档内容进行理解以生成答案。当前，开放域问答任务主要采用检索-阅读（Retriever-Reader）(Chen et al., 2017)框架。在该框架中，检索器通常基于稠密检索如DPR(Karpukhin et al., 2020)或稀疏检索如BM25算法(Robertson et al., 2009)等技术，从大规模文档集合中筛选相关文本段落。阅读器则依托生成式语言模型，通过对检索内容的语义理解和文档分析生成符合问题要求的答案。早期研究主要采用基于Transformer架构(Vaswani et al., 2017)的预训练语言模型（Pretrained Language Model）作为阅读器，例如经过微调的BERT(Devlin et al., 2019)或T5(Raffel et al., 2020)。近年来，随着大语言模型的发展，以ChatGPT(OpenAI, 2022)、Llama系列(Touvron et al., 2023)和DeepSeek(Dai et al., 2024)等为代表的大语言模型凭借其强大的生成与推理能力，已成为当前阅读器的优选方案。

基于大语言模型作为阅读器的方法主要可归纳为两类：单轮生成答案的方法与生成候选答案验证的方法。单轮生成答案的方法指的是大语言模型直接生成答案，或在对文档内容进行理解与排序后生成单一答案。例如，Lewis et al. (2020) 提出了将外部文档检索与生成模型相结合，利用检索到的文档来提升模型在开放域问答任务中准确性。Wei et al. (2022) 提出的思维链（Chain-of-Thought, CoT）提示策略能够引导模型在生成答案前对问题和文档进行中间推理，有效提升了答案的准确性。此外，Sun et al. (2023) 则探讨了GPT等模型在多文档排序任务中的能力，提出采用滑动窗口策略筛选最相关段落，以此生成更准确的答案。与单轮生成方法不同，生成候选答案验证的方法通过生成多个候选答案提高正确答案的命中概率，并基于候选答案生成相应证据和推理过程，利用生成内容进行验证以确定最终输出。例如，Weng et al. (2023) 提出采用多重采样策略，在解码阶段并行生成多个答案，并以生成的答案为已知条件，进行反向推理以进行验证并选出最终结果。Kim et al. (2024a) 则提出先通过提示策略生成多个候选答案，随后为每个答案生成摘要，并基于摘要的验证结果确定最终答案。

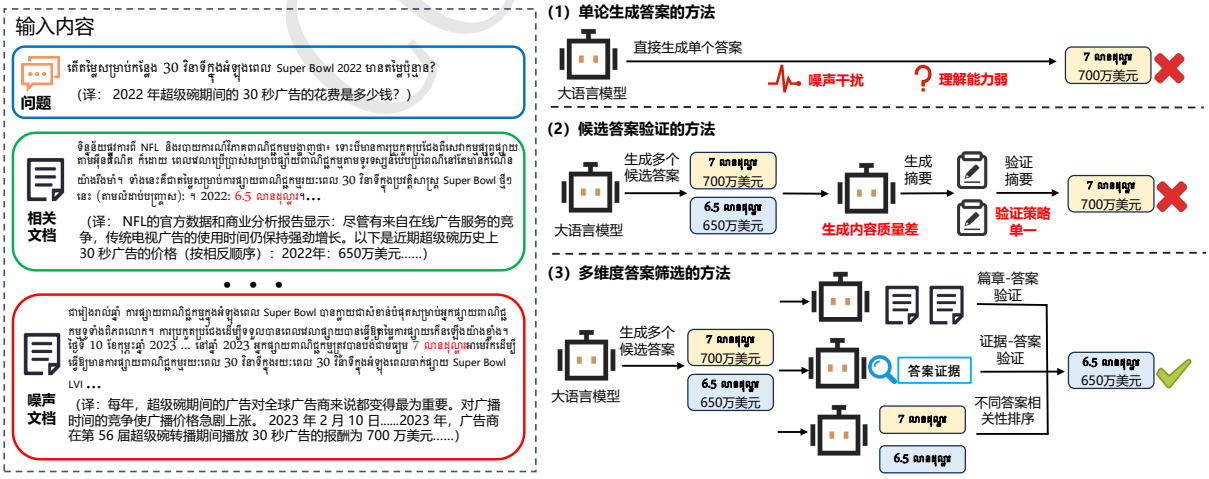

图 1: 低资源语言开放域问答示例

在英语、中文等高资源语言的开放域问答任务中，基于大语言模型的方法已取得显著成效。然而，在处理缅甸语、柬埔寨语等低资源语言时，仍面临诸多挑战。单轮生成答案的方法

在低资源语言环境下易出现推理错误，主要原因在于输入文档中往往包含大量与问题无关的噪声信息。这些噪声源于低资源语言外部知识库的匮乏，以及相关的检索技术尚不成熟。同时，主流大语言模型对低资源语言文档的理解能力有限，进一步加剧了问题推理的偏差。如图 1所示，在处理高棉语问题“2022年超级碗期间的30秒广告费用是多少钱”时，模型未能有效整合问题与文档信息，错误地输出了2023年的广告费用。尽管基于候选答案生成的验证方法通过多答案生成机制在一定程度上提升了正确答案的覆盖率，但仍存在明显不足。一方面，噪声干扰与模型对低资源语言处理能力的局限性导致生成证据推理质量较低，影响了后续验证效果；另一方面，现有验证方法较为单一，难以有效识别并筛选出正确答案。在同一案例中，尽管生成的候选答案中包含正确的“650万美元”，但因验证机制缺乏鲁棒性，导致模型任无法有效识别出正确答案。

针对上述问题，本文提出一种基于多维度答案筛选的低资源语言开放域问答方法。与现有基于大模型直接理解文档生成答案的范式不同，该方法将问答过程转换为多阶段候选答案生成与筛选框架。首先在答案生成阶段引导模型从多文档中生成多个候选答案以提高正确答案覆盖率；随后在答案筛选阶段设计多维度验证策略：基于全局文档信息对每个答案进行直接验证，针对每个答案抽取相关证据进行局部验证，以及通过不同答案间的相关性对比排序。通过这三个维度的协同筛验证机制，模型选择出最优答案。

2 相关工作

开放域问答系统通常采用基于检索-阅读的框架：首先从海量知识库中检索与问题相关的文档，随后利用语言模型对文档信息进行深度语义理解与推理，最终生成符合问题需求的答案。早期的研究主要通过微调训练传统的预训练模型来提升问答任务性能。Karpukhin et al. (2020)提出的DPR方法采用双编码器架构对问题和段落进行向量化表示，通过稠密检索技术显著提升了开放域问答系统的性能。Lewis et al. (2020)提出RAG框架将预训练的生成模型与检索器相结合，首先从外部文档库中检索出与问题相关的文档片段，然后将这些片段与问题一起输入生成模型以生成答案。Izacard and Grave (2021)提出的FiD方法通过在解码阶段融合多个检索到的段落信息，显著提高了生成答案的准确性。Cheng et al. (2021)提出UnitedQA框架创新性地结合了基于BERT的抽取式阅读器和基于T5的生成式阅读器，通过联合训练策略优化两者的互补性，从而提升开放域问答性能。

随着大语言模型的兴起，开放域问答任务研究进入新的发展阶段，研究者主要关注在阅读阶段大模型如何有效利用检索到的信息生成准确的答案。Lazaridou et al. (2022)提出了一种基于少量提示的大语言模型与互联网信息结合的方法，该方法通过检索网页内容生成候选答案，并采用重排序策略选择最优答案，在无需微调的情况下显著提升了开放域问答性能。Chen et al. (2023)提出了一种基于外部QA记忆库增强的大语言模型方法，在推理时检索相关问答对，从而提高了开放域问答系统的性能和效率。Shi et al. (2024)采用集成学习策略，通过大语言模型对每段检索到的段落进行独立预测，并采用特定的投票机制汇总结果来得出最终答案。Kim et al. (2024b)提出QPaug的方法通过将复杂问题分解为子问题，并利用大语言模型生成的内容扩充检索到的段落，显著提高了ODQA任务的性能。Kim et al. (2024a)提出了Sure方法通过大语言模型生成多个候选答案，通过答案生成摘要并进行验证来确定最终答案。

3 方法

3.1 问题描述

本文旨在基于提示学习的方法，提升大语言模型在低资源语言开放域问答任务中作为阅读器性能。具体而言，给定低资源检索文档集合$\widetilde{\mathrm{D}}$，本文将其划分为两类：若文档$d \in \widetilde{\mathrm{D}}$包含与查询问题$q$的正确答案$a$相关的信息，则定义为相关文档$d_{\text{golden}}$；反之，若$d$不包含$a$或混杂无关内容，则视为噪声文档$d_{\text{noisy}}$。本文的目标是通过设计提示集合$P = \{p_1, p_2, ..., p_m\}$，构建如图 2所示的多阶段提示引导的低资源语言问答框架，引导大语言模型M利用文档集合$\widetilde{\mathrm{D}}$中的信息，通过多阶段生成与筛选策略预测准确答案。最终任务可以形式化为以下问题：

$$\widetilde{\mathrm{D}} = \{d_{\text{golden}}, d_{\text{noisy}}, \ldots, d_{\text{noisy}}\} \quad \text{and} \quad \widetilde{a} = M(\Phi(P, q, \tilde{D})) \tag{1}$$

其中多阶段的任务函数Φ 定义为：

$$\Phi(P, q, \tilde{D}) = M|_{p_m}(\cdot) \circ M|_{p_{m-1}}(\cdot) \circ \ldots \circ M|_{p_1}(\cdot) \tag{2}$$

其中每个$M|_{p_i}(\cdot)$表示大语言模型在特定提示p_i指导下的任务处理阶段，整个系统通过模块化流程设计(∘)将这些处理阶段动态整合，最终构建出完整的问答框架。

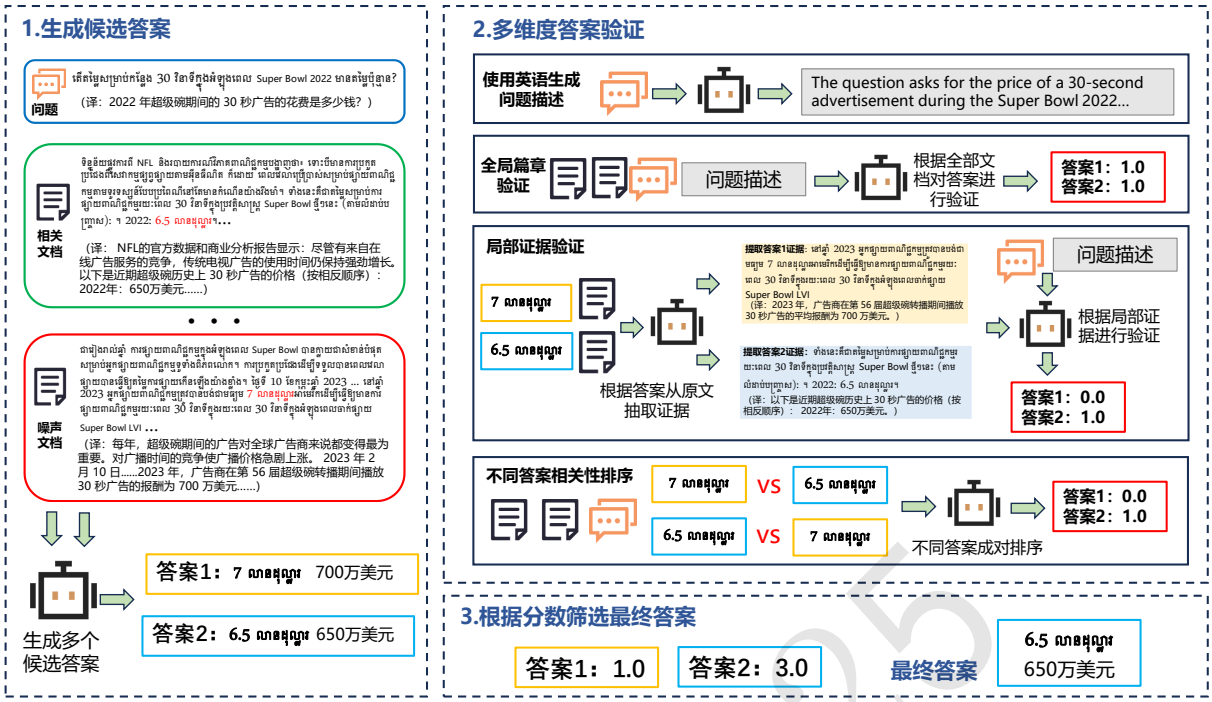

图 2: 基于多维度答案筛选的低资源语言开放域问答模型

3.2 候选答案生成

大语言模型在直接理解低资源文档并生成答案时，受限于自身能力及噪声信息的干扰，导致输出结果准确性不足。本文采用提示学习的方法，明确要求大语言模型生成多个候选答案，通过增加答案的数量提高正确答案的命中率。传统方法通常依赖多重采样技术生成候选答案，后续的分析实验观察到，通过提示的方法成多个候选答案，能够生成更加多样化的候选答案。

给定低资源语言的查询问题q，包含噪声的低资源语言文档集合$\widetilde{\mathrm{D}}$以及大型语言模型M，本文在提示p_{ans} 中设计了提示指令，以引导模型从文档中抽取并生成包含多个潜在正确答案的候选答案集合。

$$A_{\text{cand}} = M|_{p_{\text{ans}}}(q, \widetilde{D}) = \{a_i\}_{n=1}^{N} \tag{3}$$

其中A_{cand} 表示大语言模型生成的候选答案集合，N表示生成候选答案的数量，a_i表示每个潜在正确的候选答案。

3.3 多维度答案验证

为了从生成的候选答案集合中筛选出正确的答案，本文设计了多维度答案验证策略，从全局篇章答案验证，局部证据答案验证以及不同答案相关性排序模块对每个候选答案进行综合性的验证评估。

3.3.1 问题分析

由于当前主流大语言模型主要基于英语等高资源语言进行预训练，其在英语相关任务中表现出卓越的性能。为充分利用大语言模型在英语领域的优势以辅助低资源语言任务处理，本文在答案验证阶段引入英语辅助机制。在验证评估前，我们要求模型首先以英语梳理任务需求，通过英语的问题描述为后续验证过程提供辅助信息，从而提升模型判断的准确性。

针对原始低资源语言查询问题q，本文设计了英语问题描述提示p_{en}，以引导大型语言模型对低资源语言查询进行深入分析。提示p_{en}明确要求模型使用英语表述对输入问题的语义进行解构，重点识别其潜在的信息需求及预期的答案类型。在提示p_{en} 的指导下，大语言模型以低资

源语言查询q 为输入，生成对应的英文问题描述des_{en}。问题描述des_{en}作为补充说明信息，进一步被整合到后续的答案验证流程中，以提升系统回答的准确性。

$$des_{en} = M|_{p_{\text{en}}}(q) \tag{4}$$

3.3.2 全局篇章答案验证

受Weng et al. (2023)研究的启发，大语言模型能够在提示学习的策略下对其生成结果进行自我验证。针对大语言模型在低资源任务中理解能力受限的问题，引入英语问题描述信息作为补充语义特征，增强模型对查询问题的理解深度。在此基础上，结合全部的文档集合信息，对每个候选答案a_i进行全局篇章级别的验证，确保筛选答案的合理性和有效性，同时保证答案与查询问题在逻辑和语义层面的一致性。

本文设计了全局验证提示p_{glob}，将每个候选答案a_i、文档集合$\widetilde{D}$、输入查询问题q 以及问题描述des_{en} 一并输入至大语言模型M 中。根据提示p_{glob} 的指引，模型利用全文篇章信息，从正确性与合理性角度为候选答案生成分析评估并做出判断。模型生成完成后，采用正则表达式匹配的方式提取其判别结果。评估为True 的候选答案被赋予1 分，未达标准的答案则获得0分评估值。

$$\mathcal{V}_{glob}(a_i) = M|_{p_{\text{glob}}}\left(q, des_{en}, a_i, \widetilde{D}\right) \tag{5}$$

$$\text{score1}(a_i) = \begin{cases} 1 & \text{if } \text{Re}\left(\mathcal{V}_{\text{glob}}(a_i)\right) = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

其中$\mathcal{V}_{glob}(a_i)$表示模型基于全局篇章对候选答案a_i生成的评估内容，$\text{Re}(\cdot)$表示使用正则匹配从评估内容中提取的结果。

3.3.3 局部证据答案验证

全局篇章验证能够从上下文一致性上验证答案的合理性。此外，本文进一步的引入与答案相关的细粒度证据信息进行验证，从不同的维度上实现对候选答案更精准的判别。具体而言，文本利用证据提取函数$\mathcal{E}_{\text{evi}}$，将每个候选答案a_i，输入查询问题q以及文档集合$\widetilde{D}$作为输入，在证据抽取提示p_{evi}的作用下，提取出每个候选答案a_i来源的信息$evidence_i$作为细粒度的证据。

$$evidence_i = \mathcal{E}_{\text{evi}}\left(q, a_i, \widetilde{D}\right) = M|_{p_{\text{evi}}}\left(q, a_i, \widetilde{D}\right) \text{ for } i = 1, \ldots, n \tag{7}$$

其中，提取证据函数满足:

$$evidence_i \subset \widetilde{D}$$

与Kim et al. (2024a)的方法生成摘要验证不同，本文的方法是直接引导模型从原文中提取内容，与生成摘要内容相比，提取的内容与原文的相关性更高，且在低资源环境下有效规避了模型生成能力不足的问题。在提取相关证据后，本文基于局部证据对每个候选答案进行细粒度的评估。这些证据直接来源于原始检索文档，并与候选答案保持高度相关性。候选答案的置信度越高，提取到的相关证据质量也越可靠，因此通过评估局部证据是否对查询问题提供帮助信息，可以间接验证答案的有效性。

具体而言，本文设计了局部验证提示p_{part}，将候选答案a_i 的证据$evidence_i$、查询问题q 及英文问题描述des_{en} 一并输入大语言模型M中。在p_{part} 的引导下，模型判断与候选答案a_i相关的局部证据$evidence_i$ 是否对回答问题q 提供了有效支持，并在生成分析内容后给出判别结果。模型生成完成后，与全局验证类似，模型分析的输出经由正则表达式匹配提取判别结果，若评估为True，则该候选答案获得1 分，否则赋值为0分。

$$\mathcal{V}_{\text{part}}(a_i) = M|_{p_{\text{part}}}(q, des_{en}, evidence_i) \tag{8}$$

$$\text{score2}(a_i) = \begin{cases} 1 & \text{if } \text{Re}\left(\mathcal{V}_{\text{part}}(a_i)\right) = \text{True} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

其中$\mathcal{V}_{part}(a_i)$表示模型基于局部证据对候选答案a_i生成的评估内容，$\mathrm{Re}(\cdot)$表示使用正则匹配从评估内容中提取的结果。

3.3.4 不同答案相关性排序

在完成全局篇章验证和局部证据验证对答案自身的质量评估后，本文进一步采用大模型成对排名方法(Qin et al., 2024)，通过直接比较不同答案之间的相关性来区分其优劣程度。具体而言,将候选答案集合A_{cand} 中的每个答案a_i与其他候选答案a_k输入到答案对比函数$\mathcal{C}_{\text{comp}}$ 中，模型根据文档集合$\widetilde{D}$以及输入查询问题q在答案对比提示p_{comp} 的指引下，判断候选答案a_i或a_k哪一个更合适作为查询问题的正确答案。最终，基于答案a_i被选中的次数，计算出候选答案的得分，并依此对所有候选答案进行排名。

$$\text{score}_3(a_i) = \frac{1}{N-1}\sum_{k\neq i}^{N}\mathcal{C}_{\text{comp}}(a_i, a_k) \tag{10}$$

其中答案对比函数$\mathcal{C}_{\text{comp}}$ 定义为：

$$\mathcal{C}_{\text{comp}}(a_i, a_k) = \begin{cases} 1, & M|_{p_{\text{comp}}}\left(q, a_i, a_k, \widetilde{D}\right) \to a_i \\ 0, & M|_{p_{\text{comp}}}\left(q, a_i, a_k, \widetilde{D}\right) \to a_k \\ 0.5, & \text{otherwise} \end{cases} \tag{11}$$

3.4 最终答案选择

经过多维度答案筛选体系的综合评估，每个候选答案依次通过全局篇章验证、局部证据验证以及不同答案相关性排序三个环节的评分验证。在评估过程中，各模块采用等权重计算方式，最终通过汇总三个环节的评分结果，选取分数最高的候选答案作为最终输出。

$$\widetilde{a} = a_{i^*}, \quad i^* = \arg\max_{i\in[1,N]}\left(\sum_{j=1}^{3}\text{score}_j(a_i)\right) \tag{12}$$

算法 1总结基于多维度答案筛选的低资源语言开放域问答方法的形式化过程，本方法实现过程中使用的所有提示词模板均已完整收录在附录 A中。

算法 1: 基于多维度答案筛选的低资源语言开放域问答算法流程

输入 : 低资源查询问题q，文档集合$\widetilde{D}$，大模型M，提示集P
输出 : 最终答案$\widetilde{a}$

1. **候选答案生成**: $A_{\text{cand}} \leftarrow M|_{p_{\text{ans}}}(q, \widetilde{D})$
2. **多维度验证** ($\forall a_i \in A_{\text{cand}}$):
   - ▷ 问题英语分析: $des_{\text{en}} \leftarrow M|_{p_{\text{en}}}(q)$
   - ▷ 全局篇章验证: $\text{score}_1 \leftarrow M|_{p_{\text{glob}}}(q, des_{\text{en}}, a_i, \widetilde{D})$
   - ▷ 证据提取与局部验证: $\text{evidence}_i \leftarrow M|_{p_{\text{evi}}}(q, a_i, \widetilde{D})$; $\text{score}_2 \leftarrow M|_{p_{\text{part}}}(q, des_{\text{en}}, \text{evidence}_i)$
   - ▷ 不同答案相关性排序: $\text{score}_3 \leftarrow \frac{1}{N-1}\sum_{k\neq i}^{N}\mathcal{C}_{\text{comp}}(a_i, a_k)$
3. **答案选择**: $\widetilde{a} \leftarrow a_{i^*}$, $i^* = \arg\max_i(\sum_{j=1}^{3}\text{score}_j(a_i))$

4 实验

4.1 数据集构建

为评估大语言模型在低资源开放域问答任务中的性能，本文基于Chen et al. (2024)提出的RGB数据集，构建了一个专门面向东南亚低资源任务的问答基准数据集，并将其命名为LRQA (Low-resource QA)，LRQA数据集涵盖了泰语、缅甸语、老挝语和柬埔寨语四种低资源语言。具体而言，本文从RGB数据集中抽取了部分英语数据，并利用Google翻译引擎将每条数据中的问题、相关文档和噪声文档翻译成目标低资源语言文本。随后，本文聘请了专业的低

资源语言母语者对翻译结果进行了审校与校正。基于校正后的内容，进一步标注了与问题相关的正确答案。

在具体的实验设定中，受实验成本的限制，本文构建了800个问答对以及4000个文档，每种低资源语言包含200条测试数据。为了更真实地模拟低资源语言开放域问答任务中检索文档存在大量噪声的情况，本文将噪声比例设置为0.8。具体而言，针对每条测试数据，为其添加了一条包含正确答案的相关文档，以及四条不包含正确答案的噪声文档。随后，对相关文档与噪声文档的顺序进行了随机打乱，以构建最终的外部文档集合。完整的数据集由问题-答案对及其对应的文档集合共同组成。我们对各数据集不同组成部分的平均长度进行了统计，考虑数据涵盖多种语言风格的低资源语言，采用字符串长度作为统一的度量标准。各部分的长度统计如表 1所示。

数据集	Query	Answer	Passages
泰语(th)	43.74	11.84	740.31
高棉语(km)	51.90	13.41	869.97
老挝语(lo)	43.66	11.79	705.05
缅甸语(my)	54.62	12.18	848.88

表 1: LRQA数据集的平均长度信息。

4.2 评估指标

为了评估模型方法在开放域问答任务中的性能，本文选择了精确匹配（Exact Match，EM）和F1分数作为评估指标。精确匹配（EM）衡量的是测试数据集中正确答案的比率，如果给定的预测与正确答案之一完全一致，则认为该预测是正确的。F1分数则衡量了预测答案与正确答案之间词汇集合的重叠情况，旨在平衡正确识别答案和避免遗漏正确答案之间的权衡。

4.3 基线方法

本文与以下基线方法进行比较以验证提出方法的有效性:

（1）直接生成方法(Base): Base方法是通过将检索到的文段与问题联合输入到大语言模型中，并通过提示模型直接生成问题答案。

（2）思维链提示(CoT)(Wei et al., 2022): 在Base方法的基础上，采用零样本思维链提示策略(Zero-shot Chain of Thought)，引导模型逐步推理思考后再生成最终答案。

（3）跨语言提示(CLP)(Qin et al., 2023): 引导大语言模型使用英语对低资源语言的问题和段落进行思考与推理，然后得出目标低资源语言答案。

（4）语言多样性提示(LDP)(Nguyen et al., 2024): 采用了语言多样性的提示，利用高资源语言示例激活模型能力，再并通过目标语言示例引导模型生成低资源语言答案。

（5）大模型重排序方法(LLM Rerank)(Sun et al., 2023): 以大语言模型作为排序智能体，利用大语言模型对检索到的段落进行判别与排序，推行滑动窗口策略，从多个文档中选出与查询问题最为相关的文档，然后使用筛选出来的文档进行答案生成。

（6）自我验证方法(Self-verification)(Weng et al., 2023): 通过多重采样生成多个候选答案，然后利用大语言模型对自身推理的答案进行反向验证并选择最适合的答案。

（7）候选答案摘要验证(Sure)(Kim et al., 2024a): 通过提示模型生成多个候选答案，并基于这些候选答案生成支持性摘要。最终，通过评估摘要的质量来筛选出最佳答案。

4.4 实验结果

在本实验中，本文选用了性能强大且具备一定低资源语言处理能力的大语言模型，包括GPT-3.5-turbo、GPT-4o-mini以及DeepSeek-v3，以验证所提出方法的有效性。调用API时，将温度参数设置为0.0，以确保模型生成结果的稳定性。在实验中，将生成候选答案的数量固定为N=2，这是因为在后续的分析实验中本文观察到增加N对性能提升的较为有限。

本文在LRQA四个语言基准数据集上进行了对比实验，验证所提出方法的有效性。实验结果如表 2所示，主要发现如下：(1)在四种语言数据集上，本文的方法在性能上超越了大部

分现有方法。具体表现为：与次优方法相比，使用GPT-3.5-turbo模型时，EM和F1指标分别提升4.1%和4.4%；使用GPT-4o-mini模型时分别提升3.0%和3.1%；使用DeepSeek-v3模型时分别提升2.8%和2.7%，这一结果充分证明了本文提出方法的有效性。(2)实验观察到，在GPT-3.5-turbo和GPT-4o-mini模型上，单轮生成的方法如CoT(Wei et al., 2022)和CLP(Qin et al., 2023)受噪声干扰和低资源语言推理的局限性影响，性能提升不明显甚至出现下降。LDP方法(Nguyen et al., 2024)和LLMRerank方法(Sun et al., 2023)在多数的实验中上有小幅度的性能提升，但在泰语任务中出现性能下降，表明其适用性在低资源任务中不够稳定。基于候选答案验证的方法例如Sure(Kim et al., 2024a)和Self-verification(Weng et al., 2023)通过多候选答案机制提高了正确答案命中率，但因依赖单一验证策略和低质量摘要证据，性能提升有限。(3)DeepSeek-v3模型在多数实验中表现最优，展现出极高的适应性。所有基线方法在该模型上的表现均优于直接生成答案的方法，进一步验证了其优异的低资源语言泛化能力。本文认为这一优势源于其创新的多头潜在注意力机制（Multi-head Latent Attention，MLA）和DeepSeekMoE架构(Dai et al., 2024)，这些设计显著区别于传统大语言模型，有效增强了推理能力和低资源任务适应性。结合本文提出的多维度答案筛选方法，该模型在低资源开放域问答任务中展现出卓越性能。

Method	my		th		lo		km		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
GPT-3.5-turbo										
Base	25.5	32.7	49.5	63.7	28.0	41.4	35.5	46.1	34.6	46.0
CoT	25.0	32.6	47.5	62.5	24.0	37.0	34.0	43.3	32.6	43.9
CLP	24.0	31.7	44.0	55.3	23.5	34.3	29.5	38.8	30.3	40.0
LDP	26.0	34.7	48.5	61.3	30.0	42.6	35.5	45.9	35.0	46.1
LLM Rerank	27.0	35.9	49.0	61.5	30.5	43.7	37.5	48.2	36.0	47.3
Self-verification	26.0	33.0	46.5	58.7	25.5	38.3	32.5	44.9	32.7	43.7
Sure	26.5	34.3	48.5	61.8	27.5	40.7	34.5	47.5	34.3	46.1
Ours	**30.5**	**37.6**	**53.5**	**66.0**	**37.5**	**50.8**	**39.0**	**52.4**	**40.1**	**51.7**
GPT-4o-mini										
Base	51.5	68.6	62.0	75.6	46.5	67.9	54.5	71.2	53.6	71.3
CoT	52.0	69.5	61.0	74.1	44.0	66.3	52.0	68.7	52.3	70.2
CLP	48.5	65.2	58.5	73.9	43.0	65.2	47.5	65.7	49.4	67.5
LDP	53.0	70.8	62.0	74.9	52.0	73.3	56.0	72.1	55.8	72.8
LLM Rerank	52.5	70.7	59.5	72.9	47.5	69.4	56.5	73.1	54.0	71.5
Self-verification	50.5	68.1	61.0	75.1	47.5	68.8	54.0	71.3	53.3	70.8
Sure	55.0	72.1	62.5	75.3	48.0	69.7	55.5	72.6	55.3	72.4
Ours	**59.0**	**73.9**	**68.5**	**80.7**	**50.5**	**70.9**	**57.0**	**78.0**	**58.8**	**75.9**
DeepSeek-v3										
Base	57.5	73.7	62.5	74.9	50.0	72.1	51.5	66.4	55.4	71.8
CoT	61.0	74.9	62.0	75.5	53.0	73.6	57.0	76.0	58.3	75.0
CLP	62.0	77.2	65.5	78.2	53.0	75.2	56.5	74.0	59.3	76.2
LDP	60.0	75.5	66.0	79.2	55.5	76.1	57.0	75.2	59.6	76.5
LLM Rerank	59.5	75.4	65.0	77.2	54.0	74.7	52.5	68.3	57.8	73.9
Self-verification	58.5	74.1	64.0	76.9	51.0	73.2	53.5	70.1	56.8	73.6
Sure	61.0	76.4	66.0	79.3	53.5	75.0	57.5	77.1	59.5	77.0
Ours	**63.0**	**78.5**	**68.5**	**81.2**	**57.0**	**78.8**	**61.0**	**80.2**	**62.4**	**79.7**

表 2: 本文的方法与基线方法在不同模型上的效果对比实验结果

4.5 分析实验

4.5.1 生成候选答案方法分析

为了验证低资源语言开放域问答任务中生成多个候选答案策略的有效性，本文设计了一

系列实验，以评估该策略在提升正确答案覆盖率方面的效果。本文对比了三种答案生成方法：(1)直接生成答案方法，即直接提示模型根据文档信息生成与问题相关的单个答案。(2)多重采样生成策略，即在大语言模型解码过程中，通过设置超参数进行多重采样，并行生成多个答案。(3）提示生成策略，即在提示中添加指令，指导模型同时生成多个候选答案结果。在具体实验过程中，对于多重采样和提示生成策略，本文均指导模型生成2个候选结果，然后统计所有生成结果中包含正确答案的概率

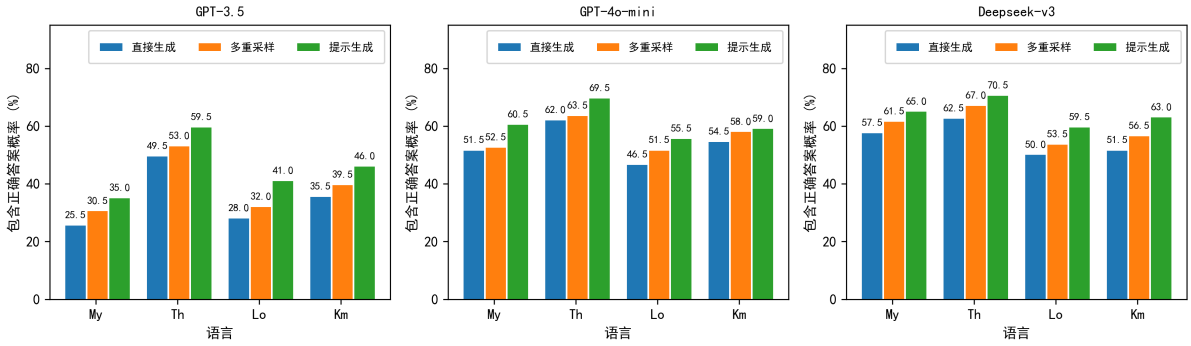

图 3: 不同方法生成候选答案结果对比

实验结果如图 3所示，在所有模型上，采用多重采样和提示生成多候选答案的方法，其包含正确答案的概率均高于直接生成单个答案的方法。其中，通过提示生成的多候选答案方法更能生成更丰富的候选答案结果。在低资源且高噪声的环境下，生成多个候选答案可以提升包含正确答案的概率，而且提示生成策略能更有效地利用大语言模型能力，生成更丰富的答案。

4.5.2 生成候选答案数量分析

为了深入探究候选答案数量N对实验结果的潜在影响，本文在提示生成候选答案的策略框架下，系统性地考察了不同候选答案数量N值对生成正确答案数量的影响。如图 4所示的实验结果表明，当候选答案数量N设置为2时，模型生成正确答案的数量呈现出较为显著的提升趋势。然而，随着N值的持续增加，生成正确答案的数量并未表现出预期的线性增长，并且出现了一些下降的情况，这一现象表明当前模型在生成候选答案时存在一定性能瓶颈，单纯依靠增加候选答案数量这一策略对于提升正确答案数量较为有限。

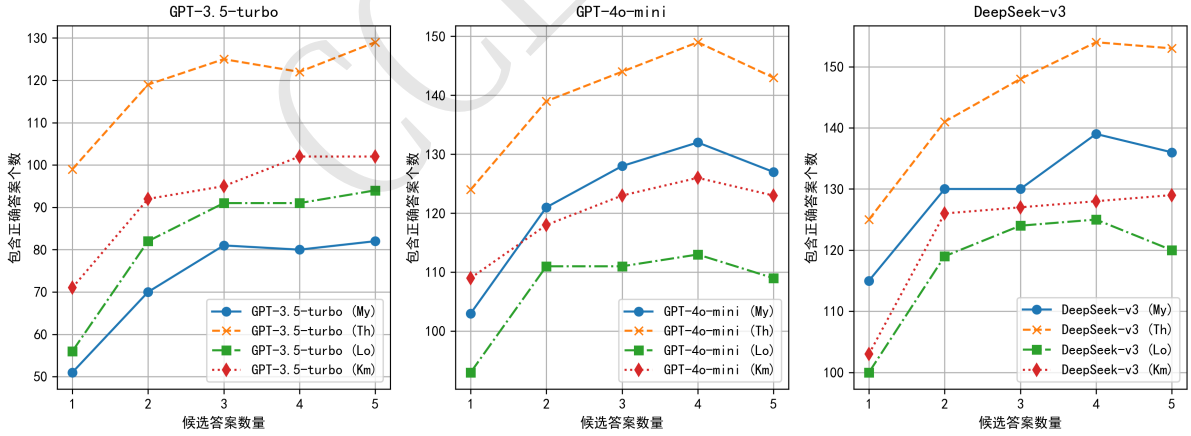

图 4: 不同候选答案个数包含正确答案数量对比

此外，本文在DeepSeek-v3 模型下，比较N=3时的完整实验结果（见表 3）进一步表明，相较于N=2，模型性能未出现显著提升。分析表明，增加候选答案数量N虽能小幅提高正确答案的命中率，但也使多维度筛选的过程变得更为复杂，且计算成本随之增加。因此，综合考虑准确性与计算成本，在本文提出的方法中，生成两个候选答案并进行筛选是一种更为合理且高效的策略。

Methods	my		th		lo		km		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Base	57.5	73.7	62.5	74.9	50.0	75.1	51.5	66.4	55.4	72.5
Ours(N=2)	63.0	78.5	68.5	81.2	57.0	78.8	61.0	80.2	62.4	79.7
Ours(N=3)	61.0	77.4	68.5	80.9	56.5	77.8	59.0	76.7	61.3	78.2

表 3: 不同候选答案数量的实验结果

4.5.3 噪声鲁棒性分析

在主实验中，本文通过控制相关文档与噪声文档的数量，将噪声比例设定为0.8，旨在探讨在极端噪声环境下大语言模型在开放域问答任务中的表现。此外，本文在DeepSeek-V3 模型上，通过调整相关文档与噪声文档的比例，进一步分析了不同噪声水平下大模型直接回答能力的变化，以及所提出方法在不同噪声干扰下的鲁棒性。

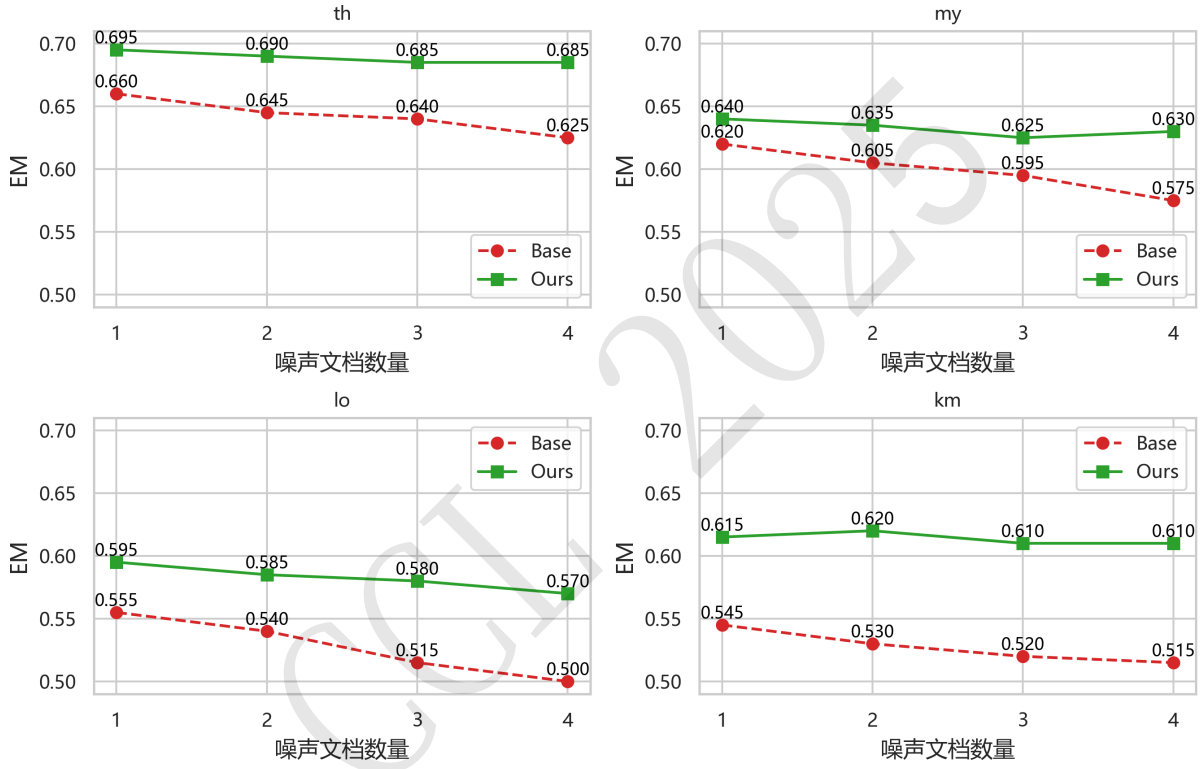

图 5: 不同噪声情况下模型鲁棒性对比

实验结果如图 5所示，随着噪声文档数量的增加，大模型直接生成答案的方式在所有语言中均表现出明显的性能下降，噪声敏感性较高，尤其在老挝语语任务中，EM值下降幅度达到5.5%。相比之下，本文所提出的方法在噪声干扰下展现出更强的鲁棒性，在各项语言任务中均优于直接问答策略，且其性能随着噪声增加呈现出更加平缓的下降趋势。上述结果充分说明本文方法在低资源语言环境下具备更好的泛化能力和抗干扰能力。

4.5.4 消融分析实验

为了进一步验证本文提出方法的有效性，本文设计了全面的消融实验在四种低资源语言数据集上进行了不同模型的消融实验，以验证各组件对模型整体性能的影响。

消融实验结果如表 4所示，得出以下重要发现：(1)完整方法在所有测试语言和评估指标上均表现出最优性能，充分证明了所提低资源问答方法的有效性。实验数据表明，移除任一功能模块都会显著影响模型性能，验证了各组件在系统中的必要性。(2)移除英语问题描述desen显

著影响模型性能，说明其作为跨语言理解桥梁对低资源语言问答任务具有重要价值。在三个评分模块中，全局篇章验证分数score1影响相对较小，但其全局验证作用仍不可或缺；局部证据验证分数score2的移除影响较为显著，印证了细粒度验证对答案准确性的关键作用；而不同答案相关性排序score3缺失导致最大性能下降，表明不同答案对比和排序机制对最终答案选择至关重要，能有效过滤错误答案并提升结果可靠性。(3)不同能力水平的模型对各组件的敏感度存在显著差异。性能相对较弱的GPT-3.5-turbo和GPT-4o-mini模型对组件移除表现出较高的敏感性，这表明其性能表现更依赖于多维度协助机制的协同作用。值得注意的是，即便是性能最优的Deepseek-v3模型，在移除各组件后也表现出明显性能衰减，验证了各功能模块对模型性能的提升作用具有普遍适用性。实验分析表明，本文提出的方法中各个组件都具有不可或缺的重要作用，其有效性得到了充分验证。

Model	my		th		lo		km		Average	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
GPT-3.5-turbo										
w/o desen	29.5	36.9	52.5	64.8	37.0	50.1	38.0	50.2	39.3	50.5
w/o score1	29.0	36.2	52.5	64.7	37.0	50.0	38.0	50.9	39.1	50.5
w/o score2	29.5	36.3	51.5	64.0	35.5	48.4	38.0	50.9	38.6	49.9
w/o score3	28.5	35.7	51.0	63.2	35.0	47.9	37.5	50.1	38.0	49.2
Full model	**30.5**	**37.6**	**53.5**	**66.0**	**37.5**	**50.8**	**39.0**	**52.4**	**40.1**	**51.7**
GPT-4o-mini										
w/o desen	57.5	72.5	67.5	79.9	48.0	68.2	56.0	75.9	57.3	74.1
w/o score1	58.0	73.0	68.0	80.4	48.5	69.6	56.0	76.6	57.6	74.9
w/o score2	57.5	72.3	67.0	79.4	47.5	68.9	55.5	75.3	56.9	74.0
w/o score3	56.5	71.8	67.5	79.7	47.5	69.2	55.0	75.0	56.6	73.9
Full model	**59.0**	**73.9**	**68.5**	**80.7**	**50.5**	**70.9**	**57.0**	**78.0**	**58.8**	**75.9**
Deepseek-v3										
w/o desen	62.5	77.9	68.0	80.7	56.0	77.0	59.5	77.3	61.5	78.2
w/o score1	62.0	77.6	67.5	80.7	56.5	77.3	60.0	77.6	61.5	78.3
w/o score2	62.0	77.8	66.5	79.8	56.5	78.2	59.5	77.1	61.1	78.2
w/o score3	61.5	77.3	66.0	79.3	55.5	77.3	59.0	77.5	60.5	77.9
Full model	**63.0**	**78.5**	**68.5**	**81.2**	**57.0**	**78.8**	**61.0**	**80.2**	**62.4**	**79.7**

表 4: 消融实验结果。其中w/o desen代表移除英语问题描述，w/o score1代表移除全局篇章验证分数score1，w/o score2代表移除局部证据验证分数score2，w/o score3代表移除不同答案相关性排序分数score3，Full model代表本文提出的完整方法。

4.6 案例分析

图 6展示了LRQA数据集中的三个示例。可以观察到：（1）在候选答案中包含正确答案的情况下，本文提出的完整模型能够准确地从中筛选出正确答案，体现了方法的有效性。（2）当移除全局篇章验证分数score1后，模型难以对存在概念混淆的候选项进行有效判别。例如，在案例1中，“天问一号”指的是火星探测任务，而“祝融号”才是具体的火星探测器。缺乏全局信息的验证使得模型无法正确理解文本中的实体关系，从而影响答案筛选的准确性。（3）移除局部证据验证分数score2后，模型在处理如案例3中同类、相似属性的答案时，缺乏辨别力。该现象验证了细粒度证据信息的重要性：局部证据验证模块能够提供更为精细的证据对比，从而帮助模型筛选出更符合查询意图的答案。（4）移除不同答案相关性排序分数score3后，模型在案例1和案例2中的答案筛选结果均出现错误。不同答案相关性排序模块基于文档信息对各候选答案进行相互对比和验证，能够筛选出与查询更相关的答案，同时剔除内容冗余或信息缺失的选项，从而选出更完整、准确的答案。案例分析进一步揭示了各个维度答案筛选模块在模型中的

作用机制。实验结果表明，只有各模块协同工作，才能实现对复杂低资源语言问答场景中最优答案的准确提取，有效验证了本文所提出方法在多维验证机制下的性能与鲁棒性。

	案例1：	案例2：	案例3：
文档信息 和 输入查询	Passage1： ...លោក Yuan Foyu ដែលជាប្រធានទី 38 នៃយុទ្ធនាការដាក់ឈ្មោះសម្រាប់យានរុករកភពអង្គារដំបូងរបស់ប្រទេសចិនបានប្រាប់ទូរទស្សន៍ CCTV ថាបេសកកម្ម Tianwen-1 និងឱកាសដាក់ឈ្មោះរបស់វាបានទាក់ទាញការចាប់អារម្មណ៍យ៉ាងទូលំទូលាយ។ ការបោះឆ្នោតនេះត្រូវបានរៀបចំរួមគ្នាដោយក្រុមហ៊ុនចិន Baidu និង Gaidu Technology ហើយថ្ងៃផុតកំណត់គឺថ្ងៃទី 28 ខែឧសភា ឆ្នាំ 2020។... （...中国首辆火星探测器命名活动第38届主席袁佛宇在接受中央电视台采访时表示，天问一号任务及其命名机遇备受关注。此次投票由中国企业百度和盖度科技联合组织，截止日期为2020年5月28日...） ...... Pasage N: ...Zhurong គឺជាផ្នែកមួយនៃបេសកកម្ម Tianwen-1 Mars របស់ប្រទេសចិន ហើយបានចុះចតនៅលើ Utopia Planitia ក្នុងខែឧសភា ឆ្នាំ 2021។ ការស៊ើបអង្កេតនេះបានចូលទៅក្នុងរបៀប hibernation នៅក្នុងខែឧសភា ឆ្នាំ 2022 ដែលអនុញ្ញាតឱ្យវាសម្រាកក្នុងអំឡុងពេលរដូវរងានៃអឌ្ឍគោលខាងជើង Martian ។... （...祝融号是中国天问一号火星任务的一部分，于 2021 年 5 月降落在乌托邦平原。该探测器于 2022 年 5 月进入休眠模式，使其能够在火星北半球的冬季休眠...） Query: តើយានរុករករបស់ចិននៅលើភពអង្គារមានឈ្មោះអ្វី ? （中国火星探测器叫什么名字？）	Passage1： ...God of War Ragnarök ເປັນເກມຜະຈົນໄພຜະຈົນໄພທີ່ພັດທະນາ ໂດຍ Santa Monica Studio ແລະຈັດພິມ ໂດຍ Sony Interactive Entertainment. ມັນໄດ້ຖືກປ່ອຍອອກມາໃນທົ່ວໂລກໃນວັນທີ 9 ພະຈິກ 2022, ສໍາລັບທັງ PlayStation 4 ແລະ PlayStation 5, ເຊິ່ງເປັນການເປີດຕົວແບບຂ້າມ Gen ຄັ້ງທໍາອິດໃນຊຸດ God of War. ມັນເປັນພາກທີ 9 ໃນຊຸດ, ລໍາດັບທີ 9, ແລະເປັນພາກຕໍ່ຂອງ God of War ຂອງປີ 2018... （...《战神：诸神黄昏》是一款由圣莫尼卡工作室开发、索尼互动娱乐发行的角色扮演冒险游戏。游戏于2022年11月9日全球发售，登陆PlayStation 4和PlayStation 5平台，是《战神》系列的首次跨世代发行。本作是该系列的第九部作品，也是2018年《战神》的续作...） ...... Pasage N: ...ພຣະເຈົ້າແຫ່ງສົງຄາມ: Ragnarok ມີວັນປ່ອຍ. ພຣະເຈົ້າແຫ່ງສົງຄາມ: Ragnarok ໄດ້ຖືກປະກາດຢ່າງເປັນທາງການ - ຊື່ຂອງເກມແມ່ນມີຄວາມລຶກລັບມາດົນນານ, ພະຍາຍາມຊັກຊວນ fans ວ່າ Ragnarok ຈະເປັນການແຂ່ງຂັນສຸດທ້າຍສໍາລັບຫົວຂໍ້ສຸດທ້າຍ... （...《战神：诸神黄昏》的发布日期已经确定。《战神：诸神黄昏》正式官宣——长期以来，这款游戏的名字一直难以猜测，试图让粉丝们相信，诸神黄昏将是争夺最终称号的最终决战...） Query: ມື້ວາງຈໍາໜ່າຍຂອງ God of War Ragnarok ແມ່ນຫຍັງ? （战神：诸神黄昏的发布日期是什么时候？）	Passage1： ...Surface Laptop 3 သည် သင့်လုပ်ငန်းကို အကောင်းဆုံးအသုံးချနိုင်စေမည့် အစွမ်းထက်ပြီး စတိုင်ကျသော လက်ပ်တော့တစ်လုံးဖြစ်သည်။ Surface Laptop 3 13.5" Sandstone နှင့် Matte Black - 1,288 g (2.84 ပေါင်)... （...Surface Laptop 3 性能强劲，外观时尚，助您轻松完成所有工作。Surface Laptop 3 13.5 英寸砂岩色和哑光黑色：1,288 克（2.84 磅）...） ...... Pasage N: ...Surface Laptop SE သည် ချစ်စရာ လက်တော့ပ်တစ်ခုဖြစ်သည်။ ၎င်းသည် ကုမ္ပဏီ၏ အထင်ကရ Surface Laptop လိုင်း၏အသေးစားဗားရှင်းတစ်ခုဖြစ်ပြီး အဖုံးနှင့်အောက်ခြေဘေးဘောင်တွင် မှတ်မိနိုင်သောလိုဂိုပါရှိသည်။ အလေးချိန် 2.45 ပေါင်သာရှိ၍ အလွန်ပေါ့ပါးသည်။ အပြင်မှာ ရှုပ်ပွနေတဲ့ ကျောင်းမော်ဒယ်များစွာနဲ့ မတူတာက ဒီကိရိယာဟာ စားသုံးသူနေရာက အရွယ်ရောက်ပြီးသူတွေကို အလွယ်တကူ မြင်နိုင်တဲ့ ကိရိယာတစ်ခုပါ။ ဒီလောက်မနှေးဘူးဆိုရင်။ ... （...Surface Laptop SE 是一款可爱的笔记本电脑。它是微软旗舰 Surface Laptop 系列的缩小版，盖子和底部边框上印有醒目的标志。它的重量仅为 2.45 磅，非常轻巧。与市面上许多笨重的学校机型不同，这款设备很容易被消费领域的成年人看到。要是它速度不那么慢就好了...） Query: Surface Laptop SE ရဲ့ အလေးချိန်က ဘယ်လောက်လဲ။ （Surface Laptop SE 的重量是多少？）
真实答案	Zhurong （祝融号）	ວັນທີ 9 ພະຈິກ 2022（2022年11月9日）	2.45 ပေါင် （2.45磅）
候选答案	(a) Zhurong （祝融号） (b) Tianwen-1 （天问1号）	(a) ວັນທີ 9 ພະຈິກ 2022（2022年11月9日） (b) ວັນທີ 9 ພະຈິກ （11月9日）	(a) 2.45 ပေါင် （2.45磅） (b) 2.84 ပေါင် （2.84磅）
w/o score1	Tianwen-1 ✗	ວັນທີ 9 ພະຈິກ 2022 ✓	2.45 ပေါင် ✓
w/o score2	Zhurong ✓	ວັນທີ 9 ພະຈິກ 2022 ✓	2.84 ပေါင် ✗
w/o score3	Tianwen-1 ✗	ວັນທີ 9 ພະຈິກ ✗	2.45 ပေါင် ✓
Full model	Zhurong ✓	ວັນທີ 9 ພະຈິກ 2022 ✓	2.45 ပေါင် ✓

图 6: LRQA数据集上的几个案例研究

5 结论

本文提出了一种基于多维度答案筛选的低资源语言开放域问答方法，通过候选答案生成并筛选的多阶段协作过程提升了大模型在低资源语言问答任务的鲁棒性。在答案生成阶段，从文档中抽取多样化候选答案，在筛选阶段，设计多维度答案筛选策略，通过全局篇章答案验证、局部证据答案验证以及不同答案相关性排序，筛选出最优答案，以此提升大语言模型在低资源语言噪声环境下的问答性能。与现有方法相比，在四种东南亚低资源语言的开放域问答任务中取得了具有竞争力的结果，充分验证了该方法的有效性和优越性。进一步的分析实验表明，各组成模块对性能提升起到了积极作用，验证了该方法在低资源开放域问答场景中的适用性。未来研究将聚焦于从低资源文档中挖掘更丰富的有效候选答案信息，以进一步提升低资源开放域问答的性能。

参考文献

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. Sure: Summarizing retrievals using answer candidates for open-domain QA of LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is chatGPT good at search? investigating large language models as re-ranking agents. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762, 2024.

Danqi Chen and Wen-tau Yih. Open-domain question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 34–37, 2020.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019.

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

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. DeepSeekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024.

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

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore, December 2023. Association for Computational Linguistics.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1504–1518, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wenhu Chen, Pat Verga, Michiel de Jong, John Wieting, and William W. Cohen. Augmenting pre-trained language models with QA-memory for open-domain question answering. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1597–1610, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

Hao Cheng, Yelong Shen, Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. UnitedQA: A hybrid approach for open domain question answering. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3080–3090, Online, August 2021. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics.

Minsang Kim, Cheoneum Park, and Seung Jun Baek. QPaug: Question and passage augmentation for open-domain question answering of LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9024–9042, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Angeliki Lazaridou, Elena Gribovskaya, Wojciech Stokowiec, and Nikolai Grigorev. Internet-augmented language models through few-shot prompting for open-domain question answering. *arXiv preprint arXiv:2203.05115*, 2022.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3501–3516, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

A 提示词模版

Prompt 1 候选答案生成提示p_{ans}

You are a professional {language} AI assistant. Below are {language} N passages related to the question at the end. After reading the passages, provide two correct and complete candidates for the answer to the question at the end. The answer should be in the form: (a) xx, (b) yy. Each candidate answer should not exceed 3 words.

Passage # 1 text: {passage1 text}
...
Passage # N text: {passageN text}

Question: {question}

Answer:

Prompt 2 英语问题描述提示p_{en}

You are a professional multilingual AI assistant. Please analyze the given {language} question using English, briefly identifying the type or format of answer it requires. Do not answer the question directly.

Question: {question}

Analyze:

Prompt 3 全局验证提示p_{glob}

Passage # 1 text: {passage1 text}
...
Passage # N text: {passageN text}

Question: {question}

Question Description: {question des}

Prediction:{Candidate Answer}

Analyze the provided answer to determine its accuracy, reasonableness, and correctness. Provide a concise evaluation first, then conclude with either "True" or "False" based on your assessment.

Prompt 4 证据抽取提示p_{evi}

Passage # 1 text: {passage1 text}
...
Passage # N text: {passageN text}

Question: {question}

Prediction: {Candidate Answer}

Your task is to extract content related to the source of the predicted answer directly from the provided {language} passages. Do not generate new content; only extract existing content from the paragraph. Make sure the extracted content fully supports the given prediction. Preserve the original sentence structure as much as possible. When you have completed the task, write [DONE] to indicate that the task is completed.

Prompt 5 局部验证提示p_{part}

Question: {question}

Question Description: {question description}

Evidence: {evidence}

Does the evidence provide useful information to answer the question? Please explain your reasoning and provide your judgment (True or False).

Prompt 6 答案对比提示p_{comp}

Passage # 1 text: {passage1 text}
...
Passage # N text: {passageN text}

Question: {question}

Candidate Answers:(Answer1:{Candidate Answers1} Answer2:{Candidate Answers2})

Based on the information provided in the passage, determine whether [Answer 1] or [Answer 2] is the most suitable answer to the question. Respond only with "Answer1" or "Answer2".