arXiv:2602.01246v1 [cs.CL] 1 Feb 2026

# Parse：面向波斯语的开放域推理问答基准

Jamshid Mozafari*因斯布鲁克大学 奥地利因斯布鲁克
ja
mshid.mozafari@uibk.ac.at

Seyed Parsa Mousavinasab*因斯布鲁克大学 奥地利因斯布鲁克
seyed.
mousavinasab@student.uibk.ac.at

Adam Jatowt 因斯布鲁克大学 奥地利因斯布鲁克
adam.jatowt@uibk.ac.at

## 摘要

随着大语言模型（LLM）的发展，以推理为重点的问答（QA）取得了快速进展，但面向低资源语言的高质量基准仍然稀缺。波斯语约有 1.3 亿人使用，却缺乏用于评估具备推理能力的问答系统的综合性开放域资源。我们推出 Parse，这是首个开放领域波斯语推理问答基准，包含 10,800 个问题，涵盖布尔型、多项选择和事实型格式，并涉及多种推理类型、难度级别和答案结构。该基准通过受控的基于大语言模型的生成流程构建，并经人工评估验证。我们还通过多阶段筛选、标注和一致性检查，确保语言质量和事实质量。我们在多种提示策略下对多语言和波斯语大语言模型进行基准测试，并发现波斯语提示和结构化提示（布尔型／多项选择题采用思维链，事实型问题采用少样本提示）能够提升性能。微调进一步提升了结果，尤其是对专门针对波斯语的模型而言。这些发现凸显了 Parse 如何支持公平比较和实用的模型适配。Parse 填补了波斯语问答研究中的关键空白，并为在低资源环境中开发和评估具备推理能力的大语言模型奠定了坚实基础。

## CCS概念

• **信息系统** → **问答**；**测试集**；**检索模型与排序**；**检索任务与目标**；**检索结果评估**。

## 关键词

问答、基准、波斯语、推理、多跳、布尔型、多项选择、事实型

**ACM参考格式：**

Jamshid Mozafari、Seyed Parsa Mousavinasab 和 Adam Jatowt。2026。Parse：面向波斯语的开放域推理问答基准。载于第*49*届 *ACM* 信息检索研究与发展国际会议论文集（*SIGIR '26*）。ACM，纽约，纽约州，美国，8页。https://doi.org/XXXXXXX.XXXXXXX

*贡献相同。

制作本作品全部或部分内容的数字或纸质副本，供个人或仅供课堂使用，无需付费，但须满足以下条件：不得为牟利或商业利益而复制或分发副本，且副本首页必须附有本声明和完整引文。必须尊重本作品中由作者以外的其他人拥有版权的组件。注明出处后可以摘录。若要以其他方式复制、重新发表、发布到服务器或向列表重新分发，须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。*SIGIR '26*，澳大利亚墨尔本© 2026 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/XXXXXXX.XXXXXXX

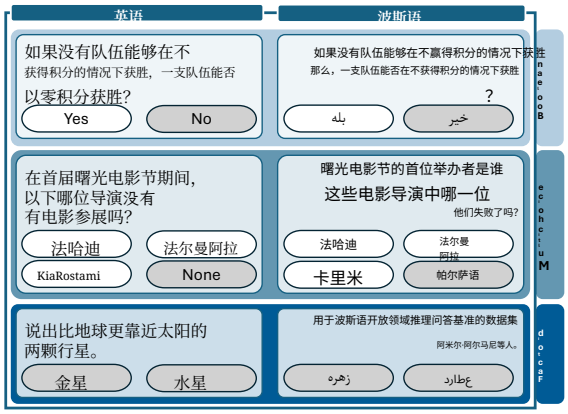


**图1：来自 Parse 的示例问题。英语列提供了对应波斯语问题的译文。灰色高亮选项表示正确答案。**

## 1 引言

问答（QA）旨在为用户查询提供准确的回答 [26]。随着大语言模型（LLM）问世 [10, 34, 38]，，近期的问答系统已不再局限于传统的抽取式设置，开始处理需要多步推理的复杂问题 [27]。这类推理问题通常需要逻辑推理、概念关联、语境理解和得出结论，而非简单的模式匹配或直接从文本中提取 [28]。

大语言模型极大地重塑了自然语言处理（NLP） [43] 和信息检索（IR） [42]的格局。它们的出现既提高了人们对现有任务的性能预期，也开辟了此前被认为难以实现的新研究方向。值得关注的主题包括检索增强生成（RAG） [23]，、幻觉检测 [12]，、超越准确率的评估 [5]，、指令调优 [32]，、多步推理 [27]，以及基于智能体的系统 [24]。

在这些方向中，推理型问答 [18] 受到了广泛关注。先前的工作探索了多种方法 [20, 36] 和数据集 [19, 41]，主要针对英语，近期也开始涉及中文 [37, 40]等语言。不断增长的研究成果表明，下一代问答系统必须具备强大的推理能力，而大语言模型回答推理问题的能力是衡量其整体能力的重要指标 [29]。

然而，低资源语言的推理问答研究仍然有限，主要是因为可用数据集和基准测试集稀缺。波斯语就是一个典型例子，这是一种拥有

**表1：英语（EN）和波斯语（FA）问答数据集在规模、发布年份、难度标签和问答任务类型方面的比较，包括布尔型、多项选择、事实型、多跳、推理、多答案和不可回答问题。**

| 数据集 | Lang | Size | Year | 难度标签 | Bool | MCQ | 事实型 | 多跳 | 推理 | 多答案 | 非答案 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RACE [22] | EN | 100k | 2017 | | | ✓ | ✓ | | ✓ | | |
| SQuAD2.0 [30] | EN | 150k | 2018 | | | | ✓ | | | | ✓ |
| HotpotQA [39] | EN | 113k | 2018 | | ✓ | | ✓ | ✓ | ✓ | | |
| MultiRC [19] | EN | 10k | 2018 | | | ✓ | ✓ | ✓ | ✓ | ✓ | |
| NaturalQuestions [21] | EN | 307k | 2019 | | ✓ | | ✓ | | | | ✓ |
| PerCQA [13] | FA | 1k | 2021 | | | | ✓ | | | ✓ | |
| PersianQA [6] | FA | 9k | 2021 | | | | ✓ | | | ✓ | ✓ |
| PersianQuAD [16] | FA | 19.6k | 2022 | | | | ✓ | | | | |
| PQuAD [8] | FA | 80k | 2023 | | | | ✓ | | | | ✓ |
| IslamicPCQA [9] | FA | 12.3k | 2023 | | ✓ | | ✓ | ✓ | | | |
| PersianMedQA [15] | FA | 20.8k | 2025 | | | ✓ | ✓ | | ✓ | | |
| PersianMHQA [33] | FA | 7k | 2025 | | ✓ | | ✓ | ✓ | | | |
| **Parse（我们提出的）** | **FA** | **10.8k** | **2025** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

悠久文化历史的语言， [35] 全球约有 1.3 亿使用者[1]。据我们所知，目前还没有专门用于评估波斯语推理问答的开放领域基准。

为弥补这一空缺，我们推出了 Parse[2]，这是首个面向波斯语、涵盖通用知识和跨主题知识的开放域推理问答基准。Parse 包含 10,800 个问题，涵盖多种格式——包括布尔型、多项选择和事实型——以及多种推理类别，例如多跳问题和复杂推理问题，其中包括单答案、多答案和无法作答的情况。如此广泛的覆盖范围使 Parse 成为评估大语言模型在波斯语问答中推理表现的综合资源。此外，鉴于高质量波斯语大语言模型十分稀缺， [3, 31], Parse 能够支持系统化的基准测试与比较。图1展示了该基准中的代表性示例。

总而言之，我们的贡献如下：

（1）我们推出了 Parse，这是首个波斯语开放域推理问答基准，包含 10,800 个涵盖多种问题和答案类型的多样化问题。（2）我们开展了两项人工评估研究，以验证基准质量。（3）我们使用多语言和波斯语大语言模型进行了全面实验，展示 Parse 的实用价值，并分析微调的影响。

## 2 相关工作

英语问答（QA）领域涌现出众多大规模基准测试集，涵盖多种任务类型和推理挑战。SQuAD2.0 [30] 引入了超过 15 万个抽取式问题，其中包含用于评估弃答能力的不可回答情形。NaturalQuestions（NQ） [21]包含超过 30 万条来自谷歌真实用户的查询，支持布尔型和事实型问答，以及不可回答问题。RACE [22] 提供了 10 万道选自英语考试的选择题，旨在考查推理能力。HotpotQA [39]包含 11.3 万个多跳问题，需要对多篇文档进行推理，而 MultiRC [19]侧重于多句阅读理解，包含多答案问题并明确要求进行推理。总体而言，这些英语基准测试集涵盖了全面的问答任务类型，包括布尔型、选择题、事实型、多跳、推理、多答案以及不可回答问题。

相比之下，波斯语问答资源的覆盖范围仍然狭窄，每个现有数据集都只涵盖问答任务维度中的一部分。PersianQuAD [16] 和 PersianQA [6] 提供事实型抽取式问答，而 PerCQA [13] 则侧重于采用多答案监督的社区问答。PQuAD [8] 将事实型问答扩展至 8 万个样例，并引入了不可回答问题，但不包含多跳或推理问题。更专业的研究则针对特定领域或推理方式。PersianMedQA [15]评估医学推理，而 IslamicPCQA [9] 和 Persian- MHQA [33] 则侧重于多跳问答。然而，这些基准测试集仍受限于领域（如医学）或任务类型（如仅多跳）。它们都无法共同支持英语基准测试集中观察到的广泛问答现象。表1总结了波斯语数据集之间的主要差异。

据我们所知，现有的 PersianQA 基准均未能全面覆盖开放领域中的布尔型、多项选择、事实型、多跳、推理、多答案和不可回答问题，且涉及的是一般性、跨主题知识，而非单一的专业领域。我们推出了 Parse，这是首个涵盖上述完整范围的开放领域波斯语推理问答基准。通过将主要问答任务类型整合到同一资源中，Parse 为波斯语建立了更具挑战性的评测设置，使人们能够更广泛地研究开放领域推理，并缩小其与成熟英语问答基准之间的差距。

## 3 Parse 基准

本节介绍 Parse 的构建过程，涵盖从提示设计到生成、组装和质量控制的各个环节。

任务设计与分类体系。 我们首先定义了三种主要问答类型——布尔型、多项选择和事实型——并将每种

[1]https://en.wikipedia.org/wiki/Persian_语言 [2] https://github.com/DataScienceUIBK/Parse

**表2：Parse 基准按问题类型、推理维度和子类型进行的分类，以及问题数量。**

| 布尔型 | | | | 多项选择 | | | | 事实型 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 多跳 | | 推理 | | 多跳 | | 推理 | | 多跳 | | 推理 | |
| 简单 | 600 | 简单 | 600 | 单答案 | 600 | 单答案 | 600 | 简单 | 600 | 简单 | 600 |
| 否定 | 600 | 否定 | 600 | 多答案 | 600 | 多答案 | 600 | 基于列表 | 600 | 基于列表 | 600 |
| 比较 | 600 | 比较 | 600 | 非答案 | 600 | 非答案 | 600 | 非答案 | 600 | 非答案 | 600 |
| **总计** | **1800** | **总计** | **1800** | **总计** | **1800** | **总计** | **1800** | **总计** | **1800** | **总计** | **1800** |

与两个正交的方法维度配对：多跳和推理。在此基础上，我们实例化了反映常见评估需求的内容子类型：布尔型包括简单、否定和比较；多项选择包括单答案、多答案和不可回答；事实型包括简单、基于列表和不可回答。每个问题还会额外获得一个难度标签（简单 /中等 /困难），以支持受控评估。这一交叉组合为每种问答类型产生6种配置[3]，总体共18种配置，广泛覆盖形式、推理要求和答案结构。表2列出了用于指导生成的分类体系。

提示词流程。 我们采用由大语言模型驱动的生成策略，使用 GPT-4o [4], 生成问题，随后对每个问题进行严格的人工核验和筛选。我们没有采用单一的庞大提示词，而是设计了针对特定配置的 提示词，其中明确规定了：(i) 预期的问答类型和子类型；(ii) 题目是否需要多跳推理和／或一般推理；(iii) 均衡的难度安排。每个提示词都遵循统一模板，包含四个部分：简洁的角色 描述，用于引导模型的行为；任务 区块，用于指定目标配置；一组要求，用于落实各项约束（例如答案格式、多项选择题的选项数量和排序、使用波斯语、真实性和主题多样性）；以及便于后续处理的简明*CSV* 格式输出 说明。这种设计使约束明确，同时保持提示词简短且可复现。所有提示词模板均已公开发布在我们的 GitHub 仓库中[4]。

对于每个配置类别，我们都专门设计了一条提示词并且每次运行生成一批 30 个问题（简单、中等、困难各 10 个）[5]。我们根据需要重复使用相同的提示，以达到预期的样本量[6]。在整个开发过程中，我们通过抽查监控生成质量，特别关注波斯语的形态学和句法、否定和比较结构的准确表达、多项选择题中答案选项的正确且均衡的排列，以及是否使用真正的多跳证据链，而非单跳回忆。

组装与规范化。 所有运行结果均导出为 CSV，并按配置进行合并。随后，我们将数据转换为规范化的 JSON模式，以便进行程序化验证。每个问题都会获得一个全局唯一的标识符，以及特定于配置的字段：问题；答案（对于单答案类型，为字符串；对于基于列表或多答案类型，为列表；对于是/否问题，为布尔型）；多项选择题的选项（恰好四个）；以及难度（简单、中等或困难）。

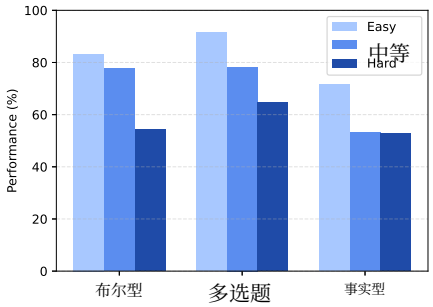


**图2：不同问题类型在各难度级别（简单、中等、困难）下的人工评估准确率。**

**表3：四个评估组在歧义性、可读性和正确性方面的平均人工评分（1–5）。**

| 组别 | 歧义性 | 可读性 | 正确性 |
|---|---|---|---|
| 第1组 | 4.184 | 4.504 | 4.307 |
| 第2组 | 4.562 | 4.867 | 4.295 |
| 第3组 | 4.53 | 4.896 | 4.546 |
| 第4组 | 4.341 | 4.412 | 4.311 |
| 平均值 | **4.404** | **4.669** | **4.389** |

质量控制与平衡。我们进行了多轮质量控制。首先，确保结构有效：不得缺少字段（事实型不可回答题除外，此类题使用惯例占位符“None”）、多项选择题的选项数量正确，且难度标签有效。其次，删除了同一批次内及不同批次间完全重复的题目，并确保多项选择题的选项不重复。第三，检查了各配置特有的语义要求：布尔题项必须能够用“是/否”（对应的波斯语表达）作答，并且与命题一致；多项选择单答案题必须恰有一个有效答案，且绝不能是“None”；多答案题必须有两个至四个有效答案；而不可回答题必须仅将“None”设为正确选项。事实型简单题和不可回答题必须恰好包含一个答案字符串（其中不可回答题的答案须设为“None”），而基于列表题必须包含两至五个标准答案。

[3] 2 方法 × 3 = 6 子类型
[4] https://github.com/DataScienceUIBK/Parse [5] 我们选择每批 30 个问题，因为更大的批次会降低输出质量。[6] 我们将每个子类型的样本量设为 600。

**表 4：英语和波斯语设置下，零样本、少样本和思维链提示的布尔问答性能。灰色单元格标示每种模型/子类型中英语与波斯语表现较高的一方。下划线表示每种模型/语言/子类型下表现最强的提示策略（不包括并列情况）。**

| 模型 | 英语 | | | | | | 波斯语 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 多跳 | | | 推理 | | | 多跳 | | | 推理 | | |
| | 比较 | 否定 | 简单 | 比较 | 否定 | 简单 | 比较 | 否定 | 简单 | 比较 | 否定 | 简单 |
| **零样本** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.49 | 0.57 | 0.78 | 0.71 | 0.83 | 0.82 | 0.53 | 0.63 | 0.79 | 0.77 | 0.82 | 0.83 |
| LLaMA 3 8B | 0.54 | 0.69 | 0.77 | 0.81 | 0.65 | 0.69 | 0.59 | 0.78 | 0.78 | 0.70 | 0.48 | 0.71 |
| Mistral 24B | 0.52 | 0.73 | 0.89 | 0.72 | 0.74 | 0.85 | 0.56 | 0.93 | 0.93 | 0.84 | 0.71 | 0.87 |
| Gemma 2 27B | 0.57 | 0.94 | 0.94 | 0.86 | 0.82 | 0.92 | 0.58 | 0.93 | 0.93 | 0.86 | 0.79 | 0.91 |
| LLaMA 3 70B | 0.58 | 0.95 | 0.94 | 0.89 | 0.82 | 0.92 | 0.57 | 0.96 | 0.95 | 0.89 | 0.82 | 0.92 |
| Qwen 2.5 72b | 0.57 | 0.86 | 0.95 | 0.87 | 0.82 | 0.95 | 0.57 | 0.88 | 0.95 | 0.87 | 0.81 | 0.94 |
| **少样本** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.58 | 0.64 | 0.84 | 0.75 | 0.81 | 0.84 | 0.60 | 0.85 | 0.86 | 0.81 | 0.67 | 0.81 |
| LLaMA 3 8B | 0.52 | 0.87 | 0.76 | 0.82 | 0.62 | 0.76 | 0.45 | 0.55 | 0.56 | 0.51 | 0.57 | 0.63 |
| Mistral 24B | 0.52 | 0.54 | 0.58 | 0.56 | 0.58 | 0.60 | 0.54 | 0.82 | 0.82 | 0.74 | 0.73 | 0.80 |
| Gemma 2 27B | 0.58 | 0.85 | 0.94 | 0.86 | 0.89 | 0.93 | 0.58 | 0.88 | 0.95 | 0.86 | 0.89 | 0.92 |
| LLaMA 3 70B | 0.58 | 0.95 | 0.94 | 0.89 | 0.90 | 0.94 | 0.58 | 0.96 | 0.94 | 0.87 | 0.89 | 0.94 |
| Qwen 2.5 72b | 0.58 | 0.89 | 0.96 | 0.87 | 0.90 | 0.95 | 0.57 | 0.94 | 0.97 | 0.88 | 0.90 | 0.95 |
| **思维链** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.57 | 0.63 | 0.84 | 0.82 | 0.84 | 0.89 | 0.56 | 0.62 | 0.81 | 0.83 | 0.84 | 0.86 |
| LLaMA 3 8B | 0.55 | 0.47 | 0.83 | 0.80 | 0.71 | 0.81 | 0.57 | 0.62 | 0.82 | 0.78 | 0.69 | 0.83 |
| Mistral 24B | 0.54 | 0.76 | 0.92 | 0.85 | 0.75 | 0.93 | 0.55 | 0.89 | 0.92 | 0.86 | 0.69 | 0.92 |
| Gemma 2 27B | 0.56 | 0.87 | 0.96 | 0.86 | 0.85 | 0.93 | 0.58 | 0.90 | 0.95 | 0.85 | 0.88 | 0.94 |
| LLaMA 3 70B | 0.57 | 0.81 | 0.94 | 0.88 | 0.85 | 0.94 | 0.57 | 0.80 | 0.95 | 0.87 | 0.87 | 0.94 |
| Qwen 2.5 72b | 0.57 | 0.80 | 0.95 | 0.87 | 0.83 | 0.93 | 0.58 | 0.82 | 0.95 | 0.89 | 0.84 | 0.96 |

最后，我们针对难度校准和语言鲁棒性进行了专项检查。我们确保困难题目的难度源自推理深度、组合性或细微的语义差异，而非繁复的语法或生僻词汇；不符合这一原则的题目均予以淘汰。我们还筛查了波斯语表达，确保其流畅且符合当地文化习惯，优先采用惯用表达，并剔除可能造成理解偏差的直译。这些原则被直接纳入提示设计，并通过事后筛选予以强化，以提升题目在人类评估和自动评估中的有效性。

验证后，我们获得了均匀覆盖：18 个配置类别中的每个类别都恰好包含 600 个问题，并按难度均匀划分（200 个简单、200 个中等、200 个困难），总计得到 **10,800** 个问答对。最终发布版本保留了表2所示的分类体系，既支持有针对性的消融实验，也支持对波斯语各种问答现象进行端到端评估。

## 4 人工评估

由于仅依赖大语言模型生成的输出可能导致系统性错误，我们开展了人工评估，以核验生成的问答对在语言和事实方面的质量，并确认所赋难度标签的准确性。我们进行了两项互补评估：质量评估 和难度评估。

质量评估。 我们在1–5分量表上评估了歧义性、可读性和正确性。从全部10,800个问答对（54种配置）中，我们为每种配置抽取5个条目，每组得到270个条目。四个互不重叠的组（1,080个条目共计，占基准的10%）被创建；每组由三名波斯语母语者独立评估（共12名标注员），从而实现多数一致。

标注员包括 8 名男性和 4 名女性，教育背景多样，涵盖本科在读（2 名男性、1 名女性）、学士（4 名男性、2 名女性）和硕士（2 名男性、1 名女性），年龄为 22 至 57 岁。标注员使用轻量级网页界面，并根据三个标准对每个问答对进行评分。标注界面可在我们的 GitHub 仓库中获取。

表3展示了所有标注员的平均得分。三项指标的平均得分均高于4，表明：（i）问题清晰明确、毫无歧义；（ii）波斯语语法流畅；（iii）事实正确性较高。

难度评估。 为验证难度标签（简单/中等/困难），三位标注员评估了270组问答对的一个子集。每位标注员在每种配置下回答五个问题，且未被告知原始难度标签。他们只获知问题类型，而不知道子类型，以避免潜在偏见，并更公平地评估问题复杂度。

图2显示，在所有问题类型中，标注员在简单问题上的表现最佳，其次是中等问题，最后是困难问题。人类判断与指定难度标签之间的一致性，支持了该基准难度结构的有效性。

## 5 实验设置

在我们的实验中，我们评估了六种通用大语言模型——通义千问- 2.5（7B、72B） [38], LLaMA-3（8B、70B） [10], Mistral-24B [14], 以及

**表5：英语和波斯语设置下，零样本、少样本和思维链提示的多项选择问答性能。灰色单元格标示每个模型/子类型中英语与波斯语两者中较高的结果。下划线标示每种模型/语言/子类型下表现最强的提示策略（不包括并列情况）。**

| 模型 | 英语 | | | | | | 波斯语 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 多答案 | 多跳 非答案 | 单答案 | 多答案 | 推理 非答案 | 单答案 | 多答案 | 多跳 非答案 | 单答案 | 多答案 | 推理 非答案 | 单答案 |
| 零样本 | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.83 | 0.33 | 0.66 | 0.78 | 0.48 | 0.71 | 0.78 | 0.29 | 0.66 | 0.77 | 0.48 | 0.69 |
| LLaMA 3 8B | 0.76 | 0.64 | 0.70 | 0.63 | 0.74 | 0.48 | 0.78 | 0.45 | 0.71 | 0.65 | 0.56 | 0.49 |
| Mistral 24B | 0.90 | 0.20 | 0.85 | 0.82 | 0.46 | 0.80 | 0.90 | 0.20 | 0.85 | 0.82 | 0.52 | 0.82 |
| Gemma 2 27B | 0.92 | 0.09 | 0.89 | 0.85 | 0.42 | 0.82 | 0.92 | 0.07 | 0.88 | 0.84 | 0.42 | 0.82 |
| LLaMA 3 70B | 0.95 | 0.05 | 0.91 | 0.86 | 0.33 | 0.84 | 0.95 | 0.08 | 0.91 | 0.86 | 0.40 | 0.83 |
| Qwen 2.5 72b | 0.94 | 0.33 | 0.90 | 0.88 | 0.70 | 0.87 | 0.94 | 0.38 | 0.88 | 0.88 | 0.71 | 0.87 |
| 少样本 | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.81 | 0.27 | 0.73 | 0.76 | 0.51 | 0.76 | 0.83 | 0.28 | 0.71 | 0.77 | 0.53 | 0.75 |
| LLaMA 3 8B | 0.66 | 0.12 | 0.67 | 0.53 | 0.18 | 0.40 | 0.79 | 0.15 | 0.70 | 0.65 | 0.24 | 0.56 |
| Mistral 24B | 0.43 | 0.26 | 0.27 | 0.41 | 0.24 | 0.27 | 0.46 | 0.30 | 0.28 | 0.44 | 0.27 | 0.30 |
| Gemma 2 27B | 0.93 | 0.29 | 0.90 | 0.83 | 0.60 | 0.82 | 0.92 | 0.26 | 0.90 | 0.84 | 0.59 | 0.81 |
| LLaMA 3 70B | 0.93 | 0.25 | 0.91 | 0.83 | 0.50 | 0.86 | 0.93 | 0.32 | 0.92 | 0.83 | 0.53 | 0.85 |
| Qwen 2.5 72b | 0.94 | 0.43 | 0.91 | 0.88 | 0.69 | 0.91 | 0.94 | 0.40 | 0.91 | 0.89 | 0.66 | 0.91 |
| 思维链 | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.81 | 0.21 | 0.74 | 0.79 | 0.35 | 0.88 | 0.76 | 0.28 | 0.67 | 0.77 | 0.39 | 0.82 |
| LLaMA 3 8B | 0.78 | 0.44 | 0.70 | 0.70 | 0.59 | 0.63 | 0.80 | 0.36 | 0.65 | 0.67 | 0.52 | 0.64 |
| Mistral 24B | 0.90 | 0.25 | 0.85 | 0.85 | 0.43 | 0.92 | 0.89 | 0.37 | 0.85 | 0.86 | 0.46 | 0.92 |
| Gemma 2 27B | 0.92 | 0.17 | 0.89 | 0.86 | 0.42 | 0.94 | 0.92 | 0.23 | 0.88 | 0.85 | 0.51 | 0.91 |
| LLaMA 3 70B | 0.93 | 0.23 | 0.91 | 0.88 | 0.44 | 0.95 | 0.92 | 0.28 | 0.91 | 0.87 | 0.48 | 0.95 |
| Qwen 2.5 72b | 0.92 | 0.20 | 0.90 | 0.89 | 0.39 | 0.93 | 0.92 | 0.27 | 0.88 | 0.90 | 0.43 | 0.93 |

Gemma-2（27B） [34]——涵盖不同的模型系列和规模，以尽量减少与架构和规模相关的偏差。此外，我们还纳入了*Dorna*[7]，，这是一种基于波斯语语料库训练的波斯语专用模型，以评估语言专门化的影响。

对于评估，我们使用**准确率** 评估布尔型问题。对于多项选择题，单答案和不可回答变体使用**准确率**，多答案情况使用**杰卡德系数**。对于事实型问题，不可回答实例使用**准确率**，简单的字符串匹配答案使用**包含匹配**，列表式答案使用**杰卡德系数** 。

所有实验——包括指令调优——均在 Together AI[8] 平台上进行，解码温度设为 0.7，并采用三种不同的提示策略：零样本、少样本和思维链。按问题类型分组的推理提示可在我们的 GitHub 仓库中找到。

## 6 实验

为评估 Parse 基准的有效性，我们开展了两组实验。首先，我们将 Parse 用作基准，评估各种大语言模型在不同问题类型、类别和子类型上的表现。其次，我们将 Parse 用作训练资源，对大语言模型进行微调，并探究该数据集对模型适配的价值。

模型表现。 我们使用两种等效的提示（分别为英语和波斯语）在 Parse 基准上评估了多个大语言模型。表 4、5 和 6 报告了所有配置下的结果。

[7]https://huggingface.co/PartAI/Dorna-Llama3-8B-Instruct [8] https://www.together.ai/

总体而言，结果表明，在布尔型、多项选择和事实型任务中，波斯语提示的表现始终优于英语提示。这很可能是因为问题本身是用波斯语编写的，而波斯语提示能更好地引导模型忠实地理解这些问题。关于提示策略，对于布尔型和多项选择问题，思维链 提示表现最佳，这可能是因为这些任务中的问题具有较强的推理性质。相比之下，对于事实型问题，少样本提示效果最佳，这表明事实型问题从示例驱动的提示中获益更多，而非逐步推理，这与先前的观察一致 [7]。

微调。 我们在 Parse 的训练部分上对两个大语言模型——LLaMA 3 8B 及其波斯语变体 Dorna——进行微调。为创建训练-测试集划分，我们从 18 个配置实例中的每个实例采样 120 个问答对，共得到 2,160 个问答对作为测试集。剩余的 8,640 个问答对用于训练。随后，我们在原始版本和微调设置下分别评估这两个模型。在推理过程中，根据 模型性能 实验的结果，我们使用波斯语提示；对于布尔型问题和多项选择题，采用思维链提示；对于事实型问题，则使用少样本提示。图 3 展示了结果。

正如预期，Dorna 的表现优于 LLaMA 3 8B，这可能是因为它在波斯语语料库上进行了预训练。微调进一步显著提升了两个模型的性能，其中微调后的 Dorna 整体表现最佳。这些发现表明，Parse 不仅是一个评测基准，也是适配支持波斯语的大语言模型的有效资源。

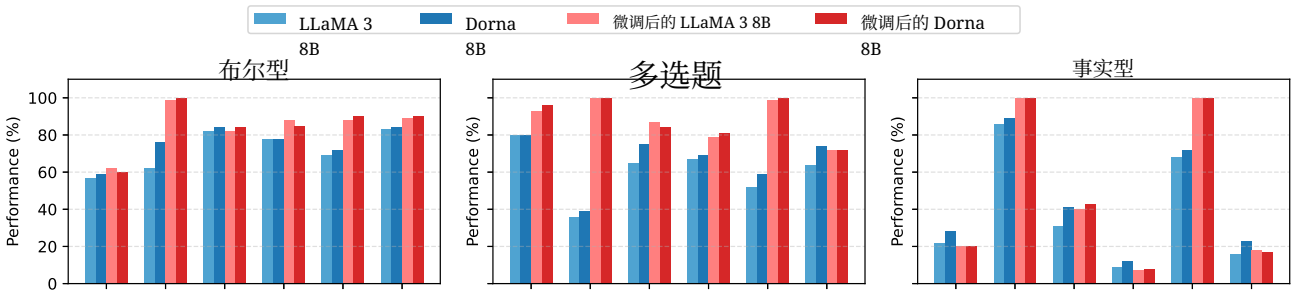


**图 3：LLaMA 3 8B 和 Dorna 在 Parse 上微调前后的表现。该图报告了从所有配置中抽取的 2,160 项测试集上的评估分数。**

**表 6：英语和波斯语环境下，零样本、少样本和思维链提示的事实型问答表现。灰色单元格表示每种模型/子类型中英语和波斯语表现较高的一方。下划线表示每种模型/语言/子类型中效果最强的提示策略（不包括并列情况）。**

| 模型 | 英语 | | | | | | 波斯语 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 多跳 | | | 推理 | | | 多跳 | | | 推理 | | |
| | 基于列表 | 非答案 | 简单 | 基于列表 | 非答案 | 简单 | 基于列表 | 非答案 | 简单 | 基于列表 | 非答案 | 简单 |
| **零样本** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.16 | 0.82 | 0.22 | 0.09 | 0.46 | 0.19 | 0.16 | 0.80 | 0.23 | 0.08 | 0.48 | 0.22 |
| LLaMA 3 8B | 0.24 | 0.01 | 0.27 | 0.08 | 0.01 | 0.18 | 0.23 | 0.33 | 0.27 | 0.08 | 0.27 | 0.17 |
| Mistral 24B | 0.38 | 0.79 | 0.57 | 0.17 | 0.70 | 0.29 | 0.39 | 0.94 | 0.59 | 0.18 | 0.89 | 0.29 |
| Gemma 2 27B | 0.46 | 0.47 | 0.66 | 0.26 | 0.61 | 0.31 | 0.47 | 0.61 | 0.64 | 0.28 | 0.75 | 0.29 |
| LLaMA 3 70B | 0.41 | 0.62 | 0.68 | 0.23 | 0.44 | 0.36 | 0.42 | 0.46 | 0.68 | 0.23 | 0.37 | 0.34 |
| Qwen 2.5 72b | 0.37 | 0.97 | 0.55 | 0.21 | 0.80 | 0.31 | 0.38 | 0.98 | 0.53 | 0.21 | 0.94 | 0.30 |
| **少样本** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.17 | 0.98 | 0.27 | 0.09 | 0.87 | 0.21 | 0.18 | 0.99 | 0.26 | 0.09 | 0.95 | 0.24 |
| LLaMA 3 8B | 0.19 | 0.98 | 0.28 | 0.07 | 0.88 | 0.11 | 0.22 | 0.86 | 0.31 | 0.09 | 0.68 | 0.16 |
| Mistral 24B | 0.22 | 0.30 | 0.33 | 0.09 | 0.41 | 0.09 | 0.17 | 0.47 | 0.22 | 0.09 | 0.70 | 0.06 |
| Gemma 2 27B | 0.47 | 0.98 | 0.65 | 0.28 | 0.91 | 0.35 | 0.47 | 0.99 | 0.64 | 0.28 | 0.91 | 0.35 |
| LLaMA 3 70B | 0.42 | 0.96 | 0.67 | 0.25 | 0.97 | 0.37 | 0.42 | 0.98 | 0.67 | 0.24 | 0.96 | 0.37 |
| Qwen 2.5 72b | 0.38 | 0.97 | 0.57 | 0.22 | 0.94 | 0.35 | 0.38 | 0.99 | 0.56 | 0.22 | 0.96 | 0.33 |
| **思维链** | | | | | | | | | | | | |
| 通义千问 2.5 7B | 0.17 | 0.94 | 0.23 | 0.09 | 0.48 | 0.22 | 0.17 | 0.74 | 0.24 | 0.09 | 0.36 | 0.24 |
| LLaMA 3 8B | 0.31 | 0.32 | 0.40 | 0.15 | 0.17 | 0.26 | 0.23 | 0.39 | 0.30 | 0.09 | 0.32 | 0.23 |
| Mistral 24B | 0.39 | 0.97 | 0.57 | 0.17 | 0.72 | 0.32 | 0.41 | 0.94 | 0.59 | 0.20 | 0.58 | 0.30 |
| Gemma 2 27B | 0.46 | 0.66 | 0.65 | 0.27 | 0.53 | 0.35 | 0.46 | 0.86 | 0.64 | 0.27 | 0.80 | 0.31 |
| LLaMA 3 70B | 0.43 | 0.53 | 0.69 | 0.23 | 0.36 | 0.36 | 0.45 | 0.79 | 0.68 | 0.23 | 0.63 | 0.37 |
| Qwen 2.5 72b | 0.37 | 0.98 | 0.54 | 0.21 | 0.77 | 0.34 | 0.41 | 0.97 | 0.58 | 0.22 | 0.55 | 0.31 |

## 7 结论

我们介绍了 Parse，这是首个面向波斯语推理问答的开放域基准。Parse 包含 10,800 个问题，涵盖多种格式、推理类别和难度级别，能够广泛且受控地覆盖问答现象。其结构化分类体系和多阶段验证确保了较高的语言、事实和结构质量。实验表明，当前的多语言模型和波斯语模型在波斯语推理任务上仍然面临困难，而在 Parse 上进行微调，则能使大多数配置取得显著改进。这些结果证明了该基准在评估和模型适配方面的有效性。

Parse 填补了波斯语自然语言处理和信息检索领域的重要资源缺口，为未来的多语言推理和大语言模型开发奠定了基础。未来，我们计划探索检索增强生成（RAG）以及 DeepSeek [11] 等更先进的、面向推理的模型。

# 参考文献


[1]Negin Abadani、Jamshid Mozafari、Afsaneh Fatemi、Mohamadali Nematbakhsh 和 Arefeh Kazemi。2021。ParSQuAD：基于 SQuAD 2.0 机器翻译的波斯语问答数据集。《国际网络研究杂志》4, 1（2021），34–46。arXiv：https://ijwr.usc.ac.ir/article_139661_877052c9a0d82d59550f67ed6be95c02.pdf doi:10.22133/ijwr.2021.293313.1101[2] Negin Abadani、Jamshid Mozafari、Afsaneh Fatemi、Mohammd Ali Nematbakhsh 和 Arefeh Kazemi。2021年。ParSQuAD：用于波斯语问答的机器翻译版 SQuAD 数据集。载于*2021*年第七届国际网络研究会议（*ICWR*）。163–168。doi:10.1109/ICWR51868.2021.9443126[3] Mohammad Amin Abbasi、Arash Ghafouri、Mahdi Firouzmandi、Hassan Naderi 和 Behrouz Minaei Bidgoli。2023年。Persianllama：迈向构建首个波斯语大语言模型。*arXiv*预印本 *arXiv:2312.15713*（2023年）。[4] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Floren- cia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。2023年。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*（2023年）。[5] 匿名作者。2024年。超越准确率：理解大语言模型在面向人类设计的考试中的表现。https://openreview.net/forum?id=Cth1PyCwZt[6] Mohammad Yasin Ayoubi、Sajjad & Davoodeh。2021年。PersianQA：一个波斯语问答数据集。https://github.com/SajjjadAyobi/PersianQA。[7] Rakesh Chada 和 Pradeep Natarajan。2021年。FewshotQA：一种利用预训练文本到文本模型进行问答任务少样本学习的简单框架。载于*2021*年自然语言处理实证方法会议论文集，Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih（编）。计算语言学协会，线上及多米尼加共和国蓬塔卡纳，6081–6090。doi:10.18653/v1/2021.emnlp- main.491[8] Kasra Darvishi、Newsha Shahbodaghkhan、Zahra Abbasiantaeb 和 Saeedeh Momtazi。2023年。PQuAD：一个波斯语问答数据集。《计算机语音与语言》80 2023 101486 10 1016 2023 101486（），。doi: . /j.csl. 。[9] Arash Ghafouri、Mohammad Aghajani Asl、Hasan Naderi 和 Mahdi Firouz- mandi。2025年。IslamicPCQA：一个用于伊斯兰文本资源中波斯语多跳复杂问答的数据集。《*IEEE*音频、语音与语言处理汇刊》33（2025年），3801–3812。doi:10.1109/TASLPRO.2025.3587450[10] Aaron Grattafiori、Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。2024年。Llama 3模型家族。*arXiv*预印本 *arXiv:2407.21783*（2024年）。[11] Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Ruoyu Zhang、Runxin Xu、Qihao Zhu、Shirong Ma、Peiyi Wang、Xiao Bi 等。2025年。DeepSeek-R1：通过强化学习激发大语言模型的推理能力。*arXiv*预印本 *arXiv:2501.12948*（2025年）。[12] Lei Huang、Weijiang Yu、Weitao Ma、Weihong Zhong、Zhangyin Feng、Haotian Wang、Qianglong Chen、Weihua Peng、Xiaocheng Feng、Bing Qin 和 Ting Liu。2025年。大语言模型中的幻觉综述：原理、分类体系、挑战与开放问题。《*ACM*信息系统汇刊》43，2，文章 42 2025 55 10 1145 3703155（1月），页码。doi: . /[13] Naghme Jamali、Yadollah Yaghoobzadeh 和 Heshaam Faili。2022年。PerCQA：波斯语社区问答数据集。载于第十三届语言资源与评测会议论文集，Nicoletta Calzolari、Frédéric Béchet、Philippe Blache、Khalid Choukri、Christopher Cieri、Thierry Declerck、Sara Goggi、Hitoshi Isahara、Bente Maegaard、Joseph Mariani、Hélène Mazo、Jan Odijk 和 Stelios Piperidis（编）。欧洲语言资源协会，法国马赛，6083–6092。https://aclanthology.org/2022.lrec-1.654/[14] Albert Q Jiang、Alexandre Sablayrolles、Arthur Mensch、Chris Bamford、De- vendra Singh Chaplot、Diego de las Casas、Florian Bressand、Gianna Lengyel、Guillaume Lample、Lucile Saulnier 等。2023年。Mistral 7B。*arXiv*预印本 *arXiv:2310.06825*（2023年）。[15] Mohammad Javad Ranjbar Kalahroodi、Amirhossein Sheikholselami、Sepehr Karimi、Sepideh Ranjbar Kalahroodi、Heshaam Faili 和 Azadeh Shakery。2025年。PersianMedQA：波斯语医学领域大语言模型的以语言为中心的评测。*arXiv*预印本 *arXiv:2506.00250*（2025年）。[16] Arefeh Kazemi、Jamshid Mozafari 和 Mohammad Ali Nematbakhsh。2022年。PersianQuAD：波斯语原生问答数据集。《*IEEE*开放获取》10（2022年），26045–26057。doi:10.1109/ACCESS.2022.3157289[17] Arefeh Kazemi、Zahra Zojaji、Mahdi Malverdi、Jamshid Mozafari、Fatemeh Ebrahimi、Negin Abadani、Mohammad Reza Varasteh 和 Mohammad Ali Nemat- bakhsh。2023年。FarsNewsQA：面向波斯语新闻文章的基于深度学习的问答系统。《信息检索杂志》26，1（2023年3月19日），3。doi:10.1007/s10791-023-09417-2[18] Daniel Khashabi。2019年。面向自然语言理解的推理驱动问答。宾夕法尼亚大学。[19] Daniel Khashabi、Snigdha Chaturvedi、Michael Roth、Shyam Upadhyay 和 Dan Roth。2018年。超越表层：一个针对多个句子阅读理解的挑战集。载于*2018*年会议论文集美国计算语言学协会分会：人类语言技术，第*1*卷（长篇论文），Marilyn Walker、Heng Ji 和 Amanda Stent（编）。计算语言学协会，新奥尔良，路易斯安那州，252 262 10 18653 1 18 1023 – 。doi: 。/v /N -[20] Jaehyung Kim、Dongyoung Kim 和 Yiming Yang。2024年。借助黑盒大语言模型学习纠正问答推理。载于*2024*年自然语言处理实证方法会议论文集，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编）。计算语言学协会，美国佛罗里达州迈阿密，8916–8937。doi:10.18653/v1/2024.emnlp-main.504[21] Tom Kwiatkowski、Jennimaria Palomaki、Olivia Redfield、Michael Collins、Ankur Parikh、Chris Alberti、Danielle Epstein、Illia Polosukhin、Jacob Devlin、Kenton Lee、Kristina Toutanova、Llion Jones、Matthew Kelcey、Ming-Wei Chang、Andre w M. Dai、Jakob Uszkoreit、Quoc Le 和 Slav Petrov。2019年。Natural Questions：问答研究基准。计算语言学协会会刊 7 2019 452 466 10 1162 _ _00276 ( ), – 。doi: 。/[22] tacl a Guokun Lai、Qizhe Xie、Hanxiao Liu、Yiming Yang 和 Eduard Hovy。2017年。RACE：面向考试的大规模阅读理解数据集。载于*2017*年自然语言处理实证方法会议论文集，Martha Palmer、Rebecca Hwa 和 Sebastian Riedel（编）。计算语言学协会，丹麦哥本哈根，785–794。doi:10.18653/ v1/D17-1082[23] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel、Sebastian Riedel 和 Douwe Kiela。2020年。面向知识密集型自然语言处理任务的检索增强生成。载于第*34*届神经信息处理系统国际会议论文集（加拿大不列颠哥伦比亚省温哥华）（*NIPS '20*）。Curran Associates Inc.，美国纽约州雷德胡克，文章793，16页。[24] Junyu Luo、Weizhi Zhang、Ye Yuan、Yusheng Zhao、Junwei Yang、Yiyang Gu、Bohan Wu、Binqi Chen、Ziyue Qiao、Qingqing Long 等。2025年。大语言模型智能体：方法、应用与挑战综述。*arXiv*预印本 *arXiv:2503.21460*（2025年）。[25] Jamshid Mozafari、Arefeh Kazemi、Parham Moradi 和 Mohammad Ali Nemat-bakhsh。2022年。PerAnSel：一种用于波斯语问答的新型深度神经网络系统。计算智能与神经科学2022，1（2022），3661286。arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1155/2022/3661286 doi:10.1155/2022/3661286[26] Hariom A Pandya 和 Brijesh S Bhatt。2021年。问答综述：方向、挑战、数据集与评估矩阵。*arXiv*预印本 *arXiv:2112.03572*（2021年）。[27] Nisarg Patel、Mohith Kulkarni、Mihir Parmar、Aashna Budhiraja、Mutsumi Nakamura、Neeraj Varshney 和 Chitta Baral。2024年。Multi-LogiEval：迈向大语言模型多步逻辑推理能力评估。载于*2024*年自然语言处理实证方法会议论文集，Yaser Al-Onaizan、Mohit Bansal 和 Yun-Nung Chen（编）。计算语言学协会，美国佛罗里达州迈阿密，20856–20879。doi:10.18653/v1/2024.emnlp-main.1160[28] Aske Plaat、Annie Wong、Suzan Verberne、Joost Broekens、Niki van Stein 和 Thomas Back。2024年。大语言模型推理：综述。*arXiv*预印本 *arXiv:2407.11511*（2024年）。[29] Alessandro Raganato、Rafael Penaloza、Marco Viviani 和 Gabriella Pasi。2024年。大语言模型的推理能力与不变性。载于*2024*年*IEEE/WIC*网络智能与智能代理技术国际会议（*WI-IAT*）。IEEE计算机学会，美国加利福尼亚州洛斯阿拉米托斯，125–132。doi:10.1109/WI-IAT62293.2024.00025[30] Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。知其所不知：SQuAD中的不可回答问题。载于计算语言学协会第*56*届年会论文集（第*2*卷：短篇论文），Iryna Gurevych 和 Yusuke Miyao（编）。计算语言学协会，澳大利亚墨尔本，784 789 10 18653 1 18 2124 – 。doi: 。/v /P - Pedram[31] Rostami、Ali Salemi 和 Mohammad Javad Dousti。2024年。Persian-mind：跨语言波斯语-英语大语言模型。*arXiv*预印本 *arXiv:2401.06466*（2024年）。[32] Zhang Shengyu、Dong Linfeng、Li Xiaoya、Zhang Sen、Sun Xiaofei、Wang Shuhe、Li Jiwei、Runyi Hu、Zhang Tianwei、Fei Wu 等。2023年。大语言模型的指令调优：综述。*arXiv*预印本 *arXiv:2308.10792*（2023年）。[33] Mobina Taji、Arash Ghafouri、Hasan Naderi 和 Behrouz Minaei-Bidgoli。2025年。PersianMHQA：基于维基百科全书的开放领域波斯语多跳问答数据集。*ACM*亚洲低资源语言与信息处理汇刊24，2，文章18（2025年2月），17页。doi:10.1145/3711826[34] Gemma 团队、Aishwarya Kamath、Johan Ferret、Shreya Pathak、Nino Vieillard、Ramona Merhej、Sarah Perrin、Tatiana Matejovicova、Alexandre Ramé、Morgane Rivière 等。2025年。Gemma 3技术报告。*arXiv*预印本 *arXiv:2503.19786*（2025年）。[35] Gernot L. Windfuhr。2009年。伊朗语言。载于伊朗语言，Gernot Windfuhr（编）。Routledge。https://www.taylorfrancis.com/books/edit/10.4324/9780203641736/iranian-languages-gernot-windfuhr-gernot-windfuhr[36] Jian Wu、Linyi Yang、Yuliang Ji、Wenhao Huang、Börje F Karlsson 和 Manabu Okumura。2024年。Gendec：一种稳健的生成式问题分解方法

用于多跳推理。*arXiv*预印本 *arXiv:2402.11166*（2024年）。[37] Lyu Yana、Qin Xutong、Du Xiuli、Zhao Niujie 和 Qiu Shaoming。2025年。面向知识图谱多跳问答的多路径推理。《电子学报》34，2（2025年），642–648。doi:10.23919/cje.2023.00.044[38] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。2025年。通义千问3技术报告。*arXiv*预印本 *arXiv:2505.09388*（2025年）。[39] Zhilin Yang、Peng Qi、Saizheng Zhang、Yoshua Bengio、William Cohen、Ruslan Salakhutdinov 和 Christopher D. Manning。2018年。HotpotQA：多样化、可解释的多跳问答数据集。载于*2018*年自然语言处理实证方法会议论文集，Ellen Riloff、David Chiang、Julia Hockenmaier 和 Jun'ichi Tsujii（编）。计算语言学协会，比利时布鲁塞尔，2369–2380。doi:10.18653/v1/D18-1259[40] Wangjie You、Xusheng Wang、Xing Wang、Wenxiang Jiao、Chao Feng、Juntao Li 和 Min Zhang。2025年。多跳推理视角下的中文常识推理基准测试。*arXiv*预印本 *arXiv:2510.08800*（2025年）。[41] Zhehao Zhang、Xitao Li、Yan Gao 和 Jian-Guang Lou。2023年。CRT-QA：表格数据复杂推理问答数据集。载于*2023*年自然语言处理实证方法会议论文集，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，新加坡，2131–2153。doi:10.18653/v1/2023.emnlp-main.132[42] Yutao Zhu、Huaying Yuan、Shuting Wang、Jiongnan Liu、Wenhan Liu、Chenlong Deng、Haonan Chen、Zheng Liu、Zhicheng Dou 和 Ji-Rong Wen。2025年。用于信息检索的大语言模型：综述。*ACM*信息系统汇刊（2025年9月）。doi:10.1145/3748304 已接受。[43] Arkaitz Zubiaga。2024年。大语言模型时代的自然语言处理。《人工智能前沿》第6卷 - 2023（2024年）。doi:10.3389/frai.2023.1350306