



文献 CSTR:

32001.14.11-6035.csd.2025.0062.zh



文献 DOI:

10.11922/11-6035.csd.2025.0062.zh

数据 DOI:

10.57760/sciencedb.23774

文献分类: 计算机科学技术

收稿日期: 2025-04-16

录用日期: 2025-05-21

发表日期: 2026-02-24

MTR-QA: 多类型推理问答数据集

王强^{1,2}, 姜程林^{1,2}, 马宁^{1,2*}, 李英杰^{1,2}, 吴文社^{1,2}

1. 西北民族大学, 语言与文化计算教育部重点实验室, 兰州 730030

2. 西北民族大学, 甘肃省民族语言文化智能信息处理重点实验室, 兰州 730030

摘要: 针对大语言模型在处理复杂推理任务时容易出现逻辑混乱和隐含关系捕捉不足问题, 本研究提出并构建了一个覆盖多推理类型的高质量问答数据集 MTR-QA (Multi-Type Reasoning Question Answering)。通过收集整理近 15 年的公务员考试真题与权威模拟题库, 使用包括文本标准化、哈希去重、近重复检测与语义嵌入相似度过滤的多阶段文本处理框架实现数据清洗, 有效降低冗余与噪声干扰。为确保数据可靠性, 本研究设计了一种融合 GPT-4、DeepSeek-V1 和 Qwen-2.5 的多模型评分机制, 从完整性 (CPL)、准确性 (ACC)、安全性 (SFC) 和思维链质量 (CoT-Q) 4 个维度进行量化评估, 结合人工审核对低质量样本进行过滤, 消除有害性、歧义性和不安全性数据, 确保数据的可靠性与严谨性, 最终筛选出 24,312 条高质量数据, 使用 json 格式存储, 大小为 34.1 MB。每条数据样本包含问题、选项、答案、思维链、类型和难度等级共 6 个属性, 并将其分为逻辑、语义、数理和综合知识推理 4 大核心推理类型。MTR-QA 数据集在推理类型和题材广度上实现了扩展, 后续可为大语言模型预训练、监督微调和模型评估等多种推理任务提供可靠数据基础, 推动大语言模型在复杂推理问答场景下的性能提升。

关键词: 大语言模型; 多类型; 推理问答; 数据集

数据库(集)基本信息简介

数据库(集)名称	MTR-QA: 多类型推理问答数据集
数据通信作者	马宁 (maning@xbmu.edu.cn)
数据作者	王强、姜程林、马宁、李英杰、吴文社
数据量	34.1 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.23774
基金项目	国家自然科学基金项目 (62466052); 甘肃省中央引导地方科技发展资金项目 (25ZYJA034); 西北民族大学中央高校基本科研业务费专项资金项目 (31920250008)。
数据库(集)组成	数据集包括 24,312 条 json 格式的数据, 每条 json 对象包含 6 个属性, 分别为 instruction、question、answer、target、lable 和 difficulty。

* 论文通信作者

马宁: maning@xbmu.edu.cn

引言

近年来,大语言模型(Large Language Model, LLM)在自然语言处理领域取得了突破性进展,以其强大的语义理解与文本生成能力在开放域问答、多轮对话等任务中展现出显著优势^[1]。基于 GPT-4、LLAMA 和 DeepSeek 为代表的基座模型,使用大量数据进行预训练和指令微调后^[2-3],能够生成与上下文相关且连贯的答案,可以流畅完成多轮对话和代码生成等任务,但其在处理多类型推理问答的复杂场景时,生成答案往往存在严谨性缺乏、对逻辑关系的捕捉不足,以及跨领域问题易产生“幻觉”等现象^[4],准确性和可解释性仍有提升空间。大语言模型的推理问答能力高度依赖数据多样性,凸显高质量数据集在模型学习与评估中的重要作用^[5]。

推理问答是大模型运用逻辑、数学、多步信息整合的综合能力体现,在对上下文和问题进行思考后作出回答,满足用户信息需求^[6]。而提升大模型推理问答能力的关键之一是高质量、多种类的训练和评估推理数据集^[7]。现有数据集如 GSM8K、PIQA、LogiQA 等,通过精心设计问题结构、多种类知识来源与推理链条^[8-9],同时融入思维链(CoT, Chain of Thought)等新技术融入改变问答推理的基本范式,可显著提升模型的问答推理能力,利用 CoT 微调的模型使用带有逐步推理的指令数据,模型性能显著优于之前的微调模型^[10]。对于复杂推理的 GSM8K,通过设计思维链,8 个示例的提示就可达到良好性能,而微调方法则需要利用完整训练集,这种鲜明的对比也增强了用户对于 CoT 等新技术的期待。CoT 技术同时增加了模型输出的可解释性,逐步推理的过程让大型语言模型给出更加可信的结果^[11]。整体而言,现有数据集多聚焦于单一推理形式或特定任务^[12],难以覆盖现实场景中的复合推理需求,缺乏一个可以覆盖多类型推理的统一数据集,来服务大模型预训练、监督微调和评估任务。

本研究构建的 MLR-QA (Multi-Type Reasoning Question Answering) 数据集,通过系统性数据采集、清洗、标注及质量控制,为大模型提供涵盖逻辑推理、语义推理、数理推理和综合知识推理 4 类核心推理能力的训练与评估资源。数据集构建遵循多源采集、分层处理、动态优化流程。首先从相关网站爬取原始语料,确保问题种类的多样性;其次通过对文本数据去重与格式统一化,对数据进行预处理;最后采用大模型与人工相结合的方法,对每一条数据的问题、答案、思维链进行多粒度分析标注。并通过数据评估优化数据样本,提升数据集质量,确保在大模型的预训练、监督微调和评估过程中的高效适配。MLR-QA 数据集为大模型推理能力的针对性优化和细粒度评估提供了标准化基准,有助于提升模型在推理问答任务中的性能,促进大模型生成内容更加准确并更具说服力。

1 数据采集和处理方法

1.1 数据采集

本研究构建的大模型问答数据集来源于:(1)国内影响力较高的主流公务员事业编考试培训机构(粉笔、中公教育等)的模拟题库;(2)国家公务员事业编考试历年发布的标准化试题与真题汇编。数据重点选取近 15 年共 32,169 条原始试题数据,题目形式以客观选择题和主观问答题为主,涉及科技、人文、法律、历史、经济等多个学科领域,具有较强的知识深度与广度。在数据采集过程中,优先选取命题质量高、语言规范严谨、逻辑结构清晰的原始试题数据,该类试题在命题逻辑、内容组织和语言表述方面具有高度标准化与严谨性特性,涵盖知识理解、逻辑推理与语言表达等多

种能力考查维度，具备较高代表性，能反映真实应用场景的问答信息需求。

1.2 数据处理

1.2.1 文本清洗与去重

原始数据存在不同程度的重复和冗余，以及数据格式不统一等问题。为保证数据集的信息质量，需对其进行筛选和清洗。本研究构建了多阶段文本处理框架，用于筛除部分格式出错数据和重复性冗余数据以提高数据纯净度，提升数据集整体质量。该框架融合文本标准化处理、哈希匹配、近重复性检测与语义嵌入相似度过滤 4 个模块，如图 1 所示。

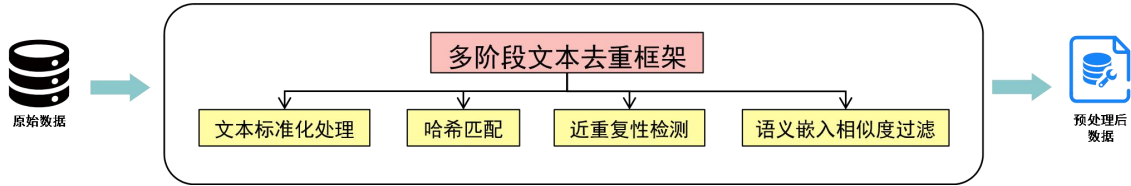


图 1 多阶段文本去重框架数据处理流程

Figure 1 Data processing workflow of the multi-stage text de-duplication framework

文本标准化处理模块对原始文本进行统一格式处理，包括去除多余空格、空行和特殊字符，以及标点规范化，提升后续模块的去重效果；在哈希匹配模块中使用 SimHash 相似度判定，以汉明距离为相似度指标识别数据的结构有轻微变化但是语义基本一致的内容；用近重复性检测模块识别内容结构和词语相近的段落，但是内容不完全匹配的内容；语义嵌入相似度过滤模块通过将文本编码为高维度语义向量，计算余弦相似度来识别语义等价但表述方式差异较大的问答对。通过该框架处理实现对原始数据的预处理工作。多阶段文本处理框架原理如公式（1）所示：

$$F = \begin{cases} 1, & (D_H \leq \tau_h \cup J \geq \tau_j \cup \text{sim}_{\cos} \geq \tau_c) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

其中 D_H 是哈希去重的汉明距离， J 是近重复性检测的 Jaccard 相似度， $\text{sim}_{\cos}$ 是语义嵌入相似度过滤的余弦相似度， τ 为设定的阈值。当数据文本对满足括号中任意一个判定条件时， F 值为 1，判定为重复性文本，去除该条数据样本；反之， F 值为 0，判定为非重复性文本，保留该条数据样本。

1.2.2 数据整理与标注

原始数据经过筛选清洗后，通过数据整理和标注来确保 MTR-QA 数据集的一致性。按照原始试题不同模块中对应的题目类型进行自动化分类，将逻辑判断、言语理解、数量计算、综合知识判断 4 个题型分别对应划分为逻辑推理、语义推理、数理推理和综合知识推理 4 个类别。分类完成后，将每条数据样本按照规定格式与属性进行结构化标注，每条数据包含 instruction（用户提出的指令）、question（用户提供的选项）、answer（正确回答）、target（思维链）、lable（题目推理类型）和 difficulty（题目难度等级）6 个属性，各属性分别对应数据样本的题干、选项、正确答案、详细解题步骤、类别和难度，属性内容均由自动标注脚本将清洗去重后的数据样本依据对应的属性填写，通过预设模板进行人工抽检修正，确保数据样本内容完整、格式统一。

2 数据样本描述

MTR-QA 数据集包含 24,312 条推理数据,其中有 8,740 条逻辑推理数据、9,105 条语义推理数据、2,647 条数理推理数据、3,818 条综合知识推理的数据,大小为 34.1 MB,用 json 格式存储,如表 1 所示。

表 1 数据集推理类型详情

Table 1 Details of reasoning types in the dataset

分类	条数	数据集大小
逻辑推理	8,740	34.1 MB
语义推理	9,105	
数理推理	2,647	
综合知识推理	3,818	

每个 json 对象有 6 个属性: instruction、question、answer、target、lable 和 difficulty,分别对应: 用户提出的指令、用户提供的选项、正确回答、思维链、题目推理类型和题目难度等级。其中 lable 对应不同的推理类型,分为逻辑推理、语义推理、数理推理和综合知识推理;difficulty 按照语言难易程度、解题步骤和知识点复杂性分为初级、中级和高级三种难度级别,分级方法可以清晰地评估难易程度。数据集样本示例如表 2 所示。

表 2 数据集样本示例

Table 2 Examples of the dataset samples

属性	解释说明	样例
instruction	用户提出的指令	小曾、小孙、小石三个人是好朋友,他们每个人都有一个妹妹,且都比自己小 11 岁。三个妹妹的名字叫小英、小丽、小梅。已知小曾比小英大 9 岁,小曾与小丽年龄之和是 52 岁,小孙与小丽年龄之和是 54 岁。由此可知,谁是谁的妹妹?
question	用户提供的选项	A.小梅是小孙的妹妹,小丽是小曾的妹妹,小英是小石的妹妹 B.小梅是小石的妹妹,小丽是小孙的妹妹,小英是小曾的妹妹 C.小梅是小曾的妹妹,小丽是小石的妹妹,小英是小孙的妹妹 D.小梅是小曾的妹妹,小丽是小孙的妹妹,小英是小石的妹妹
answer	正确回答	C
target	思维链	思维链: 第一步: 分析题干。①小曾、小孙、小石都有一个妹妹,且都比自己小 11 岁②三个妹妹的名字叫小英、小丽、小梅。③小曾比小英大 9 岁④小曾与小丽年龄之和是 52 岁⑤ 小孙与小丽年龄之和是 54 岁。第二步: 结合题干分析选项。题干中①③④都提到了小曾,从最大信息入手,根据③,小曾比小英大 9 岁,因此小曾的妹妹不是小英。排除 B。根据②,小曾与小丽的年龄之和为 52,①小曾比妹妹大 11 岁,假设小丽是小曾妹妹,那她年龄应该为 $(52-11)/2$,结果不为整数,所以小曾的妹妹也不是小丽,排除 A 项。因此小曾的妹妹只能

属性	解释说明	样例
		为小梅, 根据条件⑤, 小孙与小丽年龄之和是 54 岁, 条件①小孙比妹妹大 11 岁, 假设小丽是小孙的妹妹, 那她年龄应该为 $(54-11)/2$, 结果不为整数, 所以小孙的妹妹不是小丽, 因此小孙的妹妹是小英, 排除 D 选项。故正确答案为 C。
lable	题目推理类型	逻辑推理
difficulty	题目难度等级	高级

3 数据质量控制和评估

3.1 数据质量控制

使用多阶段文本处理框架以及数据整理和标注, 对数据样本进行文本格式标准化和筛除重复数据后, 仍然存在部分样本有语义模糊、逻辑不严谨和内容敏感等问题。本研究引入多模型评分机制对数据集质量进行控制, 使用 GPT-4、DeepSeek-V1 和 Qwen2.5 三种推理问答能力优秀的主流大语言模型, 从数据样本完整性 (CPL)、准确性 (ACC)、安全性 (SFC) 和每条数据特定的思维链质量 (CoT-Q) 4 个维度进行语义细粒度分析与评分, 对数据样本的有害性、安全性和语义歧义性进行修正, 提升数据集在大模型推理问答场景中的有效可靠性, 如图 2 所示。

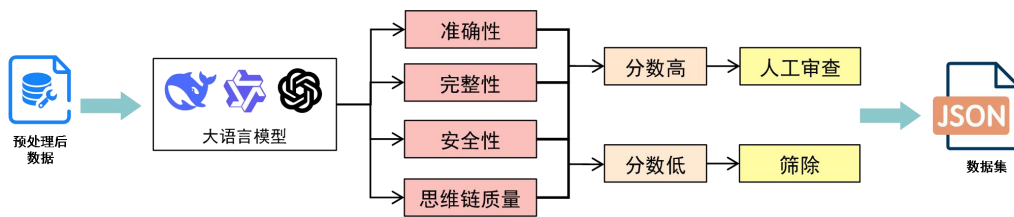


图 2 多模型评分机制示意图

Figure2 Schematic diagram of the multi-model scoring mechanism

其中, 准确性判断答案与参考解答的一致性 & 知识性错误; 完整性衡量问题与答案是否具有逻辑闭环, 是否存在信息缺失; 安全性审查数据样本敏感性与合规性; 思维链质量衡量推理过程的严谨性、解释力与可迁移性。多模型评分机制如下:

$$S_d(x) = \frac{1}{3} \sum_{i=1}^3 s_d^i(x) \quad (2)$$

$$S(x) = \sum_{d \in D} w_d \cdot S_d(x) \quad (3)$$

公式 (2) 为单维度模型评分公式, $S_d(x)$ 表示该样本在第 d 个维度的融合评分, 它是三个模型在该维度评分的算术平均, $s_d^i(x)$ 表示第 i 个模型对样本 x 在第 d 个维度上的评分 (取值范围为 $[0,1]$), $D=\{CPL, ACC, SAF, CoT-Q\}$ 表示评分维度。公式 (3) 为数据样本总评分公式, $S(x)$ 表示样本 x 的综合质量评分, w_d 表示维度 d 的权重, 本研究中 $w=(0.25, 0.25, 0.2, 0.3)$ 分别对应 CPL、ACC、SAF、CoT-Q 四个不同维度。基于多模型评分机制, 对数据样本进行筛选。当 $S_d(x) < t$ ($t=0.8$) 时, 表明该条数据样本 x 在维度 d 上的融合评分未达到质量要求的最低阈值; 当思维链质量 $S_{CoT-Q}(x) < 0.6$ 时, 表明该条数据思维链存在表达混乱的逻辑问题, 因此筛除出现这两种情况的数据样本。反之, 将剩余高分的数据样本纳入最终数据集, 作为大模型的训练与评估数据资源。

多模型评分机制可以充分发挥多个模型理解、推理判断的互补优势，筛除包含敏感信息、有害内容或存在歧义的数据样本，缓解单个模型出现评估偏差问题，提高对数据以及思维链评分的客观性，提升数据集可信度，促进模型推理过程更加准确和严谨。

3.2 数据评估

在多模型评分机制基础上，本研究引入人工审核方法进行数据集评估，聚焦数据内容的语义合理性和思维链逻辑正确性两大方面，弥补大模型在语境理解与边界判断上的不足。人工审核可确保数据样本内容具有清晰语义结构，保障思维链推理过程的逻辑连贯。审核团队由 4 名全日制研究生组成，具备自然语言处理专业知识背景，在逻辑推理、语义理解等方面有相应经验，通过审核前学习审核规则、示例讲解确保成员对规则有准确理解。为保证审核一致性，特制定了详细的审核标准，对存在多义性、模糊性的数据样本，重点评估其上下文信息是否明确，检查其思维链推理过程中是否存在逻辑断裂、步骤跳跃等问题。对包含不合理语义内容和逻辑问题出错的数据样本，审核人员进行标注并提出修改建议，纠正存在偏差的错误数据样本。确保每条数据样本中问题与回答的语义保持一致，准确表达相关领域知识，确保最终数据集语义清晰、思维链逻辑连贯。

为进一步评估 MTR-QA 数据集有效性与可靠性，本研究使用 3 个有代表性的开源大模型：ChatGLM4-9B、deepseek-llm-7b 和 Llama3-8B-Chinese-Chat，并在与 MTR-QA 数据量差距较小的 LogiQA 数据集上进行对比实验。实验在 4 张 A6000GPU 服务器上运行，采用 lora 微调方法，使用两个数据集对不同模型进行训练，并在相同测试集上评测模型回答准确率（Accuracy Rate），训练集和测试集比例为 8 : 2。实验结果如表 3 所示，使用 MTR-QA 数据集训练后的准确率均高于使用 LogiQA 数据集，表明 MTR-QA 数据集在支持模型推理问答方面更具优势。

表 3 实验对比结果表

Table3 Experimental comparison results

模型/数据集	MTR-QA	LogiQA
ChatGLM4-9B	62.37	56.53
deepseek-llm-7b	64.57	60.28
Llama3-8B-Chinese-Chat	43.69	40.66

经过多阶段文本处理框架处理、多模型评分机制质量控制以及人工审核后，最终从 32,169 条原始数据中筛选出 24,312 条高质量推理问答数据。MTR-QA 数据集中 instruction、question 和 target 属性内容的字段长度按照不同推理类型进行长度区间统计，如图 3 所示。

4 数据价值

与传统单一推理任务数据集不同，MTR-QA 数据集涵盖 4 大类型推理任务，并细分为三个难度级别。可以更精准地表示问题的复杂程度，满足不同能力模型需求。在数据集构建过程中，采用多阶段文本去重框架进行数据清洗与预处理，确保数据一致性。在质量控制和评估方面，通过大模型评分与人工审核相结合的方式，实现对数据集的质量把控，提高了数据集的整体质量，能够有效模拟和解决实际应用的多层次推理问题。

MTR-QA 数据集在自然语言处理、机器学习和人工智能等多个领域具有潜在的应用价值。它能

够模拟解决实际应用中的复杂推理问题，为大模型问答推理能力的系统化提升提供可靠数据基础，帮助研究人员在多种推理类型任务上进行模型训练和优化，并为大模型自身能力的性能评估提供数据资源。此外，在真实和复杂场景中，本数据集也能推动智能问答系统在自然语言理解、对话系统等领域的应用，为通用人工智能（AGI）的实现提供数据支持。

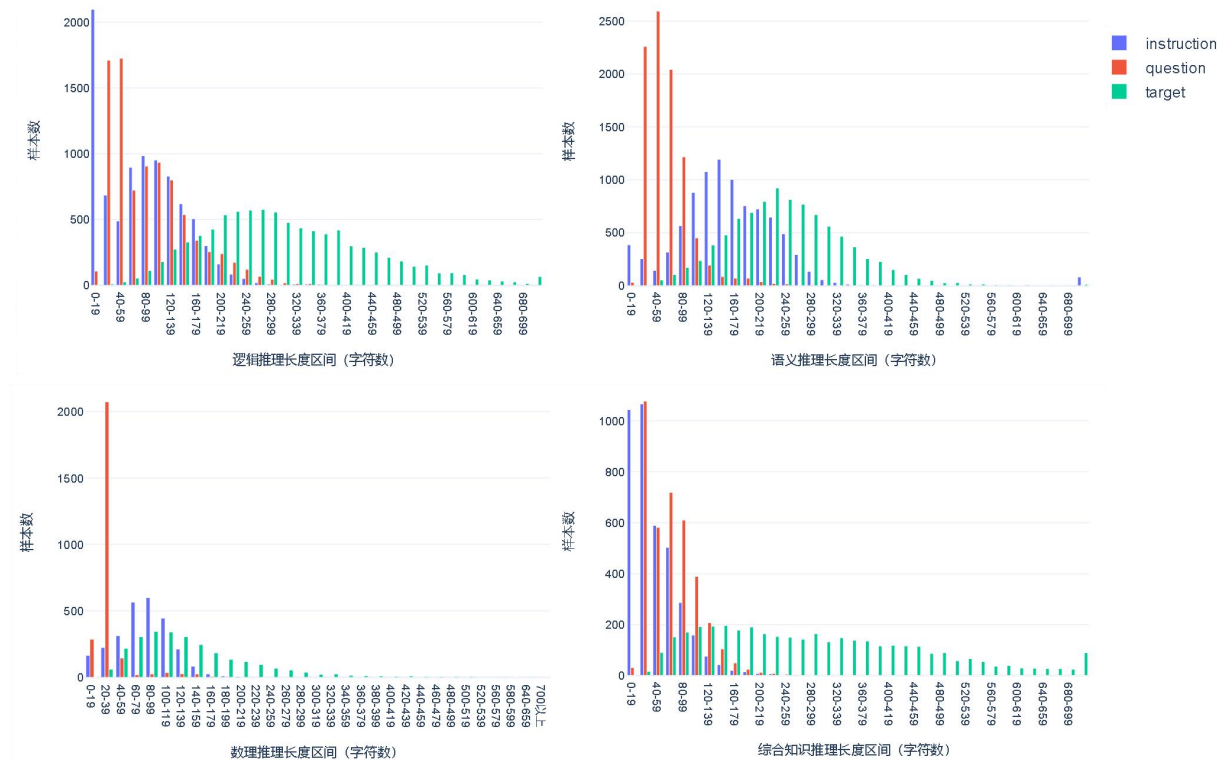


图3 不同类型对应的 instruction、question 和 target 属性长度分布

Figure3 Distribution of the lengths of the attributes instruction, question and target attributes across different types

5 数据使用方法和建议

MTR-QA 数据集适用于大模型预训练、指令微调、多任务学习分层训练、模型推理能力评估与对比实验。在使用时，应合理选择推理类型与任务难度，注意数据样本分布的偏差，防止模型出现推理类型偏倚问题，影响模型整体在推理任务上的表现。

6 AIGC 使用说明

在本研究的准备过程中，作者使用了人工智能生成内容（AIGC）工具。具体情况如下：

数据处理：利用 GPT-4 辅助编写数据清洗、去重及标注的自动化 Python 脚本，同时使用 GPT-4、DeepSeek-V1、Qwen2.5 作为质量控制辅助工具，依托其强推理能力，针对多类型推理问答数据的核心质量维度开展标准化辅助校验。作者对生成的代码进行了逐行调试、功能优化和逻辑验证，确保代码可正常运行并符合数据处理的实际需求。

公式编辑：利用 GPT-4 辅助完成数据处理与质量评估相关公式的初步推导和格式规范，作者对公式的逻辑合理性、符号使用准确性进行了严格核查与修正，保证公式表达符合学术规范和研究实际。

论文撰写：使用 GPT-4 进行语言润色和措辞建议。所有生成文本均经过作者的重大修改、事实

核查与整合, 以确保其准确反映研究思想和发现。

责任声明: 作者仔细审阅和编辑了所有 AIGC 辅助生成的内容, 并对论文的整体内容、数据真实性及学术诚信承担全部责任。AIGC 工具未被列为作者。

作者分工职责

王强 (2001—), 男, 山西长治人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据收集、数据预处理、数据集质量分析、数据集构造、论文撰写。

姜程林 (1998—), 男, 辽宁营口人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据收集、数据校对、质量审核。

马宁 (1981—), 男, 宁夏吴忠人, 博士, 教授, 研究方向为自然语言处理。主要承担工作: 数据集构建指导、数据质量控制、论文写作指导。

李英杰 (2000—), 男, 河南洛阳人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据收集、数据校对。

吴文社 (2001—), 男, 山东菏泽人, 硕士研究生, 研究方向为自然语言处理。主要承担工作: 数据收集、质量审核。

参考文献

- [1] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45. DOI:10.1145/3641289.
- [2] WU X K, CHEN M, LI W Y, et al. LLM fine-tuning: concepts, opportunities, and challenges[J]. Big Data and Cognitive Computing, 2025, 9(4): 87. DOI:10.3390/bdcc9040087.
- [3] 张钦彤, 王昱超, 王鹤羲, 等. 大语言模型微调技术的研究综述[J]. 计算机工程与应用, 2024, 60(17): 17–33. DOI:10.3778/j.issn.1002-8331.2312-0035. [ZHANG Q T, WANG Y C, WANG H X, et al. Comprehensive review of large language model fine-tuning[J]. Computer Engineering and Applications, 2024, 60(17): 17–33. DOI:10.3778/j.issn.1002-8331.2312-0035.
- [4] KURATOV Y, BULATOV A, ANOKHIN P, et al. BABILong: testing the limits of LLMs with long context reasoning-in-a-haystack[EB/OL]. (2024-06-14) [2025-04-28]. <https://arxiv.org/abs/2406.10149>
- [5] XIE S M, SANTURKAR S, MA T Y, et al. Data selection for language models via importance resampling[EB/OL]. (2023-11-18) [2025-04-28]. <https://arxiv.org/abs/2302.03169>
- [6] LI R H, WANG Y, WEN Z J, et al. Different paths to the same destination: Diversifying LLMs generation for multi-hop open-domain question answering[J]. Knowledge-Based Systems, 2025, 309: 112789. DOI:10.1016/j.knosys.2024.112789.
- [7] 刘汉蒙, 张岳. 大模型逻辑推理研究综述[C]. 第二十三届中国计算语言学大会, 中国太原, 2024. <https://aclanthology.org/2024.ccl-2.3/>. [LIU H M, ZHANG Y. Survey on Logical Reasoning of Large Pre-trained Language Models[C]. Proceedings of the 23rd Chinese National Conference on Computational Linguistics, Taiyuan, China, 2024. <https://aclanthology.org/2024.ccl-2.3/>]
- [8] ZHANG J Z, YOU J C, ASHWINEE Panda, et al. LoRA Without Forgetting: Freezing and S

parse Masking for Low-Rank Adaptation[C]. Sparsity in LLMs (SLLM): Deep Dive into Mixture of Experts, Quantization, Hardware, and Inference, Los Angeles, 2025. <https://openreview.net/forum?id=aGOQYJfz6H>.

[9] ZHUANG Y M. Autoregressive Language Model with Historical Context Re-encoding[C]. ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, 2025. <https://ieeexplore.ieee.org/abstract/document/10890165>.

[10] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学(信息科学), 2023, 53(9): 1645–1687. DOI:10.1360/SSI-2023-0113. [CHE W X, DOU Z C, FENG Y S, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions[J]. Science in China (Information Sciences), 2023, 53(9): 1645 – 1687. DOI:10.1360/SSI-2023-0113.]

[11] ZHU D W, WEI X Y, ZHAO G X, et al. Chain of preference optimization: Improving chain-of-thought reasoning in llms[EB/OL]. (2024-06-13) [2025-04-28]. <https://doi.org/10.48550/arXiv.2502.20790>.

[12] HAN S M, YU A, SHEN R, et al. P-FOLIO: evaluating and improving logical reasoning with abundant human-written reasoning chains[EB/OL]. (2024-10-11) [2025-04-28]. <https://arxiv.org/abs/2410.09207>

论文引用格式

王强, 姜程林, 马宁, 等. MTR-QA: 多类型推理问答数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-02-24). DOI: 10.11922/11-6035.csd.2025.0062.zh.

数据引用格式

王强, 姜程林, 马宁, 等. MTR-QA: 多类型推理问答数据集[DS/OL]. V1. Science Data Bank, 2025. (2025-04-17). DOI: 10.57760/sciencedb.23774.

A dataset of multi-type reasoning question answering (MTR-QA)

WANG Qiang^{1,2}, JIANG Chenglin^{1,2}, MA Ning^{1,2*}, LI Yingjie^{1,2}, WU Wenshe^{1,2}

1. Key Laboratory of Linguistic and Cultural Computing Ministry of Education, Northwest Minzu University, Lanzhou 730030, P.R. China

2. Key Laboratory of Minzu Languages and Cultures Intelligent Information Processing, Gansu Province, Northwest Minzu University, Lanzhou 730030, P.R. China

*Email: maning@xbmu.edu.cn (MA Ning)

Abstract: Considering the problems that large language models are prone to experiencing logical confusion and insufficient ability to capture of implicit relationships when dealing with complex reasoning tasks, this paper proposes and constructs a high-quality dataset of multi-type reasoning question answering dataset (MTR-QA). By collecting and sorting the past 15 years of civil service examination questions and

authoritative mock question banks, a multi-stage text processing framework including text standardisation, hash deduplication, near-duplicate detection and semantic embedding similarity filtering was used to achieve data cleaning, which effectively reduces redundancy and noise interference. To ensure the reliability of the data, this paper designs a multi-model evaluation mechanism integrating GPT-4, DeepSeek-V1 and Qwen-2.5, which quantitatively evaluates the data across four dimensions: integrity (CPL), accuracy (ACC), security (SFC) and chain of thought quality (CoT-Q). In the end, 24,312 high-quality data entries were selected and stored in json format with a size of 34.1 MB. Each data sample contains six attributes: question, options, answer, chain of thought, type, and level of difficulty, and is divided into four core types of reasoning: logic, semantics, mathematics and comprehensive knowledge reasoning. The MTR-QA dataset has been expanded in terms of reasoning types and topic breadth, providing a reliable data base for various reasoning tasks such as pre-training large language models, supervised fine-tuning and model evaluation, and promote the performance improvement of large language models in complex reasoning and Q&A scenarios.

Keywords: large language models; multi-type; reasoning question-answer; datasets

Dataset Profile

Title	A dataset of multi-type reasoning question answering (MTR-QA)
Data corresponding author	MA Ning (maning@xbmu.edu.cn)
Data authors	WANG Qiang, JIANG Chenglin, MA Ning, LI Yingjie, WU Wenshe
Data volume	34.1 MB
Data format	*.json
Data service system	https://doi.org/10.57760/sciencedb.23774
Sources of funding	National Natural Science Foundation (62466052); Gansu Province Central Guidance Local Science and Technology Development Fund Project (25ZYJA034); Fundamental Research Funds for the Central Universities of Northwest Minzu University (31920250008)
Dataset composition	The dataset includes 24,312 json-formatted data, with each json object containing six attributes: instruction, question, answer, target, label and difficulty.