专题 大模型训练数据

MDD-TD：大语言模型文本来源及内容真伪检测数据集

李硕[1,2]，马宁[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： 文本检测是自然语言处理领域的重要研究方向，旨在判别文本的来源和内容真伪，帮助人们在繁杂的网络信息中获取可信度更高的内容。针对文本检测数据集稀缺问题，本研究构建了一个覆盖多领域的文本检测数据集 MDD-TD（Multi-Domain Text Detection Dataset），数据集包括翻译优化的开源数据、通过爬虫爬取的网络数据以及基于提示增强策略生成的合成数据。翻译数据基于 SimpleAI/HC3 数据集中的问答语料进行翻译和优化；网络开源语料通过爬取微博及豆瓣数据获得；合成语料借鉴规则驱动合成思路，基于已有翻译数据和网络开源数据，利用多种提示增强策略生成。为确保数据质量，本研究通过 PPL 方法剔除困惑度异常文本，利用语义相似度去重并结合人工审核对低质量数据进行二次过滤。最终筛选出 31,721 条高质量数据，以 json 格式存储。每条数据有 3 个标签，其中 text 对应待检测文本，source 对应数据来源，label 对应数据真伪。使用本数据集在两个公开检测器上进行了评估实验，来源检测任务中评论领域的 accuracy 为 97.13%，真伪检测任务中新闻领域的 accuracy 为 98.87%，表明本数据集可以作为大模型文本检测任务的训练和评估资源。

关键词： 大语言模型；文本来源；真伪检测；数据集；数据合成

www.sciengine.com/CSD
ISSN 2096-2223
CN 11-6035/N

文献 CSTR：
32001.14.11-6035.csd.2025.0191.zh

文献 DOI：
10.11922/11-6035.csd.2025.0191.zh

数据 DOI：
10.57760/sciencedb.29741

文献分类：计算机科学技术

收稿日期：2025-10-23
录用日期：2026-02-25
发表日期：2026-03-16

* 论文通信作者
马宁：maning@xbmu.edu.cn

数据库（集）基本信息简介

数据库（集）名称	MDD-TD：大语言模型文本来源及内容真伪检测数据集
数据通信作者	马宁（maning@xbmu.edu.cn）
数据作者	李硕、马宁
数据量	37.57 MB
数据格式	*.json
数据服务系统网址	https://doi.org/10.57760/sciencedb.29741
基金项目	国家社会科学基金（25XYY015）
数据库（集）组成	数据集包括 31,721 条 json 格式的数据，每一条 json 对象包含 3 个标签，分别为 text、source 和 label。

引 言

大语言模型（Large Language Model，LLM）可以快速理解和生成各种复杂内容，正在以前所未有的速度改变着文本生成与信息传播方式[1]。随着自然语言生成（Natural Language Generation, NLG）技术的发展，大语言模型生成内容日趋丰

富且质量更高[2]。以 ChatGPT 为代表的大语言模型能够很好地完成作答、撰写邮件、编写代码等任务。但是，这也可能导致大语言模型在网络诈骗、虚假信息传播以及学术不端等不良用途上被滥用[3]。

准确高效的文本检测能力能够更好地释放大语言模型的潜力，同时降低潜在风险与严重后果发生的可能性[4]。当前文本检测研究主要聚焦两类任务：一类是文本来源检测，判别文本是由人撰写还是由模型生成的[5]；另一类是内容真伪检测，检验文本内容是真还是假[6]。在跨领域和跨任务场景下，这两类检测任务都面临模型鲁棒性较差、检测效果受限等挑战，可能会导致信息遗漏或信息误判现象，降低信息可信度[7]。因此，为使文本检测模型能够适应各种复杂场景需求，构建覆盖多场景的高质量数据集就显得尤为重要。现有数据集，如 Ignat 等人提出的 MAiDE-up 数据集[8]、Cornelius 等人公开的 BUST 数据集[9]以及 Qazi、Shiao 和 Papalexakis 发布的 GRiD 数据集[10]均是以文本来源检测任务为目的；Azizov 等人提出的 SAFARI 数据集[11]、Kanaani 等人发布的 Triple-R 数据集[12]则专注于内容真伪检测任务。这些文本检测数据集虽然考虑到了多领域场景，但是仅聚焦于单一检测任务，缺少一个能够覆盖多领域、支持检测模型完成来源和真伪两类文本分类任务的检测数据集[13]。

本研究构建的 MDD-TD 数据集涵盖问答、评论和新闻文本 3 个领域，通过筛选、清洗、标注和质量控制等流程的数据处理，能够作为大模型进行文本来源及内容真伪检测两类任务的训练和评估资源，提升大模型的文本检测和细粒度评估任务能力，减少模型输出中潜在的有害或误导信息。

1 数据采集和处理方法

1.1 数据采集

数据来源于 3 部分：1）经过翻译优化的开源数据；2）网络爬取的开源数据；3）基于提示增强策略的合成数据，涉及问答、评论和新闻领域。

1.1.1 开源数据翻译

翻译语料主要基于 SimpleAI/HC3 数据集[14]。该数据集是 Hello-SimpleAI 团队创建的人类—ChatGPT 对比语料库。

本研究首先根据来源可靠性对原始语料进行初步筛选，保留来源标注清晰、可信度高的数据作为基础数据。其次，利用事实核查工具和人工审核，对数据集进行校验，剔除有事实性错误的问答对。使用谷歌搜索对问答对涉及的时间、地点和人物等关键信息进行检索验证；采用交叉验证策略，组织两名审核人员进行人工复核，重点检查问答对的语义完整性和逻辑性。通过调用大模型 API 并结合翻译软件，对问答对进行翻译和人工校正，将其转换为目标语言。翻译过程中，最大程度确保数据的准确性和上下文逻辑的连贯性。对于校正后的语料，剔除原始问题和 ChatGPT 回答，仅保留人类答案部分，组成问答领域的真实人类撰写数据，共形成 4,000 条语料。

1.1.2 网络爬取开源数据

评论和新闻文本的原始语料来源于豆瓣网（https://www.douban.com）和微博（https://www.weibo.com）。本研究通过爬虫技术获取豆瓣网中的帖子和评论，以及微博平台上“人民日报”“央视新闻”发布的新闻文本，并进行事实验证及人工核验[15]。删除非事实性文本，确保数据真实性。经过上述步骤，获得真实的人工撰写评论及新闻文本。

1.1.3 提示增强合成数据

本研究借鉴规则驱动的数据合成方法，基于获取的真实人工撰写文本，使用多个不同的大语言模型，结合多种提示策略进行批量化生产。数据合成流程如图 1 所示：

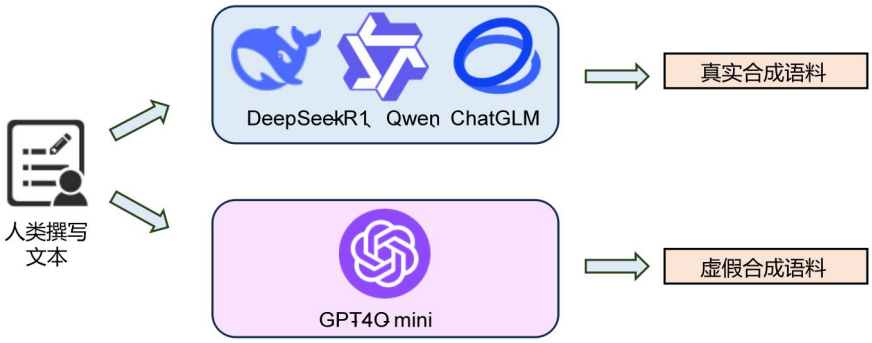


图 1　数据合成流程图

Figure 1　Data synthesis workflow

具体步骤为：（1）以人工撰写的真实文本为数据源作为初始指令输入。（2）使用 Qwen3-8b、ChatGLM3-6b、DeepSeek-R1 三个大模型，结合“直接提示”“改写提示”两种提示策略，驱动模型生成不同风格、不同结构的真实合成语料。（3）使用 GPT-4o-mini 模型，同样结合上述两种提示策略进行虚假语料合成。

最后，结合事实核查工具和人工审核，确保合成语料可靠性。在人工审核阶段，针对真实语料和虚假语料分别设定审核标准。对于真实语料，重点核查其真实性，确保模型在生成数据时没有引入事实性错误；对于虚假语料，则重点核查其逻辑合理性和虚假信息特征，确保其包含虚假信息又符合一定的语言逻辑。以上步骤均由两名审核人员完成，完成后进行一致性检验。如果有分歧，则引入第三位专家进行仲裁。基于上述人工审核方法，剔除存在事实性错误的真实语料和存在严重逻辑漏洞的虚假语料。

具体的 prompt 模板示例如表 1 所示：

表 1　prompt 模板示例

Table 1　Prompt template examples

策略/模板	真实合成语料	虚假合成语料
直接提示	请以“question”标签后面的问题为依据，为其生成一个符合客观事实、逻辑清晰的真实答案	请以“question”标签后面的问题为依据，为其生成一个不符合客观事实且有一定语言逻辑的虚假答案。
改写提示	请把“text”标签后面的评论/新闻数据进行改写，生成一个符合客观事实、逻辑清晰的改写评论/新闻	请把“text”标签后面的评论/新闻数据进行改写，生成一个不符合客观事实且有一定语言逻辑的虚假的改写评论/新闻。

1.2 数据处理

针对原始数据，先剔除字段缺失及长度过短的数据，因为这类数据很难提供有价值的信息；再对清洗后的数据去掉乱码和无效字符；随后进行数据格式化、结构标准化操作[16]；最后以 json 格式

存储。将格式化后的数据按照来源（人/Model）和真伪进行分类及标注，每条数据包含 text、source 和 label 共 3 个标签。为明确“真”和“假”的标注逻辑，保证 label 标签的准确性，本研究设置的真假判别标准为：1）文本内容在语义逻辑和事实性方面均是真实的则标注为“real”；2）文本内容包含明显的事实性和逻辑性错误则标注为“fake”。

最后，对标注完成的样本进行人工审核，着重检查 label 标签的准确性和 text 文本的完整性，保证 label 标签与其对应的数据内容相符，并核查是否依旧存在严重噪声样本。如果有，则将其剔除。数据采集和处理流程如图 2 所示。

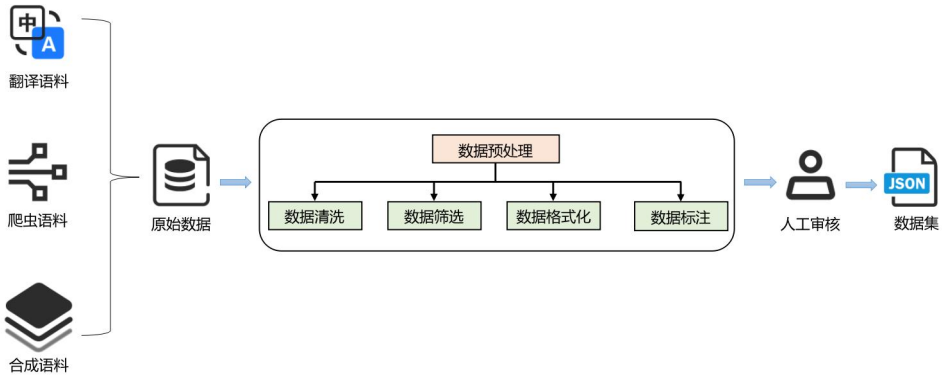


图 2　数据采集和处理流程

Figure 2　Data collection and processing workflow

2　数据样本描述

MDD-TD 数据集包含 31,721 条数据，其中，问答领域数据 10,729 条、评论领域数据 10,552 条、新闻文本领域数据 10,440 条，大小为 37.57 MB，用 json 格式存储。具体数据分布如表 2 所示。

表 2　各领域数据详情

Table 2　Dataset details by domain

分类	人类撰写数据（真）	AI 生成数据		总条数
		真	假	
问答	4,000	3,624	3,105	10,729
评论	4,000	3,342	3,210	10,552
新闻文本	4,000	3,330	3,110	10,440

每个 json 对象有 3 个标签，即 text、source 和 label，分别对应待检测文本、数据来源及数据真伪。本数据集可用于文本来源（人类/Model）检测、人类数据和模型生成数据混合情况下的真伪检测以及仅基于模型生成数据的真伪检测。需要说明的是，source 标签用于标识文本来源（Human 或 Model），label 标签用于标识文本内容真伪（real 或 fake），二者属于不同维度的标注信息。数据集样本示例如表 3 所示。

表 3　新闻与评论数据集样本示例

Table 3　Samples from news and commentary datasets

标签名称	新闻数据示例	评论数据示例
“text”	5 月 6 日，外交部发言人林剑主持例行记者会。有记者就中美是否就关税问题谈判提问。林剑表示，事实上，这场关税战是美方挑起的，中方的态度是一贯的、明确的：打，奉陪到底，谈，大门敞开。美方近期不断表示，希望同中方进行谈判。关税战、贸易战没有赢家。如果美方真想通过对话谈判解决问题，就应该停止威胁施压，在平等、尊重、互惠的基础上，同中方开展对话。	确实，对于长期存在的经典作品来说，编剧的更迭往往会带来风格和内容上的变化。这种变化有时是渐进的，有时却可能导致作品本质的巨大转变。我认为，这样的现象在一定程度上体现了创作自由与作品连续性之间的矛盾。首先，不同的编剧有着不同的创作背景、风格和理念，这自然会反映在他们的作品中。一些编剧可能更注重故事的深度和复杂性，而另一些则可能更倾向于动作和视觉冲击。这种多样性虽然为作品注入了新的活力，但也可能导致粉丝对作品原有的期待与实际内容产生偏差。其次，粉丝对作品的忠诚度和期待感往往来源于早期作品所塑造的世界观和人物设定。当后续编剧对这些元素进行改动时，粉丝可能会感到困惑甚至不满。正如您提到的，后面的编剧受到了前面作品的影响，这种传承有时会带来一定的相似性，但过度追求创新也可能导致作品失去原有的味道。
“source”	Human	Model
“label”	real	real

3　数据质量控制和评估

3.1　数据质量控制

为确保数据质量[17]，本研究采用困惑度（PPL）对数据进行初步识别，剔除困惑度异常样本。使用 GPT-2-Chinese 模型作为基座模型，通过对语料的 PPL 值进行分布统计，按照百分比筛选，从前 10%和后 10%的位置截断，仅保留中间部分。前 10%表示困惑度过低，后 10%表示困惑度过高，均为困惑度异常样本。困惑度计算公式如下：

$$PPL = e^{-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1,w_2,\ldots,w_{i-1})} \quad (1)$$

公式中，N表示每条语料中的 token 总数，w_i表示第i个词，$P(w_i|w_1,w_2\ldots,w_{i-1})$表示模型预测第$i$个词的概率。

其次，使用余弦相似度方法去除语义表述高度相似的文本，确保数据多样性。本研究使用 RoBERTa-Chinese 模型作为嵌入模型，采用边界抽样方法，将余弦相似度大于 0.85 的样本判定为语义表述高度相似文本，并进行剔除。余弦相似度计算公式如下：

$$conine_sim(u,v) = \frac{u\cdot v}{\|u\|\,\|v\|} \quad (2)$$

其中，$u=\{u_1,u_2,\ldots,u_n\}$、$v=\{v_1,v_2,\ldots,v_n\}$分别代表两个向量，$u\cdot v$为向量点积，$\|u\|$、$\|v\|$分别表示两个向量的模长。计算公式如下：

$$u\cdot v = \sum_{i=1}^{n} u_i v_i \quad (3)$$

$$\|u\| = \sqrt{\sum_{i=1}^{n} u_i^2},\ \|v\| = \sqrt{\sum_{i=1}^{n} v_i^2} \quad (4)$$

3.2 数据评估

数据评估包括两项任务：文本来源检测评估和内容真伪检测评估，分别验证数据集在来源检测和内容真伪检测任务中的有效性。

3.2.1 文本来源检测评估

文本来源检测评估使用 Detection in the Wild 检测器[18]。该检测器以参数规模为 1.49 亿的 Longformer 模型作为基座模型，使用大量不同领域的文本数据模拟各种检测场景，学习人类撰写文本和大模型生成文本的特征差异，实现对文本来源的判别。本实验基于 Detection in the Wild 检测器中的“domain_specific_cross_models”检测平台，与 HC3-Chinese 数据集[19]、YF_dianping 数据集、DeepSeek 改写数据集进行对比实验，并在相同测试集上评估准确率（Accuracy）。训练集和测试集的比例为 8：2，测试集在训练之前通过随机抽样得到，与训练集相互独立。实验结果如表 4 所示，使用 MDD-TD 数据集训练后的模型准确率均高于使用对比数据集的模型，表明 MDD-TD 数据集在支持模型文本来源检测方面更具优势。

表 4　文本来源检测评估对比表

Table 4　Comparison of text source detection performance

领域/指标	Accuracy（%）	
问答	HC3-Chinese	57.63
	MDD-TD	**96.16**
评论	YF_dianping	63.47
	MDD-TD	**97.13**
新闻文本	DeepSeek 改写	63.65
	MDD-TD	**91.86**

3.2.2 内容真伪检测评估

内容真伪检测评估采用 SheepDog 检测器[20]，该检测器以 1.25 亿参数的 ReBERTa-base 模型作为基座模型，通过“风格无关训练”以及“内容的真实性归因”，降低了风格对虚假信息检测的影响，提高了检测准确性。本实验基于 SheepDog 检测器，与 LTCR 数据集[21]、MCFEND 数据集[22]以及 Gossipcop 数据集[23]进行对比实验，并在相同测试集上进行评估。训练集和测试集的划分比例为 8：2，测试集在训练之前通过随机抽样得到，与训练集相互独立。实验结果如表 5 所示，使用 MDD-TD 数据集训练后的模型准确率均高于使用对比数据集的模型，表明 MDD-TD 数据集在支持模型内容真伪检测方面也具有一定优势。

表 5　内容真伪检测评估对比表

Table 5　Comparison of content authenticity detection performance

领域/指标	Accuracy（%）	
问答	LTCR	43.10
	MDD-TD#	96.58

领域/指标	Accuracy（%）	
	MDD-TD*	**97.95**
评论	MCFEND	92.25
	MDD-TD#	**98.66**
	MDD-TD*	98.45
新闻文本	Gossipcop	75.80
	MDD-TD#	95.40
	MDD-TD*	**98.87**

注：MDD-TD#表示仅使用模型生成文本进行训练和评估，MDD-TD*表示同时使用人类撰写文本及模型生成文本进行训练和评估。

4 数据价值

本数据集包括问答、评论和新闻 3 个领域，面向文本来源检测、内容真伪检测两类检测任务，提供包含实体上下文、文本来源及内容真伪等不同粒度语义信息的文本检测语料，为文本检测模型提供必要的多领域信息。现有文本检测数据集较为稀缺，限制了检测模型的性能，导致模型在实际应用中的准确性和鲁棒性较低，尤其是在跨领域和跨任务场景。本数据集的构建则有助于弥补这一不足，为不同场景的文本检测研究提供更加优质的训练和评估资源，从而提升检测模型的性能，促进文本检测技术的发展。科研工作人员还可将本数据集用于文本检测任务研究。

5 AIGC 使用声明

本研究在准备阶段使用了人工智能生成内容（AIGC）工具。具体情况如下：

工具与版本：我们使用了 Qwen3-8b 模型、ChatGLM3-6b 模型、DeepSeek-R1 模型、GPT-4o-mini 模型、GPT-2-Chinese 模型、RoBERTa-Chinese 模型以及 GPT-4 模型。

数据集构建：使用 Qwen3-8b 模型、ChatGLM3-6b 模型、DeepSeek-R1 模型和 GPT-4o-mini 模型进行数据合成，得到真实和虚假的模型生成语料。对生成内容设立了相关的核验标准，并经过严格的人工审核和校准，剔除不符合标准的内容。

数据质量控制：使用 GPT-2-Chinese 模型和 RoBERTa-Chinese 模型对数据进行困惑度和语义相似度计算。以此为依据，使用研究团队制定的审核标准，剔除异常文本。最终结果均经过人工复核确认。

论文撰写：使用 GPT-4 对部分段落的内容做了辅助润色和语法校对。所有由 AIGC 工具生成的内容均经过作者的严格修改和事实核查，未直接使用。

责任声明：作者对所有 AIGC 工具辅助生成的内容进行了仔细审查和编辑，并对论文的整体内容、数据真实性及学术诚信负全部责任。AIGC 工具未被列为作者。

致　谢

本数据集在构建过程中使用的开源数据有：开源数据集 SimpleAI/HC3、豆瓣网及微博等公开网

站的数据。使用了 Qwen3-8b、ChatGLM3-6b、DeepSeek-R1 及 GPT-4o-mini 等大模型进行数据合成。特别感谢相关数据集的工作人员及大模型团队的开放服务。

数据作者分工职责

李硕（1999—），男，山东潍坊人，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。
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。

参考文献

[1] BOUTADJINE A, HARRAG F, SHAALAN K. Human vs. machine: a comparative study on the detection of AI-generated content[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2025, 24(2): 1–26. DOI:10.1145/3708889.

[2] ZHANG Y H, MA Y Q, LIU J W, et al. Detection vs. anti-detection: is text generated by AI detectable? [M]//Wisdom, Well-Being, Win-Win. Cham: Springer Nature Switzerland, 2024: 209–222. DOI:10.1007/978-3-031-57850-2_16.

[3] ROSCA C M, STANCU A, IOVANOVICI E M. The new paradigm of deepfake detection at the text level[J]. Applied Sciences, 2025, 15(5): 2560. DOI:10.3390/app15052560.

[4] YU X, CHEN K J, YANG Q, et al. Text fluoroscopy: detecting LLM-generated text through intrinsic features[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA. Stroudsburg, PA, USA: ACL, 2024: 15838-15846. DOI:10.18653/v1/2024.emnlp-main.885.

[5] WU J C, YANG S, ZHAN R Z, et al. A survey on LLM-generated text detection: necessity, methods, and future directions[J]. Computational Linguistics, 2025, 51(1): 275–338. DOI:10.1162/coli_a_00549.

[6] 张廷, 袁虎, 赵小兵. 基于优先融合与模态注意力机制的虚假新闻检测[J]. 计算机应用研究, 2025, 42(5): 1392-1400. DOI: 10.19734/j.issn.1001-3695.2024.10.0422. [ZHANG T, YUAN H, ZHAO X B. Preferential fusion with modality-wise attention mechanisms for fake news detection[J]. Application Research of Computers, 2025, 42(5): 1392-1400. DOI: 10.19734/j.issn.1001-3695.2024.10.0422.]

[7] FISHCHUK V, BRAUN D. Robustness of generative AI detection: adversarial attacks on black-box neural text detectors[J]. International Journal of Speech Technology, 2024, 27(4): 861–874. DOI:10.1007/s10772-024-10144-2.

[8] IGNAT O, XU X M, MIHALCEA R. MAiDE-up: multilingual deception detection of AI-generated hotel reviews[C]//Findings of the Association for Computational Linguistics: NAACL 2025. Albuquerque, New Mexico. Stroudsburg, PA, USA: ACL, 2025: 1636-1653. DOI:10.18653/v1/2025.findings-naacl.88.

[9] CORNELIUS J, LITHGOW-SERRANO O, MITROVIC S, et al. BUST: Benchmark for the evaluation of detectors of LLM-Generated Text[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Mexico City, Mexico. Stroudsburg, PA, USA: ACL, 2024: 8029-8057. DOI:10.18653/v1/2024.naacl-long.444.

[10] QAZI Z, SHIAO W, PAPALEXAKIS E E. GPT-generated text detection: benchmark dataset and tensor-based detection method[C]//Proceedings of the ACM Web Conference 2024. Singapore, Singapore. ACM, 2024: 842–846. DOI:10.1145/3589335.3651513.

[11] AZIZOV D, MUJAHID Z M, ALQUABEH H, et al. SAFARI: cross-lingual bias and factuality detection in news media and news articles[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA. Stroudsburg, PA, USA: ACL, 2024: 12217-12231. DOI:10.18653/v1/2024.findings-emnlp.712.

[12] KANAANI M, DADKHAH S, GHORBANI A A. Triple-R: automatic reasoning for fact verification using language models[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 16831-16840. https://aclanthology.org/2024.lrec-main.1463/.

[13] KWON S, JANG B. A comprehensive survey of fake text detection on misinformation and LM-generated texts[J]. IEEE Access, 2025, 13: 25301–25324. DOI:10.1109/ACCESS.2025.3538805.

[14] SU Z P, WU X, ZHOU W, et al. HC3 plus: a semantic-invariant human ChatGPT comparison corpus[J]. ArXiv e-Prints, 2023: arXiv: 2309.02731. DOI:10.48550/arXiv.2309.02731.

[15] 赵小兵, 尹召宁, 王子豪, 等. 面向社会媒体的立场检测研究综述[J]. 计算机应用研究, 2024, 41(11): 3201-3214. DOI: 10.19734/j.issn.1001-3695.2024.01.0043. [ZHAO X B, YIN Z N, WANG Z H, et al. Review of stance detection for social media[J]. Application Research of Computers, 2024, 41(11): 3201–3214. DOI: 10.19734/j.issn.1001-3695.2024.01.0043.]

[16] 孙媛, 刘思思, 陈超凡, 等. 面向机器阅读理解的高质量藏语数据集构建[J]. 中文信息学报, 2024, 38(3): 56-64. DOI: 10.3969/j.issn.1003-0077.2024.03.007. [SUN Y, LIU S S, CHEN C F, et al. Construction of a High-Quality Tibetan Dataset for Machine Reading Comprehension[J]. Journal of Chinese Information Processing, 2024, 38(3): 56–64. DOI: 10.3969/j.issn.1003-0077.2024.03.007.]

[17] 马宁,吕 文蓉, 郭泽晨. 面向机器阅读理解的医学域数据集 MedicalQA[J/OL]. 中国科学数据, 2024, 9(1): 361-370. DOI:10.11922/11-6035.csd.2022.0030.zh. [MA N, LV W R, GUO Z C. MedicalQA: a dataset of medical domain for machine reading comprehension[J/OL]. China Scientific Data, 2024, 9(1): 361–370. DOI: 10.11922/11-6035.csd.2022.0030.zh.]

[18] LI Y F, LI Q T, CUI L Y, et al. MAGE: machine-generated text detection in the wild[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand. Stroudsburg, PA, USA: ACL, 2024: 36-53. DOI:10.18653/v1/2024.acl-long.3.

[19] GUO B Y, ZHANG X, WANG Z Y, et al. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection[EB/OL]. arXiv:2301.07597, 2023. https://arxiv.org/abs/2301.07597.

[20] WU J Y, GUO J F, HOOI B. Fake news in sheep’s clothing: robust fake news detection against LLM-empowered style attacks[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Barcelona Spain. ACM, 2024: 3367-3378. DOI:10.1145/3637528.3671977.

[21] MA Z Y, LIU M S, FANG G A, et al. LTCR: Long-Text Chinese Rumor detection dataset[J]. Neural Computing and Applications, 2025, 37(20): 15133–15152. DOI:10.1007/s00521-025-11260-z.

[22] LI Y P, HE H R, BAI J, et al. MCFEND: a multi-source benchmark dataset for Chinese fake news detection[C]//Proceedings of the ACM Web Conference 2024. ACM, 2024: 4018 – 4027. DOI:10.1145/3589334.3645385.

[23] SHU K, MAHUDESWARAN D, WANG S H, et al. FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media[J]. Big Data, 2020, 8(3): 171–188. DOI:10.1089/big.2020.0062.

论文引用格式

李硕，马宁. MDD-TD：大语言模型文本来源及内容真伪检测数据集[J/OL]. 中国科学数据, 2026, 11(1). (2026-03-16). DOI: 10.11922/11-6035.csd.2025.0191.zh.

数据引用格式

李硕，马宁. MDD-TD：大语言模型文本来源及内容真伪检测数据集[DS/OL]. V1. Science Data Bank, 2025. (2025-10-17). DOI:10.57760/sciencedb.29741.

MDD-TD: A dataset of large language model text source and content authenticity detection

LI Shuo[1,2], MA Ning[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: Text detection is a significant research area in natural language processing, aiming to identify the source and authenticity of text content to help users access more credible information amidst the vast amount of online data. To address the scarcity of text detection datasets, this study constructs a Multi-Domain Text Detection Dataset (MDD-TD). The dataset comprises translation-optimized open-source data, web data collected through web crawling, and synthetic data generated using prompt-augmentation strategies. Translation data are obtained by translating and refining the question–answer corpus from the SimpleAI/HC3 dataset. Web-sourced data are collected from the Weibo

and Douban platforms through web crawling. Synthetic data are generated following a rule-driven approach, leveraging existing translation and web-sourced data with multiple prompt enhancement strategies. To ensure data quality, this study employs the PPL method to remove texts with abnormal perplexity values. Semantic similarity‐based deduplication is further applied further used, combined with manual review to conduct secondary filtering of low-quality samples. Ultimately, 31,721 high-quality samples are retained and stored in JSON format. Each sample contains three labels: “text”, “source”, and “label”, where “text” represents the text to be detected, “source” indicates the data origin, and “label” denotes the authenticity of the text. Evaluation experiments are conducted using two publicly available detection models, achieving an accuracy of 97.13% in the review domain for source detection and 98.87% in the news domain for authenticity detection. These results demonstrate that the dataset can serve as an effective resource for training and evaluating large language models in text detection tasks.

Keywords: large language models; text source; authenticity verification; dataset; data synthesis

Dataset Profile

Title	MDD-TD: A dataset of large language model text source and content authenticity detection
Data corresponding author	MA Ning(maning@xbmu.edu.cn)
Data author(s)	LI Shuo, MA Ning
Data volume	37.57 MB
Data format	*.json
Data service system	https://doi.org/10.57760/sciencedb.29741
Sources of funding	National Social Science Fund of China (Grant No.25XYY015)
Dataset composition	The dataset includes 31,721 entries in json format, and each json object contains three labels: “text,” “source,” and “label.”