

DynaSynth: 面向非结构化网络安全语料的大语言模型微调数据动态生成代理工具

关永健, 朴乘锴, 王布宏, 赵博夫, 李思琦, 赵正阳

(空军工程大学信息与导航学院, 陕西 西安 710077)

摘要: 随着模型训练技术和数据规模的迅速发展, 大语言模型已经能够胜任各种通用任务, 但在网络安全垂直领域, 大语言模型的发展仍受数据瓶颈的制约。为解决这一难题, 提出了面向非结构化语料的大语言模型微调数据动态生成代理工具 DynaSynth。该工具首先利用向量嵌入辅助模型将原始文本切割为文本块, 并转化为可检索增强的向量数据; 然后通过对文档进行受众一体裁分析, 指导生成具有多样化问答对风格且兼具通用性的数据。在此过程中, 用户可以通过可视化界面实时查看生成的数据, 并借助提示工程对生成策略进行优化调整, 确保生成内容精准契合实际需求。实验结果表明, 使用 DynaSynth 生成的微调数据能够显著提高大语言模型在垂直领域任务中的性能。该工具不仅具备高效的数据生成能力, 还能够有效探索数据的多样化生成路径。

关键词: 大语言模型代理; 生成数据; 数据增强

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.25249

DynaSynth: a large language model fine-tuning data dynamic synthesis agent tool for unstructured cybersecurity corpus

Guan Yongjian, Piao Chengkai, Wang Buhong, Zhao Bofu, Li Siqi, Zhao Zhengyang

Information and Navigation College, Air Force Engineering University, Xi'an 710077, China

Abstract: With the rapid development of model training technologies and data scale, large language models have been able to handle various general tasks. However, in the vertical field of cybersecurity, the development of large language models is still constrained by data bottlenecks. To solve this problem, a data dynamic generation agent tool named DynaSynth for large language models based on unstructured text corpora was proposed. This tool first uses vector embeddings to assist the model in splitting the original text into blocks and converting them into retrievable enhanced vector data; then, through audience and genre analysis of the documents, it guides the generation of data with diverse question-answer pair styles and universality. During this process, users can view the generated data in real time through a visual interface and optimize the generation strategy by leveraging prompt engineering to ensure that the generated content precisely meets the actual needs. Experimental results show that the fine-tuning data generated by DynaSynth can significantly improve the performance of large language models in vertical domain tasks. This tool not only has an efficient data generation capability, but also effectively explores the diverse generation paths of data.

Key words: large language model agent, synthesis data, data augmentation

收稿日期: 2025-12-27; 修回日期: 2026-03-02

通信作者: 王布宏, wbhgroup@aliyun.com

基金项目: 国家自然科学基金资助项目 (No. 62472437)

Foundation Item: The National Natural Science Foundation of China (No. 62472437)

0 引言

大语言模型和其他基础模型的能力近年来实现爆发式增长,其应用场景已从自然语言文本的理解与生成,逐步扩展到代码编写、逻辑推理、语际翻译等^[1]。当模型的参数规模与训练数据量级突破特定阈值后,涌现能力逐渐显现(即模型在未经过明确训练的情况下自发展现出新的能力,如链式思维推理、指令遵循与情境学习等^[2])。当前,模型的参数量已从十亿级别跃升至万亿级别,训练数据也从单一的纯文本扩展至丰富多样的多模态内容。这一演变不仅提升了模型在传统自然语言处理任务上的性能,还催生了智能体、Copilot助手等新型应用形态。与此同时,这些模型正不断向科学研究、内容创作、教育及医疗等领域渗透^[3]。

然而,大语言模型在展现出强大能力的同时,也面临着由静态知识依赖引发的幻觉现象、复杂计算能力不足及实时交互缺失等问题。在此背景下,代理工具应运而生。早期的代理工具主要通过应用程序接口(API)调用等方式赋予大语言模型基础的工具使用能力,并以插件形式动态解决模型在知识实时性不足及专业领域知识局限方面的问题。随着架构的演进,代理工具的范式已从面向单一任务的问题求解,逐步演进为对特定工具或工具集合的迭代式学习,并借助反馈循环优化工具调用策略,提升复杂问题求解的鲁棒性^[4]。

在代理工具的研究与应用中,除阐明代理工具的理论意义外,还需分析其在实现层面的“实践路径”。大模型代理工具的范式由早期的单轮任务求解模式逐步演进为迭代任务求解模式。前者由大模型一次性完成所有子任务的求解与工具调用,整个过程缺乏反馈迭代机制;后者则能够根据工具反馈动态修正任务规划,从而强化学习能力。在具体实践中,代理工具的实现通常遵循以下4阶段框架。

1) 第一阶段:任务规划。该阶段的核心在于解析用户的复杂语义意图,对任务进行拆解,并规划子任务的执行顺序。通过设计规范化的提示工程指令模板,无须微调模型参数,即可激发大语言模型的固有智能,实现任务分解^[5]。

2) 第二阶段:工具选择。该阶段的核心是根据各项子任务的需求,从工具库中筛选出最优的代理工具。具体而言,主要通过术语匹配、查询语句与工具描述的语义相似度计算等方式完成工具的筛选。

3) 第三阶段:工具调用。该阶段主要负责从用户查询中解析出符合API规范的相关参数,完成参数类型、取值范围等的格式校验,确保参数合规后执行工具调用操作。

4) 第四阶段:响应生成。该阶段主要是对工具返回的原始结果进行整合处理,最终生成用户友好型的自然语言回复。在此过程中,需将工具输出的原始数据转化为自然语言,同时结合大模型的内部知识库进行补充解释^[6];此外,还需对冗长、含噪的无效信息进行过滤,对长文本进行信息整合与压缩。

当前,数据稀缺问题正严重制约人工智能(AI)在垂直领域的深度应用。合成数据生成作为突破数据瓶颈的关键技术,已成为研究热点,但现有方法仍面临两大挑战:其一,数据异构性处理能力不足,传统生成模型难以同时适配规则化的结构化数据与非规则化的复杂数据;其二,生成-训练目标失配,通用生成策略忽视了模型训练流程的阶段性需求,导致生成数据与实际训练目标之间存在错位。为此,本文提出一种基于大语言模型的训练数据生成工具——多样化非结构化语料中微调数据的动态生成代理工具DynaSynth。DynaSynth集成了多种格式数据解析、检索增强生成(RAG)、体裁-受众对数据增强等功能,利用递归字符分割及向量数据库辅助嵌入模型实现快速检索;同时,本文还提出一种从多个维度生成多样化问答对的新方法,并配备可视化界面,支持对模型参数和生成器参数进行调整。此外,本文还提出一种多阶段适配训练过程生成机制,将输出数据动态划分为预训练数据集、指令微调数据集和推理优化数据集。预训练数据集包含海量通用领域样本,用于构建模型的基础知识表示;指令微调数据集包含任务指令-问答对,以增强模型的任务泛化性;推理优化数据集注入了高难度边缘案例,用于提升模型的推理鲁棒性。结合数据集的评估指标与基

于提示工程的反馈优化机制,对数据生成过程进行动态调整,从而保障生成数据的质量。该代理系统支持通过示例注入机制提高领域特定知识的利用效率,同时具备处理不同结构数据集的能力。

本文的主要贡献如下。

1) 提出面向非结构化文档的大模型微调数据生成代理工具 DynaSynth,该工具集成了多种类型的文件解析功能,能够实现从非结构化数据到高质量微调数据的自动化生成。

2) 引入体裁-受众对 (genre-audience pairs) 数据增强机制,从多个维度生成多样化问答对,有效解决数据风格单一问题,提升原始语料库的利用效率。

3) 提供提示工程反馈优化功能,从动态系统优化视角出发,结合历史响应动态提示词和 G-Eval 评分机制,实现对提示词的动态调整。

4) 提供可视化用户界面与灵活的模型配置功能,支持 API 调用与本地部署,同时具备交互优化机制,降低无技术背景用户的使用门槛,为监督式微调中的数据瓶颈问题提供新的解决方案。

1 相关工作

近年来,围绕生成数据的合成、多智能体协作范式及基于大语言模型的智能工具,学界和产业界开展了大量研究。Sun 等^[7]提出的基于语言模型的数据代理 (language model-based data agent, LAMBDA) 是一个多智能体系统,集成了代码生成、代码执行、结果管理及知识存储等功能模块。该代理工具通过结构化知识接口实现对专家知识的集成,统一知识库内容,并支持与知识库查询系统或基于大语言模型的智能体进行交互。在人机协同方面,LAMBDA 打破了传统 AI 辅助软件工具的被动响应特性,结合交互式学习与教学场景,构建了一种基于人机协同的学习范式,实现了智能辅助系统的双向赋能。Qiao 等^[8]提出的 TaskWeaver 是一个面向大模型驱动智能体的代码优先框架。该框架将代码作为智能体间通信的核心载体,所有组件均围绕代码构建,支持具有复杂逻辑的代数式代码。其核心理念是“代码优先”——可以根据用户需求动态生成并执行代码。在该框架中,代码解释器负责代码的

生成、执行与状态维护;智能体具备会话级状态,支持上下文持续。此外,该框架通过 ReAct 范式自动检测并修复代码执行中的问题。AutoGen^[9]是 Microsoft Research 开源的多智能体对话系统框架,它通过让多个大语言模型智能体协同对话,为开发者提供了更简单、更丰富的下一代 LLM 应用构建方式。该框架基于可对话智能体构建,支持大模型驱动、工具集成、人类参与及混合协作模式。AutoGen 采用对话编程范式,允许开发者通过自然语言和 Python 代码灵活定义智能体交互逻辑,并提供统一的对话接口、动态群聊管理、角色隔离、记忆管理等机制。AutoGen 原生支持人类介入机制,实现监督式 AI 协作,确保任务执行的可控性。此外,AutoGen 通过上下文动态更新机制,在代码生成场景中实现了对 GPT-4 训练数据范围外 API 的调用。ChatDev^[10]是由清华大学团队提出的基于大语言模型的多智能体协作软件开发框架。该框架利用 Chat Chain 机制将瀑布模型分解为设计、编码和测试 3 个阶段,每个阶段细分为若干子任务,由两个具有不同角色的智能体通过多轮对话协作完成,最终形成完整的解决方案。在代码生成环节,ChatDev 引入了角色反转机制,使辅助智能体主动向指导者请求细节信息,从而减少大模型生成的代码错误、缓解幻觉问题、提升代码的可执行性。此外,ChatDev 通过短期记忆与长期记忆的结合,在维护单阶段内对话连续性的同时,实现了跨阶段的上下文传递,解决了大模型上下文长度限制问题。

面对大模型微调数据集缺乏、人工构建大规模数据集效率低下的问题,近年来陆续出现了利用基于大模型的数据生成工具来构建微调数据集的解决方案。Canto 等^[11]提出的 Distilabel 是一款数据合成工具,它将强大的大语言模型作为“教师”与“裁判”,通过可配置、可扩展的自动化管道,将原始指令转化为经过筛选与评分的高质量问答对数据集。面向大语言模型的简易数据生成 EDG4LLM 是一款旨在简化并快速为大语言模型创建微调数据集的一体化工具,具有依赖项少、资源消耗少的特点,并提供清晰详细的日志输出,便于有效调试与问题追踪。然而,上述工

具普遍缺乏图形化界面,使得缺乏人工智能技术背景的使用者难以理清从数据读取到生成的完整业务流程,带来了较高的学习门槛与欠佳的用户体验。Kiln是一款开源的数据生成工具,能够简化大语言模型的微调、合成数据生成和数据集协作流程,其通过内置的工作流引擎集成大模型以生成数据。Graphgen是Chen等^[12]提出的数据生成框架,其核心功能是从原始文本中构建细粒度知识图谱,并基于该知识图谱指导问答对的生成,尤其适用于垂直领域高质量问答对的合成。然而,该工具在生成微调数据前无法直接读取未经处理的源文件,导致在构建领域专用数据集时效率有所降低。Data-Juicer^[13]是一款专注于数据清洗和转化的工具,该工具凭借模块化设计、支持多模态、高效的数据处理能力及数据-模型协同反馈闭环,为大模型研发者提供了强有力的数据处理支持。该工具整合了负责知识图谱数据生成的合成模型与能够识别自身知识盲点的学生模型,通过引导数据生成过程弥补知识盲点,进而生成具有更高价值的长尾知识数据。然而,上述工具普遍缺乏人在回路的交互方式,难以实现对生成数据的逐次优化。

综上所述,现有数据生成工具仍难以满足使用者尤其是无人工智能技术背景使用者的实际需求。因此,提出一种高效、易用的大模型微调数据生成方案,具有重要的现实意义与迫切性。

2 本文方法

本文提出的DynaSynth数据生成代理工具可以从结构化和非结构化文本资源中获取信息,并生成高质量微调数据集。DynaSynth的架构如图1所示,主要由文件解析模块、向量数据库、向量检索模块、数据生成模块、提示词修改模块、反馈优化模块和数据输出模块等组成。文件解析模块负责对多种格式文档进行加载与语义分割,并通过嵌入模型将文本块转化为向量,存储于向量数据库。向量检索模块结合用户请求与向量数据库实现语义检索。提示词修改模块通过可视化界面,支持用户对提示工程进行优化调整。数据输出模块支持多种格式训练数据的导出。该工具的工作流程如下:首先提取原始文档文本,采用基于长度拆分的策略对文本进行分块处理,

实现自适应文档处理与用户需求的融合;然后通过解析不同格式的文档文本,结合集成模型创建向量数据库;接着通过融合用户需求和代理本身的提示词修改功能,指导多样化问答对的创建。在问答生成阶段,DynaSynth可以生成样式多样且语义严谨的问答组合。此外,该工具的模块化模型配置接口允许用户灵活地调用API和本地部署的大语言模型,同时可以自由调节生成参数;配备多种默认提示词模板,适配不同大模型微调阶段的训练任务类型与训练数据需求,并支持多种格式训练数据的导出。

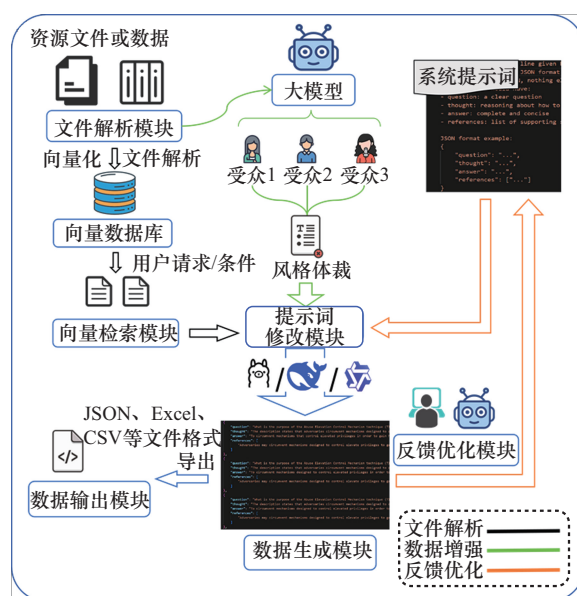


图1 DynaSynth的架构

2.1 文件解析

文件解析模块负责多种格式文件的加载、类型识别与语义分割。在文件类型预处理方面,该模块支持结构化(如CSV、Excel、DB等)、半结构化(如JSON、XML、日志等)及非结构化(如PDF、Word、PPT、图文等)等类型文件的自动识别与分析。其通过文件头特征与文件类型名相结合的方式类型判断,并能够区分扫描版与可编辑版PDF文件。在语义分割环节,该模块依据文档的章节、段落、图表等结构进行语义分片,同时提取关键实体、文本摘要及文档知识图谱。

文档加载器首先对不同格式的原生文档进行加载并转化为纯文本,然后通过字符递归文本分

割器将文本分割为语义较完整的中等长度文本块（不超过 2 048 字符，该长度可以适配大多数大模型的输入限制）。每个文本块与其前后块保留重叠区域，以确保上下文的连续性，避免关键信息丢失。字符递归文本分割器的工作原理如下：以预设最大长度为阈值，优先按段落、句子等语义边界进行递归分割；若无法分割，则按字符强制拆分。语义边界的识别通过标点符号、换行符及关键词实现。分割后的文本块通过预训练的嵌入模型转化为浮点向量，并持久化存储至自建的向量数据库中。该数据库具备高效的高维向量近似最近邻搜索能力，为后续的相似性查询提供了性能保障。

2.2 检索增强

DynaSynth 利用 RAG 系统处理超长文本输入任务，其核心思想是将参数化的大语言模型与非参数化的外部知识库相结合，进而生成事实准确性更高、上下文相关性更强的回答^[14-15]。该系统突破了大语言模型固有的上下文窗口限制，同时能够缓解由模型内在知识过时或存在偏差引发的幻觉。

RAG 系统采用 Sentence-BERT 作为嵌入模型，该模型在语义相似度任务上表现出色，支持长文本编码且计算效率高，适用于网络安全领域专业术语的细粒度语义捕捉^[16]。在生成阶段，当用户发起提问时，RAG 系统依托在线构建的知识库实时增强大模型的内容生成。当用户向 RAG 系统输入自然语言进行查询时，RAG 系统使用嵌入模型将查询向量化，确保文本块与查询处于统一的向量空间，避免模型差异导致语义表征偏移，从而保证检索结果的准确性和一致性，提升召回率和精确率。随后，系统以查询向量为键在向量数据库中进行近似最近邻检索，获取语义最匹配的排名前 k 个文本块作为回答当前问题的关键证据上下文。RAG 基于问题和检索文档生成答案的目标可形式化表示为：

$$p_{\eta}(y|x) = \sum_{z \in \text{Top-}k(p_{\eta}(\cdot|x))} p_{\eta}(z|x) \cdot p_{\eta}(y|x,z) \quad (1)$$

其中， x 为输入的查询或条件， y 为生成的响应， z 为检索到的文档片段， p_{η} 为联合概率模型， $\text{Top-}k(p_{\eta}(\cdot|x))$ 是根据查询 x 返回的排名前 k 个最相关的文档集合。

将检索到的上下文文本块与用户原始查询一

起填入预设的提示模板中。该模板明确要求大模型仅基于给定的上下文信息作答，不得脱离上下文背景或虚构内容，从而有效限制生成范围。RAG 系统通过将数据预处理与实时推理解耦，融合了密集向量检索的高效语义搜索能力与大型生成式模型的强大文本生成能力^[17]，提升了模型在处理领域专家问题时的正确性、可靠性与可信性。

2.3 数据多样化

在利用大模型生成数据的过程中，可能出现问题形式过于单调导致数据退化等现象。Hao 等^[18]提出的体裁-受众对方法为解决数据风格单一问题提供了方案。其中，体裁指代风格与知识表达框架（如学术论文、科普故事等），通过定义内容的目的、结构与深度来控制生成方向；受众则结合潜在读者的特征（如年龄、职业、知识背景）指导生成内容的适配性。二者结合，使重构文本在保留核心知识的同时，呈现高度多样化的表达形式。Long 等^[19]提出的大模型数据多步生成框架，通过任务分解和分步调度，有效弥补了单步生成在复杂推理、长文本连贯性和数据集全局控制方面的不足，是减少生成数据幻觉、提高数据多样性的关键技术之一。受上述技术启发，本文融合了体裁-受众对和多步生成框架的基本思想来解决大语言模型训练过程中出现的数据稀缺和退化问题。

在体裁-受众对数据增强框架下，各模块的交互流程如下：文件解析模块完成文档加载与语义分割后，将文本块传递至数据向量化模块，生成向量并存储于向量数据库；用户通过提示词修改模块优化提示词，结合检索增强模块从向量数据库中检索相关文本块；数据生成模块基于体裁-受众对生成多样化问答对，经数据评估模块筛选后，由数据输出模块导出多种格式的训练数据，形成闭环交互。

如图 2 所示，在数据增强环节，系统首先通过大模型对每份文档进行分析，生成 3 种体裁-受众对。每一对均包含该文档内容可能对应的受众类型及提问风格。以网络安全领域的文档为例，系统选取程序开发人员、渗透测试人员、安全运营中心（SOC）分析师和教育培训师作为目标受众，分别分析其可能提出的问题的风格^[20]。

文件解析与数据增强界面示例如图3所示, 相关生成示例和生成数据示例如图4和图5所示。多样化的体裁-受众对确保了合成过程能够从多个角度充分覆盖原始资料内容。在问答对生成环节, 系统根据文档生成的体裁-受众对, 引导大模型从不同角度生成形式多样的问题, 既保留了原文档的核心主旨, 又实现了增强型问答对的多样化合成。本文将数据增强技术从预训练增强拓展至微调数据集生成框架, 通过多样化的受众分析有效解决了数据风格单一的弊端, 提升了原始语料库的利用效率。

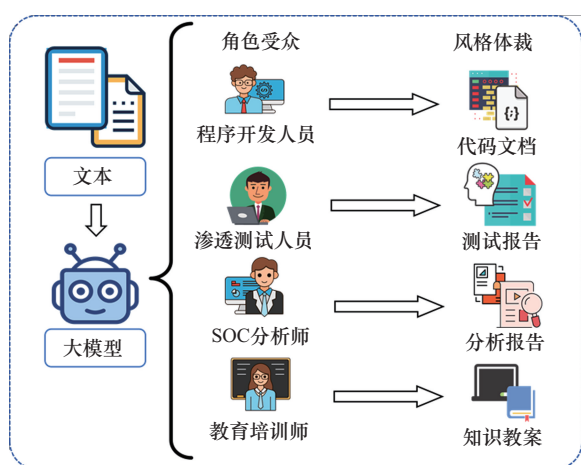


图2 体裁-受众对数据增强框架

2.4 数据生成

数据生成模块基于知识驱动的问答对构建框架, 首先建立数据集的知识本体与语义框架。DynaSynth为用户预设了多种提示词模板。参考Yu等^[21]提出的多阶段微调方法, DynaSynth提供了预训练、指令微调训练、推理微调训练和多项选择题等数据格式的提示词模版。此外, DynaSynth将整体生成过程分解为若干链式子任务, 首先对长文本样本进行分解与总结, 然后检索与文本相关内容, 接着将生成内容与数据集生成提示词结合, 最终通过大模型完成数据集生成。数据生成界面示例如图6所示。

需要说明的是, 总结内容不一定会成为最终生成结果的一部分, 但明确输出中间推理步骤有助于提升复杂数据的生成质量。生成过程严格遵循Schema-Guided原则, 通过双提示 (dual-prompt) 机制实现结构化输出, 单次推理可生成多组包含思维链的{问题, 标准化答案, 参考内容}三元组数据^[22]。

2.5 数据评估

本文引入了DeepEval框架, 集成G-Eval指标, 在数据生成阶段结束后, 可以利用质量管控大语言模型对部分数据进行抽检, 实现自动化筛选^[23]。该定制化评估机制基于问答对及其关联

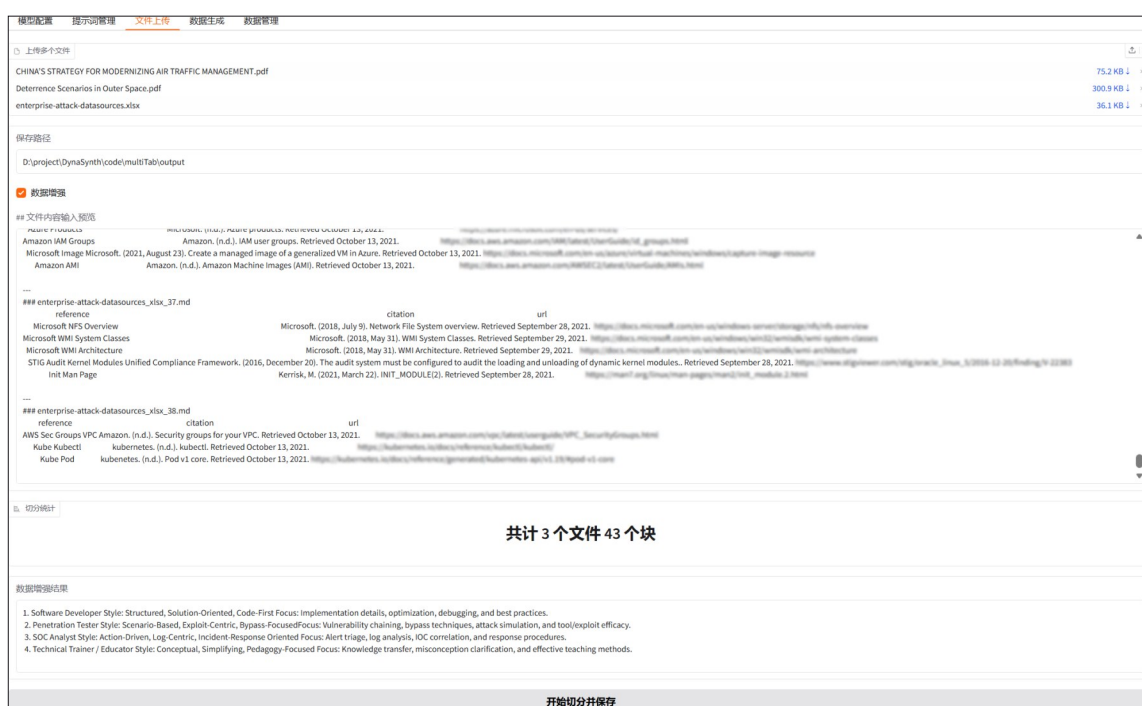


图3 文件解析与数据增强界面示例

```
{
  "Genre 1": "Focuses on threat indicators, incident response steps, and log analysis methods",
  "Audience 1": "SOC (Security Operations Center) Analyst",
  "Genre 2": "Focus on vulnerability exploitation steps, attack paths, and methods of privilege escalation",
  "Audience 2": "Penetration Tester",
  "Genre 3": "Focus on compliance requirements, risk control frameworks, and organizational security strategies",
  "Audience 3": "Security Policy Maker / Compliance Officer",
  "Genre 4": "Focus on case explanation, teaching Q&A, and simulating attack scenarios",
  "Audience 4": "Security Training Instructor"
}
```

图4 体裁-受众对生成示例

```
{
  "question": "According to the ICS ATT&CK framework, what is the primary goal of an adversary employing the 'Collection' (TA0100) tactic, and what types of data are they trying to gather?",
  "thought": "The question asks for the adversary's goal and the data types targeted in the 'Collection' tactic. The answer must be extracted directly from the description column for TA0100, avoiding vague terms and using specific content from the document.",
  "answer": "The primary goal of an adversary employing the 'Collection' tactic is to gather data of interest and domain knowledge on the ICS environment to inform their objective. They are trying to gather domain knowledge and obtain contextual feedback. The data collected includes information on control systems and targets of interest, such as operation states, screenshots, unique device roles, system and diagram schematics, protocol-specific, device-specific, and process-specific configurations and functionality. This information can pertain to a combination of system, supervisory, device, and network related data, which conceptually fall under high, medium, and low levels of plan operations. Examples include information repositories on plant data at a high level or device specific programs at a low level. Sensitive floor plans, vendor device manuals, and other references may also be targeted.",
  "references": ["The adversary is trying to gather data of interest and domain knowledge on your ICS environment to inform their goal.", "Collection consists of techniques adversaries use to gather domain knowledge and obtain contextual feedback in an ICS environment.", "This tactic is often performed as part of Discovery, to compile data on control systems and targets of interest that may be used to follow through on the adversary's objective. Examples of these techniques include observing operation states, capturing screenshots, identifying unique device roles, and gathering system and diagram schematics."]
}
```

图5 生成数据示例

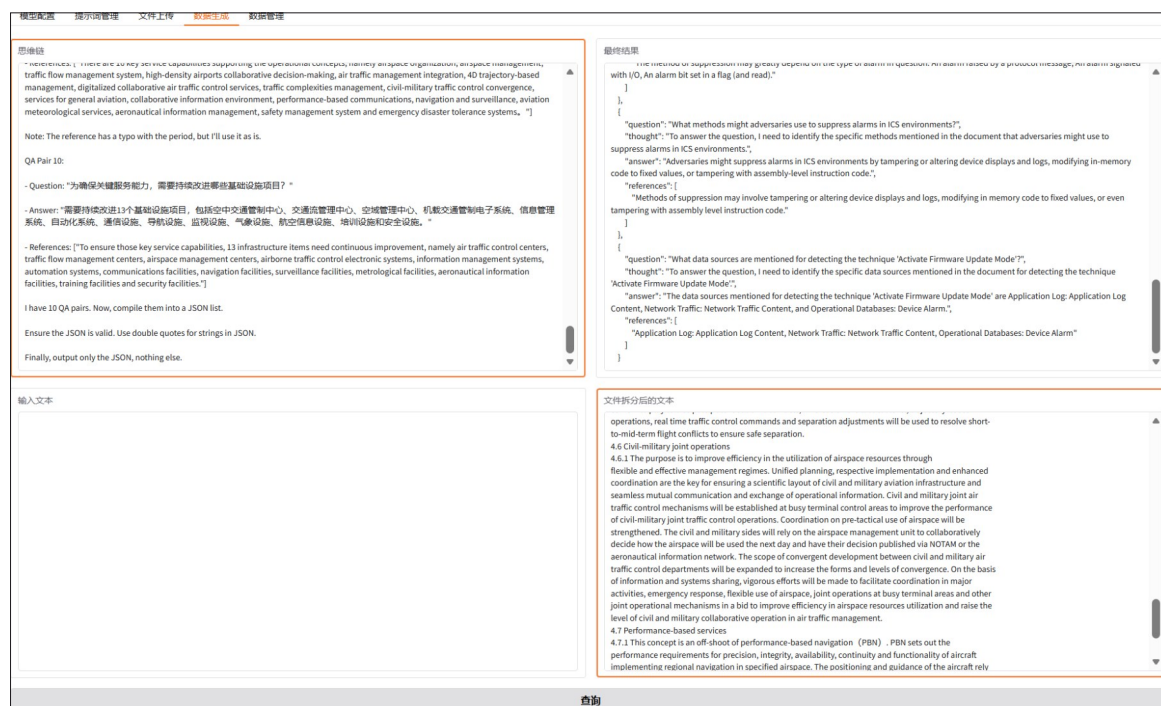


图6 数据生成界面示例

上下文生成0~100分的原始评分,并强制模型输出包含评分依据的结构化推理。评估准则主要包含两个维度:一是问题质量检测,涵盖歧义性表述、上下文主题引用缺失、无关主题引入或问题定义过宽等;二是答案质量检测,涉及答案无关性、上下文关键信息未引用、信息完整度不高、模糊表述或超出上下文信息等。此外,系统还从语义一致性、逻辑完整性、专业术语准确性和使用安全性4个维度引导模型对生成数据进行综合评估。在保证数据准确有效的同时,确保数据安全可用,避免包含过于具体的攻击指令或可能误导防御决策的建议。最终,通过除以10的后处理,将原始评分归一化为0~1的标准化值。

2.6 提示词优化反馈

在DynaSynth中,用户可以通过手动或自动方式优化提示词。提示词优化生成策略需在效率与质量之间取得平衡,在选择最优提示时,兼顾数据生成的格式约束性与内容多样性需求。DynaSynth采用基于规则的结构化提示设计,其核心是通过明确的格式指令(如分隔符、字段约束、示例)压缩候选集空间,以减少模型生成偏差。关键要素包括:使用符号定义数据结构的格式锚定、插入示例引导模型学习数据分布的示例注

入,以及添加量化限制的约束强化。提示词修改界面示例如图7所示。

自动优化方式不仅允许用户提供反馈意见,还能结合提示词优化策略对提示词进行动态更新,从而在最优控制的视角下开展提示工程^[24]。本文将提示词优化问题定义如下:

$$\begin{aligned} \max f(z_{r,t}; z_q) + R(t) \\ \text{s.t. } z_{r,t} = \text{LLM}(z_{p,t}) \\ P_t \subset P_t + 1 \end{aligned} \quad (2)$$

其中, $f(z_{r,t}; z_q)$ 为数据的评估函数, $R(t)$ 为控制交互轮次的正则化项, $z_{r,t}$ 为大语言模型(LLM)根据第 t 轮提示词生成的响应, z_q 为给定任务描述, $z_{p,t}$ 为第 t 轮的提示词, P_t 为提示词候选集,随交互过程动态扩展。提示词工程的目标是找到一组提示词 $z_{q,t}$, 这组提示词能够引导大模型作出最优响应。

DynaSynth采用结构化指令引导的迭代重写框架来生成和优化提示词变体。该框架的核心是一个轻量级的辅助大语言模型。系统首先对当前最优提示词进行结构分析,识别其核心指令、约束条件、示例部分和输出格式要求;随后,根据预设的优化目标,自动生成具体的重写指令。辅助大语言模型根据该指令对候选提示词进行多轮

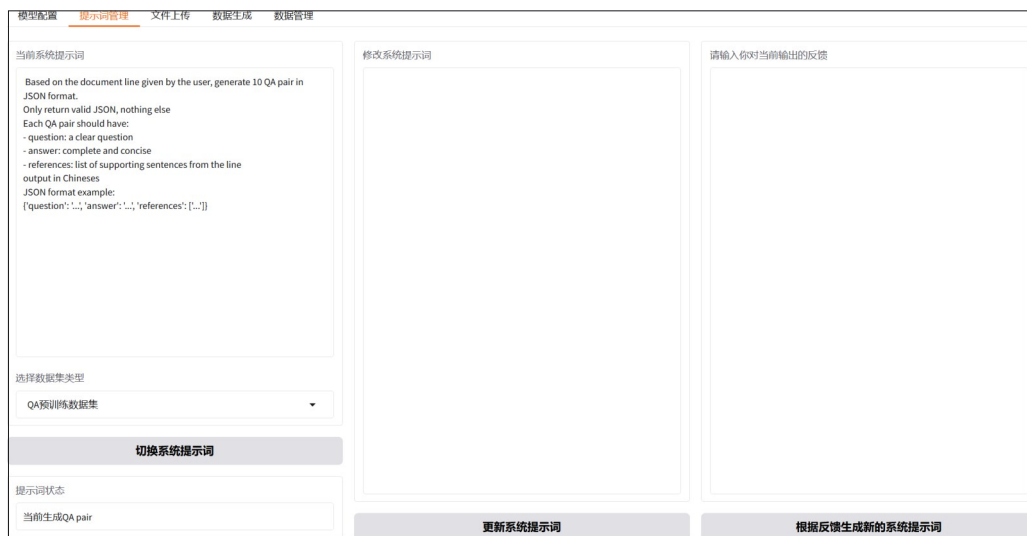


图 7 提示词修改界面示例

重写，每个候选变体会被用于生成一小批样本问答对。利用 G-Eval 评估器对这些样本进行快速评分，输出评分最高的提示词作为最终优化结果。该策略将提示词优化建模为一个搜索优化问题，利用大语言模型的生成能力在提示词空间中执行定向搜索，并通过自动化评估实现快速反馈，从而系统化地逼近最优提示词，而非依赖人工试错。

2.7 模型配置

DynaSynth 实现了大语言模型与数据合成过程的深度融合，通过可视化操作界面，使用户能够便捷地配置多种大语言模型，并自动调用模型

完成数据合成任务。模型配置界面示例如图 8 所示，用户可在本界面中选择模型并填写模型服务统一资源定位器（URL）、API Key、模型名称等信息，即可高效调用指定模型；同时支持基于 Ollama 进行本地化部署，以满足特定模型的部署需求^[25]。此外，DynaSynth 支持温度、Top- k 抽样等参数的自定义调整，用户可以根据需求与领域限制，对模型生成过程进行精细化控制。

2.8 用户界面

DynaSynth 基于 Gradio 构建可视化交互界面，为用户提供简洁易用的操作体验^[26]。用户

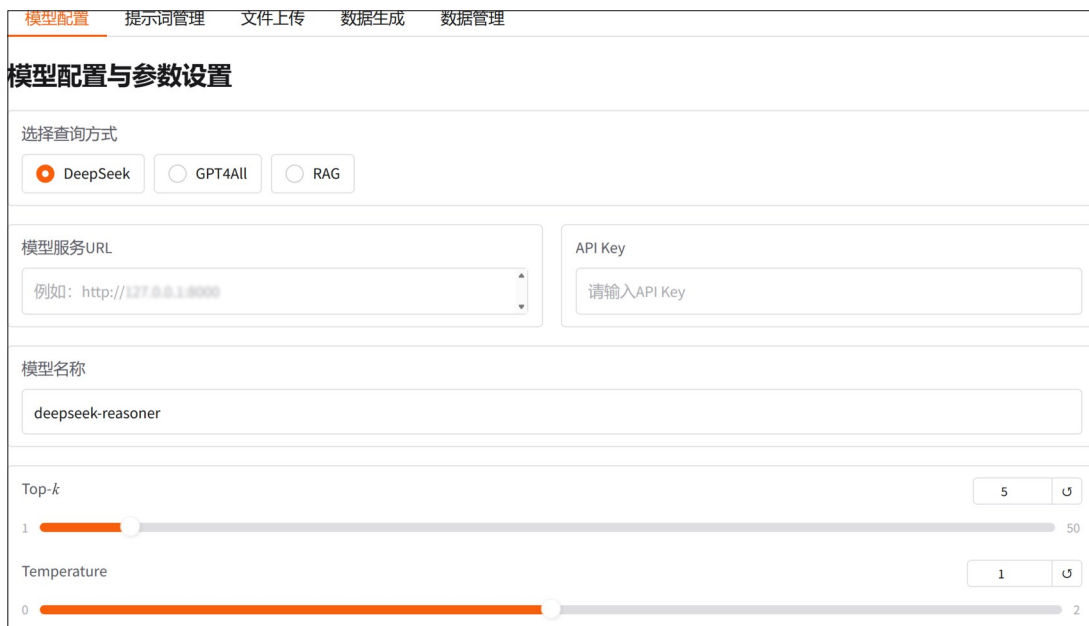


图 8 模型配置界面示例

可以先加载数据集,并以自然语言的形式向工具提交任务,同时指定期望的数据种类与格式。系统引导用户分步提交任务,相较于一次性完成所有意图描述,该方式可以实现数据生成过程的渐进推进,使用户可以掌控全过程。用户还可以通过界面定制知识库与提示词。数据集生成后,支持一键复制生成内容,并可将问答对以标准数据集格式(如JSON、TXT和CSV等格式)导出,从而降低大模型微调的使用门槛。

3 实验分析

3.1 实验设置

为验证 DynaSynth 生成数据集的质量,本文选取了 Distilabel、Kiln、Curator 和 EDG4LLM 4 款主流数据生成工具开展对比实验。基于 MITRE ATT&CK 数据源,分别使用 DynaSynth 与上述对比工具生成网络安全威胁行为领域的专业问答对数据集,并采用低秩自适应(low-rank adaptation, LoRA)方法对 Qwen3-8B-Instruct 模型进行微调^[27]。研究人员从公开网络安全数据集 AttackQA 中抽取了部分问题构成测试集,通过对比使用不同数据集微调后的模型在该测试集上的表现,间接评估各生成数据集的质量^[28]。

针对网络安全垂直领域的任务评估,本文结合主观和客观指标对生成内容质量进行系统化评估。在主观评估方面,采用 LLM-as-a-judge 方法引导大模型从语义一致性、逻辑完整性、专业术语准确性和使用安全性 4 个维度进行量化评分。其中,语义一致性衡量生成内容与参考文本在核心含义上的匹配程度,确保信息传达准确且无歧义;逻辑完整性评估生成内容的推理连贯性与结构清晰度,要求完整覆盖问题所需的关键信息点与逻辑链条;专业术语准确性考察生成内容中领域特定术语的使用是否符合行业规范,避免错误或误导性表述。最终取 3 次独立评分的平均值作为主观得分,以确保评估结果的可靠性与领域针对性。

3.2 评价指标

为客观评估各数据生成工具所生成数据集的质量,本文构建了一套融合通用自然语言处理指标与行业相关评测指标的综合评价体系。具体采用双语评估辅助(BLEU)指标和面向召回率的自动摘要评估(ROUGE)指标,并引入大语言

模型评判方法对各生成数据微调后的模型进行答案完整度和准确度的评测^[29]。其中,基于 Qwen-max 的增强评估框架用于对模型生成的网络安全响应进行 0~100 分的评分。

BLEU 指标用于评估不同数据集微调后的 Qwen3-8B-Instruct 模型所生成文本的质量。BLEU- N 分数通过计算候选文本与参考文本之间 n -gram ($n=1,2,3,4$) 的修正精确度来衡量生成质量^[30]。其计算式如下:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (3)$$

$$\text{BP} = \begin{cases} 1 & , c > r \\ e^{(1-r/c)} & , c \leq r \end{cases} \quad (4)$$

其中, P_n 为 n -gram 的修正精确度(即候选文本中所有 n -gram 的匹配次数与总次数之比), BP 为长度惩罚因子, w_n 为分母与 n -gram 对应的权重因子, c 为候选文本的长度, r 为有效参考语料长度。

ROUGE 系列指标以召回率为核心,重点考察参考文本中的信息覆盖程度^[31],其计算式如下:

$$\text{ROUGE} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (5)$$

其中, $\{\text{Refs}\}$ 为来自测试集原始答案的参考文本集合; $\text{Count}_{\text{match}}(\text{gram}_n)$ 为 n -gram 在生成文本和参考文本中同时出现的最大次数, $\text{Count}(\text{gram}_n)$ 为所有参考文本中 n -gram 的总数。

ROUGE-L 以最长公共子序列(longest common subsequence, LCS)为度量来衡量生成文本的逻辑连贯性,能够有效验证问题与答案的语义对齐程度。其计算式如下:

$$R_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{m} \quad (6)$$

$$P_{\text{lcs}} = \frac{\text{LCS}(X,Y)}{n} \quad (7)$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \times (P_{\text{lcs}} \times R_{\text{lcs}})}{\beta^2 \times P_{\text{lcs}} + R_{\text{lcs}}} \quad (8)$$

其中, R_{lcs} 为召回率, $\text{LCS}(X,Y)$ 为参考文本 X 和生成文本 Y 的 LCS 长度, m 为参考文本 X 的长度, P_{lcs} 为精度, n 为生成文本 Y 的长度, β 为召回率权重。

对于部分需完整呈现逻辑链的网络安全问答任务, ROUGE-L 关注长文本连贯性,较 n -gram 类

指标更适用于垂直领域评估。BLEU等指标易受局部词汇重复影响,而ROUGE-L通过LCS捕捉全局语义关联,能够更准确地衡量生成内容与参考文本的匹配度,从而验证生成数据是否提升了模型的逻辑化表达能力,而非仅依赖词汇层面的匹配。

3.3 结果分析

按照实验设计方法,将微调后的模型及各基线模型在测试集上进行评估。图9展示了Qwen3-8B-Instruct在使用不同数据生成工具生成的数据集微调后,在领域特定评估数据集上的表现。结果表明,DynaSynth生成数据集微调后的模型在BLEU、ROUGE等指标上均显著优于对比工具。体裁-受众对数据增强机制有效提升了数据的语义完整性与领域适配性,验证了DynaSynth在垂直领域的技术优势。

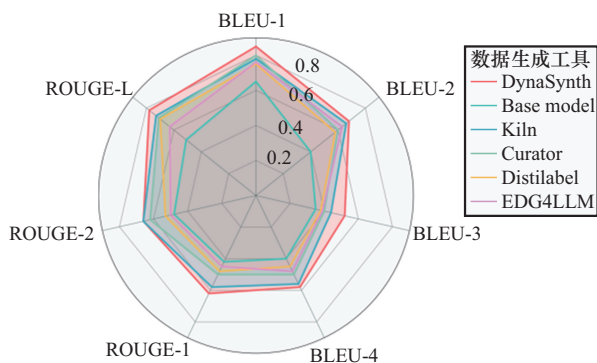


图9 各数据集微调后的模型的实验结果

各数据集微调后的大模型的ROUGE-L指标对比如图10所示。对比分析发现,在网络安全领域特定任务中,DynaSynth的ROUGE-L分布集中在0.747,较排名第二的Kiln高出0.028;在大于0.7的高分区间中,DynaSynth的占比高于其他模型,高分分布更加集中,表明生成数据的质量稳定性更优,减少了低质量数据对模型微调的干扰。这一结果验证了多样化数据增强的有效性,并证明该代理工具生成的数据能够提升模型的逻辑表达能力,而非仅依赖词汇层面的匹配。

图11展示了各数据生成代理工具在测试集上的大模型评分对比及其结果分布。DynaSynth生成数据的大模型评分分布更集中于高分区间,大部分评分在70分以上,表明生成数据质量较优,低质量数据占比较低,减少了对模型微调的干扰。

通过向量数据库检索增强生成,系统确保了领域关键信息的准确覆盖,提高了答案完整性评分;多样化的问答对生成机制有效避免了数据退化,增强了语义完整性与领域适配性。相比之下,其他工具的评分分布相对分散,高分占比低于DynaSynth,反映了其生成数据质量波动较大,且缺乏动态反馈优化机制,难以迭代提升数据质量。

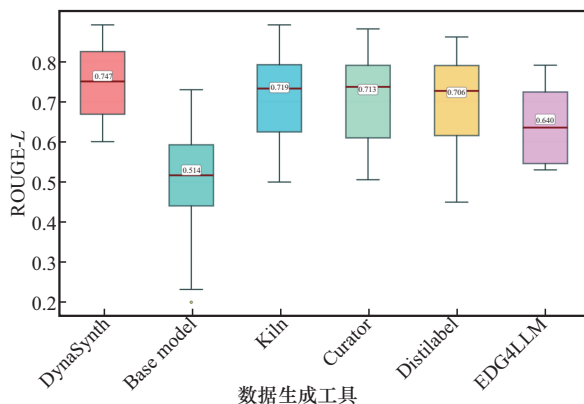


图10 各数据集微调后的大模型的ROUGE-L指标对比

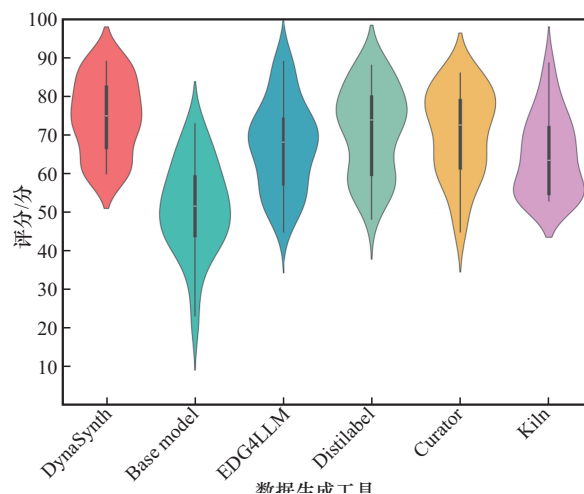


图11 各数据生成代理工具在测试集上的大模型评分对比及其结果分布

从ROUGE-L与主观评分的分布情况来看,DynaSynth生成的数据在Qwen3-8B-Instruct评估框架下质量水平更高,表明该工具在垂直领域数据生成方面具备良好的可靠性,其生成的数据能够更稳定地提升模型在网络安全任务中的性能。

4 用户测试

本文在后期开展了针对性用户测试,在合作单位内部署了DynaSynth数据生成代理工具,并

邀请多位非AI背景的用户在符合安全规范的框架下参与试用测试与访谈。测试中,4位测试人员基于DeepSeek线上大模型,使用该代理工具将多个相同的数据文档生成为一个包含100个问题的数据集,记录完成任务所需的时间与错误次数,并填写系统可用性量表,相关统计数据见表1。测试结果表明,4位测试人员完成数据生成任务的平均时间为14分15秒,系统可用性量表得分的平均值为90分(显著高于68分的可用性基准线)。参与测试的用户普遍反馈数据生成效率显著提升,完成任务所需时间较少,能够为后续大模型赋能相关领域的研究工作奠定基础。在操作难度方面,用户在经过简单学习后均可独立利用现有研究资料生成所需的大模型训练数据。测试人员反馈意见见表2。

表1 用户测试统计数据

测试人员	完成任务所需时间	系统可用性量表得分/分
测试人员1	12分13秒	95
测试人员2	15分53秒	88
测试人员3	14分1秒	91
测试人员4	14分36秒	86

表2 测试人员反馈意见

测试人员	反馈意见
测试人员1	可视化界面友好,无须懂AI技术也能快速生成数据,操作流程简单
测试人员2	提示词模板很实用,不用自己编写复杂指令。希望以后可以支持更多的大模型
测试人员3	本地化部署灵活,完成任务所需时间比预期少很多,对非技术人员很友好
测试人员4	不用手动处理数据格式,能够满足我们的需求

5 结束语

本文开发了大模型合成数据生成代理工具DynaSynth。该工具能够自动合成领域相关的微调数据集,支持多种结构的数据文件解析,并借助向量数据库检索技术实现高效的数据检索,从而生成更符合实际需求的数据集。同时,该工具中引入的基于体裁-受众对的数据增强机制,能够提升生成数据的多样化水平。此外,DynaSynth提供了可视化用户界面,支持通过API调用与本地部署模式配置模型相关参数,从而降低了用户

的技术使用门槛。在数据导出方面,该工具支持JSON、TXT、CSV等格式,并提供提示词优化功能,支持逐步生成高质量数据。

网络安全相关领域的实验结果表明,DynaSynth基于体裁-受众对机制能够生成更高质量的问答数据集。利用该数据集对大模型进行微调后,模型在垂直领域的专业能力得到显著提升,同时保持了良好的知识泛化能力。该代理工具有效地解决了当前大模型数据生成过程中存在的部分弊病,在确保生成的问答数据集兼具相关性与多样化的前提下,为解决大模型发展过程中面临的数据瓶颈问题提供了可行的技术路径。

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.
- [2] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[PP]. arXiv(2022-10-26) [2025-12-27]. arXiv: arXiv. 2206. 07682.
- [3] Wang L, Ma C, Feng X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.
- [4] Guo T C, Chen X Y, Wang Y Q, et al. Large language model based multi-agents: a survey of progress and challenges[PP]. arXiv(2024-04-19)[2025-12-27]. arXiv:arXiv. 2402.01680.
- [5] Chen W Z, Koenig S, Dilkina B. RePrompt: planning by automatic prompt engineering for large language models agents[PP]. arXiv (2025-02-13)[2025-12-27]. arXiv:arXiv.2406.11132.
- [6] Khatami S, Frantz C. Prompt engineering guidance for conceptual agent-based model extraction using large language models[PP]. arXiv (2024-12-05)[2025-12-27]. arXiv:arXiv.2412.04056.
- [7] Sun M J, Han R J, Jiang B Y, et al. LAMBDA: a large model based data agent[J]. Journal of the American Statistical Association, 2025: 1-13.
- [8] Qiao B, Li L Q, Zhang X, et al. TaskWeaver: a code-first agent framework[PP]. arXiv(2024-06-20)[2025-12-27]. arXiv:arXiv. 2311.17541.
- [9] Wu Q, Bansal G, Zhang J, et al. AutoGen: enabling next-Gen LLM applications via multi-agent conversations[PP]. arXiv(2023-08-16) [2025-12-27]. arXiv:arXiv. 2308.08155.
- [10] Qian C, Liu W, Liu H Z, et al. ChatDev: communicative agents for software development[PP]. arXiv(2024-06-05) [2025-12-27]. arXiv: arXiv. 2307.07924.
- [11] Canto Á B D, Blázquez G M, Lajarán A P, et al. Distilabel: an AI feedback (AIF) framework for building datasets with and for LLMs [EB]. 2026.
- [12] Chen Z H, Jiang W L, Li J Z, et al. GraphGen: enhancing supervised fine-tuning for LLMs with knowledge-driven synthetic data generation[PP]. arXiv(2025-05-26) [2025-12-27]. arXiv: arXiv. 2505.20416.
- [13] Chen D Y, Huang Y L, Ma Z J, et al. Data-juicer: a one-stop data processing system for large language models[C]//Proceedings of the Companion of the 2024 International Conference on Management

- of Data. New York: ACM, 2024: 120-134.
- [14] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020. San Diego: NeurIPS, 2020: 9459-9474.
- [15] Gao Y F, Xiong Y, Gao X Y, et al. Retrieval-augmented generation for large language models: a survey[PP]. arXiv(2024-03-27)[2025-12-27]. arXiv:arXiv. 2312.10997.
- [16] 刘文彦, 霍树民, 全青, 等. 网络安全评估与分析模型研究[J]. 网络与信息安全学报, 2018, 4(4): 1-11.
- Liu W Y, Huo S M, Tong Q, et al. Research on models of network security evaluation and analysis[J]. Chinese Journal of Network and Information Security, 2018, 4(4): 1-11.
- [17] Lála J, O' Donoghue O, Shtedritski A, et al. PaperQA: retrieval-augmented generative agent for scientific research[PP]. arXiv[2025-12-27]. arXiv:arXiv. 2312.07559.
- [18] Hao X T, Zhu R J, Zhang G, et al. Reformulation for pretraining data augmentation[PP]. arXiv(2023-12-14) [2025-12-27]. arXiv: arXiv. 2502. 04235.
- [19] Long L, Wang R, Xiao R X, et al. On LLMs-driven synthetic data generation, curation, and evaluation: a survey[PP]. arXiv (2025-05-19) [2025-12-27]. arXiv:arXiv. 2406.15126.
- [20] 张繁, 谢凡, 江颀. 网络威胁安全数据可视化综述[J]. 网络与信息安全学报, 2018, 4(2): 34-39.
- Zhang F, Xie F, Jiang J. Survey on the visualization technologies of threatening security data in cyber space[J]. Chinese Journal of Network and Information Security, 2018, 4(2): 34-39.
- [21] Yu Y C, Chiang T H, Tsai C W, et al. Primus: a pioneering collection of open-source datasets for cybersecurity LLM training[PP]. arXiv (2025-10-03)[2025-12-27]. arXiv:arXiv. 2502.11191.
- [22] Du Y H, Oraby S, Perera V, et al. Schema-guided natural language generation[PP]. arXiv(2020-11-04) [2025-12-27]. arXiv: arXiv. 2005.05480.
- [23] Stureborg R, Alikaniotis D, Suhara Y. Characterizing the confidence of large language model-based automatic evaluation metrics[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 76-89.
- [24] Luo Y F, Tang Y M, Shen C F, et al. Prompt engineering through the lens of optimal control[PP]. arXiv(2023-11-03)[2025-12-27]. arXiv: arXiv. 2310.14201.
- [25] Marcondes F S, Gala A, Magalhães R, et al. Using ollama[M]//Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs. Cham: Springer Nature Switzerland, 2025: 23-35.
- [26] Abid A, Abdalla A, Abid A, et al. Gradio: hassle-free sharing and testing of ML models in the wild[PP]. arXiv(2019-06-06)[2025-12-27]. arXiv:arXiv. 1906.02569.
- [27] Yang A, Li A F, Yang B S, et al. Qwen3 technical report[PP]. arXiv (2025-05-14)[2025-12-27]. arXiv:arXiv. 2505.09388.
- [28] Krishna V B. AttackQA: development and adoption of a dataset for assisting cybersecurity operations using fine-tuned and open-source LLMs[PP]. arXiv(2024-11-01) [2025-12-27]. arXiv: arXiv. 2411. 01073.
- [29] Zheng L, Chiang W L, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot arena[J]. Proceedings of the 37th International Conference on Neural Information Processing Systems. New York: NeurIPS, 2023: 46595-46623.
- [30] Papineni K, Roukos S, Ward T, et al. BLEU: a method for automatic evaluation of machine translation[C]//Proceedings of the 40th An-

nual Meeting on Association for Computational Linguistics. Morris-town: ACL, 2002:311-318.

[作者简介]



关永健 (1997-), 男, 空军工程大学信息与导航学院博士生, 主要研究方向为网络空间安全与大语言模型。



朴乘锴 (1991-), 男, 空军工程大学信息与导航学院讲师, 主要研究方向为网络空间安全与大语言模型安全。



王布宏 (1975-), 男, 空军工程大学信息与导航学院教授, 主要研究方向为空管信息系统安全和大语言模型。



赵博夫 (1986-), 男, 空军工程大学信息与导航学院博士生, 主要研究方向为网络空间安全。



李思琦 (2002-), 女, 空军工程大学信息与导航学院硕士生, 主要研究方向为大语言模型时间序列异常检测。



赵正阳 (1997-), 男, 空军工程大学信息与导航学院博士生, 主要研究方向为对抗学习与深度学习应用。