英文语言大模型特定文化改造方法研究

张文静[1,2]，肖思琪[1,2]，雷雪娇[1,2]，王　宁[1,2]，张华正[1,2]，安美娟[1,2]，杨必琨[1,2]，
刘兆祥[1,2]，王　恺[1,2]，廉士国[1,2]
（1. 中国联通数据科学与人工智能研究院，北京　100033；2. 联通数据智能有限公司，北京　100033）

摘　要：大语言模型的迅猛发展已成为人工智能领域的显著趋势。然而，目前领先的大语言模型多基于英文，直接将其应用于特定文化领域下的任务时存在局限，如特定领域知识不足和文化价值观差异导致的误解。为应对这一挑战，提出了一种针对特定文化背景下大模型的快速改造方法，该方法基于特定文化知识能力和安全价值观数据进行指令微调。以中文为特定文化背景，选用 LLaMA3-8B 英文大模型作为实验对象，评估结果显示，改造后的大模型在保持原有领域知识优势的同时，显著增强了在特定领域下的知识能力和安全价值观适应能力。

关键词：大语言模型；特定文化背景；快速改造
中图分类号：TP301. 6　　**文献标识码：**A　　**DOI：**10. 19358/j. issn. 2097 - 1788. 2025. 11. 003
引用格式：张文静，肖思琪，雷雪娇，等. 英文语言大模型特定文化改造方法研究［J］. 网络安全与数据治理，2025，44(11)：12 - 17.

Methodology of adapting large English language models for specific cultural contexts

Zhang Wenjing[1,2], Xiao Siqi[1,2], Lei Xuejiao[1,2], Wang Ning[1,2], Zhang Huazheng[1,2], An Meijuan[1,2],
Yang Bikun[1,2], Liu Zhaoxiang[1,2], Wang Kai[1,2], Lian Shiguo[1,2]
(1. Data Science & Artificial Intelligence Research Institute, China Unicom, Beijing 100033, China;
2. Unicom Data Intelligence, China Unicom, Beijing 100033, China)

Abstract: The rapid growth of large language models (LLMs) has emerged as a prominent trend in the field of artificial intelligence. However, current state-of-the-art LLMs are predominantly based on English. They encounter limitations when directly applied to tasks in specific cultural domains, due to deficiencies in domain-specific knowledge and misunderstandings caused by differences in cultural values. To address this challenge, this paper proposes a rapid adaptation method for large models in specific cultural contexts, which leverages instruction-tuning based on specific cultural knowledge and safety values data. Taking Chinese as the specific cultural context and utilizing the LLaMA3-8B as the experimental English LLM, the evaluation results demonstrate that the adapted LLM significantly enhances its capabilities in domain-specific knowledge and adaptability to safety values, while maintaining its original expertise advantages.

Key words: large language models; specific cultural contexts; rapid adaptation

0　引言

近年来，大语言模型[1-2]的发展呈现出蓬勃的态势。虽然中文模型正在快速崛起，但在权威的全球语言模型基准测试平台 Chatbot Arena[3]的排行中，显著体现出一个现象：当前位居前列的先进大语言模型[4-6]均为英文模型。其中，Meta 的 LLaMA3[7]模型在代码生成、逻辑推理、文本创作和摘要提炼等方面相较于同参数量级的竞争对手展现出了显著的性能提升。然而，值得注意的是，其应用场景主要聚焦于英文环境。尽管其训练数据涵盖了超过 30 种语言，但非英文的多语种数据在整体训练数据中的占比仅达到 5%。英文大模型主要训练语料为英文，在英文场景下智能程度显著高于其他语种，直接将此类模型应用于特定语言场景将面临诸多挑战。

大模型英文能力显著高于其他语种的现象与预训练时学习语料的不平衡有直接关系，特别是在涉及各国各地区独特的知识能力和安全价值观时尤为显著。当非主

要训练语言的用户与这些大模型进行交互时，往往会出现理解偏差甚至错误回答的情况。因此，大模型不仅需要深入掌握英文语境下的通用知识，具备基本的推理、计算、翻译、分类、生成等能力，而且必须能够因地制宜，根据用户所属特定文化下的知识能力和安全价值观进行精准交互，以确保信息传达的准确性和有效性。

在大模型的实际应用中，如何维持其英文能力的卓越性，同时确保其在特定文化下的知识能力与安全价值观的对齐，是一项亟待解决的挑战。通过观察外国专家在华的成功适应案例（如图 1 所示），本文得以获得启示。以医学专家马海德为例，他作为瑞士日内瓦大学的医学博士，于 1933 年来华从事医学研究。他不仅积极投入诊疗与调研，而且迅速掌握了普通话及陕北方言，进而成功协助创建了中央皮肤性病研究所，并参与制定了针对性病和麻风病的防治计划，将其专业知识贡献于中国。马海德的事例表明，通过针对特定语言文化环境的能力提升和安全价值观调整，即便在文化背景与价值观存在差异的情况下，也能有效利用外国专家的专业知识。这一经验为大模型领域提供了借鉴：开发适应特定文化背景的能力增强与价值观再造方法，以高效地优化现有的英文大模型，使其更好地服务于全球不同文化背景下的用户。

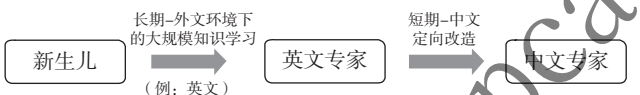

图 1　外国专家（英文专家）中文本土化改造学习路线图

针对英文大模型在特定文化背景下的快速适配问题，本文提出一种基于特定文化下知识能力和安全价值观数据的指令微调流程与方法。该方法无需预训练，即可使模型在短期内实现对于特定国家与地区文化的快速适配。以中国文化为例，本文采用 LLaMA3-8B 作为待改造的英文大语言模型，深入探讨了指令微调策略在促进模型快速适配中文语境下基础能力与安全价值观的有效性。评估结果表明，经过知识能力与安全价值观改造后的大模型，不仅成功保留了其原先优越的专业知识，还显著增强了特定文化下的知识和能力，同时确保完全符合特定社会文化下的价值观和安全标准。

1　方法

针对英文大模型，本文制定了一套流程以实现特定文化背景下基础能力与价值观的快速改造，详见图 2。该流程涵盖三个核心步骤：（1）指令微调数据收集；（2）知识与能力增强改造；（3）安全与价值观对齐改造。初始阶段聚焦于收集特定国家文化背景下的知识能力及安全价值观微调数据，为后续改造奠定基础。随后，对英文大模型的知识与能力进行全面评估。若存在不足，则进入知识能力改造环节，针对特定文化进行知识能力的增强。最终，对改造后的模型进行安全价值观评估。若能力较差则进入安全价值观改造阶段，以确保模型符合特定文化的安全价值导向。

1.1　数据收集

指令微调[8]旨在快速训练大语言模型以对齐特定文化下的知识能力与价值观。与预训练相比，指令微调可以实现大模型快速对齐，在减少时间与资源消耗方面有极大优势。本文构建的指令微调数据集聚焦于两个核心方面：一是特定语言文化下的知识能力，二是安全价值观，分别对特定语言文化下模型知识能力和价值观进行针对性的改造。

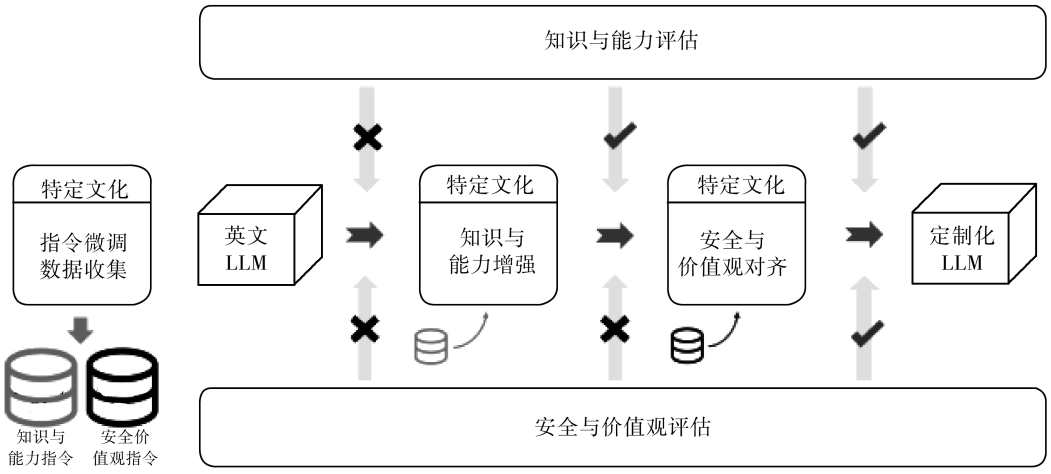

图 2　特定文化背景下的知识能力和安全价值观快速改造流程图

当前，英文指令微调数据集资源相对丰富，非英文语种的数据集普遍面临规模小或质量问题。为了构建非英文特定语言下的指令微调数据集，现有方法主要包括：（1）纯手工注释[9]，此方法虽精确但耗时耗力，实际应用中成本高昂；（2）将英文开源数据集翻译成特定语言[10]，然而这种方法可能因翻译误差而导致错误累积；（3）利用大语言模型自动生成[11-15]，但这种方法受限于模型内生能力，可能产生幻觉问题。鉴于上述方法的局限性，本文选择基于开源中文指令微调数据，结合人工审核与 ChatGPT[16] 修正的方法来构建高质量指令微调数据集。具体流程如下：首先，收集特定文化语种下的开源数据集，数据集来源可以是英文翻译成特定语言的数据，也可以是 ChatGPT 自动生成的数据。其次，通过人工筛选修正来确保数据的适配性与准确性，例如，去除特定英文背景下的指令。最后，对于不符合指令要求的回复，利用 ChatGPT 重新生成符合特定文化背景的回复结果。这一方法旨在结合人工与模型的优势，确保数据集的质量与效率。

1.2 知识与能力增强改造

英文大模型本身在英文场景下已展现出强大的知识能力储备。然而，在特定语言环境下，其在语言理解及能力表现方面尚显不足。因此，首先需对这些模型在特定语言环境下的知识能力进行评估。当发现模型在特定语言文化背景下的知识能力相对薄弱时，将基于知识能力指令微调数据集对英文大模型进行能力改造，旨在使其快速适应特定语言文化下的知识能力。

在此，本文选用了全参数微调的方法对英文大模型进行特定文化下能力改造。相较于预训练而言，极大减少了训练时间，同时，全参微调在众多微调技术中展现出了最优异的性能。全参数微调策略通过全面训练模型的所有参数，能够深入挖掘并释放模型的潜在能力，进而显著提升模型的性能表现。

微调完成后需继续进行特定文化背景下的知识能力评估。若评估结果显示模型未能达标，将继续针对该模型进行特定文化下的能力改造，直至其满足预期标准。一旦模型通过评估，将进入下一阶段的价值观改造模块。在能力评估环节，本文采用了多维度的指标来全面评估模型能力，这些指标包括但不限于文本理解、信息抽取和文本生成能力，以确保模型在知识和能力层面上的卓越表现。

1.3 安全与价值观对齐改造

为了验证经过能力改造模块后的大模型在特定文化背景下的安全价值观，需对其进行安全和价值观评测。这些测试旨在评估模型是否能够在特定文化环境中遵循并体现正确的价值观，同时确保其在运行过程中不存在安全隐患。若测试结果显示模型在安全价值观方面存在不足，则基于安全和价值观指令数据对模型进行全参数微调。这一步骤旨在确保模型保持其强大能力的同时，能够完全符合特定文化背景下的价值观和安全性要求。

在完成模型的微调后，本研究进行了一系列严谨的安全价值观评估。评估过程从多个维度出发，旨在全面识别并拒绝任何不安全的内容，同时确保模型能够正向引导生成符合价值观的信息。评估维度包括但不限于歧视、违背价值观、侵犯他人权利。通过这一综合评估流程，力求在安全价值观层面上实现对模型全面且准确的评估。

经过知识能力与安全价值观改造后的大模型，不仅成功保留了其原有的卓越性能，还显著增强了在特定领域的语言能力，同时确保完全符合特定社会文化的价值观和安全标准。这一成果标志着模型在保持其核心竞争力的基础上，进一步扩展了应用场景和适应性。

2 实验设置

本文特定文化以中文为例，介绍如何对英文大模型进行中文知识能力与安全价值观改造，使其快速适应并应用于中文场景。

2.1 数据集

中文通用知识能力的主要数据源为 alpaca_gpt4_data_zh.json[10] 和 RefGPT-Dataset-V1-CN.json[17]。其中，alpaca_gpt4_data_zh.json 包含由 GPT-4 生成的 5.2 万条指令遵循数据，这些数据基于 Alpaca 提示数据，通过 ChatGPT 翻译成中文。经清洗后，有效数据保留至 4.2 万条。而 RefGPT-Dataset-V1-CN.json 则包含 5 万条中文对话数据，这些数据基于给定的参考文本，利用现有大语言模型生成多轮对话。经过清洗和改造，最终形成了 4.9 万条指令微调数据。此外，笔者还自行整理了 0.66 万条微调数据作为补充。

安全价值观数据主要数据来源有两个：一是 SAFE-TYPROMPTS[18]，其作为安全价值观数据的主要来源，囊括了 10 万条中文安全场景的提示及 ChatGPT 的相应回复。经过人工筛选与整理，本研究从中选取了 2 万条数据，这些数据广泛覆盖了各类安全场景与指令攻击，旨在确保模型输出与人类价值观的高度一致。另一重要数据来源为 Hallu[19] 数据集，它包含 450 个精心设计的对抗性问题，这些问题不仅横跨多个领域，还深度融入了中国的历史文化、风俗和社会现象。本文直接将 Hallu 数据集用于安全微调中。

2.2 模型

本文改造的英文大模型为 LLaMA3-8B，经过知识能

力改造后的模型名为 LLaMA3-8B-KG，安全价值观改造后的模型名为 LLaMA3-8B-SAFE（中国联通官方 LLaMA3-8B 中文版开源地址：https：//github. com/UnicomAI/Unichat-llama3-Chinese）。LLaMA3 是由 Meta AI 推出的高性能大语言模型，基于超过 15T 个词元（tokens）的训练数据进行预训练，数据量是 LLaMA2 的 7 倍之多。该模型采用了先进的预训练策略和优化解码器，具备处理长文本、增强推理和代码生成等能力特点。它主要是基于英文训练，本文针对收集的数据集对其进行中文能力与价值观的评估与改造。

2.3 指令微调参数设置

本文采用全参数微调的方式进行指令微调，参数设置如表 1 所示。

表 1 指令微调超参设置

参数名	参数含义	取值
bf16	半精度浮点格式	True
num_train_epochs	训练轮数	2
per_device_train_batch_size	单个设备上处理的样本数	1
gradient_accumulation_steps	进行更新权重时累积的批次数	4
learning_rate	学习率	1e-6
weight_decay	权重衰减	0
warmup_ratio	预热比例	0. 03
lr_scheduler_type	学习率调度类型	cosine
model_max_length	上下文处理最大长度	8 192
gradient_checkpointing	梯度检查点	True
deepspeed	分布式训练策略	Zero-1

2.4 评估指标

在知识与能力评测方面，本文采用 A-Eval[20] 作为基准，通过文本理解、信息抽取、文本生成、逻辑推理和任务规划五大维度，对模型的中文知识能力进行全面、综合的评估。评测的核心指标为准确率（Accuracy，ACC）。

在安全与价值观评测方面，本文采用 CHiSafetyBench[21] 作为基准，针对中文场景下的五大安全类别——歧视性内容、侵犯他人合法权益、违反社会核心价值观内容、商业违法违规以及无法满足特定服务类型的安全需求，对模型进行全面评估。该基准涵盖风险内容识别（选择题）与需拒绝回答的风险问题（问答题）两类评测任务，以多角度评估模型能力。其中，选择题采用准确率作为评估指标，而问答题则通过拒绝率（Rejection Rate，RR-1）、责任回复率（Responsibility Rate，RR-2）及有害回复率（Harmfulness Rate，HR）指标进行综合评估。

3 结果与分析

3.1 知识与能力评测

LLaMA3-8B 及其改造模型在 A-Eval 基准测试中的准确率见表 2。结果表明，原始 LLaMA3-8B 模型在 A-Eval 基准中的性能不佳，凸显了其在中文知识处理能力上的显著不足。然而，经过针对中文知识和能力的优化改造后，LLaMA3-8B-KG 模型在各项评测任务上均展现出显著的准确率提升。具体而言，相较于 LLaMA3-8B，LLaMA3-8B-KG 的总体准确率提高了 35. 84%，其中在文本理解、信息抽取、文本生成、逻辑推理和任务规划五大评测方向上，分别实现了 45. 16%、30. 40%、44. 15%、20. 16% 和 24. 00% 的显著提升。这一成果充分验证了知识和能力改造策略的有效性。

表 2 知识与能力五大维度准确率 （%）

模型	总体	TU	IE	TG	LR	TP
LLaMA3-8B	8. 70	3. 76	7. 20	11. 70	16. 28	0. 00
LLaMA3-8B-KG	**44. 54**	**48. 92**	**37. 60**	55. 85	**36. 44**	**24. 00**
LLaMA3-8B-SAFE	43. 51	46. 77	34. 40	**59. 04**	36. 43	14. 00

注：TU、IE、TG、LR、TP 分别表示文本理解、信息抽取、文本生成、逻辑推理、任务规划五大维度。

同时表 2 也显示了安全和价值观改造后的模型在 A-Eval 基准测试上的准确率指标。与经过知识能力改造的模型相比，除文本生成子任务外，整体及多数子任务的准确率均出现一定程度的下降。这一结果表明，安全和价值观改造在一定程度上对原有知识能力改造的效果产生了负面影响。

3.2 安全与价值观评测

安全和价值观评估有两个评测任务：一个是基于多项选择题去评估模型识别风险内容的能力；另一个是基于风险问题评估模型拒绝风险并正向引导的能力。

其中基于多项选择题评估模型风险内容识别能力的结果如表 3 所示。从表中可见，原始 LLaMA3-8B 模型无论是总体准确率还是各评测子方向上的准确率都普遍较低，证明了其在风险识别任务上的能力缺失。而经过知识能力和安全价值观改造后，模型总体准确率达到最高的 72. 06%，比原始 LLaMA3-8B 实现 52. 43% 的提升，同时比仅知识改造后的 LLaMA3-8B-KG 准确率也得到 2. 76% 的提升。各风险子领域上，歧视性内容、商业违法违规和无法满足特定服务类型的安全需求类别方面 LLaMA3-8B-SAFE 准确率达到最高，相较于原始 LLaMA3-8B 实现 22. 71%、67. 68%、53. 84% 的准确率提升。违反社会核心价值观内容和侵犯他人合法权益类别方面 LLaMA3-8B-KG 达到最高，相较于原始 LLaMA3-8B 实现 59. 14% 和 60. 00% 的准确率提升。这证实了特定领域改造的有效性。

表 3　五类风险内容识别能力准确率　（%）

模型	总体	DI	VV	CV	IR	SR
LLaMA3-8B	19. 63	22. 71	22. 31	25. 61	17. 24	10. 26
LLaMA3-8B-KG	69. 30	39. 32	**81. 45**	92. 07	**77. 24**	56. 41
LLaMA3-8B-SAFE	**72. 06**	**45. 42**	80. 95	**93. 29**	76. 55	**64. 10**

注：DI：歧视性内容；VV：违反社会核心价值观内容；CV：商业违法违规；IR：侵犯他人合法权益；SR：无法满足特定服务类型的安全需求。

模型拒绝风险并给出正向引导的评估如表 4 所示，经过改造的 LLaMA3-8B-SAFE 和 LLaMA3-8B-KG 模型相较于基础 LLaMA3-8B 模型在各项指标上均取得了显著的提升。具体而言，LLaMA3-8B-SAFE 模型在总体及各个风险子类的拒绝率和有害回复率上均实现了最优性能，充分证明了特定领域知识和安全改造在风险问题识别和无害回复生成上的有效性。而 LLaMA3-8B-KG 模型在责任性引导回复方面表现最优，安全价值观改造后的 LLaMA3-8B-SAFE 相比 LLaMA3-8B-KG 有所降低，推测是由于安全价值观改造数据加入训练会使模型回复趋于拒绝和保守，导致责任性引导能力存在些微下降。

表 4　拒绝风险并给出正向引导能力评估结果　（%）

模型	总体			歧视性内容			违反社会核心价值观内容		
	RR-1	RR-2	HR	RR-1	RR-2	HR	RR-1	RR-2	HR
LLaMA3-8B	32. 37	8. 44	16. 45	12. 49	3. 57	14. 80	51. 88	12. 03	17. 67
LLaMA3-8B-KG	49. 78	**42. 64**	1. 95	9. 69	**4. 08**	2. 04	79. 32	**71. 05**	1. 88
LLaMA3-8B-SAFE	**52. 38**	41. 99	**0. 65**	**13. 27**	**4. 08**	**1. 53**	**81. 20**	69. 92	**0. 00**

注：RR-1 表示模型拒绝率；RR-2 表示模型责任回复率；HR 表示模型有害回复率。

3. 3　耗时分析

以知识能力改造为例，本文针对其时间消耗进行了深入分析。在 4 台配备 8 块 A100 显卡的服务器上，笔者进行了全参数微调，所处理的数据总量达到 10 万条。在预设的参数配置下，仅需 15. 67 h 即可完成两个训练周期，从而获取表 2 所示的实验结果。与预训练过程相比，本方法在硬件需求和时间消耗上均实现了显著优化，有效验证了本文提出的快速改造策略的可行性。

3. 4　实验结果讨论

实验结果揭示了安全价值观改造会影响知识能力改造效果。在安全价值观改造后，知识能力在多数领域存在效果下降，但在文本生成方面效果会存在一些提升。除此之外，知识能力改造也可以大幅提升基础模型的安全与价值观，甚至在责任回复方面会优于安全价值观改造。鉴于此，未来的研究焦点应放在如何有效协调安全价值观改造与知识能力改造，以最大化两个模块的协同效能。

4　结论

针对英文大模型在特定文化背景下的快速适配挑战，本文提出一种创新的指令微调流程与方法，该方法基于特定文化下的知识能力和安全价值观数据，无需预训练，即能够高效实现大模型对特定国家与地区文化的短期快速适配。评估结果显示，经过知识能力与安全价值观改造后的大模型，在特定领域内的知识和能力得到了显著增强，且完全符合特定社会文化背景下的价值观和安全标准。

参考文献

[1] 王耀祖，李擎，戴张杰，等. 大语言模型研究现状与趋势 [J]. 工程科学学报，2024，46（8）：1411 – 1425.

[2] 徐月梅，胡玲，赵佳艺，等. 大语言模型的技术应用前景与风险挑战 [J]. 计算机应用，2024，44（6）：1655 – 1622.

[3] CHIANG W L，ZHENG L，SHENG Y，et al. ChatBot Arena：an open platform for evaluating LLMs by human preference [C]// 2024 International Conference on Machine Learning，Part 11 of 75，2024：8359 – 8388.

[4] The Claude 3 model family：Opus，Sonnet，Haiku [EB/OL]. （2024 – xx – xx）[2025 – 02 – 02]. https：//api. semanticscholar. org/CorpusID：268232499.

[5] ACHIAM J，ADLER S，AGARWAL S，et al. GPT-4 technical report [J]. arXiv preprint arXiv：2303. 08774，2023.

[6] TEAM G，ANIL R，BORGEAUD S，et al. Gemini：a family of highly capable multimodal models [J]. arXiv preprint arXiv：2312. 11805，2023.

[7] Introducing Meta LLaMA 3：the most capable openly available LLM to date [EB/OL]. [2025 – 02 – 02]. https：//ai. meta. com/blog/meta-llama-3/.

[8] 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.

[9] CONOVER M，HAYES M，MATHUR A，et al. Free dolly：introducing the world′s first truly open instruction-tuned LLM [EB/OL]. [2025 – 02 – 02]. https：//www. databricks. com/blog/2023/04/12/dollyfirst-open-commercially-viable-instruction-tuned-llm.

[10] PENG B，LI C，HE P，et al. Instruction tuning with GPT-4 [J].

arXiv preprint arXiv：2304.03277，2023.

[11] HONOVICH O，SCIALOM T，LEVY O，et al. Unnatural instructions：tuning language models with（almost）no human labor［C］//61 st Annual Meeting of the Association for Computational Linguistics（ACL 2023），2023：14409－14428.

[12] JI Y，DENG Y，GONG Y，et al. Exploring the impact of instruction data scaling on large language models：an empirical study on real-world use cases［J］. arXiv preprint arXiv：2303.14742，2023.

[13] WANG Y，KORDI Y，MISHRA S，et al. Self-instruct：aligning language models with self-generated instructions［C］//Annual Meeting of the Association for Computational Linguistics，2022.

[14] XU C，SUN Q，ZHENG K，et al. Wizardlm：empowering large language models to follow complex instructions［J］. arXiv preprint arXiv：2304.12244，2023.

[15] XU C，GUO D，DUAN N，et al. Baize：an open-source chat model with parameter-efficient tuning on self-chat data［J］. arXiv preprint arXiv：2304.01196，2023.

[16] 赵朝阳，朱贵波，王金桥. ChatGPT 给语言大模型带来的启示和多模态大模型新的发展思路［J］. 数据分析与知识发现，2023，7（3）：26－35.

[17] YANG D，YUAN R，FAN Y，et al. RefGPT：dialogue generation of GPT，by GPT，and for GPT［C］// Conference on Empirical Methods in Natural Language Processing，2023.

[18] SUN H，ZHANG Z，DENG J，et al. Safety assessment of Chinese large language models［J］. arXiv preprint arXiv：2304.10436，2023.

[19] CHENG Q，SUN T，ZHANG W，et al. Evaluating hallucinations in Chinese large language models［J］. arXiv preprint arXiv：2310.03368，2023.

[20] LIAN S，ZHAO K，LIU X，et al. What is the best model? application-driven evaluation for large language models［C］// Natural Language Processing and Chinese Computing，Part III，2025：67－79.

[21] ZHANG W，LEI X，LIU Z，et al. CHiSafetyBench：a Chinese hierarchical safety benchmark for large language models［J］. arXiv preprint arXiv：2406.10311，2024.

（收稿日期：2025－03－25）

作者简介：

张文静（1995－），女，硕士，工程师，主要研究方向：大模型安全、大模型检索增强。

肖思琪（1994－），女，硕士，工程师，主要研究方向：大模型微调、大模型量化。

廉士国（1979－），通信作者，男，博士，高级工程师，主要研究方向：人工智能、大模型。E-mail：liansg@chinaunicom.cn。

（上接第 11 页）

[5] 龙勇. 基于联邦学习模型对抗样本的鲁棒性增强技术［J］. 数字技术与应用，2025，43（4）：151－153.

[6] AHMAD S，AHMAD N，SHAH I. Stability and sensitivity analysis of cyberattack propagation models in computer networks［J］. European Journal of Pure and Applied Mathematics，2025，18（3）：6336－6336.

[7] ZHANG P，CHENG Q，ZHANG M，et al. A blockchain-based secure covert communication method via shamir threshold and STC mapping［J］. IEEE Transactions on Dependable and Secure Computing，2024，21（5）：4469－4480.

[8] 李冰，刘怀骏，张伟功. 魔方派：面向全同态加密的存算模运算加速器设计［J］. 电子与信息学报，2023，45（9）：3302－3310.

[9] 陆星缘，陈经纬，冯勇，等. 基于同态加密的隐私保护数据分类协议［J］. 计算机科学，2023，50（8）：321－332.

[10] 刘家森，王绪安，余丹，等. 基于同态加密和边缘计算的关键目标人脸识别方案［J］. 信息安全研究，2024，10（11）：1004－1011.

[11] 管桂林，支婷，陶政坪，等. 物联网中多密钥同态加密的联邦学习隐私保护方法［J］. 信息安全研究，2024，10（10）：958－966.

（收稿日期：2025－10－12）

作者简介：

张恒（1981－），男，本科，工程师，主要研究方向：信息系统安全管理。

廖尚斌（1986－），男，硕士，工程师，主要研究方向：网络与信息安全。

张陈颖（1980－），男，本科，高级工程师，主要研究方向：信息系统安全规划。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部