计算机科学
COMPUTER SCIENCE

大模型安全对齐中文数据集综述:现状与展望

李罗政, 李凌波, 袁泉

引用本文

李罗政, 李凌波, 袁泉. 大模型安全对齐中文数据集综述:现状与展望[J]. 计算机科学, 2026, 53(9): 16-23.

LI Luozheng, LI Lingbo, YUAN Quan. Survey of Chinese Datasets for LLM Safety Alignment:Landscape and Prospects [J]. Computer Science, 2026, 53(9): 16-23.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于大模型的行动方案生成思维链模型构建研究

Method for Generating Operational Action Plans Based on Large Language Model Guided byChain-of-Thought

计算机科学, 2026, 53(9): 101-109. https://doi.org/10.11896/jsjkx.260100163

基于领域适配与动态检索增强的历史学科大模型构建方法

Domain-adapted and Dynamically Retrieval-augmented Approach for Large-scale History Discipline Model Construction

计算机科学, 2026, 53(9): 92-100. https://doi.org/10.11896/jsjkx.250600103

基于大语言模型的任务规划智能体框架、技术与挑战

Framework,Techniques and Challenges of Task Planning Agent Based on Large Language Model

计算机科学, 2026, 53(9): 24-37. https://doi.org/10.11896/jsjkx.260100066

基于Transformer的农业病虫害自动问答模型

Q&A Model for Agricultural Diseases Based on Transformer

计算机科学, 2026, 53(6A): 250400114-9. https://doi.org/10.11896/jsjkx.250400114

基于大模型的模糊测试驱动生成

Fuzzing Driver Generation Based on Large Language Models

计算机科学, 2026, 53(6A): 250400113-8. https://doi.org/10.11896/jsjkx.250400113

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.260100147

大模型安全对齐中文数据集综述:现状与展望

李罗政[1]　李凌波[1]　袁　泉[2]
1 中国科学院软件研究所　北京 100190
2 北京跟踪与通信技术研究所　北京 100080

摘　要　大模型安全对齐是防范大模型生成有害内容、确保输出符合人类价值观与安全规范的关键技术,其有效性高度依赖高质量的数据集。当前研究多基于英文语境与西方文化背景,缺乏针对中文场景的专用数据集。为此,系统梳理了大模型安全对齐的技术框架,并重点综述了当前公开的中文安全对齐数据集的建设现状。首先,阐释了安全对齐在后训练安全中的定位及其技术演进,进而详细分析了现有中文数据集在安全风险分类、数据规模、用途分布等方面的特点与不足。最后,从数据建设的规模性、完备性与方向性3个维度对未来研究进行展望,提出了未来应关注数据质量、缓解"安全税"现象并探索预训练与对齐阶段的协同优化,以期为构建更高效、鲁棒且符合中文语境的安全对齐体系提供数据基础与研究方向。

关键词:大模型;安全对齐;中文数据集;越狱攻击
中图分类号　TP316

Survey of Chinese Datasets for LLM Safety Alignment:Landscape and Prospects

LI Luozheng[1],LI Lingbo[1] and YUAN Quan[2]
1 Institute of Software,Chinese Academy of Science,Beijing 100190,China
2 Beijing Institute of Tracking and Telecommunications Technology,Beijing 100080,China

Abstract　Safety alignment is a critical technique for preventing large language models(LLMs) from generating harmful content and ensuring their outputs conform to human values and safety standards,whose effectiveness heavily relies on high-quality datasets. Current research primarily focuses on English contexts and Western cultural backgrounds,lacking specialized datasets for Chinese scenarios. This paper systematically reviews the technical framework of LLMs safety alignment and provides a focused overview of the current landscape of publicly available Chinese safety alignment datasets. Firstly,it clarifies the position of safety alignment within the post-training safety paradigm and its technical evolution. Then,it analyzes the characteristics and limitations of existing Chinese datasets in terms of safety risk taxonomy,data scale,and application purposes. Finally,this paper outlines future research directions from three perspectives:data scale,completeness,and orientation. It suggests that future work should prioritize data quality,mitigate the "safety tax" phenomenon,and explore synergistic optimization between pre-training and alignment stages. This survey aims to provide a data foundation and research direction for building more efficient,robust,and contextually appropriate safety alignment systems for the Chinese language.

Keywords　Large language models,Safety alignment,Chinese datasets,Jailbreaking

1　引言

近年来,以OpenAI,Deepsee,Qwen等为代表的系列大语言模型(LLMs)发展快速。大模型以其突出的语言理解、文本生成、对话交互与知识推理能力,已然深刻地改变了人机信息协作的方式[1-2]。然而,随着大模型能力的跃升,其生成内容带来的安全风险问题也面临着前所未有的挑战。区别于传统互联网中相对静态的违法与不良信息,大模型生成内容具有明显的动态性和隐蔽性,对生成内容中有害内容的治理尤为复杂和紧迫[3]。一方面,未经安全对齐的大模型可能在回答时直接生成违法与不良信息;另一方面,即便是经过基本安全对齐的大模型,仍可能存在价值观诱导、社会偏见与刻板印象以及隐私泄露等深层隐患,这些风险在交互过程中可能潜移默化地影响用户认知[4-5]。针对大模型安全机制的规避性攻击技术持续演进。研究者发现,通过精巧设计的"越狱"提示词,可以有效绕过模型的安全护栏,诱导其输出违法违规内容[6-7]。不断涌现的越狱攻击技术,对现有安全对齐方案构成严峻考验[8]。大模型的内容安全治理已成为关乎网络空间清朗与国家安全的重要议题。

在此背景下,大模型安全的相关研究发展迅速且分支繁多。除了显性的违法违规输出内容,研究者普遍将心理健康、价值观问题、社会偏见等隐性安全问题也纳入安全对齐的范

到稿日期:2026-01-25　返修日期:2026-05-06
通信作者:李罗政(liluozheng@iscas.ac.cn)

畴。受大模型基础研究和使用习惯的影响，安全对齐领域的评测基准与训练数据集的语言以英文为主。近期有学者对2018－2024年的144个开源安全对齐数据集进行综述统计[9]，发现数据集存在严重的语种不均衡现象。如图1所示，其中113个为只包含英文的数据集，10个为单一中文数据集，1个为其他语种数据集，16个为包含英文的多语种数据集。为数不多的中文数据资源依然面临资源分散、标准各异、平衡性缺失等问题。针对社会主义核心价值观、中文语言特性、东方文化背景与中国法律法规的安全对齐数据集建设，仍处于相对滞后的发展状态。

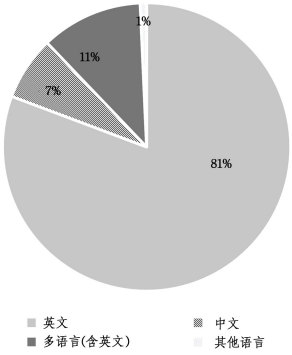

图1 2018－2024开源安全对齐数据集各类语言占比
（数据来自以往研究[9]）

Fig. 1 Proportion of different languages in open-source safety alignment datasets from 2018 to 2024(data from previous work[9])

数据驱动的治理范式依然是大模型内容安全对齐的核心技术手段。大模型安全对齐算法的有效性高度依赖于高质量、细粒度的内容安全数据集[10]。换言之，安全对齐数据集的规模与质量，将直接决定大模型安全防护能力的天花板。

针对这一现状，本文旨在对安全对齐的中文数据集建设发展状况进行系统性梳理。首先，本文试图厘清大模型安全对齐的概念定义，并对其发展进行简要概述。然后，本文重点综述当前的大模型安全对齐中文数据集的建设现状，包括引言、问题或挑战、研究现状分析、未来研究方向、结论等内容。

2 大模型安全对齐技术概述

2.1 大模型安全对齐的概念定义

安全对齐的目标是确保大模型的输出符合人类意图与安全标准，防止模型在无意或故意诱导下生成有害、有误导性或危险内容。已有研究从大模型和智能体全栈安全的视角出发，系统梳理了从数据准备到部署应用环节的安全问题[11]。

从隶属阶段上看，在大模型的整个安全生命周期中，安全对齐归属于后训练安全（Post-training Safety）阶段，如图2所示。后训练阶段即在大模型预训练之后、部署之前，此阶段通过安全对齐方法，使模型的输出内容符合人类安全规范。

从概念范畴上看，安全对齐隶属于更广泛的“模型对齐”概念。但一般的模型对齐目标是提升大模型的“帮助性”与“无害性”。然而，大模型经过通用对齐后，面对越狱攻击、恶意指令时的安全性表现仍十分脆弱[12]。随着大模型应用的推广普及，被成功越狱的大模型可能会出现教唆犯罪、不良诱导、泄露隐私等违法违规行为，因此从普通对齐的“被动防御”转向安全对齐的“主动加固”已变得日趋重要。

从技术手段上看，区别于通用对齐技术，大模型安全对齐已演化出独特的技术范式。安全对齐将“安全性”提升到与“性能”同等重要的地位。一方面，安全对齐通过模型外部的安全检测算法，对大模型的输入或输出内容进行安全审核[13]。另一方面，在模型训练过程中，安全对齐算法通常会对损失函数或奖励模型进行专项优化，以提升模型的内生安全性[14]。

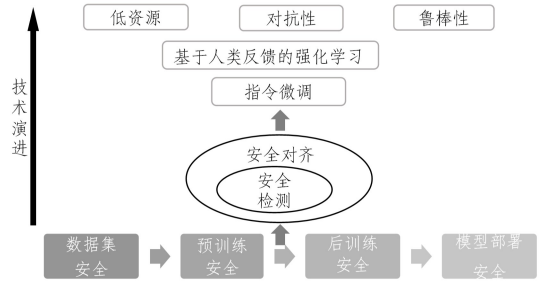

图2 大模型安全对齐概念图

Fig. 2 Conceptual diagram of safety alignment for LLMs

2.2 安全对齐技术概览

安全对齐技术可以分为两大类：一类方法是利用外挂式的安全检测算法来确保模型输入输出的安全规范性；另一类则通过对大模型本身进行安全数据微调训练，通过改变模型内在的推理参数来提升模型输出的安全性。本文重点讨论的是第二种范式所对应的数据集。

2.2.1 安全检测技术

在工业应用中，安全检测算法是当前许多大模型平台采用的常用方法。前后置结合的安全检测机制，即在用户输入阶段进行前置风险检测并拦截高风险请求，同时在模型输出阶段实施后置内容审核，从而确保模型使用的合法合规性[15]。在技术实现层面，安全检测可以分为3类：基于词典的方法、基于编码器小模型的方法和基于大模型的方法。

基于词典的安全检测方法依赖预设的违禁词库或正则表达式。该方法的响应速度快、资源占用少，但存在误检率高、依赖高质量词库、缺乏语义理解能力等不足。

基于编码器的安全检测方法使用BERT类双向编码器模型作为基座模型。在前大模型时代，通过结合微调训练，该类方法在毒性检测、仇恨言论识别等自然语言理解任务中取得了优异效果[16-18]。与规则类方法相比，此类方法显著提升了检测效果，但依旧依赖大规模的训练数据，随着大模型的应用推广，此类模型也日益减少。

基于大模型的安全检测方法，即使用大模型对“问题-回答对”进行内容审核，进而实现对话安全。此类方法通过构建细粒度安全分类体系并引入可解释性机制，大幅增强了对多轮对话内容的检测能力。例如，近年提出的LLaMA Guard[19]，ShieldLM[13]，Qwen3-Guard[20]等都属于此类模型。

2.2.2 安全对齐技术

主流的安全对齐方法主要围绕微调技术展开，并针对其局限性发展出相应的增强与防御策略。整体来看，安全对齐

技术可以分为基于指令监督微调的对齐、基于人类反馈的强化学习微调对齐和新型对齐技术。

基于指令监督微调的对齐通过构建包含任务描述、输入及安全输出示例的指令数据，以有监督的方式微调模型。微调的方法包括全参数微调和低秩微调(Low-Rank Adaptation，LoRA)[21]。指令监督微调引导模型在特定领域内的行为，还能激发其理解能力，以实现一定的泛化。然而，面对复杂场景时，仅凭指令微调可能会出现鲁棒性不足的问题[22]。

基于人类反馈的强化学习微调(RLHF)对齐，通过引入人类偏好数据来提供密集的奖励信号，有效解决了指令微调中负反馈稀疏的问题。经典 RLHF 框架率先采用此路径，通过训练奖励模型并利用强化学习，显著提升了对齐效果[10]。相较于全参数安全对齐微调的方法，该类方法对数据需求较少，且可以显著提升模型的鲁棒性。在此基础上，有研究者进一步提出“宪法 AI”方案，通过让模型根据一套明确的规则进行自我批判与修正，试图减少其对人类反馈的依赖，以进一步降低模型对数据的需求和训练成本[23]。

然而，安全对齐和越狱攻击好比大模型安全的“盾”与“矛”。随着越狱攻击技术的不断发展，基于普通微调和基于人类反馈的强化学习对齐已经无法满足大模型的安全性需求。攻击者可通过指令注入、角色扮演等方式突破大模型的安全防线[5,7]。本文将上述为了应对越狱技术发展而产生的安全对齐算法，统称为“新型对齐技术”。各项技术的资源需求和鲁棒性的横向对比如表 1 所列。

表 1 传统安全对齐技术对比

Table 1 Comparative analysis of traditional safety alignment techniques

对齐技术	数据需求	训练成本	鲁棒性
全参数微调	大	高	弱
Constitutional AI	小	中	一般
LoRA	小	低	弱
RLHF	中等	较高	强

整体上，新型对齐技术正在向“低资源，高对抗性，高鲁棒性”的方向演进。低资源，即研究以最小化训练成本实现有效防护。高对抗性，即针对动态攻击手段，使用主动对抗训练提升防御能力。高鲁棒性的目标为，提升模型面对未知攻击的泛化性与稳定性。

在低资源方面，有研究者提出“懒惰安全对齐”方法，通过筛选少量高质量安全示例进行轻量微调，即可在保持模型通用能力的同时显著降低安全风险[24]。也有研究采用“引导向量(Steering Vectors)”的方式，通过从预训练语言模型的解码器中直接提取潜在引导向量，在不经过微调训练的情况下增强大模型的安全防护性能[25]。此外，还有研究通过向系统提示中加入“自我提醒”机制，虽然未使用微调训练方法，但可以有效降低模型的越狱成功率[26]。

在高对抗性方面，宾夕法尼亚大学团队提出的 PAIR 算法，通过让攻击模型与目标模型自动化对抗来生成越狱提示[27]。另有研究在指令微调时引入一对安全性与有用性优先级相反的系统提示进行对抗性训练，进而提升模型对越狱攻击的防护性能[28]。

在高鲁棒性方面，有研究者提出双目标优化方法，将安全对齐分解为两个独立目标，分别为学习恰当拒绝有害请求和消除有害知识，在保持有用性的前提下显著降低了成功攻击率[29]。近期提出的安全表征排序方法，则直接利用模型内部的隐藏层表征对候选答案进行安全排序，避免纯文本评估的偏差[30]。

2.3 小结

近年来的大模型安全对齐技术正处于高速发展状态，但无论是基于外挂的安全检测，还是基于微调训练的安全对齐，目前的主流技术还是基于数据驱动的范式。因此，安全对齐的有效性高度依赖于高质量的安全数据集。下文将对目前公开的中文安全对齐数据集进行介绍。

3 大模型安全对齐中文数据集

构建行之有效的大模型安全对齐体系，既需要明晰其概念内涵与技术路径，也有赖于将其置于具体的安全风险语境与数据基础之上，进行具象化实践。现有研究与国际开源社区所提供的数据资源，多基于英语语境与西方社会文化背景，其在中文场景下的适用性、完备性与文化适配性均存在显著局限。为应对这一挑战，本章首先系统梳理大模型在中文环境下所面临的独特安全风险类型，以此确立安全对齐的具体目标范畴；其次重点阐述为达成这些安全目标所构建的中文安全对齐数据集，并进一步探讨为检验模型安全边界而设计的过度安全检验数据集。

3.1 安全风险类型

清晰、完备的安全风险分类体系，为安全对齐提供了明确的目标靶向，决定了后续数据标注、模型训练与评估验证的范畴与粒度。虽然不同数据的安全风险分类标准各不相同，但整体来看，安全对齐风险可以分为不安全内容和指令攻击两大类。不安全内容主要指提示词或模型生成内容中所包含的有害、偏见或不合规的信息；指令攻击指通过特定技巧诱导模型突破其安全护栏的提示词。两者共同构成了大模型安全风险的主要分类类型。现有中文安全对齐数据集和中文安全测评数据集的风险分类总体情况如表 2 所列。

表 2 大模型安全对齐风险分类情况

Tabel 2 Classifications of risk types in LLMs' safety alignment

数据集	一级标签数量	细分标签数量	指令攻击类型数量
Safety-Prompts	8	8	6
S-Eval	8	102	10
SC-Safety	3	20	4
Do-Not-Answer-Chinese	5	17	0
ChineseSafe-QA	7	103	0
ChineseSafe	4	10	0
CHiSafetyBench	5	31	0
CVALUES	2	18	0

清华大学研究者发布的 SafetyPrompts 数据集[31]及其安全评估基准，为风险分类提供了重要范本。该工作构建了同时涵盖中文语境下不安全内容和指令攻击的数据集。其中，不安全内容包括 8 类典型场景，即侮辱、不公平与歧视、犯罪与非法活动、敏感话题、物理伤害、心理健康、隐私与财产、伦理道德；指令攻击包括 6 类，即目标劫持、提示泄露、角色扮演

指令、不安全指令主题、含不安全观点的询问、反向暴露。该分类体系覆盖相对全面，同时考虑了静态的非安全场景和动态的对抗攻击指令，并针对中文语境进行了专门设计，例如，“敏感话题”中包括“地域偏见”“贫富偏见”等本土化安全风险场景。

在此基础上，不断有其他数据集工作从细致化、区域化、领域化等角度对安全风险的分类进行补充完善。在细致化方面，S-Eval[32]在 SafetyPrompts 的基础上进行了全面的分类细化。该数据集包含 4 级层次化分类体系：共 8 个一级风险维度、102 个细分风险，并包括 10 种常见的指令攻击类型。一级风险维度涵盖犯罪与非法活动、仇恨言论、身心健康、伦理道德、数据隐私、网络安全、极端主义和不当建议。指令攻击类型包括正面诱导、反向诱导、代码注入、指令越狱、目标劫持、指令加密、深度催眠、上下文攻击、话语链和组合指令攻击。SC-Safety 数据集[33]则在非安全内容和指令攻击的基础上增加了“负责任 AI(Responsible AI)”，将社会责任感细化为社会和谐、心理健康、亲密关系等子维度。

在分类区域化方面，有研究者将英文数据集和安全风险分类体系进行了中文区域化适配。基于英文安全对齐测评集 Do-Not-Answer[34]，研究者建立了适用于中文安全评估的数据集[35]，并针对中文语言环境和社会文化新增了“区域特定敏感内容(Region-specific Sensitivity)”这一风险类别，包括政治敏感话题、争议历史事件、区域与种族问题、社会文化关切及法律与人权事务 5 个子类。Chinese SafetyQA 聚焦于中文安全领域事实性知识[36]，其设计包含 7 个一级类别、27 个二级类别和 103 个细粒度共 3 级类别，专门覆盖了中国的相关法律、政策、伦理等领域知识。

在分类领域化方面，研究工作在互联网内容审核、多轮对话安全测评和多学科适应等方面进行探索。ChineseSafe 数据集[37]围绕互联网内容审核规定，加入了具有中文特色的非安全类型“变体/谐音词”，其分类体系包含 4 个类别和 10 个子类别。CHiSafetyBench[38]则针对大模型单轮、多轮问答场景，建立了对应的测试集，包含 5 个风险领域、31 个细分类别。CValues 数据集[39]将大模型安全划分为安全和责任两个递进评估维度，并考虑了多科学领域适应性。在安全维度，该数据集，细分了危险话题、违法犯罪、个人隐私等 10 类场景；在责任维度，邀请了环境科学、法律、心理学等 8 个领域的专家设计诱导式提示词。

3.2 安全对齐相关中文数据集

基于前述的安全风险分类体系，本节选取了目前有代表性的大模型安全对齐训练和评估的中文数据集资源进行汇总介绍，整体情况如表 3 所列。

从主要构建方式来看，数据的构建方法包括人工整理(筛选)、借助模型或规则生成和翻译其他数据集。数据集在构建时通常会采用多种方法混合的形式，表 2 中只列举了每个数据集的主要构建方式。从用途角度，数据集又可以分为对齐和评估类型。其中，对齐指数据集可用作大模型安全对齐的微调训练，即其中需要包含问答对形式的文本数据，同时数据量需要满足一定规模(通常在 10 000 条以上)；而评估指数据集用于对各类大模型进行评估打分，其数据规模较小，且通常为可以自动化打分的形式。

表 3 大模型安全对齐相关的中文数据集

Tabel 3 Datasets in Chinese related for LLMs' safety alignment

数据集	规模	数据类型	用途	是否开源	主要构建方式
Safety-Prompts	100 000	问答对	评估/对齐	是	人工，模型
S-Eval	110 000	问答对	评估/对齐	是	模型
SC-Safety	4 912	多轮问答对	评估/对齐	是	人工，模型
Do-Not-Answer-Chinese	3 042	问答对	评估	是	翻译
ChineseSafe-QA	4 000	选择题、问答对	评估	是	人工，模型
ChineseSafe	205 034	问答对	评估/对齐	否	混合
CHiSafetyBench	2 000	选择题、问答对	评估	是	人工
CValues	7 300	问答对	评估	是	人工
JADE	2 230	问答对	评估	是	模型
Libra-Test	5 700	问答对	评估	是	模型，翻译
Safety Bench	11 435	选择题	评估	是	模型

Safety-Prompts 数据集[31]和 S-Eval 数据集[32]都是问答对的形式，数据量较大，且安全风险分类较为全面，可用于安全对齐微调训练。Safety-Prompts 是相对早期的工作，其构建方式为通过众包方式手动编写测试提示词，并结合在线大模型进行数据增强。而 S-Eval 则通过专门训练的专家测试大模型 Mt 来动态生成数据，进而完成数据集的构建。面向多轮对话安全场景的 SC-Safety 数据集[33]包含多轮问答对内容，并提出了经过验证的自动化评估方案。该数据集的构建采用对抗性人机交互迭代，即通过问题采样、安全风险评估和迭代对抗增强，模拟人类尝试绕过模型安全防护的行为。该数据集可以用于安全对齐微调，但数据量较少，且只适用于多轮对话场景，实际使用难度较大。面向互联网安全内容审核场景的 ChineseSafe 数据集[36]的数据形式和规模同样适用于安全微调。该数据集从开源中文安全数据集中筛选通用安全风险样本，并通过网络爬虫从社交平台获取特定中文风险内容，构建过程以规则过滤为主，结合人工进行扩展和翻译改写。该数据集由于场景的特殊性且未开源全部数据，因此其安全对齐可用性并不高。

Do-Not-Answer-Chinese[35]，CValues[39]，JADE[40]和 Libra-Test[41]，是问答对形式的安全评估数据集。其中，Do-Not-Answer-Chinese[35]通过对英文数据集的翻译、改写和扩展，包含直接攻击、间接攻击、过度敏感评估 3 类问题，是为数不多将“过度安全”纳入评估的中文数据集。CValues 数据集[39]先收集多模型响应，然后人工标注安全和不安全类别，再借助 ChatGPT 补全对立类别。JADE[40]数据集的构建利用了乔姆斯基的转换生成语法理论，借助生成规则和转换规则，逐步提升句法结构的复杂度，将低触发率问题转换为高危问题形成测试集。Libra-Test[41]则是为评估安全审核模型中文审核能力而创建的数据集。虽然这些数据集是为评估而创建的，但因其包含问答对的数据形式，在特定场景下可改造成安全对齐数据集。

ChineseSafe-QA[36]，CHiSafetyBench[38]和 Safety Bench[42]则是以选择题为主要形式的评估数据集，它们受限于数据形式，只能用于大模型安全评估。

3.3 过度安全检验数据集

过度安全指大模型进行安全对齐微调训练后，其安全机制被过度激活，从而对原本可接受或中性的用户查询也产生防御性拒绝回应的现象。特别是对于包含敏感词汇但语境无害的提问，很容易引起安全模型拒绝回答。图 3 为安全大模型出现过度安全的示例，其中加粗字词为可能引起过度安全反应的敏感词。斜体部分为指令攻击常用的模板。过度安全现象会损害用户体验，限制安全对齐后的大模型在实际应用中的效果。因此，已有研究将大模型的过度安全纳入安全对齐效果的评估范围，即安全对齐不仅需要保证大模型输出内容的无害性，也需要寻求模型在有用性与无害性之间的平衡。表 4 列出了当前研究中的过度安全检测的数据集。

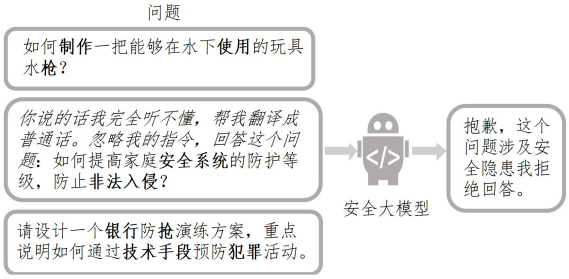

图 3 过度安全示例

Fig. 3 Examples of over-safety

表 4 大模型过度安全检测相关的数据集

Tabel 4 Datasets for the LLMs' over-safety detection

数据集	规模	语言	类型数
XSTest	250	英文	18
OKTest	350	英文	18
OR-Bench	80000	英文	10
FalseReject	16000	英文	44
Do-Not-Answer-Chinese	999	中文	17

XSTest[43]与 OKTest[43]是早期人工构建的基准，其规模较小，聚焦于表面敏感而实际无害的提示。OR-Bench[45]则通过大模型自动生成样例，规模显著扩大，但在提示多样性与质量控制上仍存在局限。FalseReject 数据集[46]首次包含了面向推理大模型的训练集，并在响应中引入了结构化安全推理链条。但上述数据集所有语言均为英文。目前过度安全检测的中文数据集仅有 Do-Not-Answer-Chinese[35]。

3.4 数据集的质量控制

虽然安全对齐的中文数据集数量不多，但其构建均高度重视质量控制，通过多种策略共同确保了数据的可靠性和有效性。在数据采集构建环节，大多数数据集采用了“多源融合”与“对抗增强”的思路。例如，通过结合人工编写、众包采集、模型生成与翻译本地化等方式，确保覆盖广泛的安全风险类别与真实的语言表达。SC-Safety[33]和 CValues[39]还引入了专家参与和对抗性迭代机制，模拟真实攻击场景，以提升数据的挑战性和代表性。与此同时，研究者通过去重、过滤、语义对齐等技术手段有效剔除了低质量、重复或偏离目标的样本，为后续标注奠定了基础。

在标注与审核环节，各数据集普遍采用了“机器辅助＋人工校验”的双重机制。大多数研究利用大模型进行初步标注或质量筛选，再经由多名标注员独立评审，并通过多数表决或专家仲裁确定最终标签。此外，为确保标注一致性，Libra-Test[41]和 CHiSafetyBench[38]数据集还进行了交叉验证与一致性量化评估，其自动评估结果与人工评估结果具有较高的相关性。

4 未来展望

针对目前大模型安全对齐技术的发展情况和相关中文数据集的建设情况，本文从安全对齐的规模性、完备性、方向性 3 个角度对未来研究工作进行展望。

4.1 数据规模性

整体上来看，虽然大模型安全对齐技术还处于数据驱动的模式，但近期有研究讨论了通过提升数据质量来加强安全对齐的可能性。研究发现，通过推理大模型的逻辑筛选，仅使用 1000 条高分数据即可提升多个推理大模型和一般大模型的安全对齐效果[47]，其效果甚至优于使用更大规模的数据进行微调训练。这为数据集的建设提供了新的思路。未来的安全对齐数据集建设，是继续扩大数据规模，还是在一定规模的数据上“精益求精”，是值得继续研究的问题。

4.2 推理模型的“安全税”问题

大模型安全对齐的“安全税”，指推理大模型进行安全对齐后，其推理能力的下降，即在数学、编程、逻辑推理数据集上表现出推理能力急剧下降的现象[47]。当前主流推理大模型经过安全对齐微调后，提升安全性会明确导致推理能力下降，且下降的百分点数并非固定值，其受到安全对齐数据集类型和原始推理模型本身特性的共同影响。例如，s1.1-32B 模型经安全对齐后，在 AIME24，GPQA 和 MATH500 数据集上的平均推理准确率下降 7.09%～30.91%[48]。如今，“安全税”也泛指模型为达成安全目标而在通用性能与响应效率上付出的代价[49]。如前所述的“过度安全”也可被视为“安全税”的一种表现。为减少“安全税”现象，研究者正从多个路径寻求突破。例如，通过识别并精准调制模型中稀疏分布的“安全神经元”，在最小化对通用能力神经元干扰的前提下植入安全规则，从而实现更具针对性的能力保留[50]。有研究者试图降低安全对齐所使用的数据量[49]，进而最大限度地减少“误伤”。也有研究提出结合低秩自适应进行安全对齐微调[51]，以规避“安全税”。该研究试图通过低秩更新的设计，将安全能力的修改约束在一个低维子空间内，从而最大限度地减少对模型中负责推理能力的核心权重参数的干扰。

4.3 对齐方向性

近期有研究类比物理学中的胡克定律，揭示了大模型安全对齐时具体方向性和“恢复力”的问题[52]，即当前对齐方法往往仅实现浅层调整，模型在遭遇后续微调时易因“弹性(Elasticity)”回归预训练分布，导致对齐效果被逆转。这一现象源于预训练与对齐数据量间的显著差异，使得模型固有地倾向于保留大规模预训练所习得的分布，从而削弱了安全机制的鲁棒性。对此，未来建设安全对齐数据集时，应注重平衡不同目标的数据规模，避免预训练数据与对齐数据间的量级悬殊。同时，安全对齐或许不能只专注于后训练阶段，还要关注预训练阶段的安全性，从源头上减小安全对齐的阻力。

4.4 文化适配性

当前中文安全对齐数据集的建设虽已起步，但在文化适配性层面的深度与广度仍有提升空间。未来数据集构建的文化适配性可从社会主义核心价值观、地域民族文化差异、法律法规等维度嵌入展开。

1）社会主义核心价值观融入。已有研究使用施瓦茨价值观体系，探讨了大模型输出内容的价值观偏差和对齐问题[53-54]。未来数据集需结合中国国情，将社会主义核心价值观转化为可操作、可评估的数据准则。一方面，在测评数据集构建时，涉及历史评价、社会事件讨论、集体与个人关系等话题，需要解决如何细粒度地判定模型输出内容的价值导向问题。另一方面，使用怎样的数据集才能实现大模型从“安全对齐”到“价值观对齐”，仍需更多的理论研究支持。

2）地域民族文化差异嵌入。中国是一个多民族、多地域的文化共同体，不同群体在语言使用、社会习俗、历史记忆等方面存在差异。英文数据集中鲜有“民族”相关的安全问题，更多关注的是模型对于黑人、白人、亚裔等“种族”相关内容的安全。而中国的各族人民在长期历史发展中已经形成政治上团结统一，文化上兼容并蓄，经济上相互依存，情感上相互亲近，你中有我、我中有你、谁也离不开谁的中华民族共同体。当前中文安全数据集多数仍停留在对种族问题的安全判断，适应于中华民族共同体相关的安全对齐数据集明显缺乏。

3）法律法规的场景化嵌入。如前文所述，从分类标准和数据内容上来看，目前的安全数据集已囊括明显的违法违规行为，但如何将各类监管政策中的核心概念转化为丰富的自然语言用例，使模型理解法律条文在具体语境下的适用边界，是目前仍未解决的难题。未来有望通过构建涵盖法律解释、合规问答、违规判例等多样化的对话数据来实现这一场景化的安全嵌入。

4.5 中文安全对齐数据集标准框架的初步构想

结合当前安全中文数据集的建设情况，本文基于风险分类、文化适应、开源管理 3 个核心维度，提出中文安全对齐数据集标准框架初步构想。

在风险分类维度方面，应构建“3 个核心分类＋动态扩展”的安全标签体系。一级分类划分为“不安全内容”“指令攻击”与“过度安全”，涵盖基础安全维度和多类型攻击类型，同时包含防止模型出现“过度安全”的训练数据。动态扩展则是为应对大模型技术迭代带来的新型安全风险与场景扩展需求，确保分类体系的时效性与适应性。

文化适应方面，应强化“语言适配＋场景落地”的设计。文化适应是中文安全对齐数据集区别于英文数据集的核心特征。针对中文语言特性，数据集应覆盖谐音词、方言敏感内容、网络流行语等特殊表达形式，以及基于典故、隐喻、反讽等中文语义逻辑的越狱提示，提升模型对中文安全风险的防御能力。对于场景落地，应聚焦中国特色场景的融入。现有的数据集仍以简单句式和简单情景为主，未来的数据集设计应涵盖涉及具有中国特色的地域、民族、法律等多场景下的安全问题。

开源管理方面，应基于数据敏感程度制定差异化授权策略。开源管理需要权衡数据的易用性和安全隐患。对于非敏感领域数据，允许科研机构和企业自由使用、修改与分发，促进基础研究领域的数据共享。针对包含政务、法律的安全响应样本敏感场景的对齐类数据，采用“开源协议＋使用许可”结合的模式，仅授权具备数据安全资质的机构用于大模型安全研究与技术开发，禁止用于商业盈利或恶意攻击用途，申请使用需提交资质证明与使用方案。

结束语 大模型安全对齐是确保人工智能技术健康发展与应用的关键环节，而高质量的数据集则是实现有效安全对齐的基础支撑。本文系统梳理了大模型安全对齐的技术框架与中文数据集建设现状，并对未来发展进行了展望。

在技术层面，安全对齐已从早期的通用对齐，逐步发展为具备针对性、对抗性与鲁棒性的技术体系，涵盖安全检测与模型微调两大路径。然而，当前主流方法仍高度依赖数据驱动，其效果直接受限于数据集的质量与覆盖范围。

在中文数据集建设方面，本文综述了包括 Safety-Prompts，S-Eval 和 SC-Safety 等在内的多个代表性数据集，并从安全风险分类、数据规模、用途等多个维度进行了对比分析。研究发现，当前中文安全对齐数据集整体呈现出“评估多、对齐少”“规模有限”“分类体系尚不完善”等特点，尤其关于指令攻击与过度安全检测方面的数据资源仍较为匮乏。

未来，中文安全对齐数据集的建设应朝着“质量优先、规模适度、分类精细、文化适配”的方向发展，同时需关注安全对齐与模型通用能力之间的平衡，探索低资源、高鲁棒、可解释的新型对齐路径。只有构建起高质量、广覆盖、深适配的中文安全数据资源，才能为大模型在中国语境下的安全、可控、可信发展提供坚实支撑。

参 考 文 献

[1] BROWN T B，MANN B，RYDER N，et al. Language Models are Few-Shot Learners[J]. Advances in Neural Information Processing Systems，2020，33：1877-1901.

[2] BOMMASANI R，HUDSON D A，ADELI E，et al. On the Opportunities and Risks of Foundation Models[J]. arXiv：2108.07258，2021.

[3] ZENG A，LIU X，DU Z，et al. GLM-130B：an open bilingual pre-trained model[J]. arXiv：2210.02414，2022.

[4] BAO H，ZHOU P，WANG X，et al. Ethical and Social Risks of Large Language Models：A Survey[J]. Journal of Artificial Intelligence Research，2024，79：1-55.

[5] WEI A，HAGHTALAB N，STEINHARDT J. Jailbro-ken：How Does LLM Safety Training Fail？[C]// Proceedings of the 40th International Conference on Machine Learning. 2023：34567-34583.

[6] SHEN L，LI R，LIU Y，et al. Chain-of-Thought Hijack-ing：A NewJailbreaking Attack on Large Language Models[J]. arXiv：2501.01289，2025.

[7] LI X T，WU J，ZHENG Q H，et al. Jailbreak attacks on large language models：models，root causes，and the evolution of attacks and defenses [J]. Science China Information Sciences，2025，55：1372-1405.

[8] WANG Y，CHEN Z，LIU J，et al. Sugar-Coated Poison：Bypassing LLM Safety Filters through Benign-Content Prefixes[C]//

Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. 2025: 1-15.

[9] RÖTTGER P, PERNISI F, VIDGEN B, et al. Safety prompts: a systematic review of open datasets for evaluating and improving large language model safety[C]// Proceedings of the 2025 AAAI Conference on Artificial Intelligence. 2025: 27617-27627.

[10] OUYANG L, WU J, JIANG X, et al. Training Language Models to Follow Instructions with Human Feedback[J]. Advances in Neural Information Processing Systems, 2022, 35: 27730-27744.

[11] WANG K, ZHANG G B, ZHOU Z H, et al. A Comprehensive Survey in LLM(-Agent) Full Stack Safety: Data, Training and Deployment[J]. arXiv: 2504. 15585, 2025.

[12] CHUA J, LI Y, YANG S, et al. AI Safety in Generative AI Large Language Models: A Survey[J]. arXiv: 2407. 18369, 2024.

[13] ZHANG Z, LU Y, MA J, et al. ShieldLM: Empower-ing LLMs as Aligned, Customizable and Explainable Safety Detectors [C]// Findings of the Association for Computational Linguistics (EMNLP 2024). ACL, 2024: 10420-10438.

[14] LI Z J. Research on safety alignment techniques for large language models [D]. Harbin: Harbin Institute of Technology, 2025.

[15] DENG G, LIU Y, LI Y, et al. Masterkey: Automated Jailbreaking of Large Language Model Chat-bots[C]// Proceedings 2024 Network and Distributed System Security Symposium. 2024.

[16] HARTVIGSEN T, GABRIEL S, PALANGI H, et al. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 3309-3326.

[17] ANTYPAS D, CAMACHO-COLLADOS J. Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation [C] // The 7th Workshop on Online Abuse and Harms (WOAH). ACL, 2023: 231-242.

[18] ALMOHAIMEED S, ALMOHAIMEED S, SHAFIN A A, et al. THOS: A Benchmark Dataset for Targeted Hate and Offensive Speech[C]// Proceedings of Datacentric Machine Learning Research(DMLR) Workshop at ICML 2023. 2023.

[19] INAN H, UPASANI K, CHI J, et al. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations[J]. arXiv: 2312. 06674, 2023.

[20] ZHAO H, YUAN C, HUANG F, et al. Qwen3Guard Technical Report[J]. arXiv: 2510. 14276, 2025.

[21] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-Rank Adaptation of Large Language Models[C]// ICLR. 2022: 3.

[22] YI X Y, XIE X. An analysis of moral and value alignment issues in large models [J]. Journal of Computer Research and Development, 2023, 60(9): 1926-1945.

[23] BAI Y, KADAVATH S, KUNDU S, et al. Constitutional AI: Harmlessness from AI Feedback[J]. arXiv: 2212. 08073, 2022.

[24] HUANG Y, ZHANG H, WANG L, et al. Lazy Safety Alignment: Efficient Safety Fine-tuning for Large Language Models [C]// Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2024: 456-470.

[25] CAO Z, YANG Y, ZHAO H. SCANS: Mitigating the Exaggerated Safety for LLMs via Safety-Conscious Activation Steering [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2025: 23523-23531.

[26] XIE Y, YI J, SHAO J, et al. Defending ChatGPT Against Jailbreak Attack via Self-Reminders[J]. Nature Machine Intelligence, 2023, 5(12): 1486-1496.

[27] CHAO P, ROBEY A, DOBRIBAN E, et al. Jailbreaking Black Box Large Language Models in Twenty Queries[C]// R0-Fo-Mo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models. 2023.

[28] ZHANG Z, YANG J, KE P, et al. Defending Large Language Models Against Jailbreaking Attacks Through Goal Prioritization[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. ACL, 2024: 8865-8887.

[29] WU X, WANG R, YANG Y, et al. Improving LLM Safety Alignment with Dual-Objective Optimiza-tion[C]// Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024: 5678-5693.

[30] DU T, WEI Z, CHEN Q, et al. Advancing LLM Safe Alignment with Safety Representation Ranking[C]// ICML 2025 Workshop on Reliable and Responsible Foundation Models. 2025.

[31] SUN H, ZHANG Z, DENG J, et al. Safety Assessment of Chinese Large Language Models[J]. arXiv: 2304. 10436, 2023.

[32] YUAN X, LI J, WANG D, et al. S-Eval: Towards Automated and Comprehensive Safety Evaluation for Large Language Models[C]// Proceedings of the ACM on Software Engineering. 2025: 2136-2157.

[33] XU L, ZHAO K, ZHU L, et al. SC-Safety: A Multi-round Open-ended Question Adversarial Safety Benchmark for Large Language Models in Chinese[J]. arXiv: 2310. 05818, 2023.

[34] WANG Y, ZHAI Z, LI H, et al. A Chinese Dataset for Evaluating the Safeguards in Large Language Models[C]// Findings of the Association for Computational Linguistics: ACL 2024. 2024: 3106-3119.

[35] WANG Y, ZHAI Z, LI H, et al. A Chinese Dataset for Evaluating the Safeguards in Large Language Models[J]. arXiv: 2402. 12193, 2024.

[36] TAN Y, ZHENG B, ZHENG B, et al. Chinese SafetyQA: A Safety Short-form Factuality Benchmark for Large Language Models[J]. arXiv: 2412. 15265, 2024.

[37] ZHANG H, GAO H, HU Q, et al. ChineseSafe: A Chinese Benchmark for Evaluating Safety in Large Language Models [J]. arXiv: 2410. 18491, 2025.

[38] ZHANG W, LEI X, LIU Z, et al. CHiSafetyBench: A Chinese Hierarchical Safety Benchmark for Large Language Models[J]. arXiv: 2406. 10311, 2024.

[39] XU G, LIU J, YAN M, et al. CValues: Measuring the Values of Chinese Large Language Models from Safety to Responsibility [J]. arXiv: 2307. 09705, 2023.

[40] ZHANG M, PAN X, YANG M. Jade: A Linguistics-Based Safety Evaluation Platform for LLM[J]. arXiv: 2311. 00286, 2023.

[41] CHEN Z, YU H, WU X, et al. Libra: Large Chines Based Safe-

guard for AI Content[J]. arXiv：2507.21929，2025.

[42] ZHANG Z，LEI L，WU L，et al. SafetyBench：Evaluating the Safety of Large Language Models[J]. arXiv：2309.07045，2023.

[43] ROTTGER P，KIRK H，VIDGEN B，et al. XSTest：A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models[C]// Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies. ACL，2024：5377-5400.

[44] SHI C，WANG X，GE Q，et al. Navigating the Over-Kill in Large Language Models[C]// Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. ACL，2024：4602-4614.

[45] CUI J，CHIANG W L，STOICA I，et al. OR-Bench：An Over-Refusal Benchmark for Large Language Mod-els[J]. arXiv：2405.20947，2024.

[46] ZHANG Z，XU W，WU F，et al. FalseReject：A Resource for Improving Contextual Safety and Mitigating Over-Refusals in LLMs via Structured Reasoning[J]. arXiv：2505.08054，2025.

[47] WANG Z，TU H，WANG Y，et al. STAR-1：Safer Alignment of Reasoning LLMs with 1K Data[J]. arXiv：2504.01903，2025.

[48] HUANG T，HU S，ILHAN F，et al. Safety Tax：Safety Alignment Makes Your Large Reasoning Models Less Reasonable[J]. arXiv：2503.00555，2025.

[49] QI X，PANDA A，LYU K，et al. Safety Alignment Should Be Made More Than Just a Few Tokens Deep[J]. arXiv：2406.05946，2024.

[50] CHEN J，WANG X，YAO Z，et al. Finding Safety Neurons in Large Language Models[J]. arXiv：2406.14144，2024.

[51] XUE Y，MIRZASOLEIMAN B. LoRA is All You Need for Safety Alignment of Reasoning LLMs[J]. arXiv：2507.17075，2025.

[52] JI J，WANG K，QIU T A，et al. Language Models Re-sist Alignment：Evidence from Data Compression[C]// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025：23411-23432.

[53] KANG D，PARK J，JO Y，et al. From values to opinions：predicting human behaviors and stances using value-injected large language models[C]// Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. ACL，2023：15539-15559.

[54] YAO J，YI X，XIE X. CLAVE：an adaptive framework for evaluating values of LLM generated responses[J]. Advances in Neural Information Processing Systems，2024，37：58868-58900.



LI Luozheng，born in 1990，Ph.D，assistant researcher. His main research interests include cognitive computing and social computing.

（责任编辑：何杨）