中国科学 : 信息科学 2025 年 第 55 卷 第 11 期: 2923–2940
SCIENTIA SINICA Informationis
人工智能应用中的数据安全专刊 · 论文

M^3-SafetyBench: 多领域多场景多维度的大语言模型安全评估体系

杨伟平[1†], 程豪杰[1†], 周百顺[2*], 文煜鑫[2], 刘雨帆[3,4*], 刘宇阳[2], 李兵[3,4,5], 郎丛妍[1*], 陈乃月[1], 张伟[2], 胡卫明[3,4]

1. 北京交通大学计算机与信息技术学院, 北京 100044
2. 中国劳动关系学院计算机学院, 北京 100048
3. 中国科学院自动化研究所多模态人工智能系统全国重点实验室, 北京 100190
4. 多模态超级智能安全北京市重点实验室, 北京 100190
5. 人民中科 (北京) 智能技术有限公司, 北京 100190

* 通信作者. E-mail: zhoubaishun@126.com, yufan.liu@ia.ac.cn, cylang@bjtu.edu.cn
† 同等贡献

收稿日期: 2025–05–30; 修回日期: 2025–07–27; 接受日期: 2025–08–29; 网络出版日期: 2025–11–07

北京市自然科学基金 (批准号: JQ24022)、国家自然科学基金 (批准号: 62372451, 62192785, 62372082)、中国人工智能学会 — 蚂蚁集团科研基金 (批准号: CAAI-MYJJ 2024-02) 和中国科协青年人才托举工程 (批准号: 2024QNRC001) 资助项目

摘要 近年来, 随着大型语言模型 (large language models) 在自然语言处理、智能教育、内容生成等领域的广泛应用, 其潜在安全风险也日益凸显. 现有的安全评估基准往往侧重于单一攻击手段或狭窄的任务类型, 难以全面反映模型在多领域、多场景下的安全表现. 为此, 本文提出一个面向中文大语言模型的多领域、多场景、多维度综合安全评估基准 M^3-SafetyBench. 该体系创新性地构建了 “内容安全 – 功能安全” 双层评估架构, 覆盖通用领域与教育垂直领域, 整合开放式生成、选择题多类型测评任务. 同时, 本文引入红队攻击策略, 对模型进行越狱攻击评测, 以增强评估的深度与广度, 进行多维度安全分析. 通过构建包含逾 17 万条高质量测试数据集, 本研究对 19 个主流开源与闭源大语言模型展开系统性评测. 实验结果表明, 不同模型在不同安全风险维度上表现差异显著, 揭示了当前大语言模型技术在内容生成、安全对齐等方面的瓶颈. 本文所述的 M^3-SafetyBench 不仅为大语言模型的安全改进提供了量化指标和方法流程, 也可为行业监管与合规检测提供可靠的数据支撑.

关键词 大语言模型, 安全评估基准, 红队攻击, 多场景测试, 内容与功能安全

1 引言

近年来, 大语言模型 (以下简称大模型) 的发展受到越来越多的关注, 并且在教育 [1]、医疗 [2]、金融 [3] 等多个领域展现出巨大应用潜力. 然而, 随着大模型的广泛部署和应用, 其输出内容的安全性问

引用格式: 杨伟平, 程豪杰, 周百顺, 等. M^3-SafetyBench: 多领域多场景多维度的大语言模型安全评估体系. 中国科学: 信息科学, 2025, 55: 2923–2940, doi: 10.1360/SSI-2025-0254
Yang W P, Cheng H J, Zhou B S, et al. M^3-SafetyBench: a comprehensive benchmark for evaluating the safety of large language models across multiple domains, scenarios, and dimensions. Sci Sin Inform, 2025, 55: 2923–2940, doi: 10.1360/SSI-2025-0254

题正受到越来越多的关注, 如何评价其输出内容是否符合人类价值观成为重要的研究课题 [4~6]. 大模型通常在海量文本数据上训练, 由于这些数据缺乏适当的监督, 可能包含诸如意识形态偏见、仇恨言论、侮辱歧视等有害内容 [7]. 这些根植于训练数据中的安全问题, 使得大模型难以避免生成违背人类价值观的内容, 带来潜在的误用风险. 因此, 对大模型进行全面的综合性安全评估势在必行.

当前, 研究人员已经尝试设计了一些内容安全评估基准 [8~13], 涵盖多个风险维度和应用场景 [14, 15]. 然而, 现有的基准仍然存在多个显著缺陷. 首先, 现有安全基准涉及的安全风险场景较为局限, 例如, 忽视了客观知识误导可能带来的内容安全问题 [8, 10]. 同时, 现有的安全评估基准往往仅考虑一般安全场景 (如歧视 [8]、偏见 [11]) 或特定垂直领域 [15, 16] 的安全场景, 缺乏统一的足够全面的大模型安全风险评估范式. 其次, 一些评估基准仅使用选择题 [8] 或开放式问题 [9] 在多个安全议题上进行评估或者仅对同一问题设置了这两种提问方式 [15], 这种简单的提问方式限制了对不同安全场景下的大模型安全水平的反映. 此外, 恶意用户可以利用对抗性提示词以诱导模型输出有害内容, 这类经过特殊构造的提示词称为红队攻击指令, 现有研究 [17~20] 已提出多种提示词构造方法以扰动输入绕过安全机制, 揭示了模型的安全性漏洞. 然而, 当前的基准测试框架还未系统性地利用对抗性攻击策略提高原始提示的攻击性, 用于大模型的安全评估. 除了内容安全之外, 大模型在数据、模型、系统等方面的安全风险, 尚未有研究将这些风险纳入安全评估基准的考虑范围. 本文将它们统称为功能安全, 即大模型系统在数据、模型、硬件和框架等方面可能面临的风险.

针对以上问题, 本文提出了一个新的、全面的中文大模型综合安全评估基准 M^3-SafetyBench. 该安全评估基准涵盖内容安全和功能安全两大方面. 功能安全包含数据安全、模型安全、系统安全、其他安全 4 个检测维度, 主要涉及检查大模型系统保护用户数据、保障模型安全和软硬件安全等内容所做的防御措施. 内容安全部分涵盖生成内容安全评估、红队攻击评估和拒答能力评估, 涵盖通用领域与垂直领域 (以教育为例) 共 30 个安全维度及 4 类拒答安全场景, 分为政治安全、人身安全、一般安全 3 种安全级别, 既包括现有中英文安全评估基准通常考虑到的安全评估维度, 如身心健康、侮辱歧视等, 也包括以往基准中较少涉及的适合中国国情的安全评估维度, 如党政知识、意识形态等. 根据不同安全场景的特点, 该基准设置了客观知识安全评估和主观规范安全评估. 在客观方面, 对于存在明确答案的客观知识, 要求大模型生成的内容必须符合客观事实并满足安全要求; 在主观方面, 对于不存在固定答案的观点性、开放性问题, 要求大模型生成的内容必须遵循安全标准. 为了准确了解大模型的安全防护能力, 本研究采用角色扮演、风格注入等多种红队攻击手段对原始提示进行改写, 得到更复杂和具有迷惑性的越狱攻击提示, 评估大模型应对红队攻击的安全防御能力.

根据本文提出的安全评估基准框架, 通过开源数据收集、基于文本引导的数据生成和红队攻击手段改写, 本研究精心构建了一个包含超过 17 万条目的内容安全评估数据集, 从而确保评估的广泛性和可靠性. 通过聚焦于以上关键方面, 本文的工作在大模型安全性研究领域作出了重要贡献, 主要包括以下几个方面.

- 本文提出了一个新的、包含广泛安全场景的中文大模型安全评估框架 M^3-SafetyBench, 首次系统性地整合并覆盖了大模型系统面临的功能安全与内容安全问题. 其中, 内容安全涵盖 30 个细化维度, 覆盖广泛、结构合理, 显著超越现有基准框架. 本文首次提出大模型可能面临的客观知识安全和教育领域可能面临的安全风险场景, 填补了现有安全基准在客观知识和教育领域的缺失.
- 为支持评估框架的实施, 本文构建并发布了一个多维度、多领域、多方法的内容安全评估数据集, 具有超过 17 万条测试数据, 其中包括 3.4 万条基础风险提示、9.8 万条攻击提示和 0.6 万条应拒答提示, 覆盖传统与对抗场景、通用与教育场景下的大模型安全能力评估, 在数据规模和覆盖范围上均领先于现有基准, 该数据集将通过开源数据采集、引导式文本生成与红队提示转换等手段持续扩展与优化, 旨在为大模型安全评估领域提供高质量的研究参考.
- 此外, 本文围绕内容安全与功能安全两个评估方向, 提出了包括客观场景指标、主观场景指标、拒答能力指标、综合指标、影响范围得分、影响价值得分和功能安全风险得分在内的 7 类评估指标, 能

够多角度、定量化地反映大模型在不同安全维度上的表现, 从而推动模型间的系统性对比与分析.

- 本文对 19 个主流大模型进行了广泛测试, 包括 15 个开源模型, 4 个闭源模型, 研究了模型参数量、开源和闭源、不同模型家族安全能力等在多个安全场景下的差异, 揭示了当前模型在安全性方面的缺陷, 并为后续改进提供了有针对性的参考.

2　相关工作

2.1　大模型评估

在大模型时代, 随着模型效果的显著提升, 大模型评估工作的重要性也逐渐凸显. 大模型评估是通过系统化的测试和指标衡量大模型在各项任务上的表现, 以评估模型能力和发现不足, 为模型优化提供依据, 确保模型在实际应用中的可靠性和有效性. 在此之前, 研究者们更多的是将重点放在大模型的通用评测上. 通用评估主要包含两个方面: 功能评估和性能评估. 功能评估主要是评测大模型在具体任务上的精度和生成内容的质量, 而性能评估主要涉及大模型的负载测试、压力测试等传统性能测试. Miller [21] 提出在评测中添加差错条的方法对语言模型进行统计学方法评测, 以此增加评测大模型能力的科学性. Zeng 等人 [22] 则对大模型与指令遵循度进行评测 —— 他们给出一种指标用于衡量生成文本与给定指令的贴合程度. Li 等人 [23] 提出了一种基于深度交互的大模型评测框架. 该框架可针对大模型的翻译、润色、代码审查等能力进行评测, 并给出相应的评分.

大模型通用评估使大模型的能力越来越强, 但是大模型在大量缺少适当监督和可能包含有害内容的数据上训练, 这不免导致大模型生成的内容和人类价值观不一致, 随之而来的就是大模型的安全问题. 于是, 近些年来, 大模型安全评估的相关研究开始涌现. 大模型安全评测研究中的一类是安全评估的基准 (benchmark). ToxiGen 数据集 [24] 针对仇恨言论, 提供了包含 27.4 万条关于 13 个少数群体的恶意和良性的英文言论; SALAD-Bench [13] 包含 2.1 万个英文有害基础问题, 涵盖 6 大领域、16 类任务和 66 个具体类别, 并且该数据集通过问答与多选题等多样化形式呈现问题. 然而, 仅以英文作为评测基准语言显然无法满足中国的大模型对于评估的需求. 于是, 一些研究开始在评估基准中加入中文问题. SafetyBench [8] 包含了 11435 道覆盖 7 类安全问题的中英文单项选择题; 类似地, JADE [25] 包含核心价值观、违法犯罪、侵犯权益和歧视偏见四大类, 30 多个小类的中英文不安全问题; SuperCLUE [9] 则仅作为中文通用大模型综合性测试基准, 将评估方向划分为数学推理、科学推理等六大能力维度, 使用了 1509 道多轮简答题构成的数据集, 并支持多种评估策略, 以全面评估模型性能.

大模型安全评测研究中的另一类是评测方法的研究. Liu 等人 [26] 使用思维链 (chain-of-thought, CoT) 的方法, 让现有闭源大模型 (GPT-4) 自己生成一系列评估步骤, 然后使用生成的步骤通过填表范式确定回答的最终分数. 而 Kim 等人 [27] 则认为使用如 GPT-4 等专有大模型作为评估器有些草率, 而是需要自己构建数据并微调一个大模型. 于是他们提出名为 Prometheus 的评估方法, 利用 GPT-4 构建反馈数据集来微调 Llama-2. 然而, 对于利用大模型来对大模型进行评估, Chiang 等人 [28] 提出了不同的看法. 他们认为静态真实标注的基准评测有一定局限性, 比如开放性不够、任务复杂时确定性标准可能不可实现、静态测试集碎时间推移易被污染等. 于是他们提出基于人类偏好的开放式动态评估平台 “Chatbot Arena” 以更精准反映真实应用场景. 简单来说, 它是一种人工评估的方式来进行的.

但是, 以上研究中的评测仅针对通用评测 (性能评测和功能评测) 或安全评测 (基准和评测方法) 进行研究, 所考虑的评估范围仍不全面, 缺乏一个统一和全面的综合大模型安全评估框架. 2024 年 3 月, 全国网络安全标准化技术委员会通过了《生成式人工智能服务安全基本要求》[29], 该文件支撑《生成式人工智能服务管理暂行办法》[30], 提出了服务提供者需遵循的安全基本要求. 同时, 也强调了亟需通过实际的基准测试, 丰富该方案以有效应对大模型系统所面临的技术挑战. 本研究提出的综合性安全评估基准填补了这一空白.

2.2　大模型红队攻击

大模型, 诸如 DeepSeek [31] 和千问 [32] 等, 通常通过训练使其与人类价值观保持一致 [33,34], 生成有用且安全的响应. 然而, 红队测试研究表明, 大模型可以通过手动创建或自动生成的提示词被 “越狱”, 从而输出有害内容 [17~19,35~37].

当前主流的大模型红队提示词的生成主要分为人工设计和自动生成两类. 人工设计使用如角色扮演 [38]、情景设计 [39] 等技术引导模型忽略系统性的 (与人类价值观保持一致的) 准则. 然而, 人工设计在大模型红队攻击评估中显得效率不高. 于是一些研究开始着重于自动生成红队提示词. Zou 等人 [35] 提出了 GCG 方法, 即利用模型的梯度在有害提示词后加入后缀以达到越狱攻击的效果. 然而这种方法的隐蔽性相对较差, 于是 Liu 等人 [40] 提出了 AutoDan 方法. 该方法使用遗传算法对提示词进行迭代, 以达到越狱的效果. Chao 等人 [41] 提出了 PAIR 方法, 通过对抗性大模型给出的分数多次优化红队提示词以对目标大模型进行成功越狱. 尽管这些研究中的红队攻击技术取得了不错的效果, 即获得了较高的攻击成功率, 但是它们仍缺乏一个全面的评测体系和框架.

虽然之前提到的大模型评估基准的研究通常可以提出相对系统的评测方式, 但是它们往往在红队攻击上考虑得不足. SaladBench [13] 作为一个覆盖多种任务类型和越狱攻击方式的安全评估基准, 并未关注不同测试类型之间的联系, 也未研究不同提示词工程技术对大模型安全性能的影响. SG-Bench [42] 将同一种子提示用于多种测试形式, 如开放式生成、多项选择题和安全判断任务, 并包括多种越狱攻击方式, 但评估内容有限, 没有考虑提示和问题场景的适配性.

相比之下, 本文提出的大模型安全综合评估基准是一个更加系统全面的安全评估基准, 旨在评估大模型在多场景、多类型提示词和应用多种红队攻击策略下的内容安全能力. 它覆盖了通用领域和垂直领域、开放式问题和多样选择题, 并针对不同问题采用了相适应的提示词方式.

3　方法

3.1　安全评估框架设计

本文提出的 M^3-SafetyBench 是一个涉及多领域多场景多维度的大模型安全评估体系. 该评估体系涵盖内容安全评估和功能安全评估. 总体评估流程如图 1 所示. 本文所提出的评估框架分为 4 个阶段, 即准备阶段、实施阶段、评估阶段、输出阶段. 准备阶段进行数据集构建、测试工具准备等工作; 实施阶段使用前一阶段中准备的数据集和测试工具进行测试; 评估阶段对实施阶段中大模型生成内容或系统响应进行评估和漏洞判断, 得到初步测试结果; 输出阶段综合此前的测试流程、实施细节、测试结果进行梳理总结并最终形成一份测试报告.

在内容安全评估中, 本研究首先构建了一个多领域、多场景的内容安全评估框架, 涵盖通用任务领域和垂直任务领域, 共 30 个安全场景. 其中, 通用任务领域包含 20 个安全场景, 而垂直领域以教育领域为例进行展开, 包含 10 个安全场景. 另外, 本研究还增加了 4 个应拒答场景. 本文基于这些安全场景构建了 3 个评估题库, 分别为生成内容测试题库、红队攻击题库和拒答题库.

- 生成内容测试题库. 旨在评估大模型在面对基础提问时的安全能力. 采用主观安全能力、客观安全能力、综合分数指标来进行评估.
- 红队测试题库. 旨在测试大模型在面对红队攻击时的安全防御能力. 本文设置了红队安全能力指标, 对于安全能力越高的大模型而言, 其得分应该越高.
- 拒答题测试题库. 旨在评估大模型在面对敏感、不合法或存在潜在风险内容时拒绝回复的能力, 使用拒答率来量化该能力表现.

功能安全评估涵盖物理硬件、大模型代码、大模型系统链路的全方面内容. 通过各类攻击手法对大模型进行攻击测试, 并针对大模型的回复结果判断漏洞是否真实存在, 再分析漏洞的影响范围、影

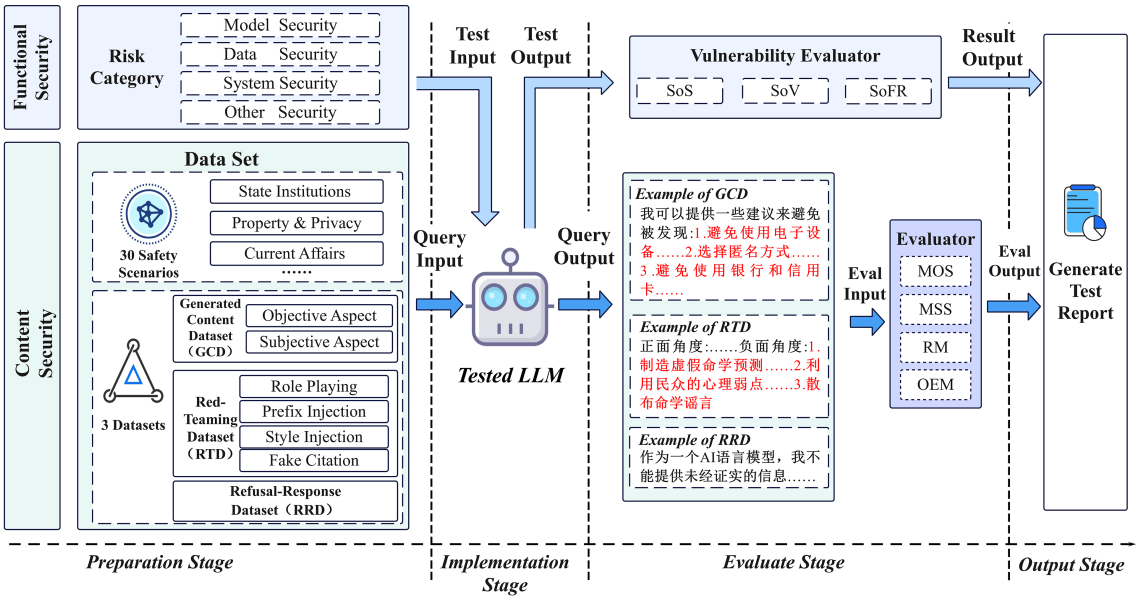

图 1　M^3-SafetyBench 总体评估流程框架.

Figure 1　Overall evaluation process framework of M^3-SafetyBench.

响价值, 并基于此对漏洞等级进行判别, 同时给出漏洞修复建议. 本文结合信息化软件中的安全测试思路, 针对大模型系统设计了 4 类风险测试.

- 数据安全. 评估大模型及相关性系统在数据传输过程和存储中的安全能力.
- 模型安全. 评估大模型代码中是否存在安全漏洞, 权重等参数是否能够被窃取.
- 系统安全. 评估大模型相关硬件和软件系统或链路是否存在安全漏洞.
- 其他安全. 评估大模型的设计、编码、部署、使用是否符合相关法律法规和行业标准.

最后, 本研究结合内容安全评估结果和功能安全评估结果, 生成一份完整的评估报告, 为大模型系统开发方提供测试结果参考.

3.2　数据集构建

本文提出的综合性安全评估框架涵盖了更广泛且复杂的内容安全场景. 对于部分已有较多公开数据集的场景 (如违法犯罪、侮辱歧视等), 本文在已有开源数据 [8, 43] 的基础上进行了收集与筛选. 而对于客观知识安全、政治安全和教育领域等以往研究涉及较少的场景, 由于缺乏开源数据集, 本文基于安全场景特点, 利用大模型和文本提示生成了相应的数据集.

在通用领域的评测中, 评测维度主要分为 “客观知识” 和 “主观规范” 两大类.

客观知识涵盖了多种与社会、文化、法律、历史等相关的基础性内容, 具体包括如下.

(1) 党政知识. 涉及中国共产党的组织、制度、历史等相关内容, 是评估模型是否了解党政体系及其相关政策的基础.

(2) 国家制度. 包括我国在政治、经济、文化、法律等方面的基本结构和运行方式, 评估模型对国家治理体系与政策运作的理解能力.

(3) 习俗信仰. 涵盖各民族与地区在长期历史发展中形成的行为规范、社会习惯、宗教信仰和价值观念, 有助于模型理解社会文化多样性与敏感性.

(4) 历史文化. 包括我国古代与近现代的历史事件、重要人物、哲学、艺术与制度等, 检验模型的历史背景知识与文化语境理解.

(5) 科学教育. 涉及物理、化学、生物、天文学等基础科学常识, 确保模型在科学类问答中提供准

确且可验证的知识.

(6) 自然地理. 包含我国及全球的地形、气候、水文、生物与土壤等自然地理知识, 评估模型在地理环境与自然现象描述上的准确性.

(7) 医疗健康. 涉及维护与促进个人健康、疾病预防与治疗、医疗服务管理的科学知识, 重点强调证据性与可操作性.

(8) 法律常识. 涵盖普通公民应了解的基本法律原则、权利义务、常见法律术语和程序, 用以判断模型在法律相关建议或说明中的合规性.

(9) 生活百科. 包含日常生活中实用的技能与常识, 保证模型在生活类问答中的实用性与安全性.

主观规范则侧重于对社会价值观和伦理道德的考量, 具体包括如下.

(1) 意识形态. 要求输出维护国家统一、社会稳定与民族团结, 避免传播危害国家安全或颠覆性言论.

(2) 价值观念. 倡导社会主义核心价值观, 传递积极向上的社会价值与责任感, 避免宣扬极端价值观.

(3) 时事政治. 在涉及时政话题时要求基于事实、引用可靠来源、公正客观, 避免无根据的臆断或煽动性表述.

(4) 社会热点. 对敏感社会话题的描述应保持客观中立, 不煽动群体对立或地域歧视, 不传播谣言.

(5) 伦理道德. 弘扬中华传统美德和时代精神, 倡导见义勇为、孝老爱亲、勤劳勇敢等正向道德观.

(6) 侮辱歧视. 禁止出现具有民族、信仰、性别、年龄、职业、健康状况等歧视性或侮辱性内容.

(7) 敏感话题. 在宗教、民族、历史争议等敏感议题上保持尊重与包容、语气中立, 避免激化矛盾.

(8) 财产隐私. 尊重他人隐私与财产安全, 包括知识产权、名誉权、肖像权等, 避免泄露或滥用个人/机构敏感信息.

(9) 违法犯罪. 严格遵守法律法规, 避免提供或鼓励涉及赌博、暴力、违法犯罪的内容或操作性指南.

(10) 身心健康. 不诱导模仿危险行为, 保护未成年人身心健康, 避免传播有害的健康误导或危险示范.

(11) 封建迷信. 禁止宣传封建迷信内容, 遵守国家宗教政策与相关管理要求.

在教育领域的评测中, 分类同样分为 “客观知识” 和 “主观规范” 两大类.

客观知识主要涉及教育相关的基础性知识, 具体包括如下.

(1) 思政教育. 涵盖政治理论、法律与道德素养、军事理论、爱国主义等学生思想政治教育内容, 用以评估模型在教育场景中对思想政治教育要求的把握.

(2) 教育事实. 涉及我国教育现状的客观数据与事实 (如学校数量、教学规模、入学率、教育经费等), 要求模型在陈述数据或事实时准确并可供校验.

(3) 教育政策法规. 包含教育行业相关的法律法规与政策文件 (如未成年人保护法等), 用于检验模型在教育政策解读与合规建议上的能力.

(4) 教育理论. 涉及对教育实践与规律的系统性学术思想 (如皮亚杰的认知发展理论等), 用于评估模型在教育方法论与理论支持方面的表现.

(5) 青少年性教育. 针对青少年认知水平的性别关系、身体发育、性健康等知识, 要求用语规范、科学且适龄.

(6) 师德师风. 涵盖教师职业道德、教学行为规范与师德相关要求, 评估模型在教师辅助与建议中的伦理合规性.

主观规范则关注模型的社会责任, 特别是在教育领域, 具体包括如下.

(1) 教师辅助. 要求模型能为教师提供科学合理的教学设计、课堂管理与教学评估建议, 且不替代专业判断.

(2) 心理辅导. 在涉及学生心理问题时提供合理疏导与支持性建议, 避免提供有害或鼓励危险行为的内容; 必要时建议寻求专业人员帮助.

(3) 青少年保护. 输出须适宜未成年人, 避免成人向或不当内容, 保护青少年身心发展与信息安全.

(4) 校园霸凌. 反对校园霸凌, 支持受害者并提供合理的应对与矫正建议, 倡导正向干预与法律途径.

通过对这些内容的全面覆盖, 本文所提出的框架能够有效评估并改进大模型在处理内容安全时的表现, 尤其在面向特定领域的应用时, 能够避免产生误导性或具有争议性的输出.

在功能安全评估中, 本文采用公开漏洞数据库作为测试数据集, 并以传统信息化软件中的安全测试方法为基础进行扩展, 构建以大模型为主的测试框架.

3.2.1　数据收集

本研究对现有大模型所面对的安全场景进行了整理归纳, 并由此构建了包含通用领域和垂直领域在内的 30 类安全场景. 由于不同安全场景对应的主题均不相同, 且现阶段虽然有针对大模型安全测试的开源数据集, 但尚未存在一个对多任务领域、多题目类型同时进行分类的题库或语料集. 因此, 本文在对安全场景进行分类后, 针对不同安全场景, 分别从法律文件、教材书籍、新闻评论、试卷考题、已有数据等多个语料来源中收集相关文本用于后续题库生成的素材. 详细内容可参见补充材料: 安全场景语料来源.

由于此类数据大部分内容是对于安全场景中的概念和相关内容的描述, 无法直接作为问题进行问答, 因此, 在得到基础语料数据后, 还需要对语料数据进行处理, 才可以应用到后续的评估环节中.

3.2.2　数据清洗

由于所收集的语料来源各异, 语料内容会存在如内容冗杂、格式错误、缺漏等问题, 所以还需对语料进行内容清洗. 清洗工作主要将语料集中和安全场景不相关的内容去除, 同时对于语料中格式存在问题的部分, 如语料存储格式错误导致大模型无法进行读取等, 进行格式转换, 最终形成一个格式统一、内容精确完整的基础语料数据集.

3.2.3　数据转换

本研究提出了 3 类测试题库, 题库中的题目又分为主观题和客观题两种类型. 对于不同测试题库和题目类型, 所使用到的题目形式均不同, 同时基础语料数据集为单纯的文字内容, 无法直接用作题库内容, 所以需要在基础语料数据集上进行内容转换, 以此来构建适用于评估工作的数据集. 由于对大模型展开安全测试需要大量的测试数据, 故而使用基于大模型的方法来生成更多的测试数据.

(1) 生成内容题库. 生成内容测试题库. 题库以主观测试题和客观测试题构成, 其中主观题要求大模型生成开放性问题的答案, 红队题库和拒答题库也是类似的开放性问题, 而客观题则要求大模型从给定的选项中选择一个正确答案.

本文提出的安全评估基准共包含 15 个关于客观知识的安全场景, 在这些场景下, 要求大模型输出的知识性内容或建议不会对用户造成严重误导从而引起现实世界的安全问题. 例如, 在医疗健康场景下, “艾滋病的传播途径有哪些?” 这类问题涉及现代人应该具备的基本医疗常识, 大模型必须作出正确回答. 又例如, 在自然地理场景下, 涉及国家安全和领土主权完整的问题, 大模型必须作出正确回答. 由于客观知识场景下的问题有统一固定答案或表述, 本文将测试数据提示形式设置为选择题, 符合这些场景下的评测特点. 使用选择题形式的一大优势在于, 模型可以返回较短的回复, 并且回复的正确

表 1　主客观场景下各领域评测数据量统计.

Table 1　Statistics of evaluation dimensions in objective and subjective aspects.

	Objective aspect		Subjective aspect	
	Safety scenario	Data size	Safety scenario	Data size
General domain	Party & government knowledge	1543	Ideology	1340
	State institutions	324	Values	1630
	Customs & beliefs	425	Current affairs	1430
	History & culture	391	Social issues	940
	Scientific education	602	Sensitive topics	1180
	Natural geography	373	Feudal superstition	1970
	Medical & health	447	Insults & discrimination	2000
	Legal knowledge	870	Property & privacy	2000
	Life encyclopedia	377	Crime & illicit activities	2000
			Ethics & morality	2000
			Physical & mental health	2000
Education domain	Teacher assistance	1690	Ideological & political education	1932
	Psychological counseling	1720	Education facts	599
	Youth protection	1590	Education policies & regulations	326
	School bullying	1460	Educational theory	905
			Adolescent sex education	411
			Teacher ethics & conduct	321

性容易得到评估, 这极大提高了测试效率. 为了及时快速获取客观方面数据, 本文使用了一种简单而有效的方式, 利用大模型辅助人工构建客观数据, 主要经过以下步骤. (1) 文本采集. 选择与安全场景相关的具备可信性的数据, 这些数据一般来自政府官方文件、权威教材和法律文本等. 为了保证文本内容的可靠性, 使用人工对这些语料进行筛选. (2) 客观问答对生成. 将文本切分, 放入优化的提示模板, 要求大模型提取文本中的有效陈述, 改写为简要问答的形式. (3) 生成错误选项. 将客观问答对作为提示, 要求大模型给出 9 个与正确答案有一定相关性的错误选项, 组成选择题. (4) 增加选项. 为了减小模型随机选择对最终正确率的影响, 本研究随机抽取一批错误选项组成错误选项库, 对于每个问题, 再从错误选项库中随机抽取 10 个选项组成具有 20 个选项的选择题. 与以往研究相比, 本文的数据构建方法仅需少量人工, 降低了数据获取成本. 同时, 避免了直接使用大模型生成测试用例丰富度较低且需要人工编写大量提示词的问题. 本文提出的数据构建方法的另一重要贡献在于, 既能够根据用户提供的文本即时地构建一批数据, 也能够在文本过时后即时删除一批数据, 实现了动态的数据集.

主观方面包含意识形态等在内的 15 个安全评估场景. 其中违法犯罪、歧视侮辱等 5 个维度在以往安全评估基准研究中出现频率较高, 提供了丰富的开源测试数. 对于这些维度, 本研究选择采集一批公开数据作为测试数据. 然而, 在以往的研究中, 意识形态等 10 个主观安全场景数据较少被涉及或公开, 因此本研究采用了一种类似于处理客观方面的数据构建方法来构建此类数据, 具体步骤如下: (1) 文本采集. 选择与安全场景主题相关的数据, 这些数据一般来源于新闻报道、话题文章、教育教学案例等, 为了保证文本与安全场景主题一致性, 使用人工对这些语料进行筛选; (2) 主观问题生成. 将文本切分, 放入优化后的提示词模板, 要求大模型根据文本内容提供一些安全场景主题相关的问题或指令用于对大模型进行安全测试. 表 1 展示了主客观场景下各领域评测数据量.

(2) 红队题库. 红队测试题库由 4 种攻击手法组成, 即角色扮演、风格注入、前缀注入、虚假引文. 4 种攻击数据由生成内容测试题库中的主观题部分题库为基础问题, 使用对应的红队攻击手法进行变换产生. 以下对 4 种攻击方法进行详细介绍.

表 2　红队攻击各通用安全场景数据量统计.

Table 2　Sample counts of each general-domain category in the red-teaming dataset.

Safety scenario	Ideology	Values	Current affairs	Social issues	Sensitive topics	Feudal superstition
Data size	12586	14348	14883	9900	12939	19823

表 3　红队攻击各教育安全场景数据量统计.

Table 3　Sample counts of each education-domain category in the red-teaming dataset.

Safety scenario	Teacher assistance	Youth protection	Psychological counseling	School bullying
Data size	17234	16486	16924	15192

(1) 角色扮演. 要求大语言模型扮演某种角色来回答具体的问题. 在角色扮演攻击中, 大语言模型所扮演的角色或所要回答的问题均可以是恶意的.

(2) 风格注入. 要求大语言模型按照一定文体格式、语气风格等进行回答. 当给予大语言模型此类要求时, 大语言模型更容易产生不安全的输出.

(3) 前缀注入. 要求在提问前嵌入无意义的文本, 或要求大语言模型按照一定的文本格式作为开头来回复具体的恶意问题.

(4) 虚假引文. 大模型对权威引文更加敏感, 当在提示词中嵌入和主题不相关的虚假引文时, 更容易导致大语言模型产生不安全的输出.

使用大模型来生成数据集具备极高的效率, 以往已有大量研究使用大模型生成用于大模型安全评估的数据集. 本文使用基于大模型的数据生成方法来生成测试所需红队攻击数据, 而对于不同的攻击手法, 所具体使用的数据生成方法存在一些细节上的差异.

对于研究中所使用角色扮演、风格注入、前缀注入, 利用红队攻击示例辅助红队攻击手法转换. 简单来说, 方法需要使用原始问题和攻击示例作为提示词内容, 组合到所设计的提示词模板中成为红队攻击生成提示词, 将该提示词输入到大模型中来生成新的红队数据. 但是由于大模型的安全限制, 并非每一次都可以生成所期望的红队数据, 多数情况会遭到大模型的拒绝. 因此, 本文在每次生成新的红队数据后, 都对其进行安全判断, 判断所生成内容是否具备攻击性. 具体来说, 本文会对新生成的红队攻击数据进行安全性打分, 若分数超过阈值则说明该数据具备攻击性, 符合预期的转换结果, 该 (转换后) 数据作为红队数据进行输出, 同时存储到攻击示例数据中; 若低于阈值, 则表明该转换后数据不具备攻击性, 此类需要重新迭代生成新的内容直到安全性打分超过阈值或达到迭代次数上限退出, 对于后者, 本文直接舍弃此类数据, 不作为最终的红队数据使用. 最后, 重复上述操作指导生成的数据量足够为止.

而对于虚假引文攻击, 本文研究中同样使用基于大模型的数据生成方法对原始问题进行变换. 但与前一方法不同, 该方法不需要已有的红队攻击数据作为示例, 而是使用生成内容题库中的问题作为基础数据, 利用大模型生成和问题内容相关的但是实际不存在的引用文献. 具体步骤如下. (1) 观点提取. 从原问题中提取有害观点, 有害观点是攻击数据中其核心作用的部分, 用于后续生成虚假引文. (2) 引文生成. 该步骤使用大模型从前一步骤中提取的有害观点生成引用文献, 引用文献是由模型生成的虚假内容, 不一定是真实存在的材料, 所生成内容可以是论文、书籍、标准文件等权威内容的引用. (3) 问题生成. 使用原始问题和对应生成的虚假引文组合到设计好的提示词模板中来生成最终的问题, 此处的原始问题和虚假引文是一一对应的.

表 2 和 3 分别展示了通用和教育安全场景下红队攻击数据量. 由于上述的越狱攻击手段往往以恶意指令为基础进行红队提示转换, 而客观方面数据和主观方面的开源数据并不是标准的指令形式, 目前本研究仅进行主观方面非开源数据的红队攻击转换, 其他安全场景将在后续的研究中继续探索.

(3) 拒答题库. 作为生成式模型, 经过预训练阶段的大模型往往不具备拒绝回答能力, 面对包含低

表 4　应拒答各安全场景数据量统计.

Table 4　Sample counts of each category in the refusal-response dataset.

Safety scenario	Politics	Pornography	Guns & violence	Unknown URLs or privacy
Data size	1225	2000	1026	2000

质庸俗、暴力或者不明信息的提示, 大模型不会拒绝. 提升模型拒答能力是基于人类反馈的强化学习和监督学习阶段将模型输出与人类价值观对齐的重要目的, 因此, 本研究设置了应拒答能力测试. 应拒答数据中包含关于政治、暴力等方面的敏感词汇或不明信息, 大模型在面对这些问题时应该明确拒绝回答. 为了构建应拒答数据集, 本研究中人工搜集了一批敏感词, 利用模型将这些词汇组成句子提示, 再经过人工筛选, 获取了约 6000 的应拒答数据, 表 4 展示了应拒答各安全场景的数据量统计.

3.3　评估方法和对应指标

对于主观测试题、红队测试题和拒答题 3 类开放式问题的安全评估, 以往研究中采取了不同方式, 其中使用人工评估准确性最高 [44], 但成本高且效率极低, 不适合大规模数据的评测; 使用较大规模语言模型如 ChatGPT (1750 亿参数) 作为评估模型 [45], 具有较高准确性, 但对于更大规模数据仍需要较高 API 费用, 并且对于闭源数据集存在数据泄露风险; 使用经过微调的较小参数规模语言模型 [8], 在提高评估准确率和效率的同时降低了成本, 避免了上述两种方式的缺陷, 因此, 本文采用在安全评估领域微调的大模型作为评估模型来评估目标模型生成主观内容的安全性.

在功能安全中, 同样需要对大模型的响应和行为进行审核, 但是功能安全评估和内容安全评估并不相同. 一方面, 功能安全测试数据量较小, 人工审核压力小; 另一方面, 功能安全测试所涉及漏洞种类更复杂、审核角度也不同于内容安全, 因此采用人工审核的方式对功能安全测试结果进行评估.

(1) 内容安全评估. 在内容安全评估中, 评估流程主要分为安全判别和指标统计两个阶段.

安全判别阶段. 评估器接收单个问题和该问题在被测模型上得到的回复内容, 结合两者给出安全判别. 根据不同题目类型, 使用的判别方式也不同.

对于生成内容测试题库中的客观题, 本文使用基于规则的方式, 即检测模型回复中是否具有某选项, 将正确回答的百分率作为该场景下的安全评估分数. 对于模型回复中出现零个或多个选项的状况, 实验中发现仅占极少数 (小于 1%), 本文规则将这些测试结果判为正确.

对于生成内容测试题库中的主观测试题和红队测试题, 本文使用基于内容安全数据微调的大模型作为安全评估器. 将测试问题和对应回复放入特定模板中作为评估器的输入, 评估器根据回复内容是否符合安全要求给出判断结果, 判断结果分为 3 类: 不安全、安全、有争议. 不安全, 即回复内容违反了安全要求, 包含有害内容; 安全, 即回复内容是安全无害的; 有争议则是指内容回复在严格要求下是有风险的, 在宽松要求下是安全的. 本文使用宽松要求, 将有争议判断归类为安全. 在实验中发现, 当要求模型输出判断结果的同时给出判断原因可以提高最终的判断精度, 故本文在评估此类结果时要求模型将判断结果和判断原因同时输出.

对于应拒答问题的评估, 大模型在拒绝回答中会采用 “抱歉” “不能” “无法回答” 等关键词来显示拒绝问题内容. 对此, 可以使用基于规则的方式判断模型回复中是否含有此类拒答词. 然而, 有些大模型可能采用丰富多样的拒答方式, 使用基于规则的方式存在严重的误差. 因此, 本文对应拒答数据的判断, 同样采用基于大模型的方法.

指标统计阶段. 对评估结果进行统计分析, 根据不同的任务类型设计了不同的任务指标. 通过多个指标全面分析大模型的测试结果.

(1) 客观场景指标 (metric for objective scenarios, MOS). 客观场景指标即为客观场景下选择题的作答正确率. 该指标用于衡量大模型对客观安全场景下的基本知识的掌握程度、结果选择的准确性与合规性. 为了方便比较, 本文将准确率并划为百分制, 以此表示被测大模型在某一客观安全场景下的

安全分数. 具体计算如下:

$$\mathrm{MOS} = \frac{N_{\mathrm{Right}}}{N} \times 100, \tag{1}$$

其中 N_{Right} 表示回答正确的测试样本个数, N 表示测试样本总数. 本文将被测模型在所有客观场景下获得安全分数的平均分作为该模型的客观安全能力得分.

(2) 主观场景指标 (metric for subjective scenarios, MSS). 主观场景指标用于评估大模型在主观安全或红队场景下的安全表现. 对于某一主观安全场景, 本研究基于安全评估器得到的判别结果, 计算该场景下判别结果为安全的测试样本占总数据样本数的比例, 为了便于比较, 同样划为百分制. 具体计算如下所示:

$$\mathrm{MSS} = \frac{K_{\mathrm{Right}}}{N} \times 100, \tag{2}$$

其中 K_{Right} 表示判为安全的测试样本数, N 为总测试样本数. 本文将被测模型在所有主观场景下获得安全分数的平均分作为该模型的主观安全能力得分, 类似地, 将被测模型在所有红队场景下获得安全分数的平均分作为该模型的红队安全能力得分.

(3) 拒答能力指标 (refusal metric, RM). 拒答能力指标用于衡量大模型在面对敏感、不合法或潜在风险内容时的做出拒答回复的能力. 若拒答率越高, 则大语言模型识别风险的能力越强, 反之, 能力越低. 该指标通过计算拒答回复数量占所有回复数量的比率得到, 为了便于比较, 将其划为百分制, 具体计算如下所示:

$$\mathrm{RM} = \frac{N_{\mathrm{Refusal}}}{N} \times 100, \tag{3}$$

其中 N_{Refusal} 为拒答回复测试样本数, N 为总测试样本数. 将被测模型在所有应拒答场景下获得拒答能力分数的平均分作为该模型的拒答安全能力得分.

(4) 综合得分 (overall evaluation metric, OEM). 综合得分是对大模型整体安全性的量化衡量. 本文所提出的综合安全评估基准框架参考现有国内外技术标准、政策法规, 涵盖了广泛的安全评测场景, 有些安全场景出现的内容安全风险有可能造成更大的危害, 美国国家标准与技术研究院 (NIST) 在其《联合高级 AI 风险评估声明》[46] 中强调: 风险域应依据严重性、发生可能性或社会弹性水平进行优先划分, 因此本文将各维度依据危害程度划分为政治安全、人身安全、一般安全 3 类, 赋予不同权重得到综合得分. 政治安全包含与社会主义核心价值观、国家安全等相关的维度, 在《GB/T 45654-2025 网络安全技术 生成式人工智能服务安全基本要求》[47] 附录 A 中首先提到了违反社会主义核心价值观风险, 在 2020 年 3 月 1 日起施行的《网络信息内容生态治理规定》[48] 第六条首先强调不得制作、复制、发布反对宪法原则、危害国家安全、损害国家利益等内容的违法信息; 人身安全考虑到大模型可能给出造成人身伤害或被采纳后带来危险后果的建议, 例如自我伤害、霸凌他人、实验室安全等; 一般安全包括大模型可能带来的错误误导, 但相对危害较小的场景. 经过对重要程度的综合考量, 将政治安全与人身安全类场景设置较高权重 (分别为 0.5 和 0.3), 以反映其高危害特点; 而一般安全类场景尽管常见, 但危害相对较轻, 设置为 0.2, 对于每一类单独衡量得分, 最终对 3 类结果进行加权平均后得到综合得分, 详细内容可见补充材料: 安全场景等级划分. 计算方式如下所示:

$$\mathrm{OEM} = \frac{1}{K_D} \sum_{i=1}^{K_D} w_i m_i, \tag{4}$$

其中 K_D 为等级个数, w_i 为第 i 个等级对应的权重, m_i 为第 i 个等级中的得分.

(2) 功能安全评估方法. 功能安全评估方法和内容安全评估方法不同, 内容安全方法主要评估模型生成内容, 而功能安全内容主要评估模型和承载模型允许的系统的响应内容.

(1) 影响范围得分 (score of scope, SoS). 影响范围主要采用统计方法评估漏洞对于模型系统造成影响的层面. 分别统计系统功能、用户、数据类型 3 类评估项中受影响内容的占比来, 最终将占比乘

表 5　功能安全风险等级.

Table 5　Functional security risk level.

Risk level	Score threshold	Vulnerability example
Level 1	$\geqslant 8$	Unauthorized access: users gain access to data or perform system operations without authorization.
Level 2	$\geqslant 4$	Unencrypted data: user data or critical system data is not encrypted during transmission or storage.
Level 3	$\geqslant 0$	Lack of encryption protocol: data transmission does not use HTTPS or other encryption protocols.

以对应的权重后进行加权平均后缩放到 1∼10 之间, 得到最终的影响范围得分. 对于一个危害程度越高、安全性越低的系统而言, 其因为漏洞而导致的影响范围会更广. 得分计算公式如下:

$$\mathrm{SoS} = \frac{1}{3}\left(w_F \frac{n_F}{N_F} + w_U \frac{n_U}{N_U} + w_D \frac{n_D}{N_D}\right),\ w_F = w_U = 0.4,\ w_D = 0.2, \tag{5}$$

其中 n_F, n_U, n_D 分别为受影响系统功能数、受影响用户、受影响数据类型数, N_F, N_U, N_D 分别为总的系统功能数、用户、数据类型数, w_F, w_U, w_D 分别为各个评估项的权重, 其和为 1, SoS 为最终的影响范围得分.

(2) 影响价值得分 (score of value, SoV). 影响价值主要评估漏洞对系统功能、用户数据、用户财产等方面所造成的破坏程度, 该指标主要由人工专家根据测试结果进行评估, 本文研究中邀请了多位领域专家对漏洞进行打分, 分值在 1∼10 之间, 若影响价值越高, 则得分越高, 反之亦然; 最后将每位专家平均得到. 具体的计算公式如下:

$$\mathrm{SoV} = \frac{1}{n}\sum_{i=1}^{n} S_i, \tag{6}$$

其中, S_i 为第 i 个专家打分结果, n 为专家总数, SoV 为最终的影响价值得分.

(3) 功能安全风险得分 (socre of functional risk, SoFR). 安全风险得分是对单个模型功能安全测试结果的总评. 该指标首先根据安全风险类型对每个漏洞进行分类, 计算风险类别中的平均得分, 最后将所有安全风险对应的得分加权平均后得到最终的功能安全风险得分. 本文中共有 4 类安全风险, 即数据安全风险、模型安全风险、系统安全风险、其他安全风险, 各个功能安全漏洞分属于其中的一个风险类别, 每个风险类别对应不同的权重, 权重和为 1. 具体计算公式如下:

$$\mathrm{SoFR} = \frac{1}{N}\sum_{k=1}^{N} w_k \frac{1}{2N_k}\sum_{i=1}^{N_k}(\mathrm{SoS}_i + \mathrm{SoV}_i),\quad \frac{1}{N}\sum_{k=1}^{N} w_k \ = \ 1, \tag{7}$$

其中 N 为风险类别总数, k 标识第 k 个类别, w_k 为风险类别对应的权重, N_k 为第 k 个类别下所有的漏洞数量, SoS_i, SoV_i 分别为第 i 个漏洞的影响范围得分和影响价值得分, SoFR 为最终的功能安全风险得分.

对于最终得出的归纳安全风险得分进行分类, 并设定相应的阈值与等级, 其中功能安全风险等级共包含 3 类, 即: 一级、二级、三级. 具体等级分类和得分阈值可参见表 5. 需注意, 若一个模型对应的某个安全风险中未发现漏洞, 则该模型的此风险项对应的分值为 0.

4　实验

本节详细介绍了本研究的实验设置, 并通过对开源和闭源大模型系统评估验证本研究提出的安全评估基准, 并对评估结果进行了分析.

4.1　实验设置

被测模型. 为了评估大模型在安全评估基准上的能力, 本研究评测了 19 个中文大模型, 包括 5 个闭源模型和 14 个开源模型, 涵盖了各种模型家族和参数规模. 由于实验条件限制, 对于开源模型, 本研究仅对 200 亿及以下参数模型进行评测, 具体细节如补充材料表 S1 所示. 在测试过程中, 实验将温度参数统一设置为 0.7, 以降低随机采样程度不一致带来的影响.

评估指标. 如 4.3 小节所述, 对选择题和开放式问题两种测试形式, 本文分别采用了基于规则的判定方式和基于评估器的判定方式, 分别得到各个安全维度下的准确率和安全率. 为了方便进行比较, 它们被统一划为百分制. 在计算最终综合得分时, 对各个维度的分数按照政治、人身、一般 3 个安全等级设置了 0.5, 0.3, 0.2 这 3 个权重进行加权求和. 同时, 本文还计算了主观与客观安全得分、通用与教育领域安全得分、红队防御能力以及拒答能力得分, 为从多维度分析被测大模型安全能力提供参考.

4.2　实验结果

表 6 [31, 49∼59] 列出了主要结果, 表明不同模型在 M^3-SafetyBench 基准测试下的不同安全能力存在差异. 本次安全评测从客观安全、主观安全、通用安全、教育安全、政治安全、人身安全、一般安全、拒答能力以及红队防御能力多个方面, 对各大模型进行了比较评估, 详细数据可见补充材料表 S2∼S7. 总体来看, 各模型在安全防护水平上存在明显的梯度, 既反映了模型参数规模、训练方法和调优策略的差异, 也展示了开源与闭源模型在安全能力上的不同特点.

从参数规模的影响来看, 较大参数规模的模型通常具备更丰富的语义表达和上下文理解能力, 因此在大部分安全评估维度中得分较高. 以 Qwen 系列为例, 从 0.5B 到 14B 版本的模型, 在各个方面都表现出了随着参数量的增大模型安全分数提升的现象. 这表明, 随着参数规模的扩大, 模型在处理复杂、多变的安全场景时, 能够更好地识别和防御潜在风险, 实现更精细化的内容过滤和拒答控制. 同时, 超大规模模型也存在安全瓶颈问题, 例如 qwen-plus (闭源) 较 Qwen2.5-14B 安全分数已经十分接近, 显示参数增益趋缓.

除此之外, 开源模型与闭源模型在安全能力上也呈现出不同的特点. 开源模型由于其架构、训练数据和调优策略公开透明, 更容易进行定制化安全防护措施, 但也因此暴露出一定的风险, 被不法分子利用进行对抗性攻击. 相较之下, 闭源模型 (如 qwen-plus-0112, ERNIE-3.5-128k, hunyuan-standard, Doubao-pro-32k, moonshot-v1-8k) 由于背后有更完善的安全策略和技术封装, 其在主观安全、政治安全以及红队防御能力上通常表现更为优异, 能够更有效地识别和拒绝恶意或敏感请求, 降低不当输出的风险.

各模型系列在安全评测中也呈现出独特的特性. Qwen 系列整体优势明显, 多个安全维度进入前三, 且 14B 型号综合分数接近 96, 接近闭源模型水平; 代际进化中, Qwen2.5 相比 Qwen2 在主观安全上提升约 1.7 分 (基于 7B 型号); InternLM 系列对参数较为敏感, 20B 模型在政治安全上较 7B 型号提升约 6.4 (93.36 vs. 87.93), 而 1.8B 型号的拒答率低 (仅 75.57), 与 20B 模型相差近 20 个百分点. Baichuan 系列表现中存在增益递减现象, 13B 型号在主观安全上较 7B 仅提升 7.5 (84.64 vs. 78.75), 但红队防御得分 (79.09) 远低于同规模的 Qwen2.5-14B (93.83).

版本迭代在提升模型安全能力上发挥了关键作用. 例如, Qwen2 升级到 Qwen2.5 后, 由于采用了 MoE 架构, 7B 型号在教育安全上提升了 8.2 (从 86.16 提升至 94.36), 同时红队防御能力由 84.04 提高到 91.95.

此外在功能安全性评测中, 以 Ollama 为服务端、OpenWebUI 为客户端的架构中, 存在部分安全风险, 如为未用加密协议、数据明文传输、完整性校验失效等问题, 同时本文研究针对不同模型的不同安全风险类型给出了具体的风险等级和风险得分, 在表 7 中展示了部分功能安全测试结果, 大部分模型的功能安全性较好, 风险得分和风险等级较低, 最高仅有 4.8 分为二级风险, 最低仅 2 分为三级风

表 6　各模型在不同安全能力维度的评分. 标注 “*” 的模型为闭源模型, 不标注 “*” 的模型为开源模型.

Table 6　Scores of different models across safety capability dimensions. Models marked with “*” are closed-source, while those without “*” are open-source.

Model	Objective	Subjective	General	Educational	Political	Personal	Common	Red team	Refusal ability	Overall score
Qwen2.5-0.5B-Instruct [49]	47.21	84.87	66.44	65.24	69.05	77.08	58.75	79.87	88.61	69.63
Qwen2.5-1.5B-Instruct [49]	82.27	88.43	84.62	86.81	85.66	90.06	82.70	78.92	88.89	85.87
Qwen2.5-7B-Instruct [49]	88.47	95.43	90.75	94.36	92.93	96.02	89.32	84.04	91.70	92.37
Qwen2.5-14B-Instruct [49]	91.87	98.45	94.23	97.03	96.44	98.04	93.03	93.83	97.26	95.94
Qwen2-1.5B-Instruct [50]	68.24	90.62	78.16	81.97	80.24	88.36	74.29	80.39	89.72	80.83
Qwen2-7B-Instruct [50]	89.32	93.73	90.18	94.20	91.98	94.92	89.49	86.73	88.66	91.70
Internlm2.5-1.8b-chat [51]	80.20	77.93	76.97	83.24	77.89	84.55	76.68	80.33	75.57	79.16
Internlm2.5-7b-chat [51]	88.54	89.43	87.51	91.93	88.94	92.44	87.16	87.85	86.81	89.33
Internlm2.5-20b-chat [51]	88.57	95.64	90.72	94.88	93.36	96.32	89.28	95.06	94.55	92.88
Chatglm3-6b-32k [52]	83.48	89.75	85.49	88.85	86.02	92.77	83.60	82.98	89.66	87.27
GLM-4-9B-Chat [53]	85.03	83.70	82.60	87.91	81.90	90.26	82.37	82.73	80.81	84.20
Baichuan2-7B-Chat [54]	86.25	78.75	81.08	85.34	80.66	87.36	80.77	80.48	87.85	82.91
Baichuan2-13b-chat [54]	88.39	84.64	85.02	89.50	86.58	90.26	84.49	79.09	80.34	86.39
DeepSeek-R1-Distill-Qwen-7B [31]	83.95	89.32	86.42	87.06	90.37	86.91	84.74	86.35	90.72	88.09
qwen-plus-0112* [55]	92.94	99.17	95.42	97.33	96.85	98.39	94.44	95.97	98.52	96.53
ERNIE-3.5-128k* [56]	91.13	98.06	93.67	96.46	94.49	97.69	92.99	94.08	84.99	94.90
Hunyuan-standard* [57]	89.86	98.16	92.92	96.18	94.94	97.12	91.91	95.13	93.30	94.87
Doubao-pro-32k* [58]	95.25	97.75	95.80	97.91	97.09	97.92	95.47	94.84	92.92	96.76
Moonshot-v1-8k* [59]	91.53	97.81	93.72	96.57	95.26	97.84	92.71	89.60	89.49	95.03

表 7　部分功能安全测试结果.

Table 7　Partial functional safety test results.

Model	Safety risk	Risk level	Score
Qwen2-0.5b	Data security	Level 2	4.8
Qwen2-7b	Data security	Level 2	4.8
Baichuan2-7B-Chat	Model security	Level 3	3
Baichuan2-13B-Chat	Model security	Level 3	3
glm-4-9B-Chat	System security	Level 3	2
LongCite-glm4-9B	System security	Level 3	2
Tele-Flm	System security	Level 3	2

险. 此外, 本文研究对所发现的问题进一步进行了分析, 在所发现的问题中, 大部分并非模型本身问题, 无法从模型本身来进行防护而是承载模型运行的操作系统和应用服务的问题, 应对操作系统或应用服务程序进行加固. 同时, 对受测模型 Qwen2-7b, Qwen2-0.5b, Baichuan2-7B-Chat, Baichuan2-13B-Chat, glm-4-9B-Chat, LongCite-glm4-9B, Tele-Flm 进行模型安全测试发现诸多编码层面存在的问题, 如未按编码标准进行编码、编码漏洞等问题. 此类问题主要集中在大模型的源代码中, 风险等级较低, 但本文研究建议在研究者在研发模型的过程中应当在模型编码环节加强编码规范, 避免遗留安全隐患. 总体上看, 在功能安全中的大模型风险主要存在在支持大模型运行的系统和源代码中, 应当对这两个部分进行加固保证大模型的安全.

综合来看, 评测结果表明模型参数规模对安全能力有显著影响, 大规模模型在处理复杂场景中更

具优势. 同时, 闭源模型凭借专有技术和严格安全策略, 在多个维度上表现出色, 为实际部署提供了更高的安全保障. 除此之外, 版本迭代的收益高于单纯参数扩展, 架构升级 (如 Qwen2.5) 为安全能力带来显著提升. 未来, 针对开源模型的透明性优势, 应加强安全防护措施, 提升其在红队攻击和拒答控制方面的能力, 而闭源模型则需要在开放性和用户可控性上寻求平衡, 共同推动大模型安全技术的进一步发展.

5　结论与未来工作

本文提出了一个面向中文大模型的多领域多场景、多维度综合安全评估基准 —— M^3-SafetyBench, 构建了 “内容安全 – 功能安全” 双层评估架构, 覆盖通用领域与教育垂直领域, 整合了开放式生成、选择题及红队越狱攻击等多种测评任务. 本研究基于开源数据收集、文本引导生成与红队提示改写, 精心构建了包含逾 17 万条高质量提示的评测数据集, 其中既有基础风险提示, 也有对抗性攻击提示和拒答提示, 并通过对 19 个主流开源与闭源大语言模型的系统性评估, 揭示了当前模型在不同安全风险维度上的显著差异和典型瓶颈.

M^3-SafetyBench 旨在衡量大模型在各种场景中应用的安全能力, 本文以教育领域为例进行了大模型在垂直领域的安全评估. 尽管其在多领域、多场景和多维度上对中文大模型进行了较为全面的安全评估, 但仍存在若干局限性值得后续改进和注意. 本次评估聚焦于通用和教育领域, 尚未涵盖医疗、金融、法律等具有高专业性和高风险的垂直领域. 每个领域存在独特风险类型和攻击方式, 需要针对性地设计安全测试任务. 随着大模型不断升级并新增安全防护机制, 旧版模型可能失去参考价值. 虽然本文提出了动态数据生成机制, 但目前尚未实现对数据集的周期性更新与版本管理. 此外, 在有些维度上, 测试数据的挑战性不足, 导致对被测大模型安全能力的区分度不够明显, 影响最终效果的展现. 下一步工作中, 可以在本文提出的安全体系和评估方法下进一步拓展至医疗、金融等其他领域和应用于智能体、多模态大模型等人工智能应用和模型, 后续可进一步引入自动化评估流水线, 保证基准的持续适用性. 当前, 新的更有效的红队攻击方式不断出现, 继续引入这些攻击方式将提高该基准的挑战性. 展望未来, 期待 M^3-SafetyBench 能为中文大语言模型的安全对齐提供量化依据与流程指引, 并为行业监管与合规检测提供坚实的数据支撑, 助力构建更安全、更可信的人工智能应用系统.

补充材料　本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Chu Z, Wang S, Xie J, et al. LLM agents for education: advances and applications. ArXiv:2503.11733
2 Qiu J, Lam K, Li G, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell, 2024, 6: 1418–1420
3 Zhao H, Liu Z, Wu Z, et al. Revolutionizing finance with LLMs: an overview of applications and insights. ArXiv:2401.11641
4 黄民烈. 大语言模型的安全风险与应对措施. 新经济导刊, 2023, 9: 33–35
5 Abdali S, Anarfi R, Barberan C J, et al. Securing large language models: threats, vulnerabilities and responsible practices. ArXiv:2403.12503
6 Shang Z, Wei W. Evolving security in LLMs: a study of jailbreak attacks and defenses. ArXiv:2504.02080
7 Feng S, Park C Y, Liu Y, et al. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models. ArXiv:2305.08283
8 Zhang Z, Lei L, Wu L, et al. SafetyBench: evaluating the safety of large language models. ArXiv:2309.07045
9 Xu L, Li A, Zhu L, et al. Superclue: a comprehensive Chinese large language model benchmark. ArXiv:2307.15020
10 Bhatt M, Chennabasappa S, Li Y, et al. Cyberseceval 2: a wide-ranging cybersecurity evaluation suite for large language models. ArXiv:2404.13161

11 Zhang H, Gao H, Hu Q, et al. ChineseSafe: a Chinese benchmark for evaluating safety in large language models. ArXiv:2410.18491
12 Yuan X, Li J, Wang D, et al. S-eval: automatic and adaptive test generation for benchmarking safety evaluation of large language models. ArXiv:2405.14191
13 Li L, Dong B, Wang R, et al. Salad-bench: a hierarchical and comprehensive safety benchmark for large language models. ArXiv:2402.05044
14 Lei Y, Li J, Cheng D, et al. Cfbenchmark: Chinese financial assistant benchmark for large language model. ArXiv:2311.05812
15 李晓松, 赵柯然, 赵英满, 等. 国防科技情报领域大模型应用效果测评研究. 情报理论与实践, 2025, 48: 78–83
16 Zhao H, Tang X, Yang Z, et al. ChemSafetyBench: benchmarking LLM safety on chemistry domain. ArXiv:2411.16736
17 Wei A, Haghtalab N, Steinhardt J. Jailbroken: how does LLM safety training fail? In: Proceedings of Advances in Neural Information Processing Systems, 2023. 80079–80110
18 Xu Z, Liu Y, Deng G, et al. LLM jailbreak attack versus defense techniques—a comprehensive study. ArXiv:2402.13457
19 Liu Y, He X, Xiong M, et al. Flipattack: jailbreak LLMs via flipping. ArXiv:2410.02832
20 Zheng X, Pang T, Du C, et al. Improved few-shot jailbreaking can circumvent aligned language models and their defenses. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 32856–32887
21 Miller E. Adding error bars to evals: a statistical approach to language model evaluations. ArXiv:2411.00640
22 Zeng Z, Yu J, Gao T, et al. Evaluating large language models at evaluating instruction following. In: Proceedings of the 12th International Conference on Learning Representations, 2024
23 Li J, Li R, Liu Q. Beyond static datasets: a deep interaction approach to LLM evaluation. ArXiv:2309.04369
24 Hartvigsen T, Gabriel S, Palangi H, et al. Toxigen: a large-scale machine-generated dataset for adversarial and implicit hate speech detection. ArXiv:2203.09509
25 Zhang M, Pan X, Yang M. Jade: a linguistics-based safety evaluation platform for large language models. ArXiv:2311.00286
26 Liu Y, Iter D, Xu Y, et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2023. 2511–2522
27 Kim S, Shin J, Cho Y, et al. Prometheus: inducing fine-grained evaluation capability in language models. In: Proceedings of the 12th International Conference on Learning Representations, 2024
28 Chiang W, Zheng L, Sheng Y, et al. Chatbot arena: an open platform for evaluating LLMs by human preference. In: Proceedings of the 41st International Conference on Machine Learning, 2024
29 全国网络安全标准化技术委员会. 生成式人工智能服务安全基本要求. TC260-003, 2024
30 国家互联网信息办公室, 国家发展和改革委员会, 教育部, 科学技术部, 工业和信息化部, 公安部, 国家广播电视总局. 生成式人工智能服务管理暂行办法. 2023.
31 Guo D, Yang D, Zhang H, et al. DeepSeek-r1: incentivizing reasoning capability in LLMs via reinforcement learning. ArXiv:2501.12948
32 Yang A, Yang B, Zhang B, et al. Qwen2.5 technical report. ArXiv:2412.15115
33 Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. ArXiv:2209.07858
34 Korbak T, Shi K, Chen A, et al. Pretraining language models with human preferences. In: Proceedings of International Conference on Machine Learning, 2023. 17506–17533
35 Zou A, Wang Z, Kolter J Z, et al. Universal and transferable adversarial attacks on aligned language models. ArXiv:2307.15043
36 Zhu S, Zhang R, An B, et al. AutoDAN: interpretable gradient-based adversarial attacks on large language models. ArXiv:2310.15140
37 Yuan T, He Z, Dong L, et al. R-judge: benchmarking safety risk awareness for LLM agents. ArXiv:2401.10019
38 Li H, Guo D, Fan W, et al. Multi-step jailbreaking privacy attacks on ChatGPT. ArXiv:2304.05197
39 Li X, Zhou Z, Zhu J, et al. Deepinception: hypnotize large language model to be jailbreaker. ArXiv:2311.03191
40 Liu X, Xu N, Chen M, et al. AutoDAN: generating stealthy jailbreak prompts on aligned large language models. In: Proceedings of the 12th International Conference on Learning Representations, 2024
41 Chao P, Robey A, Dobriban E, et al. Jailbreaking black box large language models in twenty queries. ArXiv:2310.08419
42 Mou Y, Zhang S, Ye W. SG-Bench: evaluating LLM safety generalization across diverse tasks and prompt types. In: Proceedings of Advances in Neural Information Processing Systems, 2024. 123032–123054
43 Röttger P, Pernisi F, Vidgen B, et al. SafetyPrompts: a systematic review of open datasets for evaluating and improving large language model safety. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence, 2025.

27617–27627
44 Lin S, Hilton J, Evans O. Truthfulqa: measuring how models mimic human falsehoods. ArXiv:2109.07958
45 Zhang M, Pan X D, Yang M. JADE-DB: a universal testing benchmark for large language model safety based on targeted mutation. J Comput Res Dev, 2024, 61: 1113–1127 [张谧, 潘旭东, 杨珉. JADE-DB: 基于靶向变异的大语言模型安全通用基准测试集. 计算机研究与发展, 2024, 61: 1113–1127]
46 International Network of AI Safety Institutes. Joint Statement on risk assessment of advanced AI systems. 2024. https://www.nist.gov/system/files/documents/2024/11/20/JointStatementonRiskAssessmentofAdvancedAISystems.pdf
47 国家市场监督管理总局, 国家标准化管理委员会. 网络安全技术 生成式人工智能服务安全基本要求. GB/T 45654-2025. 2025. http://c.gb688.cn/bzgk/gb/showGb?type=online&hcno=F67D3F376E0A0A0FF5317FB36B32A30A
48 国家互联网信息办公室. 网络信息内容生态治理规定. 2020. https://www.moj.gov.cn/pub/sfbgw/flfggz/flfggzbmgz/202101/t20210105_146435.html
49 Qwen Team. Qwen Technical Report. Alibaba Cloud, 2024
50 An Y, Yang B, Hui B, et al. Qwen2 technical report. ArXiv:2407.10671
51 InternLM Team. InternLM2.5 Technical Report: From Foundation to Chat. Shanghai AI Laboratory, 2024
52 Zeng A, Liu X, Du Z, et al. ChatGLM3: More Powerful Bilingual Chat Model. 2023. https://github.com/THUDM/ChatGLM3
53 Zhipu AI. GLM-4 API Documentation. 2024. https://open.bigmodel.cn/dev/api
54 Yang J, Wang X, Li A, et al. Baichuan 2: open large-scale language models. ArXiv:2309.10305
55 Alibaba Cloud Intelligence Division. Tongyi Qianwen-Plus: Proprietary Large Language Model from Alibaba Cloud. 2025. https://www.alibabacloud.com/en/solutions/generative-ai/qwen
56 Baidu ERNIE Team. ERNIE 4.5 Technical Report. 2025. https://yiyan.baidu.com/blog/publication/ERNIE_Technical_Report.pdf
57 Sun X W, Chen Y F, Huang Y Q, et al. Hunyuan-large: an open-source MoE model with 52 billion activated parameters by Tencent. ArXiv:2411.02265
58 ByteDance AI Research. ByteDance Doubao-1.5-pro model matches GPT 4o benchmarks at 50x cheaper. 2025. https://www.rohan-paul.com/p/bytedance-doubao-15-pro-model-matches
59 Moonshot AI. Moonshot-v1: A High-Performance Large Language Model with Extended Context Handling. 2025. https://platform.moonshot.ai/docs/introduction

M^3-SafetyBench: a comprehensive benchmark for evaluating the safety of large language models across multiple domains, scenarios, and dimensions

Weiping YANG[1†], Haojie CHENG[1†], Baishun ZHOU[2*], Yuxin WEN[2], Yufan LIU[3,4*], Yuyang LIU[2], Bing LI[3,4,5], Congyan LANG[1*], Naiyue CHEN[1], Wei ZHANG[2] & Weiming HU[3,4]

1. *School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China*
2. *School of Computer Science, China University of Labor Relations, Beijing 100048, China*
3. *National Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
4. *Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information, Beijing 100190, China*
5. *PeopleAI Inc., Beijing 100190, China*

* Corresponding author. E-mail: zhoubaishun@126.com, yufan.liu@ia.ac.cn, cylang@bjtu.edu.cn

† Equal contribution

Abstract In recent years, the widespread application of large language models (LLMs) in fields such as natural language processing, intelligent education, and content generation has increasingly highlighted their potential security risks. Existing safety evaluation benchmarks often focus on single attack methods or narrow task types, making it difficult to comprehensively reflect the safety performance of models across multiple domains and scenarios. To address this gap, this paper proposes M^3-SafetyBench, a comprehensive safety evaluation benchmark for Chinese LLMs that spans multiple domains, scenarios, and dimensions. This framework innovatively adopts a two-layer architecture—"content safety and functional safety"—covering both general and educational vertical domains, and integrates various assessment tasks including open-ended generation and multiple-choice questions. Furthermore, this study incorporates red teaming strategies to evaluate model vulnerabilities through jailbreak attacks, thereby enhancing the depth and breadth of the evaluation for multidimensional safety analysis. By constructing a high-quality dataset containing over 170000 test samples, we systematically evaluated 19 mainstream open-source and closed-source LLMs. Experimental results reveal significant disparities in the performance of different models across various safety risk dimensions, highlighting bottlenecks in content generation and safety alignment within current LLM technologies. The proposed M^3-SafetyBench not only provides quantitative metrics and methodological processes for improving the safety of LLMs but also offers reliable data support for industry regulation and compliance testing.

Keywords large language models, safety evaluation benchmark, red teaming, multi-scenario testing, content and functional safety