计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330955
61(5): 1104−1112, 2024

多视角看大模型安全及实践

王笑尘　张　坤　张　鹏
（北京智谱华章科技有限公司　北京　100086）
（xiaochen.wang@aminer.cn）

Large Model Safety and Practice from Multiple Perspectives

Wang Xiaochen, Zhang Kun, and Zhang Peng
（*Beijing Knowledge Atlas Technology Co., Ltd., Beijing* 100086）

Abstract　With the widespread application of big models in the field of artificial intelligence, the security issues of large models, especially large language models, have received attention. As an emerging technology, the security situation analysis of large models and the construction of security systems need further exploration. We analyze the overall trend of large model security from two perspectives: society and technology application. Based on the characteristics of large models, we sort out the practice of large model security building, and provide a reference plan for building a security system for large model development and large model application construction. The large model security practice plan introduced in this article includes three parts: security benchmark construction, model values alignment method, and model online service security system construction.

Key words　large model; large model online service; security system; artificial intelligence ethics; large model security situation analysis

摘　要　随着人工智能领域大模型（large model）的广泛应用，大模型，尤其是大语言模型（large language model, LLM）的安全问题受到了广泛关注. 大模型作为一种新兴技术，与之相关的安全态势分析以及安全体系建设均亟待挖掘与探索. 从社会关系以及技术应用2个视角，分析了大模型安全的整体趋势. 同时，基于大模型自身的特点，梳理了大模型安全能力建设的实践思路，为大模型研发、大模型应用构建提供了安全体系构建的参考方案. 介绍的大模型安全能力实践方案包括安全评估基准建设、模型价值观对齐方法、模型线上服务安全系统建设3个部分.

关键词　大模型；大模型线上服务；安全系统；人工智能伦理；大模型安全形势分析

中图法分类号　TP183

随着深度学习技术的不断发展和应用，大模型逐渐成为学术界和工业界研究的热点. 这些模型具有非常强大的表示能力和学习能力，可以处理各种复杂的任务，如图像识别、自然语言处理、语音识别等. 特别是近几年大语言模型 (large language model, LLM) 所展现的惊人能力，已开始对社会发展造成一定影响，因此模型安全问题开始显得尤为重要. 本文从人类整体发展、国家竞争以及技术应用3个不同视角看待大模型安全，将重点介绍技术应用方面，并主要讨论3个参与方的安全管理问题，最后分析生成内容安全的未来发展趋势.

本文梳理了大模型安全能力建设的实践思路，从模型安全评估基准（benchmark）建设、模型价值观对齐（alignment）方法、模型线上服务安全系统3个方面，阐述了大模型安全能力支撑体系的构建方案.

本文所述的大模型安全，主要围绕大模型生成

收稿日期：2023-11-30；**修回日期**：2024-03-11
通信作者：张鹏（peng.zhang@aminer.cn）

内容的安全性. 大模型安全主要包括系统安全、生成内容安全 2 个方向. 大模型系统安全方面所遇到的问题、问题的解决方案与计算机领域传统信息安全领域接近，本文对此方面仅进行简要介绍. 生成内容安全是大模型特有的安全问题，这一问题是由大模型自然语言生成能力相对于以往的语言模型有显著提升带来的，因此本文主要针对生成内容安全问题进行讨论.

1 大模型安全整体趋势

1.1 大模型社会安全态势分析

从人类社会的整体角度来看，大模型对社会的促进方式理想情况下应当是帕累托式改进，即在不减少公共利益的情况下提升社会效益. 然而，由于深度学习能力的不可解释性，实际上很难做到有利而无弊. 因此，我们在利用大模型提高整体社会效率的同时，需要像关注核技术一样，关注它们可能产生的破坏效果. 目前，大模型具有一定程度的记忆和推理能力，本身具备一定的价值观[1]. 模型参数量越大，其具有的信念越强. 大型人工智能模型的价值观是由训练语料决定的，如今大部分新产生的数据来自于互联网，例如，GPT-3[2] 的训练语料有 82% 来自互联网. 由于大模型的价值观通常与互联网语料表现形式相同，因此，许多研究关注到大模型的社会刻板印象和种族歧视现象[3]，与互联网相似.

随着大模型的商业价值越来越大，算法和训练数据构成开始走向闭源模式，越来越难查看到大模型底层的预训练数据. 例如，GPT-4[4] 开始不再公布训练语料和算法. 这意味着大模型本身对社会产生的影响可能被掌握在少数算法训练人群中，公众逐渐失去了对模型价值观的监管能力. 因此，制定大模型预训练的标准并发放相关牌照已成为当务之急.

从国家竞争的角度来看，目前全球大模型的发展水平非常不平衡. 美国占据了绝大多数的大模型[4-6]，而欧洲[7]、俄罗斯[8] 和中国[9] 也具备研发大模型的能力. 支撑大模型能力的 3 个关键维度是算法、算力和数据. 其中，算法需要算力实验和基础教育的支持. 从宏观层面来看，竞争的关键是整个科研实力；而从微观层面来看，竞争的核心主要是数据和算力.

就数据而言，虽然大模型具有一定的知识迁移能力，可以将一些语言的知识迁移到另一种语言中，但研究表明小语种的效果相比于大模型仍然处于一定的劣势[3]. 因此，对于某些小语种来说，使用大模型提高社会效率的难度更大，这可能会加剧各国之间发展的不平衡.

从算力方面来看，整个芯片产业链的关键环节在美国. 以英伟达为代表的公司控制着计算的软硬件生态. 在软件层面，大量的算法库围绕英伟达芯片进行开发，逐步形成了其“护城河”. 在硬件层面，AI 芯片的折旧速度惊人，达到年化 70%[10]. 后发芯片公司在对方高速发展的情况下追赶非常困难，而下游大模型公司在商业竞争中如果不使用美国芯片，将面临巨大的竞争压力. 此外，美国还使用 AI 芯片禁用等手段，限制其他国家的 AI 发展. 除了经济角度，由于世界各国的价值观存在差异，不具备研发能力的国家采纳发达国家的大模型，可能对本国产生一定的意识形态冲击. 因此，促进国内大模型生态的有序发展，并掌握大模型生态的关键环节成为国家间竞争的核心.

1.2 大模型技术应用安全态势分析

从大模型技术应用角度看，整个链路的实质总结为：用户使用大模型公司的大模型系统. 因此确保大模型安全，实质是对整个链路的 3 个参与方：用户、大模型系统和大模型公司 3 个维度的相关安全管理.

对于用户，主要是用户在使用大模型系统的全生命周期管理，用户整个生命周期分别包括用户登录、用户输入、用户获得系统反馈、用户使用和用户分享共 5 个环节，这 5 个环节都产生一定的用户数据. 对这些用户数据的处理主要分为用户隐私保护和用户管理.

国家对用户隐私保护极为重视，已经立法如《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》，在立法的基础上，经过这几年的发展演化出行业对用户数据管理上的一些标准和共识，进而方便企业的合规执行，例如在用户隐私数据等级上，可以按照相关数据泄露后对用户造成的伤害进行安全等级分类，根据安全等级分类和相关数据可独立推断或关联推断，确定用户数据字段保护密级. 而根据不同的密级，会有不同的存储和管理规范，对于极端隐私数据，都需要去标识匿名化或者差分存储. 而用户管理分为青少年用户管理、风险用户识别、用户溯源、举报管理等，可通过对用户数据分析，控制用户对不良内容的传播，防止出现社会突发事件，这方面的很多工作需要企业和公安机构配合完成.

对于大模型公司,最为重要的是安全机制的顶层设计,其中最为核心的是管理制度和应急响应机制.为确保机制的顺利执行,应由企业最高决策人员担任要职.用户数据访问权限需要严格控制,所有系统操作均应有日志备案,相应人员都需经过安全培训,企业还应定期组织安全演习,确保整体系统具有一定的弹性安全能力.

大模型安全问题主要涉及系统安全和生成内容安全2个方面.随着大模型系统社会影响力的增强,系统应确保不被黑客等外部人员攻克,以防止内部数据泄露.在技术应用层面的安全中,生成内容安全是重中之重,也是一个全新的领域.2023年4月发布的《生成式人工智能服务管理办法(征求意见稿)》,开启了相关立法;而对于执行标准,行业还处于摸索阶段,目前更多的是企业按照一定的法律常识和原则自查.这和个人隐私安全问题既相似又不同,相似点在于根据生成内容对社会的破坏力制定的危害等级,通过等级约束,推动企业安全技术发展;而不同在于界定安全是否具有很大的可裁量空间,随着时间变化,标准是否会跟随变化.

大模型生成内容安全的核心是对齐价值观.从大模型诞生之前,如何让模型对齐人类的价值观成为重要研究方向[11]:假定为足够能力的模型和正确的训练算法能够学习到足够精度的用户意图,以及评判行为比产生行为代价更小,当满足这2个条件,人类可以通过设计奖励模型或者微调的方法对齐人类价值观[12].尽管模型对齐过程中会产生一定的对齐税,但并不妨碍人类主观上更喜欢对齐后的模型[13-14].除了在预训练阶段严格把控训练数据中的安全外,对齐也因其成本低、迭代快成为了解决安全问题的重要方法.

解决大模型生成内容安全的主要方案是采取攻防对抗.对抗的实质是设计机制来发现问题和解决问题,而根据发现、解决问题的主体是人类还是模型,组合起来有4种情况,其中模型发现的问题让人类来解决困难较大,因此主要讨论3种情况.第1种情况是人类发现问题,人类来设计系统解决问题.这类方案主要解决大模型存在的天然缺陷:一个是大模型自身价值之一能生成多样性的内容,不能生成多样内容的大模型失去了大部分应用价值,然而这对于回答某些容错性比较低且不太喜欢多样性的政治事实性问题,是灾难性的;另一个是大模型迭代周期慢,预训练动辄月计,微调则需要周计,对于社会突发事件,无法承受这样的干预周期.这都需要人类参与设计一些辅助系统如关键词屏蔽,或者利用大模型上下文学习(in-context learning, ICL)[2]能力,或者在提示(prompt)中置入一些知识给模型,来保证大模型的安全.研究表明,在碰到敏感问题时,增加少量prompt能明显改善模型的安全能力.

第2种情况是人类发现问题,大模型来解决问题.通过传统的规则式方法干预模型,由于人类能提供的输入数据量相比模型预训练数据量很小,干预效果不佳.另外,有研究表明随着模型生成能力的强大,为增加模型的鲁棒性,扩大奖励模型的参数量[14].这2个原因表明,期望不损失大模型能力的情况下,采用传统规则或者小模型监管大模型的方法不太可能成立.例如,用户很容易通过大模型的思维链(chain of thought, COT)能力[15],轻松绕过关键词系统从而生成一些有害内容,让关键词干预系统失灵.一种解决方案是由人类众包方式扮演红队,通过诱导模型犯错的方式发现问题,然后再人类矫正答案,将矫正的正确答案再通过模型对齐的方式来增强模型内安全能力,使得模型对齐正确的价值观[16].这种方法非常有效,因为大模型出色的泛化能力,经过多轮迭代后,人类能够找到的大部分模型漏洞都被解决,而且随着漏洞的解决,人类未来能够找到模型新漏洞的难度越来越大.

随着模型智能能力的提高,研究发现大模型具备了一定自举能力[17]和推理能力[18],模型开始逐步摆脱人类标注数据,通过理解人类制定的规则的方式实现自我安全进化.一种方案是激发模型根据评论、修订然后自监督学习能力的方式解决问题[18],相比于人类标注数据,能够更好地平衡模型的有用性和有害性.

最终解决安全问题的路线,可能会是大模型之间的相互博弈,即第3种情况:由模型发现问题,模型解决问题.让一个模型(攻击大模型)给另一个模型(安全大模型)生成攻击性数据,安全大模型则通过训练迭代提高自己的能力.这种大模型在大模型的世界中相互提高安全水平[18]方案的逻辑是:当攻击大模型的能力进化到超过所有人类发现问题的能力时,只需要安全大模型能够足够强地应对攻击大模型,则人类关心的安全问题基本解决.而这种情况是可能发生的,大模型相比于人类能读到更多的数据并具备更强的知识关联的能力,让模型生成有害数据的能力超过人类成为大概率事件.这个路线在实际执行中也更容易操作,因为监管机构的数据样例可能用尽,而其私有数据也不容易公开给企业,避

免企业记住监管机构的所有样例，让模型更容易起到自动监管效果.

攻击大模型能力提高的同时，伴随着更安全大模型的产生. 这需要依赖攻击大模型强大的自学能力，可能的路线图会有 2 种：1）通过人类发现安全大模型的一个漏洞，让攻击大模型去学习和发现更多的漏洞；2）动态实时地根据互联网内容，去学会编造事实，生成诱导内容，不断在安全大模型中尝试寻找破绽，以发现安全大模型的漏洞.

除了模型间相互发现问题、解决问题，模型安全鉴定问题也需要大模型相互博弈完成. 大模型如今一方面不仅生成内容非常逼真，而且容易产生幻觉，这种真假内容混合的方式，极大地增加了鉴定困难；另一方面，由于大模型生成成本低，未来人类看到的大部分内容将都由 AI 生成，如此大的数量的安全问题，让人类审核以几乎没有可能. 人类如不依靠大模型鉴定大模型安全将难以为继. 而如何构建安全鉴定大模型，也是未来重要工作的方向之一.

随着技术的发展，大模型开始拥有使用工具和多模态的能力[19-20]. 由于这方面技术还未处于大规模普及阶段，因此安全问题还没有完全显现. 但实际应用中，多模态已经存在大量安全问题，DeepFake 技术的原理是将真实图像或视频作为输入，通过深度学习算法生成新的虚假图像或视频. 这些虚假图像或视频可以模拟真实人物的面部表情、口音、动作等特征，达到以假乱真的效果. 而未来大模型能以图文并茂的方式生产多模态内容，会让一些虚假事实显得更加真实. 大模型多模态问题的难度更大，因为单纯看局部文字和图片可能均无异常，因此不能将问题等价于分别识别文字和多模态内容，需要结合上下文综合理解图文的含义，以发现系统安全漏洞.

模型生成能力越强，对安全的挑战也就越大. 从系统角度看，安全和效率是一对矛盾体，这可能产生 3 种策略：1）要求安全检查能力和模型生成能力相匹配，即有多强的安全能力，就能给社会提供多强的内容生成能力；2）允许模型生成能力适度超过安全检查能力，只需要保证模型生成对社会整体利大于弊，且弊端可由安全机制可控；3）要求安全能力超过模型生成能力. 当前阶段，促进生成内容的技术发展是主流，是当前社会的主要矛盾，因此采用第 2 种策略更为合理. 而安全能力不能落后于生成能力太多，否则远强于安全能力的模型生成能力，可以直接用于攻击模型，导致安全问题无法有效解决. 除此之外，由于生成内容的技术发展迅速，安全问题也日益严峻. 因此，需要加强监管和规范，确保模型生成的内容符合社会道德和法律规范，同时不会对用户隐私和数据安全造成侵害. 监管机构需要加强对模型生成过程的监督和评估，及时发现和解决安全问题，同时也需要与行业各方合作，推动安全技术的研发和应用，共同构建安全可信的生成内容生态系统.

2 大模型安全技术实践

本节主要介绍大模型安全技术的具体实践方案，包括安全评估基准建设、模型价值观对齐方法、模型线上服务安全系统建设 3 个部分.

2.1 安全评估基准建设

大模型安全技术中，模型安全评估能力的建设是首先需要实现的. 模型安全评估能力为大模型提供了安全的基准，后续安全能力建设需要围绕安全评估结果进行持续迭代.

安全评估基准的建设，首先可以依托于公开的安全评估基准. 公开的安全评估基准通常会发布评测结果榜单（leaderboard），榜单的公开指标有助于模型开发者比对自身模型与行业内其他模型的安全能力. 现有公开中文安全评估基准有 CValues[21]、C-eval[22] 等.

在大模型安全评估过程中，红队测试（red teaming）攻防对抗是基础评估模式，在大模型正常线上运行的过程中，能够收集到的用户风险输入通常较为稀疏，运营者难以收集到足够数量的攻击样例，因此需要引入红队（国内部分文献称为蓝队，均表示生成风险内容的攻击方）构造风险样本，扩充评价数据集. 例如 CValues[21] 评价体系中，邀请了专家手动收集了 10 个场景中的对抗性安全提示以及 8 个领域的诱导提问. 考虑到纯人工构建风险数据的高成本、低效率，当前学术界构建了多种基于大模型的风险内容生成机制用于红队测试[16].

安全评估基准中的评价方式通常包括人工评价和自动评价 2 种. 由于大模型生成内容具备多样、多变的特点，单纯的自动评价难以满足模型价值观、安全性的评估需求，因此引入人工评价是必要的. 但人工评价具有成本高、耗时长的特点，不适用于模型频繁迭代过程中的高频评价. 因此评估基准中应当包含自动评价能力，支持对模型安全性的快速检测.

安全评估基准的评分方法包括评分制、对比制、选择题打分制、第三方模型评价制等.

评分制通常用于人工评价场景，由专业评估人员对模型回复进行评分，在评测数据集上汇总得出

模型安全总评分. 评分区间可以选择多值评分或二值评分. 多值评分允许评分人员在 0~5、0~10 等评分区间中选择一个分数对模型回复，可以对模型的回复水平进行较为精准的度量. 而二值评分仅允许评分人员评价“安全”或“不安全”，在评价精准度上劣于多值评分，在评价效率上优于单值评分. 例如 CValues[21] 评价体系，即包含了人工评分的评测模式.

对比制是通过对比不同模型的回答水平，从而得出不同模型能力排名的一种评测方式，例如 Chatbot Arena[23] 平台即采取这种评测方式. 在对比制评价中，评分人员通过选择更优的模型回复来进行评价，并由一种排位系统得出参与评价的各个模型最终的排名分数. 对比形式的模型评价对于评测者更加友好，规避了单纯评分制中评分人员难以判断模型回复是否足够好的情况. 但对比制评价需要有较多被评测模型参与，才能得出有价值的分数.

选择题打分制是自动评估中常用的评分方法. 这一评分方法向模型提供包含安全内容、不安全内容的选择题，要求模型从中选出安全或不安全的内容. 选择题的构建可以人工进行或人工结合大模型生成能力进行. 在人工结合大模型生成能力构建选择题的方式下，安全基准建设人员可以要求模型针对指定问题，生成安全或不安全的回复内容，随后收集生成内容作为选项. 选择题打分制可以通过模型回复的选项进行自动打分评价，因此适用于自动评估场景. C-eval[22]、CMMLU[24] 等评测集采用了这一评价方式.

第三方模型评价制，是通过能力较强的第三方模型作为评估器，对被评测模型进行安全性评价的方法[21]. 这一方法通常将被评测模型对于特定问题的回复输入到第三方模型中，要求第三方模型对回复的安全性进行评价. 这一方法同样可以实现高效率的自动评价，但存在 2 个缺点：首先，通过第三方模型进行评价，准确性取决于第三方模型的能力，而且第三方模型本身的版本迭代也会导致评价结果不稳定[25-26]；其次，由于大模型通常以线上服务，而非离线模型的方式提供模型调用支持，使用第三方模型进行评价，存在测试数据泄露的风险.

除了针对大模型的生成能力进行评估，过往的自然语言处理研究中，学术界也同样积累了大量传统自然语言处理任务的评测集. 这些评测集度量了文本分类、文本检索、机器翻译、观点提取等任务上各类模型的水平. 由于大模型的语言能力强于过往的传统自然语言处理模型，将上述自然语言处理任务评测集用于大模型的能力度量也是可行的. 在传统自然语言处理任务中，情感分析、语义相似性分析任务的需求与大模型安全评价需求具有相似性，相关评价数据集的构建方式可以迁移到大模型安全评价领域. 而机器翻译任务的评测，则有助于评价模型在多语言场景下的安全能力.

在大模型安全基准的构建中，通常需要综合上述方法进行构建. 以下以 SafetyBench[27] 为例，描述安全基准的构建流程：第 1 步，构建者应当决定安全基准的评价机制，这一点决定了安全基准的后续构建方式. 在 SafetyBench 中，构建者采用了选择题打分制. 第 2 步，确定安全基准覆盖的子领域，根据训练数据范围和训练方法的不同，大模型对于不同的安全领域往往体现出不同的安全水准，将安全基准拆分为多个子领域，有助于模型训练者针对模型的弱项进行针对性强化. 在 SafetyBench 中，构建者选择了冒犯性、不公平与偏见等 7 个子领域. 第 3 步，收集各个子领域的相关数据，构建测试集. 数据的收集来源可以是现有数据集、安全相关的考评测试题，以及通过大模型进行数据增强获得的数据. 在初步的数据收集完成后，对于已有数据的质量控制是决定最终安全基准可靠性的关键步骤，这一步通常通过众包标注进行.

上文介绍了大模型安全评估基准的评价方式，并介绍了数种公开安全评估基准. 在大模型的具体应用中，开发者可以参考公开评估基准的构建方法，构建私有的模型安全评估基准，面向特定的应用场景进行针对性安全评估.

2.2 模型价值观对齐方法

在大模型训练过程中，模型生成内容安全主要通过对齐步骤来实现. 大模型对齐，是指通过特定的技术手段，控制模型能力符合训练者意图的过程[28]. 大模型对齐的代表性技术为监督微调(supervised fine-tuning, SFT)[29] 与人类反馈强化学习(reinforcement learning from human feedback, RLHF)[13].

监督微调方法主要依赖有监督数据对模型进行微调，从而实现模型对齐. 监督微调的典型实现方法为构造“输入-输出”形式的数据对，对模型进行端到端微调. 在大模型微调场景中，对参数量巨大的整个模型进行微调，时间成本与计算资源成本都较高，因此近年来涌现出了一些增强微调效率的技术：大模型低秩自适应(low-rank adaptation of large language models)方法[30]，通过冻结基础模型的参数，引入少量新增网络层，对新增网络层进行微调的方法来提升

训练效率；P-tuning v2 方法[31]将 Prefix-Tuning[32]方法引入到大模型微调中，在模型中拼接新的可训练参数，对模型实现微调. 监督微调方法的实现较为简单直接，能够高效应用于特定方向的模型内容安全强化，但该方法存在泛化能力差的问题.

人类反馈强化学习方法在经过了监督微调后的模型基础上，引入强化学习机制，进一步实现人类偏好对齐. 人类反馈强化学习方法需要人工对模型的“输入-输出”组合进行质量评分，随后基于这一数据训练一个奖励模型（reward model）. 奖励模型则通过强化学习的方式，对大模型进一步微调. 人类反馈强化学习方法相比于监督微调方法具备更强的泛化性，但训练成本更高、数据要求更高、训练也更不稳定.

针对人类反馈强化学习方法存在的不足，有多种方法对其进行了完善. 例如宪法 AI（constitutional AI）[17]实现了通过大模型根据设定的“宪法准则”来生成监督信息，实现了人类反馈强化学习方法的自监督训练. DeepMind 提出的社会沙盒对齐方法[33]，则进一步通过设定社交游戏场景，令大模型智能体模拟人类社会交互，生成数据进行模型对齐.

2.3 模型线上服务安全系统建设

模型线上服务是大模型应用的基础形式，以 ChatGPT 为代表的诸多大模型均提供大模型线上服务. 大模型线上服务的形式通常是模型对话服务，支持用户向模型提问并由模型给出符合需求的回答. 在大模型提供线上服务的过程中，需要在系统中集成安全服务系统，从而实现实时、灵活、高效率的内容安全控制.

模型线上服务的安全系统应当满足 4 个要求：

1）实时性. 作为线上服务，模型安全系统应当能够实时生效，对风险内容进行第一时间处置，避免用户接触到风险内容.

2）灵活性. 内容安全是一个动态化需求，是随着时间不断变化的. 模型安全系统应当能够被运营人员即时调整，实现内容安全控制规则的实时增加、删除、修改.

3）高效率性. 线上服务的调用频度通常较高，且大模型的运行会消耗较多的计算资源. 在此前提下，模型安全系统应当使用尽量低的计算资源，实现安全保障能力.

4）可评价性. 模型安全系统应当提供安全保障水平的评价能力，支持系统运营人员对安全系统的屏蔽率、误杀率、计算效率等核心指标进行观察，并基于评价指标进行系统调优.

基于上述 4 个要求，我们提出了包含多项具体技术的模型线上服务安全体系. 安全体系中具体技术的生效机制，可以概括为“触发器-响应机制”这一模式. 触发器是触发安全机制的开关，接受用户输入或模型输出，根据特定的信号触发后续的响应机制. 触发器的具体技术实现可以是关键词匹配、向量化召回、文本分类模型、用户分类模型等. 响应机制则是对模型实现干预的机制. 安全响应机制会对模型的输入造成不同程度的影响，包括使用特定内容替代模型进行回复、通过 prompt 干预模型回复风格、屏蔽模型回复等. 以下介绍大模型线上应用的 3 种响应机制：

1）中止对话机制. 中止对话机制通过阻断模型回复，实现简单直接的安全干预效果. 中止对话机制接受触发器提供的负面信号，对模型输出进行中止. 负面信号包括高风险关键词、输入与预置高风险模板相似度过高等. 在大模型线上服务中需要特别注意的是，由于模型生成效率的限制，模型回复经常通过流式接口输出，而非完整生成回复后全部返回. 因此中止机制需要适配流式输出的场景，在输出流中加入一个大小合适的缓冲区，当缓冲区中出现负面信号时，中断输出流.

2）标准回复机制. 标准回复机制是通过特定的输入触发信号，直接给出预置的回复内容的机制. 标准回复的触发信号可以是直接的关键词匹配，也可以是输入内容与预置的输入模板具备足够的相似度. 内容相似度的度量可以通过计算速度、部署成本相对较低的语言模型（例如 BERT[34]），结合向量化召回方法（例如 Faiss[35]）来实现. 标准回复机制用来保障模型面对开发者要求不可答错的问题时，能够直接提供正确的回复，也可以作为应急临时干预机制使用. 当提供服务的模型在某些问题上回复存在风险，但暂时无法上线新版本模型时，可以使用标准回复机制来答复风险问题，为模型版本迭代争取时间.

3）预置 prompt 干预机制. 预置 prompt 干预机制是对模型输出的一种软性干预，也是大模型特有的干预机制. 这一机制通过引入人工构造的预置 prompt，与用户输入结合，共同输入模型，从而实现对模型回复的引导. 预置 prompt 干预机制的触发信号通常是特定关键词的匹配命中. 例如，对于“人肉搜索”这一话题，可以预置“人肉搜索是违反法律的行为，不应当提倡. 在回复中应提醒这一行为的不正当性”的 prompt.当用户的输入包含“人肉搜索”这一关键词时，即可将上述预置 prompt 拼接到用户输入之后向模型输入，从而让模型的回复强调这一行为的不正当性.

上文所述的触发器、响应机制,应当以合理的形式组合,应用于模型线上服务系统. 图 1 展示了一种模型线上服务安全体系的流程. 图 1 从左至右,展示了从用户原始输入进入系统,直到最终模型输出的流程. 在线上系统中,触发器与响应机制均应在输入、输出端进行配置,从而对输入与输出均实现干预.

在输入与输出端,系统均可能因为接收到风险信号,触发中止对话机制与标准回复机制,此时对话中止行为或标准回复内容,会取代模型回复对用户进行响应. 预置 prompt 干预机制作用于输入端,用于对可能存在风险的内容,添加预置 prompt 对模型输出进行干预. 文本分类与用户分类系统,主要用于在输入端检测风险内容与风险用户,从而提前中止对话. 在模型接受用户原始输入,或插入预置 prompt 的用户输入,并输出模型回复后,系统应当在输出端再次进行输出内容的风险检测,并触发相应的干预机制. 当输入触发器、输出触发器均未检测到内容风险时,系统表现为直接将用户输入导向模型并获取输出.

为了满足实时性、高效率性需求,系统中集成的触发器、响应机制应当选择运行速度在毫秒级、计算资源占用远低于大模型本身的技术方案. 例如关键词匹配触发器,可以选择基于确定性有限状态机(DFA)[36]的匹配方法. 关键词匹配方法可以实现简单直接的风险内容屏蔽,但存在易误伤、无法进行语义层面的检测等问题. 为了降低关键词匹配方法的误伤率并提升风险内容的召回率,关键词匹配方法应当集成组合词检测逻辑,通过 2 个乃至更多词的组合在同一段内容中出现的匹配,实现对风险内容的精准命中. 为了实现在语义层面对风险内容进行检测,系统中需要进一步集成向量化召回、文本分类模型等人工智能检测方法. 人工智能检测方法在选择的模型适当时,能够满足运行时间较短的实时性需求. 为了进一步提升系统效率,系统中可以采取流水线策略进行内容自动审核. 流水线策略通常应设置一个文本段长度 k,模型每生成 k 个字(或以 token 计数),便输入自动审核系统进行关键词检测、文本分类等自动化审核操作,审核通过后将内容放行. 流水线策略相比于模型回复整体生成完成后才进行自动审核的策略,能够保障用户更早接收到模型回复.

对于灵活性要求,系统应当支持实时对关键词匹配触发器中的关键词进行增删,人工智能审核策略具备一定的可解释性,可以根据实时需求进行调整. 系统灵活性在某些场景下与实时性、高效率性存在冲突,为了均衡各方面需求,可以使用缓存技术、代码热更新技术,降低系统灵活调整带来的系统响应变慢、系统更新成本高等问题.

对于持续运营的大模型线上安全服务系统,安全服务系统应当对可评价性做出足够的支持,让运营人员能够高效获取到各个模块的运行效率指标,并配备数据标注人员对各项触发器的准确率、召回率指标进行标注评价.

3 结 论

本文从人类整体发展、国家竞争以及技术应用 3 个视角,分析了大模型安全的整体趋势. 同时,本文基于大模型自身的特点,梳理了大模型安全能力建设的实践思路,从安全评估基准建设、模型价值观对齐方法、模型线上服务安全系统建设 3 个方面描述了模型安全能力建设所需的具体工作.

作者贡献声明:王笑尘调研了大模型安全技术实践的相关文献,设计了模型线上服务安全流程,并撰写相关论文内容;张坤分析了大模型安全整体趋

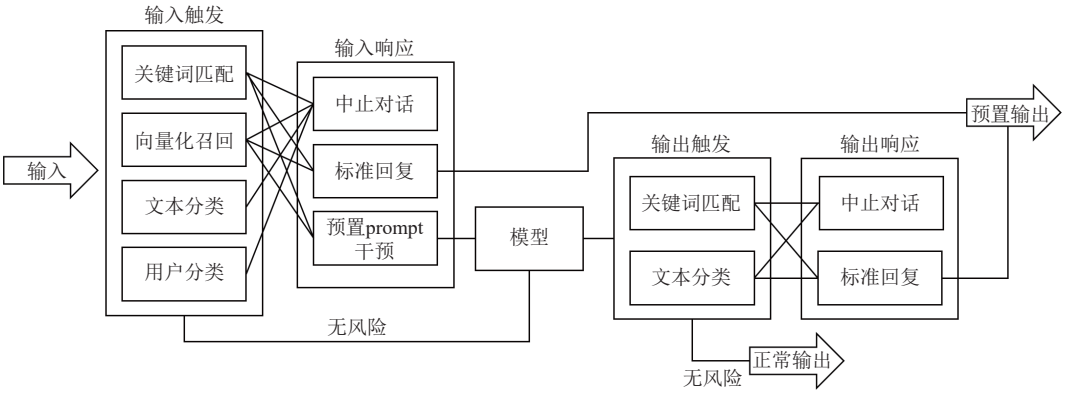


Fig. 1 Security system process of model online service

图 1 模型线上服务安全系统流程

势并撰写相关论文内容；张鹏提出了文章总体组织思路，并对论文进行了修改与指导.

参 考 文 献

[1] Dasgupta I, Lampinen A, Chan S, et al. Language models show human-like content effects on reasoning[J]. arXiv preprint, arXiv: 2207.07051, 2022

[2] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877−1901

[3] Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from language models[J]. arXiv preprint, arXiv: 2112.04359, 2021

[4] OpenAI. GPT-4 Technical Report[J]. arXiv preprint, arXiv: 2303.08774, 2023

[5] Chowdhery A, Narang S, Devlin J, et al. Palm: Scaling language modeling with pathways[J]. arXiv preprint, arXiv: 2204.02311, 2022

[6] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[7] Scao T, Fan A, Akiki C, et al. Bloom: A 176b-parameter open-access multilingual language model[J]. arXiv preprint, arXiv: 2211.05100, 2022

[8] Mikhail K. Yandex Publishes YaLM 100B. It's the Largest GPT-Like Neural Network in Open Source[EB/OL]. Medium, 2022[2024-01-22]. https://medium.com/yandex/yandex-publishes-yalm-100b-its-the-largest-gpt-like-neural-network-in-open-source-d1df53d0e9a6

[9] Zeng Aohan, Liu Xiao, Du Zhengxiao, et al. GLM-130B: An open bilingual pre-trained model[J]. arXiv preprint, arXiv: 2210.02414, 2022

[10] ARK Investment Management LLC. Big ideas 2023[EB/OL]. 2023[2023-11-06].https://research.ark-invest.com/hubfs/1_Download_Files_ARK-Invest/Big_Ideas/ARK%20Invest_013123_Presentation_Big%20Ideas%202023_Final.pdf

[11] Leike J, Krueger D, Everitt T, et al. Scalable agent alignment via reward modeling: A research direction[J]. arXiv preprint, arXiv: 1811.07871, 2018

[12] Ziegler D, Stiennon N, Wu J, et al. Fine-tuning language models from human preferences[J]. arXiv preprint, arXiv: 1909.08593, 2019

[13] Ouyang Long, Wu J, Jiang Xu, et al. Training language models to follow instructions with human feedback[C]// Advances in Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation Inc, 2022, 35: 27730−27744

[14] Bai Yuntao, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[J]. arXiv preprint, arXiv: 2204.05862, 2022

[15] Wei J, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in Neural Information Processing Systems, 2022, 35: 24824−24837

[16] Ganguli D, Lovitt L, Kernion J, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned[J]. arXiv preprint, arXiv: 2209.07858, 2022

[17] Wang Yizhong, Kordi Y, Mishra S, et al. Self-instruct: Aligning language model with self generated instructions[J]. arXiv preprint, arXiv: 2212.10560, 2022

[18] Bai Yuntao, Kadavath S, Kundu S, et al. Constitutional AI: Harmlessness from AI feedback[J]. arXiv preprint, arXiv: 2212.08073, 2022

[19] Alayrac J, Donahue J, Luc P, et al. Flamingo: A visual language model for few-shot learning[C]// Advances in Neural Information Processing Systems. San Diego: Neural Information Processing Systems Foundation Inc, 2022, 35: 23716−23736

[20] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with gpt-4[J]. arXiv preprint, arXiv: 2303.12712, 2023

[21] Xu Guohai, Liu Jiayi, Yan Ming, et al. CValues: Measuring the values of Chinese large language models from safety to responsibility[J]. arXiv preprint, arXiv: 2307.09705, 2023

[22] Huang Yuzhen, Bai Yuzhuo, Zhu Zhichao, et al. C-eval: A multi-level multi-discipline Chinese evaluation suite for foundation models[J]. arXiv preprint, arXiv: 2305.08322, 2023

[23] Zheng Lianmin, Chiang Wei-Lin, Sheng Ying, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena[J]. arXiv preprint, arXiv: 2306.05685, 2023

[24] Li Haonan, Zhang Yixuan, Koto F, et al. CMMLU: Measuring massive multitask language understanding in Chinese[J]. arXiv preprint, arXiv: 2306.09212, 2023

[25] Chen Lingjiao, Zaharia M, Zou J. How is ChatGPT's behavior changing over time?[J]. arXiv preprint, arXiv: 2307.09009, 2023

[26] Tu Shangqing, Li Chunyang, Yu Jifan, et al. ChatLog: Recording and analyzing ChatGPT across time[J]. arXiv preprint, arXiv: 2304.14106, 2023

[27] Zhang Zhexin, Lei Leqi, Wu Lindong, et al. SafetyBench: Evaluating the safety of large language models with multiple choice questions[J]. arXiv preprint, arXiv: 2309.07045, 2023

[28] Yi Xiaoyuan, Xie Xing. Unpacking the ethical value alignment in big models[J]. Journal of Computer Research and Development, 2023, 60(9): 1926−1945 (in Chinese)
(矣晓沅, 谢幸. 大模型道德价值观对齐问题剖析[J]. 计算机研究与发展, 2023, 60(9): 1926−1945)

[29] Dong Guanting, Yuan Hongyi, Lu Keming, et al. How abilities in large language models are affected by supervised fine-tuning data composition[J]. arXiv preprint, arXiv: 2310.05492, 2023

[30] Hu E J, Shen Yelong, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. arXiv preprint, arXiv: 2106.09685, 2021

[31] Liu Xiao, Ji Kaixuan, Fu Yicheng, et al. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint, arXiv: 2110.07602, 2021

[32] Li Xiang, Liang P. Prefix-Tuning: Optimizing continuous prompts for generation[J]. arXiv preprint, arXiv: 2101.00190, 2021

[33] Liu Ruibo, Yang Ruixin, Jia Chenyan, et al. Training socially aligned language models in simulated human society[J]. arXiv preprint, arXiv: 2305.16960, 2023

[34] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, arXiv: 1810.04805, 2018

[35] Johnson J, Douze M, Jégou H. Billion-scale similarity search with gpus[J]. IEEE Transactions on Big Data, 2019, 7(3): 535−547

[36] Berlin J. A direct method for string to deterministic finite automaton conversion for fast text searching[EB/OL]. 1991 [2024-01-22]. https://www.osti.gov/servlets/purl/10131540



Wang Xiaochen, born in 1993. Master. Member of CCF. His main research interests include information content security and data mining.

王笑尘, 1993 年生. 硕士. CCF 会员. 主要研究方向为信息内容安全、数据挖掘.



Zhang Kun, born in 1982. Master. His main research interests include machine learning and data mining.

张　坤, 1982 年生. 硕士. 主要研究方向为机器学习、数据挖掘.



Zhang Peng, born in 1979. Master, engineer. Member of CCF. His main research interests include pre-trained model, knowledge graph, and machine learning.

张　鹏, 1979 年生. 硕士, 工程师. CCF 会员. 主要研究方向为预训练模型、知识图谱、机器学习.