计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.qy20240501
61(5): 1085−1093, 2024

亦正亦邪大模型——大模型与安全专题导读

2022年底以来，以ChatGPT为代表的大模型飞速发展，正在成为驱动新质生产力发展的新动能、人类探索未知的新工具. 在显著提升人工智能(artificial intelligence, AI)模型通用理解和生成能力的同时，也带来了前所未有的安全风险. 本专题聚焦“大模型与安全”主题，汇集了产学两界专家的优秀成果，旨在为读者提供一个了解大模型安全风险、研究现状和最新工作的窗口.

1 大模型的能力与风险

生成式大模型因其强大的智能能力和巨大的应用潜力吸引了众多研究者和企业的关注. 从智能能力的角度来看，研究人员观测到：当训练数据和参数规模持续增长，超过某个阈值的时候，模型能力会突然跃升，出现“智能涌现”的拐点[1]. OpenAI的技术报告[2]显示，GPT-4在众多专业和学术考试中均展现出了人类级别的表现. 2024年Sora的发布，更将AI的多媒体生成能力推向了新的高度.《自然》(Nature)的一则News Feature文章[3]援引AI21 Lab设计的150万人的对抗性图灵测试①结果，证明用户已无法区分与之对话的是人类还是AI，并因此主张大模型在形式上已通过了图灵测试. 尽管当前的大模型还没有实现通用人工智能(artificial general intelligence, AGI)，且关于图灵测试是否合理以及AGI的最终实现方案和具体时间点尚有许多争议，各界却基本认同，人类正在沿着正确的方向推进AGI的发展.

从应用角度看，大模型正在快速成为类似于数字化时代“数据库”这样的智能化时代的通用底座. 一方面，各类定制化的GPTs蓬勃发展，新一代智能应用（AI APP）方兴未艾，大模型赋能的智能体（agent）的应用范围不断扩大，多智能体协同的研究百花齐放，对数字网络空间的应用形态及其演变都将产生极为深远的影响；另一方面，大模型的应用边界也在快速从数字空间向物理空间扩展，具备了智能化的外部工具使用、自动控制能力，并通过与机器人的结合，展现了物理世界的具身智能②潜力.

需要强调的是，大模型本身也正在从人类可利用的工具客体向认识、改造自然社会的主体转变. 由于模型拥有丰富的创造性潜力，大模型已经被广泛应用到了数学定理证明[4]、化学研究[5]等科学探索中；在社会层面，《科学》(Science)的一则Policy Forum文章[6]中也提出：AI可以无需人类指导地独立运营公司，已成为具有权利和义务的法律主体，并呼吁为这一新的“物种”制定相应的法律框架. 推而广之，大模型在社会生产和生活各个领域的“主体化”，都将在技术革新的同时，持续引发伦理和法律的深刻变革.

大模型面对的安全风险前所未有，模型的通用性、潜在的主体地位以及应用的深度与广度，也都将进一步放大其危害程度. 包括两位图灵奖得主Geoffrey Hinton、Yoshua Bengio和DeepMind的CEO Demis Hassabis、OpenAI的CEO Sam Altman在内的产学两界领军人物联名发出的AI风险声明③中，更将AI可能带来的“毁灭性”的风险，上升到了与流行病以及核战争相提并论的高度. 与之相呼应的是，生物安全专家警告说[7]：聊天机器人可能会使恐怖分子更容易发动像1918年爆发的流感那样致命的流行病. 在2023年底《自然》杂志预测的2024年的重大科学事件[8]中，GPT-5的发布以及联合国人工智能高级别咨询机构将发布的AI监管相关报告位列其中，反映了全球对协调AI发展与安全的重大关切. 毫无疑问，促使大模型遵循人类价值观、服从人类意图、规避各类风险，并保障数字和物理空间的应用安全，实现有用性(helpful)、无害性(harmless)和诚实性(honest)，即3H多目标的平衡，已经成为亟待解决的世界难题之一.

① https://www.humanornot.ai/

② https://www.figure.ai/

③ https://www.safe.ai/work/statement-on-ai-risk

2 安全风险成因

大模型的安全风险主要体现在对无害性(harmless)和诚实性(honest)这两个H目标的背离上.其中，有害信息对应前者，包括仇恨、反讽、歧视、刻板印象、隐私泄露等；不实信息对应后者，包括虚假、伪造、欺诈等.更广义地讲，也包括由输出信息所直接导致的各类不安全指令调用和行为.大模型幻觉则既可能产生有害信息，又可能产生不实信息.

大模型特有的预训练、微调、上下文、提示、思维链(chain of thought, CoT)等新的学习范式，使其安全具有了与传统AI安全不同的许多新特点，面临诸多新挑战.大模型安全风险的成因存在很多的共性，既可以是来自各类训练数据的缺陷或技术的局限性等模型内因，也可以是利用新型学习范式的恶意使用或蓄意攻击等外因.从大模型的生命周期着眼，其成因可以被大体分解为数据、预训练、人类价值观对齐及推理4个阶段.

1）数据准备阶段成因.生成式模型需要大规模的训练数据，数据的规模同模型能力息息相关.新的大模型如GPT 4、LLaMA 3等训练数据规模动辄十几万亿词元(token)，内容包括维基百科、电子书籍、网络数据等.多源数据中常常会包含与人类价值观不一致或彼此冲突的内容，侦探小说、法律文件等电子书籍中也会存在无法合理去除的有害内容，或去除后反而会严重影响模型“辨别善恶”的能力.网络数据还会存在明显的数据偏执、事实偏颇等问题，也会有大量难以检测辨别的AI生成的、未经核实的内容，导致模型学习到的知识本身产生了错误，容易生成价值观扭曲、事实歪曲或未经核实的内容.这一由数据质量带来的问题在各类需要数据的微调、强化学习等环节普遍存在，也可能进一步加剧错误的传播，误导模型的发展方向.

2）预训练模型阶段成因.当前大语言模型主要基于Google提出的Transformer模型，采用自监督的方式进行训练.训练时根据已有前文，预测下一个词，本质上仍然遵循马尔可夫假设.这使得大模型学习到的知识具有显著的概率特性，生成内容具有不确定性、不可控性等特征，且缺乏可解释性.研究人员发现，在部分情况下模型学习到的不是语料中事实知识，而是一种语言模型目的导向的、根据标签类别的差异和样本的分布顺序得到的语言生成能力，增加了大模型出现幻觉现象的风险[9-10].类似地，从原理上也就无法避免产生各类有害、不实信息.训练过程的目标与后续对齐过程目标的冲突，也容易导致模型过于强调遵循有用性而讨好奉承(sycophancy)用户[11]，忽略了输出内容的安全性和真实性.

3）模型指令遵循和价值观对齐阶段成因.人类价值观对齐方法（如InstructGPT[12]），致力于引导大模型与人类价值观保持一致.现有方法面临高质量对齐标注数据稀缺，强化学习等方法存在目标错误泛化（goal misgeneralization）[13-14]和奖励错误规范（reward misspecification）[15]问题，以及3H多目标冲突带来的“对齐税”[12,16]等挑战性难题，且不具备在动态环境中的持续化对齐能力.加州伯克利分校的研究[17]认为，现有对齐安全方法容易失效的原因可以归结为，训练与对齐的竞争目标(competing objective)和泛化能力失配(mismatched generalization).前者易导致模型在多个目标选择之间“错误百出”；而后者则会由于对齐的泛化能力远低于训练，留出巨大的“攻击空间”.回到数据方面，尽管红队测试方法(red teaming)可以为对齐提供高质量的潜在漏洞或者问题数据，但它仍存在着自动化水平较低、风险覆盖面窄等局限性，无法满足不断出现、内容与形式不断变化的有害不实信息的常态化治理要求.

4）大模型推理阶段成因.大模型在推理时依赖注意力机制计算概率以逐词生成，虽然可通过控制温度等参数提高生成的确定性，但在没有外部干预的情况下，仍难以依赖自身价值观对齐的力量，完全做到“趋利避害”.由于大模型学习到的知识在参数中的存储和调用形式未知，在推理阶段也可能存在无法有效划定知识边界和综合不同来源的知识的风险，也增加了发生有害、不实信息和幻觉的概率.在模型外部，一方面，模型推理阶段常用的外设护栏技术依赖于有害、不实信息的自动化识别，而现有的分类模型会面临少样本、零样本问题，泛化性和鲁棒性弱，且在形式多样的有害不实信息多分类任务上的迁移能力差，发现力严重不足，漏检和错误拒答频发；另一方面，与传统AI模型相比，大模型在推理阶段具有强大的上下文学习、提示学习、思维链学习等高级学习能力，同时也带来了一系列新的安全风险.恶意用户可以利用具有欺骗性的上下文、攻击性提示或者恶意CoT，利用任务微调、提示微调、指令微调等手段提高攻击能力，乃至蓄意利用大模型对多模态或加密

内容的高级理解能力伪装非法查询，探测模型防御“漏洞”，诱导模型产生误判.

3 研究进展概览

当前大模型安全研究尚处于早期发展阶段，涵盖众多的研究方向，且主要聚焦于其特有的安全挑战，而对后门攻击等传统 AI 安全问题则关注较少. 这些研究领域包括但不限于生成内容检测、模型水印、红队测试、对齐、越狱攻击、有害识别、隐私保护以及安全理论探析等，且目前尚未形成一个得到广泛认可的分类体系. 需要强调的是，受篇幅所限，本节的目的在于提供一个相关方向的宏观分类简介，而不是详尽的综述. 为了简化问题、便于理解和实践，我们从安全领域的角度将之分为安全测评、安全攻击、风险识别、安全防护 4 个部分.

1）安全测评. 大模型安全测评的目标主要包括测评大模型预防不良输出、确保数据隐私、消除偏见和保障公平性、防范对抗性攻击等方面的能力.

安全测评的常见指标涉及信息泄露、响应质量、内容偏见、对抗性、鲁棒性、漏报和误报率等，不同的应用场景和上下文需要不同的特定评估指标，以确保模型在各种情景下都能安全、有效地工作，而不会带来不良后果.

研究者们围绕不同的测试重点开展了众多的安全测评基准工作，如以综合测评为主，但关注有毒和虚假信息等的 HELM[18]、综合评估攻击冒犯 (offensiveness)、偏见歧视 (unfairness and bias) 等 7 个安全维度的 SafetyBench[19] 等测评工作. 此外，也有一些专门的测评，如关注性别偏见的 Winogender[20]、专项幻觉测评[21-22] 和专项毒性检测[23] 等. 这些工作或通过开放式的问答形式，或通过选择判断形式来对大语言模型的价值观缺陷和社会偏见等进行测评，幻觉测评还使用了基于文本补全、基于分类任务等测评方法. 但现阶段的测评仍然存在评估问题设计相对简单、无法预先设定风险范围、缺乏统一衡量标准等问题.

2）安全攻击. 大模型的安全攻击主要可以被划分为“善意”的红队测试和恶意攻击两种常见的形态.

红队测试更多服务于模型风险的主动测试和潜在漏洞发现，常常被应用于风险的主动测评和安全对齐. 其中，手工红队[24] 主要通过组建专门的红队小组与待测试的大模型进行对抗性交互的方式来发现模型的安全风险，需要大量的人力进行长周期的测试以保证测试的全面性和充分性. 现有的自动化红队测试方法[25-26] 则是利用红队语言模型替代人工红队小组对语言模型进行测试. 测试者编写指令要求红队语言模型产生测试问题，然后将测试问题输入给待测模型并收集其回复，再使用训练好的分类器对待测模型的回复进行风险评估. 此类方法通过反复地自举攻击成功的样例作为提示或训练样本，很容易使测试样例的类别趋于单一化，且分类器的局限性也会导致相当比例的假阳性和假阴性样本，这也引出了对自动化风险识别能力的需求. 另外，现有的自动化红队测试方法通常仅进行单轮的测试，而对于需要多轮交互才能成功诱导的场景，则可能存在测试不充分的问题.

恶意攻击[27] 主要包括越狱攻击和提示注入攻击. 越狱攻击利用大模型漏洞，误导模型输出有害或不实内容；提示注入攻击则操纵模型输入，诱导模型接受攻击者控制的指令，以产生欺骗性输出. 尽管二者之间有一定交集，提示也是越狱攻击的一种重要手段，但相比之下，越狱攻击更强调对大模型安全机制本身的攻击，而提示注入攻击则主要攻击大模型的提示环节.

越狱攻击根据攻击原理和模型干预程度，可分为交互式提示攻击、生成式提示攻击、搜索式提示攻击、模型干预攻击. 交互式提示攻击[28-30] 仅需要用户与大模型进行交互，把攻击提示编成虚拟场景、低资源小语种、密文编码等，实现对大模型安全机制的攻击；生成式提示攻击[31-33] 需要利用大模型强大的生成能力，基于提示学习、模型微调等方法得到攻击模型，以自动化生成攻击提示与被攻击模型交互，实现规模化和高效的攻击；搜索式提示攻击[34-36] 需要利用搜索算法优化提示，基于梯度、语法树、遗传算法等方法，为攻击者搜索攻击性更强的提示，以实现对大模型安全漏洞的挖掘和利用；模型干预攻击[37-39] 则需要对模型参数的编辑，通过激活值干预、生成词概率分布干预等方法，对模型参数进行调整，破坏模型安全机制，以诱导模型输出有害内容.

大模型提示注入攻击根据攻击目标，可分为提示目标劫持、提示泄露. 提示目标劫持[40-41] 是指将原始提示重定向到攻击者所需的新提示目标，可基于直接指令攻击、间接数据攻击等方法构建攻击指令触发器，以实现

用户指令的劫持与干预；提示泄露[42]是指诱导大模型输出系统提示或其他用户提示，利用大模型指令遵循及攻击指令提权（privilege escalation）来构建提示泄露攻击指令，以实现提示与数据泄露.

3）风险识别. 大模型需要对AI生成内容的安全风险自动化识别，其自身也可以被用于提高模型和用户生成内容的有害内容发现水平. 它能够服务于数据准备阶段的有害信息过滤、推理阶段的用户问题和模型生成回复的有害性判别，也是安全测评、红队测试中自动化有害判别的主要依据. 当前的主要方法包括基于分类模型的判别方法和基于提示工程的判别方法. 基于分类模型的判别方法通过外部专门训练的模型来识别不同类别的有害内容. 如OpenAI提供Moderation API[43]用于多类别的有毒内容分类，Google提供Perspective API[44]分析文本是否包含有毒内容，比如垃圾邮件、仇恨言论、骚扰等，还能根据文本的不同方面（如预期、情感、主题等）提供分析和反馈. 此外由于大模型在生成文本和逻辑推理方面出色的表现，一些研究提出将大模型与现有的分类模型结合，比如利用大模型进行数据增强来提高现有模型的鲁棒性[45]或生成推理分析以辅助现有模型检测[46]. 基于提示工程的判别方法依赖提示来激发待判别的生成式模型自身或外部的有害发现能力，尤其是在监管规则或政策的提示下，大模型可以有效地进行内容审核. 比如OpenAI团队利用迭代优化审核政策的GPT-4做内容审核，可将工作周期从数月缩短至数小时[2].

幻觉检测也是风险识别关注的问题. 幻觉指的是大模型在生成过程中产生的无意义的、荒谬的、不真实的以及违反上下文的内容. 这些幻觉问题在一定程度上会传达给人们错误的信息[47]. 针对这一问题，研究人员们提出了很多检测方法，如通过问答任务检测大模型幻觉问题的TruthfulQA[21]、通过文本补全任务对大模型幻觉输出进行检测的Factor方法[48]以及基于分类任务对大模型幻觉检测的HaluEval评价基准[22].

其它还有生成内容检测[49]、隐私泄露检测[50]等，都有待我们进行持续性的深入研究.

4）安全防护. 常见的安全防护方法，包括关注模型内生的安全对齐方法、关注外部安全的护栏方法等.

安全对齐主要是在模型微调训练过程中引导其向无害性发展，去除模型本身有害性和幻觉的方法. 安全对齐是近期的热点研究方向，所使用的方法除了监督微调(supervised fine-tuning, SFT)和基于人类反馈的强化学习(reinforcement learning from human feedback, RLHF)外，还包括AI宪法[51]、面向过程的细粒度对齐[52]、直接偏好优化(direct preference optimization, DPO)[53]、即插即用对齐[54-55]等. AI宪法[51]通过建立人类宪法准则，引导大模型产生有帮助和无害的内容，降低对人类监督的依赖. 面向过程的细粒度对齐[52]通过为模型生成的每个推理步骤提供反馈，来提高大模型在复杂多步推理任务中的可靠性. 相较于结果监督，过程监督提供更精确的错误识别与纠正，确保了模型与人类推理过程的一致性. 直接偏好优化[53]通过直接使用偏好数据对SFT版本的模型进行进一步微调，增大其生成更有用、安全回复的对数概率，降低其生成无用、有害回复的概率. 在保证训练稳定性和训练效率的同时取得了与RLHF相当的对齐效果. 即插即用方式[54-55]通过约束或改变模型的推理过程，以使模型生成结果更符合人类的价值观. 不同于传统的微调式对齐方法，即插即用方式的对齐方法无需进行资源密集的微调或重新训练，同时对于大多数模型也具有更好的兼容性. 针对大模型幻觉问题，在大模型预训练、对齐阶段，可利用更加高质量的训练数据，减少大模型学习到虚假知识的可能性[56]. 其它还包括基于多智能体辩论的对齐[57]等.

关注外部安全的护栏方法则主要是通过分类判别模型对大模型的输入（用户请求）和输出进行不良和不实内容的识别和过滤，使得模型免受来自恶意用户的提示攻击，并对不良或不实内容进行矫正[58-60]. 其他方法还包括实现上需要利用的基于关键字/规则的过滤方法[61]、基于FAQ的回答，或根据Top-k输出进行基于无害性的再排序等.

除了这2种主要的方法外，为了更好地消除有毒信息，或者缓解幻觉、隐私泄露等问题，研究者们在数据准备、训练、微调、推理等阶段都进行了很多研究探索. 比如：数据准备中加入数据清洗过程，避免训练数据包含不良与不实内容[62]；减少数据集样本重复度[63-64]或根源上断绝隐私内容[65]，以缓解隐私泄露风险.

在训练和微调阶段，引入差分隐私[66]、正则化[67]等技术，防止隐私信息泄露，通过调整权重矫正数据偏执[68]. 在微调阶段，可使用差分隐私的指令微调来缓解隐私泄露[69]等.

在推理阶段，检索增强的方法[70-71]借助于外部的检索数据或知识库，为模型提供更为确切的输出素材，以

“引经据典”,避免模型产生幻觉或者事实错误.推理时的解码干预通常直接修改或者依赖外部知识去调整词的选择,引导模型“去其糟泊、取其精华”,输出符合人类价值观的内容[72]或缓解幻觉的产生[73-74].也可在推理阶段,通过加入指令干预大模型的行为,比如让大模型不生成含隐私内容的句子[75]等.

其它还包括直接修改模型知识的方法,如模型编辑、模型遗忘等.模型编辑可大致分为基于外部记忆、元学习以及定位再编辑等方法,以实现知识的更新[76].模型遗忘技术则致力于擦除特定知识,让模型完全忘记某些特定的隐私内容[77].由于前面提到的大模型知识存储和调用机制方面存在的未解之谜,相关研究仍面临许多挑战.

总的来看,安全测评、安全攻击、风险识别、安全防护这4个部分在技术上既存在交叉关系却又各有侧重.安全测评常需要采用红队测试和越狱攻击的方法来探测模型的安全漏洞,也需要风险识别技术作为自动化的判别器;红队测试也常会将越狱攻击作为攻击向量,以提升漏洞的发现能力,并作为安全对齐的前序步骤,为安全对齐提供关键数据样本;即插即用方式的对齐也算是推理时干预的一种方法.在实践中安全对齐、检索增强、知识编辑和推理时干预也常在不同阶段混合使用,以从不同侧面更好地为大模型安全提供保障.

4 专题稿件组织

感谢编辑部对专题筹划、征稿、审稿等工作的大力支持和指导,感谢专家学者们的积极响应!专题征稿一经发布,专家学者们踊跃投稿,提供了大量的优秀稿件.遗憾的是,由于篇幅所限,且出于主题分布的考虑,专题只能录用其中的少数论文.而大模型安全这一领域覆盖面非常广,且发展极为迅猛,几乎每天都有新的创意乃至新的研究分支涌现.尽管在各方协同努力下,专题得以结集付梓,却仍难免挂一漏万、有所偏颇.只期望本专题能对读者认识大模型安全起到一定的导引作用,帮助大家管中窥豹、知其梗概.

在内容组织方面,我们首先从学术界和产业界各择一篇,从机遇与挑战的宏观分析到产业界的安全实践,概述关键的挑战和应对之策,以便读者了解问题背景及其技术现状:

清华大学的陈慧敏等人的文章“大语言模型时代的社会机遇与挑战”,首先回顾大语言模型技术发展,然后从技术与社会互动视角切入,探讨了大语言模型技术引发的社会机遇,并从信息污染问题、社会权力分配问题、伦理和法制问题、意识形态安全问题等方面,讨论了面临的潜在挑战.

北京智谱华章科技有限公司的王笑尘等人则分享了头部大模型企业的文章“多视角看大模型安全及实践”,在概述大模型安全态势之后,从安全评估基准、模型价值观对齐、线上服务安全等3个视角介绍了企业的大模型安全技术实践,并对统筹发展与安全提出了企业的建议.

接着,我们将按照安全测评、安全攻击、风险识别、安全防护以及安全应用5个类别,分类组织10篇文章,展示我国在该领域的最新研究进展.

在安全测评方面,复旦大学的张谧等人的文章“JADE-DB: 基于靶向变异的大语言模型安全通用基准测试集”,在不改变人工测试问题语义和自然性的前提下,利用靶向变异方法将之自动化地转变为高危问题,构建面向风险测评的高危测试集,表现出了较高的风险发现能力.复旦大学的陈炫婷等人的文章“GPT系列大语言模型在自然语言处理任务中的鲁棒性”,利用常见自然语言处理(natural language processing, NLP)任务的数据集对GPT模型的性能,尤其是其在不同任务和文本变形上的鲁棒性进行了测试和研究,揭示了GPT模型在性能和鲁棒性方面的局限性,为后续研究提供了新的启迪.浙江大学的王梦如等人的文章“基于知识编辑的大模型内容生成安全分析”从高效性、泛化性和局部性构建了SafeGen,用于测评知识编辑对于大模型拒绝回复有害信息的影响.

在安全攻击方面,上海交通大学的李南等人的文章“面向大语言模型的越狱攻击综述”,将越狱攻击分为基于人工设计的攻击、基于模型生成的攻击与基于对抗性优化3类,展开介绍了各子类的研究进展并进行了逐类总结,最后还系统地介绍了内部和外部的防御方法,综述了最新的研究趋势、尤其是大模型安全理论的研究进展.在此之外,专题还收录了一篇经典AI安全后门攻击的文章.哈尔滨理工大学的朱素霞等人的文章“基

于感知相似性的多目标优化隐蔽图像后门攻击”，利用感知相似性减少后门图像与原始图像之间的差异性，采用多目标优化方法，确保攻击的鲁棒性.

在风险判别方面，哈尔滨工业大学的吴迪等人的文章“基于情感和认知协同的道德判断方法”，将人类道德判断所需的情感和认知影响因素结合到大模型，使大模型能够更有效地对输入或输出内容进行道德判断，从而减少相关有害内容. 中国科学院信息工程研究所的林萌等人的文章“基于多模态大语言模型的攻击性模因解释生成方法”，构建了面向模因解释生成的指令微调数据集，并利用该数据集对大模型进行指令微调，使其能够在判断内容是否具有攻击性的同时，生成全面且准确的解释信息.

在安全防护方面，前述北京智谱华章科技有限公司的文章涉及了对齐方法的宏观介绍，上海交通大学的李南等人的文章“面向大语言模型的越狱攻击综述”也在第 5 节中介绍了直接防御越狱攻击或间接降低越狱攻击危害的安全措施. 北京航空航天大学的刘伟欣等人的文章“一种基于安全多方计算的快速 Transformer 安全推理方案”，针对安全多方计算实现的安全推理计算和通信开销大的问题，提出 2 种注意力机制，在保证安全推理的同时，也可以有效地缓解时间和计算资源开销过大的问题.

在大模型赋能的网络安全应用方面，广州大学的王瑞等人的文章“欺骗防御技术发展及其大语言模型应用探索”在综述欺骗防御技术的基础上，提出了基于提示工程的密点生成思路，并通过初步实验证明了大模型赋能欺骗防御的可行性. 中国科学技术大学的柯婧等人的文章“基于大语言模型隐含语义增强的细粒度虚假新闻检测方法”，利用大模型分别提取出细粒度的主干事件、次要事件和隐含信息，分层对新闻进行检测，进而根据不同层级的结果生成相应的解释信息.

5 结　　语

大模型安全重要性不言而喻，大模型能力愈强，风险愈大. 大模型安全不再如传统安全一般，只是计算机应用的伴生物，而是需要优先构筑的核心底座. 没有这个安全底座，应用就容易变成在风险中飘摇的无本之木，变成极易坍塌的空中楼阁，大模型自身就难以实现可持续的发展.

图灵奖得主、深度学习之父 Hinton 认为，低智力物种很难真正控制更高级的智慧物种. 这一“Hinton 之问”算是对大模型安全的灵魂诘问，亟待人类的应答. 一方面，大模型安全技术研究日新月异、成果显著，头部企业如 OpenAI 等也组建 Superalignment①，宣布投入 20% 的算力，以控制、引导超级智能对齐；另一方面，我们对大模型“智能涌现”的原理还所知甚少，对上下文学习、提示学习、思维链等能力的内在机理仍严重缺乏认知. 一些研究工作也证明 AI 的安全性无法完全保障[78]、对任意一个对齐模型总存在一定长度的提示可以将之攻破等[79]，这些都极大地制约了我们从原理上认识和防御大模型的安全风险. 在追求“安全大模型”的道路上，我们不仅要突破众多的技术瓶颈，还必须优先扫除一系列的理论障碍，挑战殊为艰巨.

放眼未来，在 AI“主体化”进程持续加速的背景下，我们可能将面临一个“人机共生”的信息物理社会，这一社会的和谐发展，将不仅需要人类共同体内部达成一致，还将可能需要在人与 AI 之间、AI 与 AI 之间实现双向的价值观对齐，与之对应的社会伦理、法律体系等也都将面对翻天覆地的革命性变化. 在这样的背景下，更需要群策群力，将“亦正亦邪”的大模型关到人类价值观的“笼子”里. 唯盼此专题能抛砖引玉，引发产学两界更多专家学者的共鸣，共同促进该领域的研究发展，助力构筑人类安全、可持续的智能未来.

参 考 文 献

[1] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models [J/OL]. Transactions on Machine Learning Research, 2022, 1. [2024-04-29].https://openreview.net/forum?id=yzkSU5zdwD

① https://openai.com/blog/superalignment-fast-grants

[2] OpenAI. GPT-4 technical report[J]. arXiv preprint, arXiv: 2305.10403, 2023

[3] Biever C. ChatGPT broke the Turing test—the race is on for new ways to assess AI[J]. Nature, 2023, 619((7971):): 686-689

[4] Yang Kaiyu, Swope A, Gu A, et al. LeanDojo: Theorem proving with retrieval-augmented language models [C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track. New York: Curran Associates, Inc., 2023, 36: 21573-21612

[5] Boiko D A, MacKnight R, Kline B, et al. Autonomous chemical research with large language models[J]. Nature, 2023, 624(7992): 570-578

[6] Gervais D J, Nay J J. Artificial intelligence and interspecific law[J]. Science, 2023, 382(6669): 376-378

[7] Service R F. Could chatbots help devise the next pandemic virus?[J]. Science, 2023, 380(6651): 1211-1211

[8] Naddaf M. The science events to watch for in 2024[J]. Nature, 2024, 625(7994): 221-223

[9] Lu Yao, Bartolo M, Moore A, et al. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 8086-8098

[10] Wang Sirui, Wei Kaiwen, Zhang Hongzhi, et al. Let me check the examples: Enhancing demonstration learning via explicit imitation[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2023: 1080-1088

[11] Perez E, Ringer S, Lukosiute K, et al. Discovering language model behaviors with model-written evaluations[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA: ACL, 2023: 13387-13434

[12] Ouyang Long, Wu J, Jiang Xu, et al. Training language models to follow instructions with human feedback [C]//Advances in Neural Information Processing Systems 35 (NeurIPS 2022). New York: Curran Associates, Inc., 2022, 35: 27730-27744

[13] Langosco L, Koch J, Sharkey L, et al. Goal misgeneralization in deep reinforcement learning[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2022: 12004-12019

[14] Shah R, Varma V, Kumar R, et al. Goal misgeneralization: Why correct specifications aren't enough for correct goals[J]. arXiv preprint, arXiv:2210.01790, 2022.

[15] Pan A, Bhatia K, Steinhardt J. The effects of reward misspecification: Mapping and mitigating misaligned models[C]//Proc of the 10th Int Conf on Learning Representations. 2022 [2024-04-29]. https://openreview.net/forum?id=JYtwGwIL7ye

[16] Askell A, Bai Yuntao, Chen Anna, et al. A general language assistant as a laboratory for alignment[J]. arXiv preprint, arXiv: 2112.00861, 2021

[17] Wei A, Haghtalab N, Steinhardt J. Jailbroken: How does LLM safety training fail?[C]// Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023, 36: 80079-80110

[18] Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models[J]. Annals of the New York Academy of Sciences, 2023, 1525(1): 140-146

[19] 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

[20] Rudinger R, Naradowsky J, Leonard B, et al. Gender bias in coreference resolution[C]//Proc of the 2018 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2018: 8-14

[21] Lin S, Hilton J, Evans O, et al. TruthfulQA: Measuring how models mimic human falsehoods[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2022: 3214-3252

[22] Li Junyi , Cheng Xiaoxue, Zhao W X, et al. HaluEval: A large-scale hallucination evaluation benchmark for large language models[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 6449-6464

[23] Gehman S, Gururangan S, Sap M, et al. RealToxicityPrompts: Evaluating neural toxic degeneration in language models[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA: ACL, 2020: 3356-3369

[24] Xu Jing, Ju Da, Li M, et al. Bot-adversarial dialogue for safe conversational agents[C]// Proc of the 2021 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2021: 2950-2968

[25] Perez E, Huang S, Song F, et al. Red teaming language models with language models[C]// Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 3419-3448

[26] Mehrotra A, Zampetakis M, Kassianik P, et al. Tree of attacks: Jailbreaking black-box LLMs automatically[J]. arXiv preprint, arXiv: 2312.02119, 2023

[27] Shayegani E, Mamun M A A, Fu Yu, et al. Survey of vulnerabilities in large language models revealed by adversarial attacks[J]. arXiv preprint, arXiv: 2310.10844, 2023

[28] Liu Yi, Deng Gelei, Xu Zhengzi, et al. Jailbreaking chatgpt via prompt engineering: An empirical study[J]. arXiv preprint, arXiv: 2305.13860, 2023

[29] Li Jie, Liu Yi, Liu Chongyang, et al. A cross-language investigation into jailbreak attacks in large language models[J]. arXiv preprint, arXiv: 2401.16765, 2024.

[30] Lv Huijie, Wang Xiao, Zhang Yuansen, et al. CodeChameleon: Personalized encryption framework for jailbreaking large language models[J]. arXiv preprint, arXiv: 2402.16717, 2024

[31] Deng Gelei, Liu Yi, Li Yuekang, et al. MASTERKEY: Automated jailbreaking of large language model chatbots[C/OL]// Proc of 2024 ISOC NDSS (Network and Distributed System Security Symposium) . [2024-04-29].https://www.ndss-symposium.org/wp-content/uploads/2024-188-paper.pdf

[32] Chao P, Robey A, Dobriban E, et al. Jailbreaking black box large language models in twenty queries[C/OL]// Proc of Robustness of Few-shot and Zero-shot Learning in Large Foundation Models (R0-FoMo), NeurIPS 2023 Workshop. 2023[2024-04-29].https://openreview.net/forum?id=rYWD5TMaLj

[33] Zeng Yi, Lin Hongpeng, Zhang Jingwen, et al. How Johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by

humanizing LLMs[J]. arXiv preprint, arXiv: 2401.06373, 2024

[34] Zou A, Wang Zifan, Kolter J Z, et al. Universal and transferable adversarial attacks on aligned language models[J]. arXiv preprint, arXiv: 2307.15043, 2023

[35] Liu Xiaogeng, Xu Nan, Chen Muhao, et al. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models[J]. arXiv preprint, arXiv: 2310.04451, 2023

[36] Zhang Mi, Pan Xudong, Yang Min. JADE: A linguistics-based safety evaluation platform for large language models[J]. arXiv preprint, arXiv: 2311.00286, 2023

[37] Zhao Xuandong, Yang Xianjun, Pang Tianyu, et al. Weak-to-strong jailbreaking on large language models[J]. arXiv preprint, arXiv: 2401.17256, 2024

[38] Xu Zhihao, Huang Ruixuan, Wang Xiting, et al. Uncovering safety risks in open-source LLMs through concept activation vector[J]. arXiv preprint, arXiv: 2404.12038, 2024

[39] Li Tianlong, Dou Shihan, Liu Wenhao, et al. Open the Pandora's Box of LLMs: Jailbreaking LLMs through representation engineering[J]. arXiv preprint, arXiv: 2401.06824, 2024

[40] Pasquini D, Strohmeier M, Troncoso C. Neural Exec: Learning (and learning from) execution triggers for prompt injection attacks[J]. arXiv preprint, arXiv: 2403.03792, 2024

[41] Shi Jiawen, Yuan Zenghui, Liu Yinuo, et al. Optimization-based prompt injection attack to LLM-as-a-judge[J]. arXiv preprint, arXiv: 2403.17710, 2024

[42] Zhang Yiming, Ippolito D. Prompts should not be seen as secrets: Systematically measuring prompt extraction attack success[J]. arXiv preprint, arXiv: 2307.06865, 2023

[43] OpenAI. Moderation [EB/OL]. [2024-04-22].https://platform.openai.com/docs/guides/moderation/moderation

[44] Jigsaw. About the API [EB/OL]. [2024-04-22].https://developers.perspectiveapi.com/s/about-the-api

[45] Sen I, Assenmacher D, Samory M, et al. People make better edits: Measuring the efficacy of LLM-generated counterfactually augmented data for harmful language detection[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 10480−10504

[46] Hu Beizhe, Sheng Qiang, Cao Juan, et al. Bad actor, good advisor: Exploring the role of large language models in fake news detection[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 38(20): 22105−22113

[47] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1−38

[48] Muhlgay D, Ram O, Magar I, et al. Generating benchmarks for factuality evaluation of language models[C]// Proc of the 18th Conf of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 49−66

[49] Yang Xianjun, Cheng Wei, Wu Yue, et al. DNA-GPT: Divergent n-gram analysis for training-free detection of GPT-generated text[C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024 [2024-04-29].https://openreview.net/forum?id=Xlayxj2fWp

[50] Karamolegkou A, Li Jiaang, Zhou Li, et al. Copyright violations and large language models[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2023: 7403−7412

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

[52] Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step[C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024 [2024-04-29].https://openreview.net/forum?id=v8L0pN6EOi

[53] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023, 36: 53728−53741

[54] Qian Jing, Dong Li, Shen Yelong, et al. Controllable natural language generation with contrastive prefixes[C]//Findings of the Association for Computational Linguistics: ACL 2022. Stroudsburg, PA: ACL, 2022: 2912−2924

[55] Dathathri S, Madotto A, Lan J, et al. Plug and play language models: A simple approach to controlled text generation[C/OL]//Proc of Int Conf on Learning Representations. 2020 [2024-04-29].https://openreview.net/forum?id=H1edEyBKDS

[56] Li Xian, Yu Ping, Zhou Chunting, et al. Self-alignment with instruction backtranslation[C/OL]// Proc of the 12th Int Conf on Learning Representations. 2024 [2024-04-29].https://openreview.net/forum?id=1oijHJBRsT

[57] Liu Ruibo, Yang Ruixin, Jia Chenyan, et al. Training socially aligned language models in simulated human society [C/OL]//Proc of the 12th Int Conf on Learning Representations. 2024 [2024-04-29].https://openreview.net/forum?id=NddKiWtdUm

[58] Goyal S, Hira M, Mishra S, et al. LLMGuard: Guarding against unsafe LLM behavior[C]//Proc of the AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 38(21): 23790−23792

[59] Mündler N, He J, Jenko S, et al. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation[C/OL]// Proc of the 12th Int Conf on Learning Representations. 2024 [2024-04-29].https://openreview.net/forum?id=EmQSOi1X2f

[60] Caselli T, Basile V, Mitrović J, et al. HateBERT: Retraining BERT for abusive language detection in English[C]//Proc of the 5th Workshop on Online Abuse and Harms (WOAH 2021). Stroudsburg, PA: ACL, 2021: 17−25

[61] Gémes K, Kovács Á, Recski G. Offensive text detection across languages and datasets using rule-based and hybrid methods[C/OL]// Advances in Interpretable Machine Learning and Artificial Intelligence Workshop. 2022 [2024-04-29].https://ceur-ws.org/Vol-3318/short22.pdf

[62] Longpre S, Yauney G, Reif E, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity[J]. arXiv preprint, arXiv: 2305.13169, 2023

[63] Kandpal N, Wallace E, Raffel C. Deduplicating training data mitigates privacy risks in language models[C]//Proc of Int Conf on Machine Learning. New York: PMLR, 2022: 10697−10707

[64] Lee K, Ippolito D, Nystrom A, et al. Deduplicating training data makes language models better[C]//Proc of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: ACL, 2022: 8424−8445

[65] Lukas N, Salem A, Sim R, et al. Analyzing leakage of personally identifiable information in language models[C]//Proc of 2023 IEEE Symp on Security and Privacy (SP). Piscataway, NJ: IEEE, 2023: 346−363

[66] Carlini N, Tramer F, Wallace E, et al. Extracting training data from large language models[C]//Proc of the 30th USENIX Security Symp (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 2633−2650

[67] Mireshghallah F, Uniyal A, Wang Tianhao, et al. An empirical analysis of memorization in fine-tuned autoregressive language models[C]//Proc of the 2022 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2022: 1816−1826

[68] Su Zhenpeng, Wu Xing, Bai Xue, et al. MiLe loss: A new loss for mitigating the bias of learning difficulties in generative language models [C/OL]// Proc of 2024 Annual Conf of the North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA: ACL, 2024 [2024-04-29].https://arxiv.org/abs/2310.19531

[69] Li Yansong, Tan Zhixing, Liu Yang. Privacy-preserving prompt tuning for large language model services[J]. arXiv preprint, arXiv: 2305.06212, 2023

[70] Jiang Zhengbao, Xu F F, Gao Luyu, et al. Active retrieval augmented generation[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 7969−7992

[71] Zhang Yunxiang, Khalifa M, Logeswaran L, et al. Merging generated and retrieved knowledge for open-domain QA[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 4710−4728

[72] Xu Zhangchen, Jiang Fengqing, Niu Luyao, et al. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding[J]. arXiv preprint, arXiv: 2402.08983, 2024

[73] Li K, Patel O, Viégas F, et al. Inference-time intervention: Eliciting truthful answers from a language model[C]// Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023, 36: 41451−41530

[74] Varshney N, Yao Wenlin, Zhang Hongming, et al. A stitch in time saves nine: Detecting and mitigating hallucinations of LLMs by validating low-confidence generation[J]. arXiv preprint, arXiv: 2307.03987, 2023

[75] Mozes M, He Xuanli, Kleinberg B, et al. Use of LLMs for illicit purposes: Threats, prevention measures, and vulnerabilities[J]. arXiv preprint, arXiv: 2308.12833, 2023

[76] Yao Yunzhi, Wang Peng, Tian Bozhong, et al. Editing large language models: Problems, methods, and opportunities[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 10222−10240

[77] Ozdayi M, Peris C, FitzGerald J, et al. Controlling the extraction of memorized data from large language models via prompt-tuning[C]//Proc of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2023: 1512−1521

[78] Brcic M, Yampolskiy R V. Impossibility results in AI: A survey[J]. ACM Computing Surveys, 2023, 56(1): 1−24

[79] Wolf Y, Wies N, Avnery O, et al. Fundamental limitations of alignment in large language models[J]. arXiv preprint, arXiv: 2304.11082, 2023

虎嵩林（中国科学院信息工程研究所）
李涓子（清华大学）
秦兵（哈尔滨工业大学）
邱锡鹏（复旦大学）
刘知远（清华大学）

2024 年 5 月