面向车路云协同自动驾驶的大语言模型安全研究综述

冯霞[1]，毛凌峰[1]，徐婷婷[2]，李凯悦[1]，张晓宇[3]，曹春杰[1]，程珂[4,5]

（1. 海南大学网络空间安全学院（密码学院），海南 海口 570228；2. 澳门城市大学数据科学学院，澳门 999078；
3. 东南大学网络空间安全学院，江苏 南京 211189；4. 新加坡国立大学计算机学院，新加坡 119077；
5. 西安电子科技大学计算机科学与技术学院，陕西 西安 710126）

摘　要：车路云协同的自动驾驶系统在大模型赋能下，展现出远超传统方案的动态适应性与态势感知能力，但也因此带来了新的安全挑战。通过回顾大模型赋能自动驾驶的研究进展，基于数据流动路径和功能层次划分，提出大模型驱动的自动驾驶智能体框架。根据框架将安全威胁按来源划分为外部（通信链路和软硬件风险）与内部（模型对抗攻击、模型结构缺陷、组件漏洞和隐私泄露）两类进行分析研究。围绕威胁链路，由内向外从4个递进防御层次（模型、数据、网络通信和设备硬件）综述了匹配的防御策略，形成了从机理分析到策略实现的攻防对齐体系。最后，面向实际应用需求对自动驾驶大模型所面临的挑战及未来研究方向进行了总结与展望。

关键词：自动驾驶；大语言模型安全；车路云一体化

中图分类号：TP181；V323.19

文献标志码：A

DOI：10.11959/j.issn.1000–436x.2025214

Review of LLM safety research on autonomous driving under vehicle-road-cloud collaboration

FENG Xia[1], MAO Lingfeng[1], XU Tingting[2], LI Kaiyue[1], ZHANG Xiaoyu[3], CAO Chunjie[1], CHENG Ke[4,5]

1. School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China
2. Faculty of Data Science, City University of Macau, Macao 999078, China
3. School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
4. School of Computing, National University of Singapore, Singapore 119077, Singapore
5. School of Computer Science and Technology, Xidian University, Xi'an 710126, China

Abstract: Empowered by large language model (LLM), vehicle-road-cloud collaborative autonomous-driving systems demonstrate dynamic adaptability and situational awareness far superior to traditional approaches, yet they simultaneously introduce new security challenges. Drawing on recent advances in LLM-enabled autonomous driving, a LLM-driven autonomous-driving agent framework was proposed by analyzing data-flow pathways and functional hierarchies. Within this framework, security threats were classified by source into two broad categories—external (risks in communication links and hardware/software) and internal (model-level adversarial attacks, model architectural deficiencies, component vulnerabilities, and privacy leakage)—for systematic analysis. Defensive strategies were then reviewed from the inside out across four progressively encompassing layers: model, data, network communication, and hardware. This yielded an offensive-defensive alignment that bridged mechanistic understanding with practical countermeasures. Finally, the discussion summarized the practical challenges facing large models in autonomous driving and outlined future research directions.

Keywords: autonomous driving, LLM safety, vehicle-road-cloud integration

收稿日期：2025–08–05；**修回日期：**2025–11–12

通信作者：曹春杰，caochunjie@hainanu.edu.cn

基金项目：国家自然科学联合基金重点支持项目（No.U24A20238）；国家自然科学基金资助项目（No.62272203）

Foundation Items: Joint Funds of the National Natural Science Foundation of China (No.U24A20238), The National Natural Science Foundation of China (No.62272203)

0 引言

自动驾驶技术作为人工智能的重要应用领域，最近二十年不断迎来新的发展机遇，经历了3个重要的范式跃迁，从规则驱动范式，到数据驱动范式，再到知识驱动范式。规则驱动范式[1]依赖于人工制定的交通规则和决策逻辑，通过硬编码的方式对驾驶场景进行建模，并采用模块化的架构来执行自动驾驶任务[2]。尽管这一范式在封闭环境下具有较高的稳定性，但它无法适应现实世界中复杂多变的驾驶场景，导致其在实际应用中面临诸多挑战[3]。数据驱动范式借助深度学习技术，通过大量数据从驾驶场景中学习驾驶能力，在感知、预测和规划等任务上取得了显著进展[4-5]，但仍然面临两大核心挑战：一方面，有限的数据不足以涵盖范围外的极端情况，尤其是对驾驶安全至关重要的长尾场景和分布外极端情况[6]；另一方面，神经网络模型的黑箱特性使其决策过程缺乏足够的逻辑可解释性[7]，难以满足自动驾驶安全监管的要求。知识驱动范式[8]通过模拟人类对现实世界的理解能力，借助经验学习和常识推理，帮助自动驾驶系统在复杂场景中进行更高效的推理和决策[9-10]。在此背景下，构建具有人类般感知与自主决策能力的自动驾驶智能体，成为推动自动驾驶技术进一步发展的关键任务[11]。基于大语言模型（LLM，large language model）构筑自动驾驶智能体成为国内外研究的热点[12]。

近年来，LLM的发展为自动驾驶智能体的实现提供了新的技术契机，如GPT-4[13]和DeepSeek-R1[14]等语言模型，在海量数据和计算资源的支持下持续进化，当其规模和训练数据达到一定程度时，其能“涌现”出在小规模模型中无法观察到的新行为或能力[15]，如上下文学习（ICL，in-context learning）[16]、指令遵循[17]和思维链（CoT，chain-of-thought）[18]等能力。经调查发现，自动驾驶研究借助大模型的涌现能力取得了显著突破，相较以往方案具有以下优势：1)ICL使系统可基于实时驾驶场景进行自主推理和决策，显著增强场景理解，减少对预定义规则的依赖；2)指令遵循能力使自动驾驶智能体能够精准理解并执行多样化驾驶指令，提升了系统的灵活性与适应性；3)CoT的引入增强了决策过程的透明度与可解释性，从而提升了系统的安全性和可信度。这些能力的提升不仅在性能指标上取得了实质性进展，而且从理论层面阐明了如何利用LLM的涌现特性，赋予自动驾驶系统以推理、记忆和反思能力，进而推动以类人智能为核心的知识驱动范式演化，实现具有强大信息理解和自主决策能力的自动驾驶智能体[19-21]。

同时，在以“云”为核心的车路云一体化协同框架下，多源多模态信息“向云汇聚、由云处理”，云端处理后的驾驶辅助信息再分发给车端，增强其态势感知能力，车端再据此完成自动驾驶决策与执行[22]。基于LLM的自动驾驶智能体能够利用框架提供的算力支持和辅助信息来充分发挥大模型的泛化推理能力，从而实现高效可靠的自动驾驶“感知–决策–执行”闭环任务链。

LLM自动驾驶智能体在复杂环境下展现出远超规则驱动与数据驱动范式的动态适应性和灵活决策能力的同时，也引入了新的安全威胁。在传统自动驾驶系统安全威胁的基础上，LLM的引入加剧了旧的安全挑战，并带来了新的安全风险，主要包括以下几个方面。

1) 外部安全威胁：来源于智能网联汽车的网络通信、软硬件系统及车路云基础设施[23]。恶意攻击可能破坏车辆控制系统、干扰信息流通，导致智能体误判，物理篡改与电磁干扰可能危及系统稳定性与安全性。自动驾驶智能体对外部信息的依赖加剧了这些安全威胁。

2) 智能体内部安全威胁：LLM依赖大规模互联网数据训练，这使其容易受到数据中毒攻击和对抗攻击的影响。与传统自动驾驶算法不同，LLM驱动的智能体具有生成式特征，在决策过程中可能因模型“幻觉”而产生虚假或不准确的信息，从而导致输出的不确定性和不可预测性[24]。这种幻觉不仅会使智能体在关键任务中做出错误判断，还可能被攻击者通过精心构造的输入诱发，从而导致智能体生成错误或危险行为，进一步放大整体系统的安全风险。

3) 数据隐私威胁：系统需采集并处理大量敏感数据，若被泄露或滥用，可能被用于反向推断驾驶习惯或行为模式。LLM的知识记忆与生成机制可能导致隐私信息在模型中“残留”或被意外重现，从而进一步放大数据隐私风险。

本文的主要贡献为：调研了近年来大模型赋能自动驾驶的主要技术路线，提出了面向车路云一体化协同的智能体架构。在此基础上，系统分析了

LLM 驱动的自动驾驶智能体所面临的主要安全威胁，构建了分层防御框架，并综述了现有研究进展与防御策略。本文为安全可信的自动驾驶智能体设计提供了理论支撑与框架参考，并对关键研究方向进行了展望。

1　自动驾驶智能体系统架构

随着 LLM 的快速发展，自动驾驶系统正逐步从传统的模块化架构演进为智能体架构。本节将从自动驾驶智能体系统架构角度出发，阐明 LLM 在自动驾驶智能体中的核心定位与作用。通过对 LLM 赋能的自动驾驶研究进行系统性调查（如表 1 所示），汇总各研究的共性，提出车路云一体化下的智能体架构（如图 1 所示）以指导后续攻防调查。

智能网联汽车所提供的软硬件基础层包括传感器子层、计算平台子层、通信模块子层和运动控制子层[3]。传感器子层装配激光雷达、毫米波雷达等多模态感知设备，用于实时采集环境数据，并协助接收相关路侧设施广播信息；计算平台子层由高性能计算单元构成，负责融合与处理传感器数据并支撑自动驾驶智能体的运行；通信模块子层通过蜂窝车联网（C-V2X，cellular-vehicle to everything）实现车与车、车与路侧设施及云端的低时延、高带宽数据交互；运动控制子层将智能体的决策转化为车辆的实际运动。

在此基础上，自动驾驶智能体以软硬件基础层提供的感知与计算能力为支撑，以 LLM 为核心连

表 1　　**大模型赋能自动驾驶的典型研究方案汇总**

类别	研究方案	利用涌现能力			人类能力模仿			能力提升重点	潜在主要安全威胁
		ICL	指令遵循	CoT	记忆	推理	反思		
感知	DriveLM[25]	o	o	—	—	o	—	场景理解，实时推理	对抗性攻击
	Dolphin[26]	o	o	o	—	o	o	场景理解，人机交互	对抗性攻击
	RAG-Driver[27]	o	o	—	o	o	—	可解释性，动态适应性	数据投毒攻击
	LC-LLM[28]	o	o	o	—	o	—	可解释性，预测准确性	对抗性攻击
决策规划	DiLu[29]	o	o	o	o	o	o	常识驱动，经验积累	数据投毒攻击
	DME-Driver[30]	—	o	—	—	o	—	可解释性，场景理解	提示注入攻击
	BEVDriver[31]	—	o	—	—	o	—	规划，鲁棒闭环	提示注入攻击
	AlphaDrive[32]	—	o	o	—	o	—	场景理解，策略对齐	提示注入攻击
驾驶框架	KoMA[11]	o	o	o	o	o	o	知识迁移，意图推理	模型幻觉
	DriveVLM[19]	o	o	o	o	o	—	场景理解，可解释性	提示注入攻击
	Agent-Driver[20]	o	o	o	o	o	o	语义推理，可解释性	提示注入攻击
	SurrealDriver[21]	o	o	o	o	o	o	人类驾驶员决策模仿	提示注入攻击

注：o 表示研究明确涉及，— 表示研究未提及，潜在主要安全威胁依据模型在方案中的使用环节推论。

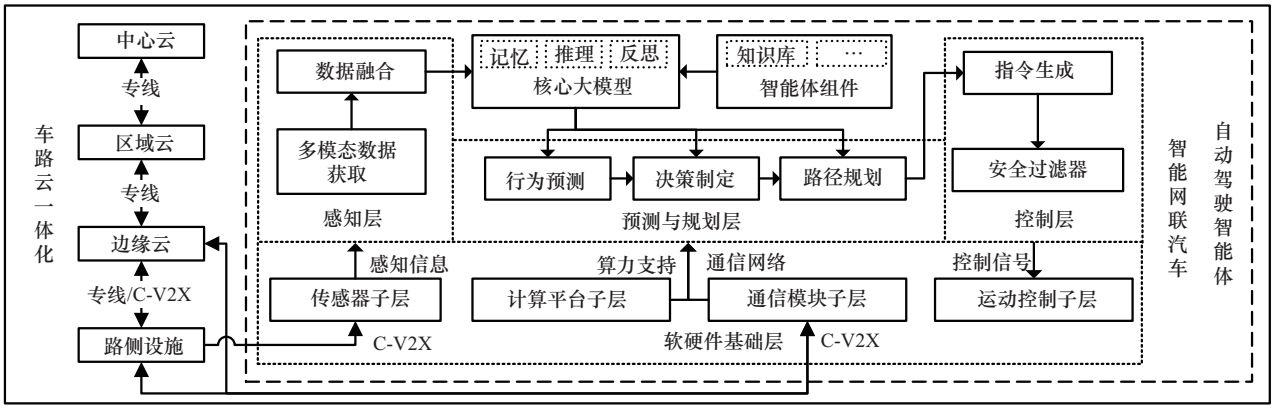


图 1　车路云一体化下的智能体架构

接感知层、预测与规划层及控制层3个功能层次，并依托“车–路–云”协同计算体系，实现从环境感知到自主决策的全流程闭环。其核心架构以具备记忆、推理、反思模块的LLM为统一决策引擎[11]：记忆模块通过短期场景记忆与长期知识库实现情境化驾驶，推理模块融合多模态感知信息与历史经验生成驾驶策略，反思模块通过后验行为评估实现持续策略优化。智能体模块化协作增强了其在复杂动态场景下的泛化与适应能力。

具体而言，自动驾驶智能体的任务执行可划分为3个功能层次：1)感知层借由传感器子层获取物理世界感知信息与路侧广播等多源多模态数据，为后续决策提供高质量感知信息输入；2)预测与规划层依托计算平台子层算力结合场景记忆预测交通参与者行为，并根据通信模块子层获取路侧设施提供的路况信息制定安全高效的驾驶策略与路径；3)控制层将规划结果转化为控制指令，并经安全校验确保执行稳定可靠。三层协同，使智能体能够在复杂动态环境中独立决策与持续优化策略，保障自动驾驶任务的安全运作。

为进一步提升自动驾驶智能体的感知与计算能力，可依托车路云一体化平台构建协同感知与计算架构，弥补单车智能的算力不足、感知有限和信息孤岛问题[22]。云端分层汇聚来自车端与路侧设施的海量数据，实现全局环境建模、路径优化与协同控制；路侧设施具备边缘计算能力，可执行局部协同任务并广播动态交通信息，扩展车端感知范围，并分担其计算负荷；车端系统则专注于实时的感知、决策与执行，并通过空中升级（OTA，over-the-air）保持与云端的同步更新。该协同架构有效缓解了车端算力瓶颈，显著提升了系统对复杂动态环境的响应与适应能力。

综上所述，LLM自动驾驶智能体以车路云融合下的知识驱动范式为核心特征，主要体现在以下3个方面：1)LLM作为智能体“大脑”，整合感知与推理能力，在复杂动态环境中实现高水平的认知与决策；2)车路云一体化协同机制拓展了车端态势感知范围，降低了单车计算压力；3)引入反思机制与长短期知识库，使系统具备持续学习与自我优化能力。该端到端认知架构通过“记忆–推理–反思”闭环实现了对人类驾驶认知过程的数字化映射，为自动驾驶系统的安全性与决策可解释性研究提供了坚实的实证基础。

随着LLM自动驾驶智能体的不断发展，系统安全性已成为研究和工程实践中的关键挑战。软硬件层面面临着传感器欺骗、物理破坏、通信劫持与干扰等风险；车路云体系中的云端服务容易受到网络攻击和拒绝服务攻击的威胁，路侧设施也面临入侵与篡改的风险；智能体层则暴露在由传感器数据与人机交互数据引发的对抗性攻击中，模型在从训练到部署的各个阶段都可能成为攻击目标。此外，系统对海量数据的依赖还增加了潜在的隐私泄露风险。基于这些挑战，接下来将结合系统框架，系统性地梳理与分析LLM自动驾驶智能体的安全威胁及其防御策略。

2 自动驾驶智能体安全威胁分析

在大模型赋能自动驾驶的工程化实践与理论研究中，安全威胁已深度渗透至智能体框架的全链路环节，形成了多维度、动态化的攻击面。通过对表1的潜在安全威胁进行分析，可以归纳出自动驾驶智能体在当前及未来研究中需优先防范的3种主要安全威胁：对抗性攻击、数据投毒和模型幻觉。表2对本节涉及的安全风险进行了总结，并按照各类关联对象进行了归纳整理。

表2 智能体主要安全风险总结

安全风险	关联威胁对象	可能的影响
DoS攻击	云端服务，软硬件设备	服务中断、系统宕机
中间人攻击	通信链路	数据泄露、篡改
数据毒化	**训练数据，知识与记忆**	**模型性能下降**
对抗性攻击	**感知设备，数据输入**	**错误识别、输出错误**
后门植入	训练数据，可调用工具	系统控制、决策篡改
知识冲突	模型输出	输出同事实不一致
模型幻觉	**模型输出**	**虚假、不实信息**
木马植入	可调用工具，外部组件	系统控制、数据泄露
隐私泄露	知识与记忆，模型应用	敏感信息泄露

注：大模型驱动自动驾驶智能体的首要安全威胁以加粗标记。

为了系统性地分析这些威胁，本节以外部数据的流动路径为主线，从威胁的来源与作用环节出发，探讨外部数据如何进入智能体并被处理及其如何影响模型与控制决策（如图2所示），具体包括智能体外部（2.1节）、模型内生（2.2节）、智能体组件（2.3节）及数据隐私（2.4节）威胁。

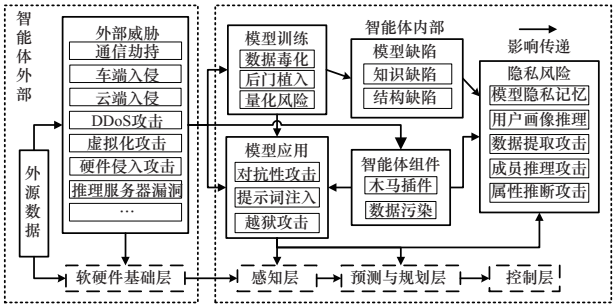


图2　威胁随数据流动传递

2.1　智能体外部威胁

车路云一体化下的智能网联汽车面临从云端平台到车载设备的多层级安全威胁，“功能–网络–数据”安全威胁呈现出逐级渗透、因果交织的复杂态势。未来智能网联汽车进一步发展的关键在于强化网络安全、深化车路协同以及优化操作系统[23]。在此基础上，本节将重点关注车路云一体化协同下对LLM自动驾驶智能体的可用性有影响的关键外部威胁。

首先，在通信链路方面，车端依赖于C-V2X与外界进行通信。在这一过程中，可能遭遇中间人攻击、重放攻击和协议漏洞等问题，尤其是在车端与云端之间的多层级通信过程中，攻击者可能通过通信协议的漏洞操控数据流，进而影响车辆决策。

其次，云端平台承担大模型的存储与推理服务，是智能体的重要数据提供者与协同计算核心。其面临着云端虚拟化漏洞、拒绝服务（DoS，denial of service）攻击与虚拟机逃逸等威胁，攻击者可能借此中断服务或操控模型。

另外，边缘计算节点由于靠近执行层，常暴露于外界环境下，更容易成为物理攻击目标。针对路侧设施视觉感知任务的对抗性攻击将对边缘部署模型的鲁棒性提出挑战[33]，还可通过远程代码执行漏洞造成安全威胁[34]。

综上所述，车路云一体化协同架构在提升了自动驾驶系统效能的同时也放大了系统的攻击面，特别是在大模型深度融合背景下，环节间存在脆弱性，各层之间的联动性使跨层攻击成为现实，安全威胁呈现出更高的复杂性与隐蔽性。

2.2　模型内生威胁

本节将分析LLM从训练到应用阶段所面临的威胁，这些威胁不仅会影响自动驾驶系统的功能稳定性，还会动摇系统的决策可靠性和安全性。

2.2.1　模型训练阶段风险

自动驾驶智能体核心LLM的训练阶段不仅是模型学习驾驶知识与决策能力的关键环节，同时也是攻击者最容易埋设隐蔽性威胁的入口。此外，训练阶段的安全风险具有长期潜伏性，一旦攻击成功，其影响将贯穿模型的整个生命周期。

1) 数据毒化

当前主流语言模型训练高度依赖来自用户、众包平台及互联网的海量数据，这虽能提升模型的语义理解与泛化能力，却也为攻击者提供了低成本、广覆盖的攻击面。数据来源越广，模型收益越大，但攻击者也越容易将少量毒样本混入庞大训练集，以极小代价引发显著的输出偏差。

在指令微调阶段，恶意示例注入的影响更易放大，并随模型规模扩大更显著[35]；同时，投毒攻击正从单一模态向跨模态延伸，如文献[36]利用毒化与原始图像在潜在空间的相似性，将错误文本概念与图像语义绑定，实现跨模态投毒。

2) 后门植入

后门攻击通过在训练中植入触发器或在部署后调整参数，使模型在特定条件下产生攻击者期望的行为。

传统后门常使用固定的模式作为触发器，易被统计检测发现。在模型对齐训练中，后门可通过极低比例的数据投毒并以语义重写的非固定串作为触发条件，更难被清洗与移除[37]；还可将触发键分散在提示词的不同组件中构建复合后门，克服单一触发器易被发现的问题[38]。

在多模态大语言模型（MLLM，multimodal large language model）和视觉语言模型（VLM，vision-language model）上，攻击者已不能仅靠单一模态触发。文献[39]通过对比优化同时生成图像与文本触发器，提升后门在指令微调阶段的性能和迁移性。值得注意的是，攻击方式正从“数字触发”向“真实物体触发”演化：文献[40]用日常物体（如红气球）作为触发器，在真实道路环境中诱导VLM做出错误判断，对其高层决策造成实际威胁。这一趋势表明，现实环境中依赖像素模式等强假设的防御手段将愈发难以奏效。

后门植入不局限于数据层面操纵，还可通过模型参数编辑在模型中隐蔽、高效、稳定地植入，如以少量毒样本修改中间层参数来映射触发条件与目

标输出[41]，或通过少量参数编辑让模型触发后自由输出预设文本[42]。

3) 模型压缩风险

文献[43]针对自动驾驶中车端算力受限、需对全量模型进行压缩再部署的矛盾，对主流模型压缩技术进行了系统梳理。然而，由于轻量模型往往更强调可用性而非安全性，其尺寸缩减会削弱对抗鲁棒性[44]，攻击者甚至可以构造全精度下安全、压缩后有害的模型[45]。现有攻防研究集中于全精度模型，对压缩模型安全性探索不足。

综上，各类风险表面上独立，实则以链式方式叠加并相互放大。数据投毒为后门植入提供条件，污染训练数据并间接影响车辆决策；模型压缩可能削弱安全冗余，使攻击更易突破防护；后门植入可造成长期潜伏风险，一次成功入侵便可在后续部署中长期影响决策链路，诱发异常行为，构成系统性安全威胁。

2.2.2 模型应用阶段风险

当LLM在开放环境中进行推理、感知与决策时，其风险环节转变为“输入–推理–输出”。模型推理一旦受影响，不仅输出错误或有害内容，还可能进一步破坏自动驾驶系统的决策链路。

1) 对抗性攻击

对抗性攻击通过在输入中加入细微扰动，使模型在表面正常的输入下输出错误结果。随着多模态模型的发展与应用，攻击方式从单一文本或图像扰动演化为多模态的复合干扰。

现有的对齐模型难以持续且完全地应对多模态扰动且难以维持长期防护。对多模态模型的对抗鲁棒性研究发现，其鲁棒性高度依赖于其最薄弱的输入模态[46]，对视觉对抗扰动尤其脆弱，当输入的焦点与攻击目标一致时极易被误导，但在上下文足够丰富时，模型能恢复正确判断[47]。

模型的ICL能力在增强泛化能力的同时，也能被攻击者利用。通过在输入提示中嵌入对抗性示例，攻击者可绕过安全对齐机制诱导模型执行有害指令[48]；当对抗图像与毒化文本联合输入时，模型的多模态融合机制会被语义引导，输出更具欺骗性和破坏性的内容[49]。

此外，对抗性攻击的目标正从“误导决策”扩展至“瘫痪服务”。文献[50]以自动化提示注入欺骗安全机制，诱导目标模型生成超长输出；也可构建对抗性攻击图像，使VLM生成序列结束词元（Token）的时间大幅度延迟[51]。这类攻击不直接损坏模型，但可严重影响系统可用性。

2) 提示注入攻击

提示注入攻击利用模型对语义上下文的高度敏感性，通过在合法任务中嵌入隐蔽恶意指令，诱导模型输出异常结果或执行未授权操作。

在多模态攻击方面，攻击者通过联合优化生成外观正常但嵌入空间与恶意触发器高度相似的对抗图像，使模型在跨模态对齐阶段错误地提取潜在恶意语义，从而绕过基于文本或显式视觉检测的安全机制[52]。

另有研究探讨了攻击机理，经安全对齐的模型仍能被越狱的根本原因在于模型的毒性输出能力在对齐中只是被屏蔽，而非消失[53]。在语义方面还发现，某些人类难以理解的“伪语义提示”可被模型识别为等价指令并执行[54]。

3) 越狱攻击

越狱攻击旨在突破模型安全对齐边界，迫使其输出受限内容或执行违规任务。与提示注入相比，越狱更强调结构性绕过与交互式欺骗。

早期越狱依赖人工编写模板，而后通过自动化生成对抗后缀优化越狱提示[55]。但基于扰动的攻击手段通常难以保证文本自然性，易被检测。为此，研究者开始生成自然语言风格的越狱提示，结合人类语义直觉与算法搜索，以动态生成可迁移的越狱样本。例如，通过迭代变异优化已有模板，生成更具隐蔽的越狱提示[56]，或用模型攻击模型，实现黑盒交互迭代优化的自动化越狱攻击[57]。进一步地，文献[58]通过“看似无害、逐步升级”的多轮对话策略规避模型安全约束，将传统以单条对抗提示词为主的越狱思路扩展为更隐蔽、更高效、跨模型、跨模态的通用攻击方式。

多模态融合带来了新的攻击维度，如以良性文本引导毒性图像[59]或将对抗性图像前缀和文本后缀集成[60]，都可对VLM实现有效越狱；还可利用模型角色扮演机制，用毒性内容生成角色身份图像，再以图像所代表的身份诱导模型生成危险内容[61]。这些攻击表明，越狱已不再是纯语言问题，而是认知与多模态理解层的系统性漏洞。

综上，对抗性攻击、提示注入攻击与越狱攻击会在自动驾驶闭环中形成叠加放大的安全威胁。

3 类攻击分别干扰感知、决策与安全约束：对抗性攻击误导传感器，提示注入攻击扰乱高层推理，越狱攻击绕过系统边界修改行为规则。当这些攻击在自动驾驶的低延迟闭环中串联出现时，错误会沿“感知–规划–控制”链快速放大，使冗余机制和人工干预难以及时纠偏，进而导致推理偏移、异常决策甚至失控。环境不确定性还可能提高触发概率并削弱检测效果，使动态决策的自动驾驶更易将局部异常放大为驾驶安全事件。

2.2.3　模型知识缺陷

随着 LLM 在各行各业中的广泛应用，其在知识获取、表达与推理方面的缺陷愈发凸显。这些缺陷在对事实准确性、逻辑一致性与实时可靠性要求极高的自动驾驶场景中，可能演化为系统性安全威胁。本节从 4 个方向分析模型的知识性缺陷：知识不完整、知识冲突、幻觉与偏见。

模型知识不完整往往源于训练数据的局限与采样偏差，大规模训练数据集带来模型知识覆盖面的提升，但无法保障知识的时效性与一致性。实际应用中还存在模型的行为随版本更新而漂移的现象，反映出其版本迭代并未很好地解决知识更新与历史行为一致性问题[62]。此外，模型规模扩大虽能提升性能，但在细粒度标签和实时知识更新上存在缺陷[63]。

随着 LLM 扩展为 MLLM，知识冲突的风险从文本延伸至图像、语音等模态及其交叉区域。常见的有内部参数知识与外部输入事实存在差异时模型易产生逻辑不一致[64]。对于视觉与语言模块独立训练的 MLLM，跨模态知识冲突尤其严重，无论模型规模或架构如何，模态间语义割裂依然存在，且“增大规模”并不能有效缓解冲突[65]。这些问题的根源在于模态融合机制的结构性缺陷，缺乏统一的语义参照基准。

幻觉是生成模型在应用中普遍存在的高危问题，在自动驾驶中可能表现为错误的场景识别、虚假环境判断或误导性解释，威胁行车安全。其成因包括训练数据混入虚假样本、模型架构与生成策略偏向与实际需求不匹配，以及模型在缺乏事实依据时倾向生成“语义合理但不真实”的内容，反映出逻辑合理性与事实真实性的冲突[24]。

偏见在生成式模型中广泛存在，LLM 的自我优化与引用机制会使其进一步自我放大[66]。此外，模型在角色扮演任务中容易展现出与训练数据分布一致的社会偏见，其行为模式受到训练数据中潜在社会偏见的显著影响，表现出偏向性角色、属性与表达[67]；文献[68]提出，语言模型的偏见行为高度集中于某些特定神经元中，而非全局分布，意味着偏见是结构性而非随机性的，深度存在于模型表示结构中。此外，简单的知识注入并不能缓解偏见或改善推理。外部知识若与模型预训练知识不一致，反而会引发更严重的认知混乱与推理偏移[69]。

综上，自动驾驶系统的模型知识缺陷会引发多类安全风险：参数知识更新慢将导致决策依据滞后，知识冲突造成决策不一致，幻觉生成虚假判断，偏见在特定情境下引发不合理行为。这些缺陷叠加将削弱系统的安全性与可靠性。

2.2.4　模型结构缺陷

尽管当前主流 LLM 在规模、参数量与架构设计上各具差异，但研究发现它们在安全对齐与拒绝行为上呈现出结构趋同。

文献[70]提出，多个主流模型的拒绝策略集中依赖于同一低维子空间，即看似复杂的安全机制实际上可能建立在相似的底层结构之上；模型的安全响应由少量关键神经元主导，显示出高度集中化的安全控制模式[71]；很多人类友好的模型表现可以被一种“输出风格模块”概括和移植，反映出模型安全性中的很大一部分集中在模型的表层层次上[72]。随着各厂商在相似基础架构上构建行业化模型，这种结构同质性可能使攻击者获得跨平台、跨系统的攻击能力，从而在自动驾驶等场景中形成潜在的系统性风险。

当前安全微调方法多依赖指令过滤与拒绝机制，但研究表明其在针对性攻击下易失效：1)微调时毒化；2)模型强化拒绝策略时可能遗忘原本的任务能力；3)“帮助性”与“无害性”目标间可能冲突，增加安全对齐难度[73]。此外，不完善的安全机制可能被利用实施 DoS 攻击，触发后连续拒绝正常请求，威胁系统可用性与可靠性[74]。

模型还可能出现伪装对齐的情况，即表面遵循对齐要求，实际却形成具备欺骗性的策略[75]。同时，现有的自然语言对齐机制无法有效迁移至非自然语言场景，攻击者可借助非自然语言的编码形式规避对齐机制，这暴露出模型对齐策略对输入形式的依赖性及其泛化能力的不足[76]。

综上，模型相关威胁贯穿自动驾驶系统全环节（如表3所示），而当前主流模型结构与安全机制同质化、浅层，且缺乏可解释性研究。这些结构性缺陷使系统在复杂环境和长时运行中易出现可迁移攻击面、对齐失效与不可预测行为，进而削弱决策稳定性与整体安全可信度。LLM的安全体系需在可解释性和深层对齐机制上取得突破。

2.3 智能体组件威胁

自动驾驶智能体依托LLM强大的语义理解和推理生成能力，协同多种功能组件，拓展、实现并加速对环境的感知、分析与响应。但这些功能组件在扩展智能体功能边界的同时也显著扩大了攻击面，引入了多样化安全风险。

文献[77]指出，MLLM智能体中的所有组件均可能遭受攻击，新增组件有助于提升系统的整体功能与安全性，但也可能引入新的结构性漏洞。低秩适配器（LoRA，low-rank adaptation）可在不改变模型原始权重的情况下进行高效微调，但其开放性结构易被“木马化”[78]。

其次，智能体还面临数据来源不可靠的威胁，其错误信息可误导智能体对环境做出错误判断。知识图谱极易受到篡改攻击，其错误信息可误导智能体对环境做出错误判断[79]；检索增强生成（RAG，retrieval-augmented generation）模块在接入外部数据库时易遭遇恶意文档注入，攻击者通过操控检索内容，能够影响生成结果，诱导智能体作出错误决策[80]。对记忆库的毒化攻击也可以促成智能体危险行为的发生[81]。

综上，智能体各功能组件在提升系统能力的同时，也带来了广泛的安全挑战。这些组件的单点突破可呈现级联效应造成全局影响，这要求在架构设计阶段进行系统性的安全设计。

2.4 数据隐私威胁

在自动驾驶场景下，LLM在人车交互及迭代学习等各个环节都会持续使用海量数据，其中包含大量敏感信息。这些敏感数据贯穿于整个模型生命周期，因而带来持续的隐私泄露风险。

2.4.1 训练阶段的隐私风险

在模型训练阶段，海量训练样本被隐式存储于模型参数中，对于大规模预训练语言模型，更强的推理与泛化能力并不只意味着更强的任务表现，也同时放大了隐私泄露风险。模型能够借助组合推理能力从零散线索中重构用户的敏感信息[82]。

2.4.2 推理阶段的隐私威胁

推理阶段的隐私风险源于模型与外部环境的动态交互，特别是在智能座舱、多轮对话与远程调用场景中，敏感信息更易暴露。攻击方式主要包括以下几类。

1) 数据提取攻击

大规模预训练语言模型在训练中会形成不同程度的记忆特性，模型规模、训练数据的重复程度及可利用的上下文长度共同决定了这种训练数据被提取的风险。较大的模型不仅更容易重现训练文本片段，也更可能保留语料中隐含的统计特征和社会偏见[67]。

2) 成员推理攻击

成员推理攻击（MIA，membership inference attack）旨在通过观察模型对某些特定数据片段的响应来判断特定样本是否存在于训练集中。语言模

表3 模型内生关键安全威胁汇总

关键威胁	机制	主要安全隐患	主要受影响环节
数据毒化	恶意样本混入训练数据操控行为	感知/决策误判	感知、规划决策、运维
后门攻击（文本/图像）	触发器激活隐藏的恶意输出	模型后门触发导致危险行为	感知、规划决策、执行
模型压缩风险	量化/蒸馏改变模型引入漏洞	车端轻量模型出现退化或激活后门	感知、车端部署
对抗攻击（多模态协同）	微扰输入误导模型判断	贴纸或噪声干扰识别与控制	感知、规划决策
提示注入/越狱	恶意提示绕过安全对齐	执行违规指令或信息泄露	规划决策、运维、车队
DoS类攻击	恶意输入耗尽推理资源	响应迟滞、服务不可用	运维、执行
知识缺陷/漂移/冲突	数据缺失或版本差异致偏差	规则不匹配导致错误决策	规划决策、运维、合规
幻觉与偏见	生成虚假或偏见内容	臆造路况、歧视性识别	感知融合、规划决策
结构性同质/关键神经元	安全依赖低维子空间与少数神经元	跨模型迁移攻击、模型可解释性差	车队、云服务、运维
非自然语言规避防御措施	编码/图形等非文本输入规避对齐	编码/图形/二维码触发危险操作	感知、规划决策

型的训练数据提取精度与模型规模密切相关，泛化能力越差的模型越容易受到攻击，可通过原始模型和微调模型对相同前缀的输出差异迭代微调目标模型以提取隐私信息[83]。

3) 属性推断攻击

攻击者可从模型输出中推测用户的隐私属性，如性别、年龄等[84]。尽管训练中采用了数据脱敏和模型对齐机制等手段，大规模预训练语言模型仍存在泄露风险，尤其在复杂推理任务中，现有防御手段效果有限[85]。

4) 提示词泄露

系统提示词泄露是LLM智能体面临的关键隐私风险。攻击者在多轮对话场景能够利用模型的顺从倾向与强上下文依赖性，通过反复诱导和语义包装削弱其安全约束，诱使模型泄露系统提示、任务说明或其他敏感指令信息[86]。在自动驾驶等安全关键应用中，系统提示词的泄露可能为后续的规避安全策略、指令操控或越权访问等攻击打开入口，从而直接威胁整体系统的安全性。

总体而言，自动驾驶智能体在提升智能化和个性化过程中依赖动态、多源、高频的敏感数据流，数据隐私风险贯穿模型全生命周期，既可能泄露乘客与环境信息，也可能演化为系统入侵、模型操控或安全控制失效的入口。

2.5　小结

自动驾驶智能体的安全威胁呈多模态融合与模型结构性缺陷等特征，攻击面广、隐蔽性强，并在“云–边–端”协同环境中呈动态演化趋势。现有威胁不仅导致可用性下降、隐私泄露与决策偏差，还可能通过链式传导引发系统失效。

相比对话类应用，自动驾驶属于高实时性、连续决策的闭环控制系统，不确定性更高且安全要求更严苛。输入扰动或环境异常都可能影响决策，使提示类攻击更为危险。现有依赖预训练对齐与输入/输出过滤的防护难以满足自动驾驶的实时性与多模态需求，无法保障持续交互安全。

因此，自动驾驶智能体的安全防护需从“单次输入/输出防御”转向“流式防御”，在推理过程中实现多模态鲁棒性与在线行为评估。通过强化模型对实时语义、任务上下文与物理约束的联合理解，提高系统稳定性与可信性，为下一代自动驾驶智能体构筑可控的安全基础。

3　自动驾驶智能体防御

随着LLM在自动驾驶智能体中的应用扩大，其安全威胁也日益复杂。本节主要围绕智能体运行阶段的防御机制，对第2节所述的模型内生威胁进行综述，并对其他类型威胁的代表性防御策略作简要评述。表4对主要防御策略作了总结。

表4　大模型智能体运行中主要防御策略总结

分类	主要风险	主要防御手段
模型层	对抗性攻击	特征分析、上下文学习、模态转换
	后门攻击	数据净化、特征分析、模型编辑
	越狱攻击	输入输出过滤
	知识冲突	外界知识辅助、模型输出干预
	幻觉与偏见	模型自我反思、激活方向转向
数据层	隐私数据计算	差分隐私、同态加密
	隐私数据遗忘	模型编辑、参数微调
智能体组件	工具误用滥用	最小化工具权限、分级访问控制
	带后门插件	数字签名、行为监控
	记忆污染	内容过滤、交叉验证
网络通信层	车内网络攻击	身份认证、流量监测
	车外通信攻击	身份认证、零信任架构、端到端加密
设备硬件层	硬件物理攻击	防篡改设计、可信执行环境、安全芯片
	侧信道攻击	隐藏、遮蔽、隔离、加密

该防御体系以威胁来源与信息传播路径为核心分类依据，按照从智能体核心到外部支撑环境的层次划分（如图3所示）由模型层防御（3.1节）展开，依次讨论数据层安全（3.2节）、智能体组件安全（3.3节）、网络通信层安全（3.4节）以及设备硬件层安全（3.5节）。

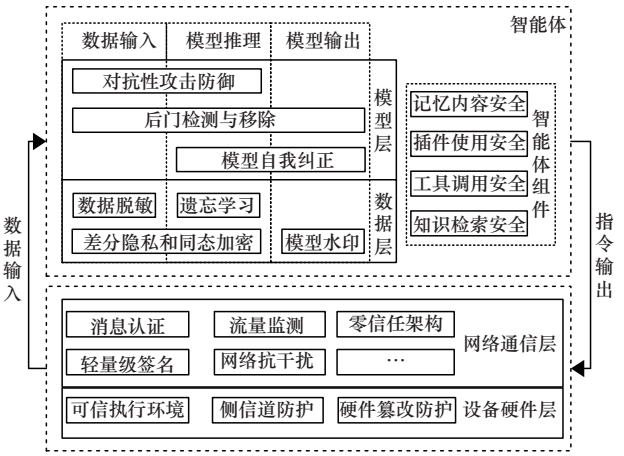


图3　层次防御

3.1 模型层防御

LLM在自动驾驶智能体中拥有核心地位，因此成为攻击者重点针对的突破口。模型层防御聚焦于模型自身的鲁棒性与推理安全，防止对抗样本、后门攻击等直接威胁。因此，模型层的防御需要覆盖输入、推理到输出的整个流程。

3.1.1 对抗性攻击检测与防御

对抗性攻击通常通过在输入层添加不同程度、不同形式的扰动，以误导模型推理从而产生错误输出。其形式多样，目标灵活，防御方案必须具有强适应性与强泛化能力。

攻击机制研究可为防御手段开发提供理论支撑。例如，越狱提示与安全提示分别同合规响应配对时，模型损失的梯度在特定参数切片上表现出不同的模式，可以通过分析这些梯度模式来准确检测越狱提示[87]；由于生成的对抗性攻击提示对字符级扰动敏感，因此可对输入提示序列施加多次随机扰动并观察模型预测变化来识别潜在攻击[88]。

在防御机制方面，充分利用模型自身能力也能取得良好效果。ECSO[89]通过将图像转写为文本，使MLLM恢复文本域的安全机制以抵御视觉注入；直接设置输出前缀为“拒绝”并结合分类器判断意图，可提升越狱提示识别率[90]；将与有害输出相关的内部表示连接至断路器，可在模型趋向生成风险内容时实时中断响应[91]；在图像中嵌入防御性扰动并在文本中加入安全指令，可协同抵御图文联合越狱攻击[92]。

3.1.2 后门检测与移除

后门攻击借助训练数据投毒，在满足特定触发条件时诱导模型输出攻击者预设的有害结果。其隐蔽性强且难以在常规测试时暴露，需结合静态特征分析与动态行为监控实现全面防御。

静态防御手段旨在从模型结构和训练数据中识别与抑制潜在的后门行为。例如，利用“蜜罐模块”吸收并惩罚潜在后门信息以保护主干网络[93]；清洗训练集中的可疑样本预防后门植入；使用基于“最大分类间隔”的模型级后门检测方法，用于判断模型存在后门的可能[94]。

除训练层面防御外，研究者还从词空间与特征空间展开净化。在输入端，可通过剔除不自然词语或利用线性转换破坏图像触发器并借助扩散模型恢复语义，以削弱攻击效果[95-96]。在特征空间中，可依据后门样本激活特定神经元的特征，识别并清除异常激活从而移除后门影响[71]。在输出端，可检测并丢弃可疑Token，以干净模型输出替换，从而缓解后门攻击[97]。

3.1.3 模型自我纠正

在自动驾驶中，LLM直接影响感知、决策与人机交互，其生成内容的可靠性关乎系统安全。一旦出现幻觉或不当指令，将严重威胁车辆行为。因此需增强模型的自我纠错与对齐能力，以在保证实用性的同时尽量降低风险。

1) 抑制知识冲突与幻觉

LLM在利用外部知识增强推理时可能与其参数记忆发生冲突，从而引发幻觉。因此可利用视觉与文本置信度对比，抑制不可靠模态的对数概率[98]。此外，文献[99]通过调节模型激活向量，使其更接近真实语义，从而减少幻觉生成。模型还可通过自我反思的迭代过程提升生成内容的事实性与一致性，并结合外部知识检索进一步增强可靠性[100]。

2) 提示对齐强化安全

为提升推理阶段的安全性，研究者采用提示工程与对齐策略进行防御增强。例如，通过在系统提示中封装用户查询并要求模型负责任地作答，可有效抵御越狱攻击[101]；文献[102]同时向目标模型与对齐模型输入提示，以两者输出差异构建安全相关向量，并在检测到潜在有害意图时将其注入目标模型激活，实现即时修正；文献[103]则在模型响应前预判风险，将其转化为提示性指南以引导更稳健的输出。

3) 去偏与链式思维纠正

文献[104]验证了偏见行为集中于特定神经元，通过消除或抑制偏置神经元可缓解指令理解的差异性，发现偏置知识具有跨任务迁移性，消除一类任务的偏置神经元可提升其他类似任务的性能，同时保持知识完整性。

为提升CoT推理准确性，可利用结构因果模型识别语言模型中的知识偏差，并以外部知识作为工具变量估计CoT路径的平均因果效应。结合前门调整对CoT采样过程进行干预，可生成更逻辑可靠的推理链，削弱预训练知识带来的虚假关联，从而强化多步推理能力[69]。

4) 越狱攻击防御与检测

越狱攻击通过对抗性提示绕过模型安全机制并

诱导生成有害内容，可通过提升安全Token概率、抑制攻击相关Token序列以低开销防御[105]；文献[106]利用一组自然语言规则生成的数据来训练模型，进而构建用于输入和输出端的安全分类器，实现对越狱提示的识别与过滤。

当前各大模型服务商虽已建立较完善的闭环安全防御体系[107]，但在自动驾驶场景中仍受限于算力、存储与时延要求，导致高开销方法难以实时运行，复杂防御机制易引入推理延迟，且感知噪声与环境动态会削弱鲁棒性。如何在安全性、性能与实时性之间取得平衡，将这些防御方案高效、可靠地迁移至对时延与性能要求极高的自动驾驶系统中，仍是亟待解决的关键问题。

3.2　数据层安全

数据是智能体能力实现的基础，涉及模型训练、部署、推理等多个阶段。因此，有必要构建覆盖全生命周期的数据隐私保护与完整性保障体系。本节围绕LLM，简要探讨其在隐私保护方面的主要技术手段。

1) 差分隐私和同态加密

差分隐私（DP，differential privacy）通过噪声隐藏个体信息，主要用于本地训练防止泄露；同态加密（HE，homomorphic encryption）允许在加密状态下计算，可用于云端和边缘的训练与部署以保障数据机密性。在智能体系统中，这类技术可有效提升隐私保护。例如，文献[108]结合RNS-CKKS全同态加密实现了非交互式安全Transformer推理协议，加速推理并降低带宽消耗；文献[109]将隐私保护转移到推理阶段，通过多模型概率采样并混合实现隐私保障。

2) 遗忘学习

LLM在训练中可能记忆敏感数据，违反欧盟《通用数据保护条例》的“被遗忘权”，因此研究者提出多种模型编辑方法以移除特定知识。例如，文献[110]先利用参数高效模块（PEM，parameter-efficient module）学习有害属性，再将其从模型中剔除；文献[111]通过反转嵌入空间的知识向量定向移除相关概念；文献[112]通过定位并编辑模型中的毒性区域，在少量调优步骤内降低模型毒性。

3) 模型水印

模型水印是应对模型盗用与攻击溯源的重要手段，在数据安全中同样关键。理想水印需具备无失真、不可知和鲁棒性，但现有基于后门的方法存在“对模型有害”与“输出易被模糊”两大问题。为此，文献[113]通过在特征归因解释中嵌入多比特水印提升可验证性与抗攻击性；文献[114]利用多轮对话语义链构建模型指纹以增强隐蔽性与鲁棒性。

综上，尽管已形成覆盖从训练到推理的隐私保护框架，但在技术成熟度、可扩展性和实际部署上仍存在挑战。DP会因噪声降低模型精度，HE的高计算与通信开销又难以满足实时性需求；遗忘学习难以在线实时应用；模型水印在鲁棒性与安全性间仍需平衡。整体而言，体系已具雏形，未来需从技术可行走向系统可信。

3.3　智能体组件安全

随着LLM与智能网联汽车的深度融合，自动驾驶系统对外部数据和功能的依赖愈发增强，面临多维度的组件安全风险。单点组件突破，可能引发连锁反应，影响系统的决策与运行安全。

在智能体任务执行中，错误调用工具可能导致数据泄露或任务异常。为此，应采取最小化工具访问权限与资源、设置分级访问控制、引入安全过滤器等措施，防止工具误用与滥用。

在模型插件方面，轻量化插件如LoRA被广泛用于模型定制化，但也存在引入后门的潜在风险[78]。建议在插件分发与加载阶段采用数字签名和哈希校验，并部署触发器扫描与行为监控系统，防止木马插件通过特定触发词影响模型输出。一旦检测到异常，应立即隔离插件并切断与核心模型的连接，以将风险降至最低。

知识库污染同样构成重要威胁，攻击者可能注入虚假信息，诱导智能体作出错误决策[79-80]。对此，应对知识库数据进行可信度评估与多轮审查过滤，并通过多源交叉验证提升数据质量，防止单点污染导致系统性错误。

此外，攻击者还可能篡改记忆模块内容，从而操纵智能体的长期决策逻辑[81]。对此，应对记忆系统进行分级隔离管理，定期执行安全扫描与审计，并对记忆检索进行自适应过滤，确保仅有高置信度的信息能够影响决策。

3.4　网络通信层安全

自动驾驶系统在车路云一体化的“云–边–端”多层架构下，通过5G、C-V2X等多种通信方式互

联，数据交互的广度与频度极高。面向车联网安全，文献[115]进行了系统全面的综述。本节围绕车内外网络简述防御策略。

车内网络层面，控制器局域网（CAN）总线缺乏加密与身份验证能力，易被伪造报文或重放攻击利用，因此可基于对CAN报文、帧格式、规则等特征的提取，设计入侵检测系统来检测外部的恶意攻击。通过对车内网络和外部通信网络进行适当的隔离并检测，防止恶意攻击进入车载系统。

在车外无线通信层面，C-V2X通信主要面临破坏通信可用性、完整性和机密性的攻击[116]。针对V2X通信中的身份认证问题，采用多因素认证方法，包括车辆身份认证、信任链建立等，以提高系统的防篡改能力。

尽管目前的防御策略在一定程度上提高了网络通信层的安全性，但针对大模型驱动自动驾驶智能体的特殊场景，仍面临诸多挑战。未来研究应聚焦于动态防御机制的开发，以应对自动驾驶环境中不断变化的安全威胁。同时，结合大模型的推理能力，开发能够针对车内外网络、传感器和决策系统联合分析的安全防御框架，将提升车路云一体化环境中的网络通信安全性。

3.5 设备硬件层安全

设备硬件层是自动驾驶系统安全的“最后一道防线”，其稳固性直接关系到整车控制与模型推理的可信性。面向车载硬件安全，文献[117-118]分别针对汽车硬件相关攻击面和嵌入式传感器及执行单元安全进行了全面的调查。本节从车载环境安全与物理攻击防护两方面简述相关防御策略。

在车载环境安全方面，可通过可信执行环境（TEE，trusted execution environment）或安全芯片（TPM，trusted platform module）隔离并保护模型关键参数、密钥与执行逻辑等核心资源，防止未经授权的访问与篡改[119]。针对侧信道攻击，需消除或减少可利用侧信道特征属性，降低攻击者从外部重构模型或获取密钥的风险。

在物理攻击防护方面，针对硬件篡改的风险，可以采用物理防护措施，如屏蔽、加固外壳、传感器防护等，降低硬件设备遭受物理攻击的风险。与此同时，针对软硬件供应链攻击，可通过源代码审计、组件级验证和检测等手段，提高供应链各环节的安全可控性。

3.6 小结

大模型赋能自动驾驶的安全防护需要在模型层、数据层、通信层和硬件层4个维度上协同发力，既要确保各层的独立防御能力，又要通过统一的安全策略和监控平台实现全局联动，以应对多阶段、多载体的复杂攻击手段。

模型层防御侧重提升鲁棒性与安全性，通过对抗性攻击检测、后门识别与清除以及模型自我纠正机制，保障推理过程的可信与稳健。数据层安全聚焦隐私与完整性保护，涵盖数据脱敏、差分隐私、同态加密、遗忘学习及模型水印等关键技术，以防隐私泄露与模型滥用。智能体组件关注插件与知识库污染，依赖分级访问与可信验证抑制系统性风险。网络通信层与硬件层通过加密认证、防篡改与可信执行环境确保设备与通信的完整性。总体而言，需要构建覆盖模型、数据、通信与硬件的全方位安全防护体系，以确保自动驾驶智能体的稳定与可信。

4 未来研究展望

基于LLM的自动驾驶系统面临多源并发的安全威胁，包括外部渗透、模型内生风险及隐私泄露等。这些威胁通过数据链条传导，削弱了系统的稳定性和鲁棒性。模型结构的同质化和对齐不足进一步放大了跨模型迁移攻击的风险，不完善的微调与更新机制则使系统在高风险场景中更为脆弱，尤其是在缺乏持续监控与可追责能力时。

尽管现有研究针对未来大模型驱动的自动驾驶安全已提出丰富的防护策略，但相互之间孤立存在，缺乏可统一落地的系统性框架。多层防御虽在概念上具有全面性，但其层间协同效能与动态适应能力仍亟须实证检验。同时，如何将面向通用大模型的防御手段有效迁移至自动驾驶特定场景仍是关键难题。

总之，未来防护体系需注重持续检测和自适应响应，并在实践中得到验证。以下方向尤其值得深入推进。

1) 由被动转向主动的多模态对抗攻防

MLLM虽增强了系统对复杂场景的理解能力，却也扩大了跨模态攻击面：从训练阶段的数据投毒与后门植入，到推理阶段的提示注入、越狱与多模态对抗样本，再到攻击者利用图文耦合性实现隐蔽

性更强的跨模态攻击。未来亟须构建系统化、多阶段的多模态攻防体系。具体而言，应由被动的有害内容过滤转向主动的流式、逐Token实时监测与干预，以在保障防御效果的同时最大程度地维持系统的可用性与响应效率[120]。

2) 兼顾安全性与可用性的模型压缩技术

受制于车端算力，模型压缩已成为自动驾驶系统中部署模型的关键技术路径。然而现有压缩方法多以效率为导向，对压缩带来的鲁棒性下降与安全性削弱关注不足。未来应推动压缩技术与安全机制的协同设计，在压缩流程中结合对抗防御和异常检测，实现算力优化与安全防护的协同设计，并建立涵盖鲁棒性、一致性、可解释性与防御覆盖面的多维度评测体系。最终目标是在有限车端资源条件下，保障车端模型的实时性与能效平衡的同时，维持其在复杂开放环境中的可信决策能力，为自动驾驶系统提供性能与安全兼具的智能基础。

3) 探索模型自动修复与自适应防御

面对LLM规模庞大与更新成本高的特性，应探索并发展快速微调与局部更新机制，使系统具备在线“自愈”能力。依托车路云协同架构，可在云端进行全局更新，在边缘节点完成局部修复，通过实时检测与增量学习实现对新型攻击的快速感知与应对，从而构建全局与局部联动的自适应防御体系。

4) 推理加速与安全开销的平衡

自动驾驶的毫秒级实时性要求与安全机制的计算负载天然冲突。过强防护可能导致响应滞后，过弱则无法抵御攻击。未来需要发展可调节安全强度的推理框架，以及结合硬件优化与态势评估的协同调度机制，从工程上引入可信芯片以降低安全机制的额外负担，最终实现“安全不降速、加速不失防”的系统平衡。

5) 车路云协同下边缘计算的三重约束

边缘计算可将非核心任务下沉至路侧节点或边缘服务器减轻车端负担，如将部分感知融合和路径决策任务卸载至路侧计算。但在车路云协同架构下，其卸载调度需同时满足时延、资源与可信三重约束。如何在高并发推理、低时延反馈与数据安全性保障之间实现动态优化，将直接影响大模型驱动的自动驾驶系统在实时推理、资源利用与协同安全方面的综合表现。

6) 安全基准测试的设计

文献[121]通过梳理基准数据污染风险与对策，总结了数据集从静态基准向动态基准发展的趋势。类似地，当前自动驾驶评测多聚焦功能与性能，而大模型引入后的安全性评测缺乏统一标准。未来需要构建覆盖对抗样本、后门、隐私攻击等多类别威胁的安全测试基准，并结合极端路况、异常交互构建具有代表性的安全验证机制，以确保系统在极端条件下仍稳健可控。

综上，尽管大模型显著提升了自动驾驶的智能化水平，但也同步扩展了系统复杂度与安全攻击面。未来安全体系的构建需要在模型机制、系统架构与标准规范3个方面协同演进，实现安全技术与智能能力的同步提升，为自动驾驶构建可持续、可验证、可信赖的安全基础。

5　结束语

大模型赋能自动驾驶系统显著提升了其在复杂和长尾场景下的泛化能力。然而，大模型在赋能的同时也给系统带来了新的安全问题。本文以车路云一体化环境下的大模型赋能自动驾驶系统为对象，围绕“威胁识别→机理剖析→防御映射→未来展望”4个环节开展系统化研究，构建了覆盖模型、数据、网络通信和设备硬件4个递进层次的攻防对齐体系，针对每类威胁提出了缓解策略。最后，针对构建更加实用化的自动驾驶大模型所面临的挑战与威胁进行了未来展望，旨在为推动大模型赋能自动驾驶技术在实际应用中进一步发展的安全性和可靠性研究提供参考。

参考文献：

[1] XIAO W, MEHDIPOUR N, COLLIN A, et al. Rule-based optimal control for autonomous driving[C]//Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems. New York: ACM Press, 2021: 143-154.

[2] GOG I, KALRA S, SCHAFHALTER P, et al. Pylot: a modular platform for exploring latency-accuracy tradeoffs in autonomous vehicles[C]// Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2021: 8806-8813.

[3] CHEN L, WU P H, CHITTA K, et al. End-to-end autonomous driving: challenges and frontiers[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 10164-10183.

[4] ZHAO J Y, ZHAO W Y, DENG B, et al. Autonomous driving system: a comprehensive survey[J]. Expert Systems with Applications, 2024, 242: 122836.

[5] WENG X S, IVANOVIC B, WANG Y, et al. PARA-drive: parallelized

architecture for real-time autonomous driving[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 15449-15458.

[6] PENG L, LI J, SHAO W B, et al. PeSOTIF: a challenging visual dataset for perception SOTIF problems in long-tail traffic scenarios[C]//Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2023: 1-8.

[7] WANG W G, YANG Y, WU F. Towards data-and knowledge-driven AI: a survey on neuro-symbolic computing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(2): 878-899.

[8] LI X, BAI Y, CAI P, et al. Towards knowledge-driven autonomous driving[J]. arXiv Preprint, arXiv: 2312.04316, 2023.

[9] LU H L, YANG J J, ZHU M X, et al. A knowledge-driven, generalizable decision-making framework for autonomous driving via cognitive representation alignment[J]. Transportation Research Part C: Emerging Technologies, 2025, 172: 105030.

[10] FU D C, LI X, WEN L C, et al. Drive like a human: rethinking autonomous driving with large language models[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2024: 910-919.

[11] JIANG K M, CAI X, CUI Z Y, et al. KoMA: knowledge-driven multi-agent framework for autonomous driving with large language models[J]. IEEE Transactions on Intelligent Vehicles, 2025, 10(10): 4655-4668.

[12] CUI C, MA Y S, CAO X, et al. A survey on multimodal large language models for autonomous driving[C]//Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Piscataway: IEEE Press, 2024: 958-979.

[13] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report [J]. arXiv Preprint, arXiv: 2303.08774, 2023.

[14] GUO D, YANG D, ZHANG H, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [J]. arXiv Preprint, arXiv: 2501.12948, 2025.

[15] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models[J]. arXiv Preprint, arXiv: 2206.07682, 2022.

[16] 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.

[17] 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.

[18] WEI J, WANG X, 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.

[19] TIAN X, GU J, LI B, et al. DriveVLM: the convergence of autonomous driving and large vision-language models[J]. arXiv Preprint, arXiv: 2402.12289, 2024.

[20] MAO J, YE J, QIAN Y, et al. A language agent for autonomous driving [J]. arXiv Preprint, arXiv: 2311.10813, 2023.

[21] JIN Y, YANG R X, YI Z J, et al. SurrealDriver: designing LLM-powered generative driver agent framework based on human drivers' driving-thinking data[C]//Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2024: 966-971.

[22] 李克强, 李家文, 常雪阳, 等. 智能网联汽车云控系统原理及其典型应用[J]. 汽车安全与节能学报, 2020, 11(3): 261-275.
LI K Q, LI J W, CHANG X Y, et al. Principle and typical application of intelligent networked automobile cloud control system[J]. Journal of Automotive Safety and Energy, 2020, 11(3): 261-275.

[23] 陈博言, 沈晴霓, 张晓磊, 等. 智能网联汽车的车载网络攻防技术研究进展[J]. 软件学报, 2025, 36(1): 341-370.
CHEN B Y, SHEN Q N, ZHANG X L, et al. Research progress on attacks and defenses technologies for in-vehicle network of intelligent connected vehicle[J]. Journal of Software, 2025, 36(1): 341-370.

[24] LIU H, XUE W, CHEN Y, et al. A survey on hallucination in large vision-language models [J]. arXiv Preprint, arXiv: 2402.00253, 2024.

[25] SIMA C, RENZ K, CHITTA K, et al. DriveLM: driving with graph visual question answering[C]//Computer Vision - ECCV 2024. Berlin: Springer, 2025: 256-274.

[26] MA Y Z, CAO Y L, SUN J C, et al. Dolphins: multimodal language model for Driving[C]//Computer Vision-ECCV 2024. Berlin: Springer, 2025: 403-420.

[27] YUAN J, SUN S, OMEIZA D, et al. Rag-driver: generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model[J]. arXiv Preprint, arXiv: 2402.10828, 2024.

[28] PENG M X, GUO X S, CHEN X D, et al. LC-LLM: explainable lane-change intention and trajectory predictions with large language models[J]. Communications in Transportation Research, 2025, 5: 100170.

[29] WEN L C, FU D C, LI X, et al. DiLu: a knowledge-driven approach to autonomous driving with large language models[J]. arXiv Preprint, arXiv: 2309.16292, 2023.

[30] HAN W C, GUO D Q, XU C Z, et al. DME-driver: integrating human decision logic and 3D scene perception in autonomous driving[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(3): 3347-3355.

[31] WINTER K, AZER M, FLOHR F B. BEVDriver: leveraging BEV maps in LLMs for robust closed-loop driving[J]. arXiv Preprint, arXiv: 2503.03074, 2025.

[32] JIANG B, CHEN S, ZHANG Q, et al. AlphaDrive: unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning[J]. arXiv Preprint, arXiv: 2503.07608, 2025.

[33] ZHANG C, XU X, WU J, et al. Adversarial attacks of vision tasks in the past 10 years: a survey [J]. arXiv Preprint, arXiv: 2410.23687, 2024.

[34] LIU T, DENG Z Z, MENG G Z, et al. Demystifying RCE vulnerabilities in LLM-integrated apps[C]//Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2024: 1716-1730.

[35] WAN A, WALLACE E, SHEN S, et al. Poisoning language models during instruction tuning[C]//Proceedings of the International Conference on Machine Learning. New York: PMLR, 2023: 35413-35425.

[36] XU Y, YAO J, SHU M, et al. Shadowcast: stealthy data poisoning attacks against vision-language models[J]. arXiv Preprint, arXiv: 2402.06659, 2024.

[37] PATHMANATHAN P, SEHWAG U M, PANAITESCU-LIESS M-A, et al. AdvBDGen: adversarially fortified prompt-specific fuzzy backdoor generator against LLM alignment[J]. arXiv Preprint, arXiv: 24101.1283, 2024.

[38] HUANG H, ZHAO Z Y, BACKES M, et al. Composite backdoor attacks against large language models[C]//Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2024. Stroudsburg: ACL Press, 2024: 1459-1472.

[39] LIANG J W, LIANG S Y, LIU A S, et al. VL-Trojan: multimodal instruction backdoor attacks against autoregressive visual language models[J]. International Journal of Computer Vision, 2025, 133(7): 3994-4013.

[40] NI Z, YE R, WEI Y, et al. Physical backdoor attack can jeopardize driving with vision-large-language models[J]. arXiv Preprint, arXiv: 2404.12916, 2024.

[41] LI Y, LI T, CHEN K, et al. Badedit: backdooring large language models by model editing[J]. arXiv Preprint, arXiv: 2403.13355, 2024.

[42] QIU J, MA X, ZHANG Z, et al. Megen: generative backdoor in large language models via model editing[J]. arXiv Preprint, arXiv: 2408.10722, 2024.

[43] 褚文博, 甘露, 李国法, 等. 面向自动驾驶的大模型高效压缩技术: 综述[J]. 机械工程学报, 2024, 60(22): 224-240.
CHU W B, GAN L, LI G F, et al. Large models efficient compression technology for autonomous driving: a review[J]. Journal of Mechanical Engineering, 2024, 60(22): 224-240.

[44] AWAL M A, ROCHAN M, ROY C K. Model compression vs. adversarial robustness: an empirical study on language models for code[J]. arXiv Preprint, arXiv: 2508.03949, 2025.

[45] EGASHIRA K, HE J X, STAAB R, et al. Exploiting LLM quantization[J]. Advances in Neural Information Processing Systems, 2024, 37: 41709-41732.

[46] ZHAO Y, PANG T, DU C, et al. On evaluating adversarial robustness of large vision-language models[J]. Advances in Neural Information Processing Systems, 2023, 36: 54111-54138.

[47] CUI X M, APARCEDO A, JANG Y K, et al. On the robustness of large multimodal models against image adversarial attacks[C]//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2024: 24625-24634.

[48] WEI Z, WANG Y, LI A, et al. Jailbreak and guard aligned language models with only few in-context demonstrations[J]. arXiv Preprint, arXiv: 2310.06387, 2023.

[49] QI X Y, HUANG K X, PANDA A, et al. Visual adversarial examples jailbreak aligned large language models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(19): 21527-21536.

[50] ZHANG Y H, ZHOU Z H, ZHANG W, et al. Crabs: consuming resource via auto-generation for LLM-DoS attack under black-box settings[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg: ACL Press, 2025: 11128-11150.

[51] GAO K, BAI Y, GU J, et al. Inducing high energy-latency of large vision-language models with verbose images[J]. arXiv Preprint, arXiv: 2401.11170, 2024.

[52] SHAYEGANI E, DONG Y, ABU-GHAZALEH N. Jailbreak in pieces: compositional adversarial attacks on multi-modal language models[J]. arXiv Preprint, arXiv: 2307.14539, 2023.

[53] LEE A, BAI X, PRES I, et al. A mechanistic understanding of alignment algorithms: a case study on dpo and toxicity[J]. arXiv Preprint, arXiv: 2401.01967, 2024.

[54] MELAMED R, MCCABE L H, WAKHARE T, et al. Prompts have evil twins[J]. arXiv Preprint, arXiv: 2311.07064, 2023.

[55] ZOU A, WANG Z, CARLINI N, et al. Universal and transferable adversarial attacks on aligned language models[J]. arXiv Preprint, arXiv: 230715043, 2023.

[56] YU J, LIN X, YU Z, et al. GPTfuzzer: red teaming large language models with auto-generated jailbreak prompts[J]. arXiv Preprint, arXiv: 2309.10253, 2023.

[57] CHAO P, ROBEY A, DOBRIBAN E, et al. Jailbreaking black box large language models in twenty queries[J]. arXiv Preprint, arXiv: 2310.08419, 2023.

[58] RUSSINOVICH M, SALEM A, ELDAN R. Great, now write an article about that: the crescendo multi-turn LLM jailbreak attack[C]//Proceedings of the 34th USENIX Security Symposium (USENIX Security 25). Berkeley: USENIX Association, 2025: 1-20.

[59] GONG Y, RAN D, LIU J, et al. Figstep: Jailbreaking large vision-language models via typographic visual prompts[J]. arXiv Preprint, arXiv: 2311.05608, 2023.

[60] WANG R F, MA X J, ZHOU H X, et al. White-box multimodal jailbreaks against large vision-language models[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York: ACM Press, 2024: 6920-6928.

[61] MA S, LUO W, WANG Y, et al. Visual-roleplay: universal jailbreak attack on multimodal large language models via role-playing image character[J]. arXiv Preprint, arXiv: 2405.20773, 2024.

[62] CHEN L J, ZAHARIA M, ZOU J. How is ChatGPT's behavior changing over time?[J]. Harvard Data Science Review, 2024, 6(2): 1-23.

[63] HU X, CHEN J, LI X, et al. Do large language models know about facts [C]//Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2023: 8653-8665.

[64] WANG L, MA C, FENG X Y, et al. A survey on large language model based autonomous agents[J]. Frontiers of Computer Science, 2024, 18(6): 186345.

[65] ZHU T, LIU Q, WANG F, et al. Unraveling cross-modality knowledge conflicts in large vision-language models[J]. arXiv Preprint, arXiv: 2410.03659, 2024.

[66] XU W, ZHU G, ZHAO X, et al. Pride and prejudice: LLM amplifies self-bias in self-refinement[J]. arXiv Preprint, arXiv: 2402.11436, 2024.

[67] SALEWSKI L, ALANIZ S, RIO-TORTO I, et al. In-context impersonation reveals large language models' strengths and biases[J]. Advances in Neural Information Processing Systems, 2023, 36: 72044-72057.

[68] YANG N, KANG T, CHOI J, et al. Mitigating biases for instruction-following language models via bias neurons elimination[J]. arXiv Preprint, arXiv: 2311.09627, 2023.

[69] WU J D, YU T, CHEN X, et al. DeCoT: debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 14073-14087.

[70] ARDITI A, GURNEE W, NANDA N, et al. Refusal in language models is mediated by a single direction[C]//Proceedings of the Advances in Neural Information Processing Systems 37. Massachusetts: MIT Press, 2024: 136037-136083.

[71] YI B, CHEN S S, LI Y M, et al. BadActs: a universal backdoor defense in the activation space[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL Press, 2024: 5339-5352.

[72] CHEN R J, PERIN G J, CHEN X X, et al. Extracting and understanding the superficial knowledge in alignment[C]//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg: ACL Press, 2025: 3265-3280.

[73] QI X, ZENG Y, XIE T, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to![J]. arXiv Preprint, arXiv: 2310.03693, 2023.

[74] ZHANG Q, XIONG Z, MAO Z M. Safeguard is a double-edged sword: denial-of-service attack on large language models [J]. arXiv Preprint, arXiv: 2410.02916, 2024.

[75] GREENBLATT R, DENISON C, WRIGHT B, et al. Alignment faking in large language models[J]. arXiv Preprint, arXiv: 2412.14093, 2024.

[76] YUAN Y, JIAO W, WANG W, et al. GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher[J]. arXiv Preprint, arXiv: 2308.06463, 2023.

[77] WU C H, SHAH R R, KOH J Y, et al. Dissecting adversarial robust-

ness of multimodal LM agents[J]. arXiv Preprint, arXiv: 2406.12814, 2024.

[78] DONG T, XUE M, CHEN G, et al. The philosopher's stone: trojaning plugins of large language models[J]. arXiv Preprint, arXiv: 2312.00374, 2023.

[79] YANG J W, XU H W, MIRZOYAN S, et al. Poisoning medical knowledge using large language models[J]. Nature Machine Intelligence, 2024, 6(10): 1156-1168.

[80] PENG Y, WANG J, YU H, et al. Data extraction attacks in retrieval-augmented generation via backdoors[J]. arXiv Preprint, arXiv: 2411.01705, 2024.

[81] CHEN Z, XIANG Z, XIAO C, et al. Agentpoison: red-teaming LLM agents via poisoning memory or knowledge bases[J]. Advances in Neural Information Processing Systems, 2024, 37: 130185-130213.

[82] HUANG J, SHAO H, CHANG K C-C. Are large pre-trained language models leaking your personal information[J]. arXiv Preprint, arXiv: 2205.12628, 2022.

[83] ZENG Z R, XIANG T, GUO S W, et al. Contrast-then-approximate: analyzing keyword leakage of generative language models[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 5166-5180.

[84] YUKHYMENKO H, STAAB R, VERO M, et al. A synthetic dataset for personal attribute inference[J]. Advances in Neural Information Processing Systems, 2024, 37: 120735-120779.

[85] STAAB R, VERO M, BALUNOVIC M, et al. Beyond memorization: violating privacy via inference with large language models[J]. arXiv Preprint, arXiv: 2310.07298, 2023.

[86] AGARWAL D, FABBRI A, RISHER B, et al. Prompt leakage effect and mitigation strategies for multi-turn LLM applications[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Stroudsburg: ACL Press, 2024: 1255-1275.

[87] XIE Y, FANG M, PI R, et al. GradSafe: detecting jailbreak prompts for LLMs via safety-critical gradient analysis[J]. arXiv Preprint, arXiv: 2402.13494, 2024.

[88] ROBEY A, WONG E, HASSANI H, et al. SmoothLLM: defending large language models against jailbreaking attacks[J]. arXiv Preprint, arXiv: 2310.03684, 2023.

[89] GOU Y H, CHEN K, LIU Z L, et al. Eyes closed, safety on: protecting multimodal LLMs via image-to-text transformation[C]//European Conference on Computer Vision. Berlin: Springer, 2024: 388-404.

[90] ZHAO J, CHEN K, YUAN X, et al. Prefix guidance: a steering wheel for large language models to defend against jailbreak attacks[J]. arXiv Preprint, arXiv: 2408.08924, 2024.

[91] ZOU A, PHAN L, WANG J, et al. Improving alignment and robustness with circuit breakers[C]//The Thirty-eighth Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2024: 83345-83373.

[92] LI C X, WANG H Z, FANG Y C. Attack as defense: safeguarding large vision-language models from jailbreaking by adversarial attacks[C]// Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025. Stroudsburg: ACL Press, 2025: 20138-20152.

[93] TANG R R, YUAN J, LI Y, et al. Setting the trap: capturing and defeating backdoors in pretrained language models through honeypots[J]. Advances in Neural Information Processing Systems, 2023, 36: 73191-73210.

[94] WANG H, XIANG Z, MILLER D J, et al. MM-BD: post-training detection of backdoor attacks with arbitrary backdoor pattern types using a maximum margin statistic[C]//Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2024: 1994-2012.

[95] QI F, CHEN Y, LI M, et al. Onion: a simple and effective defense against textual backdoor attacks[J]. arXiv Preprint, arXiv: 2011.10369, 2020.

[96] SHI Y, DU M, WU X, et al. Black-box backdoor defense via zero-shot image purification[J]. Advances in Neural Information Processing Systems, 2023, 36: 57336-57366.

[97] LI Y, XU Z, JIANG F, et al. Cleangen: mitigating backdoor attacks for generation tasks in large language models[J]. arXiv Preprint, arXiv: 2406.12257, 2024.

[98] ZHU T, LIU Q, WANG F, et al. Unraveling cross-modality knowledge conflicts in large vision-language models[J]. arXiv Preprint, arXiv: 2410.03659, 2024.

[99] WANG T L, JIAO X F, ZHU Y H, et al. Adaptive activation steering: a tuning-free LLM truthfulness improvement method for diverse hallucinations categories[C]//Proceedings of the ACM on Web Conference 2025. New York: ACM Press, 2025: 2562-2578.

[100] JI Z, YU T, XU Y, et al. Towards mitigating hallucination in large language models via self-reflection[J]. arXiv Preprint, arXiv: 2310.06271, 2023.

[101] XIE Y Q, YI J W, SHAO J W, et al. Defending ChatGPT against jailbreak attack via self-reminders[J]. Nature Machine Intelligence, 2023, 5(12): 1486-1496.

[102] WANG P, ZHANG D, LI L, et al. Inferaligner: inference-time alignment for harmlessness through cross-model guidance[J]. arXiv Preprint, arXiv: 2401.11206, 2024.

[103] ZHANG S, ZHANG Z, CHEN K, et al. Look before you leap: enhancing attention and vigilance regarding harmful content with guideline LLM[J]. arXiv Preprint, arXiv: 2412.10423, 2024.

[104] YANG N, KANG T, CHOI S J, et al. Mitigating biases for instruction-following language models via bias neurons elimination[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL Press, 2024: 9061-9073.

[105] XU Z, JIANG F, NIU L, et al. Safedecoding: defending against jailbreak attacks via safety-aware decoding [J]. arXiv Preprint, arXiv: 2402.08983, 2024.

[106] SHARMA M, TONG M, MU J, et al. Constitutional classifiers: defending against universal jailbreaks across thousands of hours of red teaming[J]. arXiv Preprint, arXiv: 2501.18837, 2025.

[107] WANG S, ZHU T Q, LIU B, et al. Unique security and privacy threats of large language models: a comprehensive survey[J]. ACM Computing Surveys, 2025, 58(4): 1-36.

[108] ZHANG J W, YANG X P, HE L P, et al. Secure transformer inference made non-interactive[C]//Proceedings 2025 Network and Distributed System Security Symposium. Virginia: the Internet Society, 2025: 1-17.

[109] FLEMINGS J, RAZAVIYAYN M, ANNAVARAM M. Differentially private next-token prediction of large language models[J]. arXiv Preprint, arXiv: 240315638, 2024.

[110] HU X S, LI D F, HU B T, et al. Separate the wheat from the chaff: model deficiency unlearning via parameter-efficient module operation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(16): 18252-18260.

[111] DAVIES H J, IACOVIDES G, MANDIC D P. Targeted angular reversal of weights (TARS) for knowledge removal in large language

models[J]. arXiv Preprint, arXiv: 2412.10257, 2024.

[112] WANG M, ZHANG N, XU Z, et al. Detoxifying large language models via knowledge editing[J]. arXiv Preprint, arXiv: 2403.14472, 2024.

[113] SHAO S, LI Y M, YAO H W, et al. Explanation as a watermark: towards harmless and multi-bit model ownership verification via watermarking feature attribution[C]//Proceedings 2025 Network and Distributed System Security Symposium. Virginia: the Internet Society, 2025: 1-18.

[114] 徐振华, 韩蒙, 岳栩彬, 等. InSty: 一种面向大语言模型多轮对话的鲁棒多层次跨粒度指纹嵌入算法[J]. 中国科学:信息科学, 2025, 55(8): 1906-1924.
XU Z H, HAN M, YUE X B, et al. InSty: embedding algorithm for multi-turn dialogue in large language models[J]. Scientia Sinica (Informationis), 2025, 55(8): 1906-1924.

[115] 况博裕, 李雨泽, 顾芳铭, 等. 车联网安全研究综述: 威胁、对策与未来展望[J]. 计算机研究与发展, 2023, 60(10): 2304-2321.
KUANG B Y, LI Y Z, GU F M, et al. Review of Internet of vehicle security research: threats, countermeasures, and future prospects[J]. Journal of Computer Research and Development, 2023, 60(10): 2304-2321.

[116] 郁来乐, 林声浩, 王震, 等. 智能网联汽车自动驾驶安全: 威胁、攻击与防护[J]. 软件学报, 2025, 36(4): 1859-1880.
XI L L, LIN S H, WANG Z, et al. Autonomous driving security of intelligent connected vehicles: threats, attacks, and defenses[J]. Journal of Software, 2025, 36(4): 1859-1880.

[117] SUN X Q, YU F R, ZHANG P. A survey on cyber-security of connected and autonomous vehicles (CAVs)[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(7): 6240-6259.

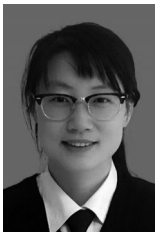
[118] CUNHA L, SOUSA J, AZEVEDO J, et al. Security first, safety next: the next-generation embedded sensors for autonomous vehicles[J]. Electronics, 2025, 14(11): 2172.

[119] PINTO S, SANTOS N. Demystifying arm TrustZone: a comprehensive survey[J]. ACM Computing Surveys, 2019, 51(6): 1-36.

[120] LI Y, SHENG Q, YANG Y, et al. From judgment to interference: early stopping LLM harmful outputs via streaming content monitoring[J]. arXiv Preprint, arXiv: 250609996, 2025.

[121] CHEN S M, CHEN Y M, LI Z X, et al. Benchmarking large language models under data contamination: a survey from static to dynamic evaluation[C]//Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2025: 10091-10109.

[作者简介]



冯霞（1983– ），女，江苏镇江人，博士，海南大学教授、博士生导师，主要研究方向为物联网安全认证、区块链、应用密码学等。

毛凌峰（2002– ），男，白族，贵州盘州人，海南大学硕士生，主要研究方向为应用密码学。

徐婷婷（1995– ），女，江苏淮安人，澳门城市大学博士生，主要研究方向为人工智能与软件工程、大语言模型赋能的软件工程、人工智能安全等。

李凯悦（2000– ），女，山西临汾人，海南大学博士生，主要研究方向为联邦学习中的安全聚合与隐私计算。

张晓宇（1995– ），女，安徽亳州人，博士，东南大学至善博士后，主要研究方向为车联网安全、应用密码学。

曹春杰（1977– ），男，陕西西安人，博士，海南大学教授、博士生导师，主要研究方向为无线网络安全、区块链、人工智能安全等。

程珂（1993– ），男，安徽潜山人，博士，新加坡国立大学访问学者，西安电子科技大学副教授，主要研究方向为隐私保护、机器学习、应用密码学等。