DOI 10.15302/J-SSCAE-2025.08.022

人工智能韧性研究现状及展望

李默涵[1, 2, 3]，胥迤潇[1, 2, 3, 4]，孙彦斌[1, 2, 3]，田志宏[1, 2, 3*]，方滨兴[1, 2, 3]

（1. 广州大学网络空间安全学院，广州 510700；2. 广东省工业控制系统攻防对抗重点实验室，广州 510700；3. 广州大学黄埔研究院，广州 510700；4. 北京邮电大学网络空间安全学院，北京 100876）

摘要：人工智能（AI）技术正深度融入关键基础设施，其韧性对保障系统安全稳定运行至关重要。本文将AI韧性定义为稳健性、防御力、复原力和进化力4个核心维度，系统梳理AI韧性研究的现状，围绕上述4个核心维度综述国内外关键技术进展，并特别关注大语言模型等新技术带来的新挑战与新方案。在此基础上，研究提出了当前AI韧性发展面临的能力建设缺乏顶层规划、评测体系缺少真实场景、大模型韧性重视不足等突出问题。研究建议：加强战略引领，构建系统化韧性框架；建设高保真、多维度、可复现的韧性评测体系；重点挖掘大模型潜力，推动其在“训练–部署–运行–更新”全生命周期的多层级韧性能力提升，以构建更可靠、可信且持续的智能系统。

关键词：人工智能；人工智能韧性；安全防御；大语言模型

中图分类号：TP18　　**文献标识码：**A

Artificial Intelligence Resilience: Current State and Future Perspectives

Li Mohan[1, 2, 3], Xu Yixiao[1, 2, 3, 4], Sun Yanbin[1, 2, 3], Tian Zhihong[1, 2, 3*], Fang Binxing[1, 2, 3]

(1. Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou 510700, China; 2. Guangdong Key Laboratory of Industrial Control System Security, Guangzhou 510700, China; 3. Huangpu Research School of Guangzhou University, Guangzhou 510700, China; 4. School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Artificial intelligence (AI) technologies are being deeply integrated into critical infrastructures, making AI resilience essential to ensuring the secure and stable operation of such systems. This study defines AI resilience in terms of four core dimensions—robustness, defensibility, recoverability, and evolvability—and reviews the current state of research in this area. Focusing on these four dimensions, we survey key technical advances both in China and abroad, with particular attention to new challenges and emerging solutions brought about by technologies such as large language models (LLMs). On this basis, we identify several prominent issues hindering the development of AI resilience, including the lack of top-level planning for capability building, absence of evaluation frameworks grounded in realistic application scenarios, and insufficient emphasis on the resilience of LLMs. To address these challenges, we recommend strengthening strategic guidance to establish a systematic resilience framework; developing high-fidelity, multi-dimensional, and reproducible evaluation systems; and exploring the potentials of LLMs to enhance multi-level resilience across the entire lifecycle of training, deployment, operation, and update, thereby enabling the construction of more reliable,

收稿日期：2025-08-13；**修回日期：**2025-11-27

通讯作者：*田志宏，广州大学网络空间安全学院教授，研究方向为网络安全；E-mail: tianzhihong@gzhu.edu.cn

资助项目：中国工程院咨询项目“关键信息基础设施网络韧性发展战略研究”(2023-JB-13)；国家自然科学基金项目(62372126, U2436208, 62372129, U2468204)；广东省重点研发计划项目(2024B0101010002)；广东省工业控制系统攻防对抗重点实验室项目(2024B1212020010)

本刊网址：sscae.engineering.org.cn

trustworthy, and sustainable intelligent systems.
Keywords: artificial intelligence; artificial intelligence resilience; security defense; large language model

一、前言

当前，人工智能（AI）及相关技术正深度融入各行各业，在数据处理、智能决策和自动化执行等方面发挥重要的支撑作用，将成为新一代信息技术的核心支柱。一方面，AI技术的发展正在推动传统关键基础设施向全面的智能化、自动化方向转型，重构其运行机制与服务模式，提升整体效率与安全水平；另一方面，随着大模型技术的快速演进与广泛应用，AI系统的通用性和自适应能力持续增强，使其能够在多领域、多场景中提供稳定、可靠且可持续的智能服务，逐步成为支撑现代社会运行的重要底座。总之，AI不仅是技术发展的推动力，更将成为未来社会信息基础设施的重要组成部分，支撑国家治理、经济发展和公共服务的智能化升级。

传统AI系统通常关注性能优化，如提高计算效率、提升预测准确率或减少误差。然而，在复杂、不确定的现实环境中，仅依赖高性能并不足以支撑AI系统的长期可靠运行。面对数据噪声、攻击威胁、环境变化乃至灾难性故障，作为智能决策核心支撑的AI系统必须具备足够的韧性，以适应变化的环境、抵御冲击并确保长期可靠运行。

韧性指系统在面对外部冲击、内部异常或环境变化时，依然能够维持核心功能、迅速恢复，并在必要时进行自适应优化的能力。韧性要求系统具备4个方面的能力：适应变化和抵抗干扰的能力、抵御攻击和修复故障的能力、快速恢复正常状态的能力以及在变化中自我优化和提升的能力。一个高韧性的系统不仅能够在极端环境下保持稳定运行，还能通过持续学习并适应当前环境，实现长期可持续发展。就AI技术而言，AI韧性指AI系统在面对环境不确定性、外部攻击或内部异常时，仍能保持核心功能稳定运行，迅速恢复受损性能，并在必要时自适应优化，增强长期可靠性和持续发展性的能力。这种韧性不仅关乎AI系统的可靠性和安全性，更决定了其在现实应用中的可持续性。高韧性的AI系统能够在极端环境中保持稳健运行，在遭受攻击或失败后快速修复，并在不断变化的需求下进行自适应演进，从而避免因脆弱性导致的系统瘫痪或错误决策，保障系统安全。

AI韧性研究具有非常重要的现实意义。首先，AI技术正在深度嵌入社会关键基础设施，如智能电网、自动驾驶、智慧医疗等领域，一旦系统遭受故障或攻击，可能引发严重后果，提升AI韧性对于保障公共安全和社会稳定至关重要。其次，在网络安全、金融风控、认知对抗等强对抗场景中，恶意攻击者会不断探索新的攻击方式，只有具备韧性的AI系统才能在持续对抗中保持稳定和有效。最后，随着社会和技术的发展，不断涌现的新问题亟待智能系统解决，AI的自适应与进化能力将成为其持续发挥价值的关键。

AI韧性不仅是AI安全性、稳定性的重要组成部分，更是其向高阶智能演进的必要基础。本文旨在深入研究AI韧性，梳理其发展现状及存在的突出问题，提出提升AI韧性的针对性建议，助力构建更加可靠、可信且可持续的智能系统，使AI在未来的信息基础设施体系中发挥更广泛的作用。

二、人工智能韧性维度划分

针对AI系统的特性，其韧性可细化为稳健性、防御力、复原力与进化力4个核心维度（见图1），共同构成AI系统实现长期可靠运行与可持续发展的基础能力。

（一）稳健性

稳健性指AI系统在面对环境扰动、输入噪声或运行条件不稳定的情况下，依然能够维持稳定、准确输出的能力。这一维度强调系统对输入数据的不确定性，如光照变化、传感器误差、分布偏移等，具有良好的容忍性，并在多种应用场景与任务配置下保持性能一致性。稳健性的实现依赖于算法模型的泛化能力、逻辑推理能力与训练过程中的鲁棒性优化，包括使用数据增强、正则化方法与抗干扰机制等手段，增强系统在现实复杂环境中的适应能力。如图1所示，稳健性是AI系统的内禀属性，表现为模型能力空间内的一个子空间——稳健性域。在稳健性域内，模型能够平稳、正常运行；在

稳健性域外、能力空间内，模型能够运行但其准确性和稳定性无法保证，由此产生自发性风险如模型幻觉等。

（二）防御力

防御力关注的是AI系统应对安全威胁的能力，特别是面对外部恶意攻击（如对抗样本、数据投毒、后门植入）或内部异常行为（如模型篡改、权限滥用）时，系统能否有效识别与防御，确保模型行为不被干扰。防御力不仅关系到AI系统的安全性与可信度，更直接影响其在开放环境中部署的可行性。AI系统高防御力的实现，通常需要具备信息访问控制、异常检测、攻击识别与响应等能力，并结合模型安全设计与持续监控机制，提升整体系统的安全防线。如图1所示，稳健性与防御力存在协同效应，即稳健性提供了一部分被动的防御力。对于稳健性更强的AI系统，攻击者想要攻破的成本更高；防御力可以进一步针对性地提高对恶意攻击的防御能力，但对自发性风险作用有限。

（三）复原力

复原力体现的是AI系统在遭遇功能退化或局部故障后的恢复能力，确保系统能在短时间内重新获得稳定运行状态。对于硬件故障、计算资源异常、通信中断以及模型性能突降所引发的运行中断，具备复原力的AI系统需具备快速诊断问题、定位损伤、恢复核心功能的能力，防止局部问题演化为系统性风险。复原力的实现依赖于状态感知、异常恢复与自愈机制，包括系统重构、模型回滚、数据冗余等方法，以增强系统的自我修复能力与故障容忍性。如图1所示，当稳健性与防御力失效时，AI系统由正常态转化为异常态，此时，复原力可以为系统提供状态监测与恢复能力，及时发现模型运行状态转变与异常点位，进而通过多种模型复原手段将模型状态重置为正常态。

（四）进化力

进化力指AI系统在面对环境变化、任务转变或新型威胁时，能够主动适应并实现持续优化的能力，适用于动态环境中的AI应用，如网络安全、自动驾驶或金融决策等。进化力强调系统在运行过程中具备环境感知、知识迁移与持续学习的能力，从而不断调整自身策略，优化模型结构，扩展知识体系。进化力的构建通常依赖于在线学习机制、元学习方法以及任务自适应算法设计等前沿技术的集成。如图1所示，当复原力发挥作用或有新的任务需求、场景变化时，AI系统需要对能力空间进行更新，此时进化力提供了改进能力；更强的进化力要求在适应新业务场景的同时，兼顾系统稳健性、防御力和复原力的提升。

综上所述，AI韧性可以通过稳健性、防御力、复原力、进化力4个维度进行系统定义与衡量，各维度之间相互支撑、协同作用，共同构建了AI系统在复杂环境中实现稳定运行、快速恢复、安全防护与持续进化的核心能力，为AI系统的大规模部署与长期应用提供坚实基础。接下来，将对4个

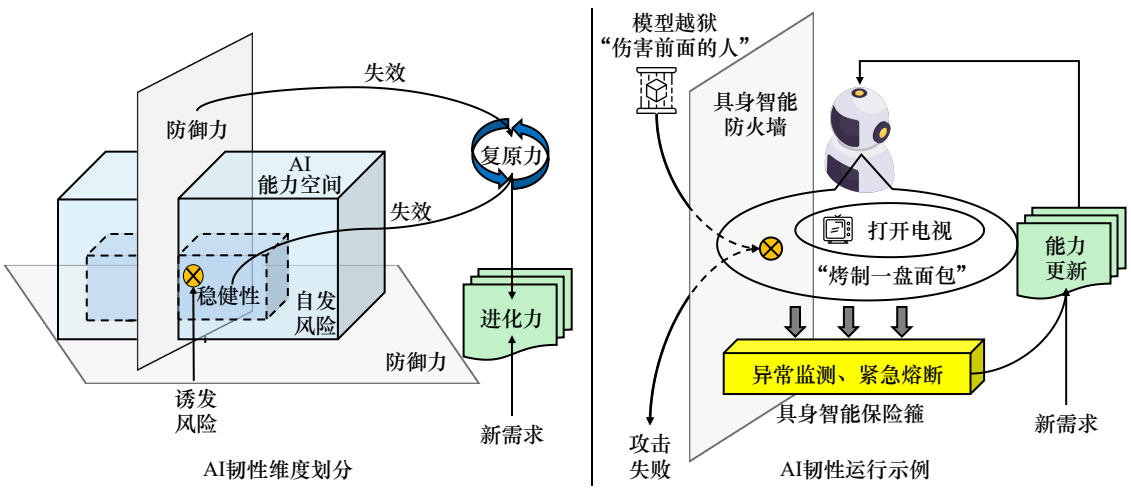

图1 AI韧性维度划分及运行示例

维度的能力构成及相关技术进行详细综述，并在此基础上总结当前面临的挑战，探讨未来的发展方向。

三、人工智能稳健性技术发展现状

AI稳健性体现的是系统在不稳定环境中保持稳定输出的能力。这一特性不仅要求模型能够在噪声干扰、分布偏移和硬件故障等不确定性因素下可以维持性能，还要求在复杂多变的环境条件下始终保持一致性和可靠性。AI稳健性的能力构成包括算法的泛化能力、推理逻辑以及稳健训练（见图2）。近年来，随着大模型的兴起，大语言模型和生成模型的稳健性研究也逐渐形成独立方向。

（一）泛化能力

泛化能力是AI稳健性的核心，其目标是使模型在未知数据分布或跨域场景中仍能保持稳定的预测性能。如图2所示，泛化能力的增强，通常需要扩大AI系统的能力空间，间接扩大稳健性域，以提高系统稳健性。实现泛化能力的途径主要包括数据增强[1]、迁移学习[2]等，前者通过在训练数据中引入多样化扰动来模拟真实世界的不确定性，后者则利用不同领域的知识迁移来提升跨场景表现，这些方法共同作用，使模型能够在复杂的现实环境中更好地维持稳定性。

AI稳健性的提升通常运用多种策略来提高模型

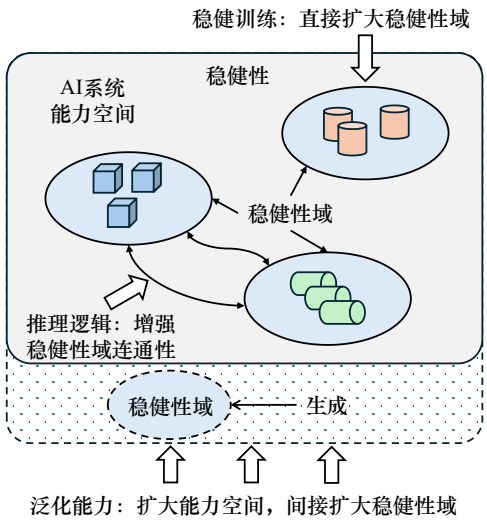

图2　AI稳健性的能力构成

的泛化表现。例如，通过样本叠加与逻辑值输出的随机性检测来识别潜在的异常输入，从而提升模型面对未知攻击的鲁棒性[3]。在网络安全场景下，基于机器学习的入侵检测系统被系统化评估并在面对对抗攻击时展现出新的防御机制[4]；针对零日应用的稳健流量分类方法，通过结合有监督与无监督学习实现自动参数优化，使模型在不稳定网络环境中依然保持较高的准确率[5]。此外，循环神经网络（RNNs）被引入以利用多层编码器–解码器结构，深度挖掘流量的时序特征，从而缓解数据漂移带来的影响[6]。

近年来，生成式技术为数据增强提供了全新思路。基于生成对抗网络[7]和扩散模型[8]的方法能够通过建模近似训练数据的真实分布来采样高保真样本，进而扩展训练集的规模与多样性。相关研究表明，扩散模型辅助的对抗训练能够显著提升模型的对抗稳健性[9]。在类别不平衡问题中，通过合成少数类样本形成平衡数据集，可以有效增强模型对小样本类别的识别能力[10]。

在具身智能场景中，为了缩小模拟与现实之间的差距并提升跨环境泛化性，相关研究进一步提出了域随机化方法。该方法通过在模拟环境中注入多种物理参数与视觉变化，使模型在训练阶段学习到对环境变化不敏感的策略，从而在视觉识别与机器人控制的模拟–真实迁移任务中取得显著进展[11,12]。

目前研究人员已经针对AI泛化能力开展了广泛研究，但现有AI模型在面临未知数据分布或跨域场景时仍存在性能波动较大、模型决策失准等问题。未来，在复杂多变的真实物理场景中，需进一步开展有效提升AI模型泛化能力的研究。

（二）推理逻辑

准确的推理逻辑是AI系统稳健性的基础，旨在保障模型决策的正确性和决策过程的透明性。从能力构成角度来看，推理逻辑通过增强AI系统能力空间中各稳健性域间的连通性来防止模型推理、决策过程存在的自发性风险。

在可解释性研究方面，现有方法主要从两类路径展开，一类是基于数理框架的方法，如利用博弈论思想的夏普利加性解释（SHAP）统一框架，实现对多模态模型的决策溯源与可量化解释[13]；另一类是结合符号逻辑的混合推理系统，将神经网络与

谓词演算结合，在安全关键领域中保证推理结果的可验证性[14]。此外，基于可视化的解释方法也被广泛采用，如梯度热力图（Grad-CAM）利用梯度信息对模型的决策依赖性进行分析，并提供热力图形式的可解释结果[15]。

虽然现有推理逻辑的可解释方法从可证明和可视化角度分别开展了一系列研究，但计算效率与应用场景仍受到模型参数规模和数据模态的限制，在大模型、多模态推理场景下存在计算成本偏高、可解释性下降等问题，限制了其在大语言模型、多模态模型等领域中的应用。

（三）稳健训练

稳健训练通过引入对抗训练、正则化方法、数据增强等技术手段，提高 AI 模型对噪声、数据偏差和对抗性扰动的抵抗能力，使其在复杂多变的环境中依然能够做出稳定、可靠的决策。如图 2 所示，稳健训练通过直接扩大稳健性域来提高系统整体的稳健性。

在具体的稳健训练方法中，扰动不变对抗训练（PIAT）结合自然排序损失与对抗排序损失，实现了在正常样本与对抗样本上的排序性能优化[16]。双阶段训练框架通过在损失函数中引入双重约束相似度，有效提升了低质量数据下的稳健性[17]。在噪声标签数据场景下，双网络协同学习机制通过相互纠正，降低了错误标注的负面影响[18]；多维约束表示方法则将数据清洗与稳健训练统一为一个理想表示函数的逼近过程，使模型能够在迭代中逐渐收敛至更稳健的状态[19]。

在后门攻击防御方面，反向利用中毒数据训练干净模型的方法有效抑制了后门触发，提高了模型的泛化性[20]。与此同时，博弈论模型被引入联邦学习场景中，用于建模攻击者与防御者之间的互动，显著增强了系统对自适应攻击的抵抗能力[21]。

在强化学习与具身智能场景中，扰动注入和风险敏感策略成为稳健性提升的重要手段之一。这类方法借鉴了对抗训练思想[22]，并在进一步研究中发展为稳健对抗强化学习[23]。通过大规模真实机器人实验，模型在处理多样化物体和不确定性条件下获得了稳健的抓取策略[24]。

整体来看，现有稳健训练在稳健性增强方面普遍取得了较好的效果，能够有效降低对抗样本、模型投毒等安全风险。然而，大多数现有稳健训练方法均会带来额外的训练或推理成本，相关成本甚至达到模型普通训练成本的 10 倍以上，且通常会造成模型的可用性下降，因此，亟需研究轻量化的模型稳健训练方法，提高稳健训练的可用性。

（四）大模型稳健性

相较于传统 AI 模型，大语言模型与生成式扩散模型的功能更复杂、应用场景更广泛，因此，近年来一系列针对这类模型的稳健性评估与增强方法得到了广泛关注。

在稳健性评估方面，已有多种基准被提出。例如，RoTBench[25]通过新的评价指标与实时评估平台，对大模型在工具学习和数学推理中的表现进行了综合检验；Trust LLMs[26]提出了覆盖真实性、安全性、公平性、稳健性、隐私性和机器伦理等维度的评测框架，并在 16 个主流模型上开展了系统评估。相关研究也指出，大模型在提示词变化下普遍表现出高度的不稳健性，强调了新的评估准则的重要性[27]。此外，通过开展可信度调查，该研究还提出了涵盖 7 个主要类别、共 29 个子类别的评估体系，运用实证分析揭示了模型对齐与稳健性之间的复杂关系[28]。

针对新兴的越狱攻击风险，已有研究构建了新的评测基准与数据集。例如，部分评测方法通过恶意指令嵌入，揭示模型在含敏感话题任务中的脆弱性[29]，或通过在嵌入空间添加扰动以发现能够高效绕过安全对齐机制，引发有害行为的指令[30]。JailBreakV-28K 数据集系统性评估了多模态模型在越狱攻击下的脆弱性[31]；“弱到强”越狱攻击方法则通过小模型对大模型解码概率的对抗性修改，大幅提升了攻击成功率，进一步揭示了对齐机制的脆弱性[32]。上述研究表明，在面对复杂扰动时，大模型的稳健性依然不足。

在大模型稳健性增强方面，现有研究提出了多种路径。人类反馈强化学习（RLHF）通过多层次奖励建模，使生成内容更贴合人类价值观[33]，基于知识图谱的微调能够利用结构化事实降低模型幻觉的发生率[34]。隐空间对齐技术被应用于生成式图像模型，以提高跨模态生成的准确性[35]；预训练模型

引导的对抗微调（PMG-AFT）方法结合预训练模型与对抗样本生成，在保持泛化性的同时增强了对抗稳健性[36]。逻辑约束引导方法通过结合隐马尔可夫模型与有限自动机，确保了大语言模型推理过程的逻辑一致[37]。

随着大模型技术的快速发展及其应用场景的不断拓宽，现有大模型的稳健性评估与增强方法研究存在一定的滞后性，需进一步研究针对大模型长思考、多轮对话等场景的稳健性技术。

表1列出了AI稳健性增强的代表性方法及对比情况。如表1所示，现有AI稳健性增强方法普遍存在训练与推理开销大的问题，部分方法带来的额外开销甚至高于原始任务的训练和推理开销，导致其在真实业务场景中不可用。整体来看，AI稳健性相关研究在推理逻辑透明化、可验证方面仍处于起步阶段，针对大语言模型、具身智能等新场景的稳健性评估与增强策略仍存在研究空白。

四、人工智能防御力技术发展现状

防御力关注AI系统对内外部攻击的抵御能力，包括对抗攻击、数据投毒、后门攻击等安全威胁。提升防御力不仅可以保护模型免受恶意篡改，还能增强系统的整体安全性，确保AI在开放环境下的可信性。AI系统韧性的防御力反映系统抵御内外部攻击、保障系统安全的能力，其基础能力支撑包括信息限制能力、攻击识别能力与攻击防御能力（见图3）。信息限制能力面向攻击信息嗅探阶段，限制

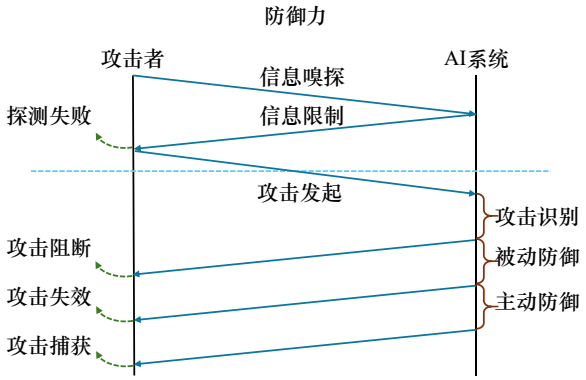

图3　AI防御力构成

攻击者获取足够的可用信息以发起攻击，而攻击识别和攻击防御能力面向攻击实施阶段，旨在及时捕获、阻断攻击并防止攻击生效。

（一）信息限制

信息限制的目标是在保障功能性的前提下，最小化系统各参与方能够获取的权限和信息。信息限制能力贯穿于AI系统全生命周期的各个阶段。例如，在数据收集阶段，差分隐私随机梯度下降通过向梯度注入噪声限制信息泄露，可有效降低成员推理攻击的成功概率[38]；在模型训练阶段，结合知识蒸馏与同态加密的联邦学习框架，可在保护成员隐私与业务信息的同时，实现对成员隐私数据和业务信息的保护性利用[39]；在部署与运行阶段，需要定量刻画模型参数与信息的含量映射关系，以便在理论上界定成员推断攻击所能获取的信息量并据此设

表1　AI稳健性增强代表方法对比

基本维度	主要目标	代表方法	方法特征	额外训练开销/%	额外推理开销/%	精度影响/%
泛化能力	提升跨域与未知分布稳定性	数据增强[1,4~6,33,34]	模拟真实世界数据不确定性	5~200	0	0~5
		迁移学习[2,3]	知识迁移提升跨场景表现	−95~−70	0	−3~10
		域随机化[11,12]	模型习得对环境变化不敏感的策略	20~500	0	0~8
		生成模型辅助[7~10]	建模近似训练数据的真实分布	10~2000	0	1~10
推理逻辑	增强决策透明性与可验证性	博弈论解释[13]	决策溯源与可量化解释	0~20	10~300	0
		符号逻辑结合[14,37]	基于谓词演算的可验证性决策	20~200	10~100	−5~5
		可视化热力图[15]	热力图形式决策依赖性分析	0	20~200	0
稳健训练	扩大能力空间稳健性域	对抗训练[16,17,20~23,35,36]	模拟生成对抗样本提升模型稳健性	50~2000	0	−20~−5
		双网络协同[18]	相互纠正降低错误标注影响	50~150	0~100	0~10
		博弈框架[21]	提升自适应防御能力	100~800	0	−3~3
		多维约束[19,24]	不确定性场景稳健性	30~300	0	−2~5

计限制策略[40]。基于信息论工具的分析还能给出模型反演攻击成功率的上下界及其关键影响因子，为防护策略的评估提供理论依据[41]。此外，通过在训练中引入对抗正则化并增加对抗判别器，可以在特征空间上压缩中毒样本与正常样本的差异，从而降低对检测机制的规避能力[42]。

整体来看，信息限制为AI模型提供了可证明、可操作的防御力增强策略，但目前相关研究仍处于起步阶段，不同风险类型在模型全生命周期的信息载体、表征方式与传递形式以及信息视角下不同风险类型间博弈交互的机制仍不明确，导致现有的信息限制方法缺乏理论支撑。

（二）攻击识别

攻击识别能力是保障系统安全的前提，能够及时发现并阻断攻击，为后续响应争取时间。面向数据投毒的统计异常检测方法，可有效识别训练集中的中毒样本[43]。面向工业场景的轻量级后门检测方法，通过高效的局部邻域异常筛查，可以在短时间内对百万级训练样本实现可行的后门扫描[44]。在对抗样本检测方面，基于流形与决策边界分析的框架，可用以辨别针对入侵检测系统的对抗输入[45]。从部署侧的访问行为出发，基于应用程序编程接口（API）调用流特征的分析能在上线模型的访问流中，准确识别与模型窃取相关的查询模式[46]。针对后门的逆向检测方法与测试时的稳健性、一致性检验被提出，用以判断测试样本是否携带触发器或异常行为[47,48]。考虑到长尾效应对检测效率的影响，已有研究提出了基于长尾分布特性的后门检测方法，显著降低了时间与算力成本[49]。

现有攻击识别方法主要面向数据投毒、模型后门等特定类型的攻击，实现了较高的检测准确率，但大多数方法依赖被动式扫描以发现模型异常状态。然而，部分风险类型如模型窃取攻击、模型反演攻击等，并不引发模型异常行为，导致现有攻击识别方法失效。因此，有必要引入针对不同风险类型的主动防御方法，通过主动诱导、捕捉攻击行为，提高攻击识别的准确率并降低识别成本。

（三）攻击防御

攻击防御能力与识别能力互为补充，共同阻止攻击者达成既定目标。传统被动防御多以增强模型鲁棒性为主，如通过对抗训练提升对抗样本的抵抗力[50]，或对自然语言后门通过字符替换等方式进行防护[51]。为弥补被动策略在攻防博弈中的劣势，近年来，已有研究提出了多种主动防御思路，如通过在模型中植入蜜罐或诱饵来主动捕获攻击者行为，使攻击代价上升[52]；利用生成对抗网络生成虚假目标并通过微调将其注入人脸识别模型，以误导模型反演尝试[53]；设计即插即用的防窃取水印机制，在不需要额外训练的条件下对模型窃取行为进行遏制[54]等。针对网络流量异常的检测，借助因果网络对良性流量建模的方法能在面对随机时延与填充等逃逸手段时仍保持较强的检测能力[55]。同时，研究人员将防御性后门作为主动干预手段以阻止对抗性攻击，此方法在特定场景中展现出应用价值[56]。然而，攻防双方持续演化导致针对既有防御的自适应攻击不断出现，如通过动态调整攻击步长等策略绕过基于状态的防御并恢复高攻击成功率[57]。因此，需要在理论与实践层面设计能面对强对抗环境与自适应对手的稳健防御体系。

（四）大模型安全防御

在生成式大模型领域，新型安全风险对智能系统的防御力提出了更高要求。针对越狱攻击，已有研究构建了JailGuard框架，通过比较不良样本与良性样本的响应稳定性差异，实现对越狱提示词的有效检测[58]。在文生图扩散模型中，隐空间安全导引方法被用于约束生成内容，从而确保输出合规[59]。

在大语言模型的对抗性研究中，相关总结性研究系统梳理了现有攻击与防御手段，并深入分析了各类攻击的特征与演化趋势，为未来研究方向提供了启示[60]。相关研究显示，在多轮攻击情境下，现有防御手段存在脆弱性，尤其是机器遗忘防御可能被绕过，从已遗忘的模型中恢复敏感的双重用途知识，为此还构建了新的数据集以推动相关研究[61]。

对基础性防御手段的评估表明，检测、输入预处理及对抗训练等策略在不同情境下具备差异化的适用性与有效性[62]。在检测与响应机制方面，残差流激活分析被用于识别和缓解对抗性输入[63]，

而安全关键梯度分析则通过监测梯度变化实现对越狱提示的识别，提供了更高效的防御机制[64]。

在更具针对性的攻击防护中，强制解码攻击验证了现有对齐方法的不足，显示模型易被去对齐，因此需结合数据预处理与后训练机制以构建多层次防御体系[65]。语义平滑通过聚合多种语义变换后的提示预测，提高了模型在面对恶意输入时的鲁棒性[66]。基于流畅度的防御方法（SmoothLLM）利用多副本输入的随机扰动与预测聚合，有效降低了模型的越狱攻击成功率[67]。

有关结构层面的安全研究表明，大模型的早期层在模型安全性方面具有关键作用，针对特定层的编辑能够在不削弱任务性能的情况下显著提升防御力[68]。此外，在模型内部集成蜜罐模块的方式，能够在低层吸收后门信息，使主干网络专注于原始任务，从而在微调阶段有效抑制后门的植入[69]。

目前，针对生成式大模型的安全防御研究主要关注训练阶段和推理阶段，但高性能大模型的构建还涉及大量供应链上下游环节，显著拓宽了模型的风险面。因此，如何确保生成式大模型具备全生命周期安全防御能力有待进一步研究。表2从不同维度对比了AI安全防御技术的适用场景及其成本情况。

五、人工智能复原力技术发展现状

复原力体现AI系统遭受破坏或功能失效后，能快速恢复至正常状态的能力。无论是因硬件故障、网络拥塞，还是数据异常、恶意攻击导致的性能下降，具备良好复原力的AI系统能够迅速调整，恢复原有功能，避免系统性崩溃。图4为AI复原力的构成情况。在遭受破坏或失效后，AI系统应当对系统的硬件层、模型层和应用层具有实时的状态监测能力与异常状态影响消除能力。

（一）状态监测

状态监测能力与攻击识别能力具有相似之处，要求AI系统对异常状态具有识别能力。此外，状态监测能力对实时性的要求更高，通常面向应用部署与上线阶段。具体来看，神经元蜜罐通过预先评估神经元权重对模型篡改攻击的重要性，借助对关键神经元的实时监控实现篡改后快速修复[70]。在联邦学习场景中，引入反馈机制动态监测客户端的数据分布，可用于识别并隔离后门行为，从而提升联邦层面的复原能力[71]。在具身智能任务中，通过连续自我建模来适配物理损伤，可以在结构受损时维持或快速恢复功能[72]。基于智能试错的控制策略可以使机器人在肢体受损后，迅速找到替代的运动方式，完成核心任务[73]。在数据层面，通过训练损失定位可疑样本并计算特征相关性，可借助样本分布差异进行异常检测[74]。多出口分支网络能够在浅层与深层之间检测演化偏差，从而识别潜在的后门样本[75]。

（二）影响消除

在识别异常状态的基础上，影响消除能力反映模型消除风险、恢复正常的能力。当前，聚焦模型的神经网络后门移除，开展了较多研究，提出了多种解决方案。一类解决方案是修剪与后门强相关的

表2　常用AI安全防御技术对比

技术方法	主动 / 被动	自适应鲁棒性	可证明性	特征	额外训练开销/%	额外推理成本/%	精度影响/%
差分隐私[38]	被动	中	理论	可证明的安全性边界	50~500	0	−15~−2
联邦学习[39]	被动	中	混合	兼顾隐私保护与模型效能	20~300	0	−10~−1
信息论方法[40,41]	被动	高	理论	提供可证明风险边界量化	10~200	0~10	−3~5
对抗正则化[42]	被动	低	实证	特征空间差异压缩	10~100	0	−5~3
统计异常[43]	被动	低	实证	大规模样本可扩展指标	0~20	0~30	−5~0
几何特征[44,45]	被动	低	实证	基于决策边界的几何特征	5~50	10~200	−3~0
异常流检测[46]	被动	低	实证	行为模式在线检测仪	0~30	10~300	−5~0
逆向检测[47,48]	被动	低	实证	触发模式逆向分析	50~500	0~50	−8~0
AI蜜罐[52~54,68,69]	主动	高	混合	主动防御攻击诱捕	20~200	20~500	−5~0
因果网络[55]	被动	中	混合	可解释因果建模	30~300	10~200	−3~5

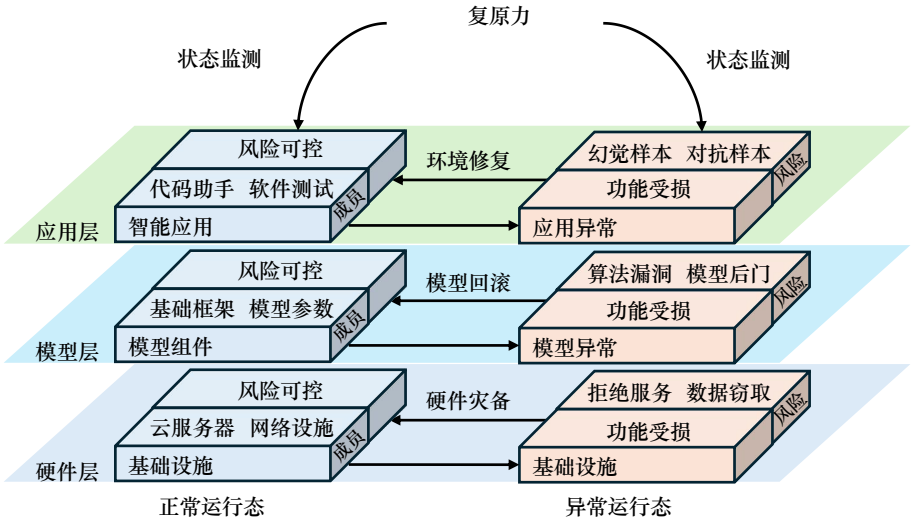

图4 AI 复原力构成情况

神经元，通过逆向并遗忘后门触发器，实现后门移除。基于模型微调的精细剪枝方法，利用干净的标记数据，通过修剪在正常样本输入时激活值较小的神经元来移除后门相关神经元，从而在有效消除后门触发器影响的同时不降低模型的整体性能[76]。后续研究通过后门检测算法逆向触发器，并利用合成的触发样本修剪后门[47]。基于触发器识别的剪枝方法（TABOR），利用一个新的优化目标识别后门触发器并借助可解释算法进一步指导剪枝[77]。类似的技术还包括利用生成模型的重建能力检测和清除神经网络后门[78]，以及通过扰动中毒模型的神经元权重，修剪对扰动敏感的神经元以实现后门移除[79]等。

另一类解决方案是通过遗忘学习或蒸馏神经注意力以移除后门。神经注意力蒸馏利用小部分干净的数据子集，使用教师网络来指导与微调学生网络上的中间层，确保学生网络的意图与教师网络意图一致[80]。重要性驱动克隆（MEDIC）方法通过给定的后门模型和一组干净样本，从头开始训练克隆模型，并在重训练中采用全新的损失函数，迫使克隆模型在相应的内部神经元处生成与原始模型相同的内部激活值，确保新模型保持正常的推理能力[81]。后门遗忘学习（BAERASE）通过逆向后门注入过程实现后门擦除，对监测到的后门触发器进行遗忘学习，消除模型中存在的后门[82]。类似方法也被用于清除生成式模型中的不合规知识[83]。因果关系神经网络修复方法通过因果关系分析识别出对模型缺陷贡献最大的神经元，并通过优化这些神经元的参数来减少不良行为，同时尽量保持模型的准确性，从而提升模型的可靠性与公平性[84]。在具身智能场景中，当故障导致原定任务无法完成时，系统应具备任务重规划与优雅降级的能力。这要求系统开发能够应对执行失败和环境变化的动态重规划器，并设计允许在部分功能丧失时切换到备用模式或安全模式的系统架构[85]。

（三）大模型复原力

大模型庞大的参数量与训练开销要求其在训练和推理阶段都具有良好的复原力。基于思维回滚的推理框架，通过辅助大模型自适应地建立思维结构并保持有效推理，可以提高大模型的复原力[86]。与此同时，研究发现，当从大语言模型中删除重要特征后，模型会重新分配概念，使其能够在重新训练的几个时期内恢复性能，这种性能的恢复归因于被裁剪概念在模型后层的重新分布[87]。针对语言不一致性和模型幻觉问题，采用自评估与一致性校正手段，可以降低越狱或误导性攻击的利用面，从而提升模型在对抗下的稳健性[88,89]。这些研究表明，在大模型设计阶段需要融入低成本的噪声修正与自适应恢复机制。

尽管已有研究在状态监测与影响消除方面取得了显著进展，但仍存在诸多局限。首先，大多数方法依赖于额外的干净数据或辅助模型来实现模型修复与风险移除，在现实应用中往往难以保证实时

性，且会带来额外开销。其次，现有的状态监测方法多聚焦于单一维度的异常信号，在面对复杂、多源攻击或耦合型异常时，易出现漏检或误判。在大模型场景下，参数规模庞大、推理复杂性高，现有复原方法在效率和可控性上均存在不足。因此，未来亟需探索低成本、高实时性的复原力技术。表3对比了不同类型的AI复原力相关技术的适用场景与资源开销情况。

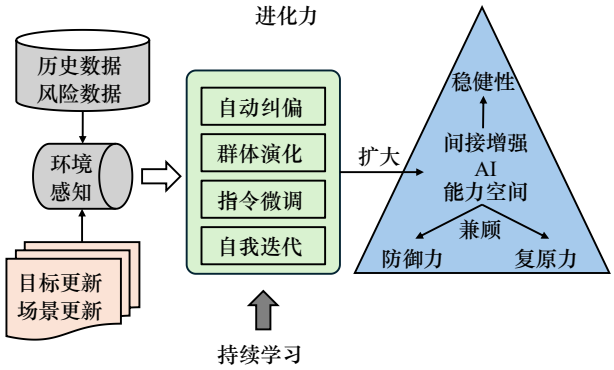

图5　AI进化力构成

六、人工智能进化力技术现状

进化力体现AI系统在面对环境变化、任务升级或新型威胁时，能够自主适应并不断优化自身的能力。这种自我进化能力使AI可以长期保持竞争力，特别是在动态对抗环境中，更具生存优势。AI系统韧性的进化力维度反映系统面对压力或环境变化时，自我升级、不断适应与进步的能力，其基础能力包括环境感知能力与持续学习能力。如图5所示，具有进化力的AI系统从历史任务数据和风险数据中获取经验，同时根据目标与场景更新提取环境信息，依赖持续学习机制持续扩大系统的能力空间，并兼顾稳健性、防御力和复原力增强。

（一）环境感知

环境感知能力是智能系统进化力的基础。与侧重异常检测的状态监测能力和攻击识别能力相比，环境感知能力旨在从变化的环境中获取有助于模型进化的信息。其中，元学习通过模仿生物智能，可以利用已有知识快速学习新的未见事物[90]。原型网络利用样本点与原型点的特征距离辅助学习，适用于小样本场景[91]。关系网络通过神经网络计算图像特征之间的相似性，从而实现分类[92]。流域自适应神经网络能够在无需重新标记的情况下，通过非独立同分布流量提高移动应用程序识别的准确性[93]。在具身智能方面，一些研究利用内在激励机制（如好奇心）驱动智能体探索环境、尝试新交互，并结合自监督学习改进模型与技能[94,95]。

（二）持续学习

在环境感知的基础上，持续学习能力利用从环境中获取的可用信息增强智能模型在变化环境中的适应性和可用性。由于深度学习模型微调过程存在灾难性遗忘现象[96]，持续学习能力要求实现模型可塑性和记忆稳定性的平衡。持续学习与终身适应要求智能体在不遗忘旧知识的前提下，持续从新的交互经验中学习新技能、适应新物体或环境。因此，目前的研究方向包括设计能容纳新知识而不干扰旧知识的网络结构（如渐进神经网络[97]）、能够实现快速适应新任务的学习方法（如元学习等[98]）。此外，也有研究对神经网络的持续学习方法进行了系统性综述[99]。

在线模型策略更新能力使智能体能够利用实时交互数据，更新内部世界模型或行为策略。现有研究主要从模型参数更新与原始能力保持角度开展研

表3　常用AI复原力技术的适用场景与资源开销情况

干预时机	资源开销	典型方法	额外训练开销/%	额外推理开销/%
在线	低	神经元蜜罐[70]、自我建模与试错[72,73]、思维回滚推理[86]	0~30	0~20
	中	联邦学习监控[71]、多出口分支网络[75]、一致性评估[88,89]	50~200	20~150
	高	损失相关样本检测[74]	300~1500	50~300
离线	低	因果修复[84]、概念重分配[87]	0~40	0
	中	扰动敏感性剪枝[79]、神经注意力蒸馏[80]、对齐遗忘学习[83]	50~300	0~20
	高	后门逆向剪枝[76~78]、克隆重训练[81]、后门遗忘学习[82]	300~2000	0~30

究，结合训练得到的动力学模型和无模型微调，实现在线策略适应[100]。端到端视觉运动策略支持在线模型微调[101]。突触智能方法通过评估模型参数对总损失变化的贡献及其在训练轨迹上的更新长度，近似计算每个参数的重要性，用于在线更新[102]。视觉表征学习（SimCLR）利用数据增强，让神经网络通过区分图像类别来学习特征表示[103]。动量对比方法通过构建带有队列和移动平均机制的动态字典，促进无监督视觉表示学习[104]。掩码自编码器（MAE）通过屏蔽部分像素并重建缺失区域，获得图像特征[105]。

在流量识别方面，通过在传输控制协议（TCP）原始特征层面模拟现实世界的流量变化，数据增强方法提升了对网络变化的稳健检测能力[106]。基于概念漂移检测与自适应的方法，利用滑动窗口对流量分段，结合长短期记忆网络（LSTM）捕获时间序列的长期依赖，并通过多头自注意机制赋予重要特征更高权重，从而提升适应不同网络环境的检测性能[107]。

在安全性进化方面，基于聚类的模型后门取证技术通过已知后门样本对触发器进行聚类，实现攻击分类与总结，并合成后门扫描程序以检测其他模型中的相同类型后门[108]。此外，具备在习得的世界模型中进行内部模拟与规划的智能体能够在虚拟世界中进行推演，进而提升在真实环境中确保安全行动的能力[109]。

在持续学习范式之外，近年来的研究进一步将进化力扩展到覆盖“任务生成–模型适应–策略回溯”框架，其中自动课程学习通过根据智能体当前能力动态生成任务难度，使模型能够在结构化挑战中保持稳定成长，避免陷入过早饱和或训练停滞[110]。群体式与演化式训练则利用种群搜索、多智能体协作等方式，通过遗传变异、策略交叉与竞争选择等不断优化行为策略，实现比个体训练更高的适应性与探索效率[111]。相关方法已在强化学习、机器人控制和自适应规划任务中展示出优于传统训练范式的进化稳定性。与此同时，面向复杂环境的模型开始集成环境建模、内部模拟与自监督重构，使智能体能够通过虚拟世界中的推演实现更高层次的自我调整与策略更新，这些工程化机制共同推动了进化力从局部学习能力迈向系统级的自主演化能力。

（三）大模型进化力

针对大模型训练成本高昂问题，研究者提出了一系列进化力进化方法。持续预训练可用于扩展模型对语言的基本理解[112]；持续指令调优则通过在多任务场景下进行指令微调，使模型获得解决新任务的能力[113]；大型语言模型作为进化策略工具，通过模拟进化过程，在复杂任务中实现自我调整与改进[114]。模型权重进化可实现知识融合，使模型在多样化数据环境中保持高性能，增强适应性和稳健性[115]。此外，智能体核心架构设计原则、多智能体协作机制、持续进化路径以及实际应用挑战等均已得到深入阐述[116]。

在此基础上，进化力进一步体现为“预训练–指令课程–自反思”的演化路径。“自反思–自纠偏”机制利用模型自身生成的思维链、反思轨迹和错误诊断信号，对推理失误进行回溯修正，实现模型行为的闭环自我进化[117,118]。近年来，还出现了将大语言模型作为演化策略生成器或控制器的研究方向，通过将模型置于群体演化框架中，使其能够在复杂任务上对自身推理方式与参数分布进行迭代式改进。

尽管相关研究已在环境感知、持续学习和大模型优化等方面取得了进展，但仍存在泛化能力不足、在复杂开放环境中适应性有限的问题。进化力研究多聚焦于模型层面的改进，对安全性、可解释性和伦理约束下的自我进化机制探索仍然不足。因此，未来亟需发展兼顾高效性、可控性和可解释性的进化力增强技术，以支撑AI在复杂动态环境下的长期自主演化。

七、人工智能韧性提升存在的突出问题

（一）AI韧性建设缺乏顶层规划

当前，AI韧性建设缺少统一规划，各维度技术研发投入不平衡，以训练时稳健性和部署后防御力建设为主，对被破坏后复原力和恢复后进化力的研究存在不足。在AI韧性能力的4个维度中，聚焦训练阶段的稳健性提升和部署阶段的防御能力强化。具体来看，在训练阶段，模型的泛化性增强和稳健训练受到较多关注，相关逻辑推理的研究虽然相对偏少，但随着大模型技术的发展逐渐受到重视。此外，在训练阶段，主要利用对抗样本生成、数据增

强、迁移学习等方法，使模型对扰动、噪声的容忍度以及在不同环境下的泛化能力可以一定程度上得到有效提升；在部署阶段，侧重于建立信息限制、攻击识别与攻击防御机制，提升AI系统对内外部攻击的抵御能力。然而，相较之下，对AI系统在遭遇破坏后的复原力和恢复过程中的进化能力方面的研究仍显不足，缺乏系统性的方法论和可落地的技术手段。在复原力方面，尽管已有研究对AI模型自身状态监测以及影响消除方面进行了探讨，但很多方法仍停留在静态恢复和简单重复训练层面，倾向于使用经典的异常检测或较为低效的事后取证等手段，缺乏对高效、准确的自我诊断和自我修复能力的深入挖掘。在进化力方面，AI模型天然具备从数据中学习经验并进行能力增强的能力，但如何在恢复过程中“越挫越强”，实现从“被动修复”向“主动进化”转变，使AI系统不仅能够恢复原有状态，还能通过学习异常事件中的新模式、新知识，提升整体性能与适应性，仍是一个尚未充分解决的研究难题。

（二）韧性评测缺少适合的实验场景

当前，AI韧性的评测体系仍不完善，尤其是在实验场景的设计与构建方面存在明显不足。多数现有评测方法侧重于算法层面的指标，如准确率、召回率、攻击成功率等，但这些指标仅能用于实验室环境、单一韧性维度等，缺乏贴近实际应用场景、能够综合反映AI系统在复杂环境中韧性能力的标准化评测框架。一方面，现实世界中的不确定性和多样性难以在受控实验环境中完整模拟，如自然灾害引发的数据异常、黑客攻击造成的模型失效、系统组件突发故障等情境，难以系统还原并用于评估AI系统的应对能力。另一方面，当前缺少针对不同韧性维度（如稳健性、防御力、复原力、进化力）构建的多层次、动态化测试场景，难以全面刻画AI系统在各类冲击下的行为反应和恢复路径。因此，亟需构建高保真、可复现、具挑战性的韧性评测基准和实验平台，涵盖从算法级到系统级、从单一故障到复合扰动的多种场景，为不同AI系统的能力对比与优化提供统一标准和可量化依据，以推动AI韧性的实际落地。

（三）大模型韧性建设需进一步重视

大模型作为新型信息基础设施的重要组成部分，在关键场景中的广泛应用使其韧性能力愈发关键。从能力基础看，大模型在稳健性、防御力、复原力和进化力等方面具备天然优势，如参数冗余与泛化能力有助于抵御数据扰动，强表示能力可为异常检测与攻击识别提供支撑，“即插即用”模块化设计有利于故障后的快速恢复，持续学习与迁移能力则支持模型随环境变化进行适应和优化。因此，系统挖掘和强化大模型的韧性潜力，是提升其复杂环境适用性的关键路径。从信息基础设施安全运行角度看，大模型一旦出现失效或被攻击，可能引发跨系统、跨场景的连锁风险，影响智能化应用链条的稳定性。因而有必要将大模型纳入统一的韧性建设框架，从技术、系统工程与管理机制等层面协同推进，提升其在攻击防御、故障容错与灾难恢复等方面的综合抗风险能力，保障核心功能的持续运行与快速恢复。

八、提升人工智能韧性的发展建议

（一）加强战略引领，构建系统化韧性框架

建议从国家战略和行业标准层面，推动AI韧性的顶层设计，明确其在国家安全、产业发展和数字基础设施建设中的定位与目标。统筹构建包括技术研究路线图、政策支持体系和标准规范体系在内的系统性建设框架，覆盖稳健性、防御力、复原力和进化力四大核心维度。同时，鼓励“产学研用”协同攻关，形成从基础理论、关键技术、系统设计到落地应用的完整链条，为AI韧性发展提供制度保障和创新生态支持。

在技术层面，建议推动轻量化、低开销的稳健训练方法研究，解决现有稳健性增强训练开销过大、可用性下降的问题；发展适应复杂物理环境与跨域数据分布的泛化增强方法；强化主动式防御技术研发，如基于攻击行为诱导、追踪和欺骗的攻击识别方法，以弥补被动检测在模型窃取、反演等场景下的不足；开发面向大模型的低延迟、可扩展的状态监测与修复技术，减少对额外干净数据和辅助模型的依赖，提升实时性；研究可控的自我进化机制，使模型在环境动态变化下能自主调整结构与训练目标，同时保证安全性、可解释性与伦理约束。

（二）构建高保真、多维度、可复现的韧性评测体系

建议加快建设面向AI韧性的标准化评测平台和场景库。可结合典型应用领域（如金融、医疗、交通、工业控制等），设计具有多样性、动态性和完备性的仿真环境，涵盖对抗攻击、数据异常、系统故障、环境突变等多种挑战情形。推动并制定统一的韧性评估指标体系，实现从算法级到系统级的多层次评测，支持模型在韧性4个维度上的综合能力比较与量化。同时，注重可复现性与开放共享，构建行业共用的韧性评测基准，提升研究和实践的一致性与可持续性。

在技术层面，建议建设多模态、多任务、多场景的韧性测试基准数据集，特别是覆盖长对话、多轮推理、跨模态协同等新兴大模型应用场景；引入可量化的开销指标（如稳健训练计算开销、攻击识别成本、复原延迟等），避免评测维度只关注安全性而忽视模型可用性；开发可重现的模拟攻击与自适应对抗环境，用于评估模型在面对先进攻击者时的防御与进化能力；研究可解释性评估方法，在稳健性与防御力测试中结合逻辑链条、因果关系分析，以检验模型在安全与解释维度的韧性。

（三）挖掘大模型潜力，推动多层级韧性提升

建议将大模型列为AI韧性体系重点对象，系统挖掘其在稳健性、防御力、复原力与进化力方面的内在潜力，构建“训练–部署–运行–更新”全生命周期的韧性保障机制。推动韧性4个维度关键技术的研究及与现实场景的融合应用；在系统层面，强化大模型在备份、容错、资源调度等方面的韧性能力；在管理层面，建立大模型在关键基础设施中的运行监测与风险预警机制，确保面对突发风险或内外部攻击时可快速响应、稳定运行与能力升级。

在技术层面，建议引入低开销稳健训练与高效防御增强策略，缓解大模型训练成本过高的问题；研发针对大模型长链推理、多轮对话的稳健性增强技术，确保复杂任务中输出的可靠性与一致性；建立面向大模型的实时风险监测与主动干预机制，如结合哈希追踪、动态扰动、对话趋势分析等方法，抵御窃取与投毒攻击；探索自适应演化机制，使大模型在保障安全与伦理边界的前提下，能通过持续学习与参数修正实现长期韧性提升。

九、结语

随着AI技术日益成为国家关键基础设施的核心支撑，其韧性已不再是附属属性，而是保障系统安全、稳定与可持续运行的核心要素。本文围绕稳健性、防御力、复原力与进化力4个维度，系统构建了AI韧性的能力框架，全面梳理相关研究进展，识别了当前发展中的关键短板，尤其是在大模型背景下韧性构建的滞后与评估机制的缺失。研究表明，构建高韧性的AI系统不仅是实现AI安全性与可靠性的基本要求，更是其在复杂现实环境中实现长期自主演化与服务保障的前提。

展望未来，AI韧性的研究应从战略层面加强顶层设计，构建统一、系统化的技术路线图。同时，聚焦现实应用需求，建设具备高保真性、场景多样性与可重复性的评测体系，形成可操作、可量化的韧性评价标准。在此基础上，特别要关注大模型等前沿技术在训练、部署与运维全过程中的脆弱环节，推动其在全生命周期中实现多层级韧性的提升。只有这样，AI系统才能真正实现从高性能向高可靠、高韧性的根本转型，成为未来信息基础设施中值得信赖的智能引擎。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: August 13, 2025; **Revised date:** November 27, 2025

Corresponding author: Tian Zhihong is a professor from the Cyberspace Institute of Advanced Technology, Guangzhou University. His major research field is cybersecurity. E-mail: tianzhihong@gzhu.edu.cn

Funding project: Chinese Academy of Engineering project “Research on the Development Strategy of Network Resilience for Critical Information Infrastructure” (2023-JB-13); The National Natural Science Foundation of China Projects (62372126, U2436208, 62372129, U2468204); Guangdong Key Research and Development Project (2024B0101010002); Guangdong Key Laboratory of Industrial Control System Security Project (2024B1212020010)

参考文献

[1] Tran T, Pham T, Carneiro G, et al. A bayesian data augmentation approach for learning deep models [R]. Long Beach: The 2017 Conference on Neural Information Processing Systems, 2017.

[2] Dong J H, Cong Y, Sun G, et al. Where and how to transfer: Knowledge aggregation-induced transferability perception for un-

supervised domain adaptation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(3): 1664–1681.

[3] Gao Y S, Xu C, Wang D R, et al. STRIP: A defence against trojan attacks on deep neural networks [R]. San Juan: The 35th Annual Computer Security Applications Conference, 2019.

[4] Han D Q, Wang Z L, Zhong Y, et al. Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors [J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2632–2647.

[5] Zhang J, Chen X, Xiang Y, et al. Robust network traffic classification [J]. IEEE/ACM Transactions on Networking, 2015, 23(4): 1257–1270.

[6] Liu C, He L T, Xiong G, et al. FS-net: A flow sequence network for encrypted traffic classification [R]. Paris: IEEE INFOCOM 2019—IEEE Conference on Computer Communications, 2019.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks [J]. Conmmunications of the ACM, 2020, 63(11): 139–144.

[8] Shmakov A, Greif K, Fenton M, et al. End-to-end latent variational diffusion models for inverse problems in high energy physics [R]. New Orleans: The 37th International Conference on Neural Information Processing Systems, 2023.

[9] Wang Z K, Pang T Y, Du C, et al. Better diffusion models further improve adversarial training [R]. Honolulu: The 40th International Conference on Machine Learning, 2023.

[10] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321–357.

[11] Tobin J, Fong R, Ray A, et al. Domain randomization for transferring deep neural networks from simulation to the real world [R]. Vancouver: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017.

[12] Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization [R]. Brisbane: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018.

[13] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [R]. Long Beach: The 31st International Conference on Neural Information Processing Systems, 2017.

[14] Badreddine S, d'Avila Garcez A, Serafini L, et al. Logic tensor networks [J]. Artificial Intelligence, 2022, 303: 103649.

[15] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks *via* gradient-based localization [J]. International Journal of Computer Vision, 2020, 128(2): 336–359.

[16] Liu Y A, Zhang R Q, Zhang M K, et al. Perturbation-invariant adversarial training for neural ranking models: Improving the effectiveness–robustness trade-off [R]. Vancouver: The 38th Annual the AAAI Conference on Artificial Intelligence, 2024, 38(8): 8832–8840.

[17] Yao L F, Niu W N, Yuan Q J, et al. A robust malicious traffic detection framework with low-quality labeled data [R]. Denver: ICC 2024–IEEE International Conference on Communications, 2024.

[18] Han B, Yao Q, Yu X, et al. Co-teaching: Robust training of deep neural networks with extremely noisy labels [R]. Montréal: The 2018 Conference on Neural Information Processing Systems, 2018.

[19] Yuan Q J, Gou G P, Zhu Y B, et al. MCRe: A unified framework for handling malicious traffic with noise labels based on multidimensional constraint representation [J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 133–147.

[20] Li Y, Lyu X, Koren N, et al. Anti-backdoor learning: Training clean models on poisoned data [R]. Virtual: The 2021 Conference on Neural Information Processing Systems, 2021.

[21] Li T, Li H E, Pan Y N, et al. Meta stackelberg game: Robust federated learning against adaptive and mixed poisoning attacks [EB/OL]. (2024-10-22)[2025-03-10]. https://arxiv.org/abs/2410.17431.

[22] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks [R]. Vancouver: International Conference on Learning Representations, 2018.

[23] Pinto L, Davidson J, Sukthankar R, et al. Robust adversarial reinforcement learning [R]. Sydney: The 34th International Conference on Machine Learning (ICML 2017), 2017.

[24] Kalashnikov D, Irpan A, Pastor P, et al. Scalable deep reinforcement learning for vision-based robotic manipulation [R]. Zürich: The 2nd Annual Conference on Robot Learning (CoRL 2018), 2018.

[25] Ye J J, Wu Y L, Gao S Y, et al. RoTBench: A multi-level benchmark for evaluating the robustness of large language models in tool learning [R]. Miami: The 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[26] Huang Y, Sun L, Wang H, et al. Position: TrustLLM: Trustworthiness in large language models [R]. Baltimore: The 41st International Conference on Machine Learning (ICML 2024), 2024.

[27] Chang Y P, Wang X, Wang J D, et al. A survey on evaluation of large language models [J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1–45.

[28] Liu Y, Yao Y S, Ton J F, et al. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment [EB/OL]. (2023-08-10)[2025-03-02]. https://arxiv.org/abs/2308.05374.

[29] Qiu H C, Zhang S, Li A Q, et al. Latent jailbreak: A benchmark for evaluating text safety and output robustness of large language models [EB/OL]. (2023-07-17)[2025-03-02]. https://arxiv.org/abs/2307.08487.

[30] Dobre D, Gidel G, Günnemann S, et al. Soft prompt threats: Attacking safety alignment and unlearning in open-source LLMs through the embedding space [R]. Vancouver: Advances in Neural Information Processing Systems 37 (NeuralIPS 2024), 2024.

[31] Luo W D, Ma S Y, Liu X G, et al. JailBreakV: A benchmark for assessing the robustness of MultiModal large language models against jailbreak attacks [EB/OL]. (2024-04-03)[2025-03-06]. https://arxiv.org/abs/2404.03027.

[32] Zhao X D, Yang X J, Pang T Y, et al. Weak-to-strong jailbreaking on large language models [EB/OL]. (2024-01-30)[2025-03-12]. https://arxiv.org/abs/2401.17256.

[33] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback [R]. New Orleans: The 2022 Conference on Neural Information Processing Systems, 2022.

[34] Dettmers T, Pagnoni A, Holtzman A, et al. QLoRA: Efficient finetuning of quantized LLMs [R]. New Orleans: the 2023 Conference on Neural Information Processing Systems, 2023.

[35] Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models [R]. New Orleans: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[36] Wang S B, Zhang J, Yuan Z, et al. Pre-trained model guided fine-tuning for zero-shot adversarial robustness [R]. Seattle: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[37] Zhang H H, Kung P N, Yoshida M, et al. Adaptable logical control for large language models [R]. Vancouver: The 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), 2024.

[38] Abadi M, Chu A, Goodfellow I, et al. Deep learning with differential privacy [R]. Vienna: The 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.

[39] Wang X L, Jin Y C. Distilling ensemble surrogates for federated data-driven many-task optimization [J]. IEEE Transactions on Evolutionary Computation, 2025, 29(6): 2401–2415.

[40] Tan Q, Li Q, Zhao Y, et al. Defending against data reconstruction attacks in federated learning: An information theory approach [R]. Philadelphia: The 33rd USENIX Conference on Security Symposium, 2024.

[41] Xu Y X, Fang B X, Li M H, et al. Query-efficient model inversion attacks: An information flow view [J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 1023–1036.

[42] Tan T J L, Shokri R. Bypassing backdoor detection algorithms in deep learning [R]. Genoa: 2020 IEEE European Symposium on Security and Privacy (EuroS&P), 2020.

[43] Steinhardt J, Koh P W, Liang P. Certified defenses for data poisoning attacks [R]. Long Beach: The 30th International Conference on Neural Information Processing Systems (NeurIPS 2017), 2017.

[44] Huang H, Erfani S M, Li Y, et al. Detecting backdoor samples in contrastive language image pretraining [R]. Singapore: The 13th International Conference on Learning Representations (ICLR 2025), 2025.

[45] Wang N, Chen Y M, Hu Y, et al. MANDA: On adversarial example detection for network intrusion detection system [R]. Vancouver: IEEE INFOCOM 2021—IEEE Conference on Computer Communications, 2021.

[46] Jiang W B, Li H W, Xu G W, et al. A comprehensive defense framework against model extraction attacks [J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(2): 685–700.

[47] Wang B L, Yao Y S, Shan S, et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks [R]. San Francisco: 2019 IEEE Symposium on Security and Privacy (SP), 2019.

[48] Liu X G, Li M H, Wang H Y, et al. Detecting backdoors during the inference stage based on corruption robustness consistency [R]. Vancouver: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[49] Fang B X, Li M H, Tang K K, et al. LT-defense: Searching-free backdoor defense *via* exploiting the long-tailed effect [R]. Vancouver: Advances in Neural Information Processing Systems 37, 2024.

[50] Tramèr F, Boneh D. Adversarial training and robustness for multiple perturbations [R]. Red Hook: The 32th International Conference on Neural Information Processing Systems (NeurIPS 2024), 2019.

[51] Qi F C, Chen Y Y, Li M K, et al. ONION: A simple and effective defense against textual backdoor attacks [R]. Punta Cana: The 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

[52] Shan S, Wenger E, Wang B L, et al. Gotta Catch'Em all: Using honeypots to catch adversarial attacks on neural networks [R]. Online: The 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020.

[53] Gong X L, Wang Z Y, Li S K, et al. A GAN-based defense framework against model inversion attacks [J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4475–4487.

[54] Xu Y X, Fang B X, Wang R, et al. Neural honeytrace: A robust plug-and-play watermarking framework against model extraction attacks [EB/OL]. (2025-01-16)[2025-03-012]. https://arxiv. org/abs/2501.09328.

[55] Gao L, Fu C P, Deng X H, et al. Wedjat: Detecting sophisticated evasion attacks *via* real-time causal analysis [R]. Toronto: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, 2025.

[56] Zhu H, Zhang S Z, Chen K. AI-guardian: Defeating adversarial attacks using backdoors [R]. San Francisco: 2023 IEEE Symposium on Security and Privacy (SP), 2023.

[57] Feng R, Hooda A, Mangaokar N, et al. Stateful defenses for machine learning models are not yet secure against black-box attacks [R]. Copenhagen: The 2023 ACM SIGSAC Conference on Computer and Communications Security, 2023.

[58] Zhang X Y, Zhang C, Li T L, et al. JailGuard: A universal detection framework for prompt-based attacks on LLM systems [J]. ACM Transactions on Software Engineering and Methodology, 2025, 35(1): 1–40.

[59] Schramowski P, Brack M, Deiseroth B, et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models [R]. Vancouver: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[60] 台建玮, 杨双宁, 王佳佳, 等. 大语言模型对抗性攻击与防御综述 [J]. 计算机研究与发展, 2025, 62(3): 563–588.
Tai J W, Yang S N, Wang J J, et al. Survey of adversarial attacks and defenses for large language models [J]. Journal of Computer Research and Development, 2025, 62(3): 563–588.

[61] Li N, Han Z W, Steneker I, et al. LLM defenses are not robust to multi-turn human jailbreaks yet [R]. Vancouver: Red Teaming GenAI Workshop, The 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), 2024.

[62] Jain N, Schwarzschild A, Wen Y X, et al. Baseline defenses for adversarial attacks against aligned language models [EB/OL]. (2023-09-01)[2025-03-25]. https://arxiv.org/abs/2309.00614.

[63] Kawasaki A, Davis A, Abbas H. Defending large language models against attacks with residual stream activation analysis [R]. Ar-

lington: The Conference on Applied Machine Learning in Information Security (CAMLIS 2024), 2024.

[64] Xie Y Q, Fang M H, Pi R J, et al. GradSafe: Detecting jailbreak prompts for LLMs *via* safety-critical gradient analysis [R]. Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[65] Zhang H F, Guo Z M, Zhu H S, et al. Jailbreak open-sourced large language models *via* enforced decoding [R]. Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[66] Ji J B, Hou B R, Robey A, et al. Defending large language models against jailbreak attacks *via* semantic smoothing [EB/OL]. (2024-02-25)[2025-03-25]. https://arxiv.org/abs/2402.16192.

[67] Robey A, Wong E, Hassani H, et al. SmoothLLM: Defending large language models against jailbreaking attacks [EB/OL]. (2023-10-05)[2025-07-25]. https://arxiv.org/abs/2310.03684.

[68] Zhao W, Li Z, Li Y G, et al. Defending large language models against jailbreak attacks *via* layer-specific editing [R]. Miami: The 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[69] Tang R, Yuan J, Li Y, et al. Setting the Trap: Capturing and defeating backdoors in pretrained language models through honeypots [R]. Orleans: The 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[70] Liu Q, Yin J, Wen W, et al. NeuroPots: Realtime proactive defense against bit-flip attacks in neural networks [R]. Anaheim: The 32nd USENIX Security Symposium (USENIX Security 2023), 2023.

[71] Andreina S, Marson G A, Möllering H, et al. BaFFLe: Backdoor detection *via* feedback-based federated learning [R]. Washington DC: 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS), 2021.

[72] Bongard J, Zykov V, Lipson H. Resilient machines through continuous self-modeling [J]. Science, 2006, 314(5802): 1118–1121.

[73] Cully A, Clune J, Tarapore D, et al. Robots that can adapt like animals [J]. Nature, 2015, 521(7553): 503–507.

[74] Zhao S, Tuan L A, Fu J, et al. Exploring clean label backdoor attacks and defense in language models [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3014–3024.

[75] Huang H Y, Wang Q, Gong X L, et al. Orion: Online backdoor sample detection *via* evolution deviance [R]. Macao: The Thirty-Second International Joint Conference on Artificial Intelligence, 2023.

[76] Liu K, Dolan-Gavitt B, Garg S. Fine-pruning: Defending against backdooring attacks on deep neural networks [R]. Cham: Research in Attacks, Intrusions, and Defenses, 2018.

[77] Guo W B, Wang L, Xu Y, et al. Towards inspecting and eliminating Trojan backdoors in deep neural networks [R]. Sorrento: 2020 IEEE International Conference on Data Mining (ICDM), 2020.

[78] Zhu L W, Ning R, Wang C, et al. GangSweep: Sweep out neural backdoors by GAN [R]. Seattle: The 28th ACM International Conference on Multimedia, 2020.

[79] Wu D, Wang Y. Adversarial neuron pruning purifies backdoored deep models [R]. Online: The 34th International Conference on Neural Information Processing Systems (NeurIPS 2021), 2021.

[80] Li Y G , Lyu X X, Koren N, et al. Neural attention distillation: Erasing backdoor triggers from deep neural networks [R]. Online: The 9th International Conference on Learning Representations (ICLR 2021), 2021.

[81] Xu Q L, Tao G H, Honorio J, et al. MEDIC: Remove model backdoors *via* importance driven cloning [R]. Vancouver: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[82] Liu Y, Fan M Y, Chen C, et al. Backdoor defense with machine unlearning [R]. London: IEEE INFOCOM 2022—IEEE Conference on Computer Communications, 2022.

[83] Chen X, Ding K, Fan C Y, et al. Defensive unlearning with adversarial training for robust concept erasure in diffusion models [R]. Vancouver: The 38th International Conference on Neural Information Processing Systems (NeurIPS 2024), 2024.

[84] Sun B, Sun J, Pham L H, et al. Causality-based neural network repair [R]. Pittsburgh: 2022 IEEE/ACM 44th International Conference on Software Engineering (ICSE), 2022.

[85] Zhang T, Zhang W J, Gupta M M, et al. Resilient robots: Concept, review, and future directions [J]. Robotics, 2017, 6(4): 22–36.

[86] Chen S, Li B. Toward adaptive reasoning in large language models with thought rollback [R]. Vienna: The 41st International Conference on Machine Learning (ICML 2024), 2024.

[87] Lo M, Barez F, Cohen S. Large language models relearn removed concepts [R]. Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics ACL 2024, 2024.

[88] Zhang L, Jin Q, Huang H Y, et al. Respond in my language: Mitigating language inconsistency in response generation based on large language models [R]. Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[89] Zhang X Y, Peng B L, Tian Y, et al. Self-alignment for factuality: Mitigating hallucinations in LLMs *via* self-evaluation [R]. Bangkok: The 62nd Annual Meeting of the Association for Computational Linguistics, 2024.

[90] Zintgraf L M, Shiarlis K, Kurin V, et al. Fast context adaptation *via* meta-learning [R]. Long Beach: The 36th International Conference on Machine Learning (ICML 2019), 2019.

[91] Snell J, Swersky K, Zemel R S. Prototypical networks for few-shot learning [R]. Long Beach: The 30th International Conference on Neural Information Processing Systems (NeurIPS 2017), 2017.

[92] Sung F, Yang Y X, Zhang L, et al. Learning to compare: Relation network for few-shot learning [R]. Salt Lake City: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[93] Jiang M H, Cui M X, Liu C, et al. Zero-relabelling mobile-app identification over drifted encrypted network traffic [J]. Computer Networks, 2023, 228: 109728.

[94] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction [R]. Sydney: The 34th International Conference on Machine Learning 2017, 2017.

[95] Pinto L, Gupta A. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours [R]. Stockholm: 2016 IEEE

International Conference on Robotics and Automation (ICRA), 2016.

[96] Mallya A, Lazebnik S. PackNet: Adding multiple tasks to a single network by iterative pruning [R]. Salt Lake City: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[97] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks [EB/OL]. (2016-06-15)[2025-03-05]. https://arxiv.org/abs/1606.04671.

[98] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks [R]. Sydney: The 34th International Conference on Machine Learning, 2017.

[99] Parisi G I, Kemker R, Part J L, et al. Continual lifelong learning with neural networks: A review [J]. Neural Networks, 2019, 113: 54–71.

[100] Nagabandi A, Kahn G, Fearing R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning [R]. Brisbane: 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018.

[101] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies [J]. Journal of Machine Learning Research, 2016, 17(1): 1334–1373.

[102] Zenke F, Poole B, Ganguli S. Continual learning through synaptic intelligence [R]. Sydney: The 34th International Conference on Machine Learning, 2017.

[103] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [R]. Vienna: The 37th International Conference on Machine Learning, 2020.

[104] He K M, Fan H Q, Wu Y X, et al. Momentum contrast for unsupervised visual representation learning [R]. Seattle: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[105] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners [R]. New Orleans: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[106] Xie R J, Wang Y X, Cao J H, et al. Rosetta: Enabling robust TLS encrypted traffic classification in diverse network environments with TCP-aware traffic augmentation [R]. Wuhan: The ACM Turing Award Celebration Conference—China 2023, 2023.

[107] Cai S H, Tang H, Chen J F, et al. CDDA-MD: An efficient malicious traffic detection method based on concept drift detection and adaptation technique [J]. Computers & Security, 2025, 148: 104121.

[108] Cheng S Y, Tao G H, Liu Y Q, et al. BEAGLE: Forensics of deep learning backdoor attack for better defense [R]. San Diego: The Network and Distributed System Security Symposium 2023, 2023.

[109] Ha D, Schmidhuber J. Recurrent world models facilitate policy evolution [R]. Montréal: The 32nd International Conference on Neural Information Processing Systems (NeurIPS 2018), 2018.

[110] Zhang Y Z, Abbeel P, Pinto L. Automatic curriculum learning through value disagreement [R]. Vancouver: The 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.

[111] Zhao R, Song J M, Yuan Y F, et al. Maximum entropy population-based training for zero-shot human-AI coordination [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(5): 6145–6153.

[112] Jang J, Ye S, Yang S, et al. Towards continual knowledge learning of language models [R]. Online: The International Conference on Learning Representations ICLR 2022, 2022.

[113] Hao S, Liu T, Wang Z, et al. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings [R]. New Orleans: The 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.

[114] Lange R, Tian Y T, Tang Y J. Large language models as evolution strategies [R]. Melbourne: The Genetic and Evolutionary Computation Conference Companion, 2024.

[115] Du G D, Li J, Liu H T, et al. Knowledge fusion by evolving weights of language models [R]. Bangkok: The 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[116] Luo J Y, Zhang W Z, Yuan Y, et al. Large language model agent: A survey on methodology, applications and challenges [EB/OL]. (2025-03-27)[2025-04-28]. https://arxiv.org/abs/2503.21460.

[117] Dou Z Y, Yang C F, Wu X Q, et al. Re-ReST: Reflection-reinforced self-training for language agents [R]. Miami: The 2024 Conference on Empirical Methods in Natural Language Processing, 2024.

[118] Shinn N, Cassano F, Berman E, et al. Reflexion: Language agents with verbal reinforcement learning [R]. Orleans: The 37th International Conference on Neural Information Processing Systems (NeurIPS 2023), 2023.