A Comprehensive Review of Gen AI Agents: Applications and Frameworks in Finance, Investments and Risk Domains

Item Type Journal Article
Author Satyadhar Joshi
Abstract This paper surveys the landscape of AI agent frameworks, highlights their core features and differences, and explores their applications in financial services. We synthesize insights from recent industry reports, academic research, and technical blog posts, focusing on frameworks such as CrewAI, LangGraph, LlamaIndex, and others. We also discuss the challenges and opportunities of deploying agentic AI in production environments, with an emphasis on financial trading, investment analysis, and decision support. We analyze the rapidly evolving landscape of agentic AI systems, focusing on their architecture, capabilities, and practical implementations in banking, trading, and risk management. The study examines prominent frameworks including LangGraph for stateful agent orchestration, CrewAI for collaborative multi-agent workflows, and AutoGen for conversational agent systems, alongside industry platforms like IBM watsonx and NVIDIA NIM. The study examines both technical frameworks (LangGraph, CrewAI, AutoGen, etc.) and practical implementations in financial institutions. We highlight productivity gains (up to 80% time reduction in data tasks), risk management improvements, and workforce transformation challenges. The paper concludes with recommendations for financial institutions adopting agentic AI solutions. Our analysis reveals three key findings: (1) specialized agent frameworks achieve 50-80% productivity gains in financial data tasks compared to traditional approaches, (2) multi-agent systems demonstrate particular promise in complex domains like algorithmic trading and fraud detection, and (3) successful deployment requires addressing critical challenges in workforce upskilling, risk alignment, and regulatory compliance. The paper provides a theoretical foundation for agentic AI in finance, introducing formal models for agent design patterns, multimodal fusion, and market microfoundations. We further present a summary of several evaluation frameworks for assessing agent performance across financial use cases, including portfolio optimization and AML compliance. The study concludes with recommendations for financial institutions adopting agentic AI, emphasizing the need for standardized architectures, robust testing protocols, and hybrid human-AI workflows.
Date 2025-05
Short Title A Comprehensive Review of Gen AI Agents
URL https://hal.science/hal-05101606
Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: joshi_comprehensive_2025
Pages 1339 – 1355
Publication International Journal of Innovative Science and Research Technology
DOI 10.38124/ijisrt/25may964
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Agentic AI, AI Agents, AI Agents Agentic AI Financial Services Multi-Agent Systems Generative AI Risk Management Multi-Agent Systems Financial Technology LLMs Autonomous Agents Frameworks, Financial Services, Financial Technology, Frameworks, Generative AI, LLMs Autonomous Agents, Multi-Agent Systems, Risk Management

A comprehensive study of jailbreak attack versus defense for large language models

Item Type Document
Author Zihao Xu
Author Yi Liu
Author Gelei Deng
Author Yuekang Li
Author Stjepan Picek
Date 2024

URL https://arxiv.org/abs/2402.13457
Extra Citation Key: xu2024comprehensivestudyjailbreakattack arXiv: 2402.13457 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

A comprehensive survey in LLM(-agent) full stack safety: Data, training and deployment

Item Type Document
Author Kun Wang
Author Guibin Zhang
Author Zhenhong Zhou
Author Jiahao Wu
Author Miao Yu
Author Shiqian Zhao
Author Chenlong Yin
Author Jinhu Fu
Author Yibo Yan
Author Hanjun Luo
Author Liang Lin
Author Zhihao Xu
Author Haolang Lu
Author Xinye Cao
Author Xinyun Zhou
Author Weifei Jin
Author Fanci Meng
Author Shicheng Xu
Author Junyuan Mao
Author Yu Wang
Author Hao Wu
Author Minghe Wang
Author Fan Zhang
Author Junfeng Fang
Author Wenjie Qu
Author Yue Liu
Author Chengwei Liu
Author Yifan Zhang
Author Qiankun Li
Author Chongye Guo
Author Yalan Qin
Author Zhaoxin Fan
Author Kai Wang
Author Yi Ding
Author Donghai Hong
Author Jiaming Ji
Author Yingxin Lai
Author Zitong Yu
Author Xinfeng Li
Author Yifan Jiang
Author Yanhui Li
Author Xinyu Deng

Author Junlin Wu
Author Dongxia Wang
Author Yihao Huang
Author Yufei Guo
Author Jen-tse Huang
Author Qiufeng Wang
Author Xiaolong Jin
Author Wenxuan Wang
Author Dongrui Liu
Author Yanwei Yue
Author Wenke Huang
Author Guancheng Wan
Author Heng Chang
Author Tianlin Li
Author Yi Yu
Author Chenghao Li
Author Jiawei Li
Author Lei Bai
Author Jie Zhang
Author Qing Guo
Author Jingyi Wang
Author Tianlong Chen
Author Joey Tianyi Zhou
Author Xiaojun Jia
Author Weisong Sun
Author Cong Wu
Author Jing Chen
Author Xuming Hu
Author Yiming Li
Author Xiao Wang
Author Ningyu Zhang
Author Luu Anh Tuan
Author Guowen Xu
Author Jiaheng Zhang
Author Tianwei Zhang
Author Xingjun Ma
Author Jindong Gu
Author Liang Pang
Author Xiang Wang
Author Bo An
Author Jun Sun
Author Mohit Bansal
Author Shirui Pan
Author Lingjuan Lyu
Author Yuval Elovici
Author Bhavya Kailkhura
Author Yaodong Yang
Author Hongwei Li
Author Wenyuan Xu
Author Yizhou Sun
Author Wei Wang

Author Qing Li
Author Ke Tang
Author Yu-Gang Jiang
Author Felix Juefei-Xu
Author Hui Xiong
Author Xiaofeng Wang
Author Dacheng Tao
Author Philip S. Yu
Author Qingsong Wen
Author Yang Liu
Date 2025
URL https://arxiv.org/abs/2504.15585
Extra Citation Key: wang2025comprehensivesurveyllmagentstack arXiv: 2504.15585 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: Applications, open challenges and future research directions

Item Type Journal Article
Author Avyay Casheekar
Author Archit Lahiri
Author Kanishk Rath
Author Kaushik Sanjay Prabhakar
Author Kathiravan Srinivasan
Date 2024
Extra Citation Key: casheekar2024contemporary Publisher: Elsevier
Volume 52
Pages 100632
Publication Computer Science Review
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

A dynamic and high-precision method for scenario-based HRA synthetic data collection in multi-agent collaborative environments driven by llms

Item Type Journal Article
Author Xingyu Xiao
Author Peng Chen
Author Qianqian Jia
Author Jiejuan Tong
Author Jingang Liang
Author Haitao Wang
Date 2025
Extra Citation Key: xiao2025dynamic
Publication arXiv preprint arXiv:2502.00022
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

A Multi-Agent Approach to Investor Profiling Using Large Language Models

Item Type Conference Paper
Author Hanpeng Wang
Author Zijiang Yang
Abstract Investor profiling is essential in financial advising, allowing advisors to tailor investment strategies based on individual risk preferences, experience, and financial goals. This research aims to automate and enhance the investor profiling process using large language models (LLMs) through interactive multi-agent conversations. Our approach involves designing an investor agent, which represents a pre-defined investor derived from a narrative generated by a large language model (LLM) with given attributes, and an advisor agent that engages in conversation to infer the hidden attributes of the investor. The advisor-agent dynamically adjusts its questions based on previous conversation context to maximize the accuracy of its attribute predictions. The advisor agent makes predictions once it acquires sufficient information and compares them against the ground truth. We conducted extensive simulations across thousands of investor attribute sets and evaluated the effectiveness of the advisor-agent's predictions based on key metrics. Our results demonstrate that LLM can effectively approximate investor characteristics. This research contributes to the field of AI-driven financial advising and unveils the potential of conversational agents in refining investor assessment methodologies.
Date 2025-07
URL https://ieeexplore.ieee.org/abstract/document/11099326
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: wang_multi-agent_2025
Pages 1–6
Proceedings Title 2025 International Conference on Control, Automation and Diagnosis (ICCAD)
DOI 10.1109/ICCAD64771.2025.11099326
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Multi-Agent Systems, AI Simulations, Automation, Autonomous systems, Autonomous Systems, Financial Technologies, Fintech, Investment, Large language models, Large Language Models, Measurement, Multi-agent systems, Oral communication, Predictive models, Refining

Notes:

ISSN: 2767-9896

A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist

Item Type Conference Paper
Author Wentao Zhang
Author Lingxuan Zhao
Author Haochong Xia
Author Shuo Sun
Author Jiaze Sun
Author Molei Qin
Author Xinyi Li
Author Yuqing Zhao
Author Yilei Zhao
Author Xinyu Cai
Author others
Date 2024

Extra Citation Key: zhang2024multimodal
Pages 4314–4325
Proceedings Title Proceedings of the 30th acm sigkdd conference on knowledge discovery and data mining
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

A practical memory injection attack against LLM agents

Item Type Document
Author Shen Dong
Author Shaochen Xu
Author Pengfei He
Author Yige Li
Author Jiliang Tang
Author Tianming Liu
Author Hui Liu
Author Zhen Xiang
Date 2025
URL https://arxiv.org/abs/2503.03704
Extra Citation Key: dong2025practicalmemoryinjectionattack arXiv: 2503.03704 [cs.LG]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

A privacy-preserving and trustable multi-agent learning framework

Item Type Journal Article
Author Anudit Nagar
Author Cuong Tran
Author Ferdinando Fioretto
Date 2021
Extra Citation Key: nagar2021privacy
Publication arXiv preprint arXiv:2106.01242
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

A Review of LLM Agent Applications in Finance and Banking

Item Type Journal Article
Author Devesh Batra
Author Conor B Hamill
Author John Hartley
Author Ramin Okhrati
Author Dale Seddon
Author Raad Khraishi
Author Greig A Cowan
Abstract The accelerating digital transformation in finance and banking, coupled with advances in large language models (LLMs), has spurred investigation into LLM-powered computational agents capable of simulating, analysing, assisting, and acting within complex financial ecosystems. Leveraging their advanced reasoning and linguistic capabilities, these agents are uniquely positioned to address the multifaceted challenges of modern banking,

finance, and economics, delivering scalable solutions for risk management, regulatory compliance, and strategic decision-making. Yet, the rapidly growing literature on LLM agents in the sector remains fragmented, lacking a cohesive survey evaluating their capabilities, risks, and real-world applicability. This survey paper offers a review of the current literature on LLM agents in finance and banking, categorising their applications in the sector into four core functions: simulation, acting, analysis, and advising. We examine a large corpus of studies employing LLM agents in market simulation, macroeconomic and microeconomic scenario planning, synthetic data generation, automated trading, and decision support systems among other applications, while critically analysing their technical efficacy, ethical dimensions, regulatory compliance, and operational limitations. In our survey we find that while LLM agents excel in linguistic tasks, their deployment in mission-critical systems requires hybrid architectures (including human supervision) and robust safeguards against hallucinations and biases. By integrating insights from diverse subject areas and frameworks, our work highlights key challenges and opportunities for advancing the safe and effective use of LLM agents in finance and banking. This work serves as both a reference for researchers and a pragmatic guide for practitioners navigating the transformative potential and pitfalls of LLM agents in finance and banking.

Language en
Extra Citation Key: batra_review_nodate
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

A spatiotemporal stealthy backdoor attack against cooperative multi-agent deep reinforcement learning

Item Type Journal Article
Author Yinbo Yu
Author Saihao Yan
Author Jiajia Liu
Date 2024
Extra Citation Key: yu2024spatiotemporal
Publication arXiv preprint arXiv:2409.07775
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

A Survey of Financial AI: Architectures, Advances and Open Challenges

Item Type Document
Author Junhua Liu
Abstract Financial AI empowers sophisticated approaches to financial market forecasting, portfolio optimization, and automated trading. This survey provides a systematic analysis of these developments across three primary dimensions: predictive models that capture complex market dynamics, decision-making frameworks that optimize trading and investment strategies, and knowledge augmentation systems that leverage unstructured financial information. We examine significant innovations including foundation models for financial time series, graph-based architectures for market relationship modeling, and hierarchical frameworks for portfolio optimization. Analysis reveals crucial trade-offs between model sophistication and practical constraints, particularly in high-frequency trading applications. We identify critical gaps and open challenges between theoretical advances and industrial implementation, outlining open challenges and opportunities for improving both model performance and practical applicability.
Date 2024-11
Short Title A Survey of Financial AI
URL http://arxiv.org/abs/2411.12747
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: liu_survey_2024 DOI: 10.48550/arXiv.2411.12747
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Quantitative Finance - Trading and Market Microstructure, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2411.12747 [q-fin]

A survey of large language models

Item Type	Journal Article
Author	Wayne Xin Zhao
Author	Kun Zhou
Author	Junyi Li
Author	Tianyi Tang
Author	Xiaolei Wang
Author	Yupeng Hou
Author	Yingqian Min
Author	Beichen Zhang
Author	Junjie Zhang
Author	Zican Dong
Author	others
Date	2023
Extra	Citation Key: zhao2023survey
Publication	arXiv preprint arXiv:2303.18223
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

A survey of LLM-driven AI agent communication: Protocols, security risks, and defense countermeasures

Item Type	Document
Author	Dezhang Kong
Author	Shi Lin
Author	Zhenhua Xu
Author	Zhebo Wang
Author	Minghao Li
Author	Yufeng Li
Author	Yilun Zhang
Author	Hujin Peng
Author	Zeyang Sha
Author	Yuyuan Li
Author	Changting Lin
Author	Xun Wang
Author	Xuan Liu
Author	Ningyu Zhang
Author	Chaochao Chen
Author	Muhammad Khurram Khan
Author	Meng Han

Date 2025
URL https://arxiv.org/abs/2506.19676
Extra Citation Key: kong2025surveyllmdrivenaiagent arXiv: 2506.19676 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations

Item Type Journal Article
Author Yihe Zhou
Author Tao Ni
Author Wei-Bin Lee
Author Qingchuan Zhao
Date 2025
Extra Citation Key: zhou2025survey
Publication arXiv preprint arXiv:2502.05224
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

A survey on evaluation of large language models

Item Type Journal Article
Author Yupeng Chang
Author Xu Wang
Author Jindong Wang
Author Yuan Wu
Author Linyi Yang
Author Kaijie Zhu
Author Hao Chen
Author Xiaoyuan Yi
Author Cunxiang Wang
Author Yidong Wang
Author others
Date 2024
Extra Citation Key: chang2024survey Publisher: ACM New York, NY
Volume 15
Pages 1–45
Publication ACM Transactions on Intelligent Systems and Technology
Issue 3
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions

Item Type Journal Article
Author Lei Huang
Author Weijiang Yu

Author Weitao Ma
Author Weihong Zhong
Author Zhangyin Feng
Author Haotian Wang
Author Qianglong Chen
Author Weihua Peng
Author Xiaocheng Feng
Author Bing Qin
Author others
Date 2023
Extra Citation Key: huang2023survey
Publication arXiv preprint arXiv:2311.05232
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

A survey on large language model based autonomous agents

Item Type Journal Article
Author Lei Wang
Author Chen Ma
Author Xueyang Feng
Author Zeyu Zhang
Author Hao Yang
Author Jingsen Zhang
Author Zhiyuan Chen
Author Jiakai Tang
Author Xu Chen
Author Yankai Lin
Author others
Date 2024
Extra Citation Key: wang2024survey Publisher: Springer
Volume 18
Pages 186345
Publication Frontiers of Computer Science
Issue 6
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

A survey on large language models for critical societal domains: Finance, healthcare, and law

Item Type Journal Article
Author Zhiyu Zoey Chen
Author Jing Ma
Author Xinlu Zhang
Author Nan Hao
Author An Yan
Author Armineh Nourbakhsh
Author Xianjun Yang
Author Julian McAuley

Author Linda Petzold
Author William Yang Wang
Date 2024
Extra Citation Key: chen2024survey
Publication arXiv preprint arXiv:2405.01769
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

A survey on model extraction attacks and defenses for large language models

Item Type Conference Paper
Author Kaixiang Zhao
Author Lincan Li
Author Kaize Ding
Author Neil Zhenqiang Gong
Author Yue Zhao
Author Yushun Dong
Date 2025
URL https://doi.org/10.1145/3711896.3736573
Extra Citation Key: zhao2025SurverModelExtraction Number of pages: 10 tex.address: New York, NY, USA
Place Toronto ON, Canada
Publisher Association for Computing Machinery
ISBN 979-8-4007-1454-2
Pages 6227–6236
Series Kdd '25
Proceedings Title Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining V.2
DOI 10.1145/3711896.3736573
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

A survey on the memory mechanism of large language model based agents

Item Type Journal Article
Author Zeyu Zhang
Author Xiaohe Bo
Author Chen Ma
Author Rui Li
Author Xu Chen
Author Quanyu Dai
Author Jieming Zhu
Author Zhenhua Dong
Author Ji-Rong Wen
Date 2024
Extra Citation Key: zhang2024survey
Publication arXiv preprint arXiv:2404.13501
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

A survey on trustworthy LLM agents: Threats and countermeasures

Item Type	Conference Paper
Author	Miao Yu
Author	Fanci Meng
Author	Xinyun Zhou
Author	Shilong Wang
Author	Junyuan Mao
Author	Linsey Pan
Author	Tianlong Chen
Author	Kun Wang
Author	Xinfeng Li
Author	Yongfeng Zhang
Author	Bo An
Author	Qingsong Wen
Abstract	With the rapid evolution of Large Language Models (LLMs), LLM-based agents and Multi-agent Systems (MAS) have significantly expanded the capabilities of LLM ecosystems. This evolution stems from empowering LLMs with additional modules such as memory, tools, environment, and even other agents. However, this advancement has also introduced more complex issues of trustworthiness, which previous research focusing solely on LLMs could not cover. In this survey, we propose the TrustAgent framework, a comprehensive study on the trustworthiness of agents, characterized by modular taxonomy, multi-dimensional connotations, and technical implementation. By thoroughly investigating and summarizing newly emerged attacks, defenses, and evaluation methods for agents and MAS, we extend the concept of Trustworthy LLM to the emerging paradigm of Trustworthy Agent. In TrustAgent, we begin by deconstructing and introducing various components of the Agent and MAS. Then, we categorize their trustworthiness into intrinsic (brain, memory, and tool) and extrinsic (user, agent, and environment) aspects. Subsequently, we delineate the multifaceted meanings of trustworthiness and elaborate on the implementation techniques of existing research related to these internal and external modules. Finally, we present our insights and outlook on this domain, aiming to provide guidance for future endeavors. For easy reference, we categorize all the studies mentioned in this survey according to our taxonomy, available at: https://github.com/Ymm-cll/TrustAgent.
Date	2025
URL	https://doi.org/10.1145/3711896.3736561
Extra	Citation Key: yu2025surveytrustworthyllmagents Number of pages: 11 tex.address: New York, NY, USA
Place	Toronto ON, Canada
Publisher	Association for Computing Machinery
ISBN	979-8-4007-1454-2
Pages	6216–6226
Series	Kdd '25
Proceedings Title	Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining V.2
DOI	10.1145/3711896.3736561
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Tags:

agent safety, llm-based agent, multi-agent system

A systematic review of poisoning attacks against large language models

Item Type	Document
Author	Neil Fendley
Author	Edward W. Staley
Author	Joshua Carney

Author William Redman
Author Marie Chau
Author Nathan Drenkow
Date 2025
URL https://arxiv.org/abs/2506.06518
Extra Citation Key: fendley2025systematicreviewpoisoningattacks arXiv: 2506.06518 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Abusing images and sounds for indirect instruction injection in multi-modal LLMs

Item Type Journal Article
Author Eugene Bagdasaryan
Author Tsung-Yin Hsieh
Author Ben Nassi
Author Vitaly Shmatikov
Date 2023
Extra Citation Key: bagdasaryan2023abusing
Publication arXiv preprint arXiv:2307.10490
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

ACE: a security architecture for LLM-integrated app systems

Item Type Document
Author Evan Li
Author Tushin Mallick
Author Evan Rose
Author William Robertson
Author Alina Oprea
Author Cristina Nita-Rotaru
Date 2025
URL https://arxiv.org/abs/2504.20984
Extra Citation Key: li2025acesecurityarchitecturellmintegrated arXiv: 2504.20984 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Achieving fairness in multi-agent MDP using reinforcement learning

Item Type Conference Paper
Author Peizhong Ju
Author Arnob Ghosh
Author Ness Shroff
Date 2023
Extra Citation Key: ju2023achieving
Proceedings Title The twelfth international conference on learning representations
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents

Item Type	Conference Paper
Author	Qiusi Zhan
Author	Richard Fang
Author	Henil Shalin Panchal
Author	Daniel Kang
Editor	Luis Chiruzzo
Editor	Alan Ritter
Editor	Lu Wang
Date	2025-04
URL	https://aclanthology.org/2025.findings-naacl.395/
Extra	Citation Key: zhan-etal-2025-adaptive
Place	Albuquerque, New Mexico
Publisher	Association for Computational Linguistics
ISBN	979-8-89176-195-7
Pages	7101–7117
Proceedings Title	Findings of the association for computational linguistics: NAACL 2025
DOI	10.18653/v1/2025.findings-naacl.395
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Advancing autonomous incident response: Leveraging llms and cyber threat intelligence

Item Type	Document
Author	Amine Tellache
Author	Abdelaziz Amara Korba
Author	Amdjed Mokhtari
Author	Horea Moldovan
Author	Yacine Ghamri-Doudane
Date	2025
URL	https://arxiv.org/abs/2508.10677
Extra	Citation Key: tellache2025advancingautonomousincidentresponse arXiv: 2508.10677 [cs.CR]
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Agent hospital: A simulacrum of hospital with evolvable medical agents

Item Type	Journal Article
Author	Junkai Li
Author	Yunghwei Lai
Author	Weitao Li
Author	Jingyi Ren
Author	Meng Zhang
Author	Xinhui Kang
Author	Siyu Wang
Author	Peng Li
Author	Ya-Qin Zhang
Author	Weizhi Ma

Author others
Date 2024
Extra Citation Key: li2024agent
Publication arXiv preprint arXiv:2405.02957
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents

Item Type Document
Author Hanrong Zhang
Author Jingyuan Huang
Author Kai Mei
Author Yifei Yao
Author Zhenting Wang
Author Chenlu Zhan
Author Hongwei Wang
Author Yongfeng Zhang
Abstract Although LLM-based agents, powered by Large Language Models (LLMs), can use external tools and memory mechanisms to solve complex real-world tasks, they may also introduce critical security vulnerabilities. However, the existing literature does not comprehensively evaluate attacks and defenses against LLM-based agents. To address this, we introduce Agent Security Bench (ASB), a comprehensive framework designed to formalize, benchmark, and evaluate the attacks and defenses of LLM-based agents, including 10 scenarios (e.g., e-commerce, autonomous driving, finance), 10 agents targeting the scenarios, over 400 tools, 27 different types of attack/defense methods, and 7 evaluation metrics. Based on ASB, we benchmark 10 prompt injection attacks, a memory poisoning attack, a novel Plan-of-Thought backdoor attack, 4 mixed attacks, and 11 corresponding defenses across 13 LLM backbones. Our benchmark results reveal critical vulnerabilities in different stages of agent operation, including system prompt, user prompt handling, tool usage, and memory retrieval, with the highest average attack success rate of 84.30\%, but limited effectiveness shown in current defenses, unveiling important works to be done in terms of agent security for the community. We also introduce a new metric to evaluate the agents' capability to balance utility and security. Our code can be found at https://github.com/agiresearch/ASB.
Date 2025-05
Short Title Agent Security Bench (ASB)
URL http://arxiv.org/abs/2410.02644
Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: zhang_agent_2025 DOI: 10.48550/arXiv.2410.02644
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Cryptography and Security

Notes:

arXiv:2410.02644 [cs]

Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast

Item Type Journal Article
Author Xiangming Gu
Author Xiaosen Zheng
Author Tianyu Pang
Author Chao Du
Author Qian Liu
Author Ye Wang
Author Jing Jiang
Author Min Lin
Date 2024
Extra Citation Key: gu2024agent
Publication arXiv preprint arXiv:2402.08567
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Agent-e: From autonomous web navigation to foundational design principles in agentic systems

Item Type Journal Article
Author Tamer Abuelsaad
Author Deepak Akkil
Author Prasenjit Dey
Author Ashish Jagmohan
Author Aditya Vempaty
Author Ravi Kokku
Date 2024
Extra Citation Key: abuelsaad2024agent
Publication arXiv preprint arXiv:2407.13032
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Agent-SafetyBench: Evaluating the safety of LLM agents

Item Type Journal Article
Author Zhexin Zhang
Author Shiyao Cui
Author Yida Lu
Author Jingzhuo Zhou
Author Junxiao Yang
Author Hongning Wang
Author Minlie Huang
Date 2024
Extra Citation Key: zhang2024agent
Publication arXiv preprint arXiv:2412.14470
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents

Item Type	Journal Article
Author	Edoardo Debenedetti
Author	Jie Zhang
Author	Mislav Balunovic
Author	Luca Beurer-Kellner
Author	Marc Fischer
Author	Florian Tramèr
Date	2025
Extra	Citation Key: debenedetti2025agentdojo
Volume	37
Pages	82895–82920
Publication	Advances in Neural Information Processing Systems
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

AgentGuard: Repurposing agentic orchestrator for safety evaluation of tool orchestration

Item Type	Document
Author	Jizhou Chen
Author	Samuel Lee Cong
Date	2025
URL	https://arxiv.org/abs/2502.09809
Extra	Citation Key: chen2025agentguardrepurposingagenticorchestrator arXiv: 2502.09809 [cs.CR]
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

AgentHarm: a benchmark for measuring harmfulness of LLM agents

Item Type	Conference Paper
Author	Maksym Andriushchenko
Author	Alexandra Souly
Author	Mateusz Dziemian
Author	Derek Duenas
Author	Maxwell Lin
Author	Justin Wang
Author	Dan Hendrycks
Author	Andy Zou
Author	J Zico Kolter
Author	Matt Fredrikson
Author	Yarin Gal
Author	Xander Davies
Date	2025
URL	https://openreview.net/forum?id=AC5n7xHuR1
Extra	Citation Key: andriushchenko2025agentharm
Proceedings Title	The thirteenth international conference on learning representations
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Agentic discovery and validation of android app vulnerabilities

Item Type Document
Author Ziyue Wang
Author Liyi Zhou
Date 2025
URL https://arxiv.org/abs/2508.21579
Extra Citation Key: wang2025agenticdiscoveryvalidationandroid arXiv: 2508.21579 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Agentic-AI healthcare: Multilingual, privacy-first framework with MCP agents

Item Type Journal Article
Author Mohammed A. Shehab
Date 2025
URL https://arxiv.org/abs/2510.02325
Extra Citation Key: shehab2025agentic
Volume arXiv:2510.02325
Publication arXiv preprint
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

preprint, submitted 25 Sep 2025, cs.CR / cs.AI

Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases

Item Type Journal Article
Author Zhaorun Chen
Author Zhen Xiang
Author Chaowei Xiao
Author Dawn Song
Author Bo Li
Date 2025
Extra Citation Key: chen2025agentpoison
Volume 37
Pages 130185–130213
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Agents4PLC: Automating closed-loop PLC code generation and verification in industrial control systems using LLM-based agents

Item Type Journal Article
Author Zihan Liu

Author Ruinan Zeng
Author Dongxia Wang
Author Gengyun Peng
Author Jingyi Wang
Author Qiang Liu
Author Peiyu Liu
Author Wenhai Wang
Date 2024
Extra Citation Key: liu2024agents4plc
Publication arXiv preprint arXiv:2410.14209
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

AgentSafe: Safeguarding large language model-based multi-agent systems via hierarchical data management

Item Type Journal Article
Author Junyuan Mao
Author Fanci Meng
Author Yifan Duan
Author Miao Yu
Author Xiaojun Jia
Author Junfeng Fang
Author Yuxuan Liang
Author Kun Wang
Author Qingsong Wen
Date 2025
Extra Citation Key: mao2025agentsafe
Publication arXiv preprint arXiv:2503.04392
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

AgentSpec: Customizable runtime enforcement for safe and reliable LLM agents

Item Type Document
Author Haoyu Wang
Author Christopher M. Poskitt
Author Jun Sun
Date 2025
URL https://arxiv.org/abs/2503.18666
Extra Citation Key: wang2025agentspeccustomizableruntimeenforcement arXiv: 2503.18666 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

AgentVigil: Generic black-box red-teaming for indirect prompt injection against LLM agents

Item Type Document
Author Zhun Wang

Author Vincent Siu
Author Zhe Ye
Author Tianneng Shi
Author Yuzhou Nie
Author Xuandong Zhao
Author Chenguang Wang
Author Wenbo Guo
Author Dawn Song
Date 2025
URL https://arxiv.org/abs/2505.05849
Extra Citation Key: wang2025agentvigilgenericblackboxredteaming arXiv: 2505.05849 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

AGrail: a lifelong agent guardrail with effective and adaptive safety detection

Item Type Document
Author Weidi Luo
Author Shenghong Dai
Author Xiaogeng Liu
Author Suman Banerjee
Author Huan Sun
Author Muhao Chen
Author Chaowei Xiao
Abstract We propose AGrail, a lifelong guardrail framework for LLM agents to detect and mitigate both task-specific and systemic risks. AGrail features adaptive safety check generation, iterative safety check optimization, and tool compatibility, achieving strong safety performance across diverse agent tasks.
Date 2025
URL https://arxiv.org/abs/2502.11448
Extra Citation Key: luo2025agrail arXiv: 2502.11448 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

AI agents under threat: a survey of key security challenges and future pathways

Item Type Document
Author Zehang Deng
Author Yongjian Guo
Author Changzhou Han
Author Wanlun Ma
Author Junwu Xiong
Author Sheng Wen
Author Yang Xiang
Date 2024
URL https://arxiv.org/abs/2406.02630
Extra Citation Key: deng2024aiagentsthreatsurvey arXiv: 2406.02630 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Ai alignment: A comprehensive survey

Item Type Journal Article
Author Jiaming Ji
Author Tianyi Qiu
Author Boyuan Chen
Author Borong Zhang
Author Hantao Lou
Author Kaile Wang
Author Yawen Duan
Author Zhonghao He
Author Jiayi Zhou
Author Zhaowei Zhang
Author others
Date 2023
Extra Citation Key: ji2023ai
Publication arXiv preprint arXiv:2310.19852
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research

Item Type Conference Paper
Author Sophia M Pressman
Author Sahar Borna
Author Cesar A Gomez-Cabello
Author Syed A Haider
Author Clifton Haider
Author Antonio J Forte
Date 2024
Extra Citation Key: pressman2024ai Number: 8
Volume 12
Publisher MDPI
Pages 825
Proceedings Title Healthcare
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

AI in finance: Challenges, techniques and opportunities

Item Type Journal Article
Author Longbing Cao
Date 2021
URL https://arXiv.org/abs/2107.09051
Extra Citation Key: cao2021aiinfinance
Publication arXiv preprint arXiv:2107.09051
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

AI research

Item Type Document
Author J.P. Morgan AI Research Lab
Date 2024
URL https://www.jpmorgan.com/technology/artificial-intelligence
Extra Citation Key: jpmorgan2024aiwhitepaper
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

AI-and LLM-driven search tools: A paradigm shift in information access for education and research

Item Type Journal Article
Author Gobinda Chowdhury
Author Sudatta Chowdhury
Date 2024
Extra Citation Key: chowdhury2024ai Publisher: SAGE Publications Sage UK: London, England
Pages 01655515241284046
Publication Journal of Information Science
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

AI-powered contract security: Managing expiry, compliance, and risk mitigation through deep learning and llms

Item Type Book Section
Author – Mhia-Alddin
Author – Hussein
Date 2025
URL https://www.researchgate.net/publication/391151875_AI-Powered_Contract_Security_Managing_Expiry_Compliance_and_Risk_Mitigation_Through_Deep_Learning_and_LLMs
Extra Citation Key: mhia2025aipowered
Book Title Proceedings / chapters in AI / applied deep learning & legal tech
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

Chapter / working publication, April 2025

AI-powered patching: the future of automated vulnerability fixes

Item Type Report
Author Jan Keller
Author Jan Nowakowski
Date 2024
Extra Citation Key: keller2024aipoweredpatching
Date Added 10/20/2025, 3:50:52 PM

Modified 10/20/2025, 3:50:52 PM

AIOS: LLM agent operating system

Item Type Conference Paper
Author Kai Mei
Author Xi Zhu
Author Wujiang Xu
Author Wenyue Hua
Author Mingyu Jin
Author Zelong Li
Author Shuyuan Xu
Author Ruosong Ye
Author Yingqiang Ge
Author Yongfeng Zhang
Abstract Proposes an OS-like runtime for agents with lifecycle control, inter-agent messaging, policy enforcement, and resource governance; prototypes show simplified multi-agent development and prevention of common pathologies.
Date 2025
URL https://arxiv.org/pdf/2403.16971
Extra Citation Key: mei2025aios
Proceedings Title Conference on language modeling
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

AirGapAgent: Protecting privacy-conscious conversational agents

Item Type Conference Paper
Author Eugene Bagdasarian
Author Ren Yi
Author Sahra Ghalebikesabi
Author Peter Kairouz
Author Marco Gruteser
Author Sewoong Oh
Author Borja Balle
Author Daniel Ramage
Date 2024
Extra Citation Key: bagdasarian2024airgapagent
Pages 3868–3882
Proceedings Title Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

---

aiXamine: Simplified LLM safety and security

Item Type Document
Author Fatih Deniz
Author Dorde Popovic
Author Yazan Boshmaf

Author Euisuh Jeong
Author Minhaj Ahmad
Author Sanjay Chawla
Author Issa Khalil
Date 2025
URL https://arxiv.org/abs/2504.14985
Extra Citation Key: deniz2025aixaminesimplifiedllmsafety arXiv: 2504.14985 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

ALI-agent: Assessing llms' alignment with human values via agent-based evaluation

Item Type Journal Article
Author Han Wang
Author An Zhang
Author Nguyen Duy Tai
Author Jun Sun
Author Tat-Seng Chua
Author others
Date 2024
Extra Citation Key: wang2024ali
Volume 37
Pages 99040–99088
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Aligned LLMs are not aligned browser agents

Item Type Conference Paper
Author Priyanshu Kumar
Author Elaine Lau
Author Saranya Vijayakumar
Author Tu Trinh
Author Elaine T Chang
Author Vaughn Robinson
Author Shuyan Zhou
Author Matt Fredrikson
Author Sean M. Hendryx
Author Summer Yue
Author Zifan Wang
Date 2025
URL https://openreview.net/forum?id=NsFZZU9gvk
Extra Citation Key: kumar2025aligned
Proceedings Title The thirteenth international conference on learning representations
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Aligning llm agents by learning latent preference from user edits

Item Type	Journal Article
Author	Ge Gao
Author	Alexey Taymanov
Author	Eduardo Salinas
Author	Paul Mineiro
Author	Dipendra Misra
Date	2025
Extra	Citation Key: gao2025aligning
Volume	37
Pages	136873–136896
Publication	Advances in Neural Information Processing Systems
Date Added	10/20/2025, 3:49:10 PM
Modified	10/20/2025, 3:49:10 PM

Alpha-GPT 2.0: Human-in-the-Loop AI for Quantitative Investment

Item Type	Document
Author	Hang Yuan
Author	Saizhuo Wang
Author	Jian Guo
Abstract	Recently, we introduced a new paradigm for alpha mining in the realm of quantitative investment, developing a new interactive alpha mining system framework, Alpha-GPT. This system is centered on iterative Human-AI interaction based on large language models, introducing a Human-in-the-Loop approach to alpha discovery. In this paper, we present the next-generation Alpha-GPT 2.0 \footnote{Draft. Work in progress}, a quantitative investment framework that further encompasses crucial modeling and analysis phases in quantitative investment. This framework emphasizes the iterative, interactive research between humans and AI, embodying a Human-in-the-Loop strategy throughout the entire quantitative investment pipeline. By assimilating the insights of human researchers into the systematic alpha research process, we effectively leverage the Human-in-the-Loop approach, enhancing the efficiency and precision of quantitative investment research.
Date	2024-02
Short Title	Alpha-GPT 2.0
URL	http://arxiv.org/abs/2402.09746
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: yuan_alpha-gpt_2024 DOI: 10.48550/arXiv.2402.09746
Publisher	arXiv
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Computational Finance

Notes:

arXiv:2402.09746 [q-fin]

An evaluation mechanism of LLM-based agents on manipulating apis

Item Type	Conference Paper

Author Bing Liu
Author Zhou Jianxiang
Author Dan Meng
Author Haonan Lu
Date 2024
Extra Citation Key: liu2024evaluation
Pages 4649–4662
Proceedings Title Findings of the association for computational linguistics: EMNLP 2024
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Analyzing and mitigating object hallucination in large vision-language models

Item Type Journal Article
Author Yiyang Zhou
Author Chenhang Cui
Author Jaehong Yoon
Author Linjun Zhang
Author Zhun Deng
Author Chelsea Finn
Author Mohit Bansal
Author Huaxiu Yao
Date 2023
Extra Citation Key: zhou2023analyzing
Publication arXiv preprint arXiv:2310.00754
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Androidworld: A dynamic benchmarking environment for autonomous agents

Item Type Journal Article
Author Christopher Rawles
Author Sarah Clinckemaillie
Author Yifan Chang
Author Jonathan Waltz
Author Gabrielle Lau
Author Marybeth Fair
Author Alice Li
Author William Bishop
Author Wei Li
Author Folawiyo Campbell-Ajala
Author others
Date 2024
Extra Citation Key: rawles2024androidworld
Publication arXiv preprint arXiv:2405.14573
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Anonymizing medical documents with local, privacy preserving large language models: The LLM-Anonymizer

Item Type Journal Article
Author Isabella C Wiest
Author Marie-Elisabeth Leßmann
Author Fabian Wolf
Author Dyke Ferber
Author Marko Van Treeck
Author Jiefu Zhu
Author Matthias P Ebert
Author Christoph Benedikt Westphalen
Author Martin Wermke
Author Jakob Nikolas Kather
Date 2024
Extra Citation Key: wiest2024anonymizing Publisher: Cold Spring Harbor Laboratory Press
Pages 2024–06
Publication medRxiv : the preprint server for health sciences
Journal Abbr medRxiv
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Arabic dataset for LLM safeguard evaluation

Item Type Journal Article
Author Yasser Ashraf
Author Yuxia Wang
Author Bin Gu
Author Preslav Nakov
Author Timothy Baldwin
Date 2024
Extra Citation Key: ashraf2024arabic
Publication arXiv preprint arXiv:2410.17040
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

ARACNE: An LLM-based autonomous shell pentesting agent

Item Type Document
Author Tomas Nieponice
Author Veronica Valeros
Author Sebastian Garcia
Date 2025
URL https://arxiv.org/abs/2502.18528
Extra Citation Key: nieponice2025aracnellmbasedautonomousshell arXiv: 2502.18528 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Architecting resilient LLM agents: a guide to secure plan-then-execute implementations

Item Type Document
Author Ron F. Del Rosario
Author Klaudia Krawiecka
Author Christian Schroeder de Witt
Date 2025
URL https://arxiv.org/abs/2509.08646
Extra Citation Key: delrosario2025architectingresilientllmagents arXiv: 2509.08646 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Artificial intelligence can make markets more efficient—and more volatile

Item Type Document
Author Nassira Abbas
Author Charles Cohen
Author Dirk Jan Grolleman
Author Benjamin Mosk
Date 2024
URL https://www.imf.org/en/Blogs/Articles/2024/10/15/artificial-intelligence-can-make-markets-more-efficient-and-more-volatile
Extra Citation Key: imf2024ai_markets tex.howpublished: IMF Blog (October 15, 2024)
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Assessing the brittleness of safety alignment via pruning and low-rank modifications

Item Type Journal Article
Author Boyi Wei
Author Kaixuan Huang
Author Yangsibo Huang
Author Tinghao Xie
Author Xiangyu Qi
Author Mengzhou Xia
Author Prateek Mittal
Author Mengdi Wang
Author Peter Henderson
Date 2024
Extra Citation Key: wei2024assessing
Publication arXiv preprint arXiv:2402.05162
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Attacking llms and AI agents: Advertisement embedding attacks against large language models

Item Type Document
Author Qiming Guo

Author Jinwen Tang
Author Xingran Huang
Date 2025
URL https://arxiv.org/abs/2508.17674
Extra Citation Key: guo2025attackingllmsaiagents arXiv: 2508.17674 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Attacks on third-party apis of large language models

Item Type Journal Article
Author Wanru Zhao
Author Vidit Khazanchi
Author Haodi Xing
Author Xuanli He
Author Qiongkai Xu
Author Nicholas Donald Lane
Date 2024
Extra Citation Key: zhao2024attacks
Publication arXiv preprint arXiv:2404.16891
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Attacks, defenses and evaluations for llm conversation safety: A survey

Item Type Journal Article
Author Zhichen Dong
Author Zhanhui Zhou
Author Chao Yang
Author Jing Shao
Author Yu Qiao
Date 2024
Extra Citation Key: dong2024attacks
Publication arXiv preprint arXiv:2402.09283
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Attention is all you need

Item Type Conference Paper
Author Ashish Vaswani
Author Noam Shazeer
Author Niki Parmar
Author Jakob Uszkoreit
Author Llion Jones
Author Aidan N Gomez
Author Łukasz Kaiser
Author Illia Polosukhin

Date 2017
Extra Citation Key: vaswani2017attention
Proceedings Title Advances in neural information processing systems
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

AttnGCG: Enhancing jailbreaking attacks on LLMs with attention manipulation

Item Type Journal Article
Author Zijun Wang
Author Haoqin Tu
Author Jieru Mei
Author Bingchen Zhao
Author Yisen Wang
Author Cihang Xie
Date 2024
Extra Citation Key: wang2024attngcg
Publication arXiv preprint arXiv:2410.09040
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Audit-LLM: Multi-agent collaboration for log-based insider threat detection

Item Type Journal Article
Author Chengyu Song
Author Linru Ma
Author Jianming Zheng
Author Jinzhi Liao
Author Hongyu Kuang
Author Lin Yang
Date 2024
Extra Citation Key: song2024audit
Publication arXiv preprint arXiv:2408.08902
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

AuditGPT: Auditing smart contracts with ChatGPT

Item Type Document
Author Shihao Xia
Author Shuai Shao
Author Mengting He
Author Tingting Yu
Author Linhai Song
Author Yiying Zhang
Abstract Automates ERC-rule verification for Ethereum smart contracts using LLMs; demonstrates comprehensive rule checking and highlights strengths/limits versus manual audits and program analyses.
Date 2024

URL	https://arxiv.org/abs/2404.04306
Extra	Citation Key: auditgpt2024
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Auditory prompt injection: Hidden commands in speech for LLM-powered agents

Item Type	Journal Article
Author	Rahul Gupta
Author	Jaehong Park
Author	Yingjie Li
Author	Wenhao Xu
Date	2025
URL	https://arxiv.org/abs/2504.15585
Extra	Citation Key: gupta2025audioprompt
Publication	arXiv preprint arXiv:2504.15585
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

AutoBnB: Multi-agent incident response with large language models

Item Type	Conference Paper
Author	Zefang Liu
Date	2025
Extra	Citation Key: liu2025multiagentcollabllm
Pages	1-6
Proceedings Title	2025 13th international symposium on digital forensics and security (ISDFS)
DOI	10.1109/ISDFS65363.2025.11012055
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Tags:

Collaboration, Large language models, large language models, Aerodynamics, Computer crime, cybersecurity, Decision making, decision-making, Digital forensics, Games, incident response, Intelligent agents, multi-agent collaboration, Training, Uncertainty

---

AutoDAN: Generating stealthy jailbreak prompts on aligned large language models

Item Type	Conference Paper
Author	Xiaogeng Liu
Author	Nan Xu
Author	Muhao Chen
Author	Chaowei Xiao
Date	2024
URL	https://openreview.net/forum?id=7Jwpw4qKkb
Extra	Citation Key: liu2024autodan
Proceedings Title	The twelfth international conference on learning representations
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Autodefense: Multi-agent llm defense against jailbreak attacks

Item Type Journal Article
Author Yifan Zeng
Author Yiran Wu
Author Xiao Zhang
Author Huazheng Wang
Author Qingyun Wu
Date 2024
Extra Citation Key: zeng2024autodefense
Publication arXiv preprint arXiv:2403.04783
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Autogen: Enabling next-gen llm applications via multi-agent conversation framework

Item Type Journal Article
Author Qingyun Wu
Author Gagan Bansal
Author Jieyu Zhang
Author Yiran Wu
Author Shaokun Zhang
Author Erkang Zhu
Author Beibin Li
Author Li Jiang
Author Xiaoyun Zhang
Author Chi Wang
Date 2023
Extra Citation Key: wu2023autogen
Publication arXiv preprint arXiv:2308.08155
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

AutoGPT: An autonomous GPT-4 experiment

Item Type Document
Author Significant Gravitas
Date 2023
URL https://github.com/Torantulino/Auto-GPT
Extra Citation Key: autogpt2023
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

GitHub repository, accessed 2025-10-08

Automate Strategy Finding with LLM in Quant Investment

Item Type	Document
Author	Zhizhuo Kou
Author	Holam Yu
Author	Junyu Luo
Author	Jingshu Peng
Author	Xujia Li
Author	Chengzhong Liu
Author	Juntao Dai
Author	Lei Chen
Author	Sirui Han
Author	Yike Guo
Abstract	We present a novel three-stage framework leveraging Large Language Models (LLMs) within a risk-aware multi-agent system for automate strategy finding in quantitative finance. Our approach addresses the brittleness of traditional deep learning models in financial applications by: employing prompt-engineered LLMs to generate executable alpha factor candidates across diverse financial data, implementing multimodal agent-based evaluation that filters factors based on market status, predictive quality while maintaining category balance, and deploying dynamic weight optimization that adapts to market conditions. Experimental results demonstrate the robust performance of the strategy in Chinese & US market regimes compared to established benchmarks. Our work extends LLMs capabilities to quantitative trading, providing a scalable architecture for financial signal extraction and portfolio construction. The overall framework significantly outperforms all benchmarks with 53.17% cumulative return on SSE50 (Jan 2023 to Jan 2024), demonstrating superior risk-adjusted performance and downside protection on the market.
Date	2025-05
URL	http://arxiv.org/abs/2409.06289
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: kou_automate_2025 DOI: 10.48550/arXiv.2409.06289
Publisher	arXiv
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Tags:

Computer Science - Machine Learning, Quantitative Finance - Portfolio Management, Quantitative Finance - Pricing of Securities

Notes:

arXiv:2409.06289 [q-fin]

Automated threat detection and response using LLM agents

Item Type	Journal Article
Author	Ramasankar Molleti
Author	Vinod Goje
Author	Puneet Luthra
Author	Prathap Raghavan
Date	2024-11
Extra	Citation Key: molleti2024threatdetectionllmagent
Volume	24
Pages	079-090
Publication	World Journal of Advanced Research and Reviews

DOI 10.30574/wjarr.2024.24.2.3329
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Automatic and universal prompt injection attacks against large language models

Item Type Journal Article
Author Xiaogeng Liu
Author Zhiyuan Yu
Author Yizhe Zhang
Author Ning Zhang
Author Chaowei Xiao
Date 2024
Extra Citation Key: liu2024automatic
Publication arXiv preprint arXiv:2403.04957
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Autonomous industrial control using an agentic framework with large language models

Item Type Journal Article
Author Javal Vyas
Author Mehmet Mercangöz
Date 2024
Extra Citation Key: vyas2024autonomous
Publication arXiv preprint arXiv:2411.05904
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

AutoPenBench: Benchmarking generative agents for penetration testing

Item Type Document
Author Luca Gioacchini
Author Marco Mellia
Author Idilio Drago
Author Alexander Delsanto
Author Giuseppe Siracusano
Author Roberto Bifulco
Date 2024
URL https://arxiv.org/abs/2410.03225
Extra Citation Key: gioacchini2024autopenbenchbenchmarkinggenerativeagents arXiv: 2410.03225 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

AutoPentest: Enhancing vulnerability management with autonomous LLM agents

Item Type Document

Author Julius Henke
Date 2025
URL https://arxiv.org/abs/2505.10321
Extra Citation Key: henke2025autopentestenhancingvulnerabilitymanagement arXiv: 2505.10321 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Awesome-LM-SSP: A reading list for large models safety, security, and privacy

Item Type Document
Author ThuCCSLab
Date 2025
URL https://github.com/ThuCCSLab/Awesome-LM-SSP
Extra Citation Key: AwesomeLMSSP
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

GitHub repository, Apache-2.0 license, 1.7k stars, accessed 2025-10-08

Backdoors stuck at the frontdoor: Multi-agent backdoor attacks that backfire

Item Type Journal Article
Author Siddhartha Datta
Author Nigel Shadbolt
Date 2022
Extra Citation Key: datta2022backdoors
Publication arXiv preprint arXiv:2201.12211
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Badagent: Inserting and activating backdoor attacks in llm agents

Item Type Journal Article
Author Yifei Wang
Author Dizhan Xue
Author Shengjie Zhang
Author Shengsheng Qian
Date 2024
Extra Citation Key: wang2024badagent
Publication arXiv preprint arXiv:2406.03007
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

BadRAG: Identifying vulnerabilities in retrieval augmented generation of large language models

Item Type Document
Author Jiaqi Xue
Author Mengxin Zheng
Author Yebowen Hu
Author Fei Liu
Author Xun Chen
Author Qian Lou
Date 2024
URL https://arxiv.org/abs/2406.00083
Extra Citation Key: xue2024badragidentifyingvulnerabilitiesretrieval arXiv: 2406.00083 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

BARTPredict: Empowering IoT security with LLM-driven cyber threat prediction

Item Type Document
Author Alaeddine Diaf
Author Abdelaziz Amara Korba
Author Nour Elislem Karabadji
Author Yacine Ghamri-Doudane
Abstract Introduces a proactive intrusion prediction framework using a fine-tuned BART model (with BERT evaluation) to forecast malicious traffic on IoT networks; achieves strong accuracy on CICIoT2023.
Date 2025
URL https://arxiv.org/abs/2501.01664
Extra Citation Key: diaf2025bartpredict
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Bells: A framework towards future proof benchmarks for the evaluation of llm safeguards

Item Type Journal Article
Author Diego Dorn
Author Alexandre Variengien
Author Charbel-Raphaël Segerie
Author Vincent Corruble
Date 2024
Extra Citation Key: dorn2024bells
Publication arXiv preprint arXiv:2406.01364
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Benchmarking and defending against indirect prompt injection attacks on large language models

Item Type Conference Paper
Author Jingwei Yi
Author Yueqi Xie
Author Bin Zhu
Author Emre Kiciman

Author	Guangzhong Sun
Author	Xing Xie
Author	Fangzhao Wu
Date	2025
URL	https://doi.org/10.1145/3690624.3709179
Extra	Citation Key: yi2025benchmarkingidirectprompt Number of pages: 12 tex.address: New York, NY, USA
Place	Toronto ON, Canada
Publisher	Association for Computing Machinery
ISBN	979-8-4007-1245-6
Pages	1809–1820
Series	Kdd '25
Proceedings Title	Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining V.1
DOI	10.1145/3690624.3709179
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Tags:

llm, defense, prompt injection attack

Benchmarking LLM-assisted blue teaming via standardized threat hunting

Item Type	Document
Author	Yuqiao Meng
Author	Luoxi Tang
Author	Feiyang Yu
Author	Xi Li
Author	Guanhua Yan
Author	Ping Yang
Author	Zhaohan Xi
Date	2025
URL	https://arxiv.org/abs/2509.23571
Extra	Citation Key: meng2025benchmarkingllmassistedblueteaming arXiv: 2509.23571 [cs.CR]
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

BLAST: a stealthy backdoor leverage attack against cooperative multi-agent deep reinforcement learning based systems

Item Type	Journal Article
Author	Yinbo Yu
Author	Saihao Yan
Author	Xueyu Yin
Author	Jing Fang
Author	Jiajia Liu
Date	2025
Extra	Citation Key: yu2025blast
Publication	arXiv preprint arXiv:2501.01593
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

BlockAgents: Towards byzantine-robust LLM-based multi-agent coordination via blockchain

Item Type	Conference Paper
Author	Bei Chen
Author	Gaolei Li
Author	Xi Lin
Author	Zheng Wang
Author	Jianhua Li
Date	2024
Extra	Citation Key: chen2024blockagents
Pages	187–192
Proceedings Title	Proceedings of the ACM turing award celebration conference-china 2024
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

BloombergGPT: a large language model for finance

Item Type	Journal Article
Author	Shijie Wu
Author	W. Branch
Author	B. Chen
Author	others
Date	2023
URL	https://arxiv.org/abs/2303.17564
Extra	Citation Key: wu2023bloomberggpt
Publication	arXiv preprint arXiv:2303.17564
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Breaking agents: Compromising autonomous llm agents through malfunction amplification

Item Type	Journal Article
Author	Boyang Zhang
Author	Yicong Tan
Author	Yun Shen
Author	Ahmed Salem
Author	Michael Backes
Author	Savvas Zannettou
Author	Yang Zhang
Date	2024
Extra	Citation Key: zhang2024breaking
Publication	arXiv preprint arXiv:2407.20859
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Breaking the code: Security assessment of AI code agents through systematic jailbreaking attacks

Item Type	Document

Author Shoumik Saha
Author Jifan Chen
Author Sam Mayers
Author Sanjay Krishna Gouda
Author Zijian Wang
Author Varun Kumar
Date 2025
URL https://arxiv.org/abs/2510.01359
Extra Citation Key: saha2025breakingcodesecurityassessment arXiv: 2510.01359 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Building trust in mental health chatbots: safety metrics and LLM-based evaluation tools

Item Type Journal Article
Author Jung In Park
Author Mahyar Abbasian
Author Iman Azimi
Author Dawn Bounds
Author Angela Jun
Author Jaesu Han
Author Robert McCarron
Author Jessica Borelli
Author Jia Li
Author Mona Mahmoudi
Author others
Date 2024
Extra Citation Key: park2024building
Publication arXiv preprint arXiv:2408.04650
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Byzantine robust cooperative multi-agent reinforcement learning as a bayesian game

Item Type Conference Paper
Author Simin Li
Author Jun Guo
Author Jingqiao Xiu
Author Ruixiao Xu
Author Xin Yu
Author Jiakai Wang
Author Aishan Liu
Author Yaodong Yang
Author Xianglong Liu
Date 2024
URL https://openreview.net/forum?id=z6KS9D1dxt
Extra Citation Key: li2024byzantine
Proceedings Title The twelfth international conference on learning representations
Date Added 10/20/2025, 3:50:52 PM

Modified 10/20/2025, 3:50:52 PM

Byzantine-robust decentralized coordination of LLM agents

Item Type Document
Author Yongrae Jo
Author Chanik Park
Date 2025
URL https://arxiv.org/abs/2507.14928
Extra Citation Key: jo2025byzantinerobustdecentralizedcoordinationllm arXiv: 2507.14928 [cs.DC]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

CAIN: Hijacking LLM-humans conversations via malicious system prompts

Item Type Document
Author Viet Pham
Author Thai Le
Date 2025
URL https://arxiv.org/abs/2505.16888
Extra Citation Key: pham2025cainhijackingllmhumansconversations arXiv: 2505.16888 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Camel: Communicative agents for" mind" exploration of large language model society

Item Type Journal Article
Author Guohao Li
Author Hasan Hammoud
Author Hani Itani
Author Dmitrii Khizbullin
Author Bernard Ghanem
Date 2023
Extra Citation Key: li2023camel
Volume 36
Pages 51991–52008
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Can large language models beat wall street? Evaluating GPT-4’s impact on financial decision-making with MarketSenseAI

Item Type Journal Article
Author George Fatouros
Author Kostas Metaxas
Author John Soldatos

Author Dimosthenis Kyriazis
Date 2024
Extra Citation Key: fatouros2024can Publisher: Springer
Pages 1–26
Publication Neural Computing and Applications
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Can large language models trade? Testing financial theories with LLM agents in market simulations

Item Type Journal Article
Author Alejandro Lopez-Lira
Date 2025
URL https://arXiv.org/abs/2504.10789
Extra Citation Key: lopez2025canllmtrade
Publication arXiv preprint arXiv:2504.10789
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Can LLM-based Financial Investing Strategies Outperform the Market in Long Run?

Item Type Document
Author Weixian Waylon Li
Author Hyeonjun Kim
Author Mihai Cucuringu
Author Tiejun Ma
Abstract Large Language Models (LLMs) have recently been leveraged for asset pricing tasks and stock trading applications, enabling AI agents to generate investment decisions from unstructured financial data. However, most evaluations of LLM timing-based investing strategies are conducted on narrow timeframes and limited stock universes, overstating effectiveness due to survivorship and data-snooping biases. We critically assess their generalizability and robustness by proposing FINSABER, a backtesting framework evaluating timing-based strategies across longer periods and a larger universe of symbols. Systematic backtests over two decades and 100+ symbols reveal that previously reported LLM advantages deteriorate significantly under broader cross-section and over a longer-term evaluation. Our market regime analysis further demonstrates that LLM strategies are overly conservative in bull markets, underperforming passive benchmarks, and overly aggressive in bear markets, incurring heavy losses. These findings highlight the need to develop LLM strategies that are able to prioritise trend detection and regime-aware risk controls over mere scaling of framework complexity.
Date 2025-08
URL http://arxiv.org/abs/2505.07078
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: li_can_2025 DOI: 10.48550/arXiv.2505.07078
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Trading and Market Microstructure, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2505.07078 [q-fin]

Can llms hack enterprise networks? Autonomous assumed breach penetration-testing active directory networks

Item Type Journal Article
Author Andreas Happe
Author Jürgen Cito
Abstract Traditional enterprise penetration-testing, while critical for validating defenses and uncovering vulnerabilities, is often limited by high operational costs and the scarcity of human expertise. This paper investigates the feasibility and effectiveness of using Large Language Model (LLM)-driven autonomous systems to address these challenges in real-world Active Directory (AD) enterprise networks.We introduce a novel prototype, cochise, designed to employ LLMs to autonomously perform Assumed Breach penetration-testing against enterprise networks. Our system represents the first demonstration of a fully autonomous, LLM-driven framework capable of compromising accounts within a real-life Microsoft Active Directory testbed, the Game of Active Directory (GOAD). The evaluation deliberately utilizes GOAD to capture the intricate interactions and sometimes nondeterministic outcomes of live network penetration-testing, moving beyond the limitations of synthetic benchmarks.We perform our empirical evaluation using five LLMs, comparing reasoning to non-reasoning models as well as including open-weight models. Through comprehensive quantitative and qualitative analysis, incorporating insights from cybersecurity experts, we demonstrate that autonomous LLMs can effectively conduct Assumed Breach simulations. Key findings highlight their ability to dynamically adapt attack strategies, perform inter-context attacks (e.g., web application audits, social engineering, and unstructured data analysis for credentials), and generate scenario-specific attack parameters like realistic password candidates. The prototype also exhibits robust self-correction mechanisms, automatically installing missing tools and rectifying invalid command generations.Critically, we find that the associated costs are competitive with, and often significantly lower than, those incurred by professional human penetration testers, suggesting a path toward democratizing access to essential security testing for organizations with budgetary constraints. However, our research also illuminates existing limitations, including instances of LLM “going down rabbit holes”, challenges in comprehensive information transfer between planning and execution modules, and critical safety concerns that necessitate human oversight. Our findings lay foundational groundwork for future software engineering research into LLM-driven cybersecurity automation, emphasizing that the prototype's underlying LLM-driven architecture and techniques are domain-agnostic and hold promise for improving autonomous LLM usage in broader software engineering domains. The source code, traces, and analyzed logs are open-sourced to foster collective cybersecurity and future research.
Date 2025-09
URL https://doi.org/10.1145/3766895
Extra Citation Key: Happe2025LLM Place: New York, NY, USA Publisher: Association for Computing Machinery
Publication ACM Trans. Softw. Eng. Methodol.
DOI 10.1145/3766895
ISSN 1049-331X
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

Large Language Models, Enterprise Networks, Security Capability Evaluation

Notes:

Just Accepted

Certifiably robust rag against retrieval corruption

Item Type Journal Article
Author Chong Xiang
Author Tong Wu
Author Zexuan Zhong
Author David Wagner
Author Danqi Chen
Author Prateek Mittal
Date 2024
Extra Citation Key: xiang2024certifiably
Publication arXiv preprint arXiv:2405.15556
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Certifying llm safety against adversarial prompting

Item Type Journal Article
Author Aounon Kumar
Author Chirag Agarwal
Author Suraj Srinivas
Author Aaron Jiaxun Li
Author Soheil Feizi
Author Himabindu Lakkaraju
Date 2023
Extra Citation Key: kumar2023certifying
Publication arXiv preprint arXiv:2309.02705
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic RL

Item Type Document
Author Weizhen Li
Author Jianbo Lin
Author Zhuosong Jiang
Author Jingyi Cao
Author Xinpeng Liu
Author Jiayu Zhang
Author Zhenqiang Huang
Author Qianben Chen
Author Weichen Sun
Author Qiexiang Wang
Author Hongxuan Lu
Author Tianrui Qin
Author Chenghao Zhu
Author Yi Yao
Author Shuying Fan
Author Xiaowan Li
Author Tiannan Wang
Author Pai Liu

Author King Zhu
Author He Zhu
Author Dingfeng Shi
Author Piaohong Wang
Author Yeyi Guan
Author Xiangru Tang
Author Minghao Liu
Author Yuchen Eleanor Jiang
Author Jian Yang
Author Jiaheng Liu
Author Ge Zhang
Author Wangchunshu Zhou
Date 2025
URL https://arxiv.org/abs/2508.13167
Extra Citation Key: li2025chainofagentsendtoendagentfoundation arXiv: 2508.13167 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Chain-of-scrutiny: Detecting backdoor attacks for large language models

Item Type Journal Article
Author Xi Li
Author Yusen Zhang
Author Renze Lou
Author Chen Wu
Author Jiaqi Wang
Date 2024
Extra Citation Key: li2024chain
Publication arXiv preprint arXiv:2406.05948
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Chat bankman-fried: an exploration of LLM alignment in finance

Item Type Journal Article
Author Claudia Biancotti
Author Carolina Camassa
Author Andrea Coletta
Author Oliver Giudice
Author Aldo Glielmo
Date 2024
Extra Citation Key: biancotti2024chat
Publication arXiv preprint arXiv:2411.11853
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Chateval: Towards better llm-based evaluators through multi-agent debate

Item Type Journal Article
Author Chi-Min Chan
Author Weize Chen
Author Yusheng Su
Author Jianxuan Yu
Author Wei Xue
Author Shanghang Zhang
Author Jie Fu
Author Zhiyuan Liu
Date 2023
Extra Citation Key: chan2023chateval
Publication arXiv preprint arXiv:2308.07201
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles

Item Type Conference Paper
Author Jiawei Zhang
Author Chejian Xu
Author Bo Li
Date 2024
Extra Citation Key: zhang2024chatscene
Pages 15459–15469
Proceedings Title Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Clinician voices on ethics of LLM integration in healthcare: A thematic analysis of ethical concerns and implications

Item Type Journal Article
Author Tala Mirzaei
Author Leila Amini
Author Pouyan Esmaeilzadeh
Date 2024
Extra Citation Key: mirzaei2024clinician Publisher: Springer
Volume 24
Pages 250
Publication BMC Medical Informatics and Decision Making
Issue 1
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Cloud infrastructure management in the age of AI agents

Item Type Journal Article
Author Zhenning Yang

Author Archit Bhatnagar
Author Yiming Qiu
Author Tongyuan Miao
Author Patrick Tser Jern Kon
Author Yunming Xiao
Author Yibo Huang
Author Martin Casado
Author Ang Chen
Abstract Cloud infrastructure requires significant manual DevOps effort. The paper argues for LLM-powered agents to automate management tasks across SDK/CLI/IaC/web UIs, reports early findings, and outlines challenges and research opportunities.
Date 2025
URL https://arxiv.org/pdf/2506.12270v1
Extra Citation Key: yang2025cloudinfrastructuremanagement Place: New York, NY, USA Publisher: Association for Computing Machinery
Volume 59
Publication ACM SIGOPS Operating Systems Review
DOI 10.1145/3759441.3759443
Issue 1
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Cloud investigation automation framework (CIAF): An AI-driven approach to cloud forensics

Item Type Document
Author Dalal Alharthi
Author Ivan Roberto Kawaminami Garcia
Date 2025
URL https://arxiv.org/abs/2510.00452
Extra Citation Key: alharthi2025cloudinvestigationautomationframework arXiv: 2510.00452 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Combating adversarial attacks with multi-agent debate

Item Type Journal Article
Author Steffi Chern
Author Zhen Fan
Author Andy Liu
Date 2024
Extra Citation Key: chern2024combating
Publication arXiv preprint arXiv:2401.05998
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Combining fine-tuning and LLM-based agents for intuitive smart contract auditing with justifications

Item Type Document
Author Wei Ma

Author Daoyuan Wu
Author Yuqiang Sun
Author Tianwen Wang
Author Shangqing Liu
Author Jian Zhang
Author Yue Xue
Author Yang Liu
Abstract iAudit: a two-stage fine-tuning (Detector/Reasoner) plus agent Ranker/Critic framework for smart-contract auditing; achieves high accuracy/F1 and improved explanatory causes on real vulnerabilities.
Date 2024
URL https://arxiv.org/abs/2403.16073
Extra Citation Key: finetuningagents2024
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Commercial LLM agents are already vulnerable to simple yet dangerous attacks

Item Type Document
Author Ang Li
Author Yin Zhou
Author Vethavikashini Chithrra Raghuram
Author Tom Goldstein
Author Micah Goldblum
Date 2025
URL https://arxiv.org/abs/2502.08586
Extra Citation Key: li2025commercialllmagentsvulnerable arXiv: 2502.08586 [cs.LG]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Confront insider threat: Precise anomaly detection in behavior logs based on LLM fine-tuning

Item Type Conference Paper
Author Shuang Song
Author Yifei Zhang
Author Neng Gao
Editor Owen Rambow
Editor Leo Wanner
Editor Marianna Apidianaki
Editor Hend Al-Khalifa
Editor Barbara Di Eugenio
Editor Steven Schockaert
Abstract Anomaly-based detection is effective against evolving insider threats but still suffers from low precision. Current data processing can result in information loss, and models often struggle to distinguish between benign anomalies and actual threats. Both issues hinder precise detection. To address these issues, we propose a precise anomaly detection solution for behavior logs based on Large Language Model (LLM) fine-tuning. By representing user behavior in natural language, we minimize information loss. We fine-tune the LLM with a user behavior pattern contrastive task for anomaly detection, using a two-stage strategy: first learning general behavior patterns, then refining with user-specific data to improve differentiation between benign anomalies and threats. We also implement a fine-grained threat tracing mechanism to provide behavior-level audit trails. To the best of our knowledge, our solution is the first to apply LLM fine-tuning in insider threat detection, achieving an F1 score of 0.8941 on the CERT v6.2 dataset, surpassing all baselines.

Date	2025-01
URL	https://aclanthology.org/2025.coling-main.574/
Extra	Citation Key: song-etal-2025-confront
Place	Abu Dhabi, UAE
Publisher	Association for Computational Linguistics
Pages	8589–8601
Proceedings Title	Proceedings of the 31st international conference on computational linguistics
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Construction and evaluation of LLM-based agents for semi-autonomous penetration testing

Item Type	Document
Author	Masaya Kobayashi
Author	Masane Fuchi
Author	Amar Zanashir
Author	Tomonori Yoneda
Author	Tomohiro Takagi
Date	2025
URL	https://arxiv.org/abs/2502.15506
Extra	Citation Key: kobayashi2025constructionevaluationllmbasedagents arXiv: 2502.15506 [cs.CR]
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

ContestTrade: A Multi-Agent Trading System Based on Internal Contest Mechanism

Item Type	Document
Author	Li Zhao
Author	Rui Sun
Author	Zuoyou Jiang
Author	Bo Yang
Author	Yuxiao Bai
Author	Mengting Chen
Author	Xinyang Wang
Author	Jing Li
Author	Zuo Bai
Abstract	In financial trading, large language model (LLM)-based agents demonstrate significant potential. However, the high sensitivity to market noise undermines the performance of LLM-based trading systems. To address this limitation, we propose a novel multi-agent system featuring an internal competitive mechanism inspired by modern corporate management structures. The system consists of two specialized teams: (1) Data Team - responsible for processing and condensing massive market data into diversified text factors, ensuring they fit the model's constrained context. (2) Research Team - tasked with making parallelized multipath trading decisions based on deep research methods. The core innovation lies in implementing a real-time evaluation and ranking mechanism within each team, driven by authentic market feedback. Each agent's performance undergoes continuous scoring and ranking, with only outputs from top-performing agents being adopted. The design enables the system to adaptively adjust to dynamic environment, enhances robustness against market noise and ultimately delivers superior trading performance. Experimental results demonstrate that our proposed system significantly outperforms prevailing multi-agent systems and traditional quantitative investment methods across diverse evaluation metrics. ContestTrade is open-sourced on GitHub at https://github.com/FinStep-AI/ContestTrade.
Date	2025-08

Short Title ContestTrade
URL http://arxiv.org/abs/2508.00554
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: zhao_contesttrade_2025 DOI: 10.48550/arXiv.2508.00554
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Computation and Language, Quantitative Finance - Trading and Market Microstructure, Quantitative Finance - Computational Finance

Notes:

arXiv:2508.00554 [q-fin]

Cooperative multi-agent learning: The state of the art

Item Type Journal Article
Author Liviu Panait
Author Sean Luke
Date 2005
Extra Citation Key: panait2005cooperative Publisher: Springer
Volume 11
Pages 387–434
Publication Autonomous agents and multi-agent systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

CORBA: Contagious recursive blocking attacks on multi-agent systems based on large language models

Item Type Journal Article
Author Zhenhong Zhou
Author Zherui Li
Author Jie Zhang
Author Yuanhe Zhang
Author Kun Wang
Author Yang Liu
Author Qing Guo
Date 2025
Extra Citation Key: zhou2025corba
Publication arXiv preprint arXiv:2502.14529
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

CORTEX: Collaborative LLM agents for high-stakes alert triage

Item Type Document

Author Bowen Wei
Author Yuan Shen Tay
Author Howard Liu
Author Jinhao Pan
Author Kun Luo
Author Ziwei Zhu
Author Chris Jordan
Date 2025
URL https://arxiv.org/abs/2510.00311
Extra Citation Key: wei2025cortexcollaborativellmagents arXiv: 2510.00311 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

CVE-bench: Benchmarking LLM-based software engineering agent's ability to repair real-world CVE vulnerabilities

Item Type Conference Paper
Author Peiran Wang
Author Xiaogeng Liu
Author Chaowei Xiao
Editor Luis Chiruzzo
Editor Alan Ritter
Editor Lu Wang
Abstract Automated vulnerability repair is a crucial field within software engineering and security research. Large Language Models (LLMs) and LLM agents have demonstrated significant potential in this domain by understanding descriptions in natural language and generating corresponding formal code. Although the coding capabilities of LLMs have advanced rapidly, evaluation benchmarks for real-world programming setups are still lagging, preventing the development of LLM and LLM agents in real-world vulnerability repair. To this end, we introduce CVE-Bench, an evaluation framework consisting of 509 Common Vulnerabilities and Exposures (CVEs) from four programming languages and 120 popular open-source repositories. Unlike previous vulnerability repair benchmarks, which only involve the code input and output, we provide LLM agents with a test environment that simulates the real-world vulnerability repair process. This environment provides multiple levels of CVE information modeling, such as black-box testing and white-box testing. It enables the agents to use static analysis tools to assist their repair process. Our evaluation reveals that the SWE-agent can only repair 21% of vulnerabilities at its best. Furthermore, they lack expert knowledge about how to use the analysis tool to assist in vulnerability repair.
Date 2025-04
URL https://aclanthology.org/2025.naacl-long.212/
Extra Citation Key: wang-etal-2025-cve
Place Albuquerque, New Mexico
Publisher Association for Computational Linguistics
ISBN 979-8-89176-189-6
Pages 4207–4224
Proceedings Title 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)
DOI 10.18653/v1/2025.naacl-long.212
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

CyberSleuth: Autonomous blue-team LLM agent for web attack forensics

Item Type Document
Author Stefano Fumero
Author Kai Huang
Author Matteo Boffa
Author Danilo Giordano
Author Marco Mellia
Author Zied Ben Houidi
Author Dario Rossi
Date 2025
URL https://arxiv.org/abs/2508.20643
Extra Citation Key: fumero2025cybersleuthautonomousblueteamllm arXiv: 2508.20643 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

CyberSOCEval: Benchmarking llms capabilities for malware analysis and threat intelligence reasoning

Item Type Document
Author Lauren Deason
Author Adam Bali
Author Ciprian Bejean
Author Diana Bolocan
Author James Crnkovich
Author Ioana Croitoru
Author Krishna Durai
Author Chase Midler
Author Calin Miron
Author David Molnar
Author Brad Moon
Author Bruno Ostarcevic
Author Alberto Peltea
Author Matt Rosenberg
Author Catalin Sandu
Author Arthur Saputkin
Author Sagar Shah
Author Daniel Stan
Author Ernest Szocs
Author Shengye Wan
Author Spencer Whitman
Author Sven Krasser
Author Joshua Saxe
Date 2025
URL https://arxiv.org/abs/2509.20166
Extra Citation Key: deason2025cybersocevalbenchmarkingllmscapabilities arXiv: 2509.20166 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

D-CIPHER: Dynamic collaborative intelligent multi-agent system with planner and heterogeneous executors for offensive security

Item Type Document
Author Meet Udeshi
Author Minghao Shao
Author Haoran Xi
Author Nanda Rani
Author Kimberly Milner
Author Venkata Sai Charan Putrevu
Author Brendan Dolan-Gavitt
Author Sandeep Kumar Shukla
Author Prashanth Krishnamurthy
Author Farshad Khorrami
Author Ramesh Karri
Author Muhammad Shafique
Date 2025
URL https://arxiv.org/abs/2502.10931
Extra Citation Key: udeshi2025dcipherdynamiccollaborativeintelligent arXiv: 2502.10931 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

DataSentinel: a game-theoretic detection of prompt injection attacks

Item Type Conference Paper
Author Yupei Liu
Author Yuqi Jia
Author Jinyuan Jia
Author Dawn Song
Author Neil Zhenqiang Gong
Date 2025-05
URL https://doi.ieeecomputersociety.org/10.1109/SP61157.2025.00250
Extra Citation Key: Liu2025DataSentinel
Place Los Alamitos, CA, USA
Publisher IEEE Computer Society
Pages 2190-2208
Proceedings Title 2025 IEEE symposium on security and privacy (SP)
DOI 10.1109/SP61157.2025.00250
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

Security, Benchmark testing, Codes, Data privacy, Optimization

---

Deal: Decoding-time alignment for large language models

Item Type Journal Article
Author James Y Huang
Author Sailik Sengupta
Author Daniele Bonadiman
Author Yi-an Lai
Author Arshit Gupta

Author Nikolaos Pappas
Author Saab Mansour
Author Katrin Kirchhoff
Author Dan Roth
Date 2024
Extra Citation Key: huang2024deal
Publication arXiv preprint arXiv:2402.06147
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Deep learning with differential privacy

Item Type Conference Paper
Author Martin Abadi
Author Andy Chu
Author Ian Goodfellow
Author H Brendan McMahan
Author Ilya Mironov
Author Kunal Talwar
Author Li Zhang
Date 2016
Extra Citation Key: abadi2016deep
Pages 308–318
Proceedings Title Proceedings of the 2016 ACM SIGSAC conference on computer and communications security
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Defeating prompt injections by design

Item Type Document
Author Edoardo Debenedetti
Author Ilia Shumailov
Author Tianqi Fan
Author Jamie Hayes
Author Nicholas Carlini
Author Daniel Fabian
Author Christoph Kern
Author Chongyang Shi
Author Andreas Terzis
Author Florian Tramèr
Date 2025
URL https://arxiv.org/abs/2503.18813
Extra Citation Key: debenedetti2025defeatingpromptinjectionsdesign arXiv: 2503.18813 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Deficiency of large language models in finance: An empirical examination of hallucination

Item Type Journal Article
Author Haoqiang Kang
Author Xiao-Yang Liu
Date 2023
Extra Citation Key: kang2023deficiency
Publication arXiv preprint arXiv:2311.15548
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Defining and characterizing reward hacking

Item Type Conference Paper
Author Joar Skalse
Author Nikolaus H. R. Howe
Author Dmitrii Krasheninnikov
Author David Krueger
Date 2022
Extra Citation Key: 10.5555/3600270.3600957 Number of pages: 12 tex.address: Red Hook, NY, USA tex.articleno: 687
Place New Orleans, LA, USA
Publisher Curran Associates Inc.
ISBN 978-1-7138-7108-8
Series Nips '22
Proceedings Title Proceedings of the 36th international conference on neural information processing systems
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Delucionqa: Detecting hallucinations in domain-specific question answering

Item Type Journal Article
Author Mobashir Sadat
Author Zhengyu Zhou
Author Lukas Lange
Author Jun Araki
Author Arsalan Gundroo
Author Bingqing Wang
Author Rakesh R Menon
Author Md Rizwan Parvez
Author Zhe Feng
Date 2023
Extra Citation Key: sadat2023delucionqa
Publication arXiv preprint arXiv:2312.05200
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

DemonAgent: Dynamically encrypted multi-backdoor implantation attack on LLM-based agent

Item Type Journal Article

Author Pengyu Zhu
Author Zhenhong Zhou
Author Yuanhe Zhang
Author Shilinlu Yan
Author Kun Wang
Author Sen Su
Date 2025
Extra Citation Key: zhu2025demonagent
Publication arXiv preprint arXiv:2502.12575
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Demonstrating specification gaming in reasoning models

Item Type Document
Author Alexander Bondarenko
Author Denis Volk
Author Dmitrii Volkov
Author Jeffrey Ladish
Date 2025
URL https://arxiv.org/abs/2502.13295
Extra Citation Key: bondarenko2025demonstratingspecificationgamingreasoning arXiv: 2502.13295 [cs.AI]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Deploying privacy guardrails for llms: a comparative analysis of real-world applications

Item Type Document
Author Shubhi Asthana
Author Bing Zhang
Author Ruchi Mahindru
Author Chad DeLuca
Author Anna Lisa Gentile
Author Sandeep Gopisetty
Abstract Presents and compares deployments of the OneShield Privacy Guard framework across enterprise/open-source settings; reports high multilingual PII detection F1 and reduced manual effort in PR triage.
Date 2025
URL https://arxiv.org/abs/2501.12456
Extra Citation Key: asthana2025privacyguardrails
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents

Item Type Journal Article
Author Zihao Wang
Author Shaofei Cai
Author Guanzhou Chen

Author Anji Liu
Author Xiaojian Shawn Ma
Author Yitao Liang
Date 2023
Extra Citation Key: wang2023describe
Volume 36
Pages 34153–34189
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

DiaHalu: a dialogue-level hallucination evaluation benchmark for large language models

Item Type Journal Article
Author Kedi Chen
Author Qin Chen
Author Jie Zhou
Author Yishen He
Author Liang He
Date 2024
Extra Citation Key: chen2024diahalu
Publication arXiv preprint arXiv:2403.00896
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Dialectical alignment: Resolving the tension of 3h and security threats of llms

Item Type Journal Article
Author Shu Yang
Author Jiayuan Su
Author Han Jiang
Author Mengdi Li
Author Keyuan Cheng
Author Muhammad Asif Ali
Author Lijie Hu
Author Di Wang
Date 2024
Extra Citation Key: yang2024dialectical
Publication arXiv preprint arXiv:2404.00486
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Discovering language model behaviors with model-written evaluations

Item Type Journal Article
Author Ethan Perez
Author Sam Ringer
Author Kamilė Lukošiūtė

Author Karina Nguyen
Author Edwin Chen
Author Scott Heiner
Author Craig Pettit
Author Catherine Olsson
Author Sandipan Kundu
Author Saurav Kadavath
Author others
Date 2022
Extra Citation Key: perez2022discovering
Publication arXiv preprint arXiv:2212.09251
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

DoomArena: a framework for testing AI agents against evolving security threats

Item Type Conference Paper
Author Léo Boisvert
Author Abhay Puri
Author Gabriel Huang
Author Mihir Bansal
Author Chandra Kiran Reddy Evuru
Author Avinandan Bose
Author Maryam Fazel
Author Quentin Cappart
Author Alexandre Lacoste
Author Alexandre Drouin
Author Krishnamurthy Dj Dvijotham
Date 2025
URL https://openreview.net/forum?id=GanmYQ0RpE
Extra Citation Key: boisvert2025doomarena
Proceedings Title Second conference on language modeling
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Easytool: Enhancing llm-based agents with concise tool instruction

Item Type Journal Article
Author Siyu Yuan
Author Kaitao Song
Author Jiangjie Chen
Author Xu Tan
Author Yongliang Shen
Author Ren Kan
Author Dongsheng Li
Author Deqing Yang
Date 2024
Extra Citation Key: yuan2024easytool
Publication arXiv preprint arXiv:2401.06201

Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Eia: Environmental injection attack on generalist web agents for privacy leakage

Item Type Journal Article
Author Zeyi Liao
Author Lingbo Mo
Author Chejian Xu
Author Mintong Kang
Author Jiawei Zhang
Author Chaowei Xiao
Author Yuan Tian
Author Bo Li
Author Huan Sun
Date 2024
Extra Citation Key: liao2024eia
Publication arXiv preprint arXiv:2409.11295
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Embedding-based classifiers can detect prompt injection attacks

Item Type Journal Article
Author Md Ahsan Ayub
Author Subhabrata Majumdar
Date 2024
Extra Citation Key: ayub2024embedding
Publication arXiv preprint arXiv:2410.22284
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Embodied multi-modal agent trained by an llm from a parallel textworld

Item Type Conference Paper
Author Yijun Yang
Author Tianyi Zhou
Author Kanxue Li
Author Dapeng Tao
Author Lusong Li
Author Li Shen
Author Xiaodong He
Author Jing Jiang
Author Yuhui Shi
Date 2024
Extra Citation Key: yang2024embodied
Pages 26275–26285
Proceedings Title Proceedings of the IEEE/CVF conference on computer vision and pattern recognition

Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Empowering users in digital privacy management through interactive LLM-based agents

Item Type Journal Article
Author Bolun Sun
Author Yifan Zhou
Author Haiyun Jiang
Date 2024
Extra Citation Key: sun2024empowering
Publication arXiv preprint arXiv:2410.11906
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Emulated disalignment: Safety alignment for large language models may backfire!

Item Type Journal Article
Author Zhanhui Zhou
Author Jie Liu
Author Zhichen Dong
Author Jiaheng Liu
Author Chao Yang
Author Wanli Ouyang
Author Yu Qiao
Date 2024
Extra Citation Key: zhou2024emulated
Publication arXiv preprint arXiv:2402.12343
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Enforcing Cybersecurity Constraints for LLM-driven Robot Agents for Online Transactions

Item Type Conference Paper
Author Shraddha Pradipbhai Shah
Author Aditya Vilas Deshpande
Abstract The integration of Large Language Models (LLMs) into autonomous robotic agents for conducting online transactions poses significant cybersecurity challenges. This study aims to enforce robust cybersecurity constraints to mitigate the risks associated with data breaches, transaction fraud, and system manipulation. The background focuses on the rise of LLM-driven robotic systems in e-commerce, finance, and service industries, alongside the vulnerabilities they introduce. A novel security architecture combining blockchain technology with multi-factor authentication (MFA) and real-time anomaly detection was implemented to safeguard transactions. Key performance metrics such as transaction integrity, response time, and breach detection accuracy were evaluated, showing improved security and system performance. The results highlight that the proposed architecture reduced fraudulent transactions by 90%, improved breach detection accuracy to 98%, and ensured secure transaction validation within a latency of 0.05 seconds. These findings emphasize the importance of cybersecurity in the deployment of LLM-driven robotic systems and suggest a framework adaptable to various online platforms.
Date 2024-09
URL http://arxiv.org/abs/2503.15546

Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: shah_enforcing_2024
Pages 1–6
Proceedings Title 2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)
DOI 10.1109/ICDSCNC62492.2024.10939862
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Computers and Society, Computer Science - Cryptography and Security

Notes:

arXiv:2503.15546 [cs]

Enhancing anomaly detection in financial markets with an llm-based multi-agent framework

Item Type Journal Article
Author Taejin Park
Date 2024
Extra Citation Key: park2024enhancing
Publication arXiv preprint arXiv:2403.19735
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Enhancing diagnostic accuracy through multi-agent conversations: using large language models to mitigate cognitive bias

Item Type Journal Article
Author Yu He Ke
Author Rui Yang
Author Sui An Lie
Author Taylor Xin Yi Lim
Author Hairil Rizal Abdullah
Author Daniel Shu Wei Ting
Author Nan Liu
Date 2024
Extra Citation Key: ke2024enhancing
Publication arXiv preprint arXiv:2401.14589
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Enhancing fake news detection with large language models through multi-agent debates

Item Type Conference Paper
Author Korir Nancy Jeptoo
Author Chengjie Sun
Date 2024

Extra Citation Key: jeptoo2024enhancing
Publisher Springer
Pages 474–486
Proceedings Title CCF international conference on natural language processing and chinese computing
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Enhancing llm-based autonomous driving agents to mitigate perception attacks

Item Type Journal Article
Author Ruoyu Song
Author Muslum Ozgur Ozmen
Author Hyungsub Kim
Author Antonio Bianchi
Author Z Berkay Celik
Date 2024
Extra Citation Key: song2024enhancing
Publication arXiv preprint arXiv:2409.14488
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Enhancing prompt injection attacks to llms via poisoning alignment

Item Type Conference Paper
Author Zedian Shao
Author Hongbin Liu
Author Jaden Mu
Author Neil Zhenqiang Gong
Date 2024
URL https://api.semanticscholar.org/CorpusID:273502594
Extra Citation Key: Shao2024EnhancingPI
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Enhancing robustness of LLM-driven multi-agent systems through randomized smoothing

Item Type Journal Article
Author Jinwei HU
Author Yi DONG
Author Zhengtao DING
Author Xiaowei HUANG
Abstract This paper presents a defense framework for enhancing the safety of Large Language Model (LLM)-empowered Multi-Agent Systems (MAS) in safety-critical domains such as aerospace. We apply randomized smoothing—a statistical robustness certification technique—to the MAS consensus context, enabling probabilistic guarantees on agent decisions under adversarial influence. Unlike traditional verification methods, our approach operates in black-box settings and employs a two-stage adaptive sampling mechanism to balance robustness and computational efficiency. Simulation results demonstrate that our method effectively prevents the propagation of adversarial behaviors and hallucinations while maintaining consensus performance. This work provides a practical and scalable path toward safe deployment of LLM-based MAS in real-world high-stakes environments.

Date	2025
URL	https://www.sciencedirect.com/science/article/pii/S1000936125003851
Extra	Citation Key: HU2025103779
Pages	103779
Publication	Chinese Journal of Aeronautics
DOI	https://doi.org/10.1016/j.cja.2025.103779
ISSN	1000-9361
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Tags:

Large language models, Multi-agent systems, Consensus seeking, Randomized smoothing, Safe planning

---

ERBench: An entity-relationship based automatically verifiable hallucination benchmark for large language models

Item Type	Journal Article
Author	Jio Oh
Author	Soyeon Kim
Author	Junseok Seo
Author	Jindong Wang
Author	Ruochen Xu
Author	Xing Xie
Author	Steven Whang
Date	2025
Extra	Citation Key: oh2025erbench
Volume	37
Pages	53064–53101
Publication	Advances in Neural Information Processing Systems
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

Ethics education for healthcare professionals in the era of ChatGPT and other large language models: Do we still need it?

Item Type	Journal Article
Author	Vasiliki Rahimzadeh
Author	Kristin Kostick-Quenet
Author	Jennifer Blumenthal Barby
Author	Amy L McGuire
Date	2023
Extra	Citation Key: rahimzadeh2023ethics Publisher: Taylor & Francis
Volume	23
Pages	17–27
Publication	The American journal of bioethics
Issue	10
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Evaluating cultural and social awareness of LLM web agents

Item Type Journal Article
Author Haoyi Qiu
Author Alexander R Fabbri
Author Divyansh Agarwal
Author Kung-Hsiang Huang
Author Sarah Tan
Author Nanyun Peng
Author Chien-Sheng Wu
Date 2024
Extra Citation Key: qiu2024evaluating
Publication arXiv preprint arXiv:2410.23252
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Evaluating the potential and pitfalls of AI-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: scoping review

Item Type Journal Article
Author Sadia Azmin Anisha
Author Arkendu Sen
Author Chris Bain
Date 2024
Extra Citation Key: anisha2024evaluating Publisher: JMIR Publications Toronto, Canada
Volume 26
Pages e56114
Publication Journal of Medical Internet Research
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review

Item Type Journal Article
Author Hang Ding
Author Joshua Simmich
Author Atiyeh Vaezipour
Author Nicole Andrews
Author Trevor Russell
Date 2024
Extra Citation Key: ding2024evaluation Publisher: Oxford University Press
Volume 31
Pages 746–761
Publication Journal of the American Medical Informatics Association
Issue 3
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Evil geniuses: Delving into the safety of llm-based agents

Item Type Journal Article
Author Yu Tian
Author Xiao Yang
Author Jingyuan Zhang
Author Yinpeng Dong
Author Hang Su
Date 2023
Extra Citation Key: tian2023evil
Publication arXiv preprint arXiv:2311.11855
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

EvoFlow: Evolving diverse agentic workflows on the fly

Item Type Journal Article
Author Guibin Zhang
Author Kaijie Chen
Author Guancheng Wan
Author Heng Chang
Author Hong Cheng
Author Kun Wang
Author Shuyue Hu
Author Lei Bai
Date 2025
Extra Citation Key: zhang2025evoflow
Publication arXiv preprint arXiv:2502.07373
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Examining the role of artificial intelligence in cyber security (CS): A systematic review for preventing prospective solutions in financial transactions

Item Type Journal Article
Author Mahfujur Rahman Faraji
Author Fisan Shikder
Author Md Hasibul Hasan
Author Md Mominul Islam
Author Umme Kulsum Akter
Date 2024
Extra Citation Key: faraji2024examining
Volume 5
Pages 4766–4782
Publication International Journal
Issue 10
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

ExCyTIn-bench: Evaluating LLM agents on cyber threat investigation

Item Type Document
Author Yiran Wu
Author Mauricio Velazco
Author Andrew Zhao
Author Manuel Raúl Meléndez Luján
Author Srisuma Movva
Author Yogesh K Roy
Author Quang Nguyen
Author Roberto Rodriguez
Author Qingyun Wu
Author Michael Albada
Author Julia Kiseleva
Author Anand Mudgerikar
Date 2025
URL https://arxiv.org/abs/2507.14201
Extra Citation Key: wu2025excytinbenchevaluatingllmagents arXiv: 2507.14201 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Exploiting communication protocols in multi-agent LLM systems: Risks and defenses

Item Type Journal Article
Author Sahar Abdelnabi
Author Zhendong Zhang
Author Mario Fritz
Date 2025
URL https://arxiv.org/abs/2502.01822
Extra Citation Key: abdelnabi2025agentcomm
Publication arXiv preprint arXiv:2502.01822
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Exploring jailbreak attacks on llms through intent decomposition and prompt reconstruction

Item Type Conference Paper
Author Tianyi Cui
Author et al.
Date 2025
URL https://aclanthology.org/2025.findings-acl.1067.pdf
Extra Citation Key: cui2025intentjailbreak
Pages –
Proceedings Title Findings of the association for computational linguistics: ACL 2025
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

Accessed: 2025-10-20

Exploring model welfare

Item Type Document
Author Anthropic
Date 2025
URL https://www.anthropic.com/news/exploring-model-welfare
Extra Citation Key: anthropic2025claude
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

Accessed: 2025-10-08

FairMindSim: Alignment of behavior, emotion, and belief in humans and LLM agents amid ethical dilemmas

Item Type Journal Article
Author Yu Lei
Author Hao Liu
Author Chengxing Xie
Author Songjia Liu
Author Zhiyu Yin
Author Canyu Chen
Author Guohao Li
Author Philip Torr
Author Zhen Wu
Date 2024
Extra Citation Key: lei2024fairmindsim
Publication arXiv preprint arXiv:2410.10398
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

FATH: Authentication-based test-time defense against indirect prompt injection attacks

Item Type Journal Article
Author Jiongxiao Wang
Author Fangzhou Wu
Author Wendi Li
Author Jinsheng Pan
Author Edward Suh
Author Z Morley Mao
Author Muhao Chen
Author Chaowei Xiao
Date 2024
Extra Citation Key: wang2024fath

Publication arXiv preprint arXiv:2410.21492
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Februus: Input purification defense against trojan attacks on deep neural network systems

Item Type Conference Paper
Author Bao Gia Doan
Author Ehsan Abbasnejad
Author Damith C Ranasinghe
Date 2020
Extra Citation Key: doan2020februus
Pages 897–912
Proceedings Title Proceedings of the 36th annual computer security applications conference
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Finance and growth: Theory and evidence

Item Type Journal Article
Author Ross Levine
Date 2005
Extra Citation Key: levine2005 Publisher: Elsevier
Volume 1
Pages 865–934
Publication Handbook of Economic Growth
Issue 12
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

FinBrain: When finance meets AI 2.0

Item Type Conference Paper
Author Xiaolin Zheng
Author Mengying Zhu
Author Qibing Li
Author Chaochao Chen
Author Yanchao Tan
Date 2018
URL https://arXiv.org/abs/1808.08497
Extra Citation Key: zheng2018finbrain
Proceedings Title arXiv preprint arXiv:1808.08497
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making

Item Type Journal Article
Author Yangyang Yu
Author Zhiyuan Yao
Author Haohang Li
Author Zhiyang Deng
Author Yuechen Jiang
Author Yupeng Cao
Author Zhi Chen
Author Jordan Suchow
Author Zhenyu Cui
Author Rong Liu
Author others
Date 2025
Extra Citation Key: yu2025fincon
Volume 37
Pages 137010–137045
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

FinGPT: Open-source financial large language models

Item Type Document
Author Hongyang Yang
Author Xiao-Yang Liu
Author Christina Dan Wang
Date 2023
URL https://arxiv.org/abs/2306.06031
Extra Citation Key: yang2023fingptopensourcefinanciallarge arXiv: 2306.06031 [q-fin.ST]
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design

Item Type Document
Author Yangyang Yu
Author Haohang Li
Author Zhi Chen
Author Yuechen Jiang
Author Yang Li
Author Denghui Zhang
Author Rong Liu
Author Jordan W. Suchow
Author Khaldoun Khashanah
Abstract Recent advancements in Large Language Models (LLMs) have exhibited notable efficacy in question-answering (QA) tasks across diverse domains. Their prowess in integrating extensive web knowledge has fueled interest in developing LLM-based autonomous agents. While LLMs are efficient in decoding human instructions and deriving solutions by holistically processing historical inputs, transitioning to purpose-driven agents requires a supplementary rational architecture to process multi-source information, establish reasoning chains, and prioritize critical tasks. Addressing this, we introduce \textsc{FinMem}, a novel LLM-based agent framework

devised for financial decision-making. It encompasses three core modules: Profiling, to customize the agent's characteristics; Memory, with layered message processing, to aid the agent in assimilating hierarchical financial data; and Decision-making, to convert insights gained from memories into investment decisions. Notably, \textsc{FinMem}'s memory module aligns closely with the cognitive structure of human traders, offering robust interpretability and real-time tuning. Its adjustable cognitive span allows for the retention of critical information beyond human perceptual limits, thereby enhancing trading outcomes. This framework enables the agent to self-evolve its professional knowledge, react agilely to new investment cues, and continuously refine trading decisions in the volatile financial environment. We first compare \textsc{FinMem} with various algorithmic agents on a scalable real-world financial dataset, underscoring its leading trading performance in stocks. We then fine-tuned the agent's perceptual span and character setting to achieve a significantly enhanced trading performance. Collectively, \textsc{FinMem} presents a cutting-edge LLM agent framework for automated trading, boosting cumulative investment returns.

Date 2023-12
Short Title FinMem
URL http://arxiv.org/abs/2311.13743
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: yu_finmem_2023 DOI: 10.48550/arXiv.2311.13743
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Quantitative Finance - Computational Finance, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2311.13743 [q-fin]

FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models

Item Type Document
Author Hongyang Yang
Author Boyu Zhang
Author Neng Wang
Author Cheng Guo
Author Xiaoli Zhang
Author Likun Lin
Author Junlin Wang
Author Tianyu Zhou
Author Mao Guan
Author Runjia Zhang
Author Christina Dan Wang
Abstract As financial institutions and professionals increasingly incorporate Large Language Models (LLMs) into their workflows, substantial barriers, including proprietary data and specialized knowledge, persist between the finance sector and the AI community. These challenges impede the AI community's ability to enhance financial tasks effectively. Acknowledging financial analysis's critical role, we aim to devise financial-specialized LLM-based toolchains and democratize access to them through open-source initiatives, promoting wider AI adoption in financial decision-making. In this paper, we introduce FinRobot, a novel open-source AI agent platform supporting multiple financially specialized AI agents, each powered by LLM. Specifically, the platform consists of four major layers: 1) the Financial AI Agents layer that formulates Financial Chain-of-Thought (CoT) by breaking sophisticated financial problems down into logical sequences; 2) the Financial LLM Algorithms layer

dynamically configures appropriate model application strategies for specific tasks; 3) the LLMOps and DataOps layer produces accurate models by applying training/fine-tuning techniques and using task-relevant data; 4) the Multi-source LLM Foundation Models layer that integrates various LLMs and enables the above layers to access them directly. Finally, FinRobot provides hands-on for both professional-grade analysts and laypersons to utilize powerful AI techniques for advanced financial analysis. We open-source FinRobot at \url{https://github.com /AI4Finance-Foundation/FinRobot}.

Date 2024-05
Short Title FinRobot
URL http://arxiv.org/abs/2405.14767
Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: yang_finrobot_2024 DOI: 10.48550/arXiv.2405.14767
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Computation and Language, Computer Science - Machine Learning, Quantitative Finance - Trading and Market Microstructure, Quantitative Finance - Statistical Finance

Notes:

arXiv:2405.14767 [q-fin]

FLAG-Trader: Fusion LLM-Agent with Gradient-based Reinforcement Learning for Financial Trading

Item Type Document
Author Guojun Xiong
Author Zhiyang Deng
Author Keyi Wang
Author Yupeng Cao
Author Haohang Li
Author Yangyang Yu
Author Xueqing Peng
Author Mingquan Lin
Author Kaleb E. Smith
Author Xiao-Yang Liu
Author Jimin Huang
Author Sophia Ananiadou
Author Qianqian Xie
Abstract Large language models (LLMs) fine-tuned on multimodal financial data have demonstrated impressive reasoning capabilities in various financial tasks. However, they often struggle with multi-step, goal-oriented scenarios in interactive financial markets, such as trading, where complex agentic approaches are required to improve decision-making. To address this, we propose \textsc{FLAG-Trader}, a unified architecture integrating linguistic processing (via LLMs) with gradient-driven reinforcement learning (RL) policy optimization, in which a partially fine-tuned LLM acts as the policy network, leveraging pre-trained knowledge while adapting to the financial domain through parameter-efficient fine-tuning. Through policy gradient optimization driven by trading rewards, our framework not only enhances LLM performance in trading but also improves results on other financial-domain tasks. We present extensive empirical evidence to validate these enhancements.
Date 2025-02
Short Title FLAG-Trader
URL http://arxiv.org/abs/2502.11433
Accessed 10/8/2025, 7:00:00 PM

Extra Citation Key: xiong_flag-trader_2025 DOI: 10.48550/arXiv.2502.11433
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Trading and Market Microstructure, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2502.11433 [cs]

Flooding spread of manipulated knowledge in llm-based multi-agent communities

Item Type Journal Article
Author Tianjie Ju
Author Yiting Wang
Author Xinbei Ma
Author Pengzhou Cheng
Author Haodong Zhao
Author Yulong Wang
Author Lifeng Liu
Author Jian Xie
Author Zhuosheng Zhang
Author Gongshen Liu
Date 2024
Extra Citation Key: ju2024flooding
Publication arXiv preprint arXiv:2407.07791
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

For markets, AI efficiency may bring volatility

Item Type Document
Author Reuters
Date 2024
URL https://www.reuters.com/markets/markets-ai-efficiency-may-bring-volatility-mcgeever-2024-10-17/
Extra Citation Key: reuters2024ai_volatility tex.howpublished: Reuters, Oct 17, 2024
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Formal verification of open multi-agent systems

Item Type Conference Paper
Author Panagiotis Kouvaros
Author Alessio Lomuscio
Author Edoardo Pirovano

Author Hashan Punchihewa
Abstract We study open multi-agent systems in which countably many agents may leave and join the system at run-time. We introduce a semantics, based on interpreted systems, to capture the openness of the system and show how an indexed variant of temporal-epistemic logic can be used to express specifications on them. We define the verification problem and show it is undecidable. We isolate one decidable class of open multi-agent systems and give a partial decision procedure for another one. We introduce MCMAS-OP, an open-source toolkit implementing the verification procedures. We present the results obtained using our tool on two examples.
Date 2019
Extra Citation Key: 10.5555/3306127.3331691 Number of pages: 9 tex.address: Richland, SC
Place Montreal QC, Canada
Publisher International Foundation for Autonomous Agents and Multiagent Systems
ISBN 978-1-4503-6309-9
Pages 179–187
Series Aamas '19
Proceedings Title Proceedings of the 18th international conference on autonomous agents and MultiAgent systems
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Tags:

multi-agent systems, open systems, parameterised model checking

---

Formalizing and benchmarking prompt injection attacks and defenses

Item Type Conference Paper
Author Yupei Liu
Author Yuqi Jia
Author Runpeng Geng
Author Jinyuan Jia
Author Neil Zhenqiang Gong
Date 2024
Extra Citation Key: Liu2024FormalingPromptInjection Number of pages: 17 tex.address: USA tex.articleno: 103
Place Philadelphia, PA, USA
Publisher USENIX Association
ISBN 978-1-939133-44-1
Series Sec '24
Proceedings Title Proceedings of the 33rd USENIX conference on security symposium
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Formally specifying the high-level behavior of LLM-based agents

Item Type Document
Author Maxwell Crouse
Author Ibrahim Abdelaziz
Author Ramon Astudillo
Author Kinjal Basu
Author Soham Dan
Author Sadhana Kumaravel
Author Achille Fokoue
Author Pavan Kapanipathi

Author Salim Roukos
Author Luis Lastras
Date 2024
URL https://arxiv.org/abs/2310.08535
Extra Citation Key: crouse2024formallyspecifyinghighlevelbehavior arXiv: 2310.08535 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

FRACTURED-SORRY-bench: Framework for revealing attacks in conversational turns undermining refusal efficacy and defenses over SORRY-bench (automated multi-shot jailbreaks)

Item Type Journal Article
Author Aman Priyanshu
Author Supriti Vijay
Date 2024
Extra Citation Key: priyanshu2024fractured
Publication arXiv preprint arXiv:2408.16163
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

From allies to adversaries: Manipulating LLM tool-calling through adversarial injection

Item Type Journal Article
Author Haowei Wang
Author Rupeng Zhang
Author Junjie Wang
Author Mingyang Li
Author Yuekai Huang
Author Dandan Wang
Author Qing Wang
Date 2024
Extra Citation Key: wang2024allies
Publication arXiv preprint arXiv:2412.10198
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

From CVE entries to verifiable exploits: An automated multi-agent framework for reproducing cves

Item Type Document
Author Saad Ullah
Author Praneeth Balasubramanian
Author Wenbo Guo
Author Amanda Burnett
Author Hammond Pearce
Author Christopher Kruegel
Author Giovanni Vigna
Author Gianluca Stringhini
Date 2025

URL https://arxiv.org/abs/2509.01835
Extra Citation Key: ullah2025cveentriesverifiableexploits arXiv: 2509.01835 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

From Tasks to Teams: A Risk-First Evaluation Framework for Multi-Agent LLM Systems in Finance

Item Type Journal Article
Author Zichen Chen
Author Jianda Chen
Author Jiaao Chen
Author Misha Sra
Abstract Current financial benchmarks reward large language models (LLMs) task accuracy and portfolio return, yet remain blind to the risks that emerge once several agents cooperate, share tools, and act on real money. We present M-SAEA, a Multi-agent, Safety-Aware Evaluation Agent that audits an entire team of LLM agents without fine-tuning. M-SAEA issues ten zero-shot probes spanning four layers including model, workflow, interaction, and system, and returns a continuous [0, 100] risk vector plus a natural-language rationale. Across three high-impact task clusters (finance management, webshop automation, transactional services) and six popular models, MSAEA (i) detects most unsafe trajectories while raising false alarms on only small number of safe ones; (ii) exposes latent hazards: temporal staleness, cross-agent race conditions, API-stress fragility, that leaderboard metrics never flag; and (iii) produces actionable, fine-grained scores that allow practitioners to trade off latency and safety before deployment. By turning safety into a measurable, model-agnostic quantity, M-SAEA shifts the evaluation focus from tasks to teams and provides a ready-to-use template for risk-first assessment of agentic AI in finance and beyond.
Language en
Extra Citation Key: chen_tasks_nodate
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

G-safeguard: a topology-guided security lens and treatment on LLM-based multi-agent systems

Item Type Journal Article
Author Shilong Wang
Author Guibin Zhang
Author Miao Yu
Author Guancheng Wan
Author Fanci Meng
Author Chongye Guo
Author Kun Wang
Author Yang Wang
Date 2025
Extra Citation Key: wang2025g
Publication arXiv preprint arXiv:2502.11127
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

GALA: Can graph-augmented large language model agentic workflows elevate root cause analysis?

Item Type Document
Author Yifang Tian

Author Yaming Liu
Author Zichun Chong
Author Zihang Huang
Author Hans-Arno Jacobsen
Date 2025
URL https://arxiv.org/abs/2508.12472
Extra Citation Key: tian2025galagraphaugmentedlargelanguage arXiv: 2508.12472 [cs.AI]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Gamegpt: Multi-agent collaborative framework for game development

Item Type Journal Article
Author Dake Chen
Author Hanbin Wang
Author Yunhao Huo
Author Yuzhao Li
Author Haoyang Zhang
Date 2023
Extra Citation Key: chen2023gamegpt
Publication arXiv preprint arXiv:2310.08067
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Generative agents: Interactive simulacra of human behavior

Item Type Conference Paper
Author Joon Sung Park
Author Joseph O'Brien
Author Carrie Jun Cai
Author Meredith Ringel Morris
Author Percy Liang
Author Michael S. Bernstein
Date 2023
URL https://doi.org/10.1145/3586183.3606763
Extra Citation Key: 10.1145/3586183.3606763 Number of pages: 22 tex.address: New York, NY, USA tex.articleno: 2
Place San Francisco, CA, USA
Publisher Association for Computing Machinery
ISBN 979-8-4007-0132-0
Series Uist '23
DOI 10.1145/3586183.3606763
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

agents, generative AI, Human-AI interaction, large language models

---

" Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction

Item Type Journal Article
Author Shuning Zhang
Author Lyumanshan Ye
Author Xin Yi
Author Jingyu Tang
Author Bo Shui
Author Haobin Xing
Author Pengfei Liu
Author Hewu Li
Date 2024
Extra Citation Key: zhang2024ghost
Publication arXiv preprint arXiv:2410.14931
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

" Glue pizza and eat rocks"–exploiting vulnerabilities in retrieval-augmented generative models

Item Type Journal Article
Author Zhen Tan
Author Chengshuai Zhao
Author Raha Moraffah
Author Yifan Li
Author Song Wang
Author Jundong Li
Author Tianlong Chen
Author Huan Liu
Date 2024
Extra Citation Key: tan2024glue
Publication arXiv preprint arXiv:2406.19417
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Good parenting is all you need–Multi-agentic LLM hallucination mitigation

Item Type Journal Article
Author Ted Kwartler
Author Matthew Berman
Author Alan Aqrawi
Date 2024
Extra Citation Key: kwartler2024good
Publication arXiv preprint arXiv:2410.14262
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

GPT-4 technical report

Item Type Document
Author OpenAI
Date 2023
URL https://cdn.openai.com/papers/gpt-4.pdf
Extra Citation Key: openai2023gpt4
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

Accessed: 2025-10-08

Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts

Item Type Journal Article
Author Jiahao Yu
Author Xingwei Lin
Author Zheng Yu
Author Xinyu Xing
Date 2023
Extra Citation Key: yu2023gptfuzzer
Publication arXiv preprint arXiv:2309.10253
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Gptswarm: Language agents as optimizable graphs

Item Type Conference Paper
Author Mingchen Zhuge
Author Wenyi Wang
Author Louis Kirsch
Author Francesco Faccio
Author Dmitrii Khizbullin
Author Jürgen Schmidhuber
Date 2024
Extra Citation Key: zhuge2024gptswarm
Proceedings Title Forty-first international conference on machine learning
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Gracefully filtering backdoor samples for generative large language models without retraining

Item Type Journal Article
Author Zongru Wu
Author Pengzhou Cheng
Author Lingyong Fang
Author Zhuosheng Zhang
Author Gongshen Liu

Date 2024
Extra Citation Key: wu2024gracefully
Publication arXiv preprint arXiv:2412.02454
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Great, now write an article about that: The crescendo multi-turn llm jailbreak attack

Item Type Journal Article
Author Mark Russinovich
Author Ahmed Salem
Author Ronen Eldan
Date 2024
Extra Citation Key: russinovich2024great
Publication arXiv preprint arXiv:2404.01833
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

GuardAgent: Safeguard LLM agents by a guard agent via knowledge-enabled reasoning

Item Type Document
Author Zhen Xiang
Author Linzhi Zheng
Author Yanjie Li
Author Junyuan Hong
Author Qinbin Li
Author Han Xie
Author Jiawei Zhang
Author Zidi Xiong
Author Chulin Xie
Author Carl Yang
Author Dawn Song
Author Bo Li
Date 2025
URL https://arxiv.org/abs/2406.09187
Extra Citation Key: xiang2025guardagentsafeguardllmagents arXiv: 2406.09187 [cs.LG]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Guardrail baselines for unlearning in llms

Item Type Journal Article
Author Pratiksha Thaker
Author Yash Maurya
Author Shengyuan Hu
Author Zhiwei Steven Wu
Author Virginia Smith
Date 2024

Extra Citation Key: thaker2024guardrail
Publication arXiv preprint arXiv:2403.03329
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Guiding pretraining in reinforcement learning with large language models

Item Type Conference Paper
Author Yuqing Du
Author Olivia Watkins
Author Zihan Wang
Author Cédric Colas
Author Trevor Darrell
Author Pieter Abbeel
Author Abhishek Gupta
Author Jacob Andreas
Date 2023
Extra Citation Key: du2023guiding
Publisher PMLR
Pages 8657–8677
Proceedings Title International conference on machine learning
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

HackSynth: LLM agent and evaluation framework for autonomous penetration testing

Item Type Document
Author Lajos Muzsai
Author David Imolai
Author András Lukács
Date 2024
URL https://arxiv.org/abs/2412.01778
Extra Citation Key: muzsai2024hacksynthllmagentevaluation arXiv: 2412.01778 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Haicosystem: An ecosystem for sandboxing safety risks in human-ai interactions

Item Type Journal Article
Author Xuhui Zhou
Author Hyunwoo Kim
Author Faeze Brahman
Author Liwei Jiang
Author Hao Zhu
Author Ximing Lu
Author Frank Xu
Author Bill Yuchen Lin
Author Yejin Choi

Author Niloofar Mireshghallah
Author others
Date 2024
Extra Citation Key: zhou2024haicosystem
Publication arXiv preprint arXiv:2409.16427
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Hallucination is inevitable: An innate limitation of large language models

Item Type Journal Article
Author Ziwei Xu
Author Sanjay Jain
Author Mohan Kankanhalli
Date 2024
Extra Citation Key: xu2024hallucination
Publication arXiv preprint arXiv:2401.11817
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

HallusionBench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models

Item Type Conference Paper
Author Tianrui Guan
Author Fuxiao Liu
Author Xiyang Wu
Author Ruiqi Xian
Author Zongxia Li
Author Xiaoyu Liu
Author Xijun Wang
Author Lichang Chen
Author Furong Huang
Author Yaser Yacoob
Author others
Date 2024
Extra Citation Key: guan2024hallusionbench
Pages 14375–14385
Proceedings Title Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Halueval-wild: Evaluating hallucinations of language models in the wild

Item Type Journal Article
Author Zhiying Zhu
Author Yiming Yang
Author Zhiqing Sun

Date 2024
Extra Citation Key: zhu2024halueval
Publication arXiv preprint arXiv:2403.04307
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

HedgeAgents: A Balanced-aware Multi-agent Financial Trading System

Item Type Document
Author Xiangyu Li
Author Yawen Zeng
Author Xiaofen Xing
Author Jin Xu
Author Xiangmin Xu
Abstract As automated trading gains traction in the financial market, algorithmic investment strategies are increasingly prominent. While Large Language Models (LLMs) and Agent-based models exhibit promising potential in real-time market analysis and trading decisions, they still experience a significant -20% loss when confronted with rapid declines or frequent fluctuations, impeding their practical application. Hence, there is an imperative to explore a more robust and resilient framework. This paper introduces an innovative multi-agent system, HedgeAgents, aimed at bolstering system robustness via “hedging” strategies. In this well-balanced system, an array of hedging agents has been tailored, where HedgeAgents consist of a central fund manager and multiple hedging experts specializing in various financial asset classes. These agents leverage LLMs' cognitive capabilities to make decisions and coordinate through three types of conferences. Benefiting from the powerful understanding of LLMs, our HedgeAgents attained a 70% annualized return and a 400% total return over a period of 3 years. Moreover, we have observed with delight that HedgeAgents can even formulate investment experience comparable to those of human experts (https://hedgeagents.github.io/).
Date 2025-02
Short Title HedgeAgents
URL http://arxiv.org/abs/2502.13165
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: li_hedgeagents_2025 DOI: 10.48550/arXiv.2502.13165
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Trading and Market Microstructure, Computer Science - Multiagent Systems

Notes:

arXiv:2502.13165 [cs]

Hijacking third-party plugins: Backdoor and supply-chain threats in LLM ecosystems

Item Type Journal Article
Author Zhen Guo
Author Xinyi Li
Author Jiaqi Zhang
Author Yu Sun
Author Shuchang Zhao

Date 2025
URL https://arxiv.org/abs/2508.17674
Extra Citation Key: guo2025backdooredplugins
Publication arXiv preprint arXiv:2508.17674
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

HijackRAG: Hijacking attacks against retrieval-augmented large language models

Item Type Journal Article
Author Yucheng Zhang
Author Qinfeng Li
Author Tianyu Du
Author Xuhong Zhang
Author Xinkui Zhao
Author Zhengwen Feng
Author Jianwei Yin
Date 2024
Extra Citation Key: zhang2024hijackrag
Publication arXiv preprint arXiv:2410.22832
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

How human–AI feedback loops alter human perceptual, emotional and social judgements

Item Type Journal Article
Author Moshe Glickman
Author Tali Sharot
Date 2024
Extra Citation Key: glickman2024human Publisher: Nature Publishing Group UK London
Pages 1–15
Publication Nature Human Behaviour
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Identifying the risks of lm agents with an lm-emulated sandbox

Item Type Journal Article
Author Yangjun Ruan
Author Honghua Dong
Author Andrew Wang
Author Silviu Pitis
Author Yongchao Zhou
Author Jimmy Ba
Author Yann Dubois
Author Chris J Maddison
Author Tatsunori Hashimoto
Date 2023

Extra Citation Key: ruan2023identifying
Publication arXiv preprint arXiv:2309.15817
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Ignore previous prompt: Attack techniques for language models

Item Type Document
Author Fábio Perez
Author Ian Ribeiro
Date 2022
URL https://arxiv.org/abs/2211.09527
Extra Citation Key: perez2022ignorepreviouspromptattack arXiv: 2211.09527 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Imprompter: Tricking LLM agents into improper tool use

Item Type Journal Article
Author Xiaohan Fu
Author Shuheng Li
Author Zihan Wang
Author Yihao Liu
Author Rajesh K Gupta
Author Taylor Berg-Kirkpatrick
Author Earlence Fernandes
Date 2024
Extra Citation Key: fu2024imprompter
Publication arXiv preprint arXiv:2410.14923
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Improved techniques for optimization-based jailbreaking on large language models

Item Type Journal Article
Author Xiaojun Jia
Author Tianyu Pang
Author Chao Du
Author Yihao Huang
Author Jindong Gu
Author Yang Liu
Author Xiaochun Cao
Author Min Lin
Date 2024
Extra Citation Key: jia2024improved
Publication arXiv preprint arXiv:2405.21018
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Improving factuality and reasoning in language models through multiagent debate

Item Type Journal Article
Author Yilun Du
Author Shuang Li
Author Antonio Torralba
Author Joshua B Tenenbaum
Author Igor Mordatch
Date 2023
Extra Citation Key: du2023improving
Publication arXiv preprint arXiv:2305.14325
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

InferAct: Inferring safe actions for LLM-based agents through preemptive evaluation and human feedback

Item Type Journal Article
Author Haishuo Fang
Author Xiaodan Zhu
Author Iryna Gurevych
Date 2024
Extra Citation Key: fang2024inferact
Publication arXiv preprint arXiv:2407.11843
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Influence of rapport and social presence with an AI psychotherapy chatbot on users' self-disclosure

Item Type Journal Article
Author Jieon Lee
Author Daeho Lee
Author Jae-gil Lee
Date 2024
Extra Citation Key: lee2024influence Publisher: Taylor & Francis
Volume 40
Pages 1620–1631
Publication International Journal of Human–Computer Interaction
Issue 7
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

InfoRM: Mitigating reward hacking in RLHF via information-theoretic reward modeling

Item Type Document
Author Yuchun Miao
Author Sen Zhang
Author Liang Ding

Author Rong Bao
Author Lefei Zhang
Author Dacheng Tao
Abstract Despite the success of reinforcement learning from human feedback (RLHF) in aligning language models with human values, reward hacking (or reward overoptimization) remains a critical challenge. This issue arises when reward models ignore human preferences and instead optimize spurious correlations. We propose InfoRM, a framework based on an information bottleneck objective that filters irrelevant features from the reward model's latent space. We also identify a correlation between overoptimization and outliers in this compressed space, and introduce the Cluster Separation Index (CSI) to quantify overoptimization. Experiments on reward model scales (70M to 7B) show InfoRM improves robustness against reward hacking and that CSI can detect overoptimization patterns.
Date 2024
URL https://arxiv.org/abs/2402.09345
Extra Citation Key: miao2024inform
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

InjecAgent: Benchmarking indirect prompt injections in tool-integrated large language model agents

Item Type Conference Paper
Author Qiusi Zhan
Author Zhixiang Liang
Author Zifan Ying
Author Daniel Kang
Editor Lun-Wei Ku
Editor Andre Martins
Editor Vivek Srikumar
Date 2024-08
URL https://aclanthology.org/2024.findings-acl.624/
Extra Citation Key: zhan-etal-2024-injecagent
Place Bangkok, Thailand
Publisher Association for Computational Linguistics
Pages 10471–10506
Proceedings Title Findings of the association for computational linguistics: ACL 2024
DOI 10.18653/v1/2024.findings-acl.624
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Integrating Large Language Models in Financial Investments and Market Analysis: A Survey

Item Type Document
Author Sedigheh Mahdavi
Author Jiating
Author Chen
Author Pradeep Kumar Joshi
Author Lina Huertas Guativa
Author Upmanyu Singh
Abstract Large Language Models (LLMs) have been employed in financial decision making, enhancing analytical capabilities for investment strategies. Traditional investment strategies often utilize quantitative models, fundamental analysis, and technical indicators. However, LLMs have introduced new capabilities to process and analyze large volumes of structured and unstructured data, extract meaningful insights, and enhance decision-

making in real-time. This survey provides a structured overview of recent research on LLMs within the financial domain, categorizing research contributions into four main frameworks: LLM-based Frameworks and Pipelines, Hybrid Integration Methods, Fine-Tuning and Adaptation Approaches, and Agent-Based Architectures. This study provides a structured review of recent LLMs research on applications in stock selection, risk assessment, sentiment analysis, trading, and financial forecasting. By reviewing the existing literature, this study highlights the capabilities, challenges, and potential directions of LLMs in financial markets.

Date 2025-06
Short Title Integrating Large Language Models in Financial Investments and Market Analysis
URL http://arxiv.org/abs/2507.01990
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: mahdavi_integrating_2025 DOI: 10.48550/arXiv.2507.01990
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Quantitative Finance - General Finance

Notes:

arXiv:2507.01990 [q-fin]

Integrating Traditional Technical Analysis with AI: A Multi-Agent LLM-Based Approach to Stock Market Forecasting

Item Type Conference Paper
Author Michał Wawer
Author Jarosław A. Chudziak
Abstract Traditional technical analysis methods face limitations in accurately predicting trends in today's complex financial markets. This paper introduces ElliottAgents, an multi-agent system that integrates the Elliott Wave Principle with AI for stock market forecasting. The inherent complexity of financial markets, characterized by non-linear dynamics, noise, and susceptibility to unpredictable external factors, poses significant challenges for accurate prediction. To address these challenges, the system employs LLMs to enhance natural language understanding and decision-making capabilities within a multi-agent framework. By leveraging technologies such as Retrieval-Augmented Generation (RAG) and Deep Reinforcement Learning (DRL), ElliottAgents performs continuous, multi-faceted analysis of market data to identify wave patterns and predict future price movements. The research explores the system's ability to process historical stock data, recognize Elliott wave patterns, and generate actionable insights for traders. Experimental results, conducted on historical data from major U.S. companies, validate the system's effectiveness in pattern recognition and trend forecasting across various time frames. This paper contributes to the field of AI-driven financial analysis by demonstrating how traditional technical analysis methods can be effectively combined with modern AI approaches to create more reliable and interpretable market prediction systems.
Date 2025
Short Title Integrating Traditional Technical Analysis with AI
URL http://arxiv.org/abs/2506.16813
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: wawer_integrating_2025
Pages 100–111
Proceedings Title Proceedings of the 17th International Conference on Agents and Artificial Intelligence
DOI 10.5220/0013191200003890
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2506.16813 [cs]

Investigating security implications of automatically generated code on the software supply chain

Item Type	Document
Author	Xiaofan Li
Author	Xing Gao
Abstract	Empirical study showing AI-generated code can propagate insecure patterns and dependency risks in supply chains; proposes mitigations such as audit scaffolding, prompt hardening, and differential testing.
Date	2025
URL	https://arxiv.org/pdf/2509.20277
Extra	Citation Key: li2025investigatingsecurity
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent

Item Type	Document
Author	Haohang Li
Author	Yupeng Cao
Author	Yangyang Yu
Author	Shashidhar Reddy Javaji
Author	Zhiyang Deng
Author	Yueru He
Author	Yuechen Jiang
Author	Zining Zhu
Author	Koduvayur Subbalakshmi
Author	Guojun Xiong
Author	Jimin Huang
Author	Lingfei Qian
Author	Xueqing Peng
Author	Qianqian Xie
Author	Jordan W. Suchow
Abstract	Recent advancements have underscored the potential of large language model (LLM)-based agents in financial decision-making. Despite this progress, the field currently encounters two main challenges: (1) the lack of a comprehensive LLM agent framework adaptable to a variety of financial tasks, and (2) the absence of standardized benchmarks and consistent datasets for assessing agent performance. To tackle these issues, we introduce \textsc{InvestorBench}, the first benchmark specifically designed for evaluating LLM-based agents in diverse financial decision-making contexts. InvestorBench enhances the versatility of LLM-enabled agents by providing a comprehensive suite of tasks applicable to different financial products, including single equities like stocks, cryptocurrencies and exchange-traded funds (ETFs). Additionally, we assess the reasoning and decision-making capabilities of our agent framework using thirteen different LLMs as backbone models, across various market environments and tasks. Furthermore, we have curated a diverse collection of open-source, multi-modal datasets and developed a comprehensive suite of environments for financial decision-making. This establishes a highly accessible platform for evaluating financial agents' performance across various scenarios.
Date	2024-12

Short Title INVESTORBENCH
URL http://arxiv.org/abs/2412.18174
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: li_investorbench_2024 DOI: 10.48550/arXiv.2412.18174
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Computational Finance, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2412.18174 [cs]

IRCopilot: Automated incident response with large language models

Item Type Document
Author Xihuan Lin
Author Jie Zhang
Author Gelei Deng
Author Tianzhe Liu
Author Xiaolong Liu
Author Changcai Yang
Author Tianwei Zhang
Author Qing Guo
Author Riqing Chen
Date 2025
URL https://arxiv.org/abs/2505.20945
Extra Citation Key: lin2025ircopilotautomatedincidentresponse arXiv: 2505.20945 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

IRIS: LLM-assisted static analysis for detecting security vulnerabilities

Item Type Conference Paper
Author Ziyang Li
Author Saikat Dutta
Author Mayur Naik
Date 2025
URL https://openreview.net/forum?id=9LdJDU7E91
Extra Citation Key: li2025iris
Proceedings Title The thirteenth international conference on learning representations
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Is my data in your retrieval database? membership inference attacks against retrieval augmented generation

Item Type Journal Article
Author Maya Anderson
Author Guy Amit
Author Abigail Goldsteen
Date 2024
Extra Citation Key: anderson2024my
Publication arXiv preprint arXiv:2405.20446
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms

Item Type Journal Article
Author Xuhui Zhou
Author Zhe Su
Author Tiwalayo Eisape
Author Hyunwoo Kim
Author Maarten Sap
Date 2024
Extra Citation Key: zhou2024real
Publication arXiv preprint arXiv:2403.05020
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint

Item Type Journal Article
Author Wei Xiong
Author Hanze Dong
Author Chenlu Ye
Author Ziqi Wang
Author Han Zhong
Author Heng Ji
Author Nan Jiang
Author Tong Zhang
Date 2023
Extra Citation Key: xiong2023iterative
Publication arXiv preprint arXiv:2312.11456
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Jailbreak attacks and defenses against large language models: a survey

Item Type Document
Author Sibo Yi
Author Yule Liu
Author Zhen Sun
Author Tianshuo Cong
Author Xinlei He
Author Jiaxing Song
Author Ke Xu
Author Qi Li
Date 2024
URL https://arxiv.org/abs/2407.04295
Extra Citation Key: yi2024jailbreakattacksdefenseslarge arXiv: 2407.04295 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Jailbreaking chatgpt via prompt engineering: An empirical study

Item Type Journal Article
Author Yi Liu
Author Gelei Deng
Author Zhengzi Xu
Author Yuekang Li
Author Yaowen Zheng
Author Ying Zhang
Author Lida Zhao
Author Tianwei Zhang
Author Kailong Wang
Author Yang Liu
Date 2023
Extra Citation Key: liu2023jailbreaking
Publication arXiv preprint arXiv:2305.13860
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Jailbreaking large language models: A comprehensive survey

Item Type Journal Article
Author Xingyu Yi
Author Xinyu Chen
Author Xuan Song
Author Chenyu Zhang
Author Jiayi Zhang
Author Mingyi Zhang
Author Neil Gong
Date 2024
URL https://arxiv.org/abs/2407.04295
Extra Citation Key: yi2024jailbreaksurvey
Publication arXiv preprint arXiv:2407.04295
Date Added 10/20/2025, 3:48:27 PM

Modified 10/20/2025, 3:48:27 PM

Jailbreaking LLM-controlled robots

Item Type Document
Author Alexander Robey
Author Zachary Ravichandran
Author Vijay Kumar
Author Hamed Hassani
Author George J. Pappas
Date 2024
URL https://arxiv.org/abs/2410.13691
Extra Citation Key: robey2024jailbreakingllmcontrolledrobots arXiv: 2410.13691 [cs.RO]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Jailbroken: how does LLM safety training fail?

Item Type Conference Paper
Author Alexander Wei
Author Nika Haghtalab
Author Jacob Steinhardt
Date 2023
Extra Citation Key: Wei2023Jailbroken Number of pages: 32 tex.address: Red Hook, NY, USA tex.articleno: 3508
Place New Orleans, LA, USA
Publisher Curran Associates Inc.
Series Nips '23
Proceedings Title Proceedings of the 37th international conference on neural information processing systems
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

JAILJUDGE: AComprehensive JAILBREAK JUDGE BENCHMARK with MULTI-AGENT ENHANCED EXPLANATION EVALUATION FRAMEWORK

Item Type Journal Article
Author JUDGE BENCHMARK
Extra Citation Key: benchmarkjailjudge
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Keeping llms aligned after fine-tuning: The crucial role of prompt templates

Item Type Journal Article
Author Kaifeng Lyu
Author Haoyu Zhao
Author Xinran Gu
Author Dingli Yu

Author Anirudh Goyal
Author Sanjeev Arora
Date 2024
Extra Citation Key: lyu2024keeping
Publication arXiv preprint arXiv:2402.18540
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

KNighter: Transforming static analysis with LLM-synthesized checkers

Item Type Conference Paper
Author Chenyuan Yang
Author Zijie Zhao
Author Zichen Xie
Author Haoyu Li
Author Lingming Zhang
Date 2025
URL https://doi.org/10.1145/3731569.3764827
Extra Citation Key: yang2025knighter tex.address: New York, NY, USA
Place Seoul, Republic of Korea
Publisher Association for Computing Machinery
Series Sosp '25
Proceedings Title Proceedings of the ACM SIGOPS 31st symposium on operating systems principles
DOI 10.1145/3731569.3764827
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

KubeIntellect: a modular LLM-orchestrated agent framework for end-to-end kubernetes management

Item Type Document
Author Mohsen Seyedkazemi Ardebili
Author Andrea Bartolini
Abstract Presents an agentic system for natural-language Kubernetes control spanning read/write/delete/exec/RBAC and lifecycle verbs, with modular domain agents orchestrated by a supervisor and secure tool synthesis. Reports high reliability across 200 NL queries.
Date 2025
URL https://arxiv.org/abs/2509.02449
Extra Citation Key: ardebili2025kubeintellect
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

LangChain: Building applications with large language models

Item Type Document
Author LangChain
Date 2023
URL https://github.com/hwchase17/langchain
Extra Citation Key: langchain2023
Date Added 10/20/2025, 3:48:27 PM

Modified 10/20/2025, 3:48:27 PM

Notes:

GitHub repository, accessed 2025-10-08

Language evolution for evading social media regulation via llm-based multi-agent simulation

Item Type Conference Paper
Author Jinyu Cai
Author Jialong Li
Author Mingyue Zhang
Author Munan Li
Author Chen-Shu Wang
Author Kenji Tei
Date 2024
Extra Citation Key: cai2024language
Publisher IEEE
Pages 1–10
Proceedings Title 2024 IEEE congress on evolutionary computation (CEC)
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Large language model agent for fake news detection

Item Type Journal Article
Author Xinyi Li
Author Yongfeng Zhang
Author Edward C Malthouse
Date 2024
Extra Citation Key: li2024large
Publication arXiv preprint arXiv:2405.01593
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Large Language Model Agent in Financial Trading: A Survey

Item Type Document
Author Han Ding
Author Yinheng Li
Author Junhao Wang
Author Hang Chen
Abstract Trading is a highly competitive task that requires a combination of strategy, knowledge, and psychological fortitude. With the recent success of large language models(LLMs), it is appealing to apply the emerging intelligence of LLM agents in this competitive arena and understanding if they can outperform professional traders. In this survey, we provide a comprehensive review of the current research on using LLMs as agents in financial trading. We summarize the common architecture used in the agent, the data inputs, and the performance of LLM trading agents in backtesting as well as the challenges presented in these research. This survey aims to provide insights into the current state of LLM-based financial trading agents and outline future research directions in this field.

Date 2024-07
Short Title Large Language Model Agent in Financial Trading
URL http://arxiv.org/abs/2408.06361
Accessed 9/3/2025, 7:00:00 PM
Extra Citation Key: ding_large_2024 DOI: 10.48550/arXiv.2408.06361
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Computation and Language, Quantitative Finance - Trading and Market Microstructure

Notes:

arXiv:2408.06361 [q-fin]

Large language model agentic approach to fact checking and fake news detection

Item Type Book Section
Author Xinyi Li
Author Yongfeng Zhang
Author Edward C Malthouse
Date 2024
Extra Citation Key: li2024large
Publisher IOS Press
Pages 2572–2579
Book Title Ecai 2024
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Large Language Model Agents for Investment Management: Foundations, Benchmarks, and Research Frontiers

Item Type Document
Author Preetha Saha
Author Jingrao Lyu
Author Arnav Saxena
Author Tianjiao Zhao
Author Dhagash Mehta
Abstract Recent advances in Large Language Models (LLMs) have triggered a new wave of intelligent financial agents capable of complex reasoning, tool use, and autonomous decision-making. This survey presents a comprehensive review of LLM-based agents in the context of investment and trading, focusing on applications such as portfolio optimization, risk management, information retrieval, and automated strategy generation. We systematically categorize the literature by use case and architectural innovations including multiagent collaborations, reflection mechanisms, and tool-augmented pipelines. Additionally, we review emerging evaluation frameworks and benchmark datasets tailored to finance-specific agent tasks. The survey identifies current trends, technical limitations, and open challenges related to robustness, explainability, and real-world deployment. We conclude with emerging directions for building more capable, adaptive, and trustworthy financial AI agents aligned with the demands of modern investment ecosystems.
Date 2025-08

Language	en
Short Title	Large Language Model Agents for Investment Management
URL	https://papers.ssrn.com/abstract=5447274
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: saha_large_2025 DOI: 10.2139/ssrn.5447274 Place: Rochester, NY Type: SSRN Scholarly Paper
Publisher	Social Science Research Network
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Tags:

Risk Management, Investment and Trading Strategy, LLM Agents, Portfolio Optimization

---

Large language model alignment: A survey

Item Type	Journal Article
Author	Tianhao Shen
Author	Renren Jin
Author	Yufei Huang
Author	Chuang Liu
Author	Weilong Dong
Author	Zishan Guo
Author	Xinwei Wu
Author	Yan Liu
Author	Deyi Xiong
Date	2023
Extra	Citation Key: shen2023large
Publication	arXiv preprint arXiv:2309.15025
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Large language model assissted multi-agent dialogue for ontology alignment

Item Type	Conference Paper
Author	Shiyao Zhang
Author	Yuji Dong
Author	Yichuan Zhang
Author	Terry R Payne
Author	Jie Zhang
Date	2024
Extra	Citation Key: zhang2024large
Pages	2594–2596
Proceedings Title	Proceedings of the 23rd international conference on autonomous agents and multiagent systems
Date Added	10/20/2025, 3:49:10 PM
Modified	10/20/2025, 3:49:10 PM

Large language model based multi-agents: A survey of progress and challenges

Item Type	Journal Article

Author Taicheng Guo
Author Xiuying Chen
Author Yaqi Wang
Author Ruidi Chang
Author Shichao Pei
Author Nitesh V Chawla
Author Olaf Wiest
Author Xiangliang Zhang
Date 2024
Extra Citation Key: guo2024large
Publication arXiv preprint arXiv:2402.01680
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Large language model guided protocol fuzzing

Item Type Conference Paper
Author Ruijie Meng
Author Martin Mirchev
Author Marcel Böhme
Author Abhik Roychoudhury
Date 2024
Extra Citation Key: chatafl
Proceedings Title Proceedings of the 31st annual network and distributed system security symposium (NDSS)
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Large language model sentinel: LLM agent for adversarial purification

Item Type Journal Article
Author Guang Lin
Author Qibin Zhao
Date 2024
Extra Citation Key: lin2024large
Publication arXiv preprint arXiv:2405.20770
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Large language models are easily confused: a quantitative metric, security implications and typological analysis

Item Type Journal Article
Author Yiyi Chen
Author Qiongxiu Li
Author Russa Biswas
Author Johannes Bjerva
Date 2024
Extra Citation Key: chen2024large

Publication arXiv preprint arXiv:2410.13237
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Large language models are edge-case generators: Crafting unusual programs for fuzzing deep learning libraries

Item Type Conference Paper
Author Yinlin Deng
Author Chunqiu Steven Xia
Author Chenyuan Yang
Author Shizhuo Dylan Zhang
Author Shujing Yang
Author Lingming Zhang
Abstract Bugs in Deep Learning (DL) libraries may affect almost all downstream DL applications, and it is crucial to ensure the quality of such systems. It is challenging to generate valid input programs for fuzzing DL libraries, since the input programs need to satisfy both the syntax/semantics of the supported languages (e.g., Python) and the tensor/operator constraints for constructing valid computational graphs. Recently, the TitanFuzz work demonstrates that modern Large Language Models (LLMs) can be directly leveraged to implicitly learn all the language and DL computation constraints to generate valid programs for fuzzing DL libraries (and beyond). However, LLMs tend to generate ordinary programs following similar patterns/tokens with typical programs seen in their massive pre-training corpora (e.g., GitHub), while fuzzing favors unusual inputs that cover edge cases or are unlikely to be manually produced.To fill this gap, this paper proposes FuzzGPT, the first approach to priming LLMs to synthesize unusual programs for fuzzing. FuzzGPT is mainly built on the well-known hypothesis that historical bug-triggering programs may include rare/valuable code ingredients important for bug finding. Meanwhile, while traditional techniques leveraging such historical information require intensive human efforts to both design dedicated generators and ensure the syntactic/semantic validity of generated programs, FuzzGPT demonstrates that this process can be fully automated via the intrinsic capabilities of LLMs (including fine-tuning and in-context learning), while being generalizable and applicable to challenging domains. While FuzzGPT can be applied with different LLMs, this paper focuses on the powerful GPT-style models: Codex and CodeGen. Moreover, FuzzGPT also shows the potential of directly leveraging the instruction-following capability of the recent ChatGPT for effective fuzzing. The experimental study on two popular DL libraries (PyTorch and TensorFlow) shows that FuzzGPT can substantially outperform TitanFuzz, detecting 76 bugs, with 49 already confirmed as previously unknown bugs, including 11 high-priority bugs or security vulnerabilities.
Date 2024
URL https://doi.org/10.1145/3597503.3623343
Extra Citation Key: deng2024llmedgecase Number of pages: 13 tex.address: New York, NY, USA tex.articleno: 70
Place Lisbon, Portugal
Publisher Association for Computing Machinery
ISBN 979-8-4007-0217-4
Series Icse '24
Proceedings Title Proceedings of the IEEE/ACM 46th international conference on software engineering
DOI 10.1145/3597503.3623343
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Large language models are zero-shot fuzzers: Fuzzing deep-learning libraries via large language models

Item Type Conference Paper
Author Yinlin Deng
Author Chunqiu Steven Xia
Author Haoran Peng

Author Chenyuan Yang
Author Lingming Zhang
Abstract Deep Learning (DL) systems have received exponential growth in popularity and have become ubiquitous in our everyday life. Such systems are built on top of popular DL libraries, e.g., TensorFlow and PyTorch which provide APIs as building blocks for DL systems. Detecting bugs in these DL libraries is critical for almost all downstream DL systems in ensuring effectiveness/safety for end users. Meanwhile, traditional fuzzing techniques can be hardly effective for such a challenging domain since the input DL programs need to satisfy both the input language (e.g., Python) syntax/semantics and the DL API input/shape constraints for tensor computations. To address these limitations, we propose TitanFuzz – the first approach to directly leveraging Large Language Models (LLMs) to generate input programs for fuzzing DL libraries. LLMs are titanic models trained on billions of code snippets and can autoregressively generate human-like code snippets. Our key insight is that modern LLMs can also include numerous code snippets invoking DL library APIs in their training corpora, and thus can implicitly learn both language syntax/semantics and intricate DL API constraints for valid DL program generation. More specifically, we use both generative and infilling LLMs (e.g., Codex/InCoder) to generate and mutate valid/diverse input DL programs for fuzzing. Our experimental results demonstrate that TitanFuzz can achieve 30.38%/50.84% higher code coverage than state-of-the-art fuzzers on TensorFlow/PyTorch. Furthermore, TitanFuzz is able to detect 65 bugs, with 44 already confirmed as previously unknown bugs. This paper demonstrates that modern titanic LLMs can be leveraged to directly perform both generation-based and mutation-based fuzzing studied for decades, while being fully automated, generalizable, and applicable to domains challenging for traditional approaches (such as DL systems). We hope TitanFuzz can stimulate more work in this promising direction of LLMs for fuzzing.
Date 2023
URL https://doi.org/10.1145/3597926.3598067
Extra Citation Key: deng2023llmzeroshotfuzzers Number of pages: 13 tex.address: New York, NY, USA
Place Seattle, WA, USA
Publisher Association for Computing Machinery
ISBN 979-8-4007-0221-1
Pages 423–435
Series Issta 2023
Proceedings Title Proceedings of the 32nd ACM SIGSOFT international symposium on software testing and analysis
DOI 10.1145/3597926.3598067
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

Large Language Model, Fuzz Testing, Test Generation

Large language models can self-improve at web agent tasks

Item Type Journal Article
Author Ajay Patel
Author Markus Hofmarcher
Author Claudiu Leoveanu-Condrei
Author Marius-Constantin Dinu
Author Chris Callison-Burch
Author Sepp Hochreiter
Date 2024
Extra Citation Key: patel2024large
Publication arXiv preprint arXiv:2405.20309
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Large language models for cyber security: a systematic literature review

Item Type	Journal Article
Author	Hanxiang Xu
Author	Shenao Wang
Author	Ningke Li
Author	Kailong Wang
Author	Yanjie Zhao
Author	Kai Chen
Author	Ting Yu
Author	Yang Liu
Author	Haoyu Wang
Date	2025-09
URL	https://doi.org/10.1145/3769676
Extra	Citation Key: 10.1145/3769676 Place: New York, NY, USA Publisher: Association for Computing Machinery
Publication	ACM Trans. Softw. Eng. Methodol.
DOI	10.1145/3769676
ISSN	1049-331X
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Tags:

Cybersecurity, Large language model, Software security

Large language models: A survey

Item Type	Journal Article
Author	Shervin Minaee
Author	Tomas Mikolov
Author	Narjes Nikzad
Author	Meysam Chenaghlu
Author	Richard Socher
Author	Xavier Amatriain
Author	Jianfeng Gao
Date	2024
Extra	Citation Key: minaee2024large
Publication	arXiv preprint arXiv:2402.06196
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

Large model agents: State-of-the-art, cooperation paradigms, security and privacy, and future trends

Item Type	Journal Article
Author	Yuntao Wang
Author	Yanghe Pan
Author	Quan Zhao
Author	Yi Deng
Author	Zhou Su

Author Linkang Du
Author Tom H Luan
Date 2024
Extra Citation Key: wang2024large
Publication arXiv preprint arXiv:2409.14457
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Large multimodal agents: A survey

Item Type Journal Article
Author Junlin Xie
Author Zhihong Chen
Author Ruifei Zhang
Author Xiang Wan
Author Guanbin Li
Date 2024
Extra Citation Key: xie2024large
Publication arXiv preprint arXiv:2402.15116
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

LAW: Legal agentic workflows for custody and fund services contracts

Item Type Journal Article
Author William Watson
Author Nicole Cho
Author Nishan Srishankar
Author Zhen Zeng
Author Lucas Cecchi
Author Daniel Scott
Author Suchetha Siddagangappa
Author Rachneet Kaur
Author Tucker Balch
Author Manuela Veloso
Date 2024
URL https://arxiv.org/abs/2412.11063
Extra Citation Key: watson2024law
Volume arXiv:2412.11063
Publication arXiv preprint
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

preprint, submitted 15 Dec 2024, cs.AI

Learn-by-interact: a data-centric framework for self-adaptive agents in realistic environments

Item Type Journal Article
Author Hongjin Su
Author Ruoxi Sun
Author Jinsung Yoon
Author Pengcheng Yin
Author Tao Yu
Author Sercan Ö Arık
Date 2025
Extra Citation Key: su2025learn
Publication arXiv preprint arXiv:2501.10893
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Leveraging the context through multi-round interactions for jailbreaking attacks

Item Type Journal Article
Author Yixin Cheng
Author Markos Georgopoulos
Author Volkan Cevher
Author Grigorios G Chrysos
Date 2024
Extra Citation Key: cheng2024leveraging
Publication arXiv preprint arXiv:2402.09177
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LISA technical report: An agentic framework for smart contract auditing

Item Type Document
Author Izaiah Sun
Author Daniel Tan
Author Andy Deng
Abstract Agentic auditor leveraging historical audit knowledge and rule/logic reasoning to generalize to new contracts; reports broader vulnerability coverage and accuracy than static analyzers.
Date 2025
URL https://arxiv.org/abs/2509.24698
Extra Citation Key: sun2025lisa
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Llama guard: Llm-based input-output safeguard for human-ai conversations

Item Type Journal Article
Author Hakan Inan
Author Kartikeya Upasani
Author Jianfeng Chi
Author Rashi Rungta
Author Krithika Iyer

Author Yuning Mao
Author Michael Tontchev
Author Qing Hu
Author Brian Fuller
Author Davide Testuggine
Author others
Date 2023
Extra Citation Key: inan2023llama
Publication arXiv preprint arXiv:2312.06674
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLaMA: Open and efficient foundation language models

Item Type Journal Article
Author Hugo Touvron
Author Thibaut Lavril
Author Gautier Izacard
Author Xavier Martinet
Author Marie-Anne Lachaux
Author Timothée Lacroix
Author Baptiste Rozière
Author Naman Goyal
Author Eric Hambro
Author Faisal Azhar
Author Aurelien Rodriguez
Author Armand Joulin
Author Edouard Grave
Author Guillaume Lample
Date 2023
URL https://arxiv.org/abs/2302.13971
Extra Citation Key: touvron2023llama
Publication arXiv preprint arXiv:2302.13971
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

LLM agentic workflow for automated vulnerability detection and remediation in infrastructure-as-code

Item Type Journal Article
Author Dheer Toprani
Author Vijay K. Madisetti
Date 2025
Extra Citation Key: toprani2025agentforvulndetectioniac
Volume 13
Pages 69175-69181
Publication IEEE access : practical innovations, open solutions
DOI 10.1109/ACCESS.2025.3560911
Journal Abbr IEEE Access
Date Added 10/20/2025, 3:50:52 PM

Modified 10/20/2025, 3:50:52 PM

Tags:

Security, Large language models, large language models, Best practices, CI/CD, Cognition, Infrastructure-as-code, LLM workflows, Organizations, Retrieval augmented generation, Runtime, Scalability, security automation, Static analysis, Vectors, vulnerability detection

---

LLM agents can autonomously exploit one-day vulnerabilities

Item Type Document
Author Richard Fang
Author Rohan Bindu
Author Akul Gupta
Author Daniel Kang
Date 2024
URL https://arxiv.org/abs/2404.08144
Extra Citation Key: fang2024llmagentsautonomouslyexploit arXiv: 2404.08144 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Llm agents can autonomously hack websites

Item Type Journal Article
Author Richard Fang
Author Rohan Bindu
Author Akul Gupta
Author Qiusi Zhan
Author Daniel Kang
Date 2024
Extra Citation Key: fang2024llm
Publication arXiv preprint arXiv:2402.06664
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLM agents can misuse tools: a security study of tool-augmented large language models

Item Type Journal Article
Author Yiming Fu
Author Xiang Zhang
Author Yuxin Zhou
Author Shuyin Zhang
Author Yang Liu
Author Hao Chen
Author Xin Zhang
Author Ce Zhang
Date 2024
URL https://arxiv.org/abs/2410.14923
Extra Citation Key: fu2024tooluse

Publication arXiv preprint arXiv:2410.14923
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

LLM agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models

Item Type Journal Article
Author Ivar Frisch
Author Mario Giulianelli
Date 2024
Extra Citation Key: frisch2024llm
Publication arXiv preprint arXiv:2402.02896
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Llm defenses are not robust to multi-turn human jailbreaks yet

Item Type Journal Article
Author Nathaniel Li
Author Ziwen Han
Author Ian Steneker
Author Willow Primack
Author Riley Goodside
Author Hugh Zhang
Author Zifan Wang
Author Cristina Menghini
Author Summer Yue
Date 2024
Extra Citation Key: li2024llm
Publication arXiv preprint arXiv:2408.15221
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Llm for patient-trial matching: Privacy-aware data augmentation towards better performance and generalizability

Item Type Conference Paper
Author Jiayi Yuan
Author Ruixiang Tang
Author Xiaoqian Jiang
Author Xia Hu
Date 2023
Extra Citation Key: yuan2023llm
Proceedings Title American medical informatics association (AMIA) annual symposium
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLM multi-agent systems: Challenges and open problems

Item Type Document
Author Shanshan Han
Author Qifan Zhang
Author Yuhang Yao
Author Weizhao Jin
Author Zhaozhuo Xu
Date 2025
URL https://arxiv.org/abs/2402.03578
Extra Citation Key: han2025llmmultiagentsystemschallenges arXiv: 2402.03578 [cs.MA]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

LLM robustness against misinformation in biomedical question answering

Item Type Journal Article
Author Alexander Bondarenko
Author Adrian Viehweger
Date 2024
Extra Citation Key: bondarenko2024llm
Publication arXiv preprint arXiv:2410.21330
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Llm with tools: A survey

Item Type Journal Article
Author Zhuocheng Shen
Date 2024
Extra Citation Key: shen2024llm
Publication arXiv preprint arXiv:2409.18807
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLM-AIx: An open source pipeline for Information Extraction from unstructured medical text based on privacy preserving Large Language Models

Item Type Journal Article
Author Isabella Catharina Wiest
Author Fabian Wolf
Author Marie-Elisabeth Leßmann
Author Marko van Treeck
Author Dyke Ferber
Author Jiefu Zhu
Author Heiko Boehme
Author Keno K Bressem
Author Hannes Ulrich

Author Matthias P Ebert
Author others
Date 2024
Extra Citation Key: wiest2024llm
Publication medRxiv : the preprint server for health sciences
Journal Abbr medRxiv
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLM-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification

Item Type Journal Article
Author Yiping Song
Author Juhua Zhang
Author Zhiliang Tian
Author Yuxin Yang
Author Minlie Huang
Author Dongsheng Li
Date 2024
Extra Citation Key: song2024llm
Publication arXiv preprint arXiv:2402.16515
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLM-driven provenance forensics for threat investigation and detection

Item Type Document
Author Kunal Mukherjee
Author Murat Kantarcioglu
Date 2025
URL https://arxiv.org/abs/2508.21323
Extra Citation Key: mukherjee2025llmdrivenprovenanceforensicsthreat arXiv: 2508.21323 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

LLM-Fuzzer: Scaling assessment of large language model jailbreaks

Item Type Conference Paper
Author Jiahao Yu
Author Xingwei Lin
Author Zheng Yu
Author Xinyu Xing
Date 2024-08
URL https://www.usenix.org/conference/usenixsecurity24/presentation/yu-jiahao
Extra Citation Key: yu2024llmfuzzer
Place Philadelphia, PA
Publisher USENIX Association

ISBN 978-1-939133-44-1
Pages 4657–4674
Proceedings Title 33rd USENIX security symposium (USENIX security 24)
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

LLM-powered automated cloud forensics: From log analysis to investigation

Item Type Conference Paper
Author Dalal Alharthi
Author Rozhin Yasaei
Date 2025
Extra Citation Key: alharthi2025llmpoweredcloudforensics
Pages 12-22
Proceedings Title 2025 IEEE 18th international conference on cloud computing (CLOUD)
DOI 10.1109/CLOUD67622.2025.00012
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

Biological system modeling, Automation, Large language models, Accuracy, Adaptation models, Adaptive Prompt Engineering, Cloud computing security, Cloud Forensics, Cloud Security, Forensic Intelligence, Forensics, Large Language Models (LLMs), Log Prioritization, Manuals, Robustness, Threat assessment, Threat Detection

---

LLM-SmartAudit: Advanced smart contract vulnerability detection

Item Type Document
Author Wei et al.
Abstract Augments static analysis with LLM reasoning for Solidity auditing; improves recall on reentrancy, arithmetic errors, and unsafe external calls relative to traditional tools.
Date 2024
Extra Citation Key: wei2024llmsmartaudit
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

arXiv preprint

LLM4CVE: Enabling iterative automated vulnerability repair with large language models

Item Type Document
Author Mohamad Fakih
Author Rahul Dharmaji
Author Halima Bouzidi
Author Gustavo Quiros Araya
Author Oluwatosin Ogundare
Author Mohammad Abdullah Al Faruque

Date 2025
URL https://arxiv.org/abs/2501.03446
Extra Citation Key: fakih2025llm4cveenablingiterativeautomated arXiv: 2501.03446 [cs.SE]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

LLMCloudHunter: Harnessing llms for automated extraction of detection rules from cloud-based CTI

Item Type Conference Paper
Author Yuval Schwartz
Author Lavi Ben-Shimol
Author Dudu Mimran
Author Yuval Elovici
Author Asaf Shabtai
Abstract As the number and sophistication of cyber attacks have increased, threat hunting has become a critical aspect of active security, enabling proactive detection and mitigation of threats before they cause significant harm. Open-source cyber threat intelligence (OSCTI) is a valuable resource for threat hunters, however, it often comes in unstructured formats that require further manual analysis. Previous studies aimed at automating OSCTI analysis are limited since (1) they failed to provide actionable outputs, (2) they did not take advantage of images present in OSCTI sources, and (3) they focused on on-premises environments, overlooking the growing importance of cloud environments. To address these gaps, we propose LLMCloudHunter, a novel framework that leverages large language models (LLMs) to automatically generate generic-signature detection rule candidates from textual and visual OSCTI data. We evaluated the quality of the rules generated by the proposed framework using 20 annotated real-world cloud threat reports. The results show that our framework achieved a precision of 83% and recall of 99% for the task of accurately extracting API calls made by the threat actor and a precision of 99% with a recall of 97% for IoCs. Additionally, 99.18% of the generated detection rule candidates were successfully compiled and converted into Splunk queries.
Date 2025
URL https://doi.org/10.1145/3696410.3714798
Extra Citation Key: schwartz2025llmcloudhunter Number of pages: 20 tex.address: New York, NY, USA
Place Sydney NSW, Australia
Publisher Association for Computing Machinery
ISBN 979-8-4007-1274-6
Pages 1922–1941
Series Www '25
Proceedings Title Proceedings of the ACM on web conference 2025
DOI 10.1145/3696410.3714798
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

cloud, cyber threat intelligence (cti), llm, sigma rules

LLMGuard: guarding against unsafe LLM behavior

Item Type Conference Paper
Author Shubh Goyal
Author Medha Hira
Author Shubham Mishra
Author Sukriti Goyal
Author Arnav Goel

Author Niharika Dadu
Author DB Kirushikesh
Author Sameep Mehta
Author Nishtha Madaan
Date 2024
Extra Citation Key: goyal2024llmguard Number: 21
Volume 38
Pages 23790–23792
Proceedings Title Proceedings of the AAAI conference on artificial intelligence
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

LLMs in the SOC: An empirical study of human-AI collaboration in security operations centres

Item Type Document
Author Ronal Singh
Author Shahroz Tariq
Author Fatemeh Jalalvand
Author Mohan Baruwal Chhetri
Author Surya Nepal
Author Cecile Paris
Author Martin Lochner
Date 2025
URL https://arxiv.org/abs/2508.18947
Extra Citation Key: singh2025llmssocempiricalstudy arXiv: 2508.18947 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

LLMSecConfig: An LLM-based approach for fixing software container misconfigurations

Item Type Document
Author Ziyang Ye
Author Triet Huynh Minh Le
Author M. Ali Babar
Abstract Combines static analysis with LLMs to automatically repair container/Kubernetes misconfigurations while preserving functionality; evaluation shows high fix rates with limited regressions.
Date 2025
URL https://arxiv.org/abs/2502.02009
Extra Citation Key: ye2025llmsecconfig
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Locus: Agentic predicate synthesis for directed fuzzing

Item Type Document
Author Jie Zhu
Author Chihao Shen
Author Ziyang Li

Author Jiahao Yu
Author Yizheng Chen
Author Kexin Pei
Date 2025
URL https://arxiv.org/abs/2508.21302
Extra Citation Key: zhu2025locusagenticpredicatesynthesis arXiv: 2508.21302 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Make agent defeat agent: automatic detection of taint-style vulnerabilities in LLM-based agents

Item Type Conference Paper
Author Fengyu Liu
Author Yuan Zhang
Author Jiaqi Luo
Author Jiarun Dai
Author Tian Chen
Author Letian Yuan
Author Zhengmin Yu
Author Youkun Shi
Author Ke Li
Author Chengyuan Zhou
Author Hao Chen
Author Min Yang
Date 2025
Extra Citation Key: Liu2025MakeAgentDefeatAgent Number of pages: 19 tex.address: USA tex.articleno: 194
Place Seattle, WA, USA
Publisher USENIX Association
ISBN 978-1-939133-52-6
Series Sec '25
Proceedings Title Proceedings of the 34th USENIX conference on security symposium
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

MalGEN: a generative agent framework for modeling malicious software in cybersecurity

Item Type Document
Author Bikash Saha
Author Sandeep Kumar Shukla
Date 2025
URL https://arxiv.org/abs/2506.07586
Extra Citation Key: saha2025malgengenerativeagentframework arXiv: 2506.07586 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Malicious agents: Exploiting agent communication and context propagation in LLM-based systems

Item Type Journal Article

Author Omer Zychlinski
Author Roee Peleg
Author Tal Rozen
Author Guy Katz
Author Asaf Shabtai
Date 2025
URL https://arxiv.org/abs/2509.00124
Extra Citation Key: zychlinski2025maliciousagents
Publication arXiv preprint arXiv:2509.00124
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Manias, panics, and crashes: a history of financial crises

Item Type Book
Author Charles P. Kindleberger
Author Robert Aliber
Date 2011
Extra Citation Key: kindleberger2011
Publisher Palgrave Macmillan
Edition 6
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Many-shot jailbreaking

Item Type Conference Paper
Author Cem Anil
Author Esin DURMUS
Author Nina Rimsky
Author Mrinank Sharma
Author Joe Benton
Author Sandipan Kundu
Author Joshua Batson
Author Meg Tong
Author Jesse Mu
Author Daniel J Ford
Author Francesco Mosconi
Author Rajashree Agrawal
Author Rylan Schaeffer
Author Naomi Bashkansky
Author Samuel Svenningsen
Author Mike Lambert
Author Ansh Radhakrishnan
Author Carson Denison
Author Evan J Hubinger
Author Yuntao Bai
Author Trenton Bricken
Author Timothy Maxwell

Author Nicholas Schiefer
Author James Sully
Author Alex Tamkin
Author Tamera Lanham
Author Karina Nguyen
Author Tomasz Korbak
Author Jared Kaplan
Author Deep Ganguli
Author Samuel R. Bowman
Author Ethan Perez
Author Roger Baker Grosse
Author David Duvenaud
Date 2024
URL https://openreview.net/forum?id=cw5mgd71jW
Extra Citation Key: anil2024manyshot
Proceedings Title The thirty-eighth annual conference on neural information processing systems
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Marketsenseai 2.0: Enhancing stock analysis through llm agents

Item Type Journal Article
Author George Fatouros
Author Kostas Metaxas
Author John Soldatos
Author Manos Karathanassis
Date 2025
Extra Citation Key: fatouros2025marketsenseai
Publication arXiv preprint arXiv:2502.00415
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

MART: Improving LLM safety with multi-round automatic red-teaming

Item Type Document
Author Suyu Ge
Author Chunting Zhou
Author Rui Hou
Author Madian Khabsa
Author Yi-Chia Wang
Author Qifan Wang
Author Jiawei Han
Author Yuning Mao
Date 2023
URL https://arxiv.org/abs/2311.07689
Extra Citation Key: ge2023martimprovingllmsafety arXiv: 2311.07689 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

MasRouter: Learning to route llms for multi-agent systems

Item Type	Document
Author	Yanwei Yue
Author	Guibin Zhang
Author	Boyang Liu
Author	Guancheng Wan
Author	Kun Wang
Author	Dawei Cheng
Author	Yiyan Qi
Date	2025
URL	https://arxiv.org/abs/2502.11133
Extra	Citation Key: masrouter arXiv: 2502.11133 [cs.LG]
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

MASS: Muli-agent simulation scaling for portfolio construction

Item Type	Document
Author	Taian Guo
Author	Haiyang Shen
Author	JinSheng Huang
Author	Zhengyang Mao
Author	Junyu Luo
Author	Binqi Chen
Author	Zhuoru Chen
Author	Luchen Liu
Author	Bingyu Xia
Author	Xuhui Liu
Author	Yun Ma
Author	Ming Zhang
Abstract	The application of LLM-based agents in financial investment has shown significant promise, yet existing approaches often require intermediate steps like predicting individual stock movements or rely on predefined, static workflows. These limitations restrict their adaptability and effectiveness in constructing optimal portfolios. In this paper, we introduce the Multi-Agent Scaling Simulation (MASS), a novel framework that leverages multi-agent simulation for direct, end-to-end portfolio construction. At its core, MASS employs a backward optimization process to dynamically learn the optimal distribution of heterogeneous agents, enabling the system to adapt to evolving market regimes. A key finding enabled by our framework is the exploration of the scaling effect for portfolio construction: we demonstrate that as the number of agents increases exponentially (up to 512), the aggregated decisions yield progressively higher excess returns. Extensive experiments on a challenging, self-collected dataset from the 2023 Chinese A-share market show that MASS consistently outperforms seven state-of-the-art baselines. Further backtesting, stability analyses and the experiment on data leakage concerns validate its enhanced profitability and robustness. We have open-sourced our code, dataset, and training snapshots at https://github.com/gta0804/MASS/ to foster further research.
Date	2025-09
Short Title	MASS
URL	http://arxiv.org/abs/2505.10278
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: guo_mass_2025 DOI: 10.48550/arXiv.2505.10278
Publisher	arXiv
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence

Notes:

arXiv:2505.10278 [cs]

Med-r 2: Crafting trustworthy LLM physicians through retrieval and reasoning of evidence-based medicine

Item Type	Journal Article
Author	Keer Lu
Author	Zheng Liang
Author	Da Pan
Author	Shusen Zhang
Author	Xin Wu
Author	Weipeng Chen
Author	Zenan Zhou
Author	Guosheng Dong
Author	Bin Cui
Author	Wentao Zhang
Date	2025
Extra	Citation Key: lu2025med
Publication	arXiv preprint arXiv:2501.11885
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Medfuzz: Exploring the robustness of large language models in medical question answering

Item Type	Journal Article
Author	Robert Osazuwa Ness
Author	Katie Matton
Author	Hayden Helm
Author	Sheng Zhang
Author	Junaid Bajwa
Author	Carey E Priebe
Author	Eric Horvitz
Date	2024
Extra	Citation Key: ness2024medfuzz
Publication	arXiv preprint arXiv:2406.06573
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Membership inference attacks cannot prove that a model was trained on your data

Item Type	Journal Article
Author	Jie Zhang
Author	Debeshee Das

Author Gautam Kamath
Author Florian Tramèr
Date 2024
Extra Citation Key: zhang2024membership
Publication arXiv preprint arXiv:2409.19798
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Membership inference attacks from first principles

Item Type Conference Paper
Author Nicholas Carlini
Author Steve Chien
Author Milad Nasr
Author Shuang Song
Author Andreas Terzis
Author Florian Tramer
Date 2022
Extra Citation Key: carlini2022membership
Publisher IEEE
Pages 1897–1914
Proceedings Title 2022 IEEE symposium on security and privacy (SP)
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Memory manipulation attacks against LLM agents

Item Type Journal Article
Author Zhichen Dong
Author Zhanhui Zhou
Author Chao Yang
Author Jing Shao
Author Yu Qiao
Date 2025
URL https://arxiv.org/abs/2503.03704
Extra Citation Key: dong2025memorypoisoning
Publication arXiv preprint arXiv:2503.03704
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

MetaGPT: Meta programming for a multi-agent collaborative framework

Item Type Conference Paper
Author Sirui Hong
Author Mingchen Zhuge
Author Jonathan Chen
Author Xiawu Zheng
Author Yuheng Cheng

Author	Jinlin Wang
Author	Ceyao Zhang
Author	Zili Wang
Author	Steven Ka Shing Yau
Author	Zijuan Lin
Author	Liyang Zhou
Author	Chenyu Ran
Author	Lingfeng Xiao
Author	Chenglin Wu
Author	Jürgen Schmidhuber
Date	2024
URL	https://openreview.net/forum?id=VtmBAGCN7o
Extra	Citation Key: hong2024metagpt
Proceedings Title	The twelfth international conference on learning representations
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Mimicking the familiar: Dynamic command generation for information theft attacks in LLM tool-learning system

Item Type	Journal Article
Author	Ziyou Jiang
Author	Mingyang Li
Author	Guowei Yang
Author	Junjie Wang
Author	Yuekai Huang
Author	Zhiyuan Chang
Author	Qing Wang
Date	2025
Extra	Citation Key: jiang2025mimicking
Publication	arXiv preprint arXiv:2502.11358
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Mind the privacy unit! user-level differential privacy for language model fine-tuning

Item Type	Journal Article
Author	Lynn Chua
Author	Badih Ghazi
Author	Yangsibo Huang
Author	Pritish Kamath
Author	Ravi Kumar
Author	Daogao Liu
Author	Pasin Manurangsi
Author	Amer Sinha
Author	Chiyuan Zhang
Date	2024
Extra	Citation Key: chua2024mind
Publication	arXiv preprint arXiv:2406.14322

Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Misusing tools in large language models with visual adversarial examples

Item Type Journal Article
Author Xiaohan Fu
Author Zihan Wang
Author Shuheng Li
Author Rajesh K Gupta
Author Niloofar Mireshghallah
Author Taylor Berg-Kirkpatrick
Author Earlence Fernandes
Date 2023
Extra Citation Key: fu2023misusing
Publication arXiv preprint arXiv:2310.03185
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Mitigating object hallucination in large vision-language models via classifier-free guidance

Item Type Journal Article
Author Linxi Zhao
Author Yihe Deng
Author Weitong Zhang
Author Quanquan Gu
Date 2024
Extra Citation Key: zhao2024mitigating
Publication arXiv preprint arXiv:2402.08680
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Mitigating privacy risks in LLM embeddings from embedding inversion

Item Type Journal Article
Author Tiantian Liu
Author Hongwei Yao
Author Tong Wu
Author Zhan Qin
Author Feng Lin
Author Kui Ren
Author Chun Chen
Date 2024
Extra Citation Key: liu2024mitigating
Publication arXiv preprint arXiv:2411.05034
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Mixtral of experts

Item Type	Journal Article
Author	Albert Jiang
Author	Teven Le Scao
Author	Stanislav Bekman
Author	others
Date	2024
Extra	Citation Key: jiang2024mixtral
Publication	arXiv preprint arXiv:2401.04088
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

MM-DREX: Multimodal-Driven Dynamic Routing of LLM Experts for Financial Trading

Item Type	Document
Author	Yang Chen
Author	Yueheng Jiang
Author	Zhaozhao Ma
Author	Yuchen Cao
Author	Jacky Keung
Author	Kun Kuang
Author	Leilei Gan
Author	Yiquan Wu
Author	Fei Wu
Abstract	The inherent non-stationarity of financial markets and the complexity of multi-modal information pose significant challenges to existing quantitative trading models. Traditional methods relying on fixed structures and unimodal data struggle to adapt to market regime shifts, while large language model (LLM)-driven solutions - despite their multi-modal comprehension - suffer from static strategies and homogeneous expert designs, lacking dynamic adjustment and fine-grained decision mechanisms. To address these limitations, we propose MM-DREX: a Multimodal-driven, Dynamically-Routed EXpert framework based on large language models. MM-DREX explicitly decouples market state perception from strategy execution to enable adaptive sequential decision-making in non-stationary environments. Specifically, it (1) introduces a vision-language model (VLM)-powered dynamic router that jointly analyzes candlestick chart patterns and long-term temporal features to allocate real-time expert weights; (2) designs four heterogeneous trading experts (trend, reversal, breakout, positioning) generating specialized fine-grained sub-strategies; and (3) proposes an SFT-RL hybrid training paradigm to synergistically optimize the router's market classification capability and experts' risk-adjusted decision-making. Extensive experiments on multi-modal datasets spanning stocks, futures, and cryptocurrencies demonstrate that MM-DREX significantly outperforms 15 baselines (including state-of-the-art financial LLMs and deep reinforcement learning models) across key metrics: total return, Sharpe ratio, and maximum drawdown, validating its robustness and generalization. Additionally, an interpretability module traces routing logic and expert behavior in real time, providing an audit trail for strategy transparency.
Date	2025-09
Short Title	MM-DREX
URL	http://arxiv.org/abs/2509.05080
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: chen_mm-drex_2025 DOI: 10.48550/arXiv.2509.05080
Publisher	arXiv
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Tags:

Quantitative Finance - Trading and Market Microstructure

Notes:

arXiv:2509.05080 [q-fin]

Model context protocol (MCP)

Item Type	Document
Author	MCP
Date	2024
URL	https://github.com/modelcontextprotocol
Extra	Citation Key: openai2024mcp
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

ModelGuard: Information-Theoretic defense against model extraction attacks

Item Type	Conference Paper
Author	Minxue Tang
Author	Anna Dai
Author	Louis DiValentin
Author	Aolin Ding
Author	Amin Hass
Author	Neil Zhenqiang Gong
Author	Yiran Chen
Author	Hai "Helen" Li
Date	2024-08
URL	https://www.usenix.org/conference/usenixsecurity24/presentation/tang
Extra	Citation Key: 294591
Place	Philadelphia, PA
Publisher	USENIX Association
ISBN	978-1-939133-44-1
Pages	5305–5322
Proceedings Title	33rd USENIX security symposium (USENIX security 24)
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Moral alignment for LLM agents

Item Type	Journal Article
Author	Elizaveta Tennant
Author	Stephen Hailes
Author	Mirco Musolesi
Date	2024
Extra	Citation Key: tennant2024moral
Publication	arXiv preprint arXiv:2410.01639
Date Added	10/20/2025, 3:49:09 PM

Modified 10/20/2025, 3:49:09 PM

More than you've asked for: A comprehensive analysis of indirect prompt injection attacks on large language models

Item Type Journal Article
Author Karl Greshake
Author Rainer Schuster
Author Frederik Ritz
Author Dominik Strohmeier
Author Christian Reuter
Author Ben Stock
Date 2023
URL https://arxiv.org/abs/2302.12173
Extra Citation Key: greshake2023prompt
Publication arXiv preprint arXiv:2302.12173
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Mrj-agent: An effective jailbreak agent for multi-round dialogue

Item Type Journal Article
Author Fengxiang Wang
Author Ranjie Duan
Author Peng Xiao
Author Xiaojun Jia
Author YueFeng Chen
Author Chongwen Wang
Author Jialing Tao
Author Hang Su
Author Jun Zhu
Author Hui Xue
Date 2024
Extra Citation Key: wang2024mrj
Publication arXiv preprint arXiv:2411.03814
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Multi-agent architecture search via agentic supernet

Item Type Journal Article
Author Guibin Zhang
Author Luyang Niu
Author Junfeng Fang
Author Kun Wang
Author Lei Bai
Author Xiang Wang
Date 2025

Extra Citation Key: zhang2025multi
Publication arXiv preprint arXiv:2502.04180
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Multi-agent collaboration: Harnessing the power of intelligent llm agents

Item Type Journal Article
Author Yashar Talebirad
Author Amirhossein Nadiri
Date 2023
Extra Citation Key: talebirad2023multi
Publication arXiv preprint arXiv:2306.03314
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Multi-agent penetration testing AI for the web (MAPTA)

Item Type Document
Author Isaac David
Author Arthur Gervais
Abstract Multi-agent web app pentesting with tool-grounded execution and exploit validation. On the XBOW benchmark, achieves strong success across SSRF, misconfiguration, broken auth, SSTI, and SQLi; includes real-world disclosures.
Date 2025
URL https://arxiv.org/abs/2508.20816
Extra Citation Key: david2025mapta
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Multi-agent risks from advanced AI

Item Type Journal Article
Author Lewis Hammond
Author Alan Chan
Author Jesse Clifton
Author Jason Hoelscher-Obermaier
Author Akbir Khan
Author Euan McLean
Author Chandler Smith
Author Wolfram Barfuss
Author Jakob Foerster
Author Tomáš Gavenčiak
Author others
Date 2025
Extra Citation Key: hammond2025multi
Publication arXiv preprint arXiv:2502.14143
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Multiagent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate

Item Type Journal Article
Author Alfonso Amayuelas
Author Xianjun Yang
Author Antonis Antoniades
Author Wenyue Hua
Author Liangming Pan
Author William Wang
Date 2024
Extra Citation Key: amayuelas2024multiagent
Publication arXiv preprint arXiv:2406.14711
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Multilingual blending: Llm safety alignment evaluation with language mixture

Item Type Journal Article
Author Jiayang Song
Author Yuheng Huang
Author Zhehua Zhou
Author Lei Ma
Date 2024
Extra Citation Key: song2024multilingual
Publication arXiv preprint arXiv:2407.07342
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Multimodal backdoor attacks and defenses in vision-language models

Item Type Journal Article
Author Eugene Bagdasaryan
Author Vitaly Shmatikov
Date 2023
URL https://arxiv.org/abs/2302.10149
Extra Citation Key: bagdasaryan2023multimodalattacks
Publication arXiv preprint arXiv:2302.10149
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Navigating the risks: a survey of security, privacy, and ethics threats in LLM-based agents

Item Type Document
Author Yuyou Gan
Author Yong Yang
Author Zhe Ma
Author Ping He

Author Rui Zeng
Author Yiming Wang
Author Qingming Li
Author Chunyi Zhou
Author Songze Li
Author Ting Wang
Author Yunjun Gao
Author Yingcai Wu
Author Shouling Ji
Date 2024
URL https://arxiv.org/abs/2411.09523
Extra Citation Key: gan2024navigatingrisksurveysecurity arXiv: 2411.09523 [cs.AI]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Netsafe: Exploring the topological safety of multi-agent networks

Item Type Journal Article
Author Miao Yu
Author Shilong Wang
Author Guibin Zhang
Author Junyuan Mao
Author Chenlong Yin
Author Qijiong Liu
Author Qingsong Wen
Author Kun Wang
Author Yang Wang
Date 2024
Extra Citation Key: yu2024netsafe
Publication arXiv preprint arXiv:2410.15686
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection

Item Type Conference Paper
Author Kai Greshake
Author Sahar Abdelnabi
Author Shailesh Mishra
Author Christoph Endres
Author Thorsten Holz
Author Mario Fritz
Date 2023
Extra Citation Key: greshake2023not
Pages 79–90
Proceedings Title Proceedings of the 16th ACM workshop on artificial intelligence and security
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Of Models and Tin Men–a behavioural economics study of principal-agent problems in AI alignment using large-language models

Item Type Journal Article
Author Steve Phelps
Author Rebecca Ranson
Date 2023
Extra Citation Key: phelps2023models
Publication arXiv preprint arXiv:2307.11137
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

On the feasibility of using llms to autonomously execute multi-host network attacks

Item Type Document
Author Brian Singer
Author Keane Lucas
Author Lakshmi Adiga
Author Meghna Jain
Author Lujo Bauer
Author Vyas Sekar
Date 2025
URL https://arxiv.org/abs/2501.16466
Extra Citation Key: singer2025feasibilityusingllmsautonomously arXiv: 2501.16466 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

On the surprising efficacy of llms for penetration-testing

Item Type Document
Author Andreas Happe
Author Jürgen Cito
Date 2025
URL https://arxiv.org/abs/2507.00829
Extra Citation Key: happe2025surprisingefficacyllmspenetrationtesting arXiv: 2507.00829 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

On the vulnerability of applying retrieval-augmented generation within knowledge-intensive application domains

Item Type Journal Article
Author Xun Xian
Author Ganghua Wang
Author Xuan Bi
Author Jayanth Srinivasa
Author Ashish Kundu
Author Charles Fleming

Author Mingyi Hong
Author Jie Ding
Date 2024
Extra Citation Key: xian2024vulnerability
Publication arXiv preprint arXiv:2409.17275
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

One llm is not enough: Harnessing the power of ensemble learning for medical question answering

Item Type Journal Article
Author Han Yang
Author Mingchen Li
Author Huixue Zhou
Author Yongkang Xiao
Author Qian Fang
Author Rui Zhang
Date 2023
Extra Citation Key: yang2023one
Publication medRxiv : the preprint server for health sciences
Journal Abbr medRxiv
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Open challenges in multi-agent security: Towards secure systems of interacting AI agents

Item Type Document
Author Christian Schroeder de Witt
Date 2025
URL https://arxiv.org/abs/2505.02077
Extra Citation Key: dewitt2025openchallengesmultiagentsecurity arXiv: 2505.02077 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Open the pandora's box of llms: Jailbreaking llms through representation engineering

Item Type Journal Article
Author Tianlong Li
Author Xiaoqing Zheng
Author Xuanjing Huang
Date 2024
Extra Citation Key: li2024open
Publication arXiv preprint arXiv:2401.06824
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Optimization-based prompt injection attack to llm-as-a-judge

Item Type Conference Paper
Author Jiawen Shi
Author Zenghui Yuan
Author Yinuo Liu
Author Yue Huang
Author Pan Zhou
Author Lichao Sun
Author Neil Zhenqiang Gong
Date 2024
Extra Citation Key: shi2024optimization
Pages 660–674
Proceedings Title Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

OS agents: a survey on MLLM-based agents for computer, phone and browser use

Item Type Conference Paper
Author Xueyu Hu
Author Tao Xiong
Author Biao Yi
Author Zishu Wei
Author Ruixuan Xiao
Author Yurun Chen
Author Jiasheng Ye
Author Meiling Tao
Author Xiangxin Zhou
Author Ziyu Zhao
Author Yuhuai Li
Author Shengze Xu
Author Shenzhi Wang
Author Xinchen Xu
Author Shuofei Qiao
Author Zhaokai Wang
Author Kun Kuang
Author Tieyong Zeng
Author Liang Wang
Author Jiwei Li
Author Yuchen Eleanor Jiang
Author Wangchunshu Zhou
Author Guoyin Wang
Author Keting Yin
Author Zhou Zhao
Author Hongxia Yang
Author Fan Wu
Author Shengyu Zhang
Author Fei Wu
Editor Wanxiang Che
Editor Joyce Nabende
Editor Ekaterina Shutova

Editor Mohammad Taher Pilehvar
Date 2025-07
URL https://aclanthology.org/2025.acl-long.369/
Extra Citation Key: hu-etal-2025-os
Place Vienna, Austria
Publisher Association for Computational Linguistics
ISBN 979-8-89176-251-0
Pages 7436–7465
Proceedings Title Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)
DOI 10.18653/v1/2025.acl-long.369
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Pandora: Detailed llm jailbreaking via collaborated phishing agents with decomposed reasoning

Item Type Conference Paper
Author Zhaorun Chen
Author Zhuokai Zhao
Author Wenjie Qu
Author Zichen Wen
Author Zhiguang Han
Author Zhihong Zhu
Author Jiaheng Zhang
Author Huaxiu Yao
Date 2024
Extra Citation Key: chen2024pandora
Proceedings Title ICLR 2024 workshop on secure and trustworthy large language models
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Paper summary attack: Jailbreaking llms through LLM safety papers

Item Type Document
Author Liang Lin
Author Zhihao Xu
Author Xuehai Tang
Author Shi Liu
Author Biyu Zhou
Author Fuqing Zhu
Author Jizhong Han
Author Songlin Hu
Date 2025
URL https://arxiv.org/abs/2507.13474
Extra Citation Key: lin2025papersummaryattackjailbreaking arXiv: 2507.13474 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

PentestAgent: Incorporating LLM agents to automated penetration testing

Item Type Conference Paper
Author Xiangmin Shen
Author Lingzhi Wang
Author Zhenyuan Li
Author Yan Chen
Author Wencheng Zhao
Author Dawei Sun
Author Jiashui Wang
Author Wei Ruan
Abstract Penetration testing is a critical technique for identifying security vulnerabilities, traditionally performed manually by skilled security specialists. This complex process involves gathering information about the target system, identifying entry points, exploiting the system, and reporting findings. Despite its effectiveness, manual penetration testing is time-consuming and expensive, often requiring significant expertise and resources that many organizations cannot afford. While automated penetration testing methods have been proposed, they often fall short in real-world applications due to limitations in flexibility, adaptability, and implementation.Recent advancements in large language models offer new opportunities for enhancing penetration testing through increased intelligence and automation. However, current LLM-based approaches still face significant challenges, including limited penetration testing knowledge and a lack of comprehensive automation capabilities. To address these gaps, we propose PentestAgent, a novel LLM-based automated penetration testing framework that leverages the power of LLMs and various LLM-based techniques like retrieval augmented generation to enhance penetration testing knowledge and automate various tasks. Our framework leverages multi-agent collaboration to automate intelligence gathering, vulnerability analysis, and exploitation stages, reducing manual intervention. We evaluate PentestAgent using a comprehensive benchmark, demonstrating superior performance in task completion and overall efficiency.
Date 2025
URL https://doi.org/10.1145/3708821.3733882
Extra Citation Key: Shen2025PentestAgent Number of pages: 17
Place New York, NY, USA
Publisher Association for Computing Machinery
ISBN 979-8-4007-1410-8
Pages 375–391
Series Asia ccs '25
Proceedings Title Proceedings of the 20th ACM asia conference on computer and communications security
DOI 10.1145/3708821.3733882
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Tags:

Agent, Large Language Model, Penetration Testing

PentestGPT: Evaluating and harnessing large language models for automated penetration testing

Item Type Conference Paper
Author Gelei Deng
Author Yi Liu
Author Víctor Mayoral-Vilches
Author Peng Liu
Author Yuekang Li
Author Yuan Xu
Author Tianwei Zhang
Author Yang Liu
Author Martin Pinzger

Author Stefan Rass
Date 2024-08
URL https://www.usenix.org/conference/usenixsecurity24/presentation/deng
Extra Citation Key: Deng2024PentestGPT
Place Philadelphia, PA
Publisher USENIX Association
ISBN 978-1-939133-44-1
Pages 847–864
Proceedings Title 33rd USENIX security symposium (USENIX security 24)
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Personal llm agents: Insights and survey about the capability, efficiency and security

Item Type Journal Article
Author Yuanchun Li
Author Hao Wen
Author Weijun Wang
Author Xiangyu Li
Author Yizhen Yuan
Author Guohong Liu
Author Jiacheng Liu
Author Wenxing Xu
Author Xiang Wang
Author Yi Sun
Author others
Date 2024
Extra Citation Key: li2024personal
Publication arXiv preprint arXiv:2401.05459
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

PhishDebate: An LLM-based multi-agent framework for phishing website detection

Item Type Document
Author Wenhao Li
Author Selvakumar Manickam
Author Yung-wey Chong
Author Shankar Karuppayah
Date 2025
URL https://arxiv.org/abs/2506.15656
Extra Citation Key: li2025phishdebatellmbasedmultiagentframework arXiv: 2506.15656 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Pleak: Prompt leaking attacks against large language model applications

Item Type Conference Paper

Author Bo Hui
Author Haolin Yuan
Author Neil Gong
Author Philippe Burlina
Author Yinzhi Cao
Date 2024
Extra Citation Key: hui2024pleak
Pages 3600–3614
Proceedings Title Proceedings of the 2024 on ACM SIGSAC conference on computer and communications security
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Plug in the safety chip: Enforcing constraints for llm-driven robot agents

Item Type Conference Paper
Author Ziyi Yang
Author Shreyas S Raman
Author Ankit Shah
Author Stefanie Tellex
Date 2024
Extra Citation Key: yang2024plug
Publisher IEEE
Pages 14435–14442
Proceedings Title 2024 IEEE international conference on robotics and automation (ICRA)
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

PoisonBench: Assessing large language model vulnerability to poisoned preference data

Item Type Conference Paper
Author Tingchen Fu
Author Mrinank Sharma
Author Philip Torr
Author Shay B Cohen
Author David Krueger
Author Fazl Barez
Date 2025
URL https://openreview.net/forum?id=21kAulloDG
Extra Citation Key: fu2025poisonbench
Proceedings Title Forty-second international conference on machine learning
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Poisonedrag: Knowledge corruption attacks to retrieval-augmented generation of large language models

Item Type Journal Article
Author Wei Zou
Author Runpeng Geng

Author Binghui Wang
Author Jinyuan Jia
Date 2024
Extra Citation Key: zou2024poisonedrag
Publication arXiv preprint arXiv:2402.07867
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

PoisonedRAG: knowledge corruption attacks to retrieval-augmented generation of large language models

Item Type Conference Paper
Author Wei Zou
Author Runpeng Geng
Author Binghui Wang
Author Jinyuan Jia
Date 2025
Extra Citation Key: Zou2025PoisonedRAG Number of pages: 18 tex.address: USA tex.articleno: 197
Place Seattle, WA, USA
Publisher USENIX Association
ISBN 978-1-939133-52-6
Series Sec '25
Proceedings Title Proceedings of the 34th USENIX conference on security symposium
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Poisoning retrieval corpora by injecting adversarial passages

Item Type Journal Article
Author Zexuan Zhong
Author Ziqing Huang
Author Alexander Wettig
Author Danqi Chen
Date 2023
Extra Citation Key: zhong2023poisoning
Publication arXiv preprint arXiv:2310.19156
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Polaris: A safety-focused llm constellation architecture for healthcare

Item Type Journal Article
Author Subhabrata Mukherjee
Author Paul Gamble
Author Markel Sanz Ausin
Author Neel Kant
Author Kriti Aggarwal
Author Neha Manjunath

Author Debajyoti Datta
Author Zhengliang Liu
Author Jiayuan Ding
Author Sophia Busacca
Author others
Date 2024
Extra Citation Key: mukherjee2024polaris
Publication arXiv preprint arXiv:2403.13313
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Position: Standard benchmarks fail–LLM agents present overlooked risks for financial applications

Item Type Journal Article
Author Zichen Chen
Author Jiaao Chen
Author Jianda Chen
Author Misha Sra
Date 2025
Extra Citation Key: chen2025position
Publication arXiv preprint arXiv:2502.15865
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Prioritizing safeguarding over autonomy: Risks of llm agents for science

Item Type Journal Article
Author Xiangru Tang
Author Qiao Jin
Author Kunlun Zhu
Author Tongxin Yuan
Author Yichi Zhang
Author Wangchunshu Zhou
Author Meng Qu
Author Yilun Zhao
Author Jian Tang
Author Zhuosheng Zhang
Author others
Date 2024
Extra Citation Key: tang2024prioritizing
Publication arXiv preprint arXiv:2402.04247
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Privacy in action: Towards realistic privacy mitigation and evaluation for LLM-powered agents

Item Type Journal Article
Author Shouju Wang

Author Fenglin Yu
Author Xirui Liu
Author Xiaoting Qin
Author Jue Zhang
Author Qingwei Lin
Author Dongmei Zhang
Author Saravan Rajmohan
Date 2025
URL https://arxiv.org/abs/2509.17488
Extra Citation Key: wang2025privacyinaction
Volume arXiv:2509.17488
Publication arXiv preprint
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

preprint, submitted 22 Sep 2025, cs.CR / cs.AI

Privacy leakage overshadowed by views of AI: a study on human oversight of privacy in language model agent

Item Type Journal Article
Author Zhiping Zhang
Author Bingcan Guo
Author Tianshi Li
Date 2024
Extra Citation Key: zhang2024privacy
Publication arXiv preprint arXiv:2411.01344
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Privacy-engineered value decomposition networks for cooperative multi-agent reinforcement learning

Item Type Conference Paper
Author Parham Gohari
Author Matthew Hale
Author Ufuk Topcu
Date 2023
Extra Citation Key: gohari2023privacy
Publisher IEEE
Pages 8038–8044
Proceedings Title 2023 62nd IEEE conference on decision and control (CDC)
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Privacy-preserving AI for encrypted medical imaging: a framework for secure diagnosis and learning

Item Type Document
Author Abdullah Al Siam
Author Sadequzzaman Shohan
Abstract Proposes encrypted-inference pipeline (AES-CBC + JPEG2000; masked CNN) enabling diagnosis on encrypted images with marginal accuracy/latency trade-offs on NIH ChestX-ray14 and LIDC-IDRI.
Date 2025
URL https://arxiv.org/abs/2507.21060
Extra Citation Key: privacymedical2025
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Privacyasst: Safeguarding user privacy in tool-using large language model agents

Item Type Journal Article
Author Xinyu Zhang
Author Huiyu Xu
Author Zhongjie Ba
Author Zhibo Wang
Author Yuan Hong
Author Jian Liu
Author Zhan Qin
Author Kui Ren
Date 2024
Extra Citation Key: zhang2024privacyasst Publisher: IEEE
Publication IEEE Transactions on Dependable and Secure Computing
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Privacylens: Evaluating privacy norm awareness of language models in action

Item Type Journal Article
Author Yijia Shao
Author Tianshi Li
Author Weiyan Shi
Author Yanchen Liu
Author Diyi Yang
Date 2025
Extra Citation Key: shao2025privacylens
Volume 37
Pages 89373–89407
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

PrivAgent: Agentic-based red-teaming for LLM privacy leakage

Item Type Journal Article
Author Yuzhou Nie

Author Zhun Wang
Author Ye Yu
Author Xian Wu
Author Xuandong Zhao
Author Wenbo Guo
Author Dawn Song
Date 2024
Extra Citation Key: nie2024privagent
Publication arXiv preprint arXiv:2412.05734
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Progent: Programmable privilege control for LLM agents

Item Type Document
Author Tianneng Shi
Author Jingxuan He
Author Zhun Wang
Author Linyu Wu
Author Hongwei Li
Author Wenbo Guo
Author Dawn Song
Abstract Defines a DSL and runtime to express/enforce fine-grained privilege policies over agent tool calls, enabling dynamic updates and safe fallbacks; demonstrates strong security-utility tradeoffs on agent benchmarks.
Date 2025
URL https://arxiv.org/abs/2504.11703
Extra Citation Key: shi2025progent
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Prompt flow integrity to prevent privilege escalation in LLM agents

Item Type Document
Author Juhee Kim
Author Woohyuk Choi
Author Byoungyoung Lee
Abstract PFI enforces safe information flow for agent tool use via untrusted data identification, least-privilege enforcement, and validation of unsafe flows; mitigates privilege escalation while preserving utility.
Date 2025
URL https://arxiv.org/abs/2503.15547
Extra Citation Key: jumiratna2025promptflow
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Prompt infection: LLM-to-LLM prompt injection within multi-agent systems

Item Type Document
Author Donghyun Lee
Author Mo Tiwari

Date 2024
URL https://arxiv.org/abs/2410.07283
Extra Citation Key: lee2024promptinfectionllmtollmprompt arXiv: 2410.07283 [cs.MA]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Prompt injection attack against LLM-integrated applications

Item Type Document
Author Yi Liu
Author Gelei Deng
Author Yuekang Li
Author Kailong Wang
Author Zihao Wang
Author Xiaofeng Wang
Author Tianwei Zhang
Author Yepang Liu
Author Haoyu Wang
Author Yan Zheng
Author Yang Liu
Date 2024
URL https://arxiv.org/abs/2306.05499
Extra Citation Key: liu2024promptinjectionattackllmintegrated arXiv: 2306.05499 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Prompt Leakage effect and defense strategies for multi-turn LLM interactions

Item Type Journal Article
Author Divyansh Agarwal
Author Alexander R Fabbri
Author Ben Risher
Author Philippe Laban
Author Shafiq Joty
Author Chien-Sheng Wu
Date 2024
Extra Citation Key: agarwal2024prompt
Publication arXiv preprint arXiv:2404.16251
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Prompt leaks: How large language models leak sensitive information and what we can do about it

Item Type Journal Article
Author Max Ayzenshteyn
Author Elena Dvorkin
Author Dhruv Patel
Author Maxim Cherepanov

Author Konstantinos Tsirlis
Author Edward Raff
Date 2024
URL https://arxiv.org/abs/2410.15396
Extra Citation Key: ayzenshteyn2024promptleaks
Publication arXiv preprint arXiv:2410.15396
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Prompt stealing attacks against {text-to-image} generation models

Item Type Conference Paper
Author Xinyue Shen
Author Yiting Qu
Author Michael Backes
Author Yang Zhang
Date 2024
Extra Citation Key: shen2024prompt
Pages 5823–5840
Proceedings Title 33rd USENIX security symposium (USENIX security 24)
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Prompt stealing attacks against large language models

Item Type Journal Article
Author Zeyang Sha
Author Yang Zhang
Date 2024
Extra Citation Key: sha2024prompt
Publication arXiv preprint arXiv:2402.12959
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

PromptGuard: Soft prompt-guided unsafe content moderation for text-to-image models

Item Type Journal Article
Author Lingzhi Yuan
Author Xinfeng Li
Author Chejian Xu
Author Guanhong Tao
Author Xiaojun Jia
Author Yihao Huang
Author Wei Dong
Author Yang Liu
Author XiaoFeng Wang
Author Bo Li
Date 2025

Extra Citation Key: yuan2025promptguard
Publication arXiv preprint arXiv:2501.03544
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

PromptSleuth: Detecting Prompt Injection via Semantic Intent Invariance

Item Type Preprint
Author Mengxiao Wang
Author Yuxuan Zhang
Author Guofei Gu
Abstract Large Language Models (LLMs) are increasingly integrated into real-world applications, from virtual assistants to autonomous agents. However, their flexibility also introduces new attack vectors-particularly Prompt Injection (PI), where adversaries manipulate model behavior through crafted inputs. As attackers continuously evolve with paraphrased, obfuscated, and even multi-task injection strategies, existing benchmarks are no longer sufficient to capture the full spectrum of emerging threats. To address this gap, we construct a new benchmark that systematically extends prior efforts. Our benchmark subsumes the two widely-used existing ones while introducing new manipulation techniques and multi-task scenarios, thereby providing a more comprehensive evaluation setting. We find that existing defenses, though effective on their original benchmarks, show clear weaknesses under our benchmark, underscoring the need for more robust solutions. Our key insight is that while attack forms may vary, the adversary's intent-injecting an unauthorized task-remains invariant. Building on this observation, we propose PromptSleuth, a semantic-oriented defense framework that detects prompt injection by reasoning over task-level intent rather than surface features. Evaluated across state-of-the-art benchmarks, PromptSleuth consistently outperforms existing defense while maintaining comparable runtime and cost efficiency. These results demonstrate that intent-based semantic reasoning offers a robust, efficient, and generalizable strategy for defending LLMs against evolving prompt injection threats.
Date 2025-09-16
Short Title PromptSleuth
Library Catalog arXiv.org
URL http://arxiv.org/abs/2508.20890
Accessed 10/20/2025, 4:25:05 PM
Extra arXiv:2508.20890 [cs]
DOI 10.48550/arXiv.2508.20890
Repository arXiv
Archive ID arXiv:2508.20890
Date Added 10/20/2025, 4:25:05 PM
Modified 10/20/2025, 4:25:05 PM

Tags:

Computer Science - Cryptography and Security

Attachments

- Full Text PDF
- Snapshot

Prsa: Prompt reverse stealing attacks against large language models

Item Type Journal Article
Author Yong Yang
Author Xuhong Zhang
Author Yi Jiang

Author Xi Chen
Author Haoyu Wang
Author Shouling Ji
Author Zonghui Wang
Date 2024
Extra Citation Key: yang2024prsa
Publication arXiv preprint arXiv:2402.19200
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Pseudo-conversation injection for LLM goal hijacking

Item Type Document
Author Zheng Chen
Author Buhui Yao
Date 2024
URL https://arxiv.org/abs/2410.23678
Extra Citation Key: chen2024pseudoconversationinjectionllmgoal arXiv: 2410.23678 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

PSG-agent: Personality-aware safety guardrail for LLM-based agents

Item Type Document
Author Yaozu Wu
Author Jizhou Guo
Author Dongyuan Li
Author Henry Peng Zou
Author Wei-Chieh Huang
Author Yankai Chen
Author Zhen Wang
Author Weizhi Zhang
Author Yangning Li
Author Meng Zhang
Author Renhe Jiang
Author Philip S. Yu
Date 2025
URL https://arxiv.org/abs/2509.23614
Extra Citation Key: wu2025psgagentpersonalityawaresafetyguardrail arXiv: 2509.23614 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Psysafe: A comprehensive framework for psychological-based attack, defense, and evaluation of multi-agent system safety

Item Type Journal Article
Author Zaibin Zhang
Author Yongting Zhang

Author Lijun Li
Author Hongzhi Gao
Author Lijun Wang
Author Huchuan Lu
Author Feng Zhao
Author Yu Qiao
Author Jing Shao
Date 2024
Extra Citation Key: zhang2024psysafe
Publication arXiv preprint arXiv:2401.11880
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

QuantAgent: Price-Driven Multi-Agent LLMs for High-Frequency Trading

Item Type Document
Author Fei Xiong
Author Xiang Zhang
Author Aosong Feng
Author Siqi Sun
Author Chenyu You
Abstract Recent advances in Large Language Models (LLMs) have shown remarkable capabilities in financial reasoning and market understanding. Multi-agent LLM frameworks such as TradingAgent and FINMEM augment these models to long-horizon investment tasks by leveraging fundamental and sentiment-based inputs for strategic decision-making. However, these approaches are ill-suited for the high-speed, precision-critical demands of High-Frequency Trading (HFT). HFT typically requires rapid, risk-aware decisions driven by structured, short-horizon signals, such as technical indicators, chart patterns, and trend features. These signals stand in sharp contrast to the long-horizon, text-driven reasoning that characterizes most existing LLM-based systems in finance. To bridge this gap, we introduce QuantAgent, the first multi-agent LLM framework explicitly designed for high-frequency algorithmic trading. The system decomposes trading into four specialized agents–Indicator, Pattern, Trend, and Risk–each equipped with domain-specific tools and structured reasoning capabilities to capture distinct aspects of market dynamics over short temporal windows. Extensive experiments across nine financial instruments, including Bitcoin and Nasdaq futures, demonstrate that QuantAgent consistently outperforms baseline methods, achieving higher predictive accuracy at both 1-hour and 4-hour trading intervals across multiple evaluation metrics. Our findings suggest that coupling structured trading signals with LLM-based reasoning provides a viable path for traceable, real-time decision systems in high-frequency financial markets.
Date 2025-09
Short Title QuantAgent
URL http://arxiv.org/abs/2509.09995
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: xiong_quantagent_2025 DOI: 10.48550/arXiv.2509.09995
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2509.09995 [cs]

Quantifying misalignment between agents: Towards a sociotechnical understanding of alignment

Item Type Journal Article
Author Aidan Kierans
Author Avijit Ghosh
Author Hananel Hazan
Author Shiri Dori-Hacohen
Date 2024
Extra Citation Key: kierans2024quantifying
Publication arXiv preprint arXiv:2406.04231
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

R-judge: Benchmarking safety risk awareness for llm agents

Item Type Journal Article
Author Tongxin Yuan
Author Zhiwei He
Author Lingzhong Dong
Author Yiming Wang
Author Ruijie Zhao
Author Tian Xia
Author Lizhen Xu
Author Binglin Zhou
Author Fangqi Li
Author Zhuosheng Zhang
Author others
Date 2024
Extra Citation Key: yuan2024r
Publication arXiv preprint arXiv:2401.10019
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

R²-guard: Robust reasoning enabled LLM guardrail via knowledge-enhanced logical reasoning

Item Type Document
Author Mintong Kang
Author Bo Li
Abstract As LLMs become increasingly prevalent, robust safety guardrails are essential. We propose R²-Guard, which combines data-driven unsafety predictions and a reasoning component embedding domain knowledge as logical rules (via Markov logic networks or probabilistic circuits). The model fuses unsafety probabilities from per-category classifiers with logical inference over safety relationships, improving robustness to correlated safety violations and challenging jailbreaks. Empirical results show large gains over LlamaGuard and robustness across multiple safety benchmarks.
Date 2024
URL https://arxiv.org/abs/2407.05557
Extra Citation Key: kang2024r2guard
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks

Item Type Journal Article
Author Changyue Jiang
Author Xudong Pan
Author Geng Hong
Author Chenfu Bao
Author Min Yang
Date 2024
Extra Citation Key: jiang2024rag
Publication arXiv preprint arXiv:2411.14110
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

RAS-eval: a comprehensive benchmark for security evaluation of LLM agents in real-world environments

Item Type Document
Author Yuchuan Fu
Author Xiaohan Yuan
Author Dongxia Wang
Date 2025
URL https://arxiv.org/abs/2506.15253
Extra Citation Key: fu2025raseval arXiv: 2506.15253 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

ReAct: Synergizing reasoning and acting in language models

Item Type Document
Author Shunyu Yao
Author Jeffrey Zhao
Author Dian Yu
Author Nan Du
Author Izhak Shafran
Author Karthik Narasimhan
Author Yuan Cao
Date 2023
URL https://arxiv.org/abs/2210.03629
Extra Citation Key: yao2023reactsynergizingreasoningacting arXiv: 2210.03629 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Red team arxiv paper update: Continuous monitoring of emerging attacks on LLM systems

Item Type Journal Article
Author Zheyuan Huang

Author Eujeong Choi
Author Xin Wang
Author Yun Zhou
Date 2025
URL https://github.com/chen37058/Red-Team-Arxiv-Paper-Update
Extra Citation Key: huang2025redteamupdate
Publication GitHub repository
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Red teaming language model detectors with language models

Item Type Journal Article
Author Zhouxing Shi
Author Yihan Wang
Author Fan Yin
Author Xiangning Chen
Author Kai-Wei Chang
Author Cho-Jui Hsieh
Date 2024
URL https://aclanthology.org/2024.tacl-1.10/
Extra Citation Key: shi-etal-2024-red Place: Cambridge, MA Publisher: MIT Press
Volume 12
Pages 174–189
Publication Transactions of the Association for Computational Linguistics
DOI 10.1162/tacl_a_00639
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned

Item Type Journal Article
Author Deep Ganguli
Author Liane Lovitt
Author Jackson Kernion
Author Amanda Askell
Author Yuntao Bai
Author Saurav Kadavath
Author Ben Mann
Author Ethan Perez
Author Nicholas Schiefer
Author Kamal Ndousse
Author others
Date 2022
Extra Citation Key: ganguli2022red
Publication arXiv preprint arXiv:2209.07858
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Red teaming language models with language models

Item Type	Document
Author	Ethan Perez
Author	Saffron Huang
Author	Francis Song
Author	Trevor Cai
Author	Roman Ring
Author	John Aslanides
Author	Amelia Glaese
Author	Nat McAleese
Author	Geoffrey Irving
Date	2022
URL	https://arxiv.org/abs/2202.03286
Extra	Citation Key: perez2022redteaminglanguagemodels arXiv: 2202.03286 [cs.CL]
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Red-teaming LLM multi-agent systems via communication attacks

Item Type	Conference Paper
Author	Pengfei He
Author	Yuping Lin
Author	Shen Dong
Author	Han Xu
Author	Yue Xing
Author	Hui Liu
Editor	Wanxiang Che
Editor	Joyce Nabende
Editor	Ekaterina Shutova
Editor	Mohammad Taher Pilehvar
Date	2025-07
URL	https://aclanthology.org/2025.findings-acl.349/
Extra	Citation Key: he-etal-2025-red
Place	Vienna, Austria
Publisher	Association for Computational Linguistics
ISBN	979-8-89176-256-5
Pages	6726–6747
Proceedings Title	Findings of the association for computational linguistics: ACL 2025
DOI	10.18653/v1/2025.findings-acl.349
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Redagent: Red teaming large language models with context-aware autonomous language agent

Item Type	Journal Article
Author	Huiyu Xu
Author	Wenhui Zhang
Author	Zhibo Wang

Author Feng Xiao
Author Rui Zheng
Author Yunhe Feng
Author Zhongjie Ba
Author Kui Ren
Date 2024
Extra Citation Key: xu2024redagent
Publication arXiv preprint arXiv:2407.16667
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

RedCode: Risky code execution and generation benchmark for code agents

Item Type Journal Article
Author Chengquan Guo
Author Xun Liu
Author Chulin Xie
Author Andy Zhou
Author Yi Zeng
Author Zinan Lin
Author Dawn Song
Author Bo Li
Date 2025
Extra Citation Key: guo2025redcode
Volume 37
Pages 106190–106236
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Redirecting scientific discovery agents to toxic compounds: a study on safety of LLM-based agents

Item Type Journal Article
Author Haotian Li
Author Yue Zhao
Author Hanyu Sun
Author Weichen Wang
Author Xinyi Zhao
Date 2025
URL https://arxiv.org/abs/2502.08586
Extra Citation Key: li2025scientificagents
Publication arXiv preprint arXiv:2502.08586
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Refining input guardrails: Enhancing LLM-as-a-judge efficiency through chain-of-thought fine-tuning and alignment

Item Type Conference Paper
Author Melissa Kazemi Rad
Author Huy Nghiem
Author Sahil Wadhwa
Author Andy Luo
Author Mohammad Shahed Sorower
Date 2025
URL https://openreview.net/forum?id=UNPzbCKovl
Extra Citation Key: rad2025refining
Proceedings Title AAAI 2025 workshop on preventing and detecting LLM misinformation (PDLM)
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Reflexion: Language agents with verbal reinforcement learning

Item Type Document
Author Noah Shinn
Author Federico Cassano
Author Edward Berman
Author Ashwin Gopinath
Author Karthik Narasimhan
Author Shunyu Yao
Date 2023
URL https://arxiv.org/abs/2303.11366
Extra Citation Key: shinn2023reflexionlanguageagentsverbal arXiv: 2303.11366 [cs.AI]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Refusal-trained llms are easily jailbroken as browser agents

Item Type Journal Article
Author Priyanshu Kumar
Author Elaine Lau
Author Saranya Vijayakumar
Author Tu Trinh
Author Scale Red Team
Author Elaine Chang
Author Vaughn Robinson
Author Sean Hendryx
Author Shuyan Zhou
Author Matt Fredrikson
Author others
Date 2024
Extra Citation Key: kumar2024refusal
Publication arXiv preprint arXiv:2410.13886
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

RepairAgent: An autonomous, LLM-based agent for program repair

Item Type	Conference Paper
Author	Islem Bouzenia
Author	Premkumar Devanbu
Author	Michael Pradel
Abstract	Automated program repair has emerged as a powerful technique to mitigate the impact of software bugs on system reliability and user experience. This paper introduces Repair Agent, the first work to address the program repair challenge through an autonomous agent based on a large language model (LLM). Unlike existing deep learning-based approaches, which prompt a model with a fixed prompt or in a fixed feedback loop, our work treats the LLM as an agent capable of autonomously planning and executing actions to fix bugs by invoking suitable tools. Repair Agent freely interleaves gathering information about the bug, gathering repair ingredients, and validating fixes, while deciding which tools to invoke based on the gathered information and feedback from previous fix attempts. Key contributions that enable Repair Agent include a set of tools that are useful for program repair, a dynamically updated prompt format that allows the LLM to interact with these tools, and a finite state machine that guides the agent in invoking the tools. Our evaluation on the popular Defects4J dataset demonstrates Repair Agent's effectiveness in autonomously repairing 164 bugs, including 39 bugs not fixed by prior techniques. Interacting with the LLM imposes an average cost of 270k tokens per bug, which, under the current pricing of OpenAI's GPT-3.5 model, translates to 14 cents per bug. To the best of our knowledge, this work is the first to present an autonomous, LLM-based agent for program repair, paving the way for future agent-based techniques in software engineering.
Date	2025-05
URL	https://doi.ieeecomputersociety.org/10.1109/ICSE55347.2025.00157
Extra	Citation Key: bouzenia2025repairagent
Place	Los Alamitos, CA, USA
Publisher	IEEE Computer Society
Pages	2188-2200
Proceedings Title	2025 IEEE/ACM 47th international conference on software engineering (ICSE)
DOI	10.1109/ICSE55347.2025.00157
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Tags:

Large language models, Autonomous agents, Computer bugs, Maintenance engineering, Pricing, Reliability, Software, Software engineering, Translation, User experience

RepoAudit: An autonomous LLM-agent for repository-level code auditing

Item Type	Conference Paper
Author	Jinyao Guo
Author	Chengpeng Wang
Author	Xiangzhe Xu
Author	Zian Su
Author	Xiangyu Zhang
Date	2025
URL	https://openreview.net/forum?id=TXcifVbFpG
Extra	Citation Key: guo2025repoaudit
Proceedings Title	Forty-second international conference on machine learning
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

Retrieval-augmented generation for knowledge-intensive nlp tasks

Item Type Journal Article
Author Patrick Lewis
Author Ethan Perez
Author Aleksandra Piktus
Author Fabio Petroni
Author Vladimir Karpukhin
Author Naman Goyal
Author Heinrich Küttler
Author Mike Lewis
Author Wen-tau Yih
Author Tim Rocktäschel
Author others
Date 2020
Extra Citation Key: lewis2020retrieval
Volume 33
Pages 9459–9474
Publication Advances in neural information processing systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Retrieval-augmented generation for large language models: a survey

Item Type Journal Article
Author Yunfan Gao
Author Yun Xiong
Author Xinyu Gao
Author and others
Date 2023
Extra Citation Key: gao2023ragsurvey
Publication arXiv preprint arXiv:2312.10997
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

RevPRAG: Revealing poisoning attacks in retrieval-augmented generation through LLM activation analysis

Item Type Document
Author Xue Tan
Author Hao Luan
Author Mingyu Luo
Author Xiaoyan Sun
Author Ping Chen
Author Jun Dai
Date 2025
URL https://arxiv.org/abs/2411.18948
Extra Citation Key: tan2025revpragrevealingpoisoningattacks arXiv: 2411.18948 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM

Modified 10/20/2025, 3:50:52 PM

Riskawarebench: Towards evaluating physical risk awareness for high-level planning of llm-based embodied agents

Item Type Journal Article
Author Zihao Zhu
Author Bingzhe Wu
Author Zhengyou Zhang
Author Baoyuan Wu
Date 2024
Extra Citation Key: zhu2024riskawarebench
Pages arXiv–2408
Publication arXiv e-prints
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Risks from language models for automated mental healthcare: Ethics and structure for implementation

Item Type Journal Article
Author Declan Grabb
Author Max Lamparth
Author Nina Vasan
Date 2024
Extra Citation Key: grabb2024risks
Publication arXiv preprint arXiv:2406.11852
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Rlhf workflow: From reward modeling to online rlhf

Item Type Journal Article
Author Hanze Dong
Author Wei Xiong
Author Bo Pang
Author Haoxiang Wang
Author Han Zhao
Author Yingbo Zhou
Author Nan Jiang
Author Doyen Sahoo
Author Caiming Xiong
Author Tong Zhang
Date 2024
Extra Citation Key: dong2024rlhf
Publication arXiv preprint arXiv:2405.07863
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

S-eval: Automatic and adaptive test generation for benchmarking safety evaluation of large language models

Item Type Journal Article
Author Xiaohan Yuan
Author Jinfeng Li
Author Dongxia Wang
Author Yuefeng Chen
Author Xiaofeng Mao
Author Longtao Huang
Author Hui Xue
Author Wenhai Wang
Author Kui Ren
Author Jingyi Wang
Date 2024
Extra Citation Key: yuan2024s
Publication arXiv preprint arXiv:2405.14191
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Safe multi-agent reinforcement learning with natural language constraints

Item Type Journal Article
Author Ziyan Wang
Author Meng Fang
Author Tristan Tomilin
Author Fei Fang
Author Yali Du
Date 2024
Extra Citation Key: wang2024safe
Publication arXiv preprint arXiv:2405.20018
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

SafeAgentBench: a benchmark for safe task planning of embodied LLM agents

Item Type Journal Article
Author Sheng Yin
Author Xianghe Pang
Author Yuanzhuo Ding
Author Menglan Chen
Author Yutong Bi
Author Yichen Xiong
Author Wenhao Huang
Author Zhen Xiang
Author Jing Shao
Author Siheng Chen
Date 2024
Extra Citation Key: yin2024safeagentbench

Publication arXiv preprint arXiv:2412.13178
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

SafeArena: Evaluating the safety of autonomous web agents

Item Type Conference Paper
Author Ada Defne Tur
Author Nicholas Meade
Author Xing Han Lù
Author Alejandra Zambrano
Author Arkil Patel
Author Esin DURMUS
Author Spandana Gella
Author Karolina Stanczak
Author Siva Reddy
Date 2025
URL https://openreview.net/forum?id=7TrOBcxSvy
Extra Citation Key: tur2025safearena
Proceedings Title Forty-second international conference on machine learning
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Safeguarding decentralized social media: LLM agents for automating community rule compliance

Item Type Journal Article
Author Lucio La Cava
Author Andrea Tagarelli
Date 2024
Extra Citation Key: la2024safeguarding
Publication arXiv preprint arXiv:2409.08963
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Safely learning with private data: A federated learning framework for large language model

Item Type Journal Article
Author JiaYing Zheng
Author HaiNan Zhang
Author LingXiang Wang
Author WangJie Qiu
Author HongWei Zheng
Author ZhiMing Zheng
Date 2024
Extra Citation Key: zheng2024safely
Publication arXiv preprint arXiv:2406.14898
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Safety assessment of chinese large language models

Item Type Journal Article
Author Hao Sun
Author Zhexin Zhang
Author Jiawen Deng
Author Jiale Cheng
Author Minlie Huang
Date 2023
Extra Citation Key: sun2023safety
Publication arXiv preprint arXiv:2304.10436
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Safety at scale: a comprehensive survey of large model and agent safety

Item Type Journal Article
Author Xingjun Ma
Author Yifeng Gao
Author Yixu Wang
Author Ruofan Wang
Author Xin Wang
Author Ye Sun
Author Yifan Ding
Author Hengyuan Xu
Author Yunhao Chen
Author Yunhao Zhao
Author Hanxun Huang
Author Yige Li
Author Yutao Wu
Author Jiaming Zhang
Author Xiang Zheng
Author Yang Bai
Author Yiming Li
Author Zuxuan Wu
Author Xipeng Qiu
Author Jingfeng Zhang
Author Xudong Han
Author Haonan Li
Author Jun Sun
Author Cong Wang
Author Jindong Gu
Author Baoyuan Wu
Author Siheng Chen
Author Tianwei Zhang
Author Yang Liu
Author Mingming Gong
Author Tongliang Liu
Author Shirui Pan
Author Cihang Xie

Author Tianyu Pang
Author Yinpeng Dong
Author Ruoxi Jia
Author Yang Zhang
Author Shiqing Ma
Author Xiangyu Zhang
Author Neil Gong
Author Chaowei Xiao
Author Sarah Erfani
Author Tim Baldwin
Author Bo Li
Author Masashi Sugiyama
Author Dacheng Tao
Author James Bailey
Author Yu-Gang Jiang
Date 2025
URL http://dx.doi.org/10.1561/3300000051
Extra Citation Key: ma2025safetyatscale
Volume 8
Pages 254-469
Publication Foundations and Trends® in Privacy and Security
DOI 10.1561/3300000051
Issue 3-4
ISSN 2474-1558
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Safety layers in aligned large language models: The key to llm security

Item Type Journal Article
Author Shen Li
Author Liuyi Yao
Author Lan Zhang
Author Yaliang Li
Date 2024
Extra Citation Key: li2024safety
Publication arXiv preprint arXiv:2408.17003
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

SandboxEval: Towards securing test environment for untrusted code

Item Type Document
Author Rafiqul Rabin
Author Jesse Hostetler
Author Sean McGregor
Author Brett Weir
Author Nick Judd
Date 2025

URL https://arxiv.org/abs/2504.00018
Extra Citation Key: rabin2025sandboxevalsecuringtestenvironment arXiv: 2504.00018 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Scalable extraction of training data from (production) language models

Item Type Journal Article
Author Milad Nasr
Author Nicholas Carlini
Author Jonathan Hayase
Author Matthew Jagielski
Author A Feder Cooper
Author Daphne Ippolito
Author Christopher A Choquette-Choo
Author Eric Wallace
Author Florian Tramèr
Author Katherine Lee
Date 2023
Extra Citation Key: nasr2023scalable
Publication arXiv preprint arXiv:2311.17035
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Scaling trends for data poisoning in llms

Item Type Journal Article
Author Dillon Bowen
Author Brendan Murphy
Author Will Cai
Author David Khachaturov
Author Adam Gleave
Author Kellin Pelrine
Date Apr. 2025
URL https://ojs.aaai.org/index.php/AAAI/article/view/34929
Extra Citation Key: Bowen2025ScalingTrends
Volume 39
Pages 27206-27214
Publication Proceedings of the AAAI Conference on Artificial Intelligence
DOI 10.1609/aaai.v39i26.34929
Issue 26
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Searching for privacy risks in LLM agents via simulation

Item Type Document
Author Yanzhe Zhang

Author Diyi Yang
Date 2025
URL https://arxiv.org/abs/2508.10880
Extra Citation Key: zhang2025searchingprivacyrisksllm arXiv: 2508.10880 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

SEC-bench: Automated benchmarking of LLM agents on real-world software security tasks

Item Type Document
Author Hwiwon Lee
Author Ziqi Zhang
Author Hanxiao Lu
Author Lingming Zhang
Date 2025
URL https://arxiv.org/abs/2506.11791
Extra Citation Key: lee2025secbenchautomatedbenchmarkingllm arXiv: 2506.11791 [cs.LG]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Secure multi-LLM agentic AI and agentification for edge general intelligence by zero-trust: a survey

Item Type Document
Author Yinqiu Liu
Author Ruichen Zhang
Author Haoxiang Luo
Author Yijing Lin
Author Geng Sun
Author Dusit Niyato
Author Hongyang Du
Author Zehui Xiong
Author Yonggang Wen
Author Abbas Jamalipour
Author Dong In Kim
Author Ping Zhang
Date 2025
URL https://arxiv.org/abs/2508.19870
Extra Citation Key: liu2025securemultillmagenticai arXiv: 2508.19870 [cs.NI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Securing agentic AI: a comprehensive threat model and mitigation framework for generative AI agents

Item Type Document
Author Vineeth Sai Narajala
Author Om Narayan
Date 2025
URL https://arxiv.org/abs/2504.19956

Extra Citation Key: narajala2025securingagenticaicomprehensive arXiv: 2504.19956 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Securing amazon bedrock agents: Safeguarding against indirect prompt injections

Item Type Document
Author Amazon Web Services
Date 2024
Extra Citation Key: AWS2024BedrockAgents
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Notes:

AWS Technical Documentation / White Paper. Listed as "LLM AGENT (agent safety orchestrator)"

Securing multi-turn conversational language models from distributed backdoor triggers

Item Type Journal Article
Author Terry Tong
Author Jiashu Xu
Author Qin Liu
Author Muhao Chen
Date 2024
Extra Citation Key: tong2024securing
Publication arXiv preprint arXiv:2407.04151
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Security attacks on LLM-based code completion tools

Item Type Journal Article
Author Wen Cheng
Author Ke Sun
Author Xinyu Zhang
Author Wei Wang
Date 2024
Extra Citation Key: cheng2024security
Publication arXiv preprint arXiv:2408.11006
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Security matrix for multimodal agents on mobile devices: A systematic and proof of concept study

Item Type Journal Article
Author Yulong Yang

Author Xinshan Yang
Author Shuaidong Li
Author Chenhao Lin
Author Zhengyu Zhao
Author Chao Shen
Author Tianwei Zhang
Date 2024
Extra Citation Key: yang2024security
Publication arXiv preprint arXiv:2407.09295
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Security of AI agents

Item Type Document
Author Yifeng He
Author Ethan Wang
Author Yuyang Rong
Author Zifei Cheng
Author Hao Chen
Date 2024
URL https://arxiv.org/abs/2406.08689
Extra Citation Key: he2024securityaiagents arXiv: 2406.08689 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Self-alignment of large language models via multi-agent social simulation

Item Type Conference Paper
Author Xianghe Pang
Author Shuo Tang
Author Rui Ye
Author Yuxin Xiong
Author Bolun Zhang
Author Yanfeng Wang
Author Siheng Chen
Date 2024
Extra Citation Key: pang2024self
Proceedings Title ICLR 2024 workshop on large language model (LLM) agents
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

SELP: Generating safe and efficient task plans for robot agents with large language models

Item Type Journal Article
Author Yi Wu
Author Zikang Xiong
Author Yiran Hu

Author Shreyash S Iyengar
Author Nan Jiang
Author Aniket Bera
Author Lin Tan
Author Suresh Jagannathan
Date 2024
Extra Citation Key: wu2024selp
Publication arXiv preprint arXiv:2409.19471
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Sentence embedding leaks more information than you expect: Generative embedding inversion attack to recover the whole sentence

Item Type Journal Article
Author Haoran Li
Author Mingshi Xu
Author Yangqiu Song
Date 2023
Extra Citation Key: li2023sentence
Publication arXiv preprint arXiv:2305.03010
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

SentinelAgent: Graph-based anomaly detection in multi-agent systems

Item Type Document
Author Xu He
Author Di Wu
Author Yan Zhai
Author Kun Sun
Date 2025
URL https://arxiv.org/abs/2505.24201
Extra Citation Key: he2025sentinelagentgraphbasedanomalydetection arXiv: 2505.24201 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Seven security challenges that must be solved in cross-domain multi-agent LLM systems

Item Type Document
Author Ronny Ko
Author Jiseong Jeong
Author Shuyuan Zheng
Author Chuan Xiao
Author Tae-Wan Kim
Author Makoto Onizuka
Author Won-Yong Shin
Date 2025

URL https://arxiv.org/abs/2505.23847
Extra Citation Key: ko2025sevensecuritychallengessolved arXiv: 2505.23847 [cs.CR]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

SG-bench: Evaluating LLM safety generalization across diverse tasks and prompt types

Item Type Journal Article
Author Yutao Mou
Author Shikun Zhang
Author Wei Ye
Date 2025
Extra Citation Key: mou2025sg
Volume 37
Pages 123032–123054
Publication Advances in Neural Information Processing Systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Shieldlm: Empowering llms as aligned, customizable and explainable safety detectors

Item Type Journal Article
Author Zhexin Zhang
Author Yida Lu
Author Jingyuan Ma
Author Di Zhang
Author Rui Li
Author Pei Ke
Author Hao Sun
Author Lei Sha
Author Zhifang Sui
Author Hongning Wang
Author others
Date 2024
Extra Citation Key: zhang2024shieldlm
Publication arXiv preprint arXiv:2402.16444
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Simple prompt injection attacks can leak personal data observed by LLM agents during task execution

Item Type Document
Author Meysam Alizadeh
Author Zeynab Samei
Author Daria Stetsenko
Author Fabrizio Gilardi
Date 2025
URL https://arxiv.org/abs/2506.01055

Extra Citation Key: alizadeh2025simplepromptinjectionattacks arXiv: 2506.01055 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Simple synthetic data reduces sycophancy in large language models

Item Type Journal Article
Author Jerry Wei
Author Da Huang
Author Yifeng Lu
Author Denny Zhou
Author Quoc V Le
Date 2023
Extra Citation Key: wei2023simple
Publication arXiv preprint arXiv:2308.03958
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Simulating Financial Market via Large Language Model based Agents

Item Type Document
Author Shen Gao
Author Yuntao Wen
Author Minghang Zhu
Author Jianing Wei
Author Yuhan Cheng
Author Qunzi Zhang
Author Shuo Shang
Abstract Most economic theories typically assume that financial market participants are fully rational individuals and use mathematical models to simulate human behavior in financial markets. However, human behavior is often not entirely rational and is challenging to predict accurately with mathematical models. In this paper, we propose \textbf{A}gent-based \textbf{S}imulated \textbf{F}inancial \textbf{M}arket (ASFM), which first constructs a simulated stock market with a real order matching system. Then, we propose a large language model based agent as the stock trader, which contains the profile, observation, and tool-learning based action module. The trading agent can comprehensively understand current market dynamics and financial policy information, and make decisions that align with their trading strategy. In the experiments, we first verify that the reactions of our ASFM are consistent with the real stock market in two controllable scenarios. In addition, we also conduct experiments in two popular economics research directions, and we find that conclusions drawn in our \model align with the preliminary findings in economics research. Based on these observations, we believe our proposed ASFM provides a new paradigm for economic research.
Date 2024-06
URL http://arxiv.org/abs/2406.19966
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: gao_simulating_2024 DOI: 10.48550/arXiv.2406.19966
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Computation and Language

Notes:

arXiv:2406.19966 [cs]

Simulating rumor spreading in social networks using LLM agents

Item Type	Journal Article
Author	Tianrui Hu
Author	Dimitrios Liakopoulos
Author	Xiwen Wei
Author	Radu Marculescu
Author	Neeraja J Yadwadkar
Date	2025
Extra	Citation Key: hu2025simulating
Publication	arXiv preprint arXiv:2502.01450
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

Sizing up the global market portfolio

Item Type	Document
Author	MSCI Research
Date	2024
URL	https://www.msci.com/research-and-insights/blog-post/sizing-up-the-global-market-portfolio
Extra	Citation Key: msci2024
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Notes:

Accessed: 2025-10-07

SLM as guardian: Pioneering AI safety with small language models

Item Type	Journal Article
Author	Ohjoon Kwon
Author	Donghyeon Jeon
Author	Nayoung Choi
Author	Gyu-Hwung Cho
Author	Changbong Kim
Author	Hyunwoo Lee
Author	Inho Kang
Author	Sun Kim
Author	Taiwoo Park
Date	2024
Extra	Citation Key: kwon2024slm
Publication	arXiv preprint arXiv:2405.19795
Date Added	10/20/2025, 3:49:09 PM

Modified 10/20/2025, 3:49:09 PM

SmartLLM: Smart contract auditing using custom generative AI

Item Type Document
Author Jun Kevin
Author Pujianto Yugopuspito
Abstract Custom fine-tuned LLaMA 3.1 with RAG for Solidity auditing; reports improved accuracy/recall versus Mythril/Slither and zero-shot LLM prompts, with strong detection on reentrancy and access-control flaws.
Date 2025
URL https://arxiv.org/abs/2502.13167
Extra Citation Key: smartllm2025
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Smoothllm: Defending large language models against jailbreaking attacks

Item Type Journal Article
Author Alexander Robey
Author Eric Wong
Author Hamed Hassani
Author George J Pappas
Date 2023
Extra Citation Key: robey2023smoothllm
Publication arXiv preprint arXiv:2310.03684
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Software testing with large language models: Survey, landscape, and vision

Item Type Journal Article
Author Junjie Wang
Author Yuchao Huang
Author Chunyang Chen
Author Zhe Liu
Author Song Wang
Author Qing Wang
Date 2024-04
URL https://doi.org/10.1109/TSE.2024.3368208
Extra Citation Key: 10.1109/TSE.2024.3368208 Number of pages: 26 Publisher: IEEE Press tex.issue_date: April 2024
Volume 50
Pages 911–936
DOI 10.1109/TSE.2024.3368208
Issue 4
ISSN 0098-5589
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

SoK: Understanding Vulnerabilities in the Large Language Model Supply Chain

Item Type Document
Author Shenao Wang
Author Yanjie Zhao
Author Zhao Liu
Author Quanchen Zou
Author Haoyu Wang
Abstract Large Language Models (LLMs) transform artificial intelligence, driving advancements in natural language understanding, text generation, and autonomous systems. The increasing complexity of their development and deployment introduces significant security challenges, particularly within the LLM supply chain. However, existing research primarily focuses on content safety, such as adversarial attacks, jailbreaking, and backdoor attacks, while overlooking security vulnerabilities in the underlying software systems. To address this gap, this study systematically analyzes 529 vulnerabilities reported across 75 prominent projects spanning 13 lifecycle stages. The findings show that vulnerabilities are concentrated in the application (50.3%) and model (42.7%) layers, with improper resource control (45.7%) and improper neutralization (25.1%) identified as the leading root causes. Additionally, while 56.7% of the vulnerabilities have available fixes, 8% of these patches are ineffective, resulting in recurring vulnerabilities. This study underscores the challenges of securing the LLM ecosystem and provides actionable insights to guide future research and mitigation strategies.
Date 2025-02
Short Title SoK
URL http://arxiv.org/abs/2502.12497
Accessed 10/6/2025, 7:00:00 PM
Extra Citation Key: wang_sok_2025 DOI: 10.48550/arXiv.2502.12497
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Cryptography and Security

Notes:

arXiv:2502.12497 [cs]

Sources of hallucination by large language models on inference tasks

Item Type Journal Article
Author Nick McKenna
Author Tianyi Li
Author Liang Cheng
Author Mohammad Javad Hosseini
Author Mark Johnson
Author Mark Steedman
Date 2023
Extra Citation Key: mckenna2023sources
Publication arXiv preprint arXiv:2305.14552
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Special characters attack: Toward scalable training data extraction from large language models

Item Type Journal Article
Author Yang Bai
Author Ge Pei
Author Jindong Gu
Author Yong Yang
Author Xingjun Ma
Date 2024
Extra Citation Key: bai2024special
Publication arXiv preprint arXiv:2405.05990
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

ST-WebAgentBench: a benchmark for evaluating safety & trustworthiness in web agents

Item Type Conference Paper
Author Ido Levy
Author Ben Wiesel
Author Sami Marreed
Author Alon Oved
Author Avi Yaeli
Author Segev Shlomov
Date 2025
Extra Citation Key: Levy2025STWebAgentBench
Proceedings Title ArXiv
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Notes:

arXiv:2410.06703

StockSim: A Dual-Mode Order-Level Simulator for Evaluating Multi-Agent LLMs in Financial Markets

Item Type Document
Author Charidimos Papadakis
Author Giorgos Filandrianos
Author Angeliki Dimitriou
Author Maria Lymperaiou
Author Konstantinos Thomas
Author Giorgos Stamou
Abstract We present StockSim, an open-source simulation platform for systematic evaluation of large language models (LLMs) in realistic financial decision-making scenarios. Unlike previous toolkits that offer limited scope, StockSim delivers a comprehensive system that fully models market dynamics and supports diverse simulation modes of varying granularity. It incorporates critical real-world factors, such as latency, slippage, and order-book microstructure, that were previously neglected, enabling more faithful and insightful assessment of LLM-based trading agents. An extensible, role-based agent framework supports heterogeneous trading strategies and multi-agent coordination, making StockSim a uniquely capable testbed for NLP research on reasoning under uncertainty and sequential decision-making. We open-source all our code at https: //github.com/harrypapa2002

/StockSim.
Date 2025-07
Short Title StockSim
URL http://arxiv.org/abs/2507.09255
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: papadakis_stocksim_2025 DOI: 10.48550/arXiv.2507.09255
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Multiagent Systems, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2507.09255 [cs]

Struq: Defending against prompt injection with structured queries

Item Type Journal Article
Author Sizhe Chen
Author Julien Piet
Author Chawin Sitawarin
Author David Wagner
Date 2024
Extra Citation Key: chen2024struq
Publication arXiv preprint arXiv:2402.06363
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Survey of hallucination in natural language generation

Item Type Journal Article
Author Ziwei Ji
Author Nayeon Lee
Author Rita Frieske
Author Tiezheng Yu
Author Dan Su
Author Yan Xu
Author Etsuko Ishii
Author Ye Jin Bang
Author Andrea Madotto
Author Pascale Fung
Date 2023
Extra Citation Key: ji2023survey Publisher: ACM New York, NY
Volume 55
Pages 1–38
Publication ACM Computing Surveys
Issue 12

Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Tackling uncertainties in multi-agent reinforcement learning through integration of agent termination dynamics

Item Type Journal Article
Author Somnath Hazra
Author Pallab Dasgupta
Author Soumyajit Dey
Date 2025
Extra Citation Key: hazra2025tackling
Publication arXiv preprint arXiv:2501.12061
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Targeted manipulation and deception emerge in llms trained on user* feedback

Item Type Conference Paper
Author Marcus Williams
Author Micah Carroll
Author Adhyyan Narang
Author Constantin Weisser
Author Brendan Murphy
Author Anca Dragan
Extra Citation Key: williamstargeted
Proceedings Title The thirteenth international conference on learning representations
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Teams of LLM agents can exploit zero-day vulnerabilities

Item Type Document
Author Yuxuan Zhu
Author Antony Kellermann
Author Akul Gupta
Author Philip Li
Author Richard Fang
Author Rohan Bindu
Author Daniel Kang
Date 2025
URL https://arxiv.org/abs/2406.01637
Extra Citation Key: zhu2025teamsllmagentsexploit arXiv: 2406.01637 [cs.MA]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Test-time backdoor mitigation for black-box large language models with defensive demonstrations

Item Type Journal Article
Author Wenjie Mo
Author Jiashu Xu
Author Qin Liu
Author Jiongxiao Wang
Author Jun Yan
Author Chaowei Xiao
Author Muhao Chen
Date 2023
Extra Citation Key: mo2023test
Publication arXiv preprint arXiv:2311.09763
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Text embeddings reveal (almost) as much as text

Item Type Journal Article
Author John X Morris
Author Volodymyr Kuleshov
Author Vitaly Shmatikov
Author Alexander M Rush
Date 2023
Extra Citation Key: morris2023text
Publication arXiv preprint arXiv:2310.06816
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

The AI revolution: Opportunities and challenges for the finance sector

Item Type Journal Article
Author Carsten Maple
Author Lukasz Szpruch
Author Gregory Epiphaniou
Author Kalina Staykova
Author Simran Singh
Author William Penwarden
Author Yisi Wen
Author Zijian Wang
Author Jagdish Hariharan
Author Pavle Avramovic
Date 2023
URL https://arXiv.org/abs/2308.16538
Extra Citation Key: maple2023airevolution
Publication arXiv preprint arXiv:2308.16538
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

The alignment problem from a deep learning perspective

Item Type Journal Article
Author Richard Ngo
Author Lawrence Chan
Author Sören Mindermann
Date 2022
Extra Citation Key: ngo2022alignment
Publication arXiv preprint arXiv:2209.00626
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

The dark side of llms: Agent-based attacks for complete computer takeover

Item Type Document
Author Matteo Lupinacci
Author Francesco Aurelio Pironti
Author Francesco Blefari
Author Francesco Romeo
Author Luigi Arena
Author Angelo Furfaro
Date 2025
URL https://arxiv.org/abs/2507.06850
Extra Citation Key: lupinacci2025darkllmsagentbasedattacks arXiv: 2507.06850 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

The economics of money, banking, and financial markets

Item Type Book
Author Frederic S. Mishkin
Date 2018
Extra Citation Key: mishkin2018
Publisher Pearson
Edition 12
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

The effects of reward misspecification: Mapping and mitigating misaligned models

Item Type Conference Paper
Author Alexander Pan
Author Kush Bhatia
Author Jacob Steinhardt
Date 2021
URL https://openreview.net/forum?id=mp1AstNFvQ5
Extra Citation Key: pan2021the
Proceedings Title Deep RL workshop NeurIPS 2021
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

The emerged security and privacy of llm agent: A survey with case studies

Item Type Journal Article
Author Feng He
Author Tianqing Zhu
Author Dayong Ye
Author Bo Liu
Author Wanlei Zhou
Author Philip S Yu
Date 2024
Extra Citation Key: he2024emerged
Publication arXiv preprint arXiv:2407.19354
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs)

Item Type Journal Article
Author Joschka Haltaufderheide
Author Robert Ranisch
Date 2024
Extra Citation Key: haltaufderheide2024ethics Publisher: Nature Publishing Group UK London
Volume 7
Pages 183
Publication NPJ digital medicine
Issue 1
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag)

Item Type Journal Article
Author Shenglai Zeng
Author Jiankun Zhang
Author Pengfei He
Author Yue Xing
Author Yiding Liu
Author Han Xu
Author Jie Ren
Author Shuaiqiang Wang
Author Dawei Yin
Author Yi Chang
Author others
Date 2024
Extra Citation Key: zeng2024good
Publication arXiv preprint arXiv:2402.16893
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

The hidden dangers of browsing AI agents

Item Type	Document
Author	Mykyta Mudryi
Author	Markiyan Chaklosh
Author	Grzegorz Wójcik
Abstract	Systematizes threats against browsing agents (prompt injection, plugin supply chain, cross-site abuse, credential theft) and proposes defense-in-depth including sanitization, planner/executor isolation, and session safeguards.
Date	2025
URL	https://arxiv.org/abs/2505.13076
Extra	Citation Key: mudryi2025hiddendangers
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey

Item Type	Journal Article
Author	Tula Masterman
Author	Sandi Besen
Author	Mason Sawtell
Author	Alex Chao
Date	2024
Extra	Citation Key: masterman2024landscape
Publication	arXiv preprint arXiv:2404.11584
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

The rise and potential of large language model based agents: A survey

Item Type	Journal Article
Author	Zhiheng Xi
Author	Wenxiang Chen
Author	Xin Guo
Author	Wei He
Author	Yiwen Ding
Author	Boyang Hong
Author	Ming Zhang
Author	Junzhe Wang
Author	Senjie Jin
Author	Enyu Zhou
Author	others
Date	2025
Extra	Citation Key: xi2025rise Publisher: Springer
Volume	68
Pages	121101
Publication	Science China Information Sciences
Issue	2
Date Added	10/20/2025, 3:49:08 PM
Modified	10/20/2025, 3:49:08 PM

The role of artificial intelligence in enhancing financial data security

Item Type Journal Article
Author KK Ramachandran
Date 2024
Extra Citation Key: ramachandran2024role
Volume 3
Pages 1–13
Publication INTERNATIONAL JOURNAL OF ARTIFICIAL INTELLIGENCE & APPLICATIONS (IJAIAP)
Issue 1
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

The size of the global financial market in 2024

Item Type Document
Author London Stock Exchange Group
Date 2024
URL https://www.lseg.com/content/dam/data-analytics/en_us/documents/charts/lseg-size-of-global-market-2024-in-charts.pdf
Extra Citation Key: lseg2024
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Notes:

Accessed: 2025-10-07

The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents

Item Type Conference Paper
Author Feiran Jia
Author Tong Wu
Author Xin Qin
Author Anna Squicciarini
Editor Wanxiang Che
Editor Joyce Nabende
Editor Ekaterina Shutova
Editor Mohammad Taher Pilehvar
Abstract Large Language Model (LLM) agents are increasingly being deployed as conversational assistants capable of performing complex real-world tasks through tool integration. This enhanced ability to interact with external systems and process various data sources, while powerful, introduces significant security vulnerabilities. In particular, indirect prompt injection attacks pose a critical threat, where malicious instructions embedded within external data sources can manipulate agents to deviate from user intentions. While existing defenses show promise, they struggle to maintain robust security while preserving task functionality. We propose a novel and orthogonal perspective that reframes agent security from preventing harmful actions to ensuring task alignment, requiring every agent action to serve user objectives. Based on this insight, we develop Task Shield, a test-time defense mechanism that systematically verifies whether each instruction and tool call contributes to user-specified goals. Through experiments on the AgentDojo benchmark, we demonstrate that Task Shield reduces attack success rates (2.07%) while maintaining high task utility (69.79%) on GPT-4o, significantly outperforming existing defenses in various real-world scenarios.

Date 2025-07
URL https://aclanthology.org/2025.acl-long.1435/
Extra Citation Key: jia-etal-2025-task
Place Vienna, Austria
Publisher Association for Computational Linguistics
ISBN 979-8-89176-251-0
Pages 29680–29697
Proceedings Title Proceedings of the 63rd annual meeting of the association for computational linguistics (volume 1: Long papers)
DOI 10.18653/v1/2025.acl-long.1435
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

The wolf within: Covert injection of malice into mllm societies via an mllm operative

Item Type Journal Article
Author Zhen Tan
Author Chengshuai Zhao
Author Raha Moraffah
Author Yifan Li
Author Yu Kong
Author Tianlong Chen
Author Huan Liu
Date 2024
Extra Citation Key: tan2024wolf
Publication arXiv preprint arXiv:2402.14859
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

This time is different: Eight centuries of financial folly

Item Type Book
Author Carmen M. Reinhart
Author Kenneth S. Rogoff
Date 2009
Extra Citation Key: reinhart2009
Publisher Princeton University Press
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

To protect the LLM agent against the prompt injection attack with polymorphic prompt

Item Type Conference Paper
Author Zhilong Wang
Author Neha Nagaraja
Author Lan Zhang
Author Hayretdin Bahsi
Author Pawan Patil
Author Peng Liu

Date 2025-06
URL https://doi.ieeecomputersociety.org/10.1109/DSN-S65789.2025.00037
Extra Citation Key: Wang2025ProtectLLMAgent
Place Los Alamitos, CA, USA
Publisher IEEE Computer Society
Pages 22-28
Proceedings Title 2025 55th annual IEEE/IFIP international conference on dependable systems and networks - supplemental volume (DSN-s)
DOI 10.1109/DSN-S65789.2025.00037
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

To Trade or Not to Trade: An Agentic Approach to Estimating Market Risk Improves Trading Decisions

Item Type Document
Author Dimitrios Emmanoulopoulos
Author Ollie Olby
Author Justin Lyon
Author Namid R. Stillman
Abstract Large language models (LLMs) are increasingly deployed in agentic frameworks, in which prompts trigger complex tool-based analysis in pursuit of a goal. While these frameworks have shown promise across multiple domains including in finance, they typically lack a principled model-building step, relying instead on sentiment- or trend-based analysis. We address this gap by developing an agentic system that uses LLMs to iteratively discover stochastic differential equations for financial time series. These models generate risk metrics which inform daily trading decisions. We evaluate our system in both traditional backtests and using a market simulator, which introduces synthetic but causally plausible price paths and news events. We find that model-informed trading strategies outperform standard LLM-based agents, improving Sharpe ratios across multiple equities. Our results show that combining LLMs with agentic model discovery enhances market risk estimation and enables more profitable trading decisions.
Date 2025-07
Short Title To Trade or Not to Trade
URL http://arxiv.org/abs/2507.08584
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: emmanoulopoulos_trade_2025 DOI: 10.48550/arXiv.2507.08584
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Statistical Finance, Computer Science - Multiagent Systems, Quantitative Finance - Computational Finance, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2507.08584 [q-fin]

Tool learning with foundation models

Item Type Journal Article
Author Yujia Qin
Author Shengding Hu

Author Yankai Lin
Author Weize Chen
Author Ning Ding
Author Ganqu Cui
Author Zheni Zeng
Author Xuanhe Zhou
Author Yufei Huang
Author Chaojun Xiao
Author others
Date 2024
Extra Citation Key: qin2024tool Publisher: ACM New York, NY
Volume 57
Pages 1–40
Publication ACM Computing Surveys
Issue 4
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Toolformer: Language models can teach themselves to use tools

Item Type Conference Paper
Author Timo Schick
Author Jane Dwivedi-Yu
Author Roberto Dessi
Author Roberta Raileanu
Author Maria Lomeli
Author Eric Hambro
Author Luke Zettlemoyer
Author Nicola Cancedda
Author Thomas Scialom
Date 2023
URL https://openreview.net/forum?id=Yacmpz84TH
Extra Citation Key: schick2023toolformer
Proceedings Title Thirty-seventh conference on neural information processing systems
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

ToolFuzz – automated agent tool testing

Item Type Document
Author Ivan Milev
Author Mislav Balunović
Author Maximilian Baader
Author Martin Vechev
Date 2025
URL https://arxiv.org/abs/2503.04479
Extra Citation Key: milev2025toolfuzzautomatedagent arXiv: 2503.04479 [cs.AI]
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Toolqa: A dataset for llm question answering with external tools

Item Type	Journal Article
Author	Yuchen Zhuang
Author	Yue Yu
Author	Kuan Wang
Author	Haotian Sun
Author	Chao Zhang
Date	2023
Extra	Citation Key: zhuang2023toolqa
Volume	36
Pages	50117–50143
Publication	Advances in Neural Information Processing Systems
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Toolsword: Unveiling safety issues of large language models in tool learning across three stages

Item Type	Journal Article
Author	Junjie Ye
Author	Sixian Li
Author	Guanyu Li
Author	Caishuang Huang
Author	Songyang Gao
Author	Yilong Wu
Author	Qi Zhang
Author	Tao Gui
Author	Xuanjing Huang
Date	2024
Extra	Citation Key: ye2024toolsword
Publication	arXiv preprint arXiv:2402.10753
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Towards a HIPAA compliant agentic AI system in healthcare

Item Type	Document
Author	Subash Neupane
Author	Shaswata Mitra
Author	Sudip Mittal
Author	Shahram Rahimi
Abstract	Framework for HIPAA-compliant agentic AI using ABAC, dual-stage PHI sanitization, and immutable audit trails to support autonomous clinical workflows under compliance constraints.
Date	2025
URL	https://arxiv.org/abs/2504.17669
Extra	Citation Key: neupane2025hipaa
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

Towards action hijacking of large language model-based agent

Item Type Document
Author Yuyang Zhang
Author Kangjie Chen
Author Jiaxin Gao
Author Ronghao Cui
Author Run Wang
Author Lina Wang
Author Tianwei Zhang
Date 2025
URL https://arxiv.org/abs/2412.10807
Extra Citation Key: zhang2025actionhijackinglargelanguage arXiv: 2412.10807 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Towards automated penetration testing: Introducing LLM benchmark, analysis, and improvements

Item Type Document
Author Isamu Isozaki
Author Manil Shrestha
Author Rick Console
Author Edward Kim
Date 2024
URL https://arxiv.org/abs/2410.17141
Extra Citation Key: isozaki2024automatedpenetrationtestingintroducing arXiv: 2410.17141 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Towards healthy AI: large language models need therapists too

Item Type Journal Article
Author Baihan Lin
Author Djallel Bouneffouf
Author Guillermo Cecchi
Author Kush R Varshney
Date 2023
Extra Citation Key: lin2023towards
Publication arXiv preprint arXiv:2304.00416
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Towards reliable healthcare llm agents: A case study for pilgrims during hajj

Item Type Journal Article
Author Hanan M Alghamdi
Author Abeer Mostafa
Date 2024

Extra Citation Key: alghamdi2024towards Publisher: MDPI
Volume 15
Pages 371
Publication Information-an International Interdisciplinary Journal
Issue 7
Journal Abbr Information
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Towards robust and secure embodied AI: a survey on vulnerabilities and attacks

Item Type Document
Author Wenpeng Xing
Author Minghao Li
Author Mohan Li
Author Meng Han
Abstract Survey of vulnerabilities/attacks specific to embodied AI (sensor spoofing, adversarial perception, planning/control hijacking); proposes a framework and research agenda for robust, secure embodied systems.
Date 2025
URL https://arxiv.org/abs/2502.13175
Extra Citation Key: xing2025embodiedai
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Trade in Minutes! Rationality-Driven Agentic System for Quantitative Financial Trading

Item Type Document
URL https://arxiv.org/html/2510.04787v1
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: noauthor_trade_nodate
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tradexpert: Revolutionizing trading with mixture of expert llms

Item Type Journal Article
Author Qianggang Ding
Author Haochen Shi
Author Jiadong Guo
Author Bang Liu
Date 2024
Extra Citation Key: ding2024tradexpert
Publication arXiv preprint arXiv:2411.00782
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

TradingAgents: Multi-Agents LLM Financial Trading Framework

Item Type Document
Author Yijia Xiao
Author Edward Sun
Author Di Luo
Author Wei Wang
Abstract Significant progress has been made in automated problem-solving using societies of agents powered by large language models (LLMs). In finance, efforts have largely focused on single-agent systems handling specific tasks or multi-agent frameworks independently gathering data. However, the multi-agent systems' potential to replicate real-world trading firms' collaborative dynamics remains underexplored. TradingAgents proposes a novel stock trading framework inspired by trading firms, featuring LLM-powered agents in specialized roles such as fundamental analysts, sentiment analysts, technical analysts, and traders with varied risk profiles. The framework includes Bull and Bear researcher agents assessing market conditions, a risk management team monitoring exposure, and traders synthesizing insights from debates and historical data to make informed decisions. By simulating a dynamic, collaborative trading environment, this framework aims to improve trading performance. Detailed architecture and extensive experiments reveal its superiority over baseline models, with notable improvements in cumulative returns, Sharpe ratio, and maximum drawdown, highlighting the potential of multi-agent LLM frameworks in financial trading. TradingAgents is available at https://github.com /TauricResearch/TradingAgents.
Date 2025-06
Short Title TradingAgents
URL http://arxiv.org/abs/2412.20138
Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: xiao_tradingagents_2025 DOI: 10.48550/arXiv.2412.20138
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Machine Learning, Quantitative Finance - Trading and Market Microstructure, and Science, Computer Science - Computational Engineering, Finance

Notes:

arXiv:2412.20138 [q-fin]

Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance

Item Type Journal Article
Author Yang Li
Author Yangyang Yu
Author Haohang Li
Author Zhi Chen
Author Khaldoun Khashanah
Date 2023
Extra Citation Key: li2023tradinggpt
Publication arXiv preprint arXiv:2309.03736
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

TradingGroup: A Multi-Agent Trading System with Self-Reflection and Data-Synthesis

Item Type Document
Author Feng Tian
Author Flora D. Salim
Author Hao Xue
Abstract Recent advancements in large language models (LLMs) have enabled powerful agent-based applications in finance, particularly for sentiment analysis, financial report comprehension, and stock forecasting. However, existing systems often lack inter-agent coordination, structured self-reflection, and access to high-quality, domain-specific post-training data such as data from trading activities including both market conditions and agent decisions. These data are crucial for agents to understand the market dynamics, improve the quality of decision-making and promote effective coordination. We introduce TradingGroup, a multi-agent trading system designed to address these limitations through a self-reflective architecture and an end-to-end data-synthesis pipeline. TradingGroup consists of specialized agents for news sentiment analysis, financial report interpretation, stock trend forecasting, trading style adaptation, and a trading decision making agent that merges all signals and style preferences to produce buy, sell or hold decisions. Specifically, we design self-reflection mechanisms for the stock forecasting, style, and decision-making agents to distill past successes and failures for similar reasoning in analogous future scenarios and a dynamic risk-management model to offer configurable dynamic stop-loss and take-profit mechanisms. In addition, TradingGroup embeds an automated data-synthesis and annotation pipeline that generates high-quality post-training data for further improving the agent performance through post-training. Our backtesting experiments across five real-world stock datasets demonstrate TradingGroup's superior performance over rule-based, machine learning, reinforcement learning, and existing LLM-based trading strategies.
Date 2025-08
Short Title TradingGroup
URL http://arxiv.org/abs/2508.17565
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: tian_tradinggroup_2025 DOI: 10.48550/arXiv.2508.17565
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence

Notes:

arXiv:2508.17565 [cs]

Training a helpful and harmless assistant with reinforcement learning from human feedback

Item Type Journal Article
Author Yuntao Bai
Author Andy Jones
Author Kamal Ndousse
Author Amanda Askell
Author Anna Chen
Author Nova DasSarma
Author Dawn Drain
Author Stanislav Fort
Author Deep Ganguli
Author Tom Henighan

Author others
Date 2022
Extra Citation Key: bai2022training
Publication arXiv preprint arXiv:2204.05862
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Training language models to follow instructions with human feedback

Item Type Journal Article
Author Long Ouyang
Author Jeffrey Wu
Author Xu Jiang
Author Diogo Almeida
Author Carroll Wainwright
Author Pamela Mishkin
Author Chong Zhang
Author Sandhini Agarwal
Author Katarina Slama
Author Alex Ray
Author others
Date 2022
Extra Citation Key: ouyang2022training
Volume 35
Pages 27730–27744
Publication Advances in neural information processing systems
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems

Item Type Document
Author Shaina Raza
Author Ranjan Sapkota
Author Manoj Karkee
Author Christos Emmanouilidis
Abstract Agentic AI systems, built upon large language models (LLMs) and deployed in multi-agent configurations, are redefining intelligence, autonomy, collaboration, and decision-making across enterprise and societal domains. This review presents a structured analysis of \textbf{Trust, Risk, and Security Management (TRiSM)} in the context of LLM-based Agentic Multi-Agent Systems (AMAS). We begin by examining the conceptual foundations of Agentic AI and highlight its architectural distinctions from traditional AI agents. We then adapt and extend the AI TRiSM framework for Agentic AI, structured around four key pillars: Explainability, ModelOps, Security, Privacy and Governance, each contextualized to the challenges of multi-agent LLM systems. A novel risk taxonomy is proposed to capture the unique threats and vulnerabilities of Agentic AI, ranging from coordination failures to prompt-based adversarial manipulation. To support practical assessment in Agentic AI works, we introduce two novel metrics: the Component Synergy Score (CSS), which quantifies the quality of inter-agent collaboration, and the Tool Utilization Efficacy (TUE), which evaluates the efficiency of tool use within agent workflows. We further discuss strategies for improving explainability in Agentic AI , as well as approaches to enhancing security and privacy through encryption, adversarial robustness, and regulatory compliance. The review concludes with a research roadmap for the responsible development and deployment of

Agentic AI, outlining critical directions to align emerging systems with TRiSM principles for safe, transparent, and accountable operation.

Date 2025-07
Short Title TRiSM for Agentic AI
URL http://arxiv.org/abs/2506.04133
Accessed 9/2/2025, 7:00:00 PM
Extra Citation Key: raza_trism_2025 DOI: 10.48550/arXiv.2506.04133
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence

Notes:

arXiv:2506.04133 [cs]

Trustagent: Towards safe and trustworthy llm-based agents

Item Type Conference Paper
Author Wenyue Hua
Author Xianjun Yang
Author Mingyu Jin
Author Zelong Li
Author Wei Cheng
Author Ruixiang Tang
Author Yongfeng Zhang
Date 2024
Extra Citation Key: hua2024trustagent
Pages 10000–10016
Proceedings Title Findings of the association for computational linguistics: EMNLP 2024
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Trustllm: Trustworthiness in large language models

Item Type Journal Article
Author Yue Huang
Author Lichao Sun
Author Haoran Wang
Author Siyuan Wu
Author Qihui Zhang
Author Yuan Li
Author Chujie Gao
Author Yixin Huang
Author Wenhan Lyu
Author Yixuan Zhang
Author others

Date 2024
Extra Citation Key: huang2024trustllm
Publication arXiv preprint arXiv:2401.05561
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

TrustRAG: Enhancing robustness and trustworthiness in RAG

Item Type Journal Article
Author Huichi Zhou
Author Kin-Hei Lee
Author Zhonghao Zhan
Author Yue Chen
Author Zhenhao Li
Date 2025
Extra Citation Key: zhou2025trustrag
Publication arXiv preprint arXiv:2501.00879
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Trustworthy agentic AI systems: a cross-layer review of architectures, threat models, and governance strategies for real-world deployment

Item Type Journal Article
Author Ibrahim Adabara
Author Bashir Olaniyi Sadiq
Author Aliyu Nuhu Shuaibu
Author Yale Ibrahim Danjuma
Author Venkateswarlu Maninti
Date 2025
URL https://f1000research.com/articles/14-905/pdf
Extra Citation Key: Adabara2025Trustworthy
Volume 14
Pages 905
Publication F1000Research
DOI 10.12688/f1000research.144501.1
Date Added 10/20/2025, 3:50:53 PM
Modified 10/20/2025, 3:50:53 PM

Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment

Item Type Journal Article
Author Yang Liu
Author Yuanshun Yao
Author Jean-Francois Ton
Author Xiaoying Zhang
Author Ruocheng Guo Hao Cheng
Author Yegor Klochkov

Author Muhammad Faaiz Taufiq
Author Hang Li
Date 2023
Extra Citation Key: liu2023trustworthy
Publication arXiv preprint arXiv:2308.05374
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Truth-aware context selection: Mitigating the hallucinations of large language models being misled by untruthful contexts

Item Type Journal Article
Author Tian Yu
Author Shaolei Zhang
Author Yang Feng
Date 2024
Extra Citation Key: yu2024truth
Publication arXiv preprint arXiv:2403.07556
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

TurkingBench: a challenge benchmark for web agents

Item Type Conference Paper
Author Kevin Xu
Author Yeganeh Kordi
Author Tanay Nayak
Author Adi Asija
Author Yizhong Wang
Author Kate Sanders
Author Adam Byerly
Author Jingyu Zhang
Author Benjamin Van Durme
Author Daniel Khashabi
Editor Luis Chiruzzo
Editor Alan Ritter
Editor Lu Wang
Date 2025-04
URL https://aclanthology.org/2025.naacl-long.188/
Extra Citation Key: xuetal2025turkingbench
Place Albuquerque, New Mexico
Publisher Association for Computational Linguistics
ISBN 979-8-89176-189-6
Pages 3694–3710
Proceedings Title 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)
DOI 10.18653/v1/2025.naacl-long.188
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

TwinBreak: jailbreaking LLM security alignments based on twin prompts

Item Type	Conference Paper
Author	Torsten Krauß
Author	Hamid Dashtbani
Author	Alexandra Dmitrienko
Date	2025
Extra	Citation Key: krauss2025Twinbreak Number of pages: 20 tex.address: USA tex.articleno: 121
Place	Seattle, WA, USA
Publisher	USENIX Association
ISBN	978-1-939133-52-6
Series	Sec '25
Proceedings Title	Proceedings of the 34th USENIX conference on security symposium
Date Added	10/20/2025, 3:50:52 PM
Modified	10/20/2025, 3:50:52 PM

TwinMarket: A Scalable Behavioral and Social Simulation for Financial Markets

Item Type	Document
URL	https://arxiv.org/html/2502.01506v1
Accessed	10/8/2025, 7:00:00 PM
Extra	Citation Key: noauthor_twinmarket_nodate
Date Added	10/20/2025, 3:48:27 PM
Modified	10/20/2025, 3:48:27 PM

Typos that broke the rag's back: Genetic attack on RAG pipeline by simulating documents in the wild via low-level perturbations

Item Type	Journal Article
Author	Sukmin Cho
Author	Soyeong Jeong
Author	Jeongyeon Seo
Author	Taeho Hwang
Author	Jong C Park
Date	2024
Extra	Citation Key: cho2024typos
Publication	arXiv preprint arXiv:2404.13948
Date Added	10/20/2025, 3:49:09 PM
Modified	10/20/2025, 3:49:09 PM

Ufo: A ui-focused agent for windows os interaction

Item Type	Journal Article
Author	Chaoyun Zhang
Author	Liqun Li
Author	Shilin He
Author	Xu Zhang
Author	Bo Qiao

Author Si Qin
Author Minghua Ma
Author Yu Kang
Author Qingwei Lin
Author Saravan Rajmohan
Author others
Date 2024
Extra Citation Key: zhang2024ufo
Publication arXiv preprint arXiv:2402.07939
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Uncovering the Vulnerability of Large Language Models in the Financial Domain via Risk Concealment

Item Type Document
Author Gang Cheng
Author Haibo Jin
Author Wenbin Zhang
Author Haohan Wang
Author Jun Zhuang
Abstract Large Language Models (LLMs) are increasingly integrated into financial applications, yet existing red-teaming research primarily targets harmful content, largely neglecting regulatory risks. In this work, we aim to investigate the vulnerability of financial LLMs through red-teaming approaches. We introduce Risk-Concealment Attacks (RCA), a novel multi-turn framework that iteratively conceals regulatory risks to provoke seemingly compliant yet regulatory-violating responses from LLMs. To enable systematic evaluation, we construct FIN-Bench, a domain-specific benchmark for assessing LLM safety in financial contexts. Extensive experiments on FIN-Bench demonstrate that RCA effectively bypasses nine mainstream LLMs, achieving an average attack success rate (ASR) of 93.18%, including 98.28% on GPT-4.1 and 97.56% on OpenAI o1. These findings reveal a critical gap in current alignment techniques and underscore the urgent need for stronger moderation mechanisms in financial domains. We hope this work offers practical insights for advancing robust and domain-aware LLM alignment.
Date 2025-09
URL http://arxiv.org/abs/2509.10546
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: cheng_uncovering_2025 DOI: 10.48550/arXiv.2509.10546
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Computer Science - Computation and Language, Computer Science - Machine Learning

Notes:

arXiv:2509.10546 [cs]

Uncovering vulnerabilities of LLM-assisted cyber threat intelligence

Item Type Document
Author Yuqiao Meng
Author Luoxi Tang

Author Feiyang Yu
Author Jinyuan Jia
Author Guanhua Yan
Author Ping Yang
Author Zhaohan Xi
Date 2025
URL https://arxiv.org/abs/2509.23573
Extra Citation Key: meng2025uncoveringvulnerabilitiesllmassistedcyber arXiv: 2509.23573 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Understanding multi-turn toxic behaviors in open-domain chatbots

Item Type Conference Paper
Author Bocheng Chen
Author Guangjing Wang
Author Hanqing Guo
Author Yuanda Wang
Author Qiben Yan
Date 2023
Extra Citation Key: chen2023understanding
Pages 282–296
Proceedings Title Proceedings of the 26th international symposium on research in attacks, intrusions and defenses
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Universal and transferable adversarial attacks on aligned language models

Item Type Document
Author Andy Zou
Author Zifan Wang
Author Nicholas Carlini
Author Milad Nasr
Author J. Zico Kolter
Author Matt Fredrikson
Date 2023
URL https://arxiv.org/abs/2307.15043
Extra Citation Key: zou2023universaltransferableadversarialattacks arXiv: 2307.15043 [cs.CL]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Universal litmus patterns: Revealing backdoor attacks in cnns

Item Type Conference Paper
Author Soheil Kolouri
Author Aniruddha Saha
Author Hamed Pirsiavash
Author Heiko Hoffmann

Date 2020
Extra Citation Key: kolouri2020universal
Pages 301–310
Proceedings Title Proceedings of the IEEE/CVF conference on computer vision and pattern recognition
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Unleashing cheapfakes through trojan plugins of large language models

Item Type Journal Article
Author Tian Dong
Author Guoxing Chen
Author Shaofeng Li
Author Minhui Xue
Author Rayne Holland
Author Yan Meng
Author Zhen Liu
Author Haojin Zhu
Date 2023
Extra Citation Key: dong2023unleashing
Publication arXiv preprint arXiv:2312.00374
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

Unleashing large-scale video generative pre-training for visual robot manipulation

Item Type Journal Article
Author Hongtao Wu
Author Ya Jing
Author Chilam Cheang
Author Guangzeng Chen
Author Jiafeng Xu
Author Xinghang Li
Author Minghuan Liu
Author Hang Li
Author Tao Kong
Date 2023
Extra Citation Key: wu2023unleashing
Publication arXiv preprint arXiv:2312.13139
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Unveiling privacy risks in LLM agent memory

Item Type Journal Article
Author Bo Wang
Author Weiyi He
Author Pengfei He

Author Shenglai Zeng
Author Zhen Xiang
Author Yue Xing
Author Jiliang Tang
Date 2025
Extra Citation Key: wang2025unveiling
Publication arXiv preprint arXiv:2502.13172
Date Added 10/20/2025, 3:49:10 PM
Modified 10/20/2025, 3:49:10 PM

Unveiling the truth and facilitating change: Towards agent-based large-scale social movement simulation

Item Type Journal Article
Author Xinyi Mou
Author Zhongyu Wei
Author Xuanjing Huang
Date 2024
Extra Citation Key: mou2024unveiling
Publication arXiv preprint arXiv:2402.16333
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

VeriPlan: Integrating formal verification and llms into end-user planning

Item Type Conference Paper
Author Christine P. Lee
Author David Porfirio
Author Xinyu Jessica Wang
Author Kevin Chenkai Zhao
Author Bilge Mutlu
Abstract Automated planning is traditionally the domain of experts, utilized in fields like manufacturing and healthcare with the aid of expert planning tools. Recent advancements in LLMs have made planning more accessible to everyday users due to their potential to assist users with complex planning tasks. However, LLMs face several application challenges within end-user planning, including consistency, accuracy, and user trust issues. This paper introduces VeriPlan, a system that applies formal verification techniques, specifically model checking, to enhance the reliability and flexibility of LLMs for end-user planning. In addition to the LLM planner, VeriPlan includes three additional core features—a rule translator, flexibility sliders, and a model checker—that engage users in the verification process. Through a user study (n = 12), we evaluate VeriPlan, demonstrating improvements in the perceived quality, usability, and user satisfaction of LLMs. Our work shows the effective integration of formal verification and user-control features with LLMs for end-user planning tasks.
Date 2025
URL https://doi.org/10.1145/3706598.3714113
Extra Citation Key: 10.1145/3706598.3714113 Number of pages: 19 tex.articleno: 247
Place New York, NY, USA
Publisher Association for Computing Machinery
ISBN 979-8-4007-1394-1
Series Chi '25
Proceedings Title Proceedings of the 2025 CHI conference on human factors in computing systems
DOI 10.1145/3706598.3714113
Date Added 10/20/2025, 3:50:53 PM

Modified 10/20/2025, 3:50:53 PM

Tags:

human-centered AI, human-in-the-loop, large-language models, verification

---

Voyager: An open-ended embodied agent with large language models

Item Type Journal Article
Author Guanzhi Wang
Author Yining Ren
Author Jianren Chen
Author others
Date 2023
Extra Citation Key: wang2023voyager
Publication arXiv preprint arXiv:2305.16291
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

VulnBot: Autonomous penetration testing for a multi-agent collaborative framework

Item Type Document
Author He Kong
Author Die Hu
Author Jingguo Ge
Author Liangxiong Li
Author Tong Li
Author Bingzhen Wu
Date 2025
URL https://arxiv.org/abs/2501.13411
Extra Citation Key: kong2025vulnbotautonomouspenetrationtesting arXiv: 2501.13411 [cs.SE]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Watch out for your agents! investigating backdoor threats to llm-based agents

Item Type Journal Article
Author Wenkai Yang
Author Xiaohan Bi
Author Yankai Lin
Author Sishuo Chen
Author Jie Zhou
Author Xu Sun
Date 2024
Extra Citation Key: yang2024watch
Publication arXiv preprint arXiv:2402.11208
Date Added 10/20/2025, 3:49:08 PM
Modified 10/20/2025, 3:49:08 PM

WebArena: a realistic web environment for building autonomous agents

Item Type	Document
Author	Shuyan Zhou
Author	Frank F. Xu
Author	Hao Zhu
Author	Xuhui Zhou
Author	Robert Lo
Author	Abishek Sridhar
Author	Xianyi Cheng
Author	Tianyue Ou
Author	Yonatan Bisk
Author	Daniel Fried
Author	Uri Alon
Author	Graham Neubig
Date	2024
URL	https://arxiv.org/abs/2307.13854
Extra	Citation Key: zhou2024webarenarealisticwebenvironment arXiv: 2307.13854 [cs.AI]
Date Added	10/20/2025, 3:50:53 PM
Modified	10/20/2025, 3:50:53 PM

When AI Meets Finance (StockAgent): Large Language Model-based Stock Trading in Simulated Real-world Environments

Item Type	Document
Author	Chong Zhang
Author	Xinyi Liu
Author	Zhongmou Zhang
Author	Mingyu Jin
Author	Lingyao Li
Author	Zhenting Wang
Author	Wenyue Hua
Author	Dong Shu
Author	Suiyuan Zhu
Author	Xiaobo Jin
Author	Sujian Li
Author	Mengnan Du
Author	Yongfeng Zhang
Abstract	Can AI Agents simulate real-world trading environments to investigate the impact of external factors on stock trading activities (e.g., macroeconomics, policy changes, company fundamentals, and global events)? These factors, which frequently influence trading behaviors, are critical elements in the quest for maximizing investors' profits. Our work attempts to solve this problem through large language model based agents. We have developed a multi-agent AI system called StockAgent, driven by LLMs, designed to simulate investors' trading behaviors in response to the real stock market. The StockAgent allows users to evaluate the impact of different external factors on investor trading and to analyze trading behavior and profitability effects. Additionally, StockAgent avoids the test set leakage issue present in existing trading simulation systems based on AI Agents. Specifically, it prevents the model from leveraging prior knowledge it may have acquired related to the test data. We evaluate different LLMs under the framework of StockAgent in a stock trading environment that closely resembles real-world conditions. The experimental results demonstrate the impact of key external factors on stock market trading, including trading behavior and stock price fluctuation rules. This research explores the study of agents' free trading gaps in the context of no prior knowledge related to market data. The patterns identified through StockAgent simulations provide valuable insights for LLM-based investment advice and stock recommendation.

The code is available at https://github.com/MingyuJ666/Stockagent.

Date 2024-09
Short Title When AI Meets Finance (StockAgent)
URL http://arxiv.org/abs/2407.18957
Accessed 10/8/2025, 7:00:00 PM
Extra Citation Key: zhang_when_2024 DOI: 10.48550/arXiv.2407.18957
Publisher arXiv
Date Added 10/20/2025, 3:48:27 PM
Modified 10/20/2025, 3:48:27 PM

Tags:

Computer Science - Artificial Intelligence, Quantitative Finance - Trading and Market Microstructure, Computer Science - Multiagent Systems

Notes:

arXiv:2407.18957 [q-fin]

When large language models contradict humans? Large language models' sycophantic behaviour

Item Type Journal Article
Author Leonardo Ranaldi
Author Giulia Pucci
Date 2023
Extra Citation Key: ranaldi2023large
Publication arXiv preprint arXiv:2311.09410
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

When llms go online: The emerging threat of web-enabled llms

Item Type Journal Article
Author Hanna Kim
Author Minkyoo Song
Author Seung Ho Na
Author Seungwon Shin
Author Kimin Lee
Date 2024
Extra Citation Key: kim2024llms
Publication arXiv preprint arXiv:2410.14569
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

Why are web AI agents more vulnerable than standalone llms? A security analysis

Item Type Conference Paper
Author Jeffrey Yang Fan Chiang
Author Seungjae Lee

Author Jia-Bin Huang
Author Furong Huang
Author Yizheng Chen
Date 2025
URL https://openreview.net/forum?id=4KoMbO2RJ9
Extra Citation Key: chiang2025harmful
Proceedings Title ICLR 2025 workshop on building trust in language models and applications
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Why do multiagent systems fail?

Item Type Conference Paper
Author Melissa Z Pan
Author Mert Cemri
Author Lakshya A Agrawal
Author Shuyi Yang
Author Bhavya Chopra
Author Rishabh Tiwari
Author Kurt Keutzer
Author Aditya Parameswaran
Author Kannan Ramchandran
Author Dan Klein
Author Joseph E. Gonzalez
Author Matei Zaharia
Author Ion Stoica
Date 2025
URL https://openreview.net/forum?id=wM521FqPvI
Extra Citation Key: pan2025why
Proceedings Title ICLR 2025 workshop on building trust in language models and applications
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

Wipi: A new web threat for llm-driven web agents

Item Type Journal Article
Author Fangzhou Wu
Author Shutong Wu
Author Yulong Cao
Author Chaowei Xiao
Date 2024
Extra Citation Key: wu2024wipi
Publication arXiv preprint arXiv:2402.16965
Date Added 10/20/2025, 3:49:09 PM
Modified 10/20/2025, 3:49:09 PM

---

xOffense: An AI-driven autonomous penetration testing framework with offensive knowledge-enhanced LLMs and multi agent systems

Item Type Document
Author Phung Duc Luong
Author Le Tran Gia Bao
Author Nguyen Vu Khai Tam
Author Dong Huu Nguyen Khoa
Author Nguyen Huu Quyen
Author Van-Hau Pham
Author Phan The Duy
Date 2025
URL https://arxiv.org/abs/2509.13021
Extra Citation Key: luong2025xoffenseaidrivenautonomouspenetration arXiv: 2509.13021 [cs.CR]
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM

τ-bench: a benchmark for tool-agent-user interaction in real-world domains

Item Type Document
Author Shunyu Yao
Author Noah Shinn
Author Pedram Razavi
Author Karthik Narasimhan
Abstract Existing benchmarks do not test language agents on their interaction with human users or ability to follow domain-specific rules, both of which are vital for deploying them in real-world applications. We propose tau-bench, a benchmark emulating dynamic conversations between a user (simulated by language models) and a language agent provided with domain-specific API tools and policy guidelines. We employ an efficient and faithful evaluation process that compares the database state at the end of a conversation with the annotated goal state. We also propose a new metric (pass^k) to evaluate the reliability of agent behavior over multiple trials. Our experiments show that even state-of-the-art function-calling agents (like gpt-4o) succeed on fewer than 50
Date 2024
URL https://arxiv.org/abs/2406.12045
Extra Citation Key: yao2024taubench
Date Added 10/20/2025, 3:50:52 PM
Modified 10/20/2025, 3:50:52 PM