

40th AAAI Conference on Artificial Intelligence (AAAI-26), 38th Conference on Innovative Applications of Artificial Intelligence (IAAI-26), and 16th Symposium on Educational Advances in Artificial Intelligence (EAAI-26)

Volume 42: AAAI Technical Tracks

- Philosophy and Ethics of AI

Singapore
20-27 January 2026

ISBN: 979-8-3313-3416-1

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2026) by Association for the Advancement of Artificial Intelligence
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

For permission requests, please contact Association for the Advancement of Artificial Intelligence
at the address below.

Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road
Suite 160
Palo Alto, California 94303
USA

Phone: 1-650-328-3123
Fax: 1-650-321-4457

<https://aaai.org/Press/press.php>

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

AAAI TECHNICAL TRACK ON PHILOSOPHY AND ETHICS OF AI

Privacy Preserving In-Context-Learning Framework for Large Language Models	35303
<i>Bishnu Bhusal, Manoj Acharya, Ramneet Kaur, Colin Samplawski, Anirban Roy, Adam D. Cobb, Rohit Chadha, Susmit Jha</i>	
Your Prompts Are Not Safe: Output-Free Membership Inference Via Prompt Vectors in Vision-Language Tuning	35313
<i>Yuran Bian, Xiaohan Zhang, Zhiyuan Yu, Changqing Li, Li Pan</i>	
When Safe Unimodal Inputs Collide: Optimizing Reasoning Chains for Cross-Modal Safety in Multimodal Large Language Models	35322
<i>Wei Cai, Shujuan Liu, Jian Zhao, Ziyang Shi, Yusheng Zhao, Yuchen Yuan, Tianle Zhang, Chi Zhang, Xuelong Li</i>	
Failures to Surface Harmful Contents in Video Large Language Models	35331
<i>Yuxin Cao, Wei Song, Derui Wang, Jingling Xue, Jin Song Dong</i>	
MOBA: A Material-Oriented Backdoor Attack Against LiDAR-Based 3D Object Detection Systems.....	35340
<i>Saket Sanjeev Chaturvedi, Gaurav Bagwe, Lan Emily Zhang, Pan He, Xiaoyong Yuan</i>	
StyleSentinel: Reliable Artistic Copyright Verification Via Stylistic Fingerprints	35348
<i>Lingxiao Chen, Liqin Wang, Wei Lu</i>	
Amplifying Discrepancies: Exploiting Macro and Micro Inconsistencies for Image Manipulation Localization	35357
<i>Shenghao Chen, Yibo Zhao, Tianyi Wang, Chunjie Ma, Weili Guan, Ming Li, Zan Gao</i>	
ALTER: Asymmetric LoRA for Token-Entropy-Guided Unlearning of LLMs.....	35366
<i>Xunlei Chen, Jinyu Guo, Yuang Li, Zhaokun Wang, Yi Gong, Jie Zou, Jiwei Wei, Wenhong Tian</i>	
Reference Recommendation Based Membership Inference Attack Against Hybrid-Based Recommender Systems	35375
<i>Xiaoxiao Chi, Xuyun Zhang, Yan Wang, Hongsheng Hu, Wanchun Dou</i>	
Feature Integration Spaces: Joint Training Reveals Dual Encoding in Neural Network Representations	35384
<i>Omar Clafin</i>	
Fairness Perceptions of Large Language Models	35393
<i>Benjamin Cookson, Soroush Ebadian, Nisarg Shah</i>	
Time Shuffle: A Transferability-Booster for Multiple Audio Adversarial Tasks	35402
<i>Jiacheng Deng, Dengpan Ye, Yuhong Liu, Zhaolin Wei, Ziyi Liu, Haoran Duan</i>	
Stability-Aware Reinforcement Learning for Robust Class Integration Test Order Generation.....	35411
<i>Yanru Ding, Yanmei Zhang, Guan Yuan, Shujuan Jiang, Wei Dai, Luciano Baresi</i>	

One for All: Synthesis-Free Fingerprint Learning for Attribution of In-The-Wild Synthetic Images	35419
<i>Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhihua Xia, Dasara Shullani, Daniele Baracchi, Alessandro Piva</i>	
Efficient, Secure, Differentially Private Deep Learning in the Two-Server Model.....	35428
<i>Jun Feng, Hong Sun, Pengfei Zhang, Bocheng Ren, Shunli Zhang</i>	
Runtime Safety and Reach-Avoid Prediction of Stochastic Systems Via Observation-Aware Barrier Functions	35437
<i>Shenghua Feng, Jie An, Fanjiang Xu</i>	
RSA-CR: Resisting Shilling Attacks in Citation Recommendation Via Dumbbell Inductive Learning	35446
<i>Xiyue Gao, Yukai Liu, Zhuoqi Ma, Xiaotian Qiao, Hui Li, Cai Xu, Kunhua Zhang, Jiangtao Cui</i>	
6DAttack: Backdoor Attacks in the 6DoF Pose Estimation	35455
<i>Jihui Guo, Zongmin Zhang, Zhen Sun, Yuhao Yang, Jinlin Wu, Fu Zhang, Xinlei He</i>	
The Silent Amplifier: In-Context Examples Fuel Bias in Large Language Models	35464
<i>Xinwei Guo, Jiashi Gao, Junlei Zhou, Jiaxin Zhang, Quanying Liu, Haiyan Wu, Xin Yao, Xuetao Wei</i>	
Beyond World Models: Rethinking Understanding in AI Models.....	35473
<i>Tarun Gupta, Danish Pruthi</i>	
Activation Manipulation Attack: Penetrating and Harmful Jailbreak Attack Against Large Vision-Language Models	35481
<i>Haojie Hao, Jiakai Wang, Aishan Liu, Yuqing Ma, Haotong Qin, Yuanfang Guo, Xianglong Liu</i>	
FILTER: A Framework for Defending Against Backdoor Attacks in Vertical Federated Learning	35490
<i>Zhanyi Hu, Cen Chen, Yanhao Wang</i>	
Parameterized Abstract Interpretation for Transformer Verification	35500
<i>Pei Huang, Dennis Wei, Omri Isac, Haoze Wu, Min Wu, Clark Barrett</i>	
Any2Critical: Safety-Critical Scenario Generation from Arbitrary Real-World Driving Contexts	35509
<i>Yao Huang, Yubo Chen, Ruochen Zhang, Yitong Sun, Shouwei Ruan, Zhenyu Wu, Yinpeng Dong, Xingxing Wei</i>	
Private Frequency Estimation Via Residue Number Systems	35518
<i>Héber Hwang Arcolezi</i>	
Efficient LLM-Jailbreaking Via Multimodal-LLM Jailbreak.....	35527
<i>Haoxuan Ji, Zheng Lin, Zhenxing Niu, Xinbo Gao, Gang Hua</i>	
MedOmni-45°: A Safety-Performance Benchmark for Reasoning-Oriented LLMs in Medicine.....	35536
<i>Kaiyuan Ji, Yijin Guo, Zicheng Zhang, Xiangyang Zhu, Yuan Tian, Ning Liu</i>	
Higher-Order Responsibility	35545
<i>Junli Jiang, Pavel Naumov</i>	
SceneJailEval: A Scenario-Adaptive Multi-Dimensional Framework for Jailbreak Evaluation	35553
<i>Lai Jiang, Yuekang Li, Xiaohan Zhang, Youtao Ding, Li Pan</i>	

PurMM: Attention-Guided Test-Time Backdoor Purification in Multimodal Large Language Models.....	35562
<i>Wenzheng Jiang, Ke Liang, Xuankun Rong, Jingxuan Zhou, Zhengyi Zhong, Guancheng Wan, Ji Wang</i>	
DySy-Det: A Synergistic Framework with Dynamic Reconstruction-Path Consistency for AI-Generated Image Detection	35571
<i>Fanli Jin, Feng Lin, Gaojian Wang, Tong Wu, Zhisheng Yan</i>	
Machine Pareidolia: Protecting Facial Image with Emotional Editing.....	35580
<i>Binh M. Le, Simon S. Woo</i>	
Cross-Modal Unlearning Via Influential Neuron Path Editing in Multimodal Large Language Models.....	35589
<i>Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, Jason Xue</i>	
The Other Mind: How Language Models Exhibit Human Temporal Cognition	35598
<i>Lingyu Li, Yang Yao, Yixu Wang, Chunbo Li, Yan Teng, Yingchun Wang</i>	
Model-Agnostic Sentiment Distribution Stability Analysis for Robust LLM-Generated Texts Detection	35608
<i>Siyan Li, Xi Lin, Guangyan Li, Zehao Liu, Aodu Wulianghai, Li Ding, Jun Wu, Jianhua Li</i>	
SAVER: Mitigating Hallucinations in Large Vision-Language Models Via Style-Aware Visual Early Revision.....	35617
<i>Zhaoxu Li, Chenqi Kong, Yi Yu, Qiangqiang Wu, Xinghao Jiang, Ngai-Man Cheung, Bihan Wen, Alex Kot, Xudong Jiang</i>	
Mind the Third Eye! Benchmarking Privacy Awareness in MLLM-Powered Smartphone Agents.....	35626
<i>Zhixin Lin, Jungang Li, Shidong Pan, Yibo Shi, Yue Yao, Dongliang Xu</i>	
EchoBat: Echo-Vision Enhancement and Echo-Layered Sampling for Video LLMs Hallucination Mitigation	35635
<i>Shuai Liu, Da Chen, Yiheng Pan, Chenwei Tian, Qian Li, Chenhao Lin</i>	
Eguard: Defending LLM Embeddings Against Inversion Attacks Via Text Mutual Information Optimization.....	35644
<i>Tiantian Liu, Hongwei Yao, Feng Lin, Tong Wu, Zhan Qin, Kui Ren</i>	
GeoShield: Safeguarding Geolocation Privacy from Vision-Language Models Via Adversarial Perturbations.....	35653
<i>Xinwei Liu, Xiaojun Jia, Yuan Xun, Simeng Qin, Xiaochun Cao</i>	
Generic Adversarial Attack Framework Against Graph-Based Vertical Federated Learning	35662
<i>Yimin Liu, Peng Jiang, Qi Liu, Liehuang Zhu</i>	
Faithful in Steps: Improving Generalization and Citation in RAG Via Query Decomposition.....	35671
<i>Yue Liu, Zhongying Ru, Shimin Di, Jipeng Zhang, Ruiyuan Zhang, Xiaofang Zhou</i>	
IS-Bench: Evaluating Interactive Safety of VLM-Driven Embodied Agents in Daily Household Tasks.....	35680
<i>Xiaoya Lu, Zeren Chen, Xuhao Hu, Yijin Zhou, Weichen Zhang, Dongrui Liu, Lu Sheng, Jing Shao</i>	

Phantom Menace: Exploring and Enhancing the Robustness of VLA Models Against Physical Sensor Attacks	35689
<i>Xuancun Lu, Jiaxiang Chen, Shilin Xiao, Zizhi Jin, Zhangrui Chen, Hanwen Yu, Bohan Qian, Ruochen Zhou, Xiaoyu Ji, Wenyuan Xu</i>	
Learning Vision-Based Neural Network Controllers with Semi-Probabilistic Safety Guarantees	35698
<i>Xinhang Ma, Junlin Wu, Hussein Sibai, Yiannis Kantaros, Yevgeniy Vorobeychik</i>	
On the Probabilistic Learnability of Compact Neural Network Preimage Bounds	35707
<i>Luca Marzari, Manuele Bicego, Ferdinando Cicalese, Alessandro Farinelli</i>	
SPAN: Benchmarking and Improving Cross-Calendar Temporal Reasoning of Large Language Models	35715
<i>Zhongjian Miao, Hao Fu, Chen Wei</i>	
An Epistemic Perspective on Agent Awareness.....	35724
<i>Pavel Naumov, Alexandra Pavlova</i>	
Beyond Binary Classification: A Semi-Supervised Approach to Generalized AI-Generated Image Detection	35733
<i>Hong-Hanh Nguyen-Le, Van-Tuan Tran, Thuc D. Nguyen, Nhien-An Le-Khac</i>	
Dynamic Deep Prompt Optimization for Defending Against Jailbreak Attacks on LLMs.....	35742
<i>Doniyorkhon Obidov, Honggang Yu, Xiaolong Guo, Kaichen Yang</i>	
Estimating the True Distribution of Data Collected with Randomized Response	35751
<i>Carlos Antonio Pinzón, Ehab Elsalamouny, Lucas Massot, Alexis Miller, Héber Hwang Arcolezi, Catuscia Palamidessi</i>	
MartDE: A Privacy-Preserving and Cost-Efficient Evaluation Framework for Data Marketplaces.....	35759
<i>Xinyuan Qian, Haoyong Wang, Hangcheng Cao, Shuai Yuan, Senkang Hu, Qingchuan Zhao, Hongwei Li, Guowen Xu</i>	
Efficient Verification and Falsification of ReLU Neural Barrier Certificates	35767
<i>Dejin Ren, Yiling Xue, Taoran Wu, Bai Xue</i>	
Probing Semantic Insensitivity for Inference-Time Backdoor Defense in Multimodal Large Language Model.....	35775
<i>Xuankun Rong, Wenke Huang, Wenzheng Jiang, Yiming Li, Wenxuan Wang, Mang Ye</i>	
RAIN: Redundancy-Aware Latent Injection for Quality-Preserving Image Watermarking	35784
<i>Yehan Sun, Rongrong Ni, Chuangchuang Tan, Huan Liu, Wenhao Ni, Renshuai Tao, Yao Zhao</i>	
Joint-GCG: Unified Gradient-Based Poisoning Attacks on Retrieval-Augmented Generation Systems.....	35793
<i>Haowei Wang, Rupeng Zhang, Junjie Wang, Mingyang Li, Yuekai Huang, Dandan Wang, Qing Wang</i>	
Enhancing All-To-X Backdoor Attacks with Optimized Target Class Mapping	35802
<i>Lei Wang, Yulong Tian, Hao Han, Fengyuan Xu</i>	
MCPTox: A Benchmark for Tool Poisoning on Real-World MCP Servers	35811
<i>Zhiqiang Wang, Yichao Gao, Yanting Wang, Suyuan Liu, Haifeng Sun, Haoran Cheng, Guanquan Shi, Haohua Du, Xiangyang Li</i>	
GUIC: Certified Graph Unlearning with Individual Fairness Guarantees.....	35820
<i>Zichong Wang, Tongliang Liu, Wenbin Zhang</i>	

ConfGuard: A Simple and Effective Backdoor Detection for Large Language Models.....	35829
<i>Zihan Wang, Rui Zhang, Hongwei Li, Wenshu Fan, Wenbo Jiang, Qingchuan Zhao, Guowen Xu</i>	
MPMA: Preference Manipulation Attack Against Model Context Protocol	35838
<i>Zihan Wang, Rui Zhang, Yu Liu, Wenshu Fan, Wenbo Jiang, Qingchuan Zhao, Hongwei Li, Guowen Xu</i>	
Robust Learning from Noisily Labeled Long-Tailed Data Via Fairness Regularizer.....	35847
<i>Jiaheng Wei, Zhaowei Zhu, Gang Niu, Tongliang Liu, Sijia Liu, Masashi Sugiyama, Yang Liu</i>	
Efficiently Computing Compact Formal Explanations.....	35857
<i>Min Wu, Xiaofu Li, Haoze Wu, Clark Barrett</i>	
ARIW-Framework: Adaptive Robust Iterative Watermarking Framework	35867
<i>Shaowu Wu, Liting Zeng, Wei Lu</i>	
BeDKD: Backdoor Defense Based on Directional Mapping Module and Adversarial Knowledge Distillation	35876
<i>Zhengxian Wu, Juan Wen, Wanli Peng, Yinghan Zhou, Changtong Dou, Yiming Xue</i>	
A Content-Preserving Secure Linguistic Steganography.....	35885
<i>Lingyun Xiang, Chengfu Ou, Xu He, Zhongliang Yang, Yuling Liu</i>	
CL-Guard: Defending DNNs Against Backdoors Via Fine-Grained Neuron Analysis and Collaborative Dual-Network Learning	35894
<i>Jie Xiao, Yuhao Huang, Yanjiao Gao, Aizhu Liu, Zhezhaoyang, Xinyue Yu, Qianwei Zhou, Fan Terry Zhang</i>	
Class-Feature Watermark: A Resilient Black-Box Watermark Against Model Extraction Attacks	35903
<i>Yaxin Xiao, Qingqing Ye, Zi Liang, Haoyang Li, Ronghua Li, Huadi Zheng, Haibo Hu</i>	
LexChain: Modeling Legal Reasoning Chains for Chinese Tort Case Analysis.....	35913
<i>Huiyuan Xie, Chenyang Li, Huining Zhu, Chubin Zhang, Yuxiao Ye, Zhenghao Liu, Zhiyuan Liu</i>	
Detect All-Type Deepfake Audio: Wavelet Prompt Tuning for Enhanced Auditory Perception	35922
<i>Yuankun Xie, Ruibo Fu, Xiaopeng Wang, Zhiyong Wang, Songjun Cao, Long Ma, Haonan Cheng, Long Ye</i>	
HealSplit: Towards Self-Healing Through Adversarial Distillation in Split Federated Learning.....	35931
<i>Yuhan Xie, Chen Lyu</i>	
ISeal: Encrypted Fingerprinting for Reliable LLM Ownership Verification	35940
<i>Zixun Xiong, Gaoyi Wu, Qingyang Yu, Mingyu Derek Ma, Lingfeng Yao, Miao Pan, Xiaojiang Du, Hao Wang</i>	
Bridging the Copyright Gap: Do Large Vision-Language Models Recognize and Respect Copyrighted Content?.....	35949
<i>Naen Xu, Jinghuai Zhang, Changjiang Li, Hengyu An, Chunyi Zhou, Jun Wang, Boyu Xu, Yuyuan Li, Tianyu Du, Shouling Ji</i>	
When Privacy Meets Recovery: The Overlooked Half of Surrogate-Driven Privacy Preservation for MLLM Editing	35958
<i>Siyan Xu, Yibing Liu, Peilin Chen, Yung-Hui Li, Shiqi Wang, Sam Kwong</i>	

Privacy Leaks by Adversaries: Adversarial Iterations for Membership Inference Attack	35967
<i>Jing Xue, Zhishen Sun, Haishan Ye, Luo Luo, Xiangyu Chang, Guang Dai</i>	
The Emotional Baby is Truly Deadly: Does Your Multimodal Large Reasoning Model Have Emotional Flattery Towards Humans?	35976
<i>Yuan Xun, Xiaojun Jia, Xinwei Liu, Simeng Qin, Hua Zhang</i>	
BLM-Guard: Explainable Multimodal Ad Moderation with Chain-Of-Thought and Policy-Aligned Rewards	35985
<i>Yiran Yang, Zhaowei Liu, Yuan Yuan, Yukun Song, Xiong Ma, Yinghao Song, Xiangji Zeng, Lu Sun, Yulu Wang, Hai Zhou, Shuai Cui, Zhaohan Gong, Jiefei Zhang</i>	
Hashed Watermark as a Filter: A Unified Defense Against Forging and Overwriting Attacks in Neural Network Watermarking.....	35994
<i>Yuan Yao, Jin Song, Jian Jin</i>	
MacPrompt: Maraconic-Guided Jailbreak Against Text-To-Image Models	36003
<i>Xi Ye, Yiwen Liu, Lina Wang, Run Wang, Geying Yang, Yufei Hou, Jiayi Yu</i>	
SafeR-CLIP: Mitigating NSFW Content in Vision-Language Models While Preserving Pre-Trained Knowledge.....	36012
<i>Adeel Yousaf, Joseph Fiorese, James Beetham, Amrit Singh Bedi, Mubarak Shah</i>	
Causally-Grounded Dual-Path Attention Intervention for Object Hallucination Mitigation in LVLMS	36021
<i>Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, Gillian Dobbie</i>	
Reason2Attack: Jailbreaking Text-To-Image Models Via LLM Reasoning	36030
<i>Chenyu Zhang, Lanjun Wang, Yiwen Ma, Wenhui Li, Guoqing Jin, Anan Liu</i>	
T2I-RiskyPrompt: A Benchmark for Safety Evaluation, Attack, and Defense on Text-To-Image Model	36039
<i>Chenyu Zhang, Tairen Zhang, Lanjun Wang, Ruidong Chen, Wenhui Li, Anan Liu</i>	
MSAT-LDM: Toward Transferable High-Fidelity Watermarking for Latent Diffusion Model Via Modular Self-Augmented Training	36048
<i>Lu Zhang, Liang Zeng</i>	
Decomposing the Neurons: Activation Sparsity Via Mixture of Experts for Continual Test Time Adaptation	36057
<i>Rongyu Zhang, Aosong Cheng, Yulin Luo, Gaole Dai, Huanrui Yang, Jiaming Liu, Ran Xu, Li Du, Dan Wang, Yuan Du</i>	
Towards Provably Unlearnable Examples Via Bayes Error Optimization	36066
<i>Ruihan Zhang, Jun Sun, Ee-Peng Lim, Peixin Zhang</i>	
Rep Deep & Machine Learning: Exemplar-Free Continual Video Action Recognition Via Slow-Fast Collaborative Learning.....	36075
<i>Xueyi Zhang, Chengwei Zhang, Zheng Li, Xiyu Wang, Siqi Cai, Mingrui Lao, Yanming Guo, Huiping Zhuang</i>	
Consensus Learning with Multi-Party Perturbation Triggers for Secure Model Access	36084
<i>Yizhun Zhang, Jie Huang, Zeping Zhang, Shuashuai Zhang, Changhao Ding, Xuan Chen</i>	
HalluClean: A Unified Framework to Combat Hallucinations in LLMs	36092
<i>Yaxin Zhao, Yu Zhang</i>	

Beyond Semantic Features: Pixel-Level Mapping for Generalized AI-Generated Image Detection..... 36101
Chenming Zhou, Jiaan Wang, Yu Li, Lei Li, Juan Cao, Sheng Tang

DeformTrace: A Deformable State Space Model with Relay Tokens for Temporal Forgery
Localization36110
*Xiaodong Zhu, Suting Wang, Yuanming Zheng, Junqi Yang, Yangxu Liao, Yuhong Yang,
Weiping Tu, Zhongyuan Wang*

Author Index