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 44: AAAI Special Track

- AI Alignment

Singapore
20-27 January 2026

ISBN: 979-8-3313-3418-5

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.

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 SPECIAL TRACK ON AI ALIGNMENT

AURA: Affordance-Understanding and Risk-Aware Alignment Technique for Large Language Models 37204
Sayantan Adak, Pratyush Chatterjee, Somnath Banerjee, Rima Hazra, Somak Aditya, Animesh Mukherjee

Beyond Patches: Mining Interpretable Part-Prototypes for Explainable AI 37213
Mahdi Alehdaghi, Rajarshi Bhattacharya, Pourya Shamsolmoali, Rafael M. O. Cruz, Eric Granger

Operationalizing Pluralistic Values in Large Language Model Alignment Reveals Trade-Offs in Safety, Inclusivity, and Model Behavior 37222
Dalia Ali, Dora Zhao, Allison Koenecke, Orestis Papakyriakopoulos

MoralReason: Generalizable Moral Decision Alignment for LLM Agents Using Reasoning-Level Reinforcement Learning 37232
Zhiyu An, Wan Du

DNR Bench: Benchmarking Over-Reasoning in Reasoning LLMs 37240
Oluwanifemi Bamgbose, Masoud Hashemi, Sathwik Tejaswi Madhusudhan, Jishnu Sethumadhavan Nair, Aman Tiwari, Vikas Yadav

ALPHA: Action-Based Learning for Pluralistic Human Alignment in Large Language Models 37249
Aanisha Bhattacharyya, Susmit Agrawal, Yaman Kumar Singla, Tarun Ram Menta, Nikitha Sr, Rajiv Ratn Shah, Changyou Chen, Balaji Krishnamurthy

A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs 37259
Trenton Chang, Tobias Schnabel, Adith Swaminathan, Jenna Wiens

MetaCipher: A Time-Persistent and Universal Multi-Agent Framework for Cipher-Based Jailbreak Attacks for LLMs 37268
Boyuan Chen, Minghao Shao, Abdul Basit, Siddharth Garg, Muhammad Shafique

On the Exponential Convergence for Offline RLHF with Pairwise Comparisons 37277
Zhirui Chen, Vincent Y. F. Tan

Preference Optimization Via Contrastive Divergence: Your Policy is Secretly an NLL Estimator 37286
Zhuotong Chen, Fang Liu, Xuan Zhu, Haozhu Wang, Jiayu Li, Yanjun Qi, Mohammad Ghavamzadeh

Deep Hidden Cognition Facilitates Reliable Chain-Of-Thought Reasoning 37295
Zijun Chen, Wenbo Hu, Richang Hong

MegaCoin: Enhancing Medium-Grained Color Perception for Vision-Language Models 37305
Ming-Chang Chiu, Shicheng Wen, Pin-Yu Chen, Xuezhe Ma

TWINFUZZ: Dual-Model Fuzzing for Robustness Generalization in Deep Learning 37314
Enze Dai, Wentao Mo, Kun Hu, Xiaogang Zhu, Xi Xiao, Sheng Wen, Shaohua Wang, Yang Xiang

A Multi-Agent Conversational Bandit Approach to Online Evaluation and Selection of User-Aligned LLM Responses 37323
Xiangxiang Dai, Yuejin Xie, Maoli Liu, Xuchuang Wang, Zhuohua Li, Huanyu Wang, John C. S. Lui

Resilience in Ambient Multi-Agent LLMs Via Decentralized Bio-Autonomic Control and Immune-Inspired Anomaly Detection 37332
Nastaran Darabi, Devashri Naik, Sina Tayebati, Dinithi Jayasuriya, Amit Ranjan Trivedi

AMaPO: Adaptive Margin-Attached Preference Optimization for Language Model Alignment 37341
Ruibo Deng, Duanyu Feng, Wenqiang Lei

Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy 37350
Alexander Duffy, Samuel J Paech, Ishana Shastri, Elizabeth Karpinski, Baptiste Alloui-Cros, Tyler Marques, Matthew Lyle Olson

ACID Test: A Benchmark for Cultural Safety and Alignment in LALMs 37360
Bikash Dutta, Adit Jain, Rishabh Ranjan, Mayank Vatsa, Richa Singh

Aligning Attention with Human Rationales for Self-Explaining Hate Speech Detection 37369
Brage Eilertsen, Røskva Bjørgfinsdóttir, Francielle Vargas, Ali Ramezani-Kebrya

The Alignment Game: A Theory of Long-Horizon Alignment Through Recursive Curation 37379
Ali Falahati, Mohammad Mohammadi Amiri, Kate Larson, Lukasz Golab

SMiLE: Provably Enforcing Global Relational Properties in Neural Networks 37387
Matteo Francobaldi, Michele Lombardi, Andrea Lodi

EssayBench: Evaluating Large Language Models in Multi-Genre Chinese Essay Writing 37396
Fan Gao, Dongyuan Li, Ding Xia, Fei Mi, Yasheng Wang, Lifeng Shang, Baojun Wang

Beyond Transcription: Mechanistic Interpretability in ASR 37407
Neta Glazer, Yael Segal-Feldman, Hilit Segev, Aviv Shamsian, Asaf Buchnick, Gill Hetz, Ethan Fetaya, Joseph Keshet, Aviv Navon

AlignTree: Efficient Defense Against LLM Jailbreak Attacks 37417
Gil Goren, Shahar Katz, Lior Wolf

Identifying Features Associated with Bias Against 93 Stigmatized Groups in Language Models and Guardrail Model Safety Mitigation 37426
Anna-Maria Gueorguieva, Aylin Caliskan

Resolving Predictive Multiplicity for the Rashomon Set 37435
Parian Haghighat, Hadis Anahideh, Cynthia Rudin

Unintended Misalignment from Agentic Fine-Tuning: Risks and Mitigation 37443
Dongyoon Hahm, Taywon Min, Woogyeol Jin, Kimin Lee

Silenced Biases: The Dark Side LLMs Learned to Refuse 37452
Rom Himelstein, Amit Levi, Brit Youngmann, Yaniv Nemcovsky, Avi Mendelson

TAPO: Dynamic Teacher and Perturbed Answer Injection for Policy Optimization 37462
Maowei Jiang, Zihang Wang, Qi Wang, Peter Búš, Moquan Cheng, Yifan Wang, Quangao Liu, Ruiqi Li, Pengyu Zeng, Ruikai Liu, Alan Liang, Yansong Xu, Yusong Hu, Chaoran Zhang, Zhiyong Dong

Uncovering and Aligning Anomalous Attention Heads to Defend Against NLP Backdoor Attacks 37472
Haotian Jin, Yang Li, Haihui Fan, Lin Shen, Xiangfang Li, Bo Li

Requirements for Aligned, Dynamic Resolution of Conflicts in Operational Constraints 37481
Steven J. Jones, Robert E. Wray, John E. Laird

Benchmarking XAI Explanations with Human-Aligned Evaluations 37491
Rémi Kazmierczak, Steve Azzolin, Eloïse Berthier, Anna Hedström, Patricia Delhomme, David Filliat, Nicolas Bousquet, Goran Frehse, Massimiliano Mancini, Baptiste Caramiaux, Andrea Passerini, Gianni Franchi

Moral Change Or Noise? on Problems of Aligning AI with Temporally Unstable Human Feedback 37501
Vijay Keswani, Cyrus Cousins, Breanna Nguyen, Vincent Conitzer, Hoda Heidari, Jana Schaich Borg, Walter Sinnott-Armstrong

Transparent Networks for Multivariate Time Series 37510
Minkyu Kim, Suan Lee, Jinho Kim

Align to Structure: Aligning Large Language Models with Structural Information 37519
Zae Myung Kim, Anand Ramachandran, Farideh Tavazoee, Joo-Kyung Kim, Oleg Rokhlenko, Dongyeop Kang

Beautiful Images, Toxic Words: Understanding and Addressing Offensive Text in Generated Images 37529
Aditya Kumar, Tom Blanchard, Adam Dziedzic, Franziska Boenisch

Cost-Minimized Label-Flipping Poisoning Attack to LLM Alignment 37538
Shigeki Kusaka, Keita Saito, Mikoto Kudo, Takumi Tanabe, Akifumi Wachi, Youhei Akimoto

Dropouts in Confidence: Moral Uncertainty in Human-LLM Alignment 37547
Jea Kwon, Luiz Felipe Vecchietti, Sungwon Park, Meeyoung Cha

Selective Weak-To-Strong Generalization 37556
Hao Lang, Fei Huang, Yongbin Li

MobileSafetyBench: Evaluating Safety of Autonomous Agents in Mobile Device Control 37565
Juyong Lee, Dongyoon Hahm, June Suk Choi, W. Bradley Knox, Kimin Lee

STELAR-VISION: Self-Topology-Aware Efficient Learning for Aligned Reasoning in Vision 37574
Chen Li, Han Zhang, Zhantao Yang, Fangyi Chen, Zihan Wang, Anudeepsekhar Bolimera, Marios Savvides

ARGH-Mark: Anchor-Synchronized Watermarking with Hamming Correction for Robust and Quality-Preserving LLM Attribution 37583
He Li, Xiaojun Chen, Jingcheng He, Zhendong Zhao, Shuguang Yuan, Xin Zhao, Yunfei Yang

StyleBreak: Revealing Alignment Vulnerabilities in Large Audio-Language Models Via Style-Aware Audio Jailbreak 37591
Hongyi Li, Chengxuan Zhou, Chu Wang, Sicheng Liang, Yanting Chen, Qinlin Xie, Jiawei Ye, Jie Wu

How Bias Binds: Measuring Hidden Associations for Bias Control in Text-To-Image Compositions 37600
Jeng-Lin Li, Ming-Ching Chang, Wei-Chao Chen

TORA: Train Once, Realign Anytime for Offline Multi-Objective Reinforcement Learning 37609
Weichen Li, Waleed Mustafa, Marcio Monteiro, Puyu Wang, Marius Kloft, Sophie Fellenz

Bolster Hallucination Detection Via Prompt-Guided Data Augmentation 37618
Wenyun Li, Zheng Zhang, Dongmei Jiang, Xiangyuan Lan

Editing as Unlearning: Are Knowledge Editing Methods Strong Baselines for Large Language Model Unlearning? 37627
Zexi Li, Xiangzhu Wang, William F. Shen, Meghdad Kurmanji, Xinchi Qiu, Dongqi Cai, Chao Wu, Nicholas D. Lane

How Much Do Large Language Model Cheat on Evaluation? Benchmarking Overestimation Under the One-Time-Pad-Based Framework 37636
Zi Liang, Liantong Yu, Zhang Shiyu, Qingqing Ye, Haibo Hu

Semantics-Preserving Adversarial Attacks on Event-Driven Stock Prediction Models 37645
Aofan Liu, Haoxuan Li, Hongjian Xing, Yuguo Yin, Zijun Li, Yiyan Qi

SRAM: Shape-Realism Alignment Metric for No Reference 3D Shape Evaluation 37654
Sheng Liu, Tianyu Luan, Phani Nuney, Xuelu Feng, Junsong Yuan

MRACL: Multi-Reward Space Guided Adaptive Curriculum Reinforcement Learning for LLMs 37663
Wenxuan Liu, Liangyu Huo, Yi Jing, Xiyuan Zhang, Jian Xie

On the Alignment of Large Language Models with Global Human Opinion 37673
Yang Liu, Masahiro Kaneko, Chenhui Chu

DarkBench+: An Extended Benchmark for Evaluating Dark Patterns in Large Language Models 37682
Yaowen Liu, Shenjia Jing, Yufei Wei, Shoumin Zhang, Jinglu Zhang, Zhen Mei, Liangliang Yue, Jiarui Wang, Peng Zhang

Targeting Misalignment: A Conflict-Aware Framework for Reward-Model-Based LLM Alignment 37692
Zixuan Liu, Siavash H. Khajavi, Guangkai Jiang, Xinru Liu

Mitigating Self-Preference by Authorship Obfuscation 37701
Taslim Mahbub, Shi Feng

DETONATE – a Benchmark for Text-To-Image Alignment and Kernelized Direct Preference Optimization 37709
Renjith Prasad Kaippilly Mana, Abhilekh Borah, Hasnat Md Abdullah, Chathurangi Shyalika, Gurpreet Singh, Ritvik Garimella, Rajarshi Roy, Harshul Raj Surana, Nasrin Imanpour, Suranjana Trivedy, Amit Sheth, Amitava Das

Misalignment from Treating Means as Ends 37719
Henrik Marklund, Alex Infanger, Benjamin Van Roy
STACK: Adversarial Attacks on LLM Safeguard Pipelines 37728
Ian R. McKenzie, Oskar John Hollinsworth, Tom Tseng, Xander Davies, Stephen Casper, Aaron David Tucker, Robert Kirk, Adam Gleave
Aligning Machiavellian Agents: Behavior Steering Via Test-Time Policy Shaping 37738
Dena Mujtaba, Brian Hu, Anthony Hoogs, Arslan Basharat
SharedRep-RLHF: A Shared Representation Approach to RLHF with Diverse Preferences 37747
Arpan Mukherjee, Marcello Bullo, Deniz Gündüz
Quiet Feature Learning in Algorithmic Tasks 37756
Prudhviraj Naidu, Zixian Wang, Leon Bergen, Ramamohan Paturi
A Tale of Two Identities: An Ethical Audit of AI-Crafted Synthetic Personas 37765
Pranav Narayanan Venkit, Jiayi Li, Yingfan Zhou, Sarah Rajtmajer, Shomir Wilson
Intrinsic Barriers and Practical Pathways for Human–AI Alignment: An Agreement-Based Complexity Analysis 37775
Aran Nayebi
CTPD: Cross Tokenizer Preference Distillation 37783
Truong Nguyen, Phi Van Dat, Ngan Nguyen, Linh Ngo Van, Trung Le, Thanh Hong Nguyen
Realist and Pluralist Conceptions of Intelligence and Their Implications on AI Research 37791
Ninell Oldenburg, Ruchira Dhar, Anders Søgaard
LieCraft: A Multi-Agent Framework for Evaluating Deceptive Capabilities in Language Models 37802
Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Tri Nguyen, Vasudev Lal, Joseph Campbell, Simon Stepputtis, Shao-Yen Tseng
Refine and Align: Confidence Calibration Through Multi-Agent Interaction in VQA 37810
Ayush Pandey, Jai Bardhan, Ishita Jain, Ramya S Hebbalaguppe, Rohan Raju Dhanakshirur, Lovekesh Vig
AdvBDGen: A Robust Framework for Generating Adaptive and Stealthy Backdoors in LLM Alignment 37820
Pankayaraj Pathmanathan, Udari Madhushani Sehwag, Michael-Andrei Panaitescu-Liess, Cho-Yu Jason Chiang, Furong Huang
Beyond I'm Sorry, I Can't: Dissecting Large-Language-Model Refusal 37830
Nirmalendu Prakash, Yeo Wei Jie, Amir Abdullah, Ranjan Satapathy, Erik Cambria, Roy Ka-Wei Lee
Towards Benchmarking Privacy Vulnerabilities in Selective Forgetting with Large Language Models 37839
Wei Qian, Chenxu Zhao, Yangyi Li, Mengdi Huai
Backdoor Attacks on Open Vocabulary Object Detectors Via Multi-Modal Prompt Tuning 37849
Ankita Raj, Chetan Arora
Chain-Of-Thought Driven Adversarial Scenario Extrapolation for Robust Language Models 37858
Md Rafi Ur Rashid, Vishnu Asutosh Dasu, Ye Wang, Gang Tan, Shagufta Mehnaz
FindTheFlaws: Annotated Errors for Detecting Flawed Reasoning and Scalable Oversight Research 37867
Gabriel Recchia, Chatrik Singh Mangat, Issac Li, Gayatri Krishnakumar
Confirmation Bias: A Challenge for Scalable Oversight 37877
Gabriel Recchia, Chatrik Singh Mangat, Jinu Nyachhyon, Mridul Sharma, Callum Canavan, Dylan Epstein-Gross, Muhammed Abdulbari
Mind the Gap: Quantifying and Aligning Human-AI Visual Attention for Accident Anticipation 37887
Hoe Sung Ryu, Christian Wallraven
Polarity-Aware Probing for Quantifying Latent Alignment in Language Models 37896
Sabrina Sadiekh, Elena Ericheva, Chirag Agarwal
Detecting Compute Structuring in AI Governance is Likely Feasible 37904
Emmanouil Seferis, Timothy Fist

Tight Robustness Certification Through the Convex Hull of ℓ_0 Attacks 37913
Yuval Shapira, Dana Drachsler-Cohen
EASE: Practical and Efficient Safety Alignment for Small Language Models 37923
Haonan Shi, Guoli Wang, Tu Ouyang, An Wang
Efficient Switchable Safety Control in LLMs Via Magic-Token-Guided Co-Training 37932
Jianfeng Si, Lin Sun, Zhewen Tan, Xiangzheng Zhang
Beyond Verdicts: Evaluating Language Model Moral Competence 37941
Aaron J Snoswell, Daniel Kilov, Seth Lazar
SMPRO: Self-Supervised Visual Preference Alignment Via Differentiable Multi-Preference Multi-Group Ranking 37951
Sirnam Swetha, Rui Meng, Shwetha Ram, Tal Neiman, Son Tran, Mubarak Shah
Persistent Instability in LLM's Personality Measurements: Effects of Scale, Reasoning, and Conversation History 37961
Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, Guillaume Dumas
Shadows in the Code: Exploring the Risks and Defenses of LLM-Based Multi-Agent Software Development Systems 37970
Xiaoqing Wang, Keman Huang, Bin Liang, Hongyu Li, Xiaoyong Du
Benchmarking Trustworthiness in Multimodal LLMs for Video Understanding 37979
Youze Wang, Zijun Chen, Ruoyu Chen, Shishen Gu, Wenbo Hu, Jiayang Liu, Yinpeng Dong, Hang Su, Jun Zhu, Meng Wang, Richang Hong
STAR-1: Safer Alignment of Reasoning LLMs with 1K Data 37988
Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Yanqing Liu, Jieru Mei, Brian R. Bartoldson, Bhavya Kailkhura, Cihang Xie
CluCERT: Certifying LLM Robustness Via Clustering-Guided Denoising Smoothing 37998
Zixia Wang, Gaojie Jin, Jia Hu, Ronghui Mu
Safe Multi-Agent Reinforcement Learning with Natural Language Constraints 38007
Ziyan Wang, Meng Fang, Tristan Tomilin, Fei Fang, Yali Du
Designing Incident Reporting Systems for Harms from General-Purpose AI 38016
Kevin Wei, Lennart Heim
HumorReject: Decoupling LLM Safety from Refusal Prefix Via a Little Humor 38030
Zihui Wu, Haichang Gao, Jiacheng Luo, Zhaoxiang Liu
MCA-Bench: A Multimodal Benchmark for Evaluating CAPTCHA Robustness Against VLM-Based Attacks 38039
Zonglin Wu, Yule Xue, Yaoyao Feng, Xiaolong Wang, Yiren Song
MedAtlas: Evaluating LLMs for Multi-Round, Multi-Task Medical Reasoning Across Diverse Imaging Modalities and Clinical Text 38048
Ronghao Xu, Zhen Huang, Yangbo Wei, Xiaoqian Zhou, Zikang Xu, Ting Liu, Zihang Jiang, S. Kevin Zhou
When Human Preferences Flip: An Instance-Dependent Robust Loss for RLHF 38057
Yifan Xu, Xichen Ye, Yifan Chen, Qiaosheng Zhang
Multi-Faceted Attack: Exposing Cross-Model Vulnerabilities in Defense-Equipped Vision-Language Models 38066
Yijun Yang, Lichao Wang, Jianping Zhang, Chi Harold Liu, Lanqing Hong, Qiang Xu
CoSPED: Consistent Soft Prompt Targeted Data Extraction and Defense 38075
Zhuochen Yang, Kar Wai Fok, Vrizlynn L. L. Thing
Fading the Digital Ink: A Universal Black-Box Attack Framework for 3DGS Watermarking Systems 38084
Qingyuan Zeng, Shu Jiang, Jiajing Lin, Zhenzhong Wang, Kay Chen Tan, Min Jiang
SL-CBM: Enhancing Concept Bottleneck Models with Semantic Locality for Better Interpretability 38093
Hanwei Zhang, Luo Cheng, Rui Wen, Yang Zhang, Lijun Zhang, Holger Hermanns

Differentiated Directional Intervention: A Framework for Evading LLM Safety Alignment..... 38102
Peng Zhang, Peijie Sun
DAVSP: Safety Alignment for Large Vision-Language Models Via Deep Aligned Visual Safety Prompt 38111
Yitong Zhang, Jia Li, Liyi Cai, Ge Li
CultureRL: Internalizing Cultural Principles in Large Language Models Via Norm-Driven Reinforcement Learning 38120
Weixiang Zhao, Haozhen Li, Yanyan Zhao, Haixiao Liu, Biye Li, Ting Liu, Bing Qin
Composable Assurance for AI Alignment: A Framework for Propagating Formal Safety Properties Through MLOps 38129
Xiaofei Zhao
Value-Aligned Prompt Moderation Via Zero-Shot Agentic Rewriting for Safe Image Generation..... 38137
Xin Zhao, Xiaojun Chen, Bingshan Liu, Zeyao Liu, Zhendong Zhao, Xiaoyan Gu
GEM: Generative Entropy-Guided Preference Modeling for Few-Shot Alignment of LLMs 38146
Yiyang Zhao, Huiyu Bai, Xuejiao Zhao
Explainable Melanoma Diagnosis with Contrastive Learning and LLM-Based Report Generation..... 38156
Junwen Zheng, Xinran Xu, Li Rong Wang, Chang Cai, Lucinda Siyun Tan, Dingyuan Wang, Hong Liang Tey, Xiuyi Fan
Can LLMs Detect Their Confabulations? Estimating Reliability in Uncertainty-Aware Language Models 38164
Tianyi Zhou, Johanne Medina, Sanjay Chawla
Not All Tokens Are Meant to Be Forgotten 38173
Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, Dongxiao Zhu
Author Index