

39th AAAI Conference on Artificial Intelligence (AAAI-25), 37th Conference on Innovative Applications of Artificial Intelligence (IAAI-25), and 15th Symposium on Educational Advances in Artificial Intelligence (EAAI-25)

Volume 26: AAAI Special Track

- AI Alignment

Philadelphia, Pennsylvania, USA
25 February - 4 March 2025

ISBN: 979-8-3313-2536-7

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© (2025) by Association for the Advancement of Artificial Intelligence
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

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 AI ALIGNMENT

SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models	27188
<i>Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, Rima Hazra</i>	
Bridging the Knowledge Gap: Understanding User Expectations for Trustworthy LLM Standards	27197
<i>Michaela Benk, Léane Wettstein, Nadine Schlicker, Florian von Wangenheim, Nicolas Scharowski</i>	
Scaling Trends for Data Poisoning in LLMs	27206
<i>Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, Kellin Pelrine</i>	
Verification of Neural Networks Against Convolutional Perturbations via Parameterised Kernels	27215
<i>Benedikt Brückner, Alessio Lomuscio</i>	
Risk Controlled Image Retrieval	27224
<i>Kaiwen Cai, Chris Xiaoxuan Lu, Xingyu Zhao, Wei Huang, Xiaowei Huang</i>	
Political Bias Prediction Models Focus on Source Cues, Not Semantics.....	27233
<i>Selin Chun, Daejin Choi, Taekyoung Kwon</i>	
Searching for Unfairness in Algorithms’ Outputs: Novel Tests and Insights.....	27242
<i>Ian Davidson, S. S. Ravi</i>	
In Search of Trees: Decision-Tree Policy Synthesis for Black-Box Systems via Search	27250
<i>Emir Demirović, Christian Schilling, Anna Lukina</i>	
Evaluate with the Inverse: Efficient Approximation of Latent Explanation Quality Distribution.....	27258
<i>Carlos Eiras-Franco, Anna Hedström, Marina M.-C. Höhne</i>	
Retrieving Versus Understanding Extractive Evidence in Few-Shot Learning	27268
<i>Karl Elbakian, Samuel Carton</i>	
LEGEND: Leveraging Representation Engineering to Annotate Safety Margin for Preference Datasets	27277
<i>Duanyu Feng, Bowen Qin, Chen Huang, Youcheng Huang, Zheng Zhang, Wenqiang Lei</i>	
SMLE: Safe Machine Learning via Embedded Overapproximation	27286
<i>Matteo Francobaldi, Michele Lombardi</i>	
MIA-Tuner: Adapting Large Language Models as Pre-training Text Detector	27295
<i>Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, Tao Jiang</i>	
The Partially Observable Off-Switch Game	27304
<i>Andrew Garber, Rohan Subramani, Linus Luu, Mark Bedaywi, Stuart Russell, Scott Emmons</i>	
UFID: A Unified Framework for Black-box Input-level Backdoor Detection on Diffusion Models	27312
<i>Zihan Guan, Mengxuan Hu, Sheng Li, Anil Kumar Vullikanti</i>	
Robust Multi-Objective Preference Alignment with Online DPO.....	27321
<i>Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, Abhinav Rastogi</i>	

Token Highlighter: Inspecting and Mitigating Jailbreak Prompts for Large Language Models.....	27330
<i>Xiaomeng Hu, Pin-Yu Chen, Tsung-Yi Ho</i>	
Joint Scoring Rules: Competition Between Agents Avoids Performative Prediction	27339
<i>Rubi Hudson</i>	
ChatBug: A Common Vulnerability of Aligned LLMs Induced by Chat Templates	27347
<i>Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, Radha Poovendran</i>	
Dynamic Algorithm Termination for Branch-and-Bound-based Neural Network Verification	27356
<i>Konstantin Kaulen, Matthias König, Holger H. Hoos</i>	
Quantifying Misalignment Between Agents: Towards a Sociotechnical Understanding of Alignment.....	27365
<i>Aidan Kierans, Avijit Ghosh, Hananel Hazan, Shiri Dori-Hacohen</i>	
On the Consideration of AI Openness: Can Good Intent Be Abused?.....	27374
<i>Yeeun Kim, Hyunseo Shin, Eunkyung Choi, Hongseok Oh, Hyunjun Kim, Wonseok Hwang</i>	
Dynamic Back-Substitution in Bound-Propagation-Based Neural Network Verification	27383
<i>Panagiotis Kouvaros, Benedikt Brückner, Patrick Henriksen, Alessio Lomuscio</i>	
Maximizing Signal in Human-Model Preference Alignment	27392
<i>Kelsey Kraus, Margaret Kroll</i>	
Sequential Decision Making in Stochastic Games with Incomplete Preferences over Temporal Objectives.....	27401
<i>Abhishek Ninad Kulkarni, Jie Fu, Ufuk Topcu</i>	
Debate Helps Weak-to-Strong Generalization.....	27410
<i>Hao Lang, Fei Huang, Yongbin Li</i>	
JailPO: A Novel Black-Box Jailbreak Framework via Preference Optimization Against Aligned LLMs.....	27419
<i>Hongyi Li, Jiawei Ye, Jie Wu, Tianjie Yan, Chu Wang, Zhixin Li</i>	
Internal Activation Revision: Safeguarding Vision Language Models Without Parameter Update.....	27428
<i>Qing Li, Jiahui Geng, Derui Zhu, Zongxiong Chen, Kun Song, Lei Ma, Fakhri Karray</i>	
Strong Empowered and Aligned Weak Mastered Annotation for Weak-to-Strong Generalization	27437
<i>Yongqi Li, Xin Miao, Mayi Xu, Tiejun Qian</i>	
Retention Score: Quantifying Jailbreak Risks for Vision Language Models.....	27446
<i>Zaitang Li, Pin-Yu Chen, Tsung-Yi Ho</i>	
Exploring Intrinsic Alignments Within Text Corpus	27455
<i>Zi Liang, Pinghui Wang, Ruofei Zhang, Haibo Hu, Shuo Zhang, Qingqing Ye, Nuo Xu, Yaxin Xiao, Chen Zhang, Lizhen Cui</i>	
Is Your Autonomous Vehicle Safe? Understanding the Threat of Electromagnetic Signal Injection Attacks on Traffic Scene Perception.....	27464
<i>Wenhao Liao, Sineng Yan, Youqian Zhang, Xinwei Zhai, Yuanyuan Wang, Eugene Fu</i>	
Single Character Perturbations Break LLM Alignment.....	27473
<i>Leon Lin, Hannah Brown, Kenji Kawaguchi, Michael Shieh</i>	
Data with High and Consistent Preference Difference Are Better for Reward Model	27482
<i>Qi Lin, Hengtong Lu, Caixia Yuan, Xiaojie Wang, Huixing Jiang, Wei Chen</i>	

Bias Unveiled: Investigating Social Bias in LLM-Generated Code	27491
<i>Lin Ling, Fazle Rabbi, Song Wang, Jinqiu Yang</i>	
Stream Aligner: Efficient Sentence-Level Alignment via Distribution Induction	27500
<i>Hantao Lou, Jiaming Ji, Kaile Wang, Yaodong Yang</i>	
Sequential Preference Optimization: Multi-Dimensional Preference Alignment with Implicit Reward Modeling	27509
<i>Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, Kaiqi Huang</i>	
MAPLE: A Framework for Active Preference Learning Guided by Large Language Models	27518
<i>Saaduddin Mahmud, Mason Nakamura, Shlomo Zilberstein</i>	
SYNAPSE: SYmbolic Neural-Aided Preference Synthesis Engine	27529
<i>Sadanand Modak, Noah Tobias Patton, Isil Dillig, Joydeep Biswas</i>	
Neural Continuous-Time Supermartingale Certificates.....	27538
<i>Grigory Neustroev, Mirco Giacobbe, Anna Lukina</i>	
Text-Diffusion Red-Teaming of Large Language Models: Unveiling Harmful Behaviors with Proximity Constraints.....	27547
<i>Jonathan Nöther, Adish Singla, Goran Radanovic</i>	
Is Poisoning a Real Threat to DPO? Maybe More So Than You Think	27556
<i>Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, Furong Huang</i>	
Do Transformer Interpretability Methods Transfer to RNNs?.....	27565
<i>Gonçalo Paulo, Thomas Marshall, Nora Belrose</i>	
Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems.....	27573
<i>Pierre Peigné, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, Esben Kran</i>	
Increased Compute Efficiency and the Diffusion of AI Capabilities.....	27582
<i>Konstantin F. Pilz, Lennart Heim, Nicholas Brown</i>	
Neurons to Words: A Novel Method for Automated Neural Network Interpretability and Alignment.....	27591
<i>Lukas-Santo Puglisi, Fabio Valdés, Jakob Johannes Metzger</i>	
SEAL: Systematic Error Analysis for Value ALignment.....	27599
<i>Manon Revel, Matteo Cargnelutti, Tyna Eloundou, Greg Leppert</i>	
ME: Modelling Ethical Values for Value Alignment	27608
<i>Eryn Rigley, Adriane Chapman, Christine Evers, Will McNeill</i>	
SafetyPrompts: A Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety.....	27617
<i>Paul Röttger, Fabio Pernisi, Bertie Vidgen, Dirk Hovy</i>	
Reinforcement Learning Platform for Adversarial Black-box Attacks with Custom Distortion Filters.....	27628
<i>Soumyendu Sarkar, Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Ricardo Luna Gutiérrez, Antonio Guillen, Desik Rengarajan</i>	

Partial Identifiability in Inverse Reinforcement Learning for Agents with Non-Exponential Discounting	27636
<i>Joar Max Viktor Skalse, Alessandro Abate</i>	
Generalizing Alignment Paradigm of Text-to-Image Generation with Preferences Through f-Divergence Minimization	27644
<i>Haoyuan Sun, Bo Xia, Yongzhe Chang, Xueqian Wang</i>	
Align-Pro: A Principled Approach to Prompt Optimization for LLM Alignment	27653
<i>Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, George K. Atia</i>	
Can Go AIs Be Adversarially Robust?	27662
<i>Tom Tseng, Euan McLean, Kellin Pelrine, Tony Tong Wang, Adam Gleave</i>	
ERCI: An Explainable Experience Replay Approach with Causal Inference for Deep Reinforcement Learning	27671
<i>Jingwen Wang, Dehui du, Lili Tian, Yikang Chen, Yida Li, YiYang Li</i>	
Towards a Theory of AI Personhood	27680
<i>Francis Rhys Ward</i>	
MMJ-Bench: A Comprehensive Study on Jailbreak Attacks and Defenses for Vision Language Models	27689
<i>Fenghua Weng, Yue Xu, Chengyan Fu, Wenjie Wang</i>	
LLMs in the Classroom: Outcomes and Perceptions of Questions Written with the Aid of AI.....	27698
<i>Gavin Witsken, Igor Crk, Eren Gultepe</i>	
DR-Encoder: Encode Low-rank Gradients with Random Prior for Large Language Models Differentially Privately	27706
<i>Huiwen Wu, Deyi Zhang, Xiaohan Li, Xiaogang Xu, Jiafei Wu, Zhe Liu</i>	
IBAS: Imperceptible Backdoor Attacks in Split Learning with Limited Information	27715
<i>Peng Xi, Shaoliang Peng, Wenjuan Tang</i>	
Evaluating Mathematical Reasoning Beyond Accuracy	27723
<i>Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, Pengfei Liu</i>	
Measuring Error Alignment for Decision-Making Systems	27731
<i>Binxia Xu, Antonis Bikakis, Daniel F.O. Onah, Andreas Vlachidis, Luke Dickens</i>	
Enhance Modality Robustness in Text-Centric Multimodal Alignment with Adversarial Prompting	27740
<i>Yun-Da Tsai, Ting-Yu Yen, Keng-Te Liao, Shou-De Lin</i>	
Aligning Large Language Models for Faithful Integrity Against Opposing Argument	27748
<i>Yong Zhao, Yang Deng, See-Kiong Ng, Tat-Seng Chua</i>	
CALM: Curiosity-Driven Auditing for Large Language Models	27757
<i>Xiang Zheng, Longxiang Wang, Yi Liu, Xingjun Ma, Chao Shen, Cong Wang</i>	
Sequence to Sequence Reward Modeling: Improving RLHF by Language Feedback	27765
<i>Jiayi Zhou, Jiaming Ji, Josef Dai, Yaodong Yang</i>	

Author Index