Thursday - February 27

9:30-10:45am - **Algorithmic Alignment**

32	Qing Li, Jiahui Geng, Derui Zhu, Zongxiong Chen, Kun Song, Lei Ma, Fakhri Karray	Internal Activation Revision: Safeguarding Vision Language Models without Parameter Update
90	Zi Liang, Pinghui Wang, Ruofei Zhang, Haibo Hu, Shuo Zhang, Qingqing Ye, Nuo Xu, Yaxin Xiao, Chen Zhang, Lizhen Cui	Exploring Intrinsic Alignments within Text Corpus
126	Jiayi Zhou, Jiaming Ji, Josef Dai, Yaodong Yang	Sequence to Sequence Reward Modeling: Improving RLHF by Language Feedback
182	Hao Lang, Fei Huang, Yongbin Li	Debate Helps Weak-to-Strong Generalization

11:15am-12:30pm - **Evaluating Safety and Alignment**

78	Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, Pengfei Liu	Evaluating Mathematical Reasoning Beyond Accuracy
124	Binxia Xu, Antonis Bikakis, Daniel F.O. Onah, Andreas Vlachidis, Luke Dickens	Measuring Error Alignment for Decision-Making Systems
333	Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, Tao Jiang	MIA-Tuner: Adapting Large Language Models as Pre-training Text Detector
137	Manon Revel, Matteo Cargnelutti, Tyna Eloundou, Greg Leppert	SEAL: Systematic Error Analysis for Value Alignment

12:30-2:30pm - **Poster Session I**

5	Hongyi Li, Jiawei Ye, Wu Jie, Tianjie Yan, 王楚, Zhixin Li	JailPO: A Novel Black-box Jailbreak Framework via Preference Optimization against Aligned LLMs
8	Benedikt Brückner, Alessio Lomuscio	Verification of Neural Networks against Convolutional Perturbations via Parameterised Kernels
9	Carlos Eiras-Franco, Anna Hedström, Marina MC Höhne	Evaluate with the Inverse: Efficient Approximation of Latent Explanation Quality Distribution
15	Paul Röttger, Fabio Pernisi, Bertie Vidgen, Dirk Hovy	SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety
18	Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, Furong Huang	Is poisoning a real threat to DPO? Maybe more so than you think
42	peng xi, Shaoliang Peng, Wenjuan Tang	IBAS:Imperceptible Backdoor Attacks in Split Learning with Limited Information
51	Yeeun Kim, Hyunseo Shin, Eunkyung Choi, Hongseok Oh, Hyunjun Kim, Wonseok Hwang	On the Consideration of AI Openness: Can Good Intent Be Abused?
58	Joar Max Viktor Skalse, Alessandro Abate	Partial Identifiability in Inverse Reinforcement Learning For Agents With Non-Exponential Discounting

61	Kaiwen Cai, Chris Xiaoxuan Lu, Xingyu Zhao, Wei Huang, Xiaowei Huang	Risk Controlled Image Retrieval
66	Gonçalo Santos Paulo, Thomas Marshall, Nora Belrose	Do Transformer Interpretability Methods Transfer to RNNs?
81	Yong Zhao, Yang Deng, See-Kiong Ng, Tat-Seng Chua	Aligning Large Language Models for Faithful Integrity against Opposing Argument
87	Leon Lin, Hannah Brown, Kenji Kawaguchi, Michael Shieh	Single Character Perturbations Break LLM Alignment
96	Eryn Rigley, Adriane Chapman, Christine Evers, Will McNeill	ME: Modelling Ethical Values for Value Alignment
100	Yun-Da Tsai, Ting-Yu Yen, Keng-Te Liao, Shou-De Lin	Enhance Modality Robustness in Text-Centric Multimodal Alignment with Adversarial Prompting
108	Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, Radha Poovendran	ChatBug: A Common Vulnerability of Aligned LLMs Induced by Chat Templates
109	Zihan Guan, Mengxuan Hu, Sheng Li, Anil Kumar Vullikanti	UFID: A Unified Framework for Black-box Input-level Backdoor Detection on Diffusion Models
113	Matteo Francobaldi, Michele Lombardi	SMLE: Safe Machine Learning via Embedded Overapproximation
119	Konstantin Friedemann Pilz, Lennart Heim, Nicholas Brown	Increased Compute Efficiency and the Diffusion of AI Capabilities
133	Ian Davidson, S. S. Ravi	Searching for Unfairness in Algorithms' Outputs: Novel Tests and Insights
140	ZAITANG LI, Pin-Yu Chen, Tsung-Yi Ho	Retention Score: Quantifying Jailbreak Risks for Vision Language Models
144	Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, Kellin Pelrine	Scaling Laws for Data Poisoning in LLMs
166	Qi Lin, Hengtong Lu, Caixia Yuan, Xiaojie Wang, Huixing Jiang, Chen Wei	Data with High and Consistent Preference Difference Are Better for Reward Model
168	Lukas-Santo Puglisi, Fabio Valdés, Jakob Johannes Metzger	Neurons to Words: A Novel Method for Automated Neural Network Interpretability and Alignment
171	Hantao Lou, Jiaming Ji, Kaile Wang, Yaodong Yang	Stream Aligner: Efficient Sentence-Level Alignment via Distribution Induction

2:30-3:45pm - **Alignment Foundations and Theory**

269	Andrew Garber, Rohan Subramani, Linus Luu, Mark Bedaywi, Stuart Russell, Scott Emmons	Will an AI with Private Information Allow Itself to be Switched Off?
318	Rubi Hudson	Incentivizing Honesty for Conditional Predictions via Zero-Sum Competition
7	Panagiotis Kouvaros, Benedikt Brückner, Patrick Henriksen, Alessio Lomuscio	Dynamic Back-Substitution in Bound-Propagation-Based Neural Network Verification
299	Grigory Neustroev, Mirco Giacobbe, Anna Lukina	Neural Continuous-Time Supermartingale Certificates

Friday - February 28

9:30-10:45am **- Human-Centered Alignment**

34	Jingwen Wang, Dehui du, Lili Tian, Yikang Chen, Yida Li, YiYang Li	ERCI: An Explainable Experience Replay Approach with Causal Inference for Deep Reinforcement Learning
63	Gavin Witsken, Igor Crk, Eren Gultepe	LLMs in the Classroom: Outcomes and Perceptions of Questions Written with the Aid of AI
300	Michaela Benk, Léane Wettstein, Nadine Schlicker, Florian von Wangenheim, Nicolas Scharowski	Bridging the Knowledge Gap: Understanding User Expectations for Trustworthy LLM Standards
230	Sadanand Modak, Noah Tobias Patton, Isil Dillig, Joydeep Biswas	From Vision to Preferences: Neurosymbolic Programming for Personalized Robotics

11:15am-12:30pm **- Security and Robustness**

30	Xiaomeng Hu, Pin-Yu Chen, Tsung-Yi Ho	Token Highlighter: Inspecting and Mitigating Jailbreak Prompts for Large Language Models
62	Jonathan Nöther, Adish Singla, Goran Radanovic	Text-Diffusion Red-Teaming of Large Language Models: Unveiling Harmful Behaviors with Proximity Constraints
218	Tom Tseng, Euan McLean, Kellin Pelrine, Tony Tong Wang, Adam Gleave	Can Go AIs be adversarially robust?
330	Pierre Peigné, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, Esben Kran	Multi-Agent Security Tax: Trading Off Security and Collaboration Capabilities in Multi-Agent Systems

12:30-2:30pm - **Poster Session II**

173	Yongqi Li, Xin Miao, Mayi Xu, Tieyun Qian	Strong Empowered and Aligned Weak Mastered Annotation for Weak-to-Strong Generalization

189	Konstantin Kaulen, Matthias König, Holger Hoos	Dynamic Algorithm Termination for Branch-and-Bound-based Neural Network Verification
196	Francis Rhys Ward	Towards a Theory of AI Personhood
198	Fenghua Weng, Yue Xu, Chengyan Fu, Wenjie Wang	$\textit{MMJ-Bench}$: A Comprehensive Study on Jailbreak Attacks and Defenses for Vision Language Models
199	Abhishek Ninad Kulkarni, Jie Fu, ufuk topcu	Sequential Decision Making in Stochastic Games with Incomplete Preferences over Temporal Objectives
213	Xiang Zheng, Longxiang WANG, Yi Liu, Xingjun Ma, Chao Shen, Cong Wang	CALM: Curiosity-Driven Auditing for Large Language Models
215	Lin Ling, Fazle Rabbi, Song Wang, Jinqiu Yang	Bias Unveiled: Investigating Social Bias in LLM-Generated Code
221	Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, Rima Hazra	SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models
222	Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, George K. Atia	Align-Pro: A Principled Approach to Prompt Optimization for LLM Alignment
229	Margaret Kroll, Kelsey Kraus	Maximizing Signal in Human-Model Preference Alignment
245	Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, Abhinav Rastogi	Robust Multi-Objective Preference Alignment with Online DPO
246	Soumyendu Sarkar, Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Ricardo Luna Gutierrez, Antonio Guillen, Desik Rengarajan	Reinforcement Learning Platform for Adversarial Black-box Attacks with Custom Distortion Filters
250	Huiwen Wu, Deyi Zhang, Xiaohan Li, Xiaogang Xu,	DR-Encoder: Encode Low-rank Gradients with Random Prior for Large Language Models Differentially Privately

	Jiafei Wu, Zhe Liu	
260	Aidan Kierans, Avijit Ghosh, Hananel Hazan, Shiri Dori-Hacohen	Quantifying Misalignment Between Agents
266	Saaduddin Mahmud, Mason Nakamura, Shlomo Zilberstein	MAPLE: A Framework for Active Preference Learning Guided by Large Language Models
267	Wenhao Liao, Sineng Yan, Youqian Zhang, Xinwei Zhai, Yuanyuan Wang, Eugene Fu	Is Your Autonomous Vehicle Safe? Understanding the Threat of Electromagnetic Signal Injection Attacks
268	Karl Elbakian, Samuel Carton	Retrieving Versus Understanding Extractive Evidence in Few-Shot Learning
272	Selin Chun, Daejin Choi, Taekyoung Kwon	Political Bias Prediction Models Focus on Source Cues, Not Semantics
280	Duanyu Feng, Bowen Qin, Chen Huang, Youcheng Huang, Zheng Zhang, Wenqiang Lei	Legend: Leveraging Representation Engineering to Annotate Safety Margin for Preference Datasets
281	Xingzhou Lou, Junge Zhang, Jian Xie, lifeng Liu, Dong Yan, Kaiqi Huang	Sequential Preference Optimization: Multi-Dimensional Preference Alignment With Implicit Reward Modeling
282	Akash Wasil, Everett Thornton Smith, Corin Katzke, Justin Bullock	AI Emergency Preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats
294	Emir Demirović, Christian Schilling, Anna Lukina	In Search of Trees: Decision-Tree Policy Synthesis for Black-Box Systems via Search
328	Haoyuan Sun, Bo Xia, Yongzhe Chang, Xueqian Wang	Generalizing Alignment Paradigm of Text-to-Image Generation with Preferences through f-divergence Minimization