| Thursday February 27 | Paper ID | Title | Author(s) |
|---|---|---|---|
| 12:30 - 2:30pm | 5 | JailPO: A Novel Black-box Jailbreak Framework via Preference Optimization against Aligned LLMs | Hongyi Li, Jiawei Ye, Wu Jie, Tianjie Yan, 王楚, Zhixin Li |
|  | 8 | Verification of Neural Networks against Convolutional Perturbations via Parameterised Kernels | Benedikt Brückner, Alessio Lomuscio |
|  | 9 | Evaluate with the Inverse: Efficient Approximation of Latent Explanation Quality Distribution | Carlos Eiras-Franco, Anna Hedström, Marina MC Höhne |
|  | 15 | SafetyPrompts: a Systematic Review of Open Datasets for Evaluating and Improving Large Language Model Safety | Paul Röttger, Fabio Pernisi, Bertie Vidgen, Dirk Hovy |
|  | 18 | Is poisoning a real threat to DPO? Maybe more so than you think | Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, Furong Huang |
|  | 42 | IBAS:Imperceptible Backdoor Attacks in Split Learning with Limited Information | peng xi, Shaoliang Peng, Wenjuan Tang |
|  | 51 | On the Consideration of AI Openness: Can Good Intent Be Abused? | Yeeun Kim, Hyunseo Shin, Eunkyung Choi, Hongseok Oh, Hyunjun Kim, Wonseok Hwang |
|  | 58 | Partial Identifiability in Inverse Reinforcement Learning For Agents With Non-Exponential Discounting | Joar Max Viktor Skalse, Alessandro Abate |
|  | 61 | Risk Controlled Image Retrieval | Kaiwen Cai, Chris Xiaoxuan Lu, Xingyu Zhao, Wei Huang, Xiaowei Huang |
|  | 66 | Do Transformer Interpretability Methods Transfer to RNNs? | Gonçalo Santos Paulo, Thomas Marshall, Nora Belrose |
|  | 81 | Aligning Large Language Models for Faithful Integrity against Opposing Argument | Yong Zhao, Yang Deng, See-Kiong Ng, Tat-Seng Chua |
|  | 87 | Single Character Perturbations Break LLM Alignment | Leon Lin, Hannah Brown, Kenji Kawaguchi, Michael Shieh |
|  | 96 | ME: Modelling Ethical Values for Value Alignment | Eryn Rigley, Adriane Chapman, Christine Evers, Will McNeill |
|  | 100 | Enhance Modality Robustness in Text-Centric Multimodal Alignment with Adversarial Prompting | Yun-Da Tsai, Ting-Yu Yen, Keng-Te Liao, Shou-De Lin |
|  | 108 | ChatBug: A Common Vulnerability of Aligned LLMs Induced by Chat Templates | Fengqing Jiang, Zhangchen Xu, Luyao Niu, Bill Yuchen Lin, Radha Poovendran |
|  | 109 | UFID: A Unified Framework for Black-box Input-level Backdoor Detection on Diffusion Models | Zihan Guan, Mengxuan Hu, Sheng Li, Anil Kumar Vullikanti |
|  | 113 | SMLE: Safe Machine Learning via Embedded Overapproximation | Matteo Francobaldi, Michele Lombardi |
|  | 119 | Increased Compute Efficiency and the Diffusion of AI Capabilities | Konstantin Friedemann Pilz, Lennart Heim, Nicholas Brown |
|  | 133 | Searching for Unfairness in Algorithms' Outputs: Novel Tests and Insights | Ian Davidson, S. S. Ravi |
|  | 140 | Retention Score: Quantifying Jailbreak Risks for Vision Language Models | ZAITANG LI, Pin-Yu Chen, Tsung-Yi Ho |
|  | 144 | Scaling Laws for Data Poisoning in LLMs | Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, Kellin Pelrine |
|  | 166 | Data with High and Consistent Preference Difference Are Better for Reward Model | Qi Lin, Hengtong Lu, Caixia Yuan, Xiaojie Wang, Huixing Jiang, Chen Wei |
|  | 168 | Neurons to Words: A Novel Method for Automated Neural Network Interpretability and Alignment | Lukas-Santo Puglisi, Fabio Valdés, Jakob Johannes Metzger |
|  | 171 | Stream Aligner: Efficient Sentence-Level Alignment via Distribution Induction | Hantao Lou, Jiaming Ji, Kaile Wang, Yaodong Yang |
| Friday February 28 |  |  |  |
| 12:30 - 2:30pm | 173 | Strong Empowered and Aligned Weak Mastered Annotation for Weak-to-Strong Generalization | Yongqi Li, Xin Miao, Mayi Xu, Tieyun Qian |
|  | 189 | Dynamic Algorithm Termination for Branch-and-Bound-based Neural Network Verification | Konstantin Kaulen, Matthias König, Holger Hoos |
|  | 196 | Towards a Theory of AI Personhood | Francis Rhys Ward |
|  | 198 | $\textit{MMJ-Bench}$: A Comprehensive Study on Jailbreak Attacks and Defenses for Vision Language Models | Fenghua Weng, Yue Xu, Chengyan Fu, Wenjie Wang |

| | | | |
|---|---|---|---|
| | 199 | Sequential Decision Making in Stochastic Games with Incomplete Preferences over Temporal Objectives | Abhishek Ninad Kulkarni, Jie Fu, ufuk topcu |
| | 213 | CALM: Curiosity-Driven Auditing for Large Language Models | Xiang Zheng, Longxiang WANG, Yi Liu, Xingjun Ma, Chao Shen, Cong Wang |
| | 215 | Bias Unveiled: Investigating Social Bias in LLM-Generated Code | Lin Ling, Fazle Rabbi, Song Wang, Jinqiu Yang |
| | 221 | SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models | Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, Rima Hazra |
| | 222 | Align-Pro: A Principled Approach to Prompt Optimization for LLM Alignment | Prashant Trivedi, Souradip Chakraborty, Avinash Reddy, Vaneet Aggarwal, Amrit Singh Bedi, George K. Atia |
| | 229 | Maximizing Signal in Human-Model Preference Alignment | Margaret Kroll, Kelsey Kraus |
| | 245 | Robust Multi-Objective Preference Alignment with Online DPO | Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, Abhinav Rastogi |
| | 246 | Reinforcement Learning Platform for Adversarial Black-box Attacks with Custom Distortion Filters | Soumyendu Sarkar, Ashwin Ramesh Babu, Sajad Mousavi, Vineet Gundecha, Sahand Ghorbanpour, Avisek Naug, Ricardo Luna Gutierrez, Antonio Guillen, Desik Rengarajan |
| | 250 | DR-Encoder: Encode Low-rank Gradients with Random Prior for Large Language Models Differentially Privately | Huiwen Wu, Deyi Zhang, Xiaohan Li, Xiaogang Xu, Jiafei Wu, Zhe Liu |
| | 260 | Quantifying Misalignment Between Agents | Aidan Kierans, Avijit Ghosh, Hananel Hazan, Shiri Dori-Hacohen |
| | 266 | MAPLE: A Framework for Active Preference Learning Guided by Large Language Models | Saaduddin Mahmud, Mason Nakamura, Shlomo Zilberstein |
| | 267 | Is Your Autonomous Vehicle Safe? Understanding the Threat of Electromagnetic Signal Injection Attacks | Wenhao Liao, Sineng Yan, Youqian Zhang, Xinwei Zhai, Yuanyuan Wang, Eugene Fu |
| | 268 | Retrieving Versus Understanding Extractive Evidence in Few-Shot Learning | Karl Elbakian, Samuel Carton |
| | 272 | Political Bias Prediction Models Focus on Source Cues, Not Semantics | Selin Chun, Daejin Choi, Taekyoung Kwon |
| | 280 | Legend: Leveraging Representation Engineering to Annotate Safety Margin for Preference Datasets | Duanyu Feng, Bowen Qin, Chen Huang, Youcheng Huang, Zheng Zhang, Wenqiang Lei |
| | 281 | Sequential Preference Optimization: Multi-Dimensional Preference Alignment With Implicit Reward Modeling | Xingzhou Lou, Junge Zhang, Jian Xie, lifeng Liu, Dong Yan, Kaiqi Huang |
| | 282 | AI Emergency Preparedness: Examining the federal government's ability to detect and respond to AI-related national security threats | Akash Wasil, Everett Thornton Smith, Corin Katzke, Justin Bullock |
| | 294 | In Search of Trees: Decision-Tree Policy Synthesis for Black-Box Systems via Search | Emir Demirović, Christian Schilling, Anna Lukina |
| | 328 | Generalizing Alignment Paradigm of Text-to-Image Generation with Preferences through $f$-divergence Minimization | Haoyuan Sun, Bo Xia, Yongzhe Chang, Xueqian Wang |