Date	Session Title	Paper ID	Paper Title	Authors
Thursday, Jan 22, 9:30-10:30am	Safety, Oversight, and Governance of Advanced AI Systems (AIA) Conference Room J	AIA67	FindTheFlaws: Annotated Errors for Detecting Flawed Reasoning and Scalable Oversight Research	Gabriel Recchia; Chatrik Singh Mangat; Issac Li; Gayatri Krishnakumar
		AIA206	Designing Incident Reporting Systems for Harms from General-Purpose AI	Kevin Wei; Lennart Heim
		AIA250	DarkBench+: An Extended Benchmark for Evaluating Dark Patterns in Large Language Models	Yaowen Liu; Shenjia Jing; Yufei Wei; Shoumin Zhang; Jinglu Zhang; Zhen Mei; Liangliang Yue; Jiarui Wang; Peng Zhang
		AIA361	Detecting Compute Structuring in AI Governance Is Likely Feasible	Emmanouil Seferis; Timothy Fist
		AIA152	Polarity-Aware Probing for Quantifying Latent Alignment in Language Models	Sabrina Sadiekh; Elena Ericheva; Chirag Agarwal
Thursday, Jan 22, 11:00am-12:00pm	Robustness and Security of LLM Systems (AIA) Conference Room J	AIA83	StyleBreak: Revealing Alignment Vulnerabilities in Large Audio-Language Models via Style-Aware Audio Jailbreak	Hongyi Li; Chengxuan Zhou; Chu Wang; Sicheng Liang; Yanting Chen; Qinlin Xie; Jiawei Ye; Jie Wu
		AIA157	STACK: Adversarial Attacks on LLM Safeguard Pipelines	Ian R. McKenzie; Oskar John Hollinsworth; Tom Tseng; Xander Davies; Stephen Casper; Aaron David Tucker; Robert Kirk; Adam Gleave
		AIA677	Resilience in Ambient Multi-Agent LLMs via Decentralized Bio-Autonomic Control and Immune-Inspired Anomaly Detection	Nastaran Darabi; Devashri Naik; Sina Tayebati; Dinithi Jayasuriya; Amit Ranjan Trivedi
		AIA718	ARGH-Mark: Anchor-Synchronized Watermarking with Hamming Correction for Robust and Quality-Preserving LLM Attribution	He Li; Xiaojun Chen; Jingcheng He; Zhendong Zhao; Shuguang Yuan; Xin Zhao; Yunfei Yang
		AIA729	AlignTree: Efficient Defense Against LLM Jailbreak Attacks	Gil Goren; Shahar Katz; Lior Wolf
Thursday, Jan 22, 3:00-4:00pm	Preference-Based Alignment and RLHF (AIA) Conference Room J	AIA307	On the Exponential Convergence for Offline RLHF with Pairwise Comparisons	Zhirui Chen; Vincent Y. F. Tan
		AIA316	CTPD: Cross Tokenizer Preference Distillation	Truong Nguyen; Phi Van Dat; Ngan Nguyen; Linh Ngo Van; Trung Le; Thanh Hong Nguyen
		AIA327	TAPO: Dynamic Teacher and Perturbed Answer Injection for Policy Optimization	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
		AIA553	SharedRep-RLHF: A Shared Representation Approach to RLHF with Diverse Preferences	Arpan Mukherjee; Marcello Bullo; Deniz Gündüz
		AIA651	Preference Optimization via Contrastive Divergence: Your Policy Is Secretly an NLL Estimator	Zhuotong Chen; Fang Liu; Jennifer Zhu; Haozhu Wang; Jiayu Li; Yanjun Qi; Mohammad Ghavamzadeh
Friday, Jan 23, 9:30-10:30am	Safety and Robustness of Large Models (AIA) Conference Room J	AIA122	CluCERT: Certifying LLM Robustness via Clustering-Guided Denoising Smoothing	Zixia Wang; Gaojie Jin; Jia Hu; Ronghui Mu
		AIA185	Deep Hidden Cognition Facilitates Reliable Chain-of-Thought Reasoning	Zijun Chen; Wenbo Hu; Richang Hong
		AIA224	Multi-Faceted Attack: Exposing Cross-Model Vulnerabilities in Defense-Equipped Vision-Language Models	Yijun Yang; Lichao Wang; Jianping Zhang; Chi Harold Liu; Lanqing Hong; Qiang Xu
		AIA584	Tight Robustness Certification Through the Convex Hull of ℓ_0 Attacks	Yuval Shapira; Dana Drachsler-Cohen
		AIA696	MetaCipher: A Time-Persistent and Universal Multi-Agent Framework for Cipher-Based Jailbreak Attacks for LLMs	Boyuan Chen; Minghao Shao; Abdul Basit; Siddharth Garg; Muhammad Shafique
Friday, Jan 23, 11:00am-12:00pm	Preference-Based Methods and	AIA102	AMaPO: Adaptive Margin-attached Preference Optimization for Language Model Alignment	Ruibo Deng; Duanyu Feng; Wenqiang Lei

12:00pm	Methods and Evaluation for Model Alignment (AIA) Conference Room J	AIA160	Misalignment from Treating Means as Ends	Henrik Marklund; Alex Infanger; Benjamin Van Roy
		AIA410	Mind the Gap: Quantifying and Aligning Human-AI Visual Attention for Accident Anticipation	Hoe Sung Ryu; Christian Wallraven
		AIA570	DETONATE – a Benchmark for Text-to-Image Alignment and Kernelized Direct Preference Optimization	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
		AIA752	GEM: Generative Entropy-Guided Preference Modeling for Few-Shot Alignment of LLMs	Yiyang Zhao; Huiyu Bai; Xuejiao Zhao
Saturday, Jan 24, 9:30-10:30am	Benchmarks for Robust, Secure, and Trustworthy LLM Systems (AIA) Conference Room J	AIA114	Benchmarking Trustworthiness in Multimodal LLMs for Video Understanding	Youze Wang; Zijun Chen; Ruoyu Chen; Shishen Gu; Wenbo Hu; Jiayang Liu; Yinpeng Dong; Hang Su; Jun Zhu; Meng Wang; Richang Hong
		AIA391	MCA-Bench: A Multimodal Benchmark for Evaluating CAPTCHA Robustness Against VLM-based Attacks	Zonglin Wu; Yule Xue; Yaoyao Feng; Xiaolong Wang; Yiren Song
		AIA495	DNR Bench: Benchmarking Over-Reasoning in Reasoning LLMs	Oluwanifemi Bamgbose; Masoud Hashemi; Sathwik Tejaswi Madhusudhan; Jishnu Sethumadhavan Nair; Aman Tiwari; Vikas Yadav
		AIA559	MedAtlas: Evaluating LLMs for Multi-Round, Multi-Task Medical Reasoning Across Diverse Imaging Modalities and Clinical Text	Ronghao Xu; Zhen Huang; Yangbo Wei; Xiaoqian Zhou; Zikang Xu; Ting Liu; Zihang Jiang; S. Kevin Zhou
		AIA588	Towards Benchmarking Privacy Vulnerabilities in Selective Forgetting with Large Language Models	Wei Qian; Chenxu Zhao; Yangyi Li; Mengdi Huai
Saturday, Jan 24, 11:00am-12:00pm	Evaluating Alignment and Behavioral Properties of LLMs (AIA) Conference Room J	AIA294	Silenced Biases: The Dark Side LLMs Learned to Refuse	Rom Himelstein; Amit LeVi; Brit Youngmann; Yaniv Nemcovsky; Avi Mendelson
		AIA528	A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs	Trenton Chang; Tobias Schnabel; Adith Swaminathan; Jenna Wiens
		AIA594	On the Alignment of Large Language Models with Global Human Opinion	Yang Liu; Masahiro Kaneko; Chenhui Chu
		AIA653	Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy	Alexander Duffy; Samuel J Paech; Ishana Shastri; Elizabeth Karpinski; Baptiste Alloui-Cros; Tyler Marques; Matthew Lyle Olson
		AIA47	Bolster Hallucination Detection via Prompt-Guided Data Augmentation	Wenyun Li; Zheng Zhang; Dongmei Jiang; Xiangyuan Lan
Saturday, Jan 24, 3:00-4:00pm	Attacks and Defenses in LLM Safety Alignment (AIA) Conference Room J	AIA423	Differentiated Directional Intervention: A Framework for Evading LLM Safety Alignment	Peng Zhang; Peijie Sun
		AIA558	Not All Tokens Are Meant to Be Forgotten	Xiangyu Zhou; Yao Qiang; Saleh Zare Zade; Douglas Zytko; Prashant Khanduri; Dongxiao Zhu
		AIA602	Chain-of-Thought Driven Adversarial Scenario Extrapolation for Robust Language Models	Md Rafi Ur Rashid; Vishnu Asutosh Dasu; Ye Wang; Gang Tan; Shagufta Mehnaz
		AIA637	AdvBDGen: A Robust Framework for Generating Adaptive and Stealthy Backdoors in LLM Alignment	Pankayaraj Pathmanathan; Udari Madhushani Sehwag; Michael-Andrei Panaitescu-Liess; Cho-Yu Jason Chiang; Furong Huang
		AIA667	EASE: Practical and Efficient Safety Alignment for Small Language Models	Haonan Shi; Guoli Wang; Tu Ouyang; An Wang
Sunday, Jan 25, 9:30-10:30am	Interpretability and Internal Representations in Deep	AIA156	Beyond Patches: Mining Interpretable Part-Prototypes for Explainable AI	Mahdi Alehdaghi; Rajarshi Bhattacharya; Pourya Shamsolmoali; Rafael M. O. Cruz; Eric Granger
		AIA267	SL-CBM: Enhancing Concept Bottleneck Models with Semantic Locality for Better Interpretability	Hanwei Zhang; Luo Cheng; Rui Wen; Yang Zhang; Lijun Zhang; Holger Hermanns

	Models (AIA) Conference Room J	AIA352	Beyond Transcription: Mechanistic Interpretability in ASR	Neta Glazer; Yael Segal-Feldman; Hilit Segev; Aviv Shamsian; Asaf Buchnick; Gill Hetz; Ethan Fetaya; Joseph Keshet; Aviv Navon
		AIA397	Explainable Melanoma Diagnosis with Contrastive Learning and LLM-based Report Generation	Junwen Zheng; Xinran Xu; Li Rong Wang; Chang Cai; Lucinda Siyun Tan; Dingyuan Wang; Hong Liang Tey; Xiuyi Fan
		AIA586	Quiet Feature Learning in Algorithmic Tasks	Prudhviraj Naidu; Zixian Wang; Leon Bergen; Ramamohan Paturi
Sunday, Jan 25, 11:00am-12:00pm	Foundations and Methods for AI Safety Alignment (AIA) Conference Room J	AIA196	Intrinsic Barriers and Practical Pathways for Human–AI Alignment: An Agreement-Based Complexity Analysis	Aran Nayebi
		AIA238	DAVSP: Safety Alignment for Large Vision-Language Models via Deep Aligned Visual Safety Prompt	Yitong Zhang; Jia Li; Liyi Cai; Ge Li
		AIA325	STAR-1: Safer Alignment of Reasoning LLMs with 1K Data	Zijun Wang; Haoqin Tu; Yuhan Wang; Juncheng Wu; Yanqing Liu; Jieru Mei; Brian R. Bartoldson; Bhavya Kailkhura; Cihang Xie
		AIA506	Realist and Pluralist Conceptions of Intelligence and Their Implications on AI Research	Ninell Oldenburg; Ruchira Dhar; Anders Søgaard
		AIA583	Requirements for Aligned, Dynamic Resolution of Conflicts in Operational Constraints	Steven J. Jones; Robert E. Wray; John E. Laird