PlurVA-LLM 2026

The First Workshop on Pluralistic Value Alignment of LLMs

Proceedings of the Workshop

November 10, 2026

The PlurVA-LLM organizers gratefully acknowledge the support from the following sponsors.

Sponsors





©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-453-8

Introduction

The First Workshop on Pluralistic Value Alignment of LLMs (PlurVA-LLM), held in conjunction with AACL-IJCNLP 2026 on November 10, 2026, in Hengqin, China, provides an interdisciplinary forum for research on aligning large language models (LLMs) with diverse human values. Its research agenda connects three complementary directions: theoretical foundations and representations of pluralistic values, alignment methods across training and inference, and context-sensitive evaluation and applications. Additional themes include human–AI collaborative data construction, interpretability, multimodal alignment, and the inclusion of low-resource languages and underrepresented cultures.

The accepted contributions examine pluralistic alignment through concrete questions about data, language and model behaviour. They include a dataset for studying regional biases in English and code-mixed social media comments, an analysis of language-conditioned ideological divergence in LLM interpretations of contested political documents, and a study of how training-data structure relates to the forms of pluralism retained after fine-tuning. These contributions bring together resource construction, empirical analysis and application-oriented evaluation, examining pluralistic value alignment in specific modelling and deployment settings. They collectively give the programme a concrete focus on how value differences are represented in data, expressed in model outputs, and affected by training choices and the contexts in which models are used.

We finally received 36 direct submissions and two submissions through ACL Rolling Review (ARR), accepting 19 direct submissions (four short and 15 long, including 14 archival and five non-archival) and two ARR submissions (one long archival and one short archival), corresponding to an acceptance rate of 55.26%. Additionally, we received 12 already accepted submissions.

A distinctive component of PlurVA-LLM is its shared task on locally grounded value judgments across China, Indonesia and Sri Lanka. The task brings together benchmarks that reflect different cultural contexts and value frameworks, with Resource-Constrained and Open Tracks that accommodate different computational resources and system designs. It attracted 20 registered teams from six countries, with five teams completing valid final submissions, all outperforming the organizer-provided baseline. The task provides a concrete setting for investigating cross-cultural alignment and the challenges of evaluating systems across heterogeneous value contexts.

The program features four keynote speakers: Rostam J. Neuwirth (University of Macau), Pak-Hang Wong (Hong Kong Baptist University), Nancy F. Chen (A*STAR, Singapore), and Roy Ka-Wei Lee (University of British Columbia). Their perspectives span AI law and governance, intercultural ethics, socially grounded AI, and pluralistic approaches to LLM safety. Together with contributed presentations and shared-task presentations, these talks connect methodological advances with broader questions about whose values models represent and how alignment decisions should be justified.

We thank our authors, program committee, reviewers, invited speakers, student co-organizers, sponsors, and the AACL-IJCNLP 2026 organizing team for making this first edition possible. We warmly welcome all participants and hope that the workshop will foster sustained collaboration toward language technologies that better understand and respect the plurality of human values.

Co-Organizers: Deyi Xiong, António Branco, Hongming Zhang, Yue Dong, Benyou Wang, Wenxuan Zhang, Li Zhou, and Jingting Zheng

Organizing Committee

Program Chairs

Deyi Xiong, Tianjin University
António Branco, University of Lisbon
Hongming Zhang, Macau University of Science and Technology
Yue Dong, University of California, Riverside
Benyou Wang, The Chinese University of Hong Kong, Shenzhen
Wenxuan Zhang, Singapore University of Technology and Design
Li Zhou, The Chinese University of Hong Kong, Shenzhen
Jingting Zheng, Tianjin University

Student Co-organizers

Jianxiang Peng, Tianjin University
Junru Wu, Tianjin University
Liping Yang, Tianjin University
Nethmi Muthugala, Tianjin University
Siyi Hao, Tianjin University
Supryadi, Tianjin University
Aoxiang Qin, The Chinese University of Hong Kong, Shenzhen
Geng Zhao, The Chinese University of Hong Kong, Shenzhen
Zichen Xu, The Chinese University of Hong Kong, Shenzhen

Program Committee

Adnan Al Ali, Charles University Prague
Edward Y Chang, Computer Science Department, Stanford University
Inseong Choi, Dongguk University
Scott E. Friedman, SIFT
Yu Fu, University of California, Riverside
Fengyu Guo, Tianjin Normal University
Jongwook Han, Seoul National University
Daniel Hershcovich, University of Copenhagen
Amir Homayounirad, Delft University of Technology
Hongjin Kim, Electronics and Telecommunications Research Institute
Alina Landowska, Kozminski University and USWPS
Hongliang Li, Xiamen University
Jingxuan Li, Northwestern University
Yang Liu, Kyoto University
JiaTang Luo, University of the Chinese Academy of Sciences
Nailia Mirzakhmedova, Bauhaus Universität Weimar
Nethmi Muthugala, Tianjin University
Malvina Nikandrou, University of Edinburgh
Gianluca De Ninno, University of Pisa
Ike Obi, Purdue University
Jianxiang Peng, Tianjin University
Irina Proskurina, Université Jean Monnet
Xin Quan, University of Manchester
Yuqi Ren, Tianjin University
Naama Rozen, Tel Aviv University
Maciej Skorski, University of Luxemburg and University of Warsaw
Woojung Song, Seoul National University
Ao Sun, Southeast University
Supryadi, Tianjin University
Mingjian Tang, University of Technology Sydney
Ayla Rigouts Terryn, Université de Montréal
Songsheng Wang, Tianjin Normal University
Junru Wu, Tianjin University
Lexing Xie, Australian National University
Hefei Xu, Hefei University of Technology
Shaoyang Xu, Singapore University of Technology and Design
Zichen Xu, The Chinese University of Hong Kong
Liping Yang, Tianjin University
Yuning Yang, Amazon
Jaeyoun You, Hanyang Universty
Tongxuan Zhang, Tianjin Normal University
Geng Zhao, Medizinische Universität Innsbruck
Jingting Zheng, Tianjin University

Keynote Talk

Pluralism and the Paradox of AI Alignment: Rethinking Human Intelligence in the Age of AI

Rostam J. Neuwirth
University of Macau

Abstract: AI alignment seeks to ensure that intelligent systems act in accordance with human values and societal goals. Yet alignment raises a deeper question: what exactly is the human intelligence against which AI is measured? As AI systems become increasingly capable of learning, adapting, and making autonomous decisions, intelligence itself emerges as a contested and plural concept. This keynote explores the bi-directional nature of human-AI alignment: while AI must adapt to human values, humans must also develop the literacy and cognitive skills needed to work effectively with AI. This creates a paradox: how do we cultivate forms of intelligence that are only now emerging?

Bio: Rostam J. Neuwirth is Distinguished Professor of Law and Head for Department of Global Legal Studies at the University of Macau. Previously, he taught at the West Bengal University of Juridical Sciences (NUJS) in Kolkata and the Hidayatullah National Law University (HNLU) in Raipur (India) and worked as a legal adviser in the Department of European Law of the International Law Bureau of the Austrian Federal Ministry for Foreign Affairs. He received his PhD degree from the European University Institute (EUI) in Florence (Italy) and also holds a master's degree in law (LLM) from the Faculty of Law of McGill University in Montreal (Canada). As an undergraduate he studied at the University of Graz (Austria) and the Université d'Auvergne (France). He is the author of The EU Artificial Intelligence Act: Regulating Subliminal AI Systems (Routledge 2023), which was also translated into Chinese, as well as Law in the Time of Oxymora: A Synaesthesia of Language, Logic and Law (Routledge 2018). He has also published extensively in international journals on contemporary global legal issues, exploring the intrinsic linkages between law, on the one hand, and language, cognition, art, culture, society, and technology, on the other.

Keynote Talk

Whose Values, Whose Safety? Building LLM Safety for a Pluralistic World

Roy Ka-Wei Lee
University of British Columbia

Abstract: Large language models increasingly mediate how people communicate, access information, and encounter harmful content online. Yet many current approaches to alignment and safety still assume relatively uniform preferences, cultural norms, or definitions of harm. In practice, societies are pluralistic: people differ in their values, languages, social identities, and interpretations of what is harmful or acceptable.

Drawing on recent work on subgroup value alignment, distributional evaluation, culturally grounded interaction, low-resource language safety, and online harm detection, this keynote examines what pluralism means for LLM safety. I will discuss how models can better represent legitimate differences within and across cultures, how safety should be evaluated when there is no single agreed-upon norm, and how current systems may systematically underserve less represented communities. I will conclude by outlining a research agenda for building LLMs that are not only safer, but also more equitable and responsive to the diverse societies in which they are deployed.

Bio: Roy Ka-Wei Lee is Associate Professor in the Department of Computer Science at the University of British Columbia. His work sits at the intersection of responsible AI, social computing, NLP, online safety, hate speech, and low-resource LLM safety. His research highlights cyber social safety and NLP/social computing, and his recent speaking record includes keynotes on multimodal AI for low-resource/social-impact contexts and AI for safer online spaces. His group also has recent work on safety alignment in low-resource English languages, including a best-paper recognition at the AAAI 2026 LM4UC workshop.

Keynote Talk

Grounding AI in Society

Nancy F. Chen
Center for Frontier AI Research, A*STAR

Abstract: I study communication across human–human, human–machine, and human–AI contexts, with particular attention to how expectations of communication shift across them. We expect humans to understand context, intent, emotion, and social norms; machines to be predictable, efficient, and reliable; and AI systems increasingly to satisfy both sets of expectations. This convergence creates new questions about trust, agency, accountability, and what constitutes meaningful communication and collaboration between humans and AI.

Against this backdrop, my work develops computational models of human communication and context, builds socially grounded technologies for trustworthy human–AI interaction, and investigates how language-model use reshapes learning, reasoning, and professional expertise. The unifying goal is to augment human capability and potential while preserving human agency and cognitive resilience. I illustrate this research agenda through case studies spanning healthcare, education, eldercare, collaborative learning, and digital services.

Bio: Nancy F. Chen is a Lead Principal Investigator and leads the Ethical and Trustworthy AI pillar at Center for Frontier AI Research, A*STAR. Her work focuses on building human-centered frontier AI that can reason, communicate, and act responsibly in diverse social, cultural, and linguistic contexts. She works across trustworthy and agentic AI, multimodal and multilingual foundation models, speech and conversational intelligence, and human–AI collaboration. Her research has increasingly explored how AI systems understand not only language and perception, but also people, values, emotions, culture, and social context. She has led work on culturally grounded AI, multimodal reasoning, AI safety and evaluation, educational AI, and socially intelligent interactive systems, with applications spanning public-sector and real-world deployments in Singapore. Chen received her PhD at MIT and Harvard. She is an ISCA Fellow and AAIA Fellow and has held leadership roles at major AI conferences, including serving as Program Chair of NeurIPS 2025 and ICLR 2023.

Keynote Talk

Pluralistic Alignment Always, Sometimes, or Never? Taking Value Pluralism Seriously

Pak-Hang Wong
Hong Kong Baptist University

Abstract: In an influential account of pluralistic alignment, Sorensen and colleagues argue that there are diverse individual goals, intentions, values, and perspectives, and that this diversity entails the need for AI systems to be capable of representing a diverse set of human values and perspectives. This research agenda is motivated, at least in part, by value pluralism associated with Isaiah Berlin and others, which holds that some fundamental human values can conflict and be incommensurable. However, relatively little work in pluralistic alignment has explored what follows if value pluralism is taken seriously, in particular what "plurality" or "pluralistic" means for pluralistic alignment.

More specifically, taking value pluralism seriously requires researchers to distinguish diversity in observed preferences or perspectives from disagreements that are irreducible, reasonable, or normatively worthy of preservation. Indeed, as a recent Roadmap by Poole-Dayan and colleagues points out, the fact that queries are subjective, open-ended, or value-laden does not in itself warrant pluralistic consideration; pace Poole-Dayan, I contend that user studies on when they think pluralistic answers are warranted do not by themselves address a prior normative problem, namely whether the diversity (or disagreement) is reasonable rather than the result of error, ignorance, bias, or merely heterogeneous preference—or, more generally, whether it constitutes a normatively worthy disagreement that deserves pluralistic consideration. Relatedly, there is an additional question concerning when such reasonable disagreement should affect the behaviour of an AI system.

Drawing on Rawls's idea of overlapping consensus as a response to the fact of reasonable pluralism, I argue that taking pluralism seriously does not require every decision—or decisions at every level—to be pluralistic. Accordingly, Rawls illustrates one possible way to draw boundaries around when pluralistic consideration is warranted. Moreover, building on work on the epistemology of disagreement, scientific pluralism, and moral disagreement, I identify conditions under which disagreement has a legitimate claim to persistence. These conditions, in turn, will allow us to decide which queries deserve a pluralistic treatment, and when. In short, I propose an account of warranted pluralistic alignment: pluralistic treatment should not be activated merely by diversity, but by disagreement that satisfies specific epistemic and normative conditions, as well as other relevant considerations of whether the disagreement should remain open in the particular context of use.

In short, my objective is to propose an account of pluralistic alignment that avoids over-pluralisation and under-pluralisation. Yet, insofar as whether disagreement remains legitimately open often depends on roles, jurisdictions, epistemic standards, authoritative rules, and prior institutional arrangements, I conclude by proposing an institutional approach to pluralistic alignment in which normative and epistemic structures play an important role in determining what should be settled, what should remain plural, and when pluralistic treatment is appropriate.

Bio: Pak-Hang Wong is a philosopher and ethicist of technology based in the Academy of Chinese, History, Religion and Philosophy and a Research Fellow at the Centre for Applied Ethics at Hong Kong Baptist University. His research explores the social, ethical, and political dimensions of AI and emerging digital technologies, particularly from an intercultural perspective. He is Editor-in-Chief (Ethics, AI & Data) of Ethics and Society and Associate Editor (Epistemology and Technology) of Philosophy & Technology. He co-edited Harmonious Technology: A Confucian Ethics of Technology, which offers an alternative, non-Western perspective on the philosophy and ethics of technology.

Table of Contents

StealthRL: Reinforcement Learning Paraphrase Attacks for Multi-Detector Evasion of AI-Text Detectors
Suraj Ranganath and Atharv Ramesh 1

MasalBench: A Benchmark for Contextual and Cross-Cultural Understanding of Persian Proverbs in LLMs
Ghazal Kalhor and Behnam Bahrak 17

Performative Alignment: Preference Learning under Deployment-Dependent Utility
Tarun Chinta 24

IndRegBias: A Dataset for Studying Indian Regional Biases in English and Code-Mixed Social Media Comments
Debasmita Panda, Akash Anil and Neelesh Kumar Shukla 43

EduValueBench: Benchmarking Situated Value Alignment of LLMs in Chinese Educational Decision-Making
Junru Wu, Yuqi Ren, Yongqi Leng, Zhuowen Han, Xiaoyang Gong, Zheyu Liu, Ruwei Jin, Cong Xiao, Wenjie Tao, Xiao Liu, Lei Guan, Zhaoxia Zhang, Dezhi Tong and Deyi Xiong 63

Concepts at Work: Testing Rival Conceptions of Liberty in Large Language Models
Mark J. Hill 78

PlurVA-LLM-2026 Shared Task Track-2: PluriGuide—Principle Distillation and Identifier-Bound Demonstrations for Pluralistic Value Alignment
Jingyi Kang, Junyu Lu, Bo Xu and Hongfei Lin 93

PlurVA-LLM-2026 Shared Task Track-1 and 2: Preference Learning for Pluralistic Value Alignment
Joydeb Mondal 98

The Language You Ask In: Language-Conditioned Ideological Divergence in LLM Analysis of Contested Political Documents
Oleg Smirnov 104

PlurVA-LLM-2026 Shared Task Track-1: Pluralistic Value Alignment in LLMs via Multilingual Fine-Tuning and Threshold Calibration
Vihindi Kotalawala and Nevidu Jayatilleke 143

Findings of the PlurVA-LLM 2026 Shared Task on Pluralistic Value Alignment Across China, Indonesia, and Sri Lanka
Supryadi, Nethmi Muthugala, Siyi Hao, Aoxiang Qin, Jingting Zheng, Li Zhou, Wenxuan Zhang, Benyou Wang, Hongming Zhang, António Branco and Deyi Xiong 154

C-VoicesCompass: Evaluating Value Alignment of LLMs with Chinese Social Values
Yuemei Xu and Han Zhang 165

Country Name or Cultural Evidence? A Causal Audit of Population-Grounded Steering in Large Language Models
Ziyan Yang and Kuo Zhao 176

PlurVA-LLM-2026 Shared Task Track-1: Curriculum Fine-Tuning with Soft Labels and Positional Consistency for Cross-Cultural Value Alignment
Amasha Widanagamage, Hansika Seneviratne, Naween Fernando, Dilushri Pavithra, Ruvan Weerasinghe and Deshan Sumanathilaka 187

NepPlural: A Human-Anchored Benchmark for Persona Annotation in Nepali Migration Discourse
Jinu Nyachhyon, Anish Khatiwada and Bimal Timilsina 197

CompOrca: Corpus-Scale Compliance Labelling of Instruction-Tuning Data
Philipp E. Glass and Alina Miron 211

Program

Tuesday, November 10, 2026

09:00 - 09:05 *Welcome Remarks*

09:05 - 09:40 *Keynote Talk 1: Pluralism and the Paradox of AI Alignment: Rethinking Human Intelligence in the Age of AI (Rostam J. Neuwirth, University of Macau)*

09:40 - 10:30 *Oral Session 1*

1. *EduValueBench: Benchmarking Situated Value Alignment of LLMs in Chinese Educational Decision-Making*
2. *Denser, Not Wider: Training-Data Structure Selects the Kind of Pluralism a Fine-Tuned Model Retains*
3. *Effects of Post-Training on Cultural Anchoring in LLMs*
4. *Guided Relational Analysis Improves Social Intelligence in LLMs*

10:30 - 11:00 *Coffee Break*

11:00 - 11:35 *Keynote Talk 2: Whose Values, Whose Safety? Building LLM Safety for a Pluralistic World (Roy Ka-Wei Lee, University of British Columbia)*

11:35 - 12:30 *Oral Session 2*

1. *StealthRL: Reinforcement Learning Paraphrase Attacks for Multi-Detector Evasion of AI-Text Detectors*
2. *Performative Alignment: Preference Learning under Deployment-Dependent Utility*
3. *Country Name or Cultural Evidence? A Causal Audit of Population-Grounded Steering in Large Language Models*
4. *NepPlural: A Human-Anchored Benchmark for Persona Annotation in Nepali Migration Discourse*

12:30 - 14:00 *Lunch*

14:00 - 14:35 *Keynote Talk 3: Grounding AI in Society (Nancy F. Chen, Center for Frontier AI Research, A*STAR)*

14:35 - 15:30 *Oral Session 3*

1. *Findings of the PlurVA-LLM 2026 Shared Task on Pluralistic Value Alignment Across China, Indonesia, and Sri Lanka*
2. *PlurVA-LLM-2026 Shared Task Track-1 and 2: Preference Learning for Pluralistic Value Alignment*
3. *PlurVA-LLM-2026 Shared Task Track-1: Pluralistic Value Alignment in LLMs via Multilingual Fine-Tuning and Threshold Calibration*
4. *MasalBench: A Benchmark for Contextual and Cross-Cultural Understanding of Persian Proverbs in LLMs*

15:30 - 16:00 *Coffee Break*

16:00 - 16:35 *Keynote Talk 4: Pluralistic Alignment Always, Sometimes, or Never? Taking Value Pluralism Seriously (Pak-Hang Wong, Hong Kong Baptist University)*

16:35 - 17:35 *Poster Session*

17:35 - 17:40 *Closing Session*