CAIT2026 Special Session IV

Special Session Basic Information:

Session Title	中文：大语言模型的推理、对齐与可信部署 英文：Reasoning, Alignment, and Trustworthy Deployment of Large Language Models

Introduction and topics

中文： 大语言模型（LLM）已从文本生成工具迅速演变为能执行复杂多步推理、工具调用与自主决策的通用智能引擎。随着 LLM 被广泛部署于医疗、教育、法律分析和软件工程等高风险领域，确保其推理准确性、价值对齐能力与部署可信性，已成为人工智能研究的核心挑战。
本专题会议诚邀围绕 LLM 推理与可信部署全链路的原创性研究成果，涵盖：思维链（Chain-of-Thought）与多智能体推理架构、基于人类反馈的强化学习（RLHF）与宪法 AI 对齐方法、幻觉抑制与事实锚定机制，以及可解释性推理等核心议题。本 Session 旨在弥合基础模型研究与现实工程实践之间的鸿沟，促进自然语言处理、AI 安全与系统工程领域研究者之间的跨学科交流。

英文： Large language models (LLMs) have rapidly evolved from text-generation tools into general-purpose reasoning engines capable of complex multi-step problem solving, tool use, and autonomous decision-making. As LLMs are deployed in high-stakes domains - including healthcare, education, legal analysis, and software engineering - ensuring their reasoning fidelity, value alignment, and operational trustworthiness has become a central challenge for the AI research community.
This special session invites original research contributions addressing the full pipeline of LLM reasoning and deployment: from chain-of-thought and multi-agent reasoning architectures, to reinforcement learning from human feedback (RLHF) and Constitutional AI for alignment, to hallucination mitigation, factual grounding, and interpretable inference. The session aims to bridge foundational model research with real-world engineering practice, fostering interdisciplinary exchange between NLP researchers, AI safety scholars, and system engineers.

Special Session Chair(s):

姓名 **Name**	陈光 Guang Chen
职称 **Prefix**	副教授 Associate Professor
部门 **Department**	人工智能学院 School of Artificial Intelligence
单位 **Organization**	北京邮电大学 Beijing University of Posts and Telecommunications
城市/地区 **City/Region**	北京 Beijing, China
Email	chenguang@bupt.edu.cn

Organizer's Brief Biography

中文： 陈光，现任北京邮电大学人工智能学院副教授，曾作为核心骨干深度参与包括国家自然科学基金在内的多项国家级科研项目，在国际顶尖期刊及学术会议上发表了多篇高水平论文，并编著有专业教材和畅销书籍。其主要研究方向为机器学习和语言智能。

英文： Guang Chen is an Associate Professor at the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He has participated in multiple national-level projects, including the National Natural Science Foundation of China, as a key researcher. He has published numerous papers in leading journals and international conferences and authored a textbook. His research interests include Machine Learning and Language AI.