



How Transformers Learn to Generalize, Adapt, and Reason

SPEAKER Wei Huang

Research Scientist, Deep Learning
Theory Team
RIKEN Center for Advanced
Intelligence Project (AIP), Tokyo,
Japan Researcher The Institute of
Statistical Mathematics, Tachikawa,
Tokyo, Japan

DATE 15 Oct, 2026 (Thu)

TIME 2:00 PM - 3:00 PM

VENUE CS Seminar Room, Y6405, 6th Floor,
Yellow Zone, Yeung Kin Man Academic
Building, City University of Hong Kong, 83
Tat Chee Avenue, Kowloon Tong

ABSTRACT

Transformers can fit complex data, adapt their behavior from context, and solve tasks that require planning. What aspects of the data and training process make these capabilities possible? In this talk, I will present three studies that connect data, optimization dynamics, and the mechanisms learned by Transformers. First, I will discuss benign overfitting in a signal-plus-noise model, showing how the interaction between attention and value learning leads to distinct generalization regimes governed by sample size and signal strength. Second, I will examine in-context learning through task-vector retrieval, explaining how training data organization determines whether a model extracts transferable task information or retains interfering low-level features. Third, I will study planning through multi-token prediction, showing how future-token objectives can provide learning signals for intermediate computations, supported by controlled experiments and a staged training analysis in a minimal Transformer. Together, these results highlight a common question: which representations and computations does training select, and when do they transfer beyond the training examples? I will conclude with a research agenda on learned task structure and reliable reasoning in modern AI systems.

BIOGRAPHY

Wei Huang is a Research Scientist at the RIKEN Center for Advanced Intelligence Project (AIP) and a Researcher at the Institute of Statistical Mathematics, Japan. He received his Ph.D. in Computer Science from the University of Technology Sydney in 2021, following bachelor's and master's degrees in physics at the University of Science and Technology of China. His research focuses on the theoretical foundations of deep learning, particularly optimization, generalization, in-context learning, and reasoning in foundation models. He received the CPAL 2025 Rising Star Award and was selected for the KAUST Rising Stars in AI Symposium in 2025. He is the principal investigator of a JST BOOST project on efficient and safe foundation models, and his research has also been supported by a JSPS Grant-in-Aid for Early-Career Scientists. He serves as an Action Editor for Transactions on Machine Learning Research (TMLR) and Program Chair of CPAL 2027, and has served as an Area Chair for NeurIPS, ICML, ICLR, and AISTATS.

All are welcome!



In case of questions, please contact Prof WONG Hau San and Prof ZHANG Qingfu at cshswong@cityu.edu.hk and qingfu.zhang@cityu.edu.hk, or visit the CS Departmental Seminar Web at <https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums>.