

Koshiro Aoki

M.S. Student in Computer Science, Waseda University

✉ aokikoshiro@akane.waseda.jp

🌐 aokikoshiro.github.io

🐙 AokiKoshiro

🌐 koshiro-aoki

Research Interests

(Mechanistic) interpretability | Introspection | Metacognition | NeuroAI

Education

Waseda University

M.S. in Computer Science
Adviser: Daisuke Kawahara

Apr. 2026 - Present

📍 Tokyo, Japan

Waseda University

B.S. in Computer Science
GPA: 3.8/4. Dean's Award.

Apr. 2022 - Mar. 2026

📍 Tokyo, Japan

Work Experience

Waseda University

Research Assistant

Research on differences between the internal representations of the human brain and language models as part of a JST CREST project.

Apr. 2026 - Present

📍 Tokyo, Japan

The University of Tokyo

Research Contractor

Research on interpretability and defenses of adversarial attacks on robot foundation models.

May 2025 - Present

📍 Tokyo, Japan

Research and Development Center for Large Language Models, National Institute of Informatics (NII)

Technical Assistant

Research on bottom-up interpretability of training dynamics in large language models.

May 2025 - Mar. 2026

📍 Tokyo, Japan

Publications

Conference

Bottom-Up Interpretability of Pretraining Dynamics via Loss Curve Decomposition

2026

Koshiro Aoki, Masaru Isonuma, Yusuke Oda, Hirokazu Kiyomaru, Takashi Kodama, Chaoran Liu, Yohei Oseki, Yusuke Miyao, Max Müller-Eberstein, Daisuke Kawahara

Third Conference on Language Modeling (COLM 2026)

Workshop

Crosscoders Identify Shared or Specific Features between the Human Brain and Language Models

2026

Koshiro Aoki, Itsuki Hamada, Naho Orita, Daisuke Kawahara, Hiromu Sakai

Mechanistic Interpretability Workshop @ ICML 2026

In-Context Neurofeedback: Can Large Language Models Control Their Internal Representations through Privileged Access?

2026

Koshiro Aoki, Ryota Takatsuki, Gouki Minegishi, Yusuke Haruki, Daisuke Kawahara

Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026

Spotlight

Evaluating the Impact of SAE-based Language Steering on LLM Performance

2026

Sebastian Zwirner, Wentao Hu, Koshiro Aoki, Daisuke Kawahara

EACL 2026 Student Research Workshop

Testing Simulation Theory in LLMs' Theory of Mind Koshiro Aoki, Daisuke Kawahara <i>IJCNLP-AACL 2025 Student Research Workshop</i>	2025
---	------

Oral

Domestic Journal

Mechanistic Interpretability: A New Trend in Interpretability Research Koshiro Aoki, Ryota Takatsuki, Gouki Minegishi, Hakaze Cho <i>Transactions of the Japanese Society for Artificial Intelligence</i>	2026
--	------

Best Paper Award

Honors and Awards

Spotlight Paper <i>Trustworthy AI for Good (AI4GOOD) Workshop @ ICML 2026</i>	2026
Best Paper Award <i>The Japanese Society for Artificial Intelligence 40th Anniversary Commemorative Paper Award</i>	2026
Dean's Award for Excellence <i>School of Fundamental Science and Engineering, Waseda University</i>	2026
Young Researcher Encouragement Award <i>The 32nd Annual Meeting of the Association for Natural Language Processing</i>	2026
Oral Presentation <i>IJCNLP-AACL 2025 Student Research Workshop</i>	2025
Encouragement Award <i>The 20th Young Researcher Association for NLP Studies</i>	2025
Sponsored Award <i>The 31st Annual Meeting of the Association for Natural Language Processing</i>	2025