Hakaze Cho / Yufeng Zhao / 趙羽風

yfzhao@jaist.ac.jp | hakaze-c.com | hc495 | Born in 1999, Beijing, China

Resume

RIKEN AIP 2026.4 – 2029.3
Special Postdoctoral Research Fellow Sendai, Miyagi, Japan
Principal Investigator of Project: Towards Mechanistic Controllability: Circuit-based Behavior Correction for Large Language Models
Mentor: Prof. Kentaro Inui

Tohoku University 2026.4 – 2029.3
Academic Researcher Sendai, Miyagi, Japan
Mentor: Prof. Kentaro Inui

Japan Advanced Institute of Science and Technology 2023.10 – 2026.3
Ph.D. in Computer Science (GPA: 3.8/4); Research Assistant Nomi, Ishikawa, Japan
Major Courses: Advanced Machine Learning; Advanced Natural Language Processing
Dissertation Title: "The Mechanistic Basis of In-context Learning"
Mentor: Associate Prof. Naoya Inoue

Beijing Institute of Technology 2021.9 – 2023.6
M.Eng. in Software Engineering (GPA: 3.5/4) Haidian, Beijing, China
Major Courses: Numerical Analysis; Software Theory; Data Engineering
Thesis Title (Translated): "Fine-tuning with Randomly Initialized Downstream Network: Finding a Stable Convex-loss Region in Parameter Space"

Beijing Institute of Technology 2017.8 – 2021.6
B.Eng. in Material Chemistry (GPA: 3.18/4) Haidian, Beijing, China
Major Courses: Calculus; Linear Algebra; Probability and Statistics; Basic Physics; (Inorganic / Organic / Physical / Analytical) Chemistry; Chemistry Experiments; C Language Programming
Thesis Title (Translated): "Synthesis and Self-Assembly of Aggregation-induced Emission Compounds"
Mentor: Associate Prof. Jianbing Shi

Research Activities

Research Interests: Representation Learning, Mechanistic Interpretability, In-context Learning

- **Interpretability for Artificial Neural Network:** Mechanistic Interpretability (especially for Transformer)
- **Controllability for Artificial Neural Network:** Low-resource Model Behavior Improvement / Controlling from Mechanistic Perspective
- **Misc.:** Manifold Learning, Low-precision Neural Networks, Neural Network Training Dynamics

Peer-review: NeurIPS 2025; ICLR 2025, 2026; ICML 2026; ACL Rolling Review: 2025 May, Jun, Oct, 2026 Jan; ICML 2025 Actionable Interpretability Workshop; ACL 2025 Student Research Workshop.

Affiliated Society: The Japanese Association for Natural Language Processing; The Japanese Society for Artificial Intelligence; Association for Computational Linguistics.

Selected Publications

Hakaze Cho has authored over 30 publications, with a selection of notable works listed below:

- **Mechanism of Task-oriented Information Removal in In-context Learning** 2026
  Hakaze Cho, Haolin Yang, Gouki Minegishi, Naoya Inoue
  *The Thirteenth International Conference on Learning Representations (**ICLR**)* (h5=362)
- **Revisiting In-context Learning Inference Circuit in Large Language Models** 2025
  Hakaze Cho, Mariko Kato, Yoshihiro Sakai, Naoya Inoue
  *The Thirteenth International Conference on Learning Representations (**ICLR**)* (h5=362)
- **Token-based Decision Criteria Are Suboptimal in In-context Learning** 2025
  Hakaze Cho, Yoshihiro Sakai, Mariko Kato, Kenshiro Tanaka, Akira Ishii, Naoya Inoue
  In Proceedings of *the 2025 Annual Conference of NAACL **(NAACL main conference)*** (h5=126)
- **Understanding Token Probability Encoding in Output Embeddings** 2025
  Hakaze Cho, Yoshihiro Sakai, Kenshiro Tanaka, Mariko Kato, Naoya Inoue
  In Proceedings of *the 31st International Conference on Computational Linguistics (**COLING**)* (h5=81)

Awards

- **Outstanding Paper (優秀賞)** (14 in 765)
  The 31st Annual Conference of the Japanese Association for Natural Language Processing. 2025
- **Research Award for Young Scholars (若手奨励賞)**
  The 260th SIG for Natural Language, Information Processing Society of Japan. 2024
- **SB Intuitions Awards**
  The 30th Annual Conference of the Japanese Association for Natural Language Processing. 2024
- **Monbukagakusho Honors Scholarship**
  Japanese Ministry of Education, Culture, Sports, Science and Technology. 2023
- **Outstanding Oral Presentation**
  2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology. 2022
- **Annual Outstanding Academic (GPA) Scholarship**
  Beijing Institute of Technology. 2018, 2019, 2021, 2022, 2023
- **First Prize**
  30th Chinese (High School) Chemistry Olympiad. 2016.
- **Second Prize**
  29th Chinese (High School) Chemistry Olympiad. 2015.

Research Projects

Mechanistic Interpretation of In-context Learning in Large Language Models

This research series focuses on decomposing the **I**n-**c**ontext **L**earning (ICL) process in large language models into human-interpretable, atomic operations. Specifically, we aim to describe ICL as a sequential procedure (or pseudocode), providing a clearer understanding of its internal mechanisms and inspiring practical applications. The project builds upon my ICLR 2025 and 2026 paper [1, 2] and extends to both applied research for enhancing ICL capabilities [3, 4, 5] and theoretical extensions [4, 24, 3, 2]. To date, this research series has fostered extensive academic collaboration both domestically and internationally.

Global Mechanistic Interpretability for Large Language Models

This project focuses on a broader scope of mechanistic interpretability, aiming to uncover the roles of individual components within large language models and how they interact with one another, and ultimately reconstructing the LLM as a pipeline of functional modules. As an initial step, we have explored how the language modeling head encode output probabilities [6]. Currently, the project is dedicated to developing more efficient automated tools to boost the mechanistic interpretability research. For example, we developed a binary SAE variant [7] to reduce the dense features.

Publication List

(**I**mpact **F**actor (IF) source: arxiv.org/pdf/2310.08037; h5 index source: Google Scholar)

International Conferences

1. Hakaze Cho, Haolin Yang, Gouki Minegishi, and Naoya Inoue. Mechanism of task-oriented information removal in in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026 (h5 index=362, IF=48.9)

2. Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *The Fourteenth International Conference on Learning Representations*, 2026 (h5 index=362, IF=48.9)

3. Haolin Yang, Hakaze Cho, Kaize Ding, and Naoya Inoue. Task vectors, learned not extracted: Performance gains and mechanistic insight. In *The Fourteenth International Conference on Learning Representations*, 2026 (h5 index=362, IF=48.9)

4. Haolin Yang, Hakaze Cho, Yiqiao Zhong, and Naoya Inoue. Unifying attention heads and task vectors via hidden state geometry in in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025 (h5 index=371, IF=23.3)

5. Hakaze Cho, Peng Luo, Mariko Kato, Rin Kaenbyou, and Naoya Inoue. Mechanistic fine-tuning for in-context learning. In *8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP @ EMNLP 2025*, 2025

6. Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. Revisiting in-context learning inference circuit in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025 (h5 index=362, IF=48.9)

7. Hakaze Cho, Yoshihiro Sakai, Mariko Kato, Kenshiro Tanaka, Akira Ishii, and Naoya Inoue. Token-based decision criteria are suboptimal in in-context learning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2025 (h5 index=126, IF=16.5)

8. Hakaze Cho, Yoshihiro Sakai, Kenshiro Tanaka, Mariko Kato, and Naoya Inoue. Understanding token probability encoding in output embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025 (h5 index=81, IF=7.7)

9. Yuting Shi, Naoya Inoue, Houjing Wei, Yufeng Zhao, and Tao Jin. Find-the-common: A benchmark for explaining visual patterns from images. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024 (h5 index=68)

10. Yufeng Zhao, Evelyn Soerjodjojo, and Haiying Che. Methods to enhance bert in aspect-based sentiment classification. In *2022 Euro-Asia Conference on Frontiers of Computer Science and Information Technology (FCSIT)*, pages 21–27. IEEE, 2022

Pre-prints

11. Hakaze Cho, Haolin Yang, Brian M. Kurkoski, and Naoya Inoue. Binary autoencoder for mechanistic interpretability of large language models. *arXiv preprint arXiv:2509.20997*, 2025

12. Takuya Kataiwa, Hakaze Cho, and Tetsushi Ohki. Measuring intrinsic dimension of token embeddings. *arXiv preprint arXiv:2503.02142*, 2025

13. Mariko Kato, Hakaze Cho, Yoshihiro Sakai, and Naoya Inoue. Affinity and diversity: A unified metric for demonstration selection via internal representations. *arXiv preprint arXiv:2502.14380*, 2025

14. Hakaze Cho and Naoya Inoue. Staicc: Standardized evaluation for classification task in in-context learning. *arXiv preprint arXiv:2501.15708*, 2025

15. Yufeng Zhao, Yoshihiro Sakai, and Naoya Inoue. Noisyicl: A little noise in model parameters calibrates in-context learning. *arXiv preprint arXiv:2402.05515*, 2024

16. Yufeng Zhao and Haiying Che. Skin: Skimming-intensive long-text classification using bert for medical corpus. *arXiv preprint arXiv:2209.05741*, 2022

Domestic Conferences / Journal / Miscellaneous

17. 趙羽風, 楊昊霖, 峰岸剛基, 井之上直也. 文脈内学習におけるタスク指向情報除去のメカニズム. 言語処理学会第 32 回年次大会. 2026

18. 趙羽風, 楊昊霖, Brian M. Kurkoski, 井之上直也. 二値化スパースオートエンコーダ. 言語処理学会第 32 回年次大会. 2026

19. 趙羽風, 井之上直也. 学会記事: Token-based Decision Criteria Are Suboptimal in In-context Learning. 自然言語処理. 2025.

20. 片岩拓也, 趙羽風, 大木哲史. トークン埋め込みの内在次元を測る. 人工知能学会第 39 回全国大会. 2025

21. 佐藤魁, 高橋良允, Benjamin Heinzerling, 田中健史朗, 趙羽風, 坂井吉弘, 井之上直也, 乾健太郎. 言語モデルにおける知識の既知性判断の内部表象. 人工知能学会第 39 回全国大会. 2025

22. 田中健史朗, 坂井吉弘, 趙羽風, 井之上直也, 佐藤魁, 高橋良允, Benjamin Heinzerling, 乾健太郎. 既知性を示す言語表現を伴う知識に関する内部表象の分析. 人工知能学会第 39 回全国大会. 2025

23. 趙羽風, 加藤万理子, 坂井吉弘, 井之上直也. 大規模言語モデルにおける In-context Learning の推論回路. 言語処理学会第 31 回年次大会. 2025 (優秀賞)

24. 趙羽風, 井之上直也. Beyond the Induction Circuit: A Mechanistic Prototype for Out-of-domain In-context Learning. 言語処理学会第 31 回年次大会. 2025

25. 片岩拓也, 趙羽風, 大木哲史. 埋め込み表現の内在次元を測る. 言語処理学会第 31 回年次大会. 2025

26. 加藤万理子, 趙羽風, 坂井吉弘, 井之上直也. 文脈内学習におけるデモの親和性と多様性の提案. 言語処理学会第 31 回年次大会. 2025

27. 趙羽風, 坂井吉弘, 加藤万理子, 井之上直也. StaICC: 文脈内学習における分類タスクの標準的なベンチマーク. 言語処理学会第 19 回 YANS シンポジウム. 2024

28. 加藤万理子, 趙羽風, 闍真竺, Yuting Shi, 井之上直也. 画像特徴ベクトルは重みを固定した言語モデルで情報豊かなトークンである. 言語処理学会第 19 回 YANS シンポジウム. 2024

29. 趙羽風, 坂井吉弘, 加藤万理子, 田中健史朗, 石井晶, 井之上直也. In-Context Learning におけるトークンベース較正手法の用いる決定境界は最適でない. 情報処理学会 NL 研第 260 回研究発表会. 2024 (若手奨励賞)

30. 趙羽風, 坂井吉弘, 井之上直也. NoisyICL: A Little Noise in Model Parameters Can Calibrate In-context Learning. 言語処理学会第 30 回年次大会. 2024

31. 坂井吉弘, 趙羽風, 井之上直也. In-context Learning において LLM はフォーマットを学べるか. 言語処理学会第 30 回年次大会. 2024 (SB Intuitions Awards)

32. Yuting Shi, Houjing Wei, Jin Tao, Yufeng Zhao, Naoya Inoue. Find-the-Common: Benchmarking Inductive Reasoning Ability on Vision-Language Models. 言語処理学会第 30 回年次大会. 2024