# Junheng Hao

| | | |
|---|---|---|
| CONTACT INFORMATION | Email: junhenghao@microsoft.com <br> Email (Personal): haojh.ucla@gmail.com <br> Mobile: +1 (424)355-5950 | *Homepage* <br> *LinkedIn* <br> *Google Scholar* |

**CURRENT EMPLOYMENT**

**Researcher / Member of Technical Staff, Microsoft**
Microsoft GenAI Research

**RESEARCH INTERESTS & AREAS**

- Large language model (Phi-3, GPT-4o) Training: Pre-training, Post-training, RLHF, etc
- Systematic data strategies for LLM: Data selection, Synthetic data generation
- Customized LLM development: Domain-specific LLM, Reasoning/Coding
- LLM + Knowledge Graph (KG)
- LLM Benchmarking/Evaluation

**EDUCATION**

**University of California Los Angeles (UCLA)**, Los Angeles, CA, United States     Sept 2022
Ph.D. in Computer Science
**Thesis:** Incorporating ontological information in knowledge graph learning and applications
**Advisors**: Yizhou Sun, Wei Wang
**Labs**: Scalable Analytics Institute (ScAi), UCLA Data Mining Lab

**Tsinghua University**, Beijing, China     May 2017
B. Eng. in School of Information Science and Technology
B. Sc (Econ, Minor) in School of Economics and Management

**PROFESSIONAL EXPERIENCE (INDUSTRY)**

**Researcher** at Microsoft GenAI, Redmond, WA     Oct 2022 - Current
- **Project:** SLM/LLM (Phi-Series, GPT-4/4o/4.1) Training and Customization
  - **Overview:** Training open-source small language model (SLM) such as Phi-model family from scratch (pre-training, post-training), LLM customization on OpenAI GPT models.
  - **Deliverables:** (1) Released pioneering open-source Phi model families: Phi-3/Phi-3.5/Phi-4-mini with two technical reports [1,2] and shipped to products including Windows and Microsoft Edge (such as Phi-Silica for Windows 11 Copilot+ PCs); (2) Customized LLM (GPT-4o, etc) for Microsoft internal (VSCode) and business partners (customer service supporting 17M+ consumers and customized AI agents for industry verticals).

**Research Intern** at Microsoft Research (MSR), Redmond, WA     June 2021 - Sept 2021
- **Mentors:** Chieh-Han Wu, Zhihong (Iris) Shen, Ye-Yi Wang, Jennifer Neville
- **Project:** KG-enhanced document representation learning
  - **Overview:** Enhancing document pretrained representations with infused document knowledge graph (DocKG), including Microsoft Academic Graph (Content + Graph).
  - **Deliverables:** One research technical preprint [8].

**PhD Research Intern** at IBM Research AI, San Jose, CA     June 2020 - Sept 2020
- **Mentor:** Chuan Lei, Berthold Reinwald, Fatma Ozcan
- **Project:** Ontology Matching by Utilizing Graph Neural Networks
  - **Overview:** Empowering hybrid graph neural networks (GNNs) for ontology matching between relational databases and standard healthcare ontologies.
  - **Deliverables:** One paper [7] published at KDD 2021 with corresponding invention filed and partially shipped in *IBM Micromedex solutions (Watson Health)*.

**Applied Scientist Intern/Student Researcher** at Amazon, Seattle, WA   June 2019 - Dec 2019
- **Mentors:** Tong Zhao, Luna Xin Dong, Christos Faloutsos
- **Project:** Diversified Complementary Recommendation on Product Graph
  - **Overview:** Enabling diversified complementary recommendation from web-scale product graphs and hierarchical product ontology.

- **Deliverables:** One paper [11] published at CIKM 2020 and deployed in *Amazon-wide product complementary recommendation engine*.

**Research Intern** at **NEC Lab America**, Princeton, NJ             June 2018 - Sept 2018
- **Mentors:** Lu-An Tang, Zhichun Li, Haifeng Chen
- **Project:** Enterprise Knowledge Graph Fusion and Adaptation
  - **Overview:** Graph-based Multi-source graph knowledge transfer and network fusion on enterprise engines for malicious process detection.
  - **Deliverables:** One research technical preprint [12] under review.

PUBLICATION SUMMARY

As of May 2025: I have accomplished 10+ Papers published papers and technical reports, among which I first-authored papers published on top-tier venues (KDD, CIKM, BCB) from interdisciplinary domains across machine learning, large language models, data mining, and bioinformatics, with 2000+ Citations. More up-to-date publication record can be found in *Google Scholar* and/or *Semantic Scholar*. Several research works at Microsoft and previous internship projects (such as Phi-4-mini [1], Phi-3.5 [2], P-Companion [11]) have been successfully deployed in Windows, Microsoft Edge and Amazon Retail services, while contributing to open-source LLM communities.

PUBLICATIONS & PREPRINTS

[1] *Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs*
Microsoft Azure GenAI team
Microsoft Blog Post: Empowering innovation: The next generation of the Phi family

[2] *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*
Microsoft GenAI team.
Microsoft Blog Post: Tiny but mighty: The Phi-3 small language models with big potential

[3] *SciAgent: A Tool-augmented LLM for Scientific Reasoning*
Yubo Ma, **Junheng Hao**, Ruochen Xu, Shuohang Wang, Zhibin Gou, Liangming Pan, Yujiu Yang, Yixin Cao, Aixin Sun, Hany Hassan Awadalla, Weizhu Chen
The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)

[4] *Language Models can be Logical Solvers*
Jiazhan Feng, Ruochen Xu, **Junheng Hao**, Hiteshi Sharma, Dongyan Zhao
2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)

[5] *Data Science Tasks Using Scripts versus GUI-based Workflows: The Good, the Bad, and the Ugly*
Alexander Taylor, Yicong Huang, **Junheng Hao**, Xinyuan Lin, Xiusi Chen, Wei Wang, Chen Li.
IEEE International Conference on Data Engineering (ICDE), Co-located with the 3rd International Workshop on Data Platform Design, Management, and Optimization (DataPlat)

[6] *Incorporating Ontological Information in Knowledge Graph Learning and Empowered Interdisciplinary Applications*
**Junheng Hao**
Doctoral committee: Wei Wang, Yizhou Sun, Kai-Wei Chang, Ying-Nian Wu
Doctoral Thesis & Dissertations

[7] *Multi-source Inductive Graph Knowledge Transfer*
**Junheng Hao**, Lu-An Tang, Yizhou Sun, Zhengzhang Chen, Haifeng Chen, Junghwan Rhee, Zhichun Li and Wei Wang.
Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD). September 2022.

[8] *Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification*
Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, **Junheng Hao**, Ye-Yi Wang, Kuansan Wang and Jiawei Han.
Proceedings of The Web Conference (WWW) 2022. April 2022.

[9] *MEDTO: Medical Data to Ontology Matching using Hybrid Graph Neural Networks*
**Junheng Hao**, Chuan Lei, Abdul Quamar, Vasilis Efthymiou, Fatma Ozcan, Yizhou Sun, Wei Wang.
Proceedings of 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD, Applied Data Science Track). August 2021.

[10] *JEDI: Circular RNA Prediction based on Junction Encoders and Deep Interaction among Splice Sites*
Jyun-Yu Jiang, Chelsea J.-T. Ju, **Junheng Hao**, Muhao Chen, Wei Wang.
Proceedings of the 29th annual international conference on Intelligent Systems for Molecular Biology and the 20th annual European Conference on Computational Biology (ISMB-ECCB). September 2021.

[11] *P-Companion: Framework for Diversified Complementary Product Recommendation*
**Junheng Hao**, Tong Zhao, Jin Li, Luna Xin Dong, Christos Faloutsos, Yizhou Sun, Wei Wang.
Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM), Applied Research Track. October 2020.

[12] *Bio-JOIE: Joint Representation Learning of Biological Knowledge Bases*
**Junheng Hao**, Chelsea J.-T. Ju, Muhao Chen, Yizhou Sun, Carlo Zaniolo, Wei Wang.
Proceedings of The 11th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB 2020), September 2020.
**Best Student Paper Award**.

[13] *Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts*
**Junheng Hao**, Muhao Chen, Wenchao Yu, Yizhou Sun, Wei Wang.
Proceedings of 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD, Research Track). August 2019.

[14] *Role of road network features in the evaluation of incident impacts on urban traffic mobility*
Chenshuo Sun, Xin Pei, **Junheng Hao**, Yewen Wang, Zuo Zhang, SC Wong
Transportation Research Part B: Methodological (TRB), 2018.

[15] *Normal/Abnormal Heart Sound Recordings Classification Using Convolutional Neural Network*
Tanachat Nilanon, Jiayu Yao, **Junheng Hao**, Yan Liu.
Proceedings of the 43rd Computing in Cardiology Conference (CinC). December 2016.

[16] *A Data Driven Approach for Evaluation of Urban Accident Impacts*
Chen-Shuo Sun, **Junheng Hao**, Xin Pei, Zuo Zhang.
Proceedings of IEEE Conference on Intelligent Transportation Systems (ITSC), December 2016.

[17] *EHIA: Empowering Homicide Analytics with MurderBook Knowledge Graphs and Domain-specific Language Models*
**Junheng Hao**, Jingyue Shen, Craig D. Uchida, Yizhou Sun, P. Jeffrey Brantingham, Wei Wang.
Prerpint, under review.

| PATENT APPLICATIONS | [18] *OntoGNN: Hybrid Graph Neural Networks for Ontology Matching.* |
|---|---|

[18] *OntoGNN: Hybrid Graph Neural Networks for Ontology Matching.*
Chuan Lei, **Junheng Hao**, Vasilis Efthymiou, Fatma Ozcan, Abdul Quamar. U.S. Patent Application. (Sept. 2021)

ACADEMIC SERVICES

### Associate Editor in Editorial Board
- Journal of Medical Internet Research (JMIR) — 2023

### Conference Area Chair / Senior Program Comittee
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) — 2023
- Conference on Information and Knowledge Management (CIKM) — 2023

### Conference Program Committee / Reviewer
- Annual Meeting of the Association for Computational Linguistics (ACL) — 2023
- Association for Computational Linguistics (ACL) Rolling Review — 2022
- Conference on Neural Information Processing Systems (NeurIPS) — 2021,2022
- SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) — 2019-2021
- International Conference on Machine Learning (ICML) — 2020-2023
- AAAI Conference on Artificial Intelligence (AAAI) — 2021-2024
- TheWebConf / International World Wide Web Conference (WWW) — 2022
- International Conference on Learning Representations (ICLR) — 2020-2024
- International Joint Conferences on Artificial Intelligence (IJCAI) — 2020-2024
- IEEE International Conference on Data Engineering (ICDE) — 2020
- Conference on Empirical Methods in Natural Language Processing (EMNLP) — 2019,2023
- SIAM International Conference on Data Mining (SDM) — 2019-2022
- Conference on Information and Knowledge Management (CIKM) — 2019-2022
- European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD) — 2019-2022
- Learning on Graphs Conference — 2022

### Program Committee Board
- International Joint Conferences on Artificial Intelligence (IJCAI) — 2022-2024

### Journal Reviewer
- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- IEEE Transactions on Big Data (TBD)
- ACM Transactions on Intelligent Systems and Technology (TIST)
- ACM Transactions on Knowledge Discovery from Data (TKDD)
- IEEE Transactions on Big Data (IEEE Big Data)
- BMC Bioinformatics

### Conference Volunteer
- International Conference on Learning Representations (ICLR) — 2021
- SIKKDD Conference on Knowledge Discovery and Data Mining (KDD), — 2019, 2020
- Conference on Empirical Methods in Natural Language Processing (EMNLP) — 2020, 2021
- Conference on Neural Information Processing Systems (NeurIPS) — 2018, 2020

### Membership
- Association for Computing Machinery (ACM)
- Institute of Electrical and Electronics Engineers (IEEE)
- Association for Computational Linguistics (ACL)
- Association for Computing Machinery's Special Interest Group on Information Retrieval (ACM SIGIR)
- Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD)
- Institute of Electrical and Electronics Engineers (IEEE) Young Professionals
- International Society for Computational Biology (ISCB)

| | |
|---|---|
| INVITED TALKS | • May 2023: Knowledge Graph Conference<br>Invited presentation: *"Ontology-aware Knowledge Graphs and Empowered Multidisciplinary Applications in Industry"* |
| | • July 2022: Google Brain.<br>Invited tech talk: *Recommendations on Documents, Products and More? A Knowledge Graph Approach* |
| | • June 2022: Microsoft, Azure Cognitive Services Research Group.<br>Invited tech talk: *Knowledge Augmented Applications: NLP, Bioinformatics and Recommendation* |
| | • Dec 2021: Coupang, Ranking, Discovery and Personalization.<br>Invited tech talk: *Knowledge Graphs Meets Product Recommendation: One Deep Learning Solution* |
| | • Aug 2021: Microsoft Search, Assistant and Intelligence (MSAI).<br>Graph learning session: *Knowledge Graph with Ontology Learning and Applications*. [Slides] |
| | • July 2019: Amazon, Product Graph.<br>Invited talk: *Representation Learning on Knowledge Graphs: Embedding, Logic Rules and Graph Neural Networks* (with Yizhou Sun). [Slides] |

| | | |
|---|---|---|
| HONORS AND AWARDS | • **Best Student Paper Award (ACM BCB)** | 2020 |
| | • NAACL D&I Subsidy Award | 2022 |
| | • ACL DEI Award | 2022 |
| | • SIGIR Student Travel Grant (CIKM) | 2020 |
| | • Student Travel Award (KDD) | 2019, 2020 |
| | • UCLA Graduate Division Fellowship | 2018-2019 |

| | |
|---|---|
| TEACHING | • CS M146: Introduction to Machine Learning (Instructor: Sriram Sankararaman).<br>Teaching Associate / Head TA, Winter 2021      Evaluation Score: **8.2**/9.0 |
| | • CS145: Introduction to Data Mining (Instructor: Yizhou Sun)<br>Teaching Associate / Head TA, Fall 2020      Evaluation Score: **8.0**/9.0 |
| | • CS32: Introduction to Computer Science II, Data Structures (Instructor: David Smallberg & Carey Nachenberg)<br>Teaching Assistant, Spring 2019 & Winter 2019 |
| | • CS145: Introduction to Data Mining (Instructor: Yizhou Sun)<br>Teaching Assistant, Fall 2018 |

| | |
|---|---|
| SKILLS | • **Programming:** Python (PyTorch), C/C++<br>• **Language:** Mandarin (Native), English (Proficient) |