

Yoshinari Fujinuma

Office Address
Virtual

fujinumay@gmail.com
<http://akkikiki.github.io>
Twitter: @akkikiki

Professional Experience

Member of Technical Staff, Cantina Labs, USA Dec 2024 - Jan 2025

- LLM Training: Ran full-parameter SFT and LoRA direct preference optimization (DPO) of mixture-of-experts (MoE) LLMs on AWS EKS Hyperpod, non-AWS k8s clusters and achieved ~8x faster multi-node training by unlocking Elastic Fabric Adapter, Infiniband for faster inter-node communication using Megatron.
- LLM Inference: Implemented, evaluated, and delivered custom sampling techniques on an internal vLLM fork to encourage creativity of LLM outputs.

Applied Scientist / Senior Applied Scientist, Amazon.com, Inc., USA Sep 2021 - Dec. 2024

- [AWS Bedrock Knowledge Base](#): Evaluated and developed extending Bedrock Knowledge Base (RAG on Bedrock) to be multimodal.
- [AWS Bedrock Titan](#): Worked on data cleaning and supervised fine-tuning for building an LLM from scratch using the AWS Elastic Kubernetes Service (EKS) cluster using DeepSpeed and NVIDIA NeMo.
- [AWS Comprehend Document AI](#): Built a model for extracting information on document images/PDFs. Made the inference time ~3x faster using [FasterTransformers](#).

Graduate Part-Time Instructor, Univ. of Colorado Boulder, USA Jan 2021 - May 2021

- CSCI 4622 Machine Learning ([Github Repo](#))

Teaching Assistant, Univ. of Colorado Boulder, USA Jan-Dec 2020, May-Jul 2021

Applied Scientist Intern, Amazon.com, Inc., USA May 2020 - Aug 2020

Research Assistant, Univ. of Colorado Boulder, USA Aug 2016 - Dec 2019

Applied Scientist Intern, Amazon.com, Inc., USA May 2018 - Aug 2018

Software Dev Engineer, Amazon Japan K.K., Japan Oct 2014 - Aug 2016

- Built a JA/ZH language detector for search queries with recall > 0.8.
- Built a semi-supervised CRF-based Japanese query label tagger with F1 = 0.91.

Software Engineer Intern, Amazon Japan K.K., Japan Nov 2013 - Feb 2014

Part-time Engineer, Atilika, Japan Aug- Nov 2013, Apr- Sep 2014

- Built a dictionary with 8000+ product names for JA named entity extractor.
- Built an EN to JA transliteration generator. Achieved accuracy@1 \approx 60%.

Software Engineer Intern, Cookpad, Japan July 2013 (one month)

Education

University of Colorado Boulder, USA Dec 2021
Ph.D. in Computer Science

Advisors

- Katharina Kann(2020-2021), Jordan Boyd-Graber(-2021), Michael J. Paul (-2019)

University of Tokyo, Japan Sep 2014

M.S. in Computer Science; Advisor: Akiko Aizawa

International Christian University, Japan

Mar 2012

B.A. in Computer Science and Mathematics; Advisor: Grant Pogosyan

Publications

- Benjamin Hsu, Xiaoyu Liu, Huayang Li, Yoshinari Fujinuma, Maria Nadejde, Xing Niu, Ron Litman, Yair Kittenplon, Raghavendra Pappagari: “[M³T: A New Benchmark Dataset for Multi-Modal Document-Level Machine Translation](#)”, NAACL (short) 2024
- Yoshinari Fujinuma*, Siddharth Varia*, Nishant Sankaran, Srikar Appalaraju, Bonan Min, Yogarshi Vyas: “[A Multi-Modal Multilingual Benchmark for Document Image Classification](#)”, EMNLP Findings (long) 2023
- Sharon Levy, Neha Anna John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth: “[Comparing Biases and the Impact of Multilingual Training across Multiple Languages](#)”, EMNLP (long) 2023
- Pietro Lesci, Yoshinari Fujinuma, Momchil Hardalov, Chao Shang, Lluís Marquez: “[Diable: Efficient Dialogue State Tracking as Operations on Tables](#)”, ACL Findings (long) 2023
- Yoshinari Fujinuma, Jordan Boyd-Graber, Katharina Kann: “[Match the Script, Adapt if Multilingual: Analyzing the Effect of Multilingual Pretraining on Cross-lingual Transferability](#)”, ACL (long) 2022
- Yoshinari Fujinuma, Masato Hagiwara: “[Semi-Supervised Joint Estimation of Word and Document Readability](#)”, TextGraphs-15@NAACL (short) 2021
- Mozhi Zhang*, Yoshinari Fujinuma*, Michael J. Paul, Jordan Boyd-Graber: “[Why Overfitting Isn’t Always Bad: Retrofitting Cross-Lingual Word Embeddings to Dictionaries](#)”, ACL (short) 2020
- Mozhi Zhang, Yoshinari Fujinuma, Jordan Boyd-Graber: “[Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification](#)”, AAAI 2020
- Yoshinari Fujinuma, Jordan Boyd-Graber, Michael J. Paul: “[A Resource-Free Evaluation Metric for Cross-Lingual Word Embeddings based on Graph Modularity](#)”, ACL (long) 2019
- Dasha Pruss, Yoshinari Fujinuma, Ashlynn R. Daughton, Michael J. Paul, Brad Arnot, Danielle Albers Szafr, Jordan Boyd-Graber: “[Zika discourse in the Americas: A multilingual topic analysis of Twitter](#)”, PLOS ONE 2019
- Mozhi Zhang, Yoshinari Fujinuma, Jordan Boyd-Graber: “Exploiting Cross-Lingual Subword Similarities in Low-Resource Document Classification”, Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing, 2018
- Yoshinari Fujinuma, Alvin Grissom II: “[Substring Frequency Features for Segmentation of Japanese Katakana Words with Unlabeled Corpora](#)”, IJCNLP (short) 2017
- Yoshinari Fujinuma, Hikaru Yokono, Pascual Martínez-Gómez, Akiko Aizawa: “[Distant-supervised Language Model for Detecting Emotional Upsurge on Twitter](#)”, PACLIC (long) 2015

*denotes equal contribution

- | | | |
|--------------------------|--|-------------|
| Selected Projects | Analysis on Multilingual Pretraining | 2020 - 2021 |
| | • Investigating the effect of using different pretraining languages. (Paper) | |
| | Intrinsic Evaluation Measure for Cross-Lingual Embeddings | 2017 - 2019 |
| | • Developed a graph-based intrinsic measure to evaluate the quality of cross-lingual word embeddings. (Paper) | |
| | Finite State Transducer (FST) for Kuromoji | 2015 |
| | • Replaced a double-array trie to an FST for Kuromoji, a java-based Japanese tokenizer used in Lucene, Solr, and Elastic Search. (Github Repo) | |

Academic Service	Area Chair:	
	<ul style="list-style-type: none"> • 2023 ACL (Multilingualism & Cross-lingual NLP track) • 2023 EMNLP (Multilingualism & Linguistic Diversity Track) • 2023 LREC-COLING (Information Extraction, Knowledge Extraction, and Text Mining Track) 	
	Program Committee/Reviewer:	
	<ul style="list-style-type: none"> • 2025 AAAI, ACL Rolling Review, COLM, LREC • 2024 AAAI, ACL Rolling Review, COLM • 2023 AAAI, ACL Rolling Review (Feb., Apr.), EACL, NAACL, W-NUT, NLP4HR • 2022 AAAI, ACL Rolling Review (Jan., Mar., Jun., Oct., Dec.), ACL, EMNLP, NAACL SRW, W-NUT, CoNLL • 2021 NAACL, ACL, EMNLP, ACL Rolling Review (Sept., Oct., Nov.), CoNLL, NAACL SRW, ACL SRW, W-NUT, Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP) • 2019 Workshop on Noisy User-Generated Text (W-NUT) 	
Academic Honors	Secondary Reviewer: 2019 ACL, 2017 EMNLP, 2017 WWW	
	Student Volunteer: 2020 ACL	
	<ul style="list-style-type: none"> • Outstanding Reviewer for CoNLL 2021 • Departmental Travel Grant from CU Boulder (300 USD) 2017 • Graduate School Travel Grant from CU Boulder (500 USD) 2017 • Dean's Fellowship from CU Boulder (1/2 of the yearly tuition and stipend) 2016 • Graduate School Travel Grant from Univ. of Tokyo (80,000 JPY) 2013 • Gödel Foundation Prize: Best Bachelor thesis in CS and Math (50,000 JPY) 2012 • Horie Takematsu and Koh Scholarship (2,000 USD) 2010 • Student Scholarship from ICU (1/3 of the yearly tuition) 2008-2012 	
Computer and Language Skills	<u>Languages:</u>	Proficient: Python; Intermediate: C++, Java
	<u>Libraries:</u>	vLLM, PyTorch, Megatron-core, Deepspeed, MLFlow, Git, Vim, L ^A T _E X
	<u>English:</u>	TOEFL iBT 101 (2015)
	<u>Domain-specific:</u>	machine learning, natural language processing