PhD Candidate, Computer Science
University of California, Santa Barbara
✉ *alonalbalak@gmail.com*
Personal Webpage
GitHub *LinkedIn*
Twitter *Scholar*

Alon Albalak

About Me

I am a Ph.D candidate in the Computer Science Department at the University of California, Santa Barbara, and a member of the NLP Group, co-advised by William Yang Wang and Xifeng Yan.
My primary research focus is on applying ML methods to NLP to improve data efficiency and model performance. In my research I have explored the use of methods including multi-armed bandits, data selection, multitask learning, transfer learning, reinforcement learning, and neuro-symbolic methods. Additionally, I have a wide array of interests in other topics including model efficiency, logic and reasoning, conversational AI, information retrieval, and multilingual models.

Education

2018–present *Ph.D, Computer Science*, *University of California, Santa Barbara*.
UCSB NLP Group
Advisors: William Yang Wang and Xifeng Yan

2016–2018 *B.S., Mathematics*, *Wayne State University*.

Selected Publications (Full publication list)

2024 *A Survey on Data Selection for Language Models*.
Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, William Yang Wang
Preprint [Github]

2024 *The Foundation Model Development Cheatsheet*.
Shayne Longpre, Stella Biderman, Alon Albalak, Gabriel Ilharco, Sayash Kapoor, Kevin Klyman, ...
Preprint [Website]

2023 *Improving Few-Shot Generalization by Exploring and Exploiting Auxiliary Data*.
Alon Albalak, Colin Raffel, William Yang Wang
NeurIPS, Main Conference, Paper [code] [presentation]

2023 *Efficient Online Data Mixing For Language Model Pre-Training*.
Alon Albalak, Liangming Pan, Colin Raffel, William Yang Wang
NeurIPS, Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models, Preprint

2023 *RWKV: Reinventing RNNs for the Transformer Era*.
Bo Peng*, Eric Alcaide*, Quentin Anthony*, Alon Albalak, ...
EMNLP, Findings, Paper [code]

2023 *Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning*.
Liangming Pan, Alon Albalak, Xinyi Wang, William Yang Wang
EMNLP, Findings, Paper [code]

2023 *CausalDialogue: Modeling Utterance-level Causality in Conversations*.
Yi-Lin Tuan, Alon Albalak, Wenda Xu, Michael Saxon, Connor Pryor, Lise Getoor, William Yang Wang
ACL, Findings, Paper [code]

2023 *Addressing Issues of Cross-Linguality in Open-Retrieval Question Answering Systems For Emergent Domains*.
Alon Albalak, Sharon Levy, William Yang Wang.
EACL, Demonstration Track. Paper [code]

2023 *NeuPSL: Neural Probabilistic Soft Logic*.
Connor Pryor, Charles Dickens, Eriq Augustine, Alon Albalak, William Wang, L. Getoor
IJCAI, Main Conference, Paper [code]

2022 *FETA: A Benchmark for Few-Sample Task Transfer in Open-Domain Dialogue*.
Alon Albalak, Yi-Lin Tuan, Pegah Jandaghi, Connor Pryor, Luke Yoffe, Deepak Ramachandran, Lise Getoor, Jay Pujara, William Yang Wang.
EMNLP, Main Conference. Paper [code]

2022 *An Exploration of Methods for Zero-shot Transfer in Small Language Models*.
Alon Albalak, Akshat Shrivastava, Chinnadhurai Sankar, Adithya Sagar, Mike Ross
NeurIPS, Efficient Natural Language and Speech Processing Workshop. Paper

2022 *Efficient Learning Losses for Deep Hinge-Loss Markov Random Fields*.
Charles Dickens, Connor Pryor, Eriq Augustine, Alon Albalak, Lise Getoor
UAI, Workshop on Tractable Probabilistic Modeling. Paper

2022 *Making Something out of Nothing: Building Robust Task-oriented Dialogue Systems from Scratch*.
Zekun Li, Hong Wang, Alon Albalak, Yingrui Yang, Jing Qian, Shiyang Li, Xifeng Yan
Alexa Prize Taskbot Challenge 2022. Paper

2022 *D-REX: Dialogue Relation Extraction with Explanations*.
Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, William Yang Wang.
ACL, NLP for Conversational AI Workshop. Paper [code]

2021 *Systems and methods for determining and using semantic relatedness to classify segments of text*.
Rohit Jain, Devin H. Redmond, Richard B. Sutton, Alon Albalak, Sharon Huffner.
US Patent US20210279420A1

2021 *Modeling Disclosive Transparency in NLP Application Descriptions*.
Michael Saxon, Sharon Levy, Alon Albalak, Xinyi Wang, William Yang Wang
EMNLP, Main Conference. Paper

Selected Projects

February 2021 – present *Recommender Dialogue Systems, in collaboration with UCSC, USC, Google.*
- Actively collaborating with researchers across institutions to solve problems in dialogue systems such as explainability, information extraction, and zero- or few-shot dialogue classification tasks
- **Resulting Publications**: FLAD, FETA, NeuPSL, D-REX

Advisors : Industry - William W. Cohen and Tania Bedrax-Weiss
Academic - William Yang Wang (UCSB), Lise Getoor (UCSC), and Jay Pujara (USC)

June 2021 – June 2022 *Alexa Prize Taskbot Challenge, Team Lead.*
- 8% acceptance rate
- Led and advised UCSB's "Team GauchoBot" in developing an agent that assists real Alexa customers to complete cooking and do-it-yourself projects that require multiple steps and complex decision making
- Designed algorithms for intent classification and question answering as well as the communication architecture between modules
- **Resulting Publication**: Making Something out of Nothing

May 2021 – October 2021 *COVID(ATACK), in collaboration with IARPA and Peraton Labs.*

- Mentored a visiting undergraduate researcher
- Built a multilingual open-retrieval question answering system for COVID-related journal articles and a clinical trials database
- Designed and implemented:
  - a multilingual deep semantic indexing method to retrieve relevant documents
  - a multilingual reading comprehension system to find answers within a document
- **Resulting Publication**: Paper/code

Professional Experience

June 2022 – September 2022 *Research Science Intern*, *Meta AI.*

- Directed and executed on 2 projects in collaboration with researchers across the company
- Explored data-efficiency through the use of multi-task learning and various prompting methods for small language models
- Explored the use of parameter-efficient methods for zero-shot generalization
- **Resulting Publications**: Data-Efficiency with a Single GPU

June 2019 – September 2020 *Research Associate*, *Theta Lake.*

- Built classifiers for automated risk detection in regulated industries through the use of natural language processing and other machine learning techniques
- Took multiple projects from inception to production, and developed 2 patent pending methods along the way
- **Resulting Patent**: US Patent US20210279420A1

December 2017 – September 2018 *Machine Learning Research Associate*, *Machine Vision and Pattern Recognition Lab*, Wayne State University.

- Research funded by the Epilepsy Foundation, titled "The Sound of Seizures"
- Built computer vision based CNN-LSTM model predicting the onset of seizures with 91% accuracy
- Optimized neural network in Keras/TensorFlow for portability to mobile devices

Fellowships & Awards

2023 *Neurips Scholar Award*, *37th Conference on Neural Information Processing Systems.*

2018 *Integrative Graduate Education and Research Traineeship (IGERT) Fellow*, *University of California, Santa Barbara.*

2018 *Academic Excellence Fellowship*, *University of California, Santa Barbara.*

2018 *Chia Kuei Tsao Award*, *Wayne State University.*
For outstanding academic achievement in the undergraduate mathematics program

Service & Outreach

ACL 2023-24 Workshop Organizer - NLP For Conversational AI (NLP4ConvAI)

ACL 2023 Social Organizer - Mindfulness meditation in a time of NLP hyperactivity

NeurIPS 2022 Workshop Organizer - Transfer Learning for NLP (TL4NLP): Insights and Advances on Positive and Negative Transfer. Proceedings.

2022-2024 Program Committee: ACL, NAACL, EMNLP, AAAI

Technical skills

Tools Python, C++, Shell, AWS, Azure

Packages PyTorch, TensorFlow, Keras, NumPy, SciPy

Machine Learning Natural language processing (NLP), computer vision (CV), transformers, sequence to sequence models, statistical analysis, regression, clustering

Teaching Experience

Spring, 2020 — *CS 165a: Artificial Intelligence - Lead TA*.

Fall 2020 – Spring 2021 — *CS 9: Object Oriented Programming*.

Military Experience

2012 – 2015 — *Reconnaissance Sabotage Unit*, *Israel Defense Forces*.

- Engineering, demolitions, and reconnaissance specialty training
- Battalion lead navigator