Marc Marone

marcmarone.com • mmarone1@jhu.edu • Google Scholar

Education

- **Johns Hopkins University** — **Baltimore, MD**
  Ph.D. in Computer Science, Advisor: Benjamin Van Durme — *08/2020 – Present*
- **Georgia Institute of Technology** — **Atlanta, GA**
  B.S. in Computer Science, Highest Honors — *08/2015 – 12/2018*

Experience

My research interests center around building and understanding large datasets for language understanding systems – from knowledge grounding in large language models to machine translation. Lately I've become interested in understanding datasets for LLMs using efficient datastructures and tools. Check out dataportraits.org

- **Databricks / MosaicML Research** — **New York, NY**
  Research Scientist Intern — *05/2024 – Present*
  Datasets and data quality for large language models. Analyzed instruction data, deduplication effects, and evaluation. Trained models with billions of parameters on trillions of tokens.
- **Microsoft - Semantic Machines** — **Bellevue, WA**
  Research Scientist Intern — *05/2021 – 08/2021*
  Researched multilingual encoders for semantic parsing in task oriented dialogue (function calling/API tool-use). Applied structured grammars and text to code models to improve multilingual performance.
- **Johns Hopkins** — **Baltimore, MD**
  Research Assistant — *08/2020 – Present*
  Research towards my PhD. Topics include efficient dataset documentation [1], [2], code generation models [3], large language model factuality [4], [5], federated learning [6], cross & multilingual semantic understanding [7].
- **Microsoft - Machine Translation Research** — **Redmond, WA**
  Researcher — *02/2019 – 08/2020*
  Worked on research projects for production quality transformer models, semi-supervised learning, and data to text generation [8] under Hany Hassan. Built recipes for efficient domain adaptation of transformer models to customer specific translation data.
- **Microsoft** — **Bellevue, WA**
  SWE Intern — *05/2018 – 08/2018*
  Built a real time analysis system for diagnosing Bing service performance anomalies using Spark, Redis, d3.js
- **Quantlab Financial** — **Houston, TX**
  SWE Intern — *05/2017 – 08/2017*
  Built an anomaly detection service for analyzing quantitative trading system data using Elasticsearch
- **Georgia Tech** — **Atlanta, GA**
  Research Assistant — *08/2017 – 12/2018*
  Undergraduate Research Assistant with Jacob Eisenstein. Researched document level translation and summarization models. Developed methods for analyzing neural representations in multilingual tagger models.

Awards and Grants

- **Outstanding paper award at the Conference on Language Models (CoLM, 0.3% of Submissions)** — **2024**
  For work on knowledge cutoffs in LLM pretraining data [2]
- **Hopkins Engineering Applications & Research Tutorials** — **2023**
  Designed course on AI Safety and Security topics
- **Amazon Initiative for Interactive Artificial Intelligence, grant co-author** — **2022**
  Rapid Multilingual Dataset Creation with Automatic Projection and Human Supervision [9]

Teaching

- **AI Safety and Security** — **Johns Hopkins University**
  Instructor — *Fall 2023*
  Designed and taught a seminar course on AI Safety and Security. Topics include hallucinations in ChatGPT, data privacy, AI Art, and adversarial attacks. Planned lectures accessible to students with minimal CS backgrounds.
- **Teaching Assistant** — **Georgia Institute of Technology**
  Introduction to OOP; Introduction to AI — *7 Semesters*
  Taught recitations, held office hours, and designed assignments for Intro. AI & Object Oriented Programming.

Service and Organizing

External

- **Reviewing:** ACL ARR (ACL, NAACL, EMNLP, etc.), NeurIPS, CoLM
- **BigCode:** Open source initiative to build billion parameter large language models for code (StarCoder V1 & V2 [3], [10]). I worked on building the dataset and related attribution tools.

Internal

- **JHU Pre-Application Support:** Mentor PhD applicants by providing application guidance
- **CLSP Director Search Committee:** Student representative on committee to select a center director
- **Dean Search Committee:** Sole undergraduate representative on committee to select the Dean of Computing
- **HackGT:** Organizer for a 1000+ person hackathon, authored workshops attended by 100+ students

Skills

Programming Languages: Python, Bash, Java, OCaml, HTML/CSS, Javascript

Software: PyTorch, Hugging Face, Redis, Marian, fairseq, DyNet

See Google Scholar and my personal site for all publications and recent news!

Selected Publications

[1] M. **Marone** and B. Van Durme, "Data portraits: Recording foundation model training data," *NeurIPS*, 2023.

[2] J. Cheng, M. **Marone**, O. Weller, D. Lawrie, D. Khashabi, and B. Van Durme, "Dated data: Tracing knowledge cutoffs in large language models," *CoLM* (***Outstanding Paper, 0.3% of Submissions***), 2024.

[3] R. Li, L. B. Allal, Y. Zi, N. Muennighoff, D. Kocetkov, C. Mou, M. **Marone**, C. Akiki, J. Li, J. Chim, *et al.*, "Starcoder: May the source be with you!" *Transactions on Machine Learning Research*, 2023.

[4] O. Weller*, M. **Marone***, N. Weir, D. Lawrie, D. Khashabi, and B. V. Durme, ""According to ..." Prompting language models improves quoting from pre-training data," *EACL*, 2024.

[5] J. Zhang, M. **Marone**, T. Li, B. Van Durme, and D. Khashabi, "Verifiable by design: Aligning language models to quote from pre-training data," *ArXiv*, 2024.

[6] O. Weller*, M. **Marone***, V. Braverman, D. Lawrie, and B. Van Durme, "Pretrained models for multilingual federated learning," *NAACL*, 2022.

[7] M. Yarmohammadi, S. Wu, M. **Marone**, H. Xu, S. Ebner, G. Qin, Y. Chen, J. Guo, C. Harman, K. Murray, A. S. White, M. Dredze, and B. Van Durme, "Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction," *EMNLP*, 2021.

[8] L. Miculicich*, M. **Marone***, and H. Hassan, "Selecting, planning, and rewriting: A modular approach for data-to-document generation and translation," *3rd WNGT at EMNLP*, 2019.

[9] S. Behzad, S. Ebner, M. **Marone**, B. Van Durme, and M. Yarmohammadi, "The effect of alignment correction on cross-lingual annotation projection," *LAW-XVII at ACL*, 2023.

[10] A. Lozhkov, R. Li, L. B. Allal, F. Cassano, J. Lamy-Poirier, N. Tazi, A. Tang, D. Pykhtar, J. Liu, Y. Wei, *et al.*, "Starcoder 2 and the stack v2: The next generation," *ArXiv*, 2024.