Aonan Zhang

📱 (+1) 929-461-6023 | ✉ szyagn@gmail.com | in aonan-zhang-4208bb67

Experience

Apple — Foundation Model (AFM) Team
Seattle, WA

Staff Machine Learning Research Scientist *Feb. 2023 – Present*

- **Core contributor to the development of Apple Intelligence's foundation language models (2023 – Present).**
- **Developed AFM's text pretraining corpus on math, including large-scale real & synthetic data.**
  - *Partner with crawling and extraction teams to expand coverage reach and improve extraction precision.*
  - *Elevate data quality by implementing robust filtering, categorization, and curation of high-value math content.*
  - *Design and curate synthetic training data from diverse sources including QA, forums, and research publications.*
- **Research on foundation models, synthetic data generation, etc.**
  - *Synthetic Bootstrapped Pretraining: an end-to-end data synthesis pipeline that generates trillions of synthetic data points with minimal human intervention, achieving data self-bootstrapping in pretraining.*
- **Improved AFM's model efficiency.**
  - *Designed, developed, and launched recurrent draft models as a speculative decoding solution into Apple's serving pipeline.*
  - *Investigated performance trade-offs in scaling sparse attention layers versus FFN in MoE.*

ByteDance
Bellevue, WA

Senior Research Scientist *Oct. 2019 – Feb. 2023*

- Designed, implemented, and deployed cutting-edge data subsampling methods with theoretical guarantees.
- Delivered 10+ launches in ByteDance's Ads. & Rec. system that reduced ∼ 20M USD resource compensation per year in total.
- Published papers on data subsampling, uncertainty estimation, graph generation, and federated learning.

Google
New York, NY

Research Intern *May. 2018 – Dec. 2018*

- Developed Unbounded Interleaved-State RNN (UIS-RNN), a sequence model for segmenting and clustering sequences.
- Applied UIS-RNN to speaker diarization, achieving state-of-the-art error rates.

Education

Columbia University in the City of New York
New York, NY

Ph.D. in Electrical Engineering *Aug. 2014 – Oct. 2019*

- Thesis: Composing deep learning and Bayesian nonparametric methods.
- Committee: Shih-Fu Chang, David Blei, John Cunningham, John Paisley.

Tsinghua University
Beijing, China

B.Eng., M.Eng. in Computer Science and Technology *Aug. 2008 – Jul. 2012, Aug. 2012 – Jul. 2014*

- Thesis: Max-margin infinite hidden Markov models. (Advisor: Jun Zhu)

Selected Publications

[1] Zitong Yang*, **Aonan Zhang***, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, and Ruoming Pang. Synthetic bootstrapped pretraining. *arXiv preprint arXiv:2509.15248*, 2025.

[2] Hanzhi Zhou, Erik Hornberger, Pengsheng Guo, Xiyou Zhou, Saiwen Wang, Xin Wang, Yifei He, Xuankai Chang, Rene Rauch, Louis D'hauwe, et al. Apple intelligence foundation language models: Tech report 2025. *arXiv preprint arXiv:2507.13575*, 2025.

[3] Tom Gunter, Zirui Wang, Chong Wang, Ruoming Pang, Andy Narayanan, **Aonan Zhang**, Bowen Zhang, Chen Chen, Chung-Cheng Chiu, David Qiu, et al. Apple intelligence foundation language models. *arXiv preprint arXiv:2407.21075*, 2024.

[4] Yunfei Cheng*, **Aonan Zhang***, Xuanyu Zhang, Chong Wang, and Yi Wang. Recurrent drafter for fast speculative decoding in large language models. *arXiv preprint arXiv:2403.09919*, 2024.

[5] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Anton Belyi, et al. Mm1: methods, analysis and insights from multimodal llm pre-training. In *European Conference on Computer Vision*, pages 304–323. Springer, 2024.