

# Stanford

---

## Qizheng Zhang

Ph.D. Student in Computer Science, admitted Autumn 2022

### Publications

---

#### PUBLICATIONS

- **CARAVAN: Practical Online Learning of In-Network ML Models with Labeling Agents**  
Zhang, Q., Imran, A., Bardhi, E., Swamy, T., Zhang, N., Shahbaz, M., Olukotun, K., USENIX  
USENIX ASSOC.2024: 325-345
- **CacheGen: KV Cache Compression and Streaming for Fast Large Language Model Serving**  
Liu, Y., Li, H., Cheng, Y., Ray, S., Huang, Y., Zhang, Q., Du, K., Yao, J., Lu, S., Ananthanarayanan, G., Maire, M., Hoffmann, H., Holtzman, et al  
ASSOC COMPUTING MACHINERY.2024: 38-56
- **The Dataflow Abstract Machine Simulator Framework**  
Zhang, N., Lacouture, R., Sohn, G., Mure, P., Zhang, Q., Kjolstad, F., Olukotun, K., IEEE COMPUTER SOC  
IEEE COMPUTER SOC.2024: 532-547