Guanghui Qin

🌐 https://gqin.me ✉ gqin2@jhu.edu 🎓 Google Scholar

Education

Johns Hopkins University Maryland, US
Ph.D. in Computer Science (Advisor: Benjamin Van Durme) *Aug 2019 – Summer 2024 (expected)*

Peking University Beijing, China
B.S. in Physics & Computer Science *Sept 2015 – Jun 2019*

Experience

Microsoft Research Lab (MSR) Washington, US
Research Intern (Mentor: Corby Rosset) *May 2023 – Aug 2023*
Keywords: Large language model (LLM), Compressed text representation, Retrieval-augmented generation (RAG).
I researched efficient methods for LLMs. I proposed a method to compress the context of LLaMA with a compression ratio of up to 20x with minimal performance tradeoffs. It worked on retrieval-augmented generation (RAG).

Microsoft Semantic Machines Remote, US
Research Intern (Mentor: Anthony Platanios) *May 2022 – Aug 2022*
Keywords: Dataset, Graph neural networks (GNNs), User action prediction.
I studied a new research problem for user action predictions. I built a 2TiB dataset from GitHub and implemented a GNN model to predict the user actions (e.g. commit and pull request).

Johns Hopkins University Maryland, US
Visiting Researcher (Mentor: Hongyuan Mei and Jason Eisner) *Jun 2018 – Oct 2018*
Keywords: Time-series models, Stochastic process, Datalog.
I worked on temporal event stream modeling. We proposed a particle smoothing solution to sample events from a neural Hawkes process. The stochastic process may interact with a deductive temporal database such as Datalog.

Microsoft Research-Asia (MSRA) Beijing, China
Research Intern (Mentor: Jin-Ge Yao and Chin-Yew Lin) *Nov 2017 – Jun 2018*
Keywords: Grounded language learning, Data-to-text generation.
I proposed a Semi-HMMs-based statistics model for grounding natural language to structured data, which can be used to induce templates for data-to-text generation.

Awards

Best Short Paper Awardee in NAACL	Association for Computational Linguistics (ACL), *2021*
Outstanding Reviewer in EMNLP	Association for Computational Linguistics (ACL), *2019*
Silver Medalist	Chinese Physics Olympiad (CPhO), *2014*

Skills

- Programming languages: Python, Rust, JAVA, and C/C++. Other languages: Shell, LaTeX, SQL.
- Experience in fine-tuning large language models (LLMs), including distributed training, the use of LoRA, and customized transformer architectures.
- Experience in information retrieval, including retrieval-augmented generation (RAG).
- Machine learning tools: PyTorch, Lightning AI, DeepSpeed, FAISS, and PEFT.
- Network/Web: I have been hosting a proxy service (WallessPKU) since 2017 with more than 13k daily active users. I implement the proxy protocols and front-/back-end and maintain the database.

Academic Service

I serve as a reviewer for the conferences of NeurIPS (2019 and 2020 as secondary; 2021 to 2023), ICLR (2019 and 2020 as secondary; 2021, 2023, and 2024), ICML (2020 and 2021), ACL (2021), EMNLP (2019 to 2022; *outstanding reviewer award* in 2019), NAACL (2024), AAAI (2021), and AKBC (2020 as secondary).

Selected Publications

- Dodo: Dynamic Contextual Compression for Decoder-only LMs.
  Guanghui Qin, Corby Rosset, Ethan C Chau, Nikhil Rao, Benjamin Van Durme. In *arXiv*. 2024.
- Researchy Questions: A dataset of multi-perspective, decompositional questions for LLM web agents.
  Corby Rosset, Ho-Lam Chung, **Guanghui Qin**, Ethan C Chau, Zhuo Feng, Ahmed Hassan Awadallah, Jennifer Neville, Nikhil Rao. In *arXiv*. 2024.
- Streaming Sequence Transduction through Dynamic Compression.
  Wenting Tan, Yunmo Chen, Tongnfei Chen, **Guanghui Qin**, Haoran Xu, Heidi C Zhang, Benjamin Van Durme, Phillip Koehn. In *arXiv*. 2024.
- Nugget: Neural Agglomerative Embeddings of Text.
  Guanghui Qin and Benjamin Van Durme. In *Proceedings of the Conference on International Conference on Machine Learning (ICML)*. 2023.
- The NLP Task Effectiveness of Long-Range Transformers.
  Guanghui Qin, Yukun Feng, and Benjamin Van Durme. In *European Chapter of the Association for Computational Linguistics (EACL, oral)*. 2023.
- Learning How to Ask: Querying LMs with Mixtures of Soft Prompts.
  Guanghui Qin and Jason Eisner. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL, short)*. 2021. 🏆 **Best Short Paper Award**
- Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction.
  Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, **Guanghui Qin**, Yunmo Chen, J. Guo, Craig Harman, K. Murray, Aaron S. White, Mark Dredze, and Benjamin Van Durme. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, oral)*. 2021.
- LOME: Large Ontology Multilingual Extraction.
  Patrick Xia*, **Guanghui Qin***, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. In *Proceedings of Conference of the European Chapter of the Association for Computational Linguistics (EACL, demo)*. 2021.
- Iterative Paraphrastic Augmentation with Discriminative Span-based Alignment.
  Ryan Culkin, J Edward Hu, Elias Stengel-Eskin, **Guanghui Qin**, and Benjamin Van Durme. In *Transactions of the Association for Computational Linguistics (TACL), 9:494-509*. 2021.
- Neural Datalog Through Time: Informed Temporal Modeling via Logical Specification.
  Hongyuan Mei, **Guanghui Qin**, Minjie Xu, and Jason Eisner. In *Proceedings of the Conference on International Conference on Machine Learning (ICML, oral)*. 2020.
- Imputing Missing Events in Continuous-Time Event Streams.
  Hongyuan Mei, **Guanghui Qin**, and Jason Eisner. In *Proceedings of the Conference on International Conference on Machine Learning (ICML, oral)*. 2019.
- Learning Latent Semantic Annotations for Grounding Natural Language to Structured Data.
  Guanghui Qin, Jin-Ge Yao, Xuening Wang, Jinpeng Wang, and Chin-Yew Lin. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, oral)*. 2018.
- Data2Text Studio: Automated Text Generation from Structured Data.
  Longxu Dou, **Guanghui Qin**, Jinpeng Wang, Jin-Ge Yao, and Chin-Yew Lin. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP, demo)*. 2018.

(* indicates equal contribution) *Last updated on Mar 12, 2024.*