Saurabh Kulshreshtha

Website: kshre.github.io
Github: github.com/kshre
Google Scholar: vvStfcUAAAAJ
Email: saurkulsh@gmail.com
Mobile: +1-978-996-6764
LinkedIn: linkedin.com/in/kshre

Education

- **University of Massachusetts Lowell** — Lowell, MA
  PhD, Computer Science — *2018 - 2022*
  Research Focus: Large Language Models (LLMs), Applied NLP/NLU, Reasoning, Deep Learning Question Answering, Information Retrieval, Transfer Learning
  Publications: First author publications at **ACL 2021, 2022, 2023 and EMNLP 2020**

- **University of Massachusetts Lowell** — Lowell, MA
  Masters of Science, Computer Science — *2016 - 2018*
  Research Focus: Deep Learning, Applied NLP for Social Media Analysis and Medical domain
  Publications: NeurIPS Workshop

- **Guru Gobind Singh Indraprastha University** — Delhi, India
  Bachelor of Technology, Computer Science and Engineering — *2011 - 2015*
  Among top 100 engineering schools in India as ranked by National Institutional Ranking Framework

Skills Summary

- **Languages**: Python, C++, Java, PHP
- **Frameworks**: PyTorch, TensorFlow, Transformers, HuggingFace, CUDA
- **Tools**: Docker, Git, Spark, matplotlib, SQL, Shell, UNIX
- **ML**: LLMs, Recommendation Systems, Retrieval Augmented Generation, Search, Ranking, Deep Learning, Evaluations, Benchmarking

Experience

Meta — New York, NY
Machine Learning Software Engineer — *Dec 2023 - Present*

- **Brand Ads Ranking**:
  - Lead research and successfully deployed state-of-the-art modeling improvements to Ad Ranking models in brand ads.
  - Drove significant incremental revenue growth for a $1B+/year product.
  - Designed significant improvement in data quality through deep audits and infrastructure overhaul.
  - Owned end-to-end ad delivery, including monitoring and maintenance post-deployment.
- **New Monetization Initiative**:
  - Developing a new monetization opportunity at Meta a 0 to 1 project.
  - Design and build the entire ML ranking stack.
  - Part of a small cross-functional team.

Fidelity Investments — Jersey City, NJ
Senior Data Scientist, AI Center of Excellence — *Feb 2023 - Dec 2023*

- **Automated Report Generation**: Designed and deployed system that generates high-quality reports from financial filings through a combination of Retrieval Augmented Generation (RAG) in Large Language Models. Leads to significant decrease in time to generate reports by human experts.
- **Investor Propensity Modeling and Segmentation**: Developed and deployed predictive analytics models to forecast investor behavior and capital flow patterns. Architected an automated segmentation framework that identifies high-value investors and at-risk accounts, enabling strategic resource allocation and proactive client engagement. This solution transformed ambiguous investor data into actionable intelligence for relationship managers.
- **Cross-Modal Document Search Analysis Engine**: Developed and deployed a Natural Language search and Analysis system through combination of VLM, LLM and Retrieval Augmented Generation (RAG). Supports multiple modalities: tabular, text and image. Applied prompt engineering and Chain-of-Thought Prompting.

Meta — New York, NY
Research Scientist — *Sept 2022 - Feb 2023*

- **User Cohort Modeling - ANN Benchmarking**: Developed pipelines to benchmark performance of approximate kNN indices to retrieve and model user cohort for aggregated and anonymous data for ad ranking.
- **Advertising Ranking Performance - ML Production Data Pipeline**: Engineered data pipelines to calculate offline ad ranking metrics on Aggregated and Anonymized data for Ad ranking teams.

Amazon — Cambridge, UK
Applied Scientist 1, Alexa, Natural Language Understanding (NLU) Team — *Aug 2019 - Mar 2020*

- **Zero-Shot Cross-Lingual Transfer in Large Language Models(LLM)**: Implemented current art and compared various transfer methods by evaluation on entity extraction and semantic slot-filling tasks. Proposed new method to improve cross-lingual transfer in large language models. **Published at EMNLP 2020.**

Facebook New York, NY
Machine Learning PhD Intern *Jun 2021 - Aug 2021*

- **Multitask Transfer Learning Framework for Personalized News Recommendation**: Implemented two tower multi-task learning recommendation system with PyTorch, worked in the News Signals team with Kevin Shi and the team was lead by Joe Isaacson. Implemented latest research, relevant research paper from ICML 2020. Cross functionally worked with two teams to collect similar tasks those teams own and integrated them as additional tasks into our multi-task experiments. Experimented with permutations of user survey tasks and found positive transfer gain of up to +2.0 AUC and +0.5 AUC for two tasks.

Q-Centrix San Diego, CA
NLP Software Engineer, PhD Intern *Jun 2019 – Aug 2019*

- **Question Answering and Entity Extraction from Electronic Health Records and Health Analytics**: Prototyped QA system that highlights relevant information in patient notes helping experts faster answer questions on medical forms. Lead effort to create a dataset of relevant entity annotations in Electronic Health Records (EHR).

Research Projects

- **Few Shot Multi-hop Question Generation**: Designed **LLM** framework to model relations between entities and themes across multiple documents with few-shot annotation. Performance improved by +20 percentage points absolute in complexity of generated questions in human evaluations. **Presented at ACL 2023 Workshop.**
- **New Factual Question Answering Benchmark for LLMs**: Developed new NLP task and benchmark of solving crosswords. Released a new large-scale question-answering benchmark sourced from crossword clues and answers. Evaluated performance with seq-2-seq and retrieval augmented generative transformers and found our benchmark to be quite challenging to most current art, RAG models are found stronger. **Published at ACL 2022 (20.8% Acceptance Rate).**
- **Anomalies in LLM Vector Spaces**: Discovered high magnitude anomalous dimensions across embeddings of pre-trained Transformer models. Investigated the effect of pruning handful parameters responsible for outliers found massive performance degradation in downstream tasks and MLM loss. **Published at ACL 2021 (Findings). (38% Acceptance Rate)**
- **Cross-lingual Alignment Methods for Multilingual LLMs**: Assesses cross-lingual supervision and alignment methods to improve Multilingual BERT transfer capability. Proposes novel normalization. Identifies task and alignment method biases, highlighting areas of strength. **Published at EMNLP 2020 (Findings). (38% Acceptance Rate)**

Publications

1. **Kulshreshtha S.**, Rumshisky A., "Reasoning Circuits: Few-shot Multihop Question Generation with Structured Rationales", 1st Workshop on Natural Language Reasoning and Structured Explanations at **ACL 2023**
2. **Kulshreshtha S.**, Kovaleva O., Shivagunde N., Rumshishky A., "Down and Across: Introducing Crossword-Solving as an NLP Benchmark", Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), **ACL 2022**
3. Kovaleva O.*, **Kulshreshtha S.***, Rogers A. and Rumshisky A., "BERT Busters: Outlier Dimensions that Disrupt Transformers", Findings of the Association for Computational Linguistics, Findings of the Association for Computational Linguistics, **ACL 2021**
4. **Kulshreshtha S.**, Redondo-Garcia J. L., Chang C. Y., "Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study", Findings of the Association for Computational Linguistics, **EMNLP 2020**.
5. Boag W., Sergeeva E., **Kulshreshtha S.**, Szolovits P., Rumshisky A. and Naumann T. "Cliner 2.0: Accessible and accurate clinical concept extraction", ML4H: Machine learning for Health Workshop, **NeurIPS 2017**.
6. Rumshisky A., Gronas M., Potash P., Dubbo M., Romanov A., **Kulshreshtha S.** and Gribov A. "Combining Network and Language Indicators for Tracking Conflict Intensity", in **International Conference on Social Informatics 2017**, SocInfo 2017.

* - Shared authorship

Professional Activities

- **Served as reviewer at top NLP conferences: ACL (2021, 2022), EMNLP (2020, 2022, 2023), AACL 2022 and ACL Rolling Review 2023**
- **Served as reviewer at ICNLSP23 and AICS23**