Abhishek Malali

abhishekmalali@gmail.com (617)-982-4783

PROFESSIONAL EXPERIENCE

Amazon AGI Boston, MA
Senior Applied Scientist Jul 2023 - Present

- Technical lead specializing in deploying large language models (LLM) in production workflows. Collaborated with over 50 scientists and led the launch of Alexa Conversational LLM.
- Conducting research and implementation of parameter-efficient fine-tuning techniques for task, domain, and language adaptation of foundational models improving expert performance by 4-6% on domain datasets.
- Designed and built scalable novel methods for generating high-quality synthetic conversational data using LLMs, incorporating self-reflection, dynamic instruction tuning, and structured validation of generated datasets.
- Expanded conversational data generation framework to create dynamic datasets and queries based on LLM agent responses. Developed into an internal tool actively used for training and evaluation data generation.
- Led the development of novel iterative, data-driven prompt optimization strategies, leveraging error attribution, prompt tagging, and refinement techniques to reduce prompt sizes while improving API precision by 7% relative.

Amazon Alexa Cambridge, MA
Applied Scientist Oct 2018 - Jun 2023

- Technical lead focused on enhancing automated speech recognition (ASR) accuracy for named entities, user preferences, and dynamic content.
- Developed offline teacher ASR models leveraging bi-directional audio encoders, task-specific language models, and deep neural language models for re-scoring, and deployed these models to generate ASR query corrections which improved 1% of existing defects.
- Improved the speech model's ability to process personalized entertainment requests and rare tokens by leveraging constrained optimization, personalized biasing artifacts, and advanced data augmentation techniques. The improved model reduced ASR defects by 11%.
- Led a team of four scientists, providing individual contributions while mentoring and coaching to optimize team performance and foster professional growth.

Tribe Dynamics San Francisco, CA
Data Scientist Jul 2017 - Sep 2018

- Developed and implemented text classification models to link social media content with brands, and built optimized ETL pipelines to support high-throughput model inference for social media data.

Harvard University Cambridge, MA
Graduate Student Researcher Dec 2015 - May 2017

- Enhanced the LSTM module for predicting irregular time series by developing an autoregressive model and a joint optimization approach that integrates autocorrelation as a regularization term, effectively mitigating autocorrelated residuals and achieving a 7% reduction in test RMSE on astronomical datasets.

H2O.ai Cambridge, MA
Data Science Intern May - Aug 2016

- Integrated hyper-parameter optimization with H2O and bench-marked different hyper-parameter optimization methods for H2O AutoML.

EDUCATION

Harvard University Cambridge, MA
Master of Engineering in Computational Science and Engineering, GPA: 3.8/4.0 Aug 2015 - May 2017

- IACS scholarship awardee for 2016, Harvard Graduate Leadership cohort for 2017, Data Science TA

National Institute of Technology - Karnataka Surathkal, India
Bachelor of Technology in Electrical and Electronics Engineering Jul 2010 - May 2014

- Awards: Institute Gold Medal 2014, O.P. Jindal Engineering and Management Scholarship 2010, 2012, 2013

PROGRAMMING SKILLS

- **Languages** : Python, PySpark, Spark, SQL, Unix, Latex
- **Tools** : Pytorch, HuggingFace, AWS (SageMaker, Bedrock, Batch, Lambda), H2O, Tensorflow, Pandas