tl;dr I am valued for discerning what matters and delivering it fast, I am looking to continue working on on-device (mobile) AI optimizations, want AI models to work correctly and blazingly fast on-device.

I'm experienced with

- On-device AI
- Server side AI - training optimizations and inference serving
- Backend engineering (distributed, devops, model serving)
- Frontend dev
- A little bit of design and lots of software engineering.

I am your person if you want someone who

- Relishes hacking on low-level optimizations.
- Is experienced and rewarded for making right technical tradeoffs and decisions.
- Will save you loads of time and money by probing forward and making decisions based on solid data.
- Is good at documenting code, processes and meetings.
- Has a penchant for profiling and optimizing, selecting the right place to optimize.
- Writes code not just for machines but humans too.
- Has experience with NLP research (papers published) and writing production code for server and mobile devices (android).
- Does all of the above quickly.

More about me

I have been fascinated with and working in Natural Language AI since 2016, since the times when seq-to-seq RNNs were the SOTA and using a Convolutional net in NLP was the hot new thing. It started with a desire to have my own personal assistant and led to me building one, one which became my thesis application project.

I have been with Samsung Research America since 2017 and have worked on a full pipeline of ML projects. I have worked closely with researchers to build a solution for the problems we solve, this includes:

- Extensive work on writing native code (C++) for Android devices. This includes writing a tokenizer for bert, full inference pipelines dealing with audio and language LLMs, writing openCL kernels, porting pytorch operators to GGML (C++).
- Identifying bottlenecks with systematic profiling (To the level of figuring that the PCIe bandwidth is the bottleneck in Amazon L40s GPU instances.).

- Built the production training pipeline for Samsung Bixby's top level classifier. Rewrote research training code for production use, sped up training by 2x (1hr -> 30 mins, ask me how).
- Also have experience with low level implementation of ML solutions (C++), developed the on-device (mobile) version of a part of Bixby running the top level classifier on mobile.
- Have extensively profiled LLMs on GPU, mobile with a variety of available solutions. Have done GPU profiling of models to identify bottlenecks and guide optimization efforts in team.
- Comfortable with selecting right trade-offs. In our recent paper, MoDeGPT: Modular Decomposition for Large Language Model Compression, did a memory - performance tradeoff to speed up an operation by **100x**.
- Comfortable with good solutions when perfect not possible. For an internal process involving human annotators in an NLP problem, proposed and built a system that tries to prefill information, to correctly as it can, even though not perfectly. Turns out, correcting some amount of info is way faster than entering everything from scratch. Reduced human annotators' time from 45 mins to 5 mins. But moreover, reduced their mental stress from 100 to almost 0.
- Working on algorithm side and publishing papers (See resume).
- Conducting large scale experiments, keeping checks for correctness.
- Speeding up these experiments in terms of compute time and human effort time as well.

What drives me

Ever since my masters and through my work experience at Samsung Research America, I have believed in the union of research and engineering. I have been drawn to Natural Language AI ever since I wanted to control music with my voice but not give my voice data to BigTech. This led me to develop my own little voice assistant and became my thesis application project.

I think that solving a problem and then being able to execute on that solution is what makes a complete engineer. Hence for me, there are no lines between research and engineering as such. It is just a full pipeline of solving problems, and then executing on it. And that has been my career trajectory as well. I work on end-to-end pipeline of a project, collaborating with researchers to solve a problem, learning the scientific and deliberate approach to solve a previously unsolved problem; and then being a bridge to the product team working closely with product engineers to translate our research solution into production quality code.