

# Insoo Chung

+1-979-326-0931 | [insoooo.chung@gmail.com](mailto:insoooo.chung@gmail.com) | [insoonotes.com](https://insoonotes.com) | [linkedin.com/in/insoochung](https://linkedin.com/in/insoochung)

## PROFESSIONAL EXPERIENCE

---

### Meta

Menlo Park, California

*Machine Learning Engineer | Python, PyTorch*

*Jun. 2024 – Present*

#### ♦ On-Device AI for Meta Glasses

- Develop and evaluate low-bit quantization techniques for efficient on-device inference on Meta glasses, focusing on software for hardware–software co-design.
- Translate quantization research into production-ready software across numerical formats, kernels, and hardware constraints.

#### ♦ Reinforcement Learning for Ads Conversion Optimization

- Led R&D of fine-tuned LLMs for target audience-aware ad creation, resulting in stat-sig increases in clicks, conversions, and revenue.

### Qualcomm AI Research

San Diego, California

*Senior Machine Learning Engineer | Python, PyTorch*

*Jul. 2023 – May 2024*

#### ♦ Large Language Model Optimization & Evaluation

- Executed low-bit mixed precision integer-based optimization tasks to enhance on-target execution of LLMs.
- Developed a standardized LLM evaluation pipeline, facilitating rapid training and optimization iterations

### Apple

Seattle, Washington

*AIML Intern @ Siri NL Platform | Objective-C, Python*

*May 2022 – Aug. 2022*

#### ♦ Siri Performance Analysis

- Developed an on-device performance benchmarking tool for Siri.

### Samsung Research

Seoul, South Korea

*Research Engineer @ NLP Lab | C++, Python, Tensorflow*

*Aug. 2017 – Jul. 2021*

#### ♦ On-Device Neural Machine Translation Research

- Suggested a low-bit (i.e. under 3 bit) quantization method for transformers – 11.8× model size compression, 8.3× memory footprint reduction, 3.5× inference speed-up within -0.5 BLEU in three translation directions
- Led the development and maintenance of a tensorflow-based training and compression codebase and a C++-based on-device inference code for quantized transformer models that powered numerous Samsung services

## RESEARCH EXPERIENCE [ ]

---

♦ **Reinforcement Learning:** Fine-tuned LLMs using custom reward models to optimize for user preferences, validating improvements in production environments.

♥ **Model Optimization and On-Device AI:** Developing quantization and ML systems for Meta glasses; previously proposed extremely low-bit transformer quantization for on-device inference. Published in **EMNLP (1st author)**.

♣ **Neural Machine Translation & Speech:** Enhanced sequence generation models via hard attention, simultaneous translation, and meta-learning. Published in **ACL (2nd author)**, **WMT**, **ICASSP**.

## EDUCATION

---

### Texas A&M University

College Station, TX, US

*M.C.S., Computer Science (GPA: 4.0/4.0)*

*Sep. 2021 – May 2023*

- Served as teaching assistant (ML, AI) and research assistant (ML-assisted HW design verification)

### Yonsei University

Seoul, South Korea

*B.S., Computer Science*

*Mar. 2011 – Aug. 2017*

- Leave of absence due to military service: *Mar. 2013 – Dec. 2014*

## AWARDS

---

- Bronze Prize in Samsung Best Paper Award** (*for ACL 2019 “Look Harder” ♣*), Samsung *Nov. 2019*  
**National Scholarship for Science and Engineering** (*full stipend*), KOSAF *Mar. 2011 – Jun. 2017*  
**Gold Prizes in Capstone Design Award**, Yonsei University *Nov. 2016*
  - Won prizes in all 3 divisions for an image captioning-based photo search engine

## TALKS

---

- Productizing Neural Machine Translation**, CSCE625, Texas A&M University *Mar. 2023*  
**Big Idea in Natural Language Processing**, CSCE181, Texas A&M University *Nov. 2022*  
**Extremely low-bit quantization for on-device transformers**, SustainNLP @ EMNLP 2020 *Nov. 2020*

## OTHER EXPERIENCE

---

- DAC Young Fellow**, 2021 58th ACM/IEEE Design Automation Conference (DAC) *Dec. 2021*  
**Military Service, ROK Army** | *Radio Comms Soldier* *Mar. 2013 – Dec. 2014*
  - Squad leader duties: *Apr. 2014 – Dec. 2014*