# Haofei Yu

haofeiy@cs.cmu.edu | +1 (412)-537-0110 | www.haofeiyu.me

## Education

**Carnegie Mellon University - Language Technologies Institute** Pittsburgh, PA
Master of Science in Intelligence Information Systems May 2024

- Current Coursework: Multimodel Machine Learning, Natural Language Processing, Advanced Natural Language Processing

**Zhejiang University - ChuKochen Honors College** Hangzhou, China
Bachelor of Engineering in Computer Science and Technology Jun. 2022

- GPA: **3.96**/4.00 | Rank: **7**/134
- Awards: Provincial Scholarship (**top 5%**)
- Selected Coursework: Operating System, Computer Network, Computer Architecture, Compiler Principle, Theory of Computation

## Professional Experience

**Tencent** Shenzhen, China
AI Lab Intern (research) Feb. 2022 – Jul.2022

- Responsible for a research project related to Long-range Language Modeling.
- Proposed a unified framework for Long-range Language Modeling which considers Transformer-XL, Routing Transformer, ExpireSpan, and other long-range transformer variants to be special cases.
- Made modifications based on K-means clustering and improved Transformer-XL to become a general form.
- Achieved 0.3 Perplexity drop on Wikitext-103 dataset with a constant extra time cost.

## Research Experience

**An Encoder as Partial Re-ranker for Open-Domain Question Answering**
Supervisor: Prof. Yue Zhang *Westlake University* | Sep. 2021 – Feb. 2022

- Considered the encoder part in the Fusion-in-Decoder architecture to be partially responsible for re-ranking and enhanced its re-ranking ability by modifying the training objective.
- Conceptually modeled a new latent variable to describe the correlation between retrieved candidate passages and answers and used labeled golden passages as supervision signals for training.
- Achieved more than 1.5 points EM improvements on both NaturalQuestions and TriviaQA datasets.

**Empirical Study on Personalized Dialog Generation**
Supervisor: Asst. Prof. Diyi Yang *Georgia Institute of Technology* | Jul. 2021 – Sep. 2021

- Reviewed personalized response generation papers from speaker model (using persona embedding ) to P-square Bot (using RL methods to model mutual persona perception).
- Utilized tensor factorization to explicitly learn the latent representation of users' embedding in order to tackle the data sparsity problem of explicit and high-quality persona or demographic information.
- Proposed to add user-user-generation signals to the existing Transformer-based Dialog Generation framework to achieve mutual persona perception in the dialog.

**Uni-Encoder: A Fast and Accurate Response Selection Paradigm for Generation-Based Dialogue Systems**
Supervisor: Asst. Prof. Zhenzhong Lan *Westlake University* | Mar. 2021 – Jul. 2021

- Developed a new paradigm called Uni-Encoder, that keeps the full attention over each pair as in Cross-Encoder while only encoding the context once, as in Poly-Encoder.
- Designed in-batch negatives mechanism that makes other responses in one batch as negative samples for one history-response pair.
- Reached state-of-the-art performance on 4 benchmark datasets (Ubuntu v1, Ubuntu v2, ConvAI, and Douban).

## Skills

**Programming Languages:** Python/C/C++(**Advanced**), Verilog/SQL(**Intermediate**).
**Technologies /Frameworks:** PyTorch/Huggingface Transformers(**Advanced**), Tensorflow/React(**Intermediate**).