Kellen Cheng

kellentan@princeton.edu 4074094268 Website kellentan U.S. Citizen

Education

Princeton University **Aug. 2022 - Nov. 2024**

M.A. Electrical and Computer Engineering, GPA: 4.000
PhD Candidate (on-leave), Specialization: Natural Language Processing — Princeton, NJ

University of California, Los Angeles (UCLA) **Sept. 2018 - Jun. 2022**

B.S. Electrical Engineering, Technical Breadth Computer Science, GPA: 3.926 — Los Angeles, CA

Experience

Microsoft: Applied Scientist II **Oct. 2025 - Present**

Manager: Dr. Ji Li — Mountain View, CA

- Create ML/PPT team.

IBM Research: AI Research Scientist Intern **May. 2025 - Aug. 2025**

Mentor: Dr. Anna Lisa Gentile — San Jose, CA

- Formulated a method to dynamically generate type-constrained negative samples to enhance an LLM's factuality.
- Constructed an automatic pipeline which exploits structured source data, generating synthetic data which is used to seed longer context generation.
- Designed a framework to better embed misinformation into negative samples by utilizing GRPO along with a custom-defined reward function.

Samsung Research America: NLP Research Scientist Intern **Jan. 2025 - May. 2025**

Mentor: Dr. Ganesh Ramesh — Mountain View, CA

- Designed a multi-agentic LLM framework to improve conversation summarization through iterative text feedback.
- Adapted the framework for on-device local inference with Apple Silicon using Ollama and MLX-LM.
- Performed quantized LoRA (QLoRA) instruction fine-tuning for a range of language models spanning 0.5B to 9B parameters.

IBM Research: AI Research Scientist Intern **Jun. 2024 - Sept. 2024**

Mentor: Dr. Anna Lisa Gentile — San Jose, CA

- Synthesized and curated an evaluation benchmark for health advice guardrails from Common Crawl web text.
- Designed and implemented a sparse human-in-the-loop system for semi-automatic annotation of synthetic data at scale.
- Formulated a method to automatically generate synthetic data using compact LLMs for health advice guardrails.
- Fine-tuned scalable and compact detector models on a blend of synthetic and open-source training data, beating GPT-4o by 3.73% in accuracy and 1.54% in F1-score, despite containing 400x less parameters.
- Created an internal Rest API that integrated my detector model and automated internal model evaluations for the team.
- Work published in EMNLP industry track (first-author), with another work currently in submission (first-author).
- Filed a patent detailing a continual learning framework with model version-control and knowledge distillation for AI safety guardrail detector development.

Princeton: NLP Researcher **Nov. 2022 - Nov. 2024**

Advisor: Dr. Suma Bhat — Princeton, NJ

- Created an end-to-end two-mask infilling fine-tuning objective for idiomatic knowledge injection using the IEKG dataset.
- Implemented two-stage fine-tuning with transfer learning to achieve new state-of-the-art performance of 83.75% accuracy on the IMPLI benchmark, an improvement of 12% compared to previous state-of-the-art.

- Conducted ablation and data perturbation studies to gauge contextual reasoning capabilities for off-the-shelf language models ranging from 0.5B to 7B parameters, uncovering that they actually perform *better* without the context.
- Work published in NAACL (first-author) and EMNLP (second-author) main conferences.

Selected Publications

- **Kellen Tan Cheng**, Anna Lisa Gentile, Chad DeLuca, Guang-Jie Ren. *Backprompting: Leveraging Synthetic Production Data for Health Advice Guardrails.* ArXiV 2025.
- Chad DeLuca, Anna Lisa Gentile, Shubhi Asthana, Bing Zhang, Pawan Chowdhary, **Kellen Tan Cheng**, Basel Shbita, Pengyuan Li, Guang-Jie Ren, Sandeep Gopisetty. *OneShield - the Next Generation of LLM Guardrails.* ArXiV 2025.
- **Kellen Tan Cheng**, Anna Lisa Gentile, Pengyuan Li, Chad DeLuca, Guang-Jie Ren. *Don't Be My Doctor! Recognizing Healthcare Advice in Large Language Models.* EMNLP 2024 Industry Track.
- **Kellen Tan Cheng**, Suma Bhat. *No Context Needed: Contextual Quandary In Idiomatic Reasoning With Pre-Trained Language Models.* NAACL 2024 Main.
- Ziheng Zeng, **Kellen Tan Cheng**, Srihari Venkat Nanniyur, Jianing Zhou, Suma Bhat. *IEKG: A Commonsense Knowledge Graph for Idiomatic Expressions.* EMNLP 2023 Main.

Awards & Organizations

Toby & Jack Wolf Travel Grant **2024**
Bede Liu Travel Grant **2023**
Princeton ECE Departmental Fellowship **2022**
Tau Beta Pi **2020 - Present**
IEEE Eta Kappa Nu (HKN) **2019 - Present**
UCLA Dean's Honor List **2019 - 2022**

Skills

Languages Python, C++, MATLAB
Frameworks PyTorch, Tensorflow, MLX, Transformers
Tools Ollama, Slurm, LSF, AWS EC2, Anaconda/Mamba, Jupyter, LaTeX, MS Office