COMP 3361 Natural Language Processing

Lecture 16: Intro to Advanced Topics

Spring 2024
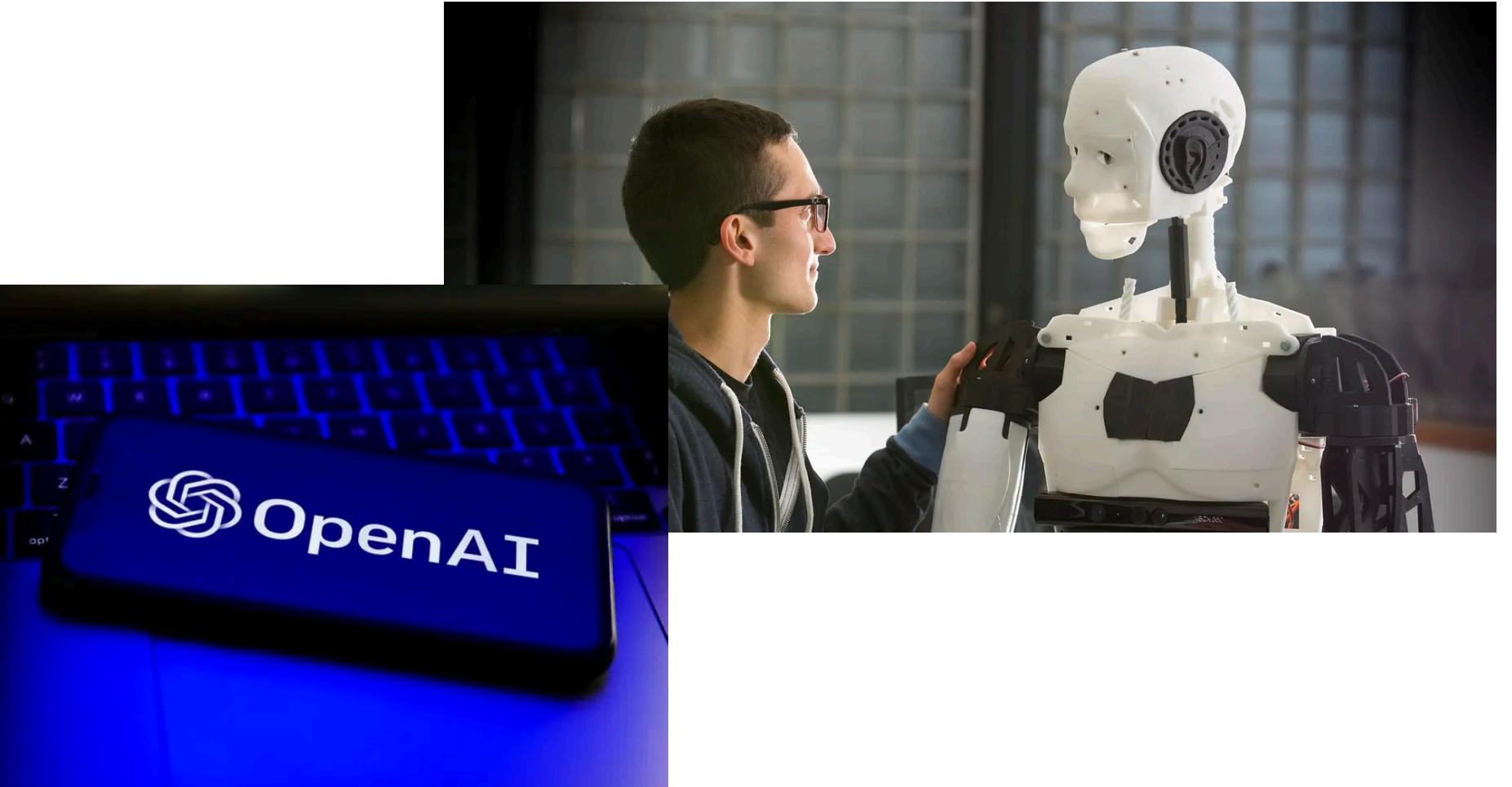
Tentative schedule

- Participate in two for the class participation + 3 more for 3% extra credits.
  - Paper readings
  - Attend the talk in person
 - Ask questions

授人以魚，不如授之以漁
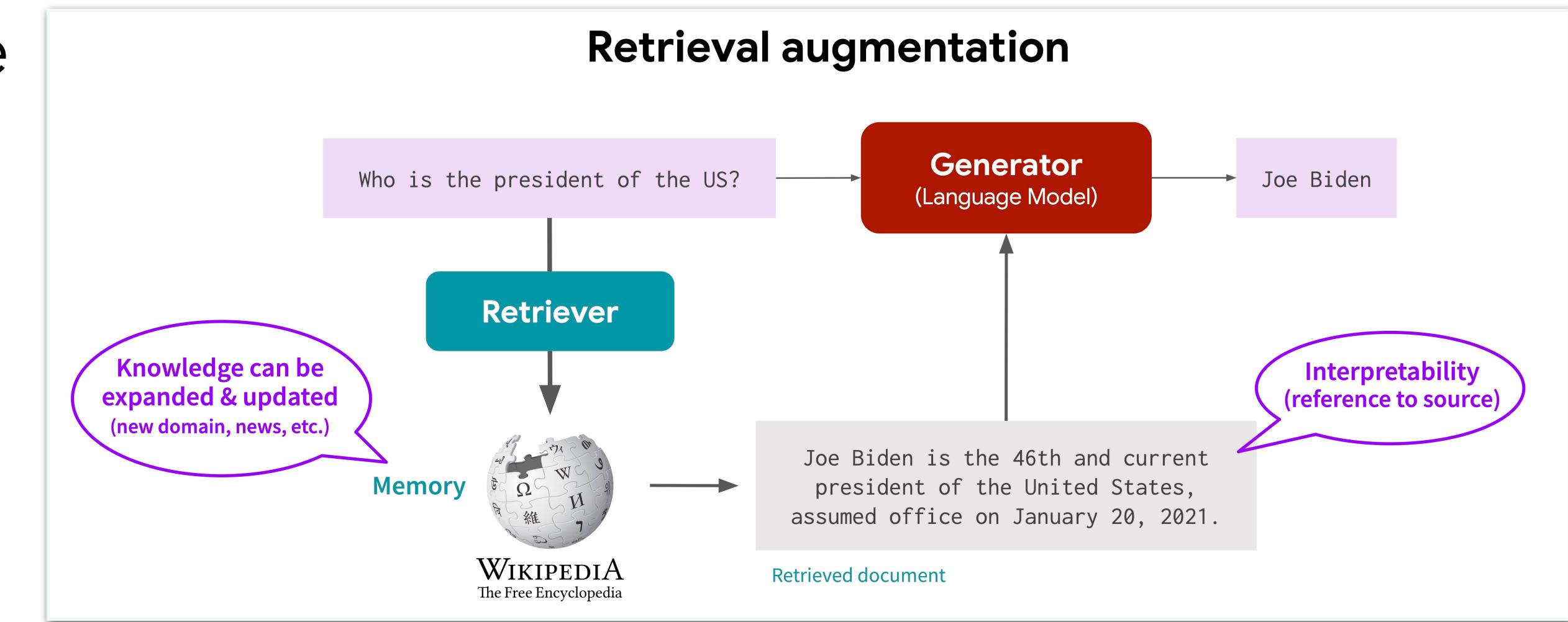
授人以魚，不如授之以漁

- Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime

- As a CS major, instead of memorizing thousands of coding examples, you can learn to code in a new programming language by reading docs and tutorials, searching Q&A on StackOverflow…

- How to teach LLMs to learn new skills?
  - Retrieval-augmented LLMs
  - Instruction tuning for LLMs
  - LLM alignment …

Retrieval-augmented LLMs

- How do we keep LLMs update-to-date without further training?

- Alleviate problems of hallucinations, lack of attributions, copyright in LLMs

- Smaller size, better performance?

Retrieval augmentation

Who is the president of the US? → **Generator** (Language Model) → Joe Biden

Retriever

Knowledge can be expanded & updated (new domain, news, etc.)

Memory

WIKIPEDIA
The Free Encyclopedia

Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.

Retrieved document

Interpretability (reference to source)

Weijia Shi

Pronouns: she/her

My name in Chinese: 施 惟佳

Email: swj0419@uw.edu

👋 **Hi!**

I am a PhD student in Computer Science at the University of Washington advised by Prof. Luke Zettlemoyer and Prof. Noah A. Smith. I have been a visiting research at Meta AI, working with Scott Yih. Prior to UW, I graduated from UCLA with a B.S. in Computer Science and Minor in Math. I am happy to mentor undergraduate or master students interested in research.

🌋 **Research Interests**

My main research focuses on natural language processing and machine learning. I am particularly interested in retrieval-augmented LMs and trustworthy AI. My goal is to build LMs that are able to communicate with external knowledge and personal data securely and robustly.
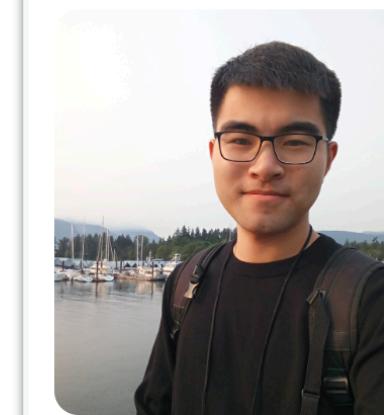
🌱 **What's NEW**

☑️ **Office hours**: Starting November 2023, I will be holding office hours (1~2 hours a week) dedicated to offering mentorship and advice to undergraduate/master students. If you want to chat about research and grad school application, please fill out the form

☑️ Honored to be selected as 2023 Machine Learning Rising Star

☑️ Two workshops accepted to *CL conferences. Stay tuned!

https://swj0419.github.io

https://thesequence.substack.com/p/edge-301-retrieval-augmented-language

Instruction tuning for LLMs

- How do we make LLMs more controllable by following specific instructions?

- Why following human instructions is important?

- That is the reason why you can prompt a single LLM to perform many tasks !!!

Yizhong Wang

PhD student

Paul G. Allen School of Computer Science & Engineering

University of Washington, Seattle, WA

Email: yizhongw [at] cs.washington.edu

Short Bio

I am a fifth-year PhD student at the Paul G. Allen School of Computer Science & Engineering, University of Washington. I am very fortunate to be co-advised by Hannaneh Hajishirzi and Noah Smith. I am also a part-time research intern at Allen Institute for Artificial Intelligence. I have previously interned at Meta AI, Microsoft Research, and Baidu NLP. Prior to UW, I obtained my Master's degree from Peking University and Bachelor's degree from Shanghai Jiao Tong University.

My primary research interests lie in natural language processing and machine learning. I am excited about the **generality** of large language models (LLMs). In particular, I have been thinking over the following topics these

https://homes.cs.washington.edu/~yizhongw

LLM alignment
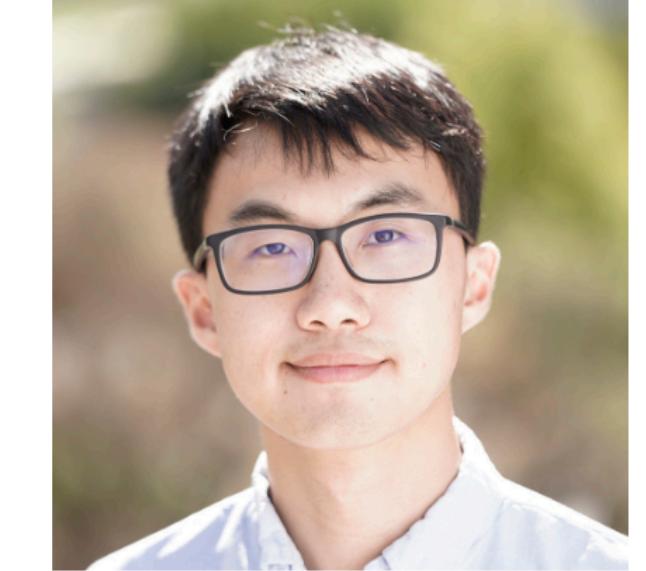
- How do we create LLMs that behaves in accordance with what a human wants?

- Who is the human? What is "a human wants"?

- If we have AI systems with skills that we find important or useful, how do we adapt those systems to reliably use those skills to do things we want? - Sam Bowman

- How do we align their (implicit) goals with the goals and values of their users?

Ruiqi Zhong

I am currently a PhD student in the UC Berkeley EECS department, and a part-time research scientist at Anthropic. I am co-advised by Prof. Jacob Steinhardt and Prof. Dan Klein.

Before coming to Berkeley, I finished my undergrad at Columbia University, where I worked with Prof. Kathleen McKeown.

Email / Google Scholar / Twitter / Github

Research Overview
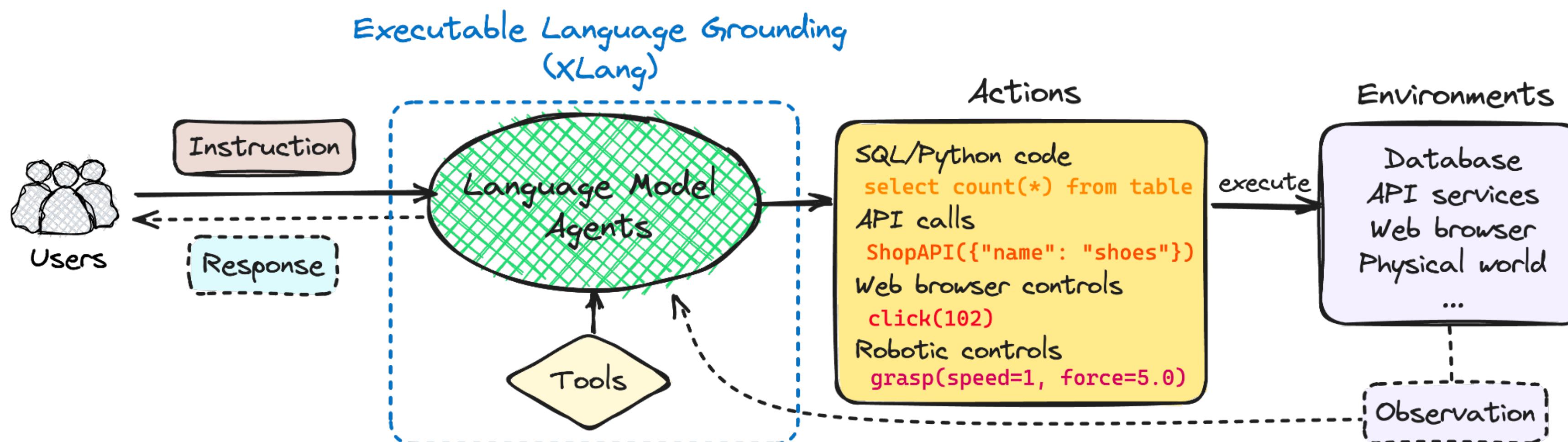
I work on scalable oversight -- supervising AI systems to accomplish tasks where humans alone struggle to determine the ground truth. Doing so requires human-AI collaborations, a better epistemic foundation, and new algorithmic tools. I currently work on concrete related problems in Natural Language Processing, Machine Learning, and Programming Language. See presentation slides here and my talk here to get a sense of my research interests.

https://ruiqi-zhong.github.io

LLMs/VLMs beyond chatting
Embodied LLMs/VLMs

LLMs/VLMs beyond chatting
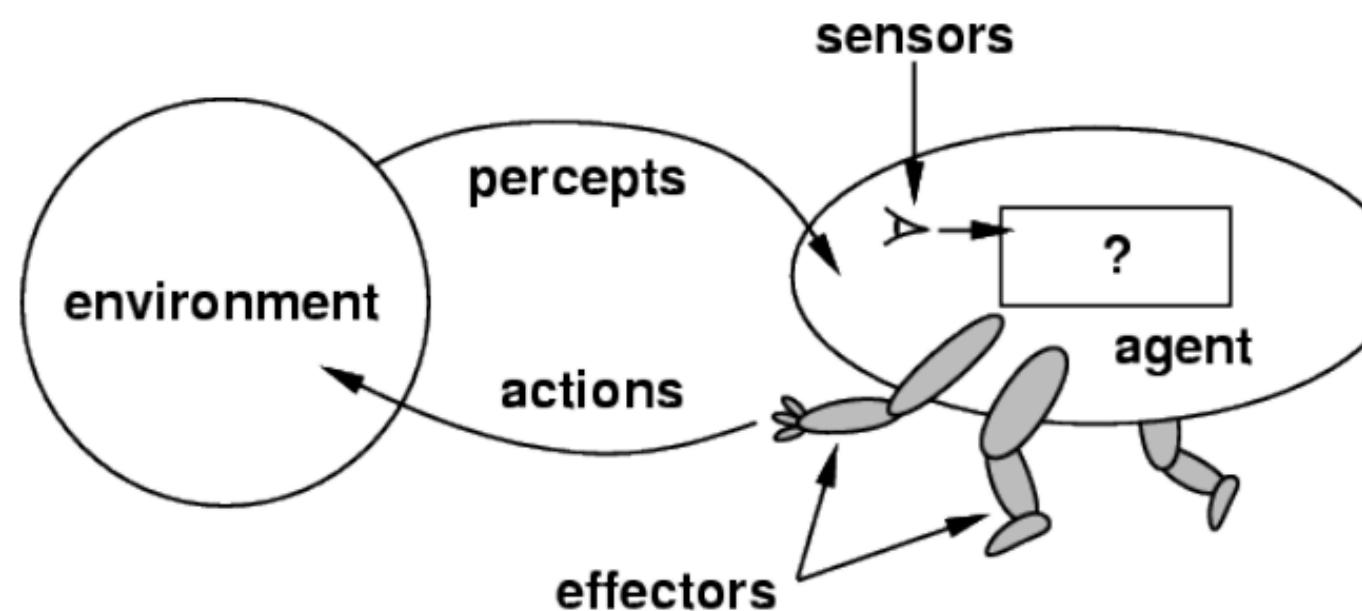
- Not just chatting with you, can we use LLMs/VLMs as brains of intelligent agents that can interact with and learn from humans and real-world environments (database, web browser, systems, physical world)?

LLMs/VLMs as agents

- The LLM/VLM agents need to make decisions for solving complex/abstract problems.

- How LLMs/VLMs as agents?

• Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.

• A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.

• Properties
  –Autonomous
  –Reactive to the environment
  –Pro-active (goal-directed)
  –Interacts with other agents
     via the environment

Shunyu Yao
姚顺雨

About

I am a PhD student at Princeton University, advised by Karthik Narasimhan.

In my work, I study methods, tasks, theories, and social impact of language agents.

In my life, I read, travel, rap, and play basketball.

I dedicate 30 minutes per week to chat with students. Just paper plane me!

Selected papers
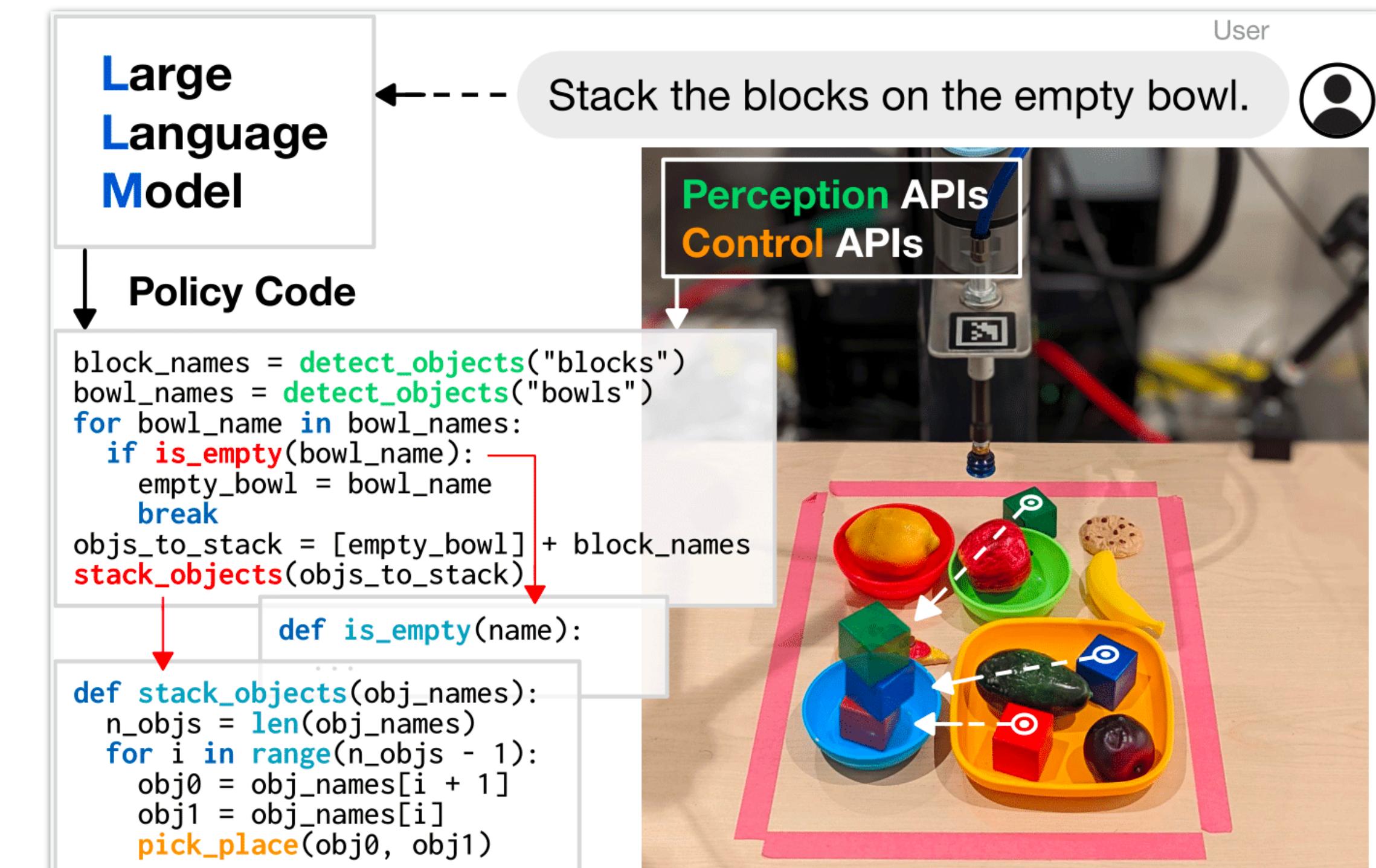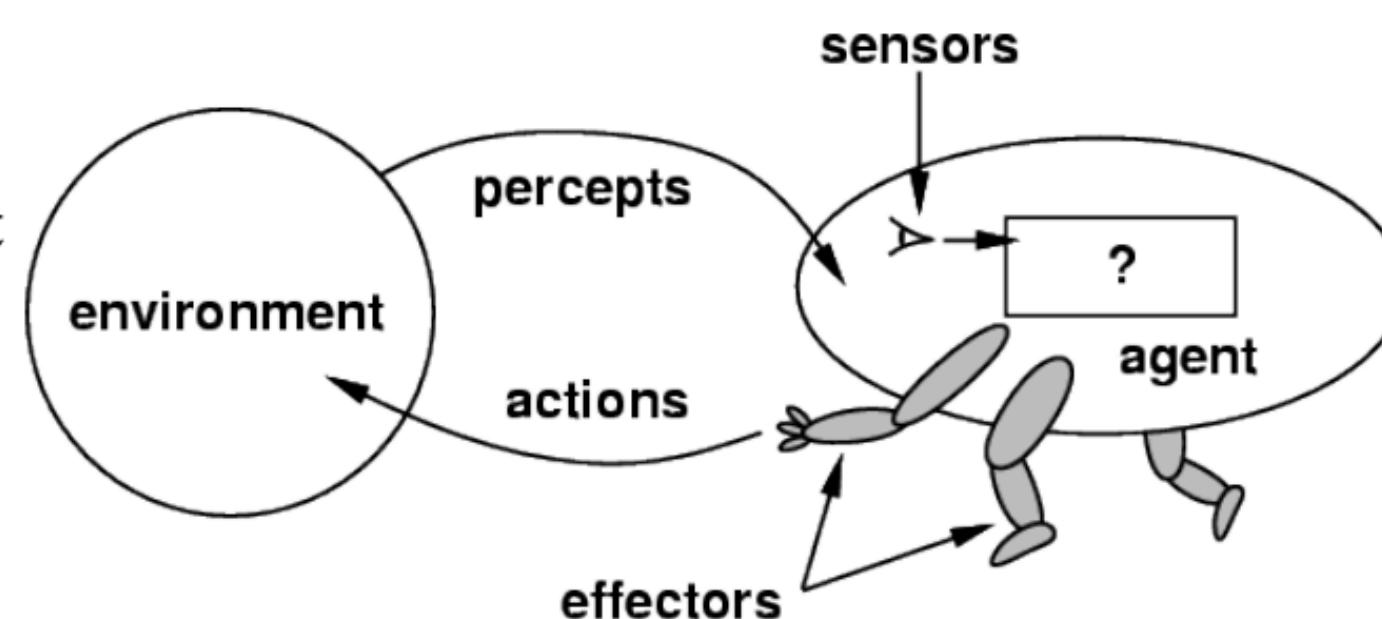
https://ysymyth.github.io
To be confirmed

LLMs/VLMs + robotics/embodied AI

- LLMs/VLMs + robotic agents enables LLMs/VLMs to take actions in real-world environments

- Multimodal LMs

- Definition: An **intelligent agent** perceives its environment via **sensors** and acts rationally upon that environment with its **effectors**.
- A discrete agent receives **percepts** one at a time, and maps this percept sequence to a sequence of discrete **actions**.
- Properties
  - Autonomous
  - Reactive to the environment
  - Pro-active (goal-directed)
  - Interacts with other agents via the environment

Large Language Model

User: Stack the blocks on the empty bowl.

Perception APIs
Control APIs

Policy Code

```
block_names = detect_objects("blocks")
bowl_names = detect_objects("bowls")
for bowl_name in bowl_names:
    if is_empty(bowl_name):
        empty_bowl = bowl_name
        break
objs_to_stack = [empty_bowl] + block_names
stack_objects(objs_to_stack)

def is_empty(name):

def stack_objects(obj_names):
    n_objs = len(obj_names)
    for i in range(n_objs - 1):
        obj0 = obj_names[i + 1]
        obj1 = obj_names[i]
        pick_place(obj0, obj1)
```

Foundation models for code and math

- How do we train LLMs that perform better on coding and math problems?

- Automatically write programs is one of the oldest and hardest problems in AI and CS

- How to use LLMs in an interactive software engineering development environment?

About Me

Hi! I am **Ansong Ni (倪安松)**, a final-year Ph.D. student in the Computer Science Department at Yale University, and my advisors are Prof. Arman Cohan, and (the late) Prof. Dragomir Radev. Prior to Yale, I obtained my M.S. in CS degree from School of Computer Science at Carnegie Mellon University and B.Eng. from Nanjing University in China.

I worked as a research intern at Google DeepMind (Summer 2023), Meta AI (Summer 2022), MSR Redmond (Summer 2021), AI2 (Summer 2020), MSR Asia (Summer, Fall 2017).

I am currently on the job market for industrial research labs & startups, and my expected graduation/start date is July 2024. You can obtain a copy of my CV here, and feel free to reach out.

Research Interest

I am generally interested in machine learning and natural language processing. My current research focus is on **teaching LLMs to write code and reason about program execution like human programmers**. More specifically, I've interested in these topics:

Ansong Ni

PhD student at Yale University. I study natural language and programming language

- New Haven, CT
- Email
- Twitter
- Github
- Google Scholar

https://niansong1996.github.io

LLM efficient training and tuning methods

Efficient LLM methods

- The computational and storage costs of LLM tuning and inference are usually too high, how can we reduce the costs?

- Parameter-efficient LM tuning approaches only fine-tune a small number of (extra) model parameters while freezing most parameters of the pretrained LLMs, thereby greatly decreasing the computational and storage costs.

About me

🧑‍💻 I am a postdoc at MIT, working with Yoon Kim. I am currently interested in developing more controllable and efficient sequence models. Free feel to reach out if you'd like to chat! 👋

Bailin Wang

Keep parsing papers on parsing ;-)

📍 Cambridge, MA

✉ Email

🐦 Twitter

⛓ Github
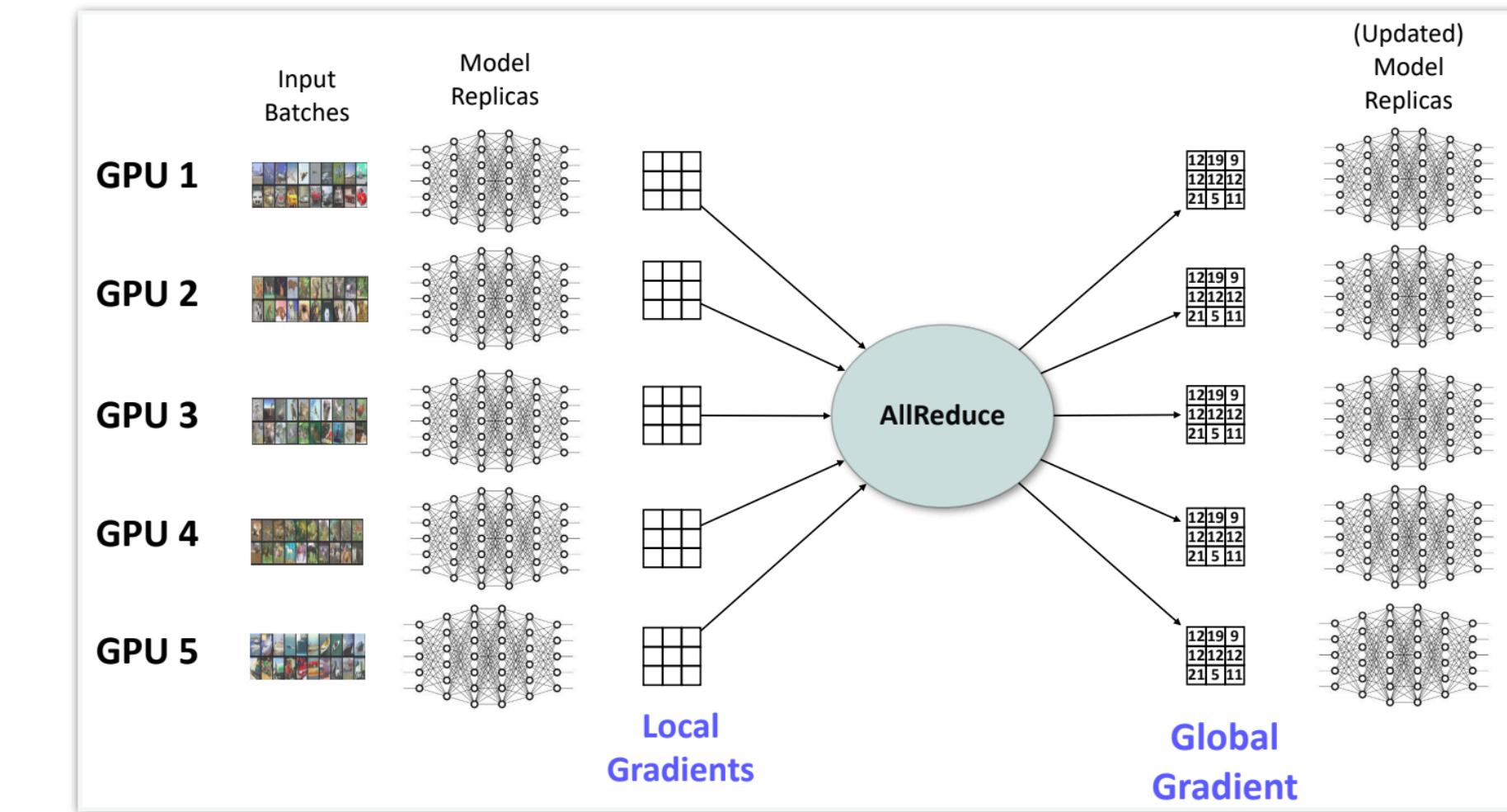
🎓 Google Scholar

I finished my PhD at the University of Edinburgh, advised by Ivan Titov and Mirella Lapata. Prior to my PhD, I worked on structured prediction with Wei Lu. During my PhD, I primarily focus on addressing several generalization challenges that arise in executable semantic parsing (e.g., text-to-SQL parsing), namely *domain generalization*, *learning from weak supervision* and *systematic generalization*, based on methodologies of **latent discrete structure learning** (e.g., separable permutation) and **specialized learning objectives** (e.g., meta-learning).

https://berlino.github.io/

LLM parallel pretraining

- Bigger models means more compute to train them. How to conduct data and model parallel training?

- Split the data and distribute data batches among replicas of the model. Partition the model across GPUs

- DeepSpeed architecture

Quentin Anthony

NOWLAB @ The Ohio State University

I am a PhD student under Dr. DK Panda's NOWLAB at The Ohio State University. My research is focused on the intersection of deep learning frameworks and high performance computing. Specifically, my research focuses on resolving distributed deep learning training bottlenecks such as checkpointing, model/optimizer compression, and deep/machine learning framework co-design. I actively contribute to the MVAPICH2 project and its subprojects such as MVAPICH2-GDR (High Performance MPI for GPU clusters), and HiDL (High Performance Deep Learning).

https://quentin-anthony.github.io
To be confirmed

Other topics

- LM evaluation, data, and benchmarking

- Bias, toxicity, and privacy in LLMs

- …