CSCE 689: Special Topics in Trustworthy NLP

Lecture 22: Tool-Augmented Language Models

Kuan-Hao Huang
khhuang@tamu.edu

Guest Lecture (Online)

- **Time:** November 24, 4:10pm-5:25pm
- **Title:** From Risk to Resilience: Addressing Misalignment in (Multimodal) Large Language Models
- **Speaker:** Fei Wang
- **Zoom Link:** https://tamu.zoom.us/my/khhuang?pwd=oAdWOKVOCGPApqDbJnVtktdW2AE6nb.1

Guest Lecture (Online)

Abstract: As (multimodal) large language models (LLMs) become central to intelligent systems, their use is expanding from everyday applications to high-stakes domains. Alignment plays a crucial role in the successful development of LLMs, ensuring that model behavior matches our expectations and remains consistent with various objectives. However, misalignment persists as a significant challenge that undermines the trustworthiness and reliability of these models. This talk will explore methods to tackle the misalignment problem by addressing three key research questions: (1) How to mitigate the risk of a misaligned LLM with only limited model accessibility? (2) How to ensure a reliable alignment process in multimodal scenarios? (3) How to integrate missing or customized alignment objectives to achieve precise control over model behavior in diverse contexts? Particularly, this talk will systematically address these challenges with resilient retrieval-augmented generation, conditional alignment, and constraint integration. In addition, this talk will shed light on the responsible development of LLMs across various scenarios and interdisciplinary contexts.

Guest Lecture (Online)

Speaker Bio:

Fei Wang is a Research Scientist at Google. Previously, he obtained his Ph.D. from University of Southern California. Fei's research focuses on developing post-training methods for robust and reliable LLMs and multimodal LLMs. His work has been recognized with an Amazon ML PhD Fellowship and an Annenberg PhD Fellowship. Additionally, he has instructed tutorials and served as area chairs at top tier NLP and ML conferences, including EMNLP, NAACL, ACL, and NeurIPS.

Project Presentations

W14	11/24	Invited Talk (Remote)	Title: *From Risk to Resilience: Addressing Misalignment in (Multimodal) Large Language Models* Speaker: *Fei Wang, Research Scientist at Google*
	11/26	Reading day (No Class)	
W15	12/1	Project Presentations (Remote)	
	12/3	Project Presentations (Remote)	
W16	12/9	Final Report Due	

Course Project – Project Presentation

- 8-min presentations + 1-min Q&A
  - Introduction to the topic you choose and problem definition
  - Existing progress and challenges
  - Proposed solutions, novelty, and contributions
  - Implementation details
  - Experimental settings, datasets, models, short introduction to baselines, evaluation metrics
  - Results and conclusion
- Clarity is the most important thing
  - Teach your classmate about your topic
- Time control is also important

Presentation Order

12/1

1. Yuqi Fan
2. Muhan Gao
3. Logon Wen
4. Kunal Jain
5. Oscar Chew, Serge Honcharenko
6. Aaron Xu
7. Yihong Yang

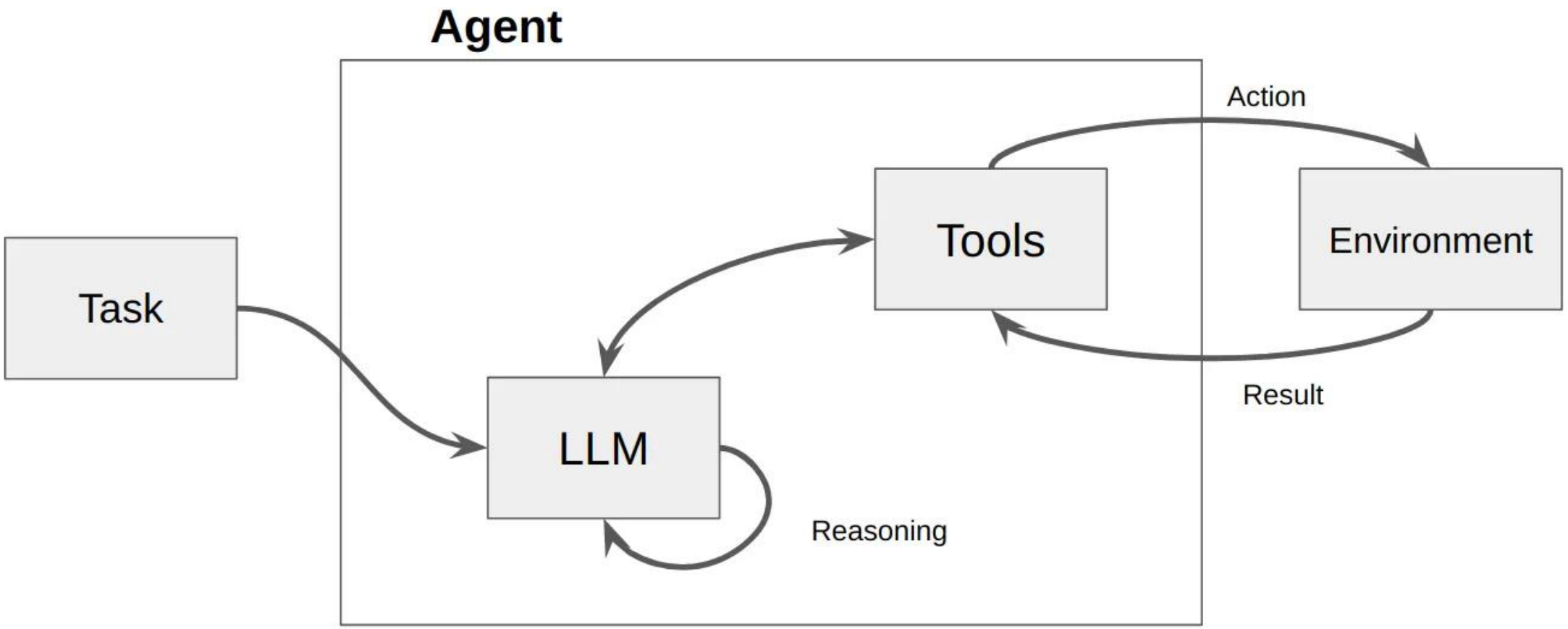
12/3

1. Sicong Liang, Ming Yang
2. Yi Wen
3. Quang Minh Nguyen
4. Bhaskar Ruthvik Bikkina
5. Junggeun Do
6. Jiongran Wang, Kowsalya Balamuralei Umamaheswari
7. Himanshu Parida

Course Project – Final Report

- Due: 12/9
- Page limit: 8 pages (excluding reference)
- The report should include
  - Introduction to the topic you choose and problem definition
  - Related literature and overview of existing progress and challenges
  - Proposed solutions, novelty, and contributions
  - Implementation details
  - Experimental settings, including dataset, models, baseline descriptions, evaluation metrics
  - Results and conclusion
- Make it a complete report!
- Codebase

Tool Using for LLMs



https://medium.com/@yasir_siddique/tool-calling-for-llms-a-detailed-tutorial-a2b4d78633e2

Why Tools?

- Improve accuracy and reliability

User

Task: Together Lily, David, and Bodhi collected 43 insects. Lily found 7 more than David. David found half of what Bodhi found. How many insects did Lily find? Solution output format: an integer.

LLM

Thought: To solve this problem, let's start by setting up equations [...] assume that David found x insects. Based on the given information, Lily found 7 more than David, so Lily found x + 7 insects [...]

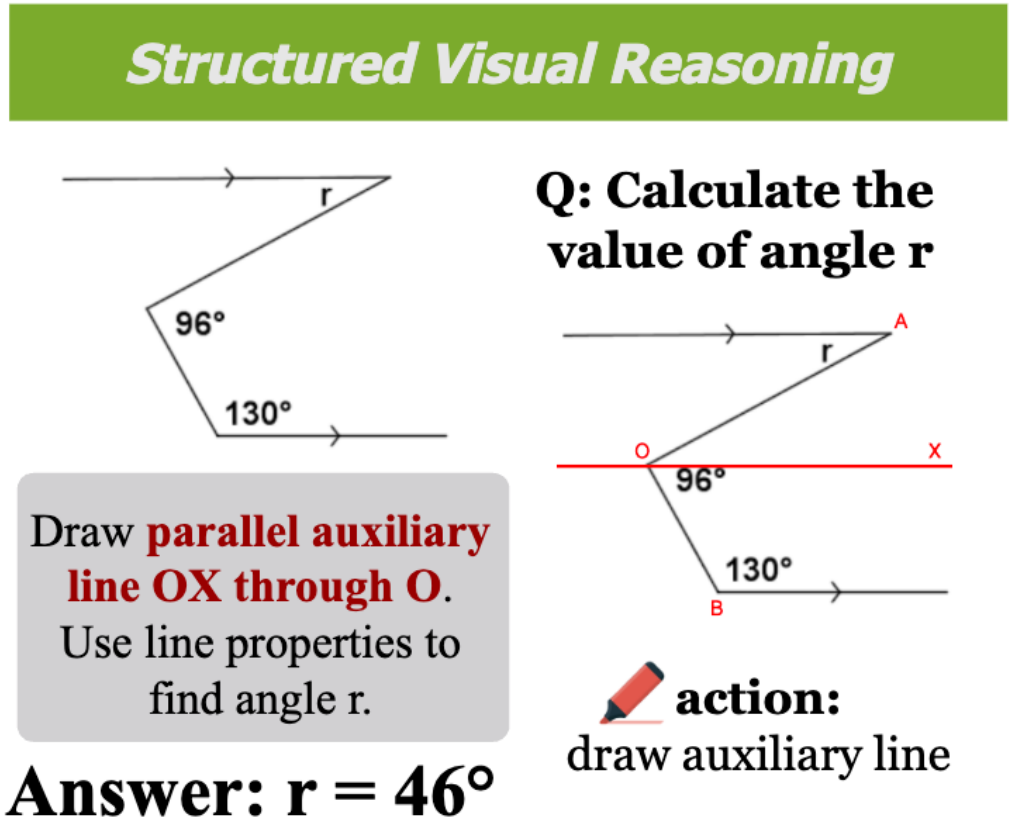
Execute:

```
from sympy import symbols, Eq, solve
x, b = symbols('x b')
# set up the equation
equation = Eq(x + (x + 7) + (1/2) * b, 43)
# solve the equation
solution = solve(equation, x)
solution[0]
```


MINT: Evaluating LLMs in Multi-turn Interaction with Tools and Language Feedback, 2024

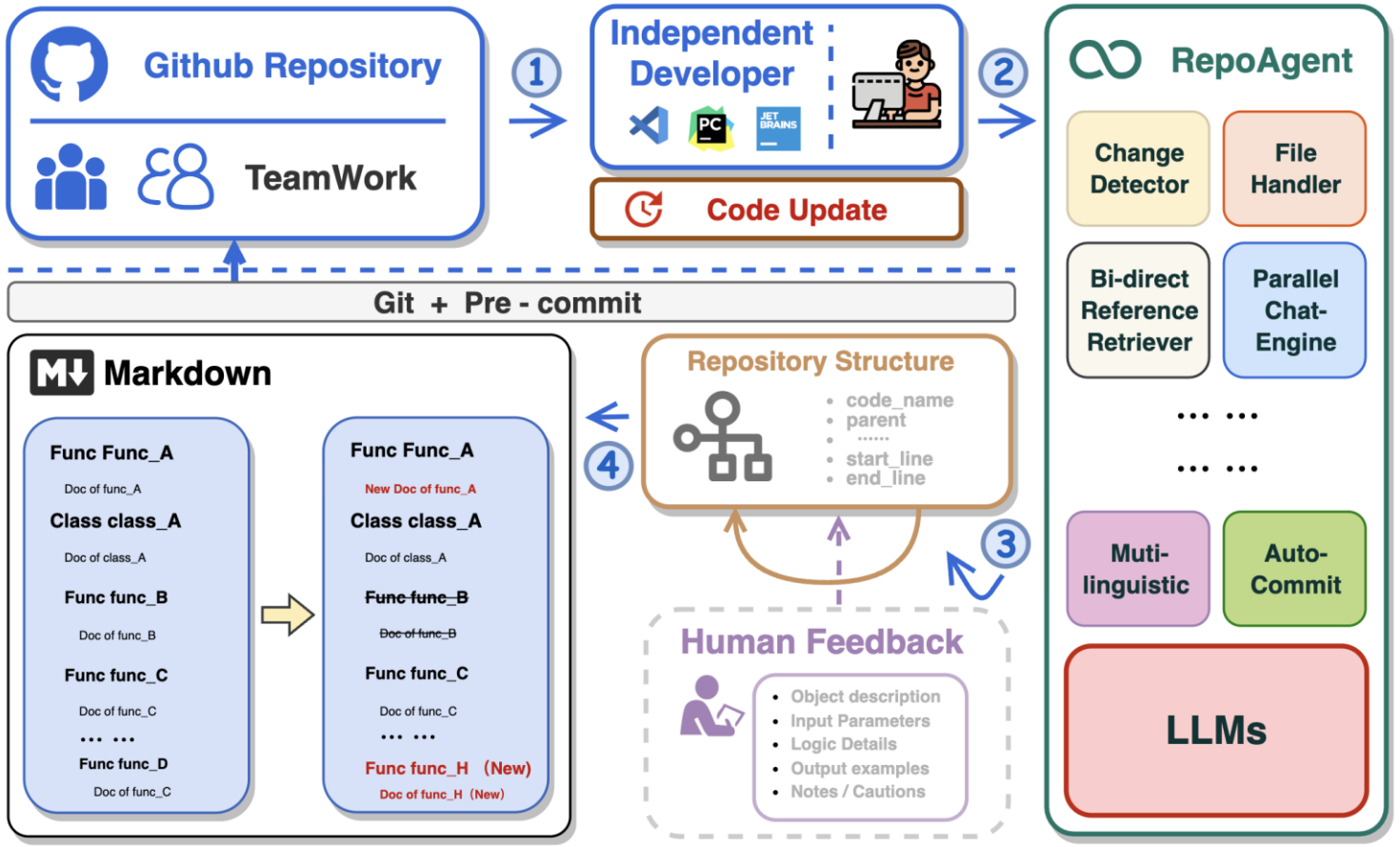
Why Tools?

- Extend capabilities beyond text



Why Tools?

- Enable real-world automation and agents



RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation, 2024

Why Tools?

- Generalize to new tasks
  - LLMs only need to learn to plan