Work In Progress: Incorporating Instructive Feedback in Human-AI Interaction

Feiyu (Gavin) Zhu Advised by Prof. Reid Simmons

Motivation

Learning Human Preference for Interaction:
- Humans have different preferences
- Accurate human model improves collaboration fluency & performance
- Respect human autonomy

Main Challenges:
- Reward function cannot be fully specified
- Limited data sample / demonstrations

Mainstream Approach:
Human-in-the-loop learning (HiLL)
- Attempt to reconstruct human's internal reward function from observation
- Mainly instance-based feedback
- Multiple interaction required for conveying simple ideas

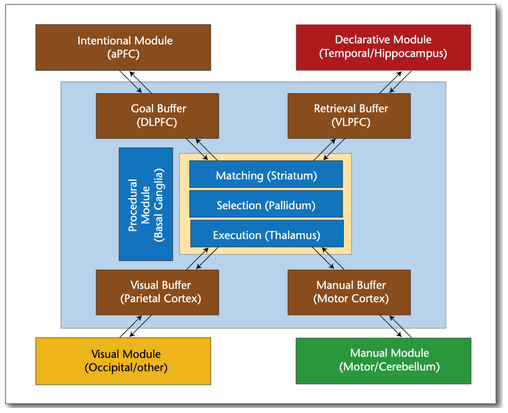
Research Question:

*Can we accelerate HiLL agent learning by incorporating **general** instructions from humans?*

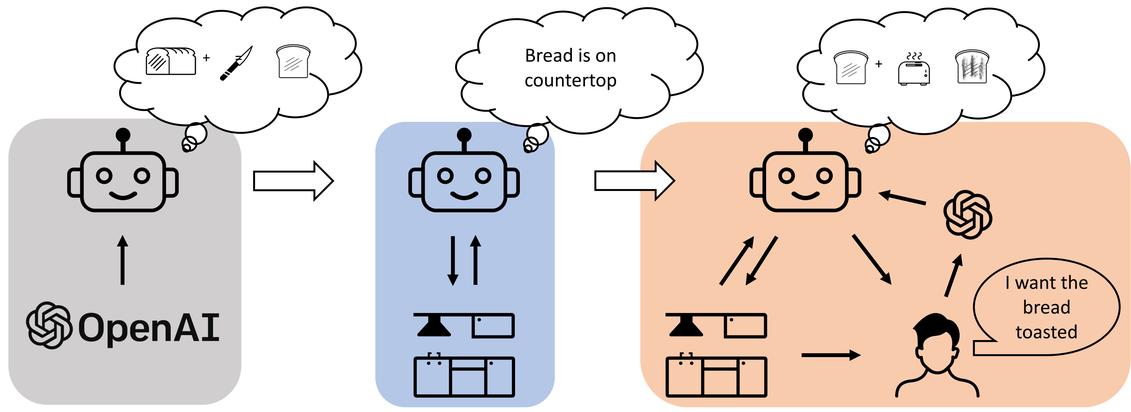
Background

Cognitive Architectures
- Consist of a goal stack, a set of declarative memories, a set of productions rules
- Focus on high-level decision-making process

✓ Model human cognition
✓ Support explicit memorization learning
× Labor-intensive design process



Approach Overview: Combine Cognitive Architecture and Large Language Models



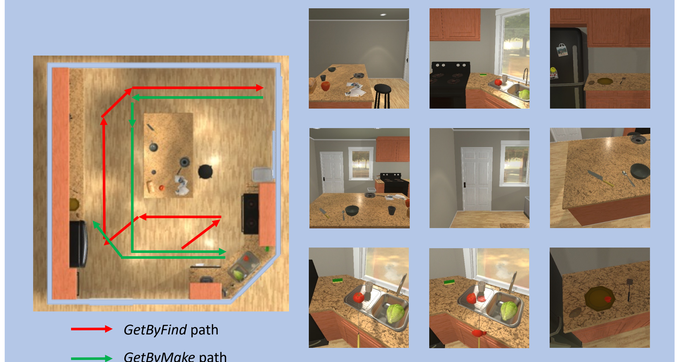
Phase 1: General Knowledge
Query large language models to initialize cognitive model
Focus on basic action primitives & general knowledge

Phase 2: Instance-based Knowledge
Interact with a specific environment
Learn by reinforcement on existing productions rules

Phase 3: Human Feedback
For error recovery & personal preference
Update robot decision-making by editing the productions

Case Study

Task 1: Put a slice of tomato on the plate



Learning from experience:
Downweigh *GetByFind* (as it fails the objective)
Upweight *GetByMake*
Memorize the locations of objects

Task 2: Put a slice of bread on the plate

Task 2 feedback:

When I say "get me a slice of bread", I mean "toasted bread"

I have two procedures for "get", but none of them apply
GetByMake: no recipe for toasted bread found
GetByFind: no toasted bread found in the kitchen

You can make a toasted bread by putting the bread in the toaster and turning the toaster on

Okay! Adding the following to my memory*:
1) When you want a slice of bread, you want it toasted
2) The recipe of making a slice of toast

Task 2 trial 2: Put a slice of bread on the plate

Future Work

Large Language Model Integration:
- Code generation for cognitive model initialization
- Translation between free-form natural language to production rule updates

User Study:
- Extend to collaborative task
- Task definition for human-AI teaming
- Participant recruitment
- Simulator supports

Summary

- Learning **human model** efficiently is crucial to human-AI collaboration
- Developing a framework to incorporate **general feedback** from human
- Cognitive architecture supports **fast adaption** and has good interpretability
- Large language models contain **general world knowledge**, and can **bridge** human and machine languages
- Preliminary kitchen case study and system workflow & learning capabilities

Icons in model diagram: uxwing.com
Kitchen simulator: ai2thor.allenai.org
Cognitive architecture: https://doi.org/10.1609/aimag.v38i4.2744