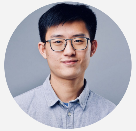
EMNLP
2024

Language Agents:
Foundations, Prospects, and Risks

Yu Su, OSU, @ysu_nlp

Diyi Yang, Stanford, @diyi_yang


Shunyu Yao, OpenAI, @ShunyuYao12

Tao Yu, HKU, @taoyds

Part II: Introduction

Yu Su

The rise, and the divide

Bill Gates

Agents are bringing about the **biggest revolution in computing** since we went from typing commands to tapping on icons.

Andrew Ng

I think AI agentic workflows will drive **massive AI progress** this year.

Sam Altman

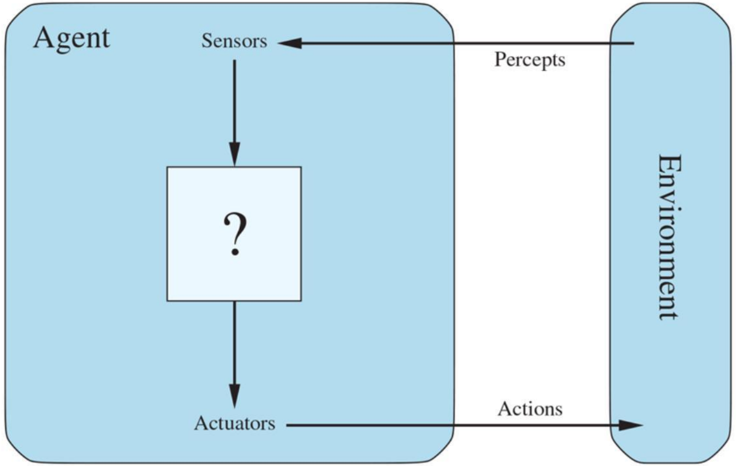
2025 is when **agents will work**.

Current agents are just **thin wrappers around LLMs**.

Autoregressive LLMs **can never reason or plan**.

Auto-GPT's limitations in … reveal that it is **far from being a practical solution**.

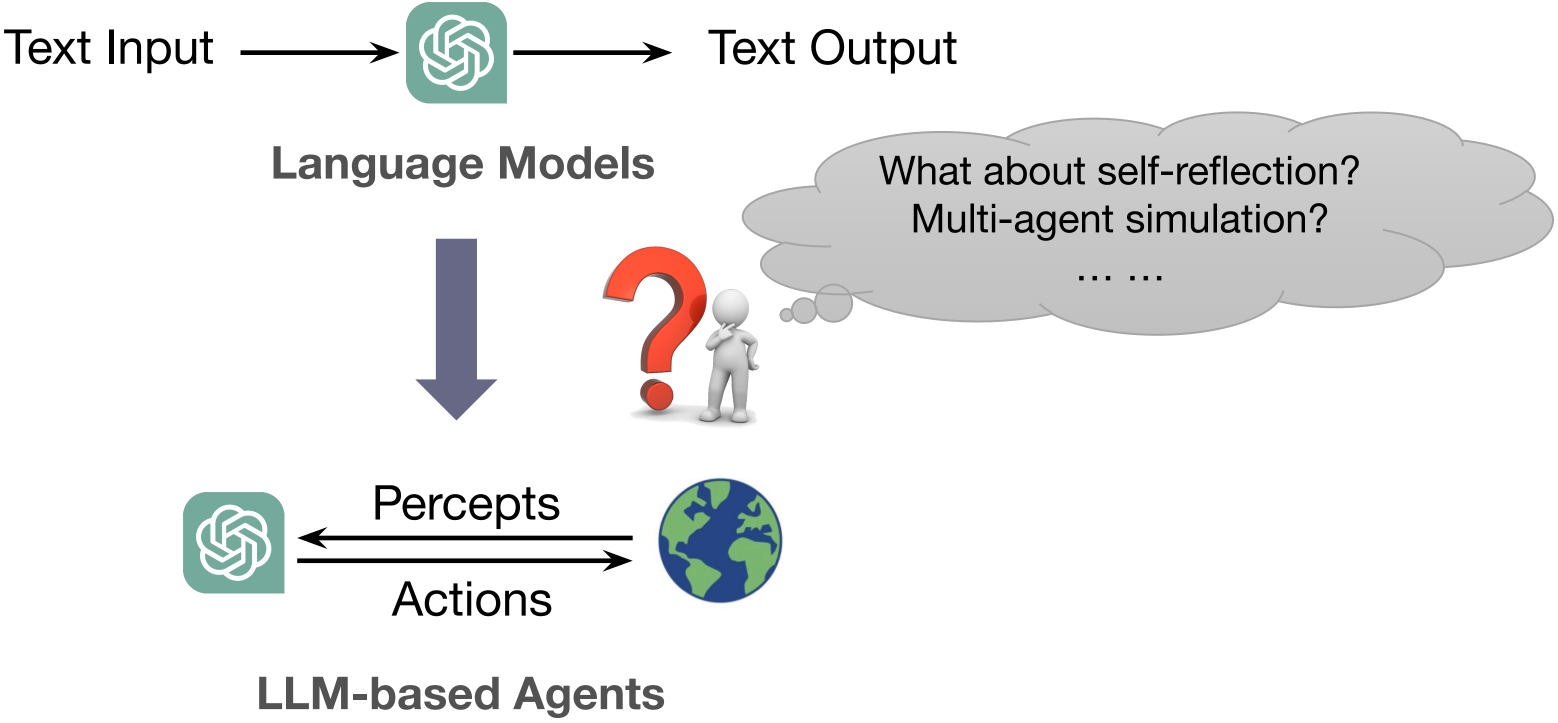
Why agents again?



“An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through actuators.”

–– Russell & Norvig, *AI: A Modern Approach* (2020)

‘Modern’ agent = LLM + external environment?



Text Input → Text Output

Language Models

What about self-reflection?
Multi-agent simulation?
… …

Percepts

Actions

LLM-based Agents

Two competing views

LLM-first view: We make an LLM into an agent!

- Implications: scaffold on top of LLMs, prompting-focused, heavy on engineering

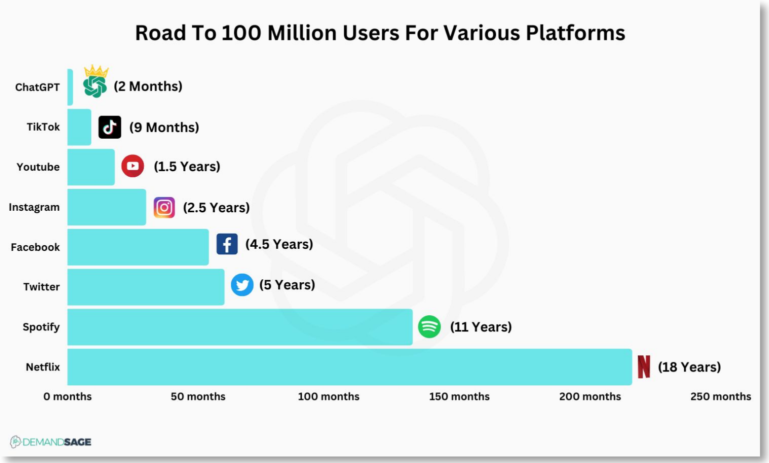
Agent-first view: We integrate LLMs into AI agents so they can use language for reasoning and communication!

- Implications: All the same challenges faced by previous AI agents (e.g., perception, reasoning, world models, planning) still remain, but we need to **re-examine them through the new lens of LLMs** and tackle new ones (e.g., synthetic data, self-reflection, internalized search)

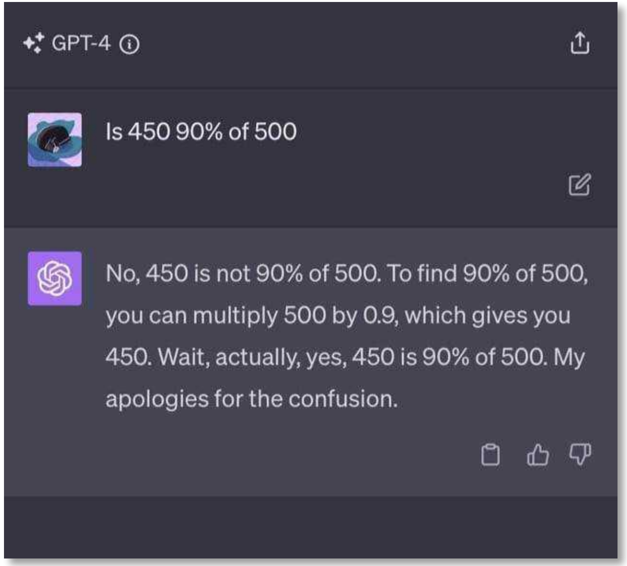
What’s fundamentally different now?

Contemporary AI agents, with integrated LLM(s), can *use language as a vehicle for reasoning and communication*

- Instruction following, in-context learning, output customization
- Reasoning (for better acting): state inferences, self-reflection, replanning, etc.



https://www.demandsage.com/chatgpt-statistics/

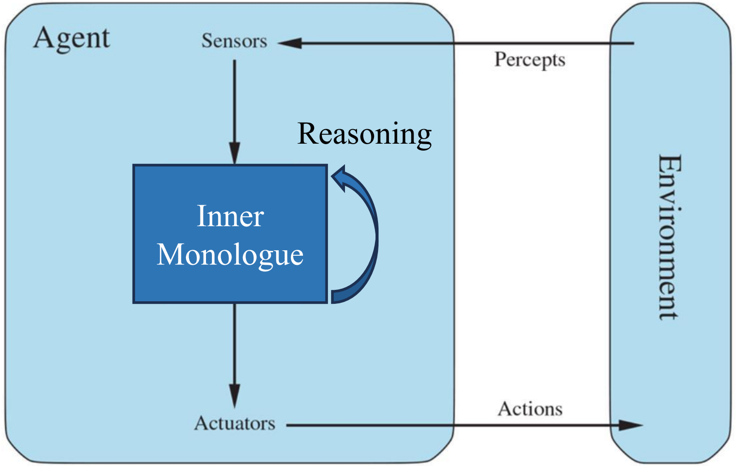


Is 450 90% of 500

No, 450 is not 90% of 500. To find 90% of 500, you can multiply 500 by 0.9, which gives you 450. Wait, actually, yes, 450 is 90% of 500. My apologies for the confusion.

https://www.reddit.com/r/ChatGPT/comments/16jvl4x/wait_actually_yes/

Reconciling with classic view of agents

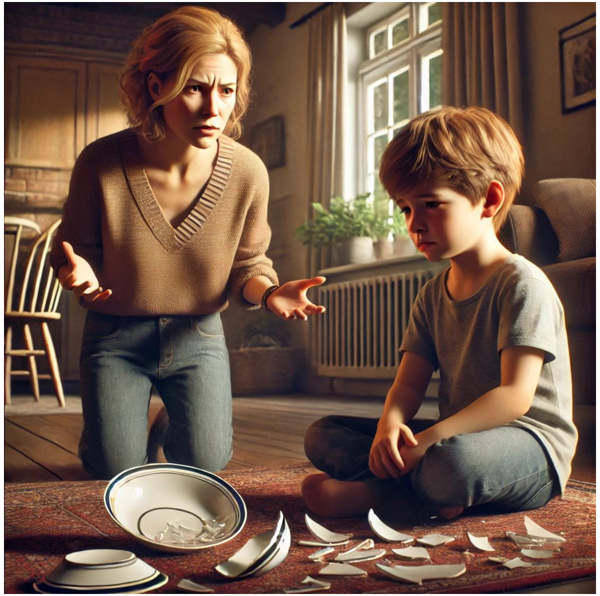


Adapted from Russell & Norvig (2020)

- **Reasoning by generating tokens** is a new type of action (*vs.* actions in external environments)
- **Internal environment**, where reasoning takes place in an inner monologue fashion
- **Self-reflection** is a 'meta' reasoning action (i.e., reasoning over the reasoning process), akin to metacognitive functions
- **Reasoning is for better acting**, by inferring environmental states, retrospection, dynamic replanning, etc.
- **Percept** and **external action spaces** are substantially expanded, thanks to using language for communication and multimodal perception

A generalized notion of ‘reasoning’

Unlike humans, LLMs (mostly) only have one mechanism (token generation) for perception, intuitive inferences, and symbolic reasoning; everything is effortful and takes a forward pass

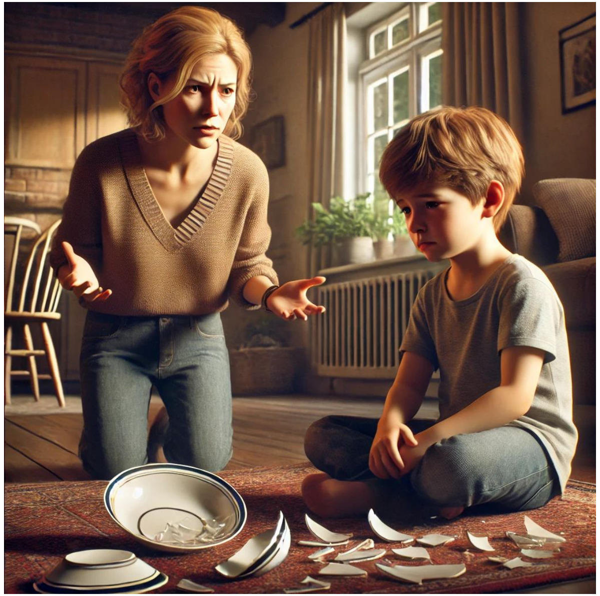


GPT-4o

In this image, we see a young boy sitting on the floor, looking down at a pile of broken dishes. He appears to be sad or remorseful. In front of him, a woman, likely his mother or caregiver, is kneeling and looking at him with a concerned, questioning expression. Her body language, with her hands open and her facial expression, suggests that she's asking him something or expressing surprise or frustration about the broken dishes.

A generalized notion of 'reasoning'

Unlike humans, LLMs (mostly) only have one mechanism (token generation) for perception, intuitive inferences, and symbolic reasoning; everything is effortful and takes a forward pass



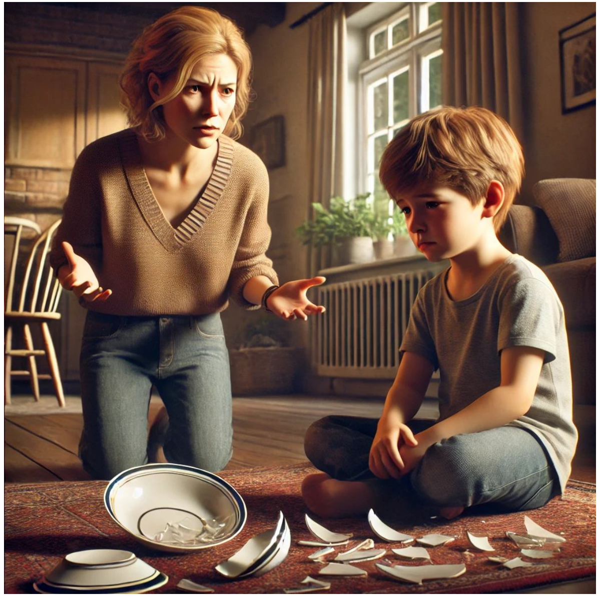
GPT-4o

In this image, we see a young boy sitting on the floor, looking down at a pile of broken dishes. He appears to be sad or remorseful. In front of him, a woman, likely his mother or caregiver, is kneeling and looking at him with a concerned, questioning expression. Her body language, with her hands open and her facial expression, suggests that she's asking him something or expressing surprise or frustration about the broken dishes.

Perception

A generalized notion of ‘reasoning’

Unlike humans, LLMs (mostly) only have one mechanism (token generation) for perception, intuitive inferences, and symbolic reasoning; everything is effortful and takes a forward pass



GPT-4o

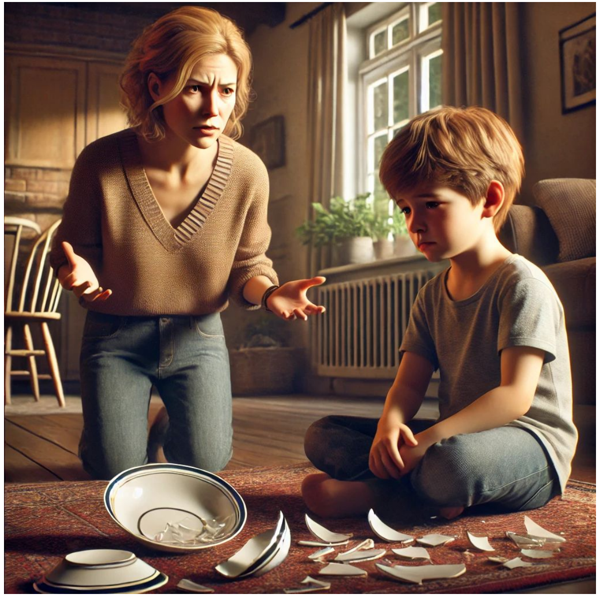
In this image, we see a young boy sitting on the floor, looking down at a pile of broken dishes. He appears to be sad or remorseful. In front of him, a woman, likely his mother or caregiver, is kneeling and looking at him with a concerned, questioning expression. Her body language, with her hands open and her facial expression, suggests that she's asking him something or expressing surprise or frustration about the broken dishes.

Perception

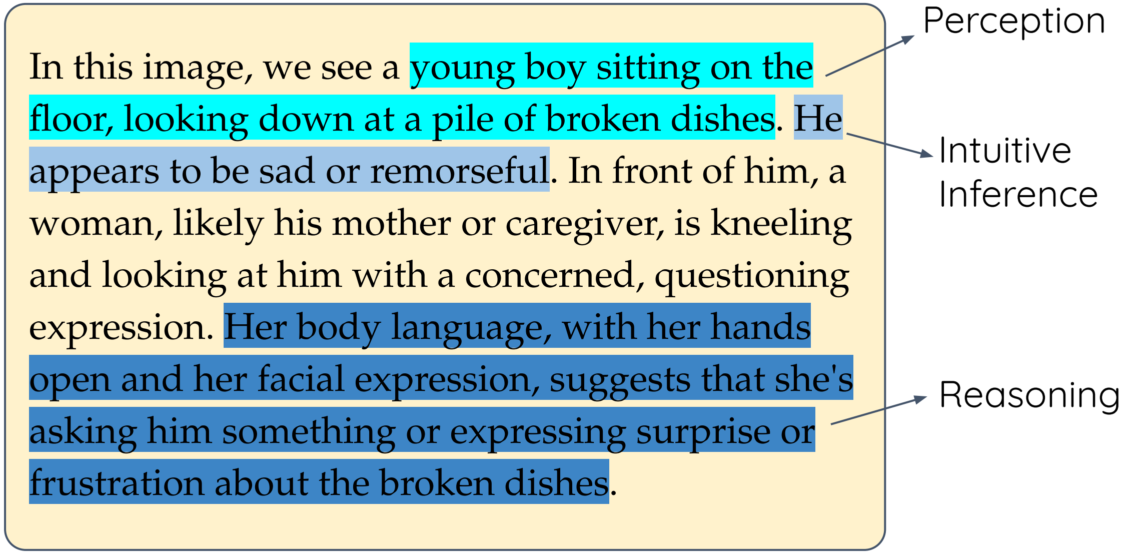
Intuitive Inference

A generalized notion of ‘reasoning’

Unlike humans, LLMs (mostly) only have one mechanism (token generation) for perception, intuitive inferences, and symbolic reasoning; everything is effortful and takes a forward pass



GPT-4o



In this image, we see a young boy sitting on the floor, looking down at a pile of broken dishes. He appears to be sad or remorseful. In front of him, a woman, likely his mother or caregiver, is kneeling and looking at him with a concerned, questioning expression. Her body language, with her hands open and her facial expression, suggests that she's asking him something or expressing surprise or frustration about the broken dishes.

Perception

Intuitive Inference

Reasoning

One may alternatively call this ‘thought’ to avoid the over-loaded term of ‘reasoning,’ at the risk of further anthropomorphizing machines

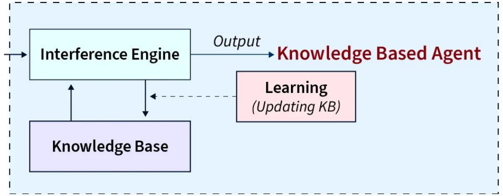
Language agents: a new type of AI agents

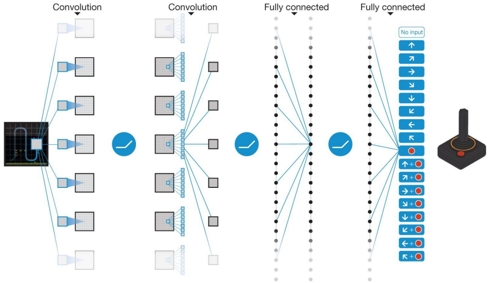
- These contemporary AI agents capable of using language for reasoning and communication are best called "**language agents**." They are qualitatively a different type of AI agents with language being their most distinct trait.

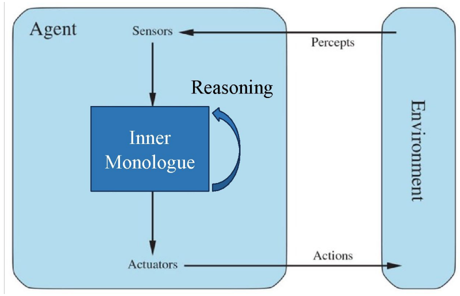
- What about *multimodal agents*?
  - While there's perception of other input modalities, language is still doing the heavy lifting (i.e., reasoning and communication)

- What about simply *LLM agents*?
  - The key is using language for reasoning and communication, but that doesn't have to come from an LLM; that may turn out to be a means to an end
  - Maybe in a few years, we will move beyond LLMs, but the need for universal language understanding and production in agents will remain

Evolution of AI agents


Interference Engine
Output
Knowledge Based Agent
Learning
(Updating KB)
Knowledge Base

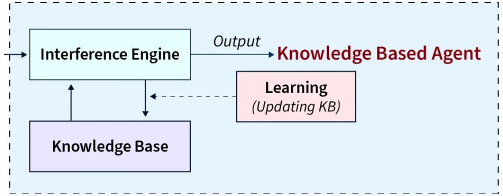


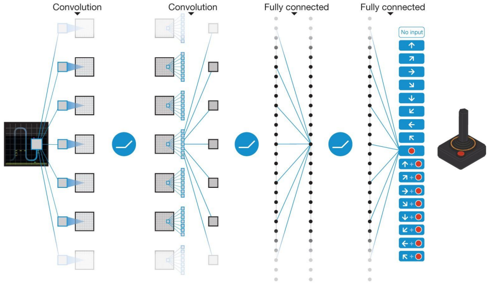

Agent
Sensors
Percepts
Reasoning
Inner Monologue
Environment
Actuators
Actions

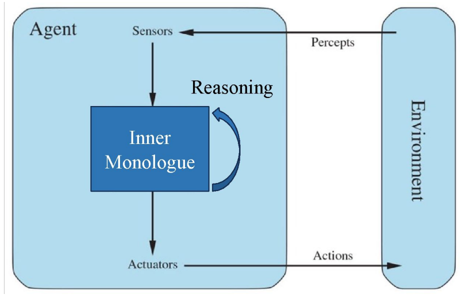
	Logical Agent	Neural Agent	Language Agent
Expressiveness			
Reasoning			
Adaptivity			

Image sources: https://www.scaler.com/topics/artificial-intelligence-tutorial/knowledge-based-agent/,
Mnih et al., "Human-level control through deep reinforcement learning." Nature (2015)

Evolution of AI agents



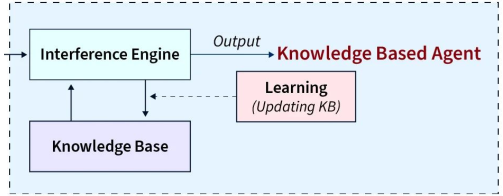


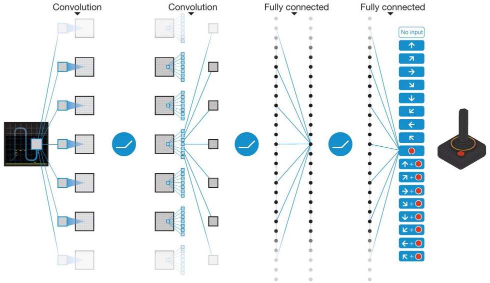


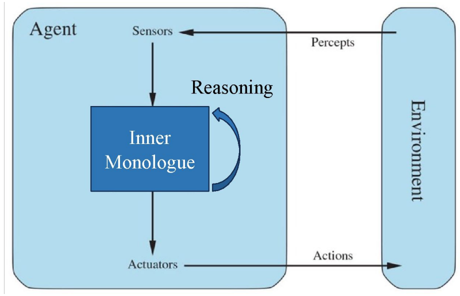
	Logical Agent	Neural Agent	Language Agent
Expressiveness	Low bounded by the logical language		
Reasoning	Logical inferences sound, explicit, rigid		
Adaptivity	Low bounded by knowledge curation		

Image sources: https://www.scaler.com/topics/artificial-intelligence-tutorial/knowledge-based-agent/,
Mnih et al., “Human-level control through deep reinforcement learning.” Nature (2015)

Evolution of AI agents







Logical Agent

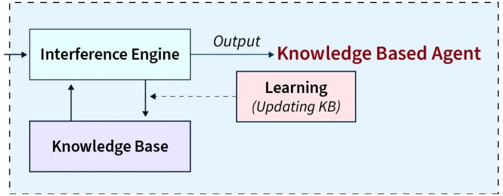
Neural Agent

Language Agent

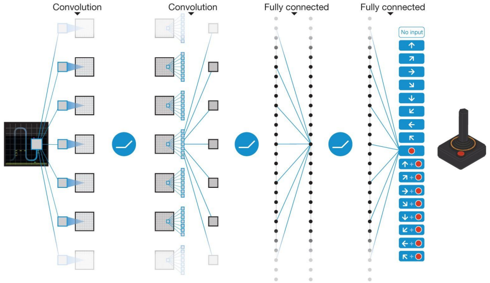
Expressiveness	Low bounded by the logical language	Medium anything a (small) NN can encode	
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	
Adaptivity	Low bounded by knowledge curation	Medium data-driven but sample inefficient	

Image sources: https://www.scaler.com/topics/artificial-intelligence-tutorial/knowledge-based-agent/, Mnih et al., “Human-level control through deep reinforcement learning.” Nature (2015)

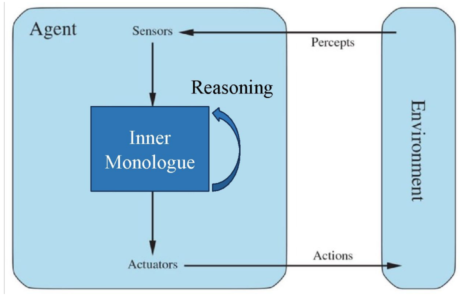
Evolution of AI agents


Interference Engine
Output
Knowledge Based Agent
Learning
(Updating KB)
Knowledge Base

Logical Agent



Neural Agent

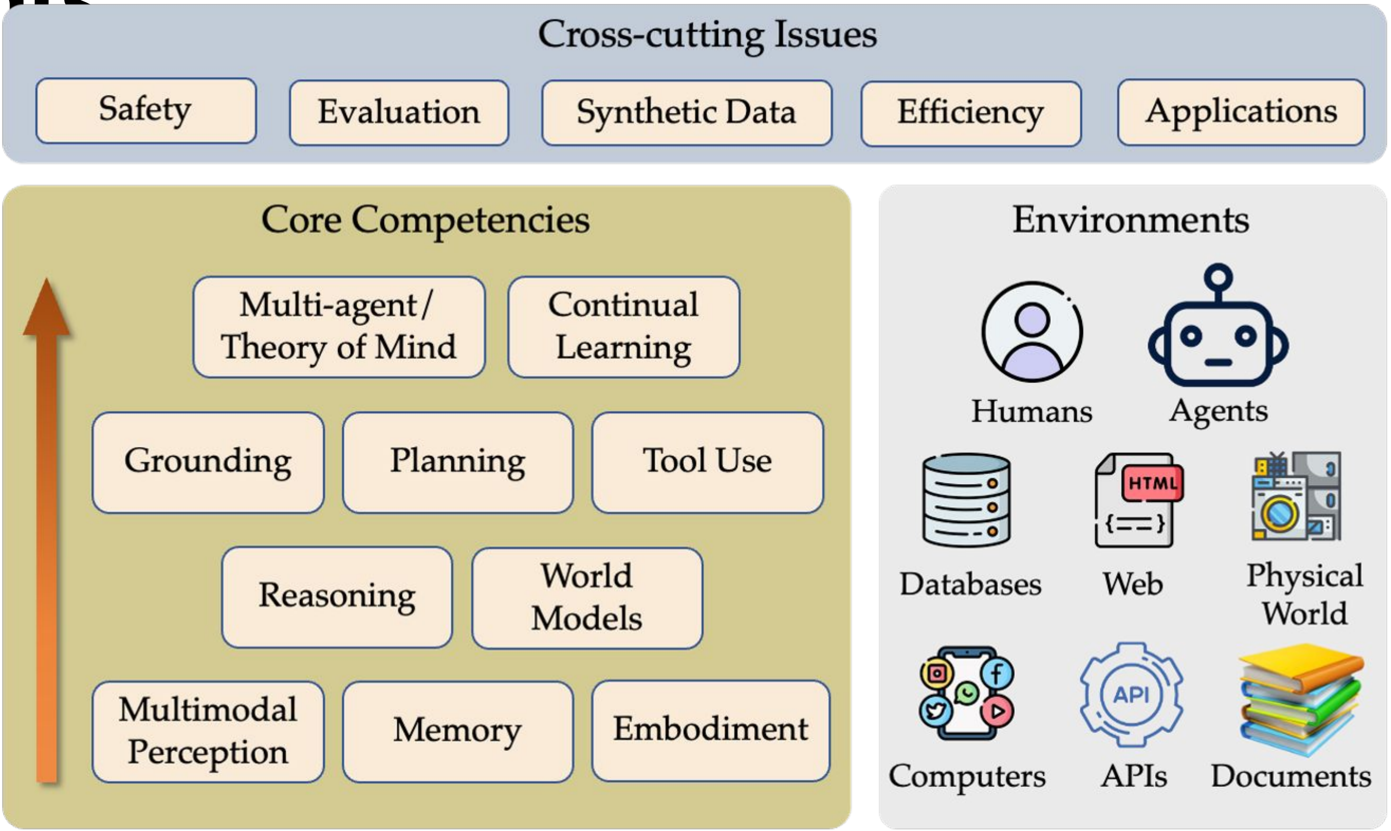

Agent
Sensors
Percepts
Reasoning
Inner Monologue
Environment
Actuators
Actions

Language Agent

Expressiveness	Low bounded by the logical language	Medium anything a (small) NN can encode	High almost anything, esp. verbalizable parts of the world
Reasoning	Logical inferences sound, explicit, rigid	Parametric inferences stochastic, implicit, rigid	Language-based inferences fuzzy, semi-explicit, flexible
Adaptivity	Low bounded by knowledge curation	Medium data-driven but sample inefficient	High strong prior from LLMs + language use

Image sources: https://www.scaler.com/topics/artificial-intelligence-tutorial/knowledge-based-agent/,
Mnih et al., “Human-level control through deep reinforcement learning.” Nature (2015)

A conceptual framework for language agents



Overview

2:00-2:20 **Part I:** Introduction [20 mins]

2:20-3:20 **Part II:** Foundations: Reasoning, Memory, and Planning [60 mins]

3:20-3:30 **Q&A** [10 mins]

3:30-4:00 **Coffee Break** [30 mins]

4:00-4:45 **Part III:** Applications, Data, and Evaluation [45 mins]

4:45-5:20 **Part IV:** Emerging Topics: Multi-Agent Systems, Safety, and Social Impact [35 mins]

5:20-5:30 **Part V:** Final Remarks and Outlook + Q&A [10 mins]

Disclaimers

This tutorial is

- to offer a **clear definition** and **conceptual framework** for language agents, properly situated in the historical context
- to have a careful discussion on **foundational competencies**, **exciting applications**, and **emerging issues**
- really, our sincere attempt to **get you excited about language agents** as much as we do!

This tutorial is *not*

- a **comprehensive survey** covering as much related work as possible
- a **practitioner's guide** that focuses on code frameworks and best practices

This is a still an emerging topic and we welcome constructive discussions!