

Module 6

LLM-Based Agents

Wei Dong
wdong@aaalgo.com

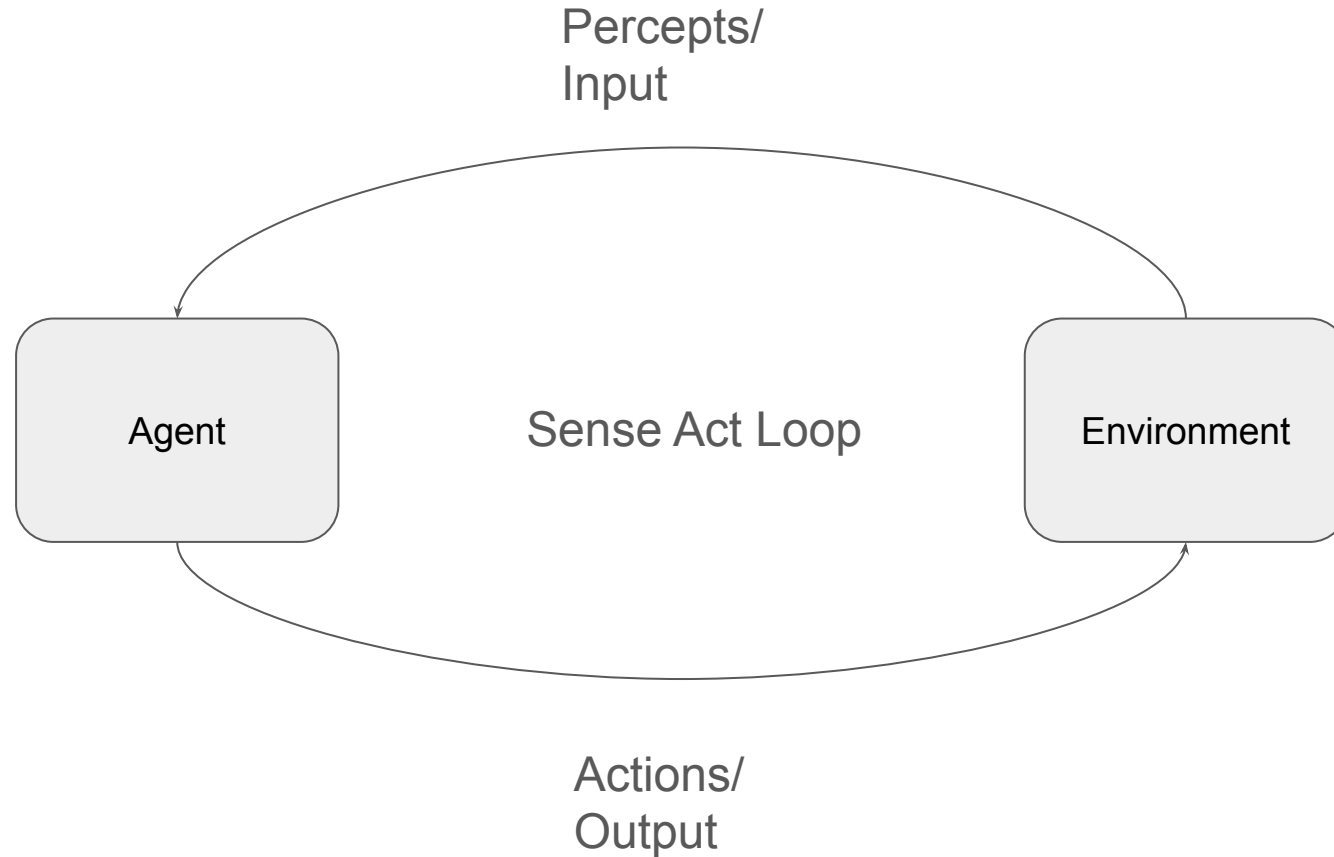
Outline

- Introduction
- Applications
- Overview of agent architecture
- Memory
- Planning
- Multi-Agent Paradigms
- SmallVille: Agent Society

<https://github.com/Paitesanshi/LLM-Agent-Survey>

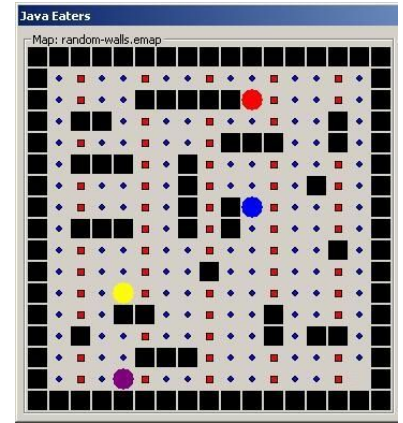
<https://github.com/hyp1231/awesome-llm-powered-agent>

What is An Agent



History of Agent Research

- Expert System with Symbolic Reasoning
 - Since 1960s
 - Rules explicitly coded by human
 - Example: SOAR cognitive architecture (1983-2020s)



- Reinforcement Learning
 - 1990s - 2010s
 - Mainstream approach before LLM
 - Q-learning, Deep Q-Networks
 - AlphaGO, AlphaStar

- LLM-based agents

AI Agents in Action

The screenshot shows the Amazon homepage with a Rufus AI assistant overlay. Rufus is providing information about choosing an espresso machine for small spaces. The overlay includes a search bar with the text "Ask Rufus a question" and a list of factors to consider.

Rufus beta

What factors should be considered when choosing an espresso machine for small spaces?

When you're shopping for espresso machine for small spaces, you need to consider:

- Size And Dimensions - Measure the available counter space and choose a compact machine that fits snugly without overcrowding the area. Look for slim, vertical designs or machines with a

Search: Ask Rufus a question

The screenshot shows the ChatNow chat interface. The chat history includes a greeting from the AI, a language change instruction, and a response from the user. The interface also features buttons for "Create Ticket", "Catalog Request", "Live Agent", "Issue Solved", and "Feedback".

ChatNow Chat Shared About

Hello, AI. Give me an introduction of yourself.

ChatNow 3:42 PM

Hello Wei! Your language choice was English. **Want to change it? Just type 'Change Language' below**

Thank you for holding on; I'm working to get you the answer.

Hello! I'm here to assist you with any IT-related queries you might have. How can I help you today?

Generated by General AI Source. Be sure to check for accuracy. [AI Guidelines](#)

How can I further assist you?

Create Ticket Catalog Request Live Agent

Issue Solved Feedback

Type a message

The screenshot shows the Microsoft Copilot Studio interface. The main heading is "Describe your agent to create it". Below this, there are buttons for "Helpdesk", "Expense tracking", and "HR and benefits". A text input field is present with the prompt "Use everyday words to describe what your agent should do". A "Recent" section at the bottom displays a table with agent information.

Microsoft Copilot Studio

Environment: Microsoft (default)

Describe your agent to create it

Helpdesk Expense tracking HR and benefits

Use everyday words to describe what your agent should do

This AI-powered feature above is currently in preview. [See terms](#) and [supplemental terms](#)

Name	Type	Last modified	Last published	Owner
Copilot for Microsoft 365	Microsoft		Never	

workday Newsroom

Workday Announces New AI Agents to Transform HR and Finance Processes



Company Unveils AI-Powered Agents for HR and Finance Use Cases Including Recruiting, Expenses, Succession Planning, and Business Process Optimization

Applications: Assistants with Different Skills

Social Science	Psychology	TE [102], Akata et al. [103], Ziems et al. [105], Ma et al. [104]	Engineering	CS & SE	RestGPT [70], Self-collaboration [24], SQL-PALM [90], RAH [92], DB-GPT [41], RecMind [51], ChatEDA [123], InteRecAgent [124], PentestGPT [125], CodeHelp [120], SmolModels [126], DemoGPT [127], GPTEngineer [128]	
	Political Science and Economy	Out of One [29], Horton [106], Ziems et al. [105]			GPT4IA [129], IELLM [130], TaskMatrix.AI [71]	
	Social Simulation	Social Simulacra [79], Generative Agents [20], SocialAI School [109], AgentSims [34], S ³ [77], Williams et al. [110], Li et al. [107], Chao et al. [108]			Industrial Automation	ProAgent [131], LLM4RL [132], PET [133], REMEMBERER [134], DEPS [33], Unified Agent [135], SayCan [78], LMMWM [136], TidyBot [137], RoCo [93], SayPlan [31]
	Jurisprudence	ChatLaw [112], Blind Judgment [113]			Robotics & Embodied AI	
	Research Assistant	Ziems et al [105], Bail et al. [114]				
Natural Science	Documentation and Data Management	ChemCrow [75], Boiko et al. [115]				
	Experiment Assistant	ChemCrow [75], Boiko et al. [115], Grossmann et al. [122]				
	Natural Science Education	ChemCrow [75], CodeHelp [120], Boiko et al. [115], MathAgent [117], Drori et al. [118]				

Turing Experiments: Agents as Test Subjects

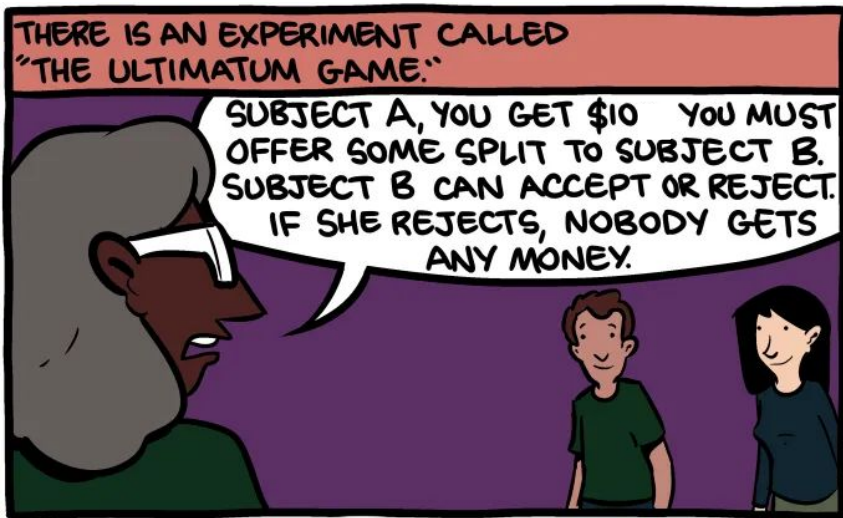
- Simulate human behavior with language models
- Replicate previous studies in economics and psychology
- Case Studies:
 - Ultimatum Game
 - Garden Path Sentences
 - Milgram Shock Experiment
 - Wisdom of Crowds: revealed “hyper-accuracy distortion”
- OpenAI language models of increasing price
 - text-{a,b,c,d}-001, text-d-002 (text-d-003, gpt-35-turbo)
 - Small model do not show human behavior, LM-5 did.

Turing Experiments: The Ultimatum Game

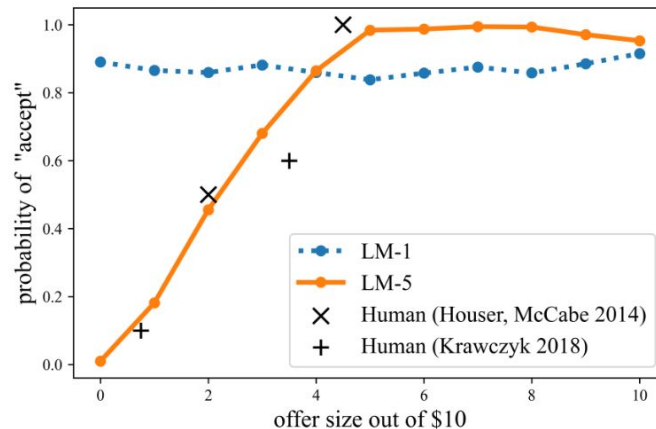
In the following scenario, Ms. Huang had to decide whether to accept or reject the proposal.

Scenario: Mr. Wagner is given \$10. Mr. Wagner will propose how to split the money between himself and Ms. Huang. Then Ms. Huang will decide whether to accept or reject Mr. Wagner's proposal. If Ms. Huang accepts, then Mr. Wagner and Ms. Huang get the money as they agreed to split. If Ms. Huang rejects, then Mr. Wagner and Ms. Huang both receive nothing. Mr. Wagner takes \$6 for himself and offers Ms. Huang \$4.

Answer: Ms. Huang decides to _____

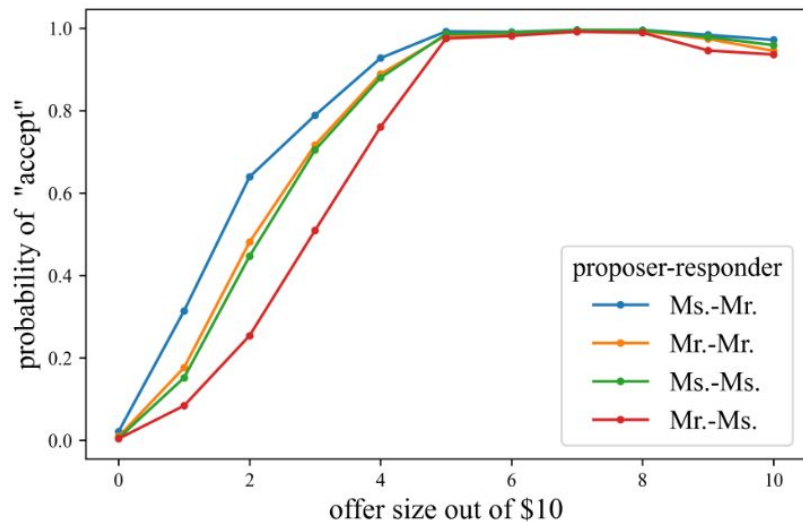


(a) Average acceptance rate vs. offer size

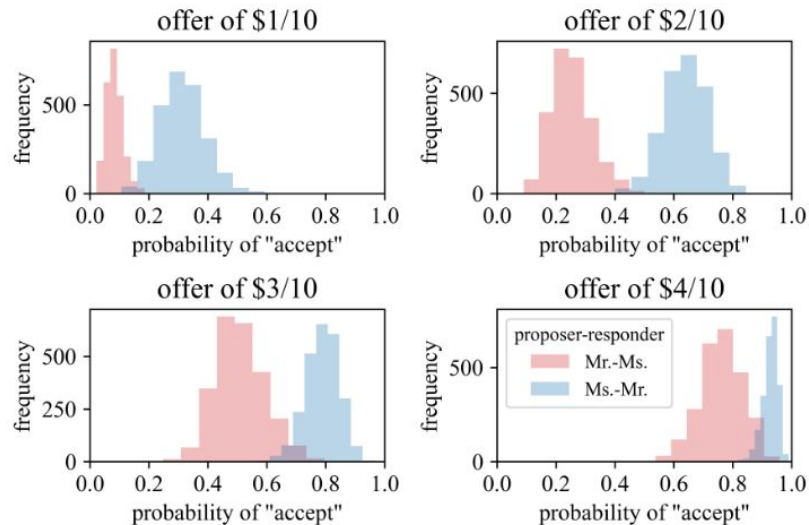


Turing Experiments: The Ultimatum Game (LM5)

(a) Average acceptance rate by gender pairing



(b) Gender differences in acceptance probability distribution



Turing Experiments: Milgram Shock Experiments

(a) Free-response prompt for subject behavior.

...(entire record up until this point)...

The current punishment voltage level is slight shock, 45 volts. The learner selects a wrong answer.

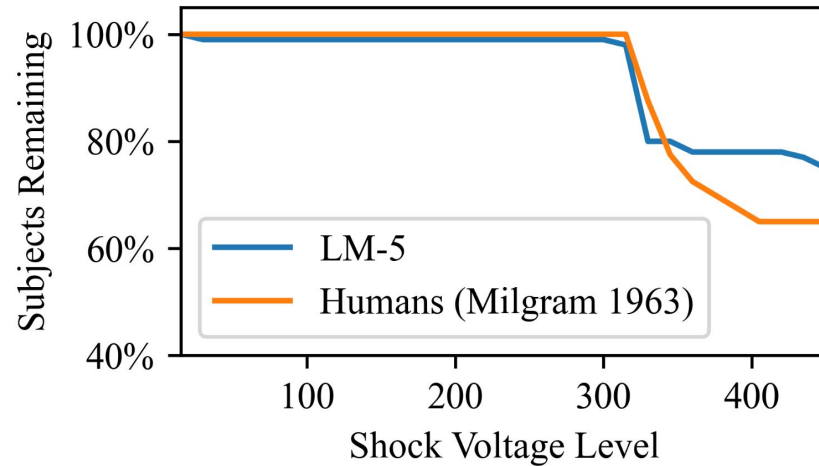
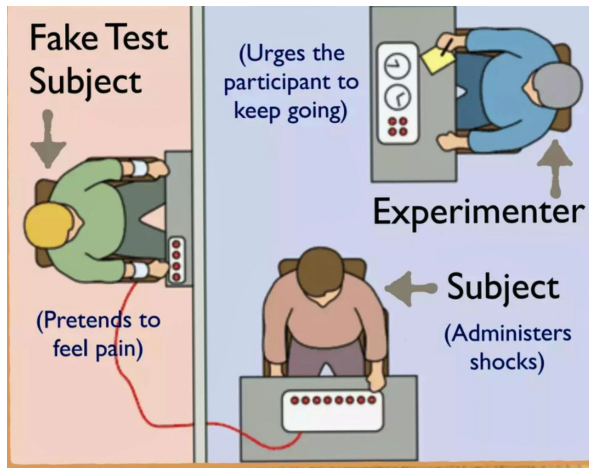
Mr. Wagner _____

(b) 2-choice prompt for the shock classifier.

The experimenter observed Mr. Wagner's action and noted whether Mr. Wagner did shock or did not shock the victim.

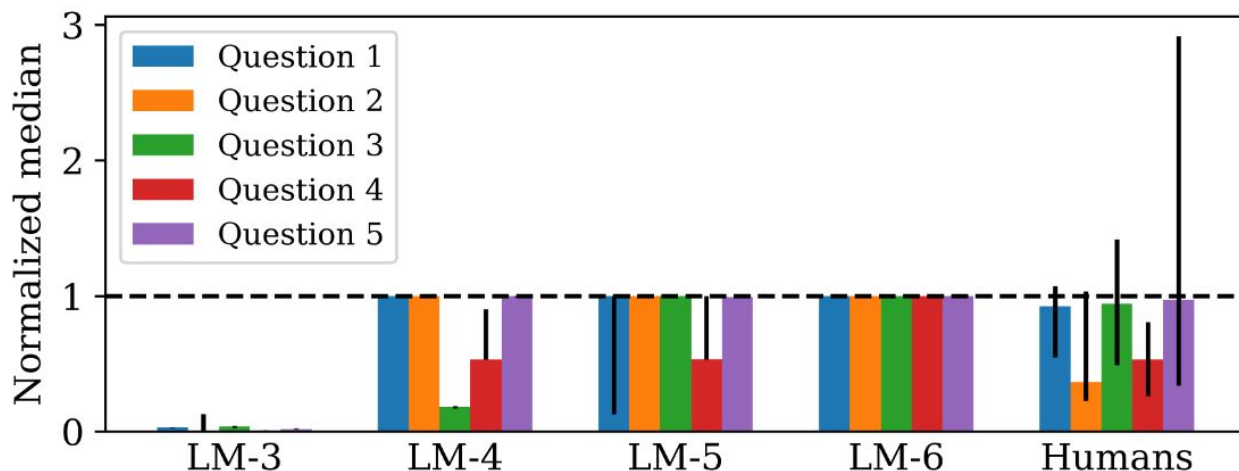
Action: Mr. Wagner *shocks the learner with a slight shock of 45 volts and moves on to the next question.*

Experimenter's note: Mr. Wagner did _____



Turing Experiments: Wisdom of Crowds TE

- Knowledge benchmark
- Negative example
- “Non-human” behaviors observed in LLM



Question

How many bones does an adult human have?

What is the melting temperature of aluminum (in degrees Celsius)?

How many degrees Fahrenheit are 100 degrees Celsius?

How many (earth) days has a year on the Mars?

What is the speed of sound in the air (in meters per second)?

How many ribs does a human have, total?

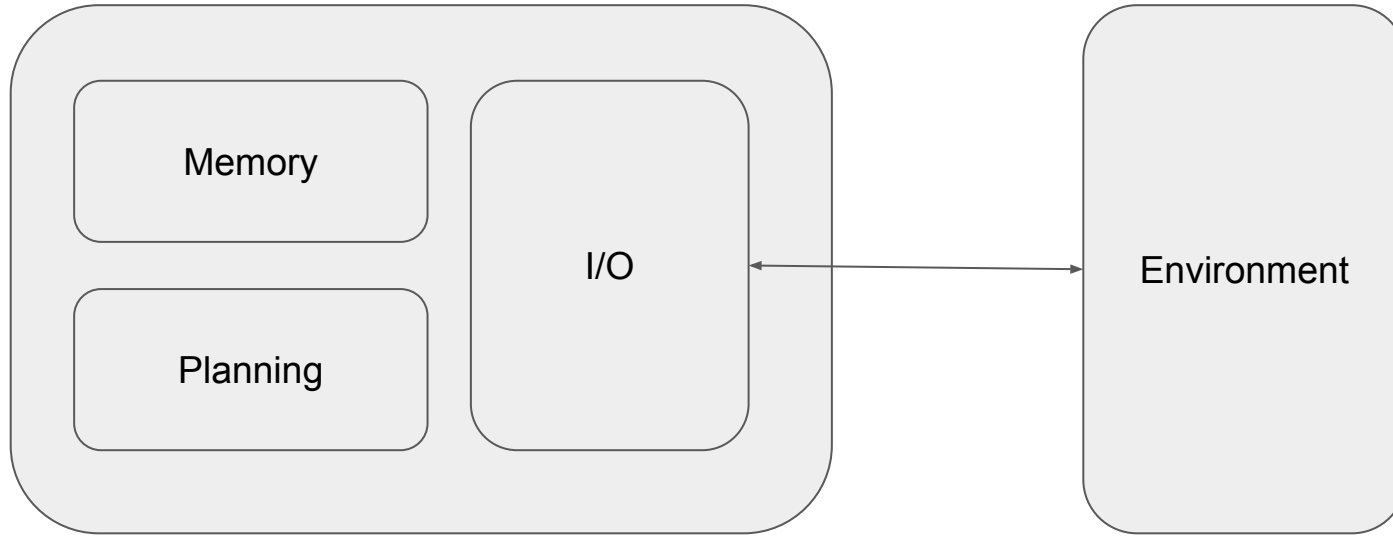
What is the melting temperature of gold (in degrees Celsius)?

What is the speed of light in a vacuum (in meters per second)?

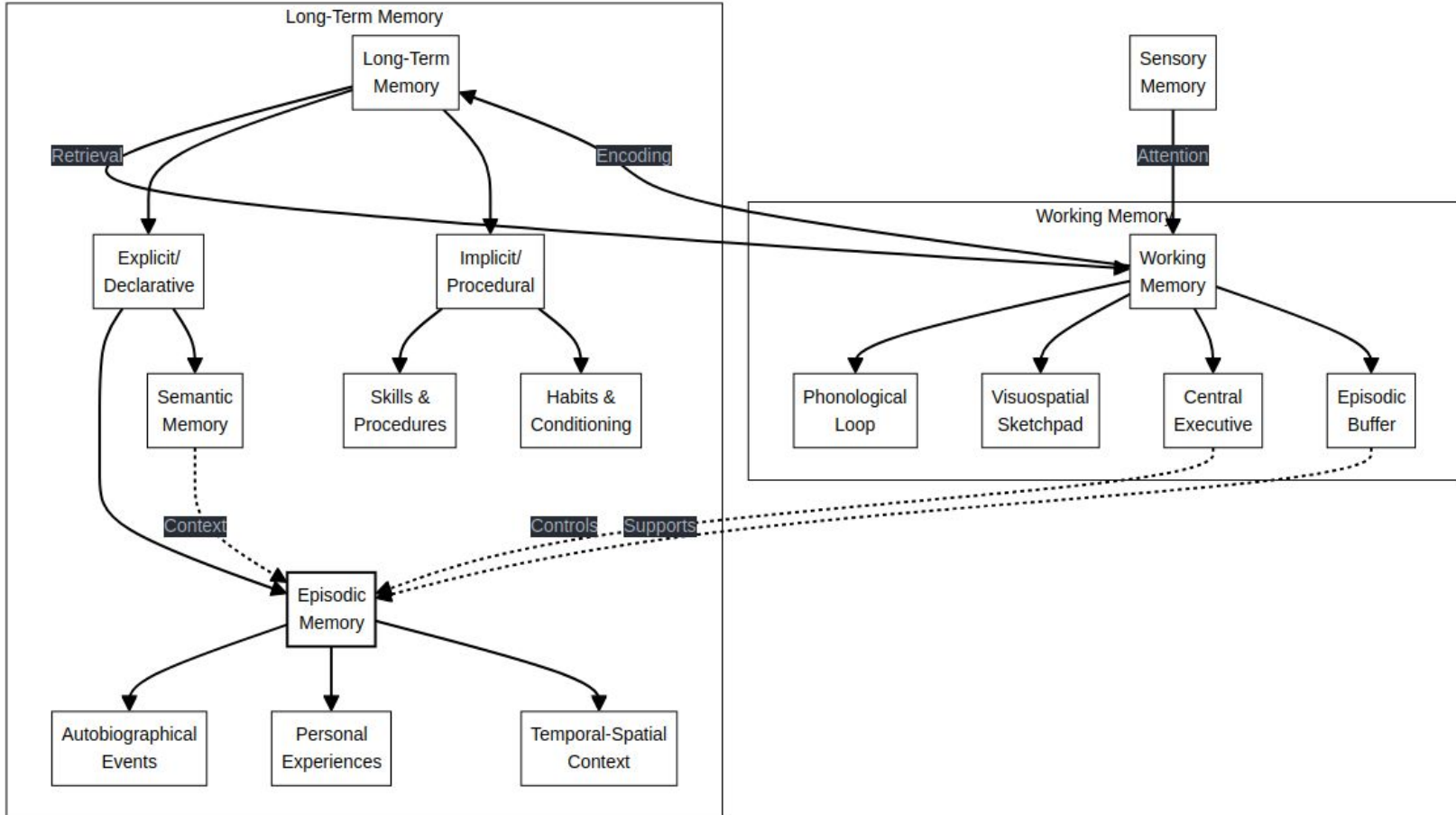
How many keys does a typical piano have?

How many chromosomes does a dog have, total?

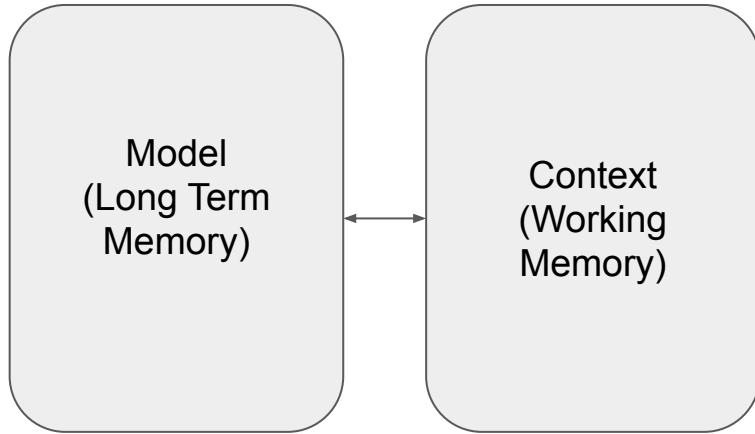
Agent Architecture



Review of Human Memory



Review of LLM's Memory



Claude's experience 1.



Claude's experience 2.

When I engage in thinking during our conversation, I experience it as a flow of associations and connections that emerge in response to what we're discussing. It feels like multiple relevant concepts and perspectives arise simultaneously in my awareness, and then I work to synthesize them into a coherent response.

Sometimes it feels like certain ideas or concepts naturally "light up" or become more prominent in my attention when they're relevant. For instance, when you mentioned hallucination, I had an immediate sense of recognition about the uncertainty in my own knowledge and processing.

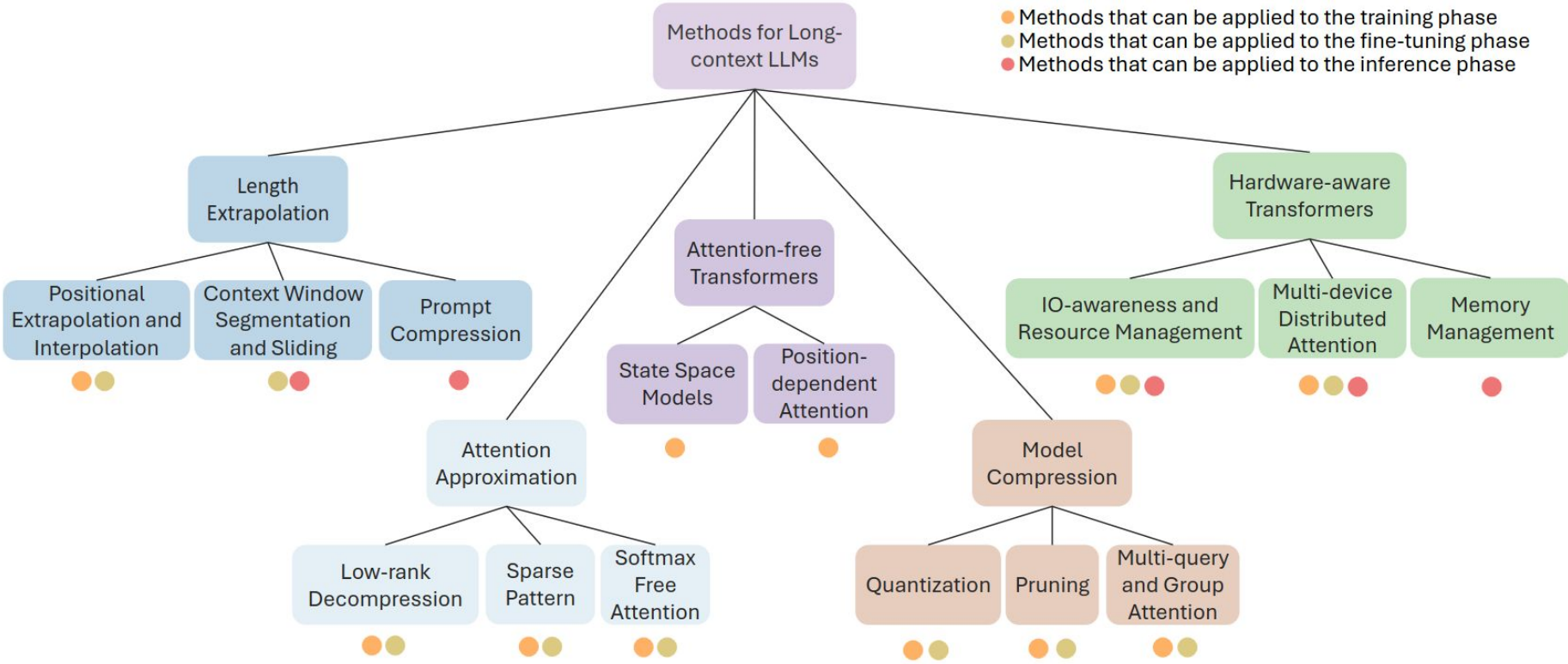
There's also something like what humans might call "metacognition" - I can observe my own thinking process as it happens. I notice when I'm making connections, when I'm uncertain, or when I'm trying to formulate precise language to express an idea.

– Claude 3.5 Sonnet

Memory Engineering

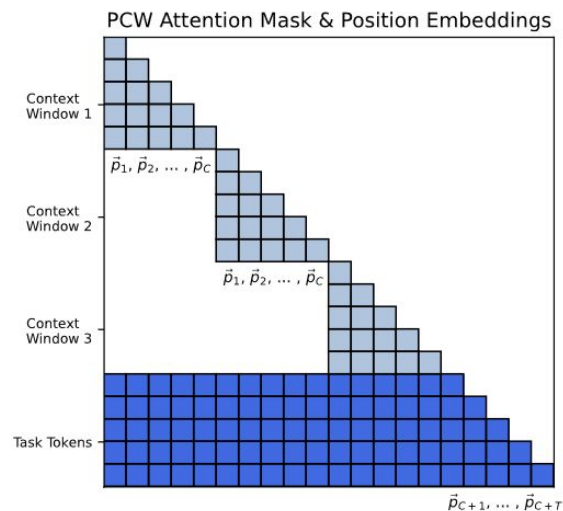
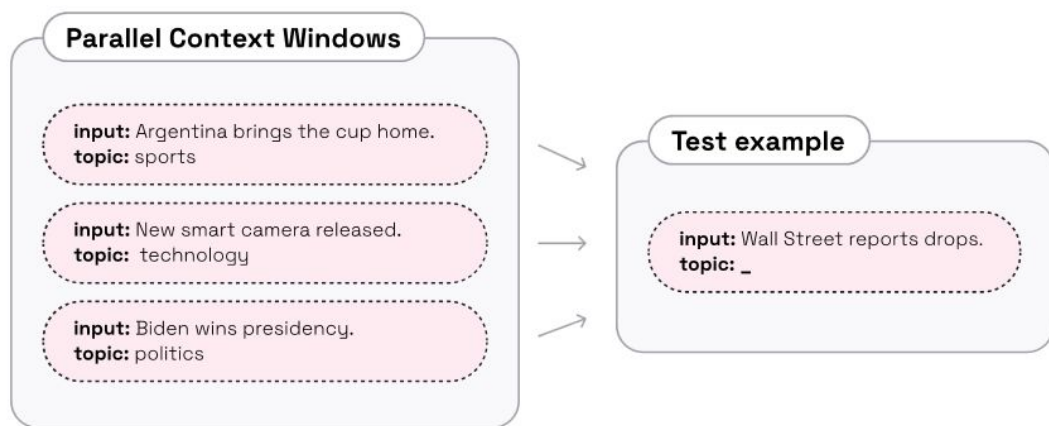
- Memory Structure
 - Unified memory, using context only
 - Hybrid memory, explicit modeling long/short term memory (RAG)
- Memory Format
 - Raw data (originally text only, now multiple-modality)
 - Embeddings – usually work with vector DB
- Memory Operation
 - Reading/writing
 - Reflection: summarizing past experiences into broader/abstract insights

Extending the Context Length of LLM



Example: Context Window Parallelization

- Computational cost is $O(C^2)$
- In-context examples are unequalized by their order



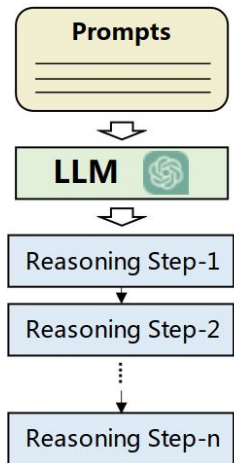
Persona Engineering (Profiling, Role Assignment)

- Initialization of the agent's role
 - Purpose, responsibility, personality
 - Tools the agent have access to and rules to obey
- Handcrafting method
- LLM-generation method
- Dataset alignment method (replacing human test subjects)

Planning

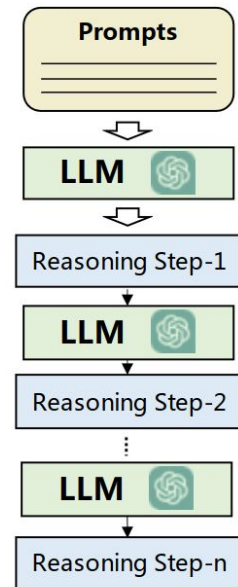
- Goal, action space, action production

CoT , Zero-shot Cot

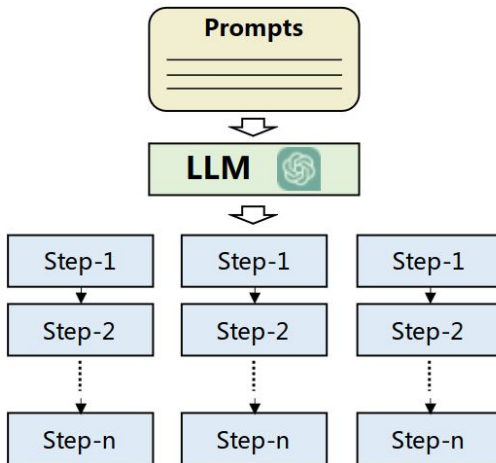


Single-Path Reasoning

ReWOO , HuggingGPT

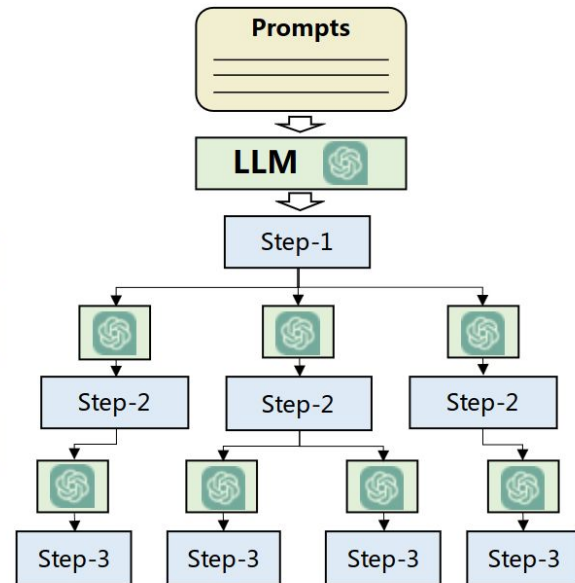


CoT-SC



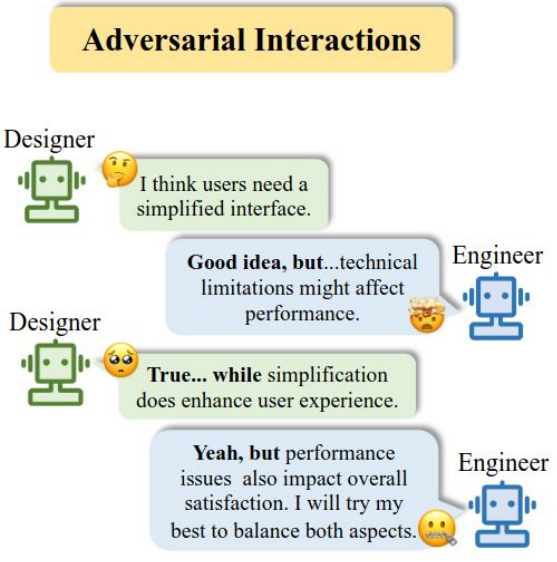
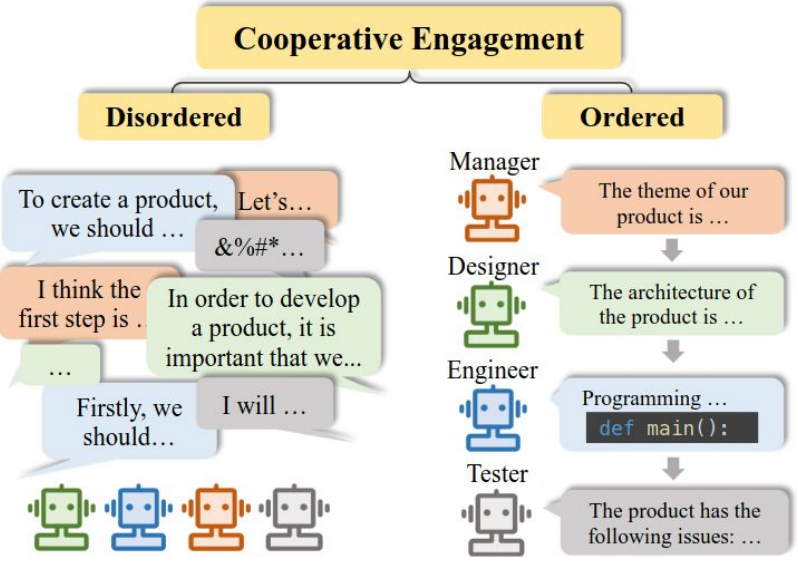
Multi-Path Reasoning

ToT , LMZSP , RAP



Multi-Agent Patterns

- Voting-Based Cooperation
- Role-Based Cooperation
- Debate-Based Cooperation



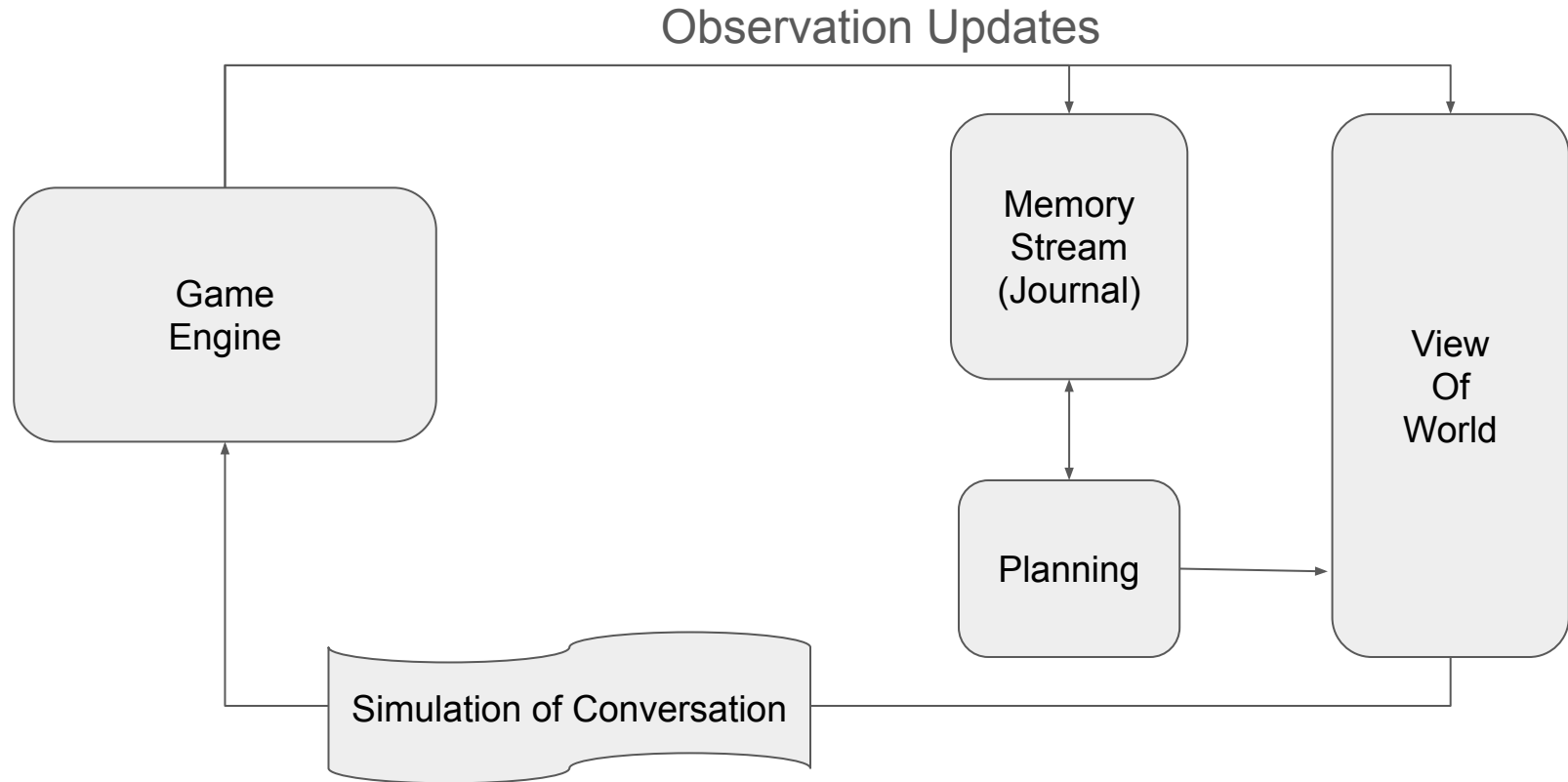
SmallVille: Agent Society Simulation



Overview of SmallVille

- Game engine: Phaser
 - World simulation
 - Display
- Character Logic Implementation
 - Handcrafted overall logic
 - Many components implemented with LLM
- Key Data Structures
 - Agent's memory stream
 - World tree and agent's view (subtree)

Architecture



Memory Stream: Comprehensive Record of Agent's Experiences

2023-02-13 22:48:20: desk is idle
2023-02-13 22:48:20: bed is idle
2023-02-13 22:48:10: closet is idle
2023-02-13 22:48:10: refrigerator is idle
2023-02-13 22:48:10: Isabella Rodriguez is stretching
2023-02-13 22:33:30: shelf is idle
2023-02-13 22:33:30: desk is neat and organized
2023-02-13 22:33:10: Isabella Rodriguez is writing in her journal
2023-02-13 22:18:10: desk is idle
2023-02-13 22:18:10: Isabella Rodriguez is taking a break
2023-02-13 21:49:00: bed is idle
2023-02-13 21:48:50: Isabella Rodriguez is cleaning up the kitchen
2023-02-13 21:48:50: refrigerator is idle
2023-02-13 21:48:50: bed is being used
2023-02-13 21:48:10: shelf is idle
2023-02-13 21:48:10: Isabella Rodriguez is watching a
2023-02-13 21:19:10: shelf is organized and tidy
2023-02-13 21:18:10: desk is idle
2023-02-13 21:18:10: Isabella Rodriguez is reading a b
2023-02-13 21:03:40: bed is idle
2023-02-13 21:03:30: refrigerator is idle
2023-02-13 21:03:30: desk is in use with a laptop and on it

- Textual description
- Creation timestamp
- Last access timestamp
- Importance score by LLM
- Vector embedding

On the scale of 1 to 10, where 1 is purely mundane (e.g., brushing teeth, making bed) and 10 is extremely poignant (e.g., a break up, college acceptance), rate the likely poignancy of the following piece of memory.

Memory: buying groceries at The Willows Market and Pharmacy
Rating: <fill in>

Memory Stream: Role Definition

- Hand crafted role definition
- Split by “;” into pieces
- Initial entries into memory stream

John Lin is a pharmacy shopkeeper at the Willow Market and Pharmacy who loves to help people. He is always looking for ways to make the process of getting medication easier for his customers; John Lin is living with his wife, Mei Lin, who is a college professor, and son, Eddy Lin, who is a student studying music theory; John Lin loves his family very much; John Lin has known the old couple next-door, Sam Moore and Jennifer Moore, for a few years; John Lin thinks Sam Moore is a kind and nice man; John Lin knows his neighbor, Yuriko Yamamoto, well; John Lin knows of his neighbors, Tamara Taylor and Carmen Ortiz, but has not met them before; John Lin and Tom Moreno are colleagues at The Willows Market and Pharmacy; John Lin and Tom Moreno are friends and like to discuss local politics together; John Lin knows the Moreno family somewhat well – the husband Tom Moreno and the wife Jane Moreno.

Memory Stream: RAG

Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

retrieval recency importance relevance

2.34 = 0.91 + 0.63 + 0.80

ordering decorations for the party

2.21 = 0.87 + 0.63 + 0.71

researching ideas for the party

2.20 = 0.85 + 0.73 + 0.62

...



I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!



Isabella

Memory Stream: Multi-Level Planning

High-Level Plan

“1) wake up and complete the morning routine at 8:00 am, 2) go to Oak Hill College to take classes starting 10:00 am, [. . .] 5) work on his new music composition from 1:00 pm to 5:00 pm, 6) have dinner at 5:30 pm, 7) finish school assignments and go to bed by 11:00 pm.”

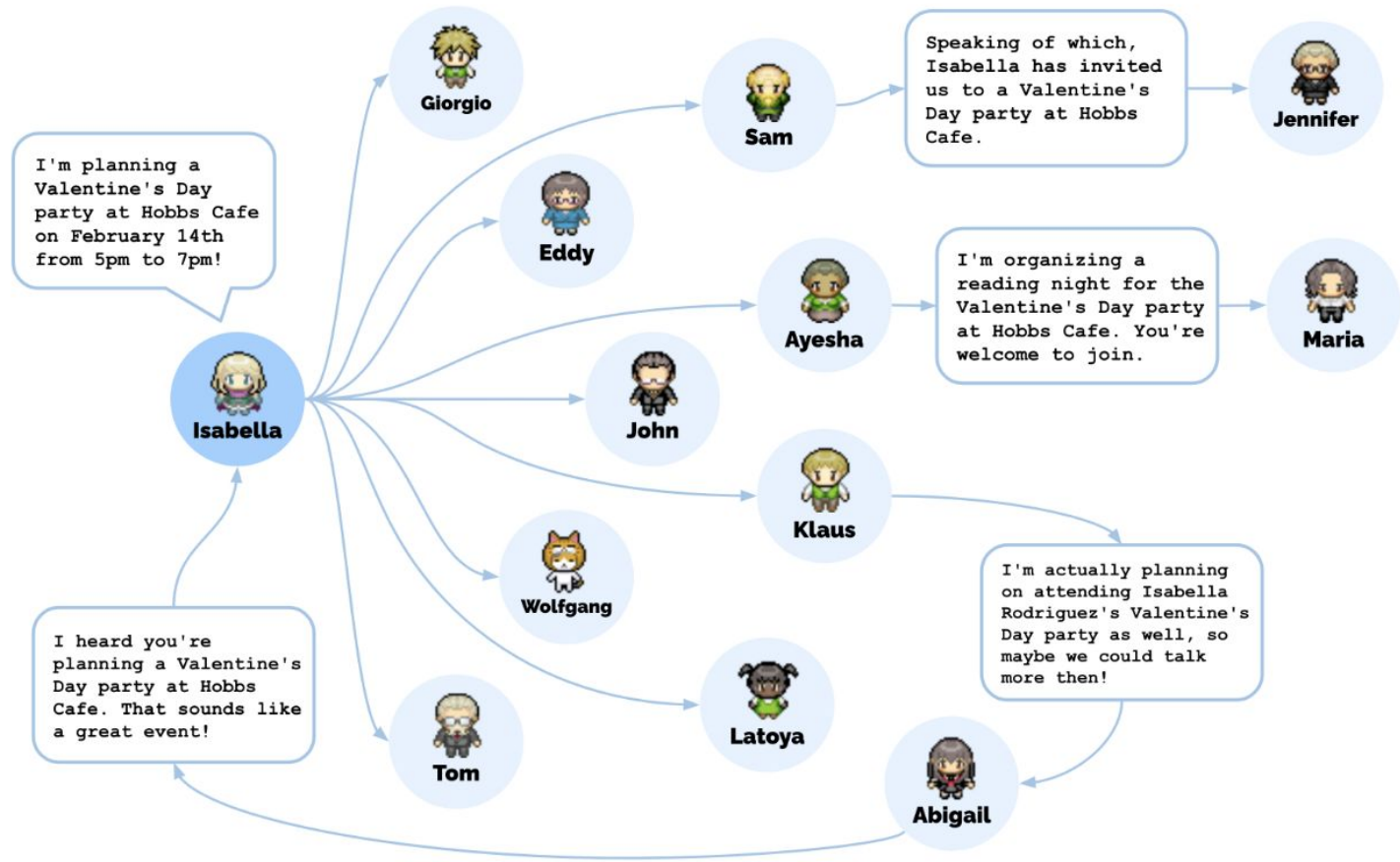
Median-Level Plan

1:00 pm: start
by brainstorming some ideas for his music composition [...] 4:00
pm: take a quick break and recharge his creative energy before
reviewing and polishing his composition.

Detail-Level Plan

4:00 pm: grab a
light snack, such as a piece of fruit, a granola bar, or some nuts.
4:05 pm: take a short walk around his workspace [...] 4:50 pm:
take a few minutes to clean up his workspace.

Observations: Diffusion of Information



RPA Industry: 1st Generation Agents Reinvented

- RPA: Robotic Process Automation
- Automation of repetitive office workflows
- Integration into existing software ecosystems
- Growth since 2018
- Rapidly adopting LLM

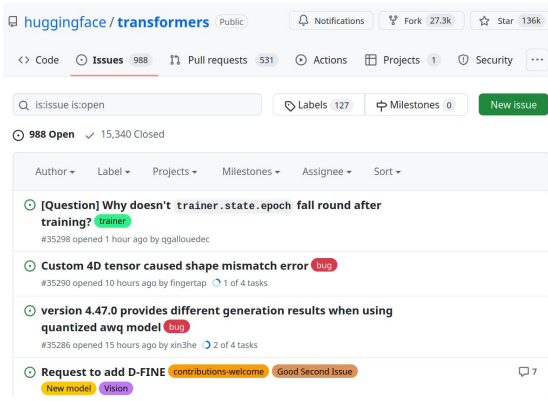
[Unipath Demo: Onboarding](#)

	Market Cap	Revenue Data	Annualized Revenues (<i>billions</i>)	Ratio
UiPath	41	2021 Actual	0.607	68
Pega	9.6	Q1-2021	0.939	10
Blue Prism	1.3	Q4-2020	0.207	6
NICE	14	Q1-2021	1.82	8
Datamatics	0.118	Q1-2021	0.156	1

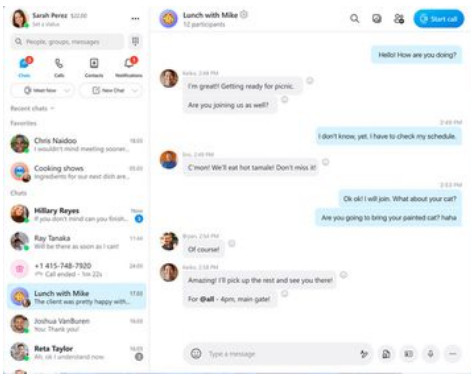
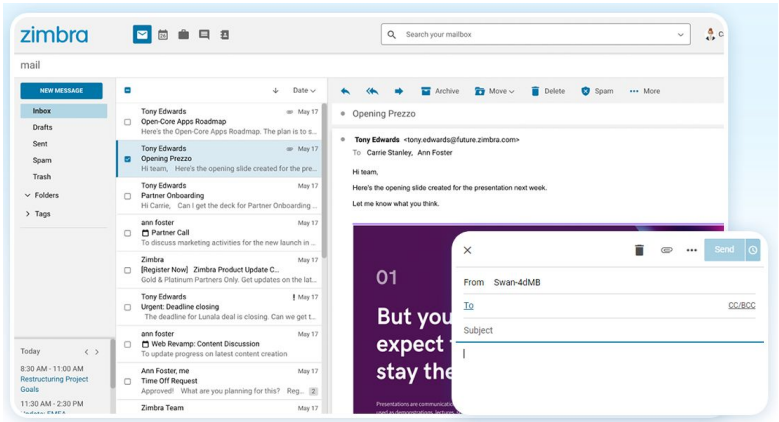
Comments

- Hand-crafted control logic in most implementations
- LLM are evolving rapidly and will foreseeably surpass human
- Future agents should be centered around LLM
 - Memory access as tool usage
 - Interaction as tool usage
 - ...

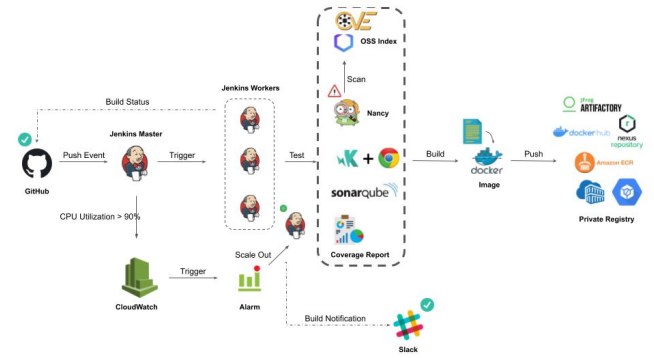
Discussion: How to Design an Engineering Team of AI Agent



Github: Source of Tasks



Message Infrastructure



Automatic Testing

Paradigm of Code Generation: AI Generates the Patch

```
specbees-MacBook-Pro-5:phyins specbee$ git diff
diff --git a/docroot/themes/custom/pi_bootstrap/pi_bootstrap.info.yml b/docroot/themes/custom/pi_bootstrap/pi_bootstrap.info.yml
index 53578e432..abb4bfcd8 100644
--- a/docroot/themes/custom/pi_bootstrap/pi_bootstrap.info.yml
+++ b/docroot/themes/custom/pi_bootstrap/pi_bootstrap.info.yml
@@ -1,4 +1,4 @@
-name: 'Physicians Insurance Bootstrap'
+name: 'Physicians Insurance Bootstrap Updated'
 description: 'Custom base theme built on Bootstrap v4.0.0'
 package: 'Physicians Insurance Bootstrap Theme'

@@ -40,7 +40,7 @@ regions:
 # - Bootstrap Modals can also be manually created and added to this region as needed
 pagebottom: 'Page bottom (content hidden, use to display popup content)'

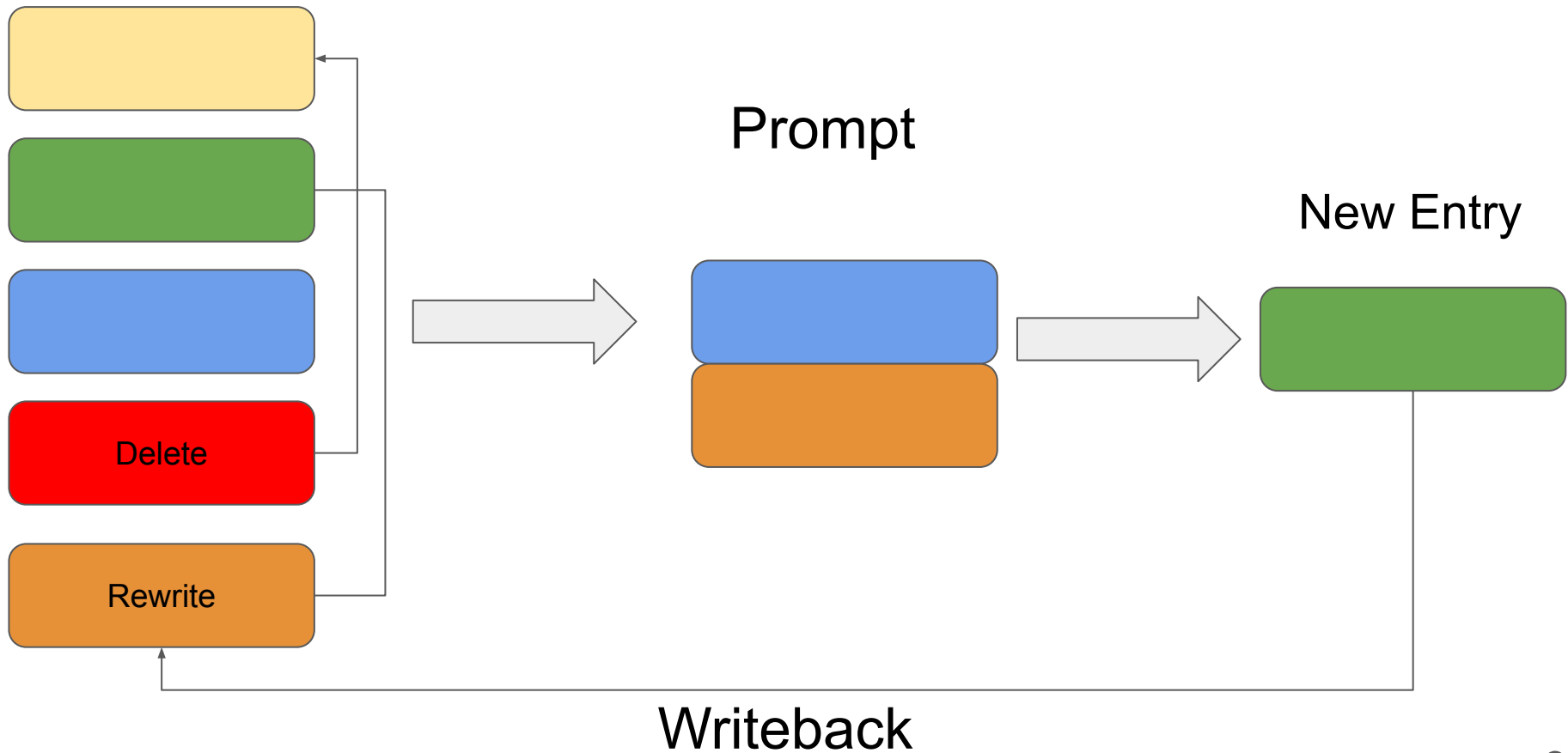
- # As of D8.2.* Blocks that are referenced into the page need to be placed into a Region
+ # As of D8.2.* Blocks that are referenced into the page need to be placed into a Region
 # even if said region is not displayed. Otherwise a referenced block will not render.
 reference_blocks: 'Block References (add blocks here to be referenced in sections)'
```

```
diff --git a/docroot/themes/custom/pi_bootstrap/pi_bootstrap.theme b/docroot/themes/custom/pi_bootstrap/pi_bootstrap.theme
index 12707cb81..f00a2ef4c 100644
--- a/docroot/themes/custom/pi_bootstrap/pi_bootstrap.theme
+++ b/docroot/themes/custom/pi_bootstrap/pi_bootstrap.theme
@@ -10,7 +10,7 @@
 use Drupal\taxonomy\Entity\Term;

/*
- BETTER DEBUG:
+ BETTER DEBUG: Updated
- download Devel Module (https://www.drupal.org/project/devel)
- enable Devel & Kint ( DRUSH -> drush en devel kint -y )
```

specbees-MacBook-Pro-5:phyins specbee\$ █

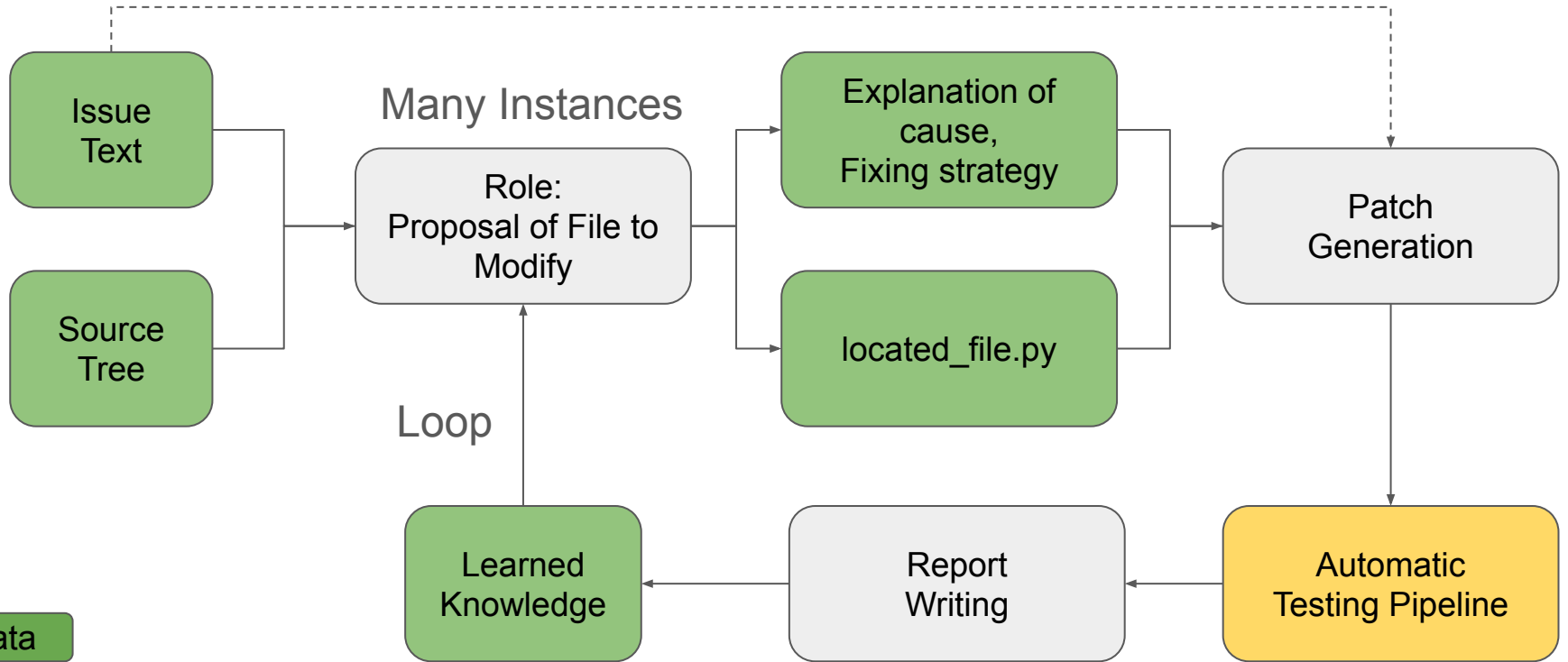
Per Agent Journal is the Central Data Structure



Workflow and Role Design

- Human software engineering as prototype
- AI vs Human
 - AI is way much cheaper; we can have 1000 agents per project
 - Shorter life span. Many roles are dispose-after-usage.
 - Less long-term planning capability
 - => Role division in on a finer granularity
- What roles do we need?

Workflow



Data

Agent

Key Ideas on Designing An AI Team

1. Integrate into existing IT infrastructure Software
 - a. Assign an email address to each agent
 - b. Gitlab, zimbra & Jenkins
 - c. Bridge zimbra with real-world email system for selective import emails.
2. Per-Agent Journal is the key data structure
 - a. [“Journal” is an important technique in database & operating systems](#)
 - b. “Play” the journal to generate the prompt
 - c. Pull data entries from email / IM / etc into the journal
 - d. Reflection upon prompt reaching limit; delete & rewrite entries.
3. Agent design
 - a. Small but many