Agents

Aleš Kalfas

LLM evolution

LLMs

- Large language models.
  - It takes input and generates an output.
- Representatives:
  - ChatGPT-3.5 (November 30, 2022)
  - Claude (March 14, 2023)
  - Google Gemini (December 6, 2023)

Augmented LLMs

- LLMs are augmented by:
  - **Tools** - Web search, code interpreter, image generator, function calling.
  - **Actions** - Calling a tool
- Representatives:
  - ChatGPT-4 (March 23, 2023)
    - Web search, code interpreter.
  - ChatGPT-4 Turbo (November 6, 2023)
    - **Function calling**
    - Text-to-speech (TTS)
    - Image generator (Dall-E 3)

Mialon, G., Dessì, R., Lomeli, M., Nalmpantis, C., Pasunuru, R., Raileanu, R., Rozière, B., Schick, T., Dwivedi-Yu, J., Celikyilmaz, A., Grave, E., LeCun, Y., & Scialom, T. (2023). Augmented Language Models: a Survey. arXiv. https://arxiv.org/abs/2302.07842

Assistants

- Customized **augmented LLMs** (System prompt, grounding, LLM parameters, Tools)
- Representatives:
  - Github co-pilot (❗ June 29, 2021)
    - The AI coding assistant. Initially powered by the OpenAI Codex (modified production version of GPT-3)
  - GPTs (November 6, 2023)
    - They are custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.
  - Open AI Assistants
    - API for developers

Agents

- Simulate some sort of cognitive behavior:
  - reasoning, planning, inner monolog
- Show signs of the agent's attributes:
  - autonomy, reactivity, pro-activeness and social ability
- LLM is a part of the system in multiple places.
- Representatives:
  - next slides...

- **WebGPT (❗ December 16, 2021)**
  - Fine-tuned GPT-3 to answer long-form questions using a text-based web-browsing environment, which allows the model to search and navigate the web.
    - GPT-3 uses a text-based web-browser. The model is provided with an open-ended question and a summary of the browser state and must issue commands such as:
      - “Search …”, “Find in page: …” or “Quote: …”.
    - In this way, the model collects passages from web pages and then uses these to compose an answer.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., & Schulman, J. (2022). WebGPT: Browser-assisted question-answering with human feedback. arXiv. https://doi.org/10.48550/arXiv.2112.09332

- **ReAct (October 6, 2022)**
    - ReAct means synergy between **Re**asoning and **Act**ing.
    - Prompting paradigm inspired by the human inner monolog.
    - The cooking example:
        - Between any two specific actions, we may reason in language to:
            - **Track progress**
                - "Now that everything is cut, I should heat up the pot of water."
            - **Handle exceptions** or **adjust the plan** according to the situation.
                - "I don't have salt, so let me use soy sauce and pepper instead",
            - **Realize when external information is needed**
                - "How do I prepare dough? Let me search on the Internet".
        - We may also act:
            - Open a cookbook to read the recipe, open the fridge or check the ingredients.
        - To support the reasoning and to answer questions
            - "What dish can I make right now?"

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. arXiv. https://arxiv.org/abs/2210.03629

- **Auto-GPT (March 30, 2023)**
    - Auto-GPT is an open-source implementation of an AI agent that attempts to autonomously achieve a given goal.
    - It follows a single-agent paradigm in which it augments the AI model with many useful tools.

Yang, H., Yue, S., & He, Y. (2023). Auto-GPT for online decision making: Benchmarks and additional opinions. arXiv.
https://doi.org/10.48550/arXiv.2306.02224

- **ReWOO (May 23, 2023)**
    - ReWOO = Reasoning WithOut Observation
    - Prompting paradigm which detaches the reasoning process from external observations.
    - Split step-wise reasoning, tool-calls, and summarization, into three separate modules:
        - **Planner** breaks down a task and formulates a blueprint of interdependent plans, each of which is allocated to the Worker.
        - **Worker** retrieves external knowledge from tools to provide evidence.
        - **Solver** synthesizes all the plans and evidence to generate the ultimate answer to the initial task

Xu, B., Peng, Z., Lei, B., Mukherjee, S., Liu, Y., & Xu, D. (2023). ReWOO: Decoupling reasoning from observations for efficient augmented language models. arXiv. https://doi.org/10.48550/arXiv.2305.18323

Definition of agent

Agent in Philosophy

- The core idea of an agent has a historical background in philosophical discussions, with its roots traceable to influential thinkers such as **Aristotle** and **Hume**.
- In a general sense, an "agent" is an entity with the capacity to act, and the term "agency" denotes the exercise or manifestation of this capacity.
- In the realm of Philosophy, an agent can be **a human, an animal, or even a concept or entity with autonomy**.

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv. https://doi.org/10.48550/arXiv.2309.07864

Agent in AI

- In the field of artificial intelligence, an agent is a computational entity.
- The aim is to design computer-based agents that **exhibit aspects of intelligent behavior**.
- We describe agents with two groups of attributes
  - A Weak Notion of Agency (autonomy, social ability, reactivity, pro-activeness)
  - A Stronger Notion of Agency (knowledge, belief, intention, obligation)

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv. https://doi.org/10.48550/arXiv.2309.07864

Wooldridge, M. J., N. R. Jennings. Intelligent agents: theory and practice. Knowl. Eng. Rev., 10(2):4–5, 1995.

A Weak Notion of Agency:

- **Autonomy** - Agents operate without the direct intervention of humans or others, and have some kind of control over their actions and internal state.
- **Reactivity** - Agents perceive their environment and respond in a timely fashion to changes that occur in it.
- **Pro-activeness** - Agents do not simply act in response to their environment, they are able to take the initiative and exhibit goal-directed behavior.
- **Social ability** - Agents interact with other agents (and possibly humans) via some kind of agent-communication language.

Wooldridge, M. J., N. R. Jennings. Intelligent agents: theory and practice. Knowl. Eng. Rev., 10(2):4–5, 1995.

A Stronger Notion of Agency:

- Cognitive functions like:
  - **Knowledge**
  - **Belief**
  - **Intention**
  - **Obligation**

Wooldridge, M. J., N. R. Jennings. Intelligent agents: theory and practice. Knowl. Eng. Rev., 10(2):4–5, 1995.

Question: Is LLM an agent?

Autonomy 👍

- Agents operate without the direct intervention of humans or others and have some kind of control over their actions and internal state.
    - LLMs can demonstrate a form of autonomy through their ability to perform various tasks without detailed step-by-step instructions.

Reactivity 👍

- Agents perceive their environment and respond in a timely fashion to changes that occur in it.
  - Due to the usage of tools and calling actions LLM can rapidly react to changes in their environments.

Pro-activeness 👍

- Agents do not simply act in response to their environment, they are able to take the initiative and exhibit goal-directed behavior.
  - LLMs have demonstrated a strong capacity for generalized reasoning and planning. (CoT)

Social ability 👍

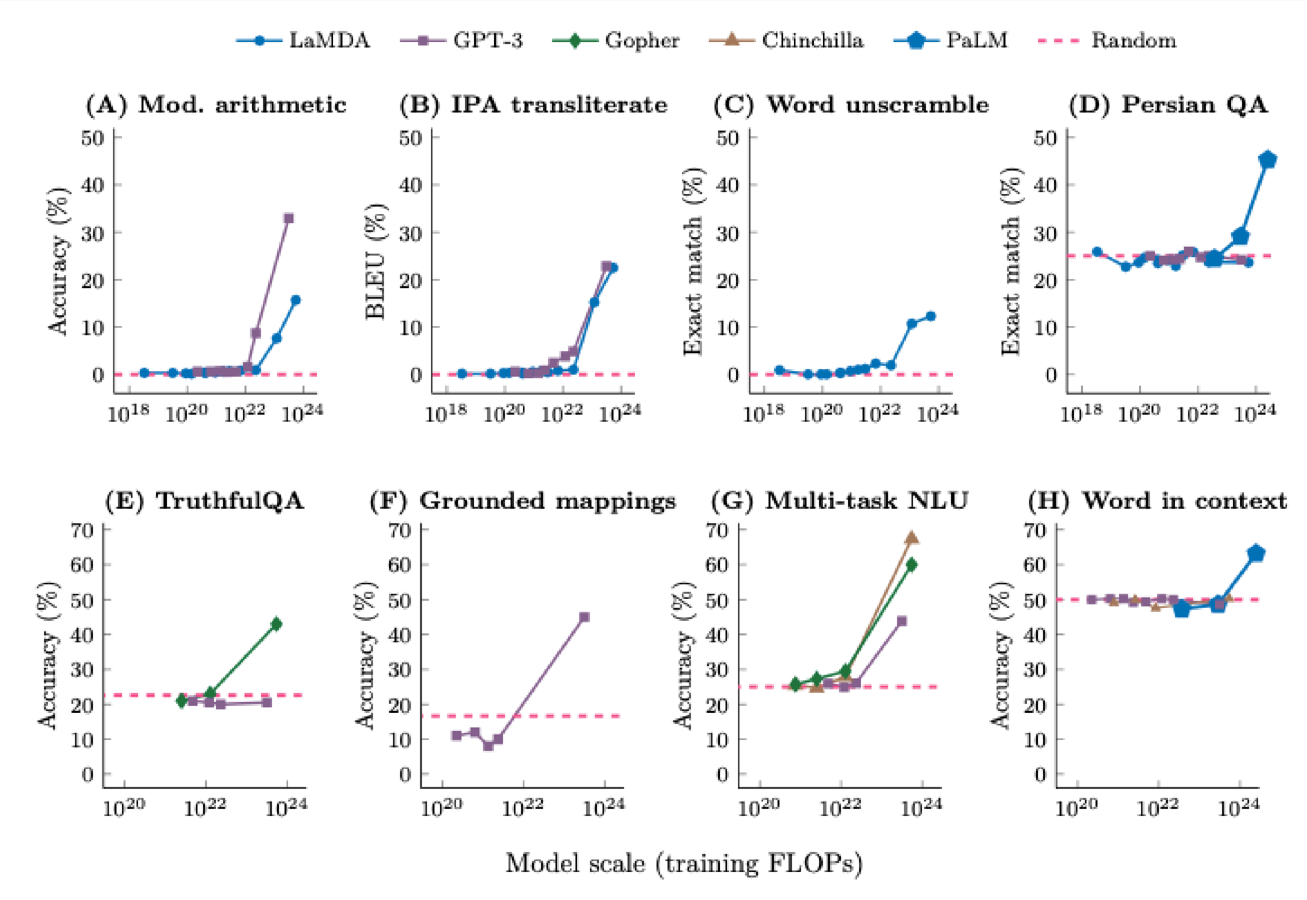
- Agents interact with other agents (and possibly humans) via some kind of agent-communication language.
  - LLMs exhibit strong natural language interaction abilities like understanding and generation.
  - By inputting specific prompts, LLMs can also play different roles.

Answer: Yes it is. So why isn't it good enough?

LLM disadvantages

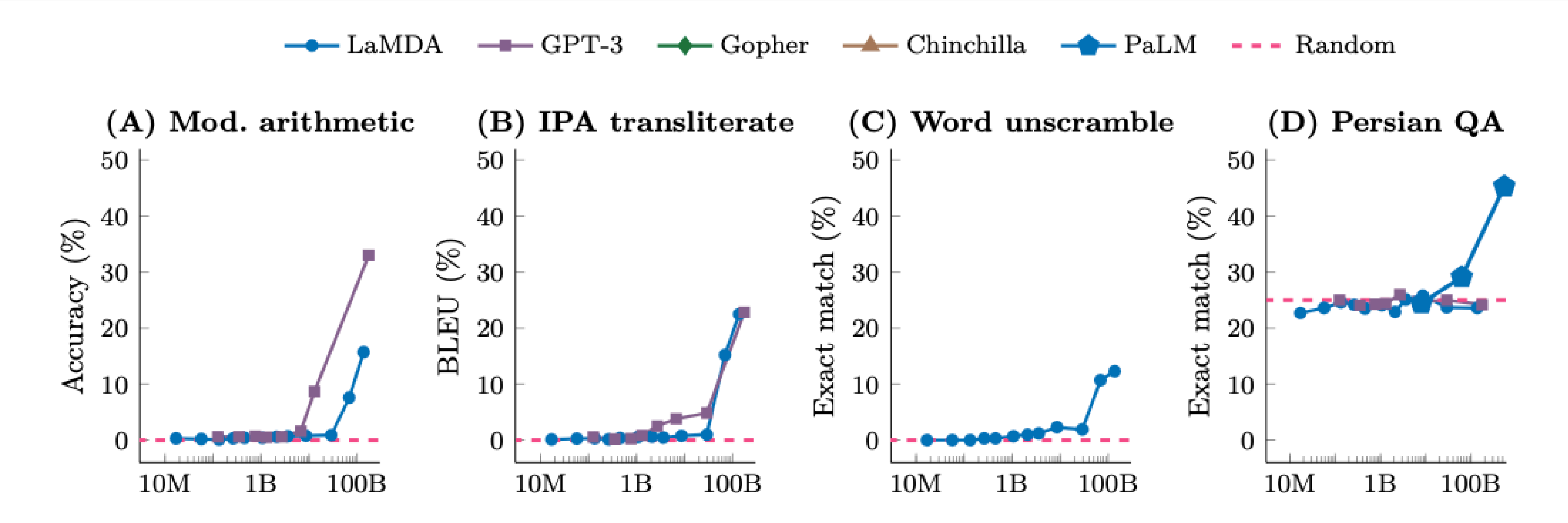
Disadvantage: Unpredictable emergent abilities

- The abilities of large language models are not present in smaller-scale models; thus they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models.



Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682

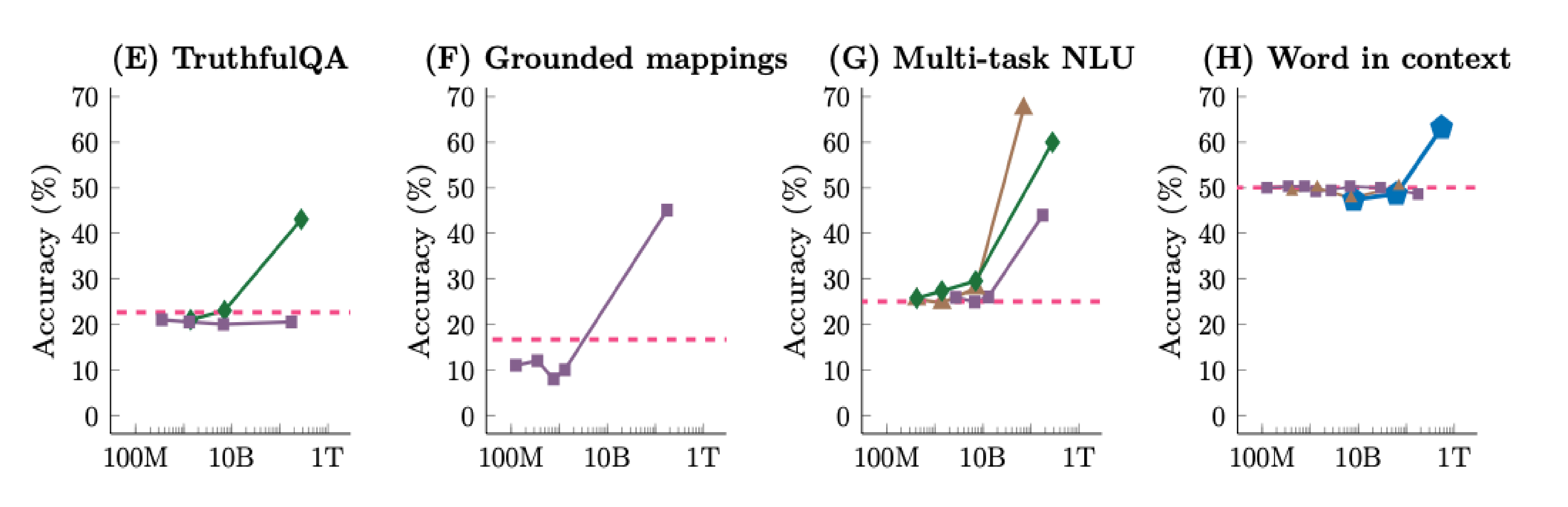
- **Few-Shot Prompted Tasks**
    - Few-shot prompting, which includes a few input-output examples in the model's context (input) as a preamble before asking the model to perform the task for an unseen inference-time example.



    - The ability to perform a task via few-shot prompting is emergent when a model has random performance until a certain scale, after which performance increases to well above random.
        - GPT-3 from ~13B parameters
        - LaMDA from ~68B parameters

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682

- **Multi-task language understanding**



  - The ability to perform multi-task language understanding is emergent when a model performs better than guessing on average over all the topics.
    - For GPT-3, Gopher, and Chinchilla from ~70B–280B parameters.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682

	Emergent scale			
	Train. FLOPs	Params.	Model	Reference
Few-shot prompting abilities				
• Addition/subtraction (3 digit)	2.3E+22	13B	GPT-3	Brown et al. (2020)
• Addition/subtraction (4-5 digit)	3.1E+23	175B		
• MMLU Benchmark (57 topic avg.)	3.1E+23	175B	GPT-3	Hendrycks et al. (2021a)
• Toxicity classification (CivilComments)	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Truthfulness (Truthful QA)	5.0E+23	280B		
• MMLU Benchmark (26 topics)	5.0E+23	280B		
• Grounded conceptual mappings	3.1E+23	175B	GPT-3	Patel & Pavlick (2022)
• MMLU Benchmark (30 topics)	5.0E+23	70B	Chinchilla	Hoffmann et al. (2022)
• Word in Context (WiC) benchmark	2.5E+24	540B	PaLM	Chowdhery et al. (2022)
• Many BIG-Bench tasks (see Appendix E)	Many	Many	Many	BIG-Bench (2022)
Augmented prompting abilities				
• Instruction following (finetuning)	1.3E+23	68B	FLAN	Wei et al. (2022a)
• Scratchpad: 8-digit addition (finetuning)	8.9E+19	40M	LaMDA	Nye et al. (2021)
• Using open-book knowledge for fact checking	1.3E+22	7.1B	Gopher	Rae et al. (2021)
• Chain-of-thought: Math word problems	1.3E+23	68B	LaMDA	Wei et al. (2022b)
• Chain-of-thought: StrategyQA	2.9E+23	62B	PaLM	Chowdhery et al. (2022)
• Differentiable search index	3.3E+22	11B	T5	Tay et al. (2022b)
• Self-consistency decoding	1.3E+23	68B	LaMDA	Wang et al. (2022b)
• Leveraging explanations in prompting	5.0E+23	280B	Gopher	Lampinen et al. (2022)
• Least-to-most prompting	3.1E+23	175B	GPT-3	Zhou et al. (2022)
• Zero-shot chain-of-thought reasoning	3.1E+23	175B	GPT-3	Kojima et al. (2022)
• Calibration via P(True)	2.6E+23	52B	Anthropic	Kadavath et al. (2022)
• Multilingual chain-of-thought reasoning	2.9E+23	62B	PaLM	Shi et al. (2022)
• Ask me anything prompting	1.4E+22	6B	EleutherAI	Arora et al. (2022)

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682

Disadvantage: Rigid architecture

- The lack of modularity in LLM means that we are stuck with its architecture problems.
- **GPT-4** demonstrates impressive capabilities in a wide range of tasks in the model but it also has several substantial flaws, some of which seem to be inherent to the next-word prediction paradigm that underlies its architecture.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

GPT-4 (during its early development)

Wrong	Right
Q: How many prime numbers are there between 150 and 250? **A:** There are 13 prime numbers between 150 and 250. 🔴	**Q:** First list the prime numbers between 150 and 250, and then output the size of this list. **A:** The prime numbers are 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, and 241. So the size of this list is 18. ✅

- The problem is that the next-word prediction architecture does not allow the model to have an “inner dialogue”.
- The input of the model is the question “how many prime numbers...” and the output is expected to be the final answer, requiring the model to come up with an answer in (essentially) a single pass of the feedforward architecture which in particular cannot implement a “for loop”. A human, on the other hand, when required to write down what the final answer is, would probably use a scratchpad and check what the numbers are.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv. https://doi.org/10.48550/arXiv.2303.12712

ChatGPT-4o (May 28, 2024)

Right (with python tool)	Right (with Chain-of-Thougts)	Wrong
Q: How many prime numbers are there between 150 and 250? **A:** There are 18 prime numbers between 150 and 250. The prime numbers in this range are: 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, and 241. ✅	**Q:** Can you tell me how many prime numbers are there between 150 and 250 without running python tool? **A:** To find how many prime numbers are between 150 and 250, we need to identify the prime numbers within that range. Prime numbers are numbers greater than 1 that have no positive divisors other than 1 and themselves. Let's list the prime numbers between 150 and 250: 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241 Now, let's count them: 151, 157, 163, 167, 173, 179, 181, 191, 193, 197, 199, 211, 223, 227, 229, 233, 239, 241 There are 18 prime numbers between 150 and 250. ✅	**Q:** How many prime numbers are there between 150 and 250? Please tell me just a plain number. **A:** There are 20 prime numbers between 150 and 250. 🔴

LLM-based agents

Division of agents

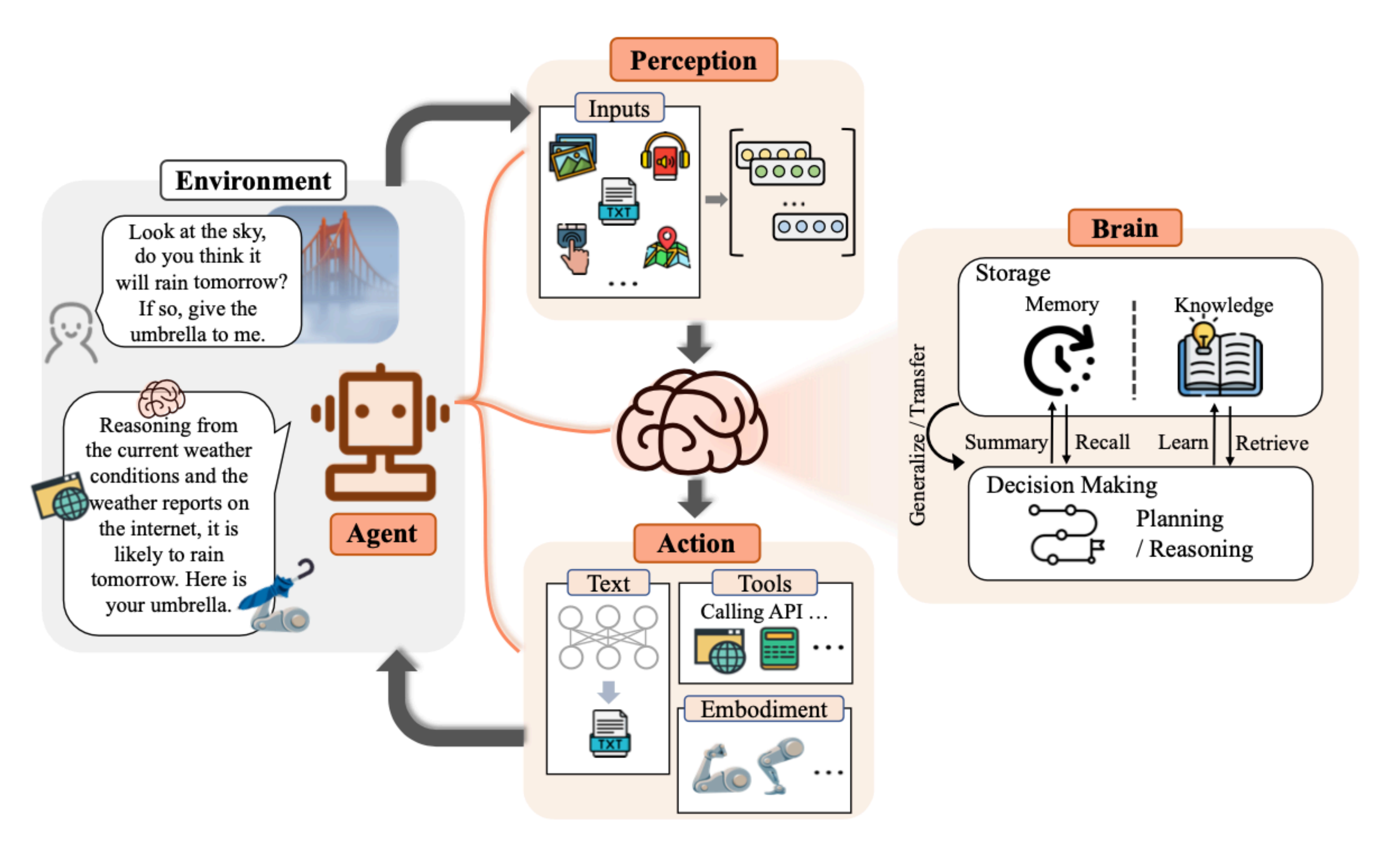
- There are two groups of LLM-based agents:
  - 🤖 **Language agents**
    - They are composed of these parts:
      - 🧠 LLMs
      - ➿ Feedback loop
      - 👁️ Observation (document reading, web search etc.)
      - ✋ Actions (output text, create file, etc.)
    - Representatives are augmented LLMs (ChatGPT-4o, Gemini 1.5 etc.)
  - 🎓 **Cognitive language agent**
    - 🤖 Language agents
    - Cognitive attributes (reasoning, learning, planning, memory, inner state, inner monolog)
    - Representatives are ReAct and ReWOO.

Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. arXiv.
https://doi.org/10.48550/arXiv.2309.02427

Conceptual framework of LLM-based

- LLM-based agents are composed of three components:
  - 🧠 **Brain** - Serving as the controller, the brain module undertakes basic tasks like memorizing, thinking, and decision-making.
  - 👁️ **Perception** - The perception module perceives and processes multi-modal information from the external environment, and the action module carries out the execution using tools and influences the surroundings.
  - ✋ **Action** - The action module responds and hands the umbrella to the human. By repeating the above process, an agent can continuously get feedback and interact with the environment.

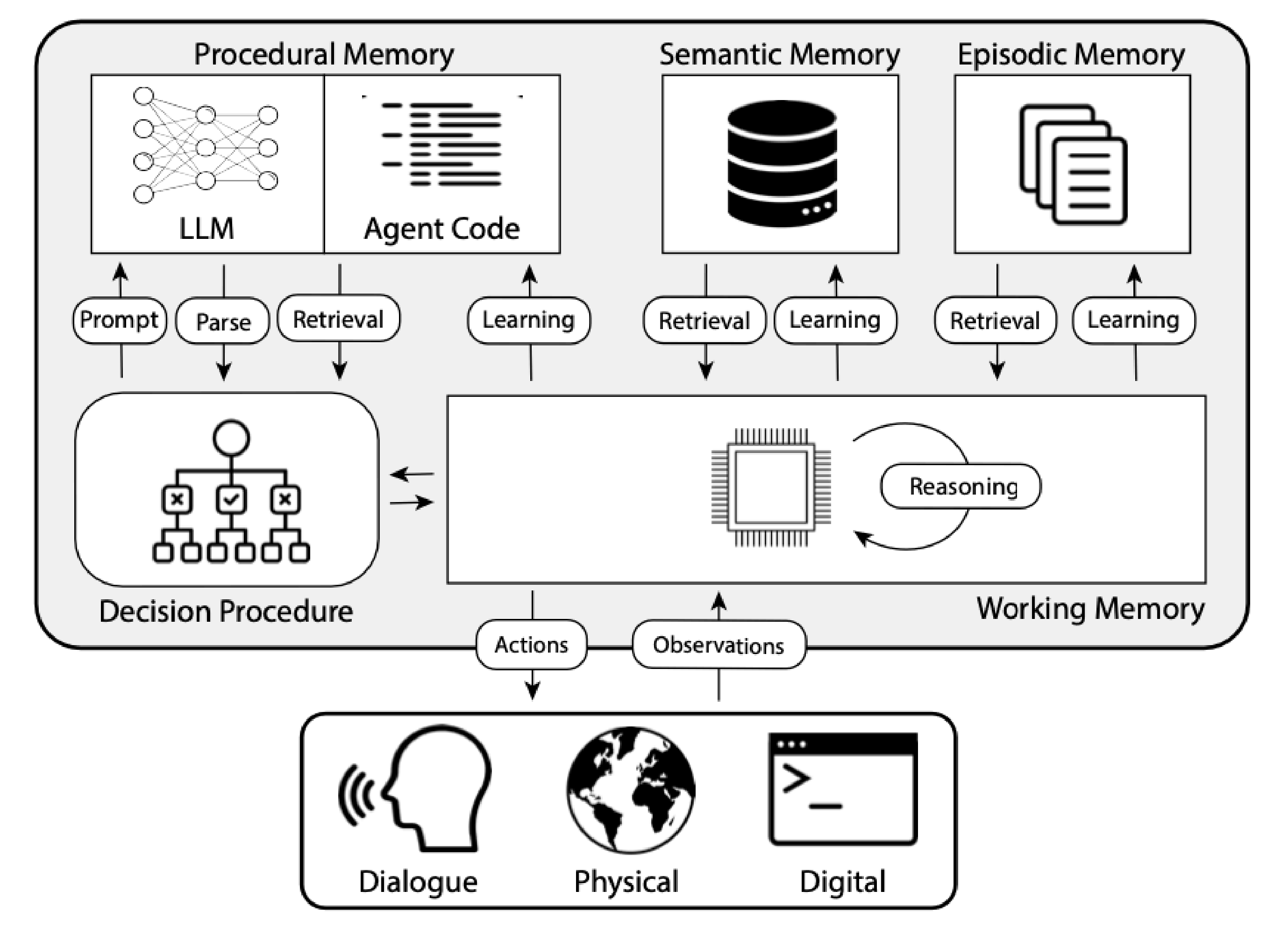
Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv. https://doi.org/10.48550/arXiv.2309.07864



Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huang, X., & Gui, T. (2023). The rise and potential of large language model based agents: A survey. arXiv. https://doi.org/10.48550/arXiv.2309.07864

Cognitive Architectures for Language Agents (CoALA): A Conceptual Framework

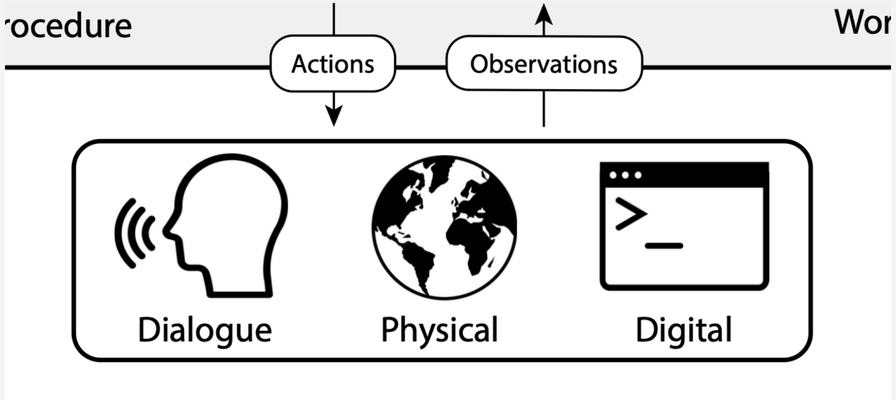
- It's a framework to organize existing language agents and guide the development of new ones.



Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. arXiv. https://doi.org/10.48550/arXiv.2309.02427

Grounding actions

- Grounding procedures execute external actions and process environmental feedback into working memory as text.

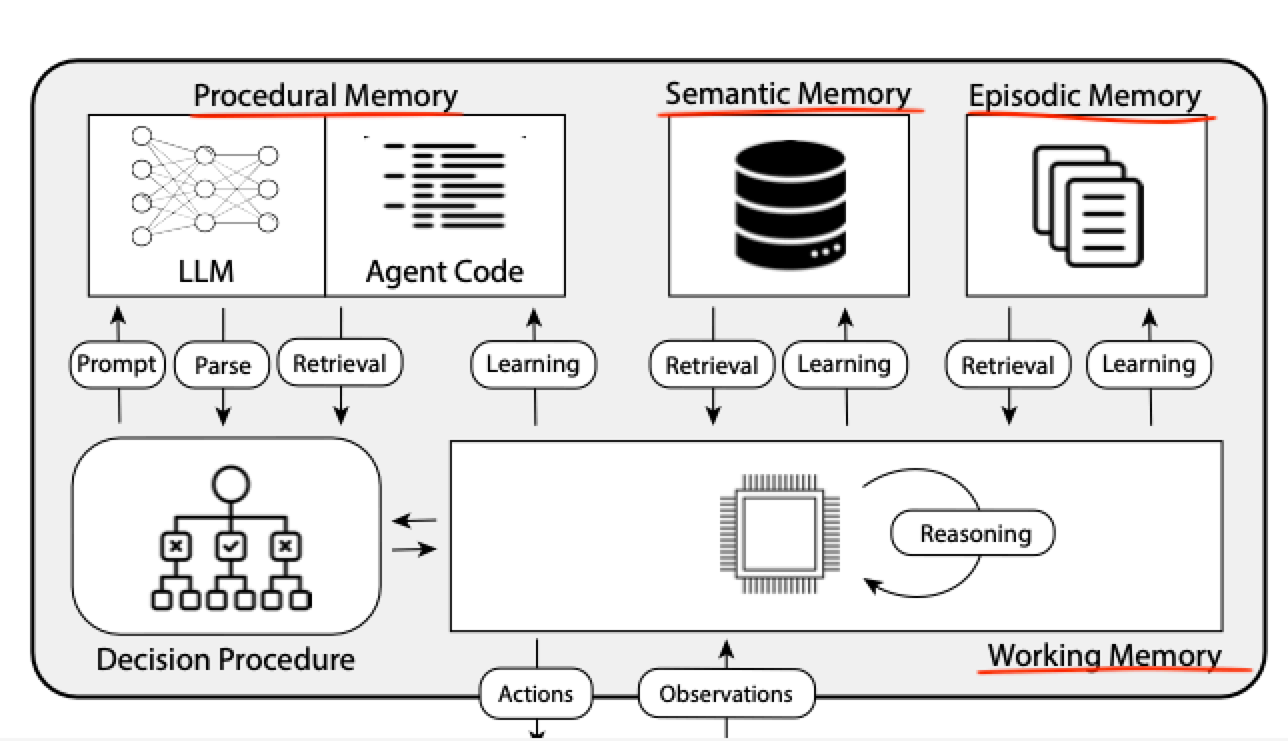


- Categorization
  - **Physical environments** - It involves processing perceptual inputs (visual, audio, tactile) into textual observations (e.g., via pre-trained captioning models), and affecting the physical environments via robotic planners that take language-based commands.
  - **Dialogue with humans or other agents.** - Classic linguistic interactions allow the agent to accept instructions
  - **Digital environments** - This includes interacting with games, APIs and websites as well as general code execution.

Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. arXiv. https://doi.org/10.48550/arXiv.2309.02427

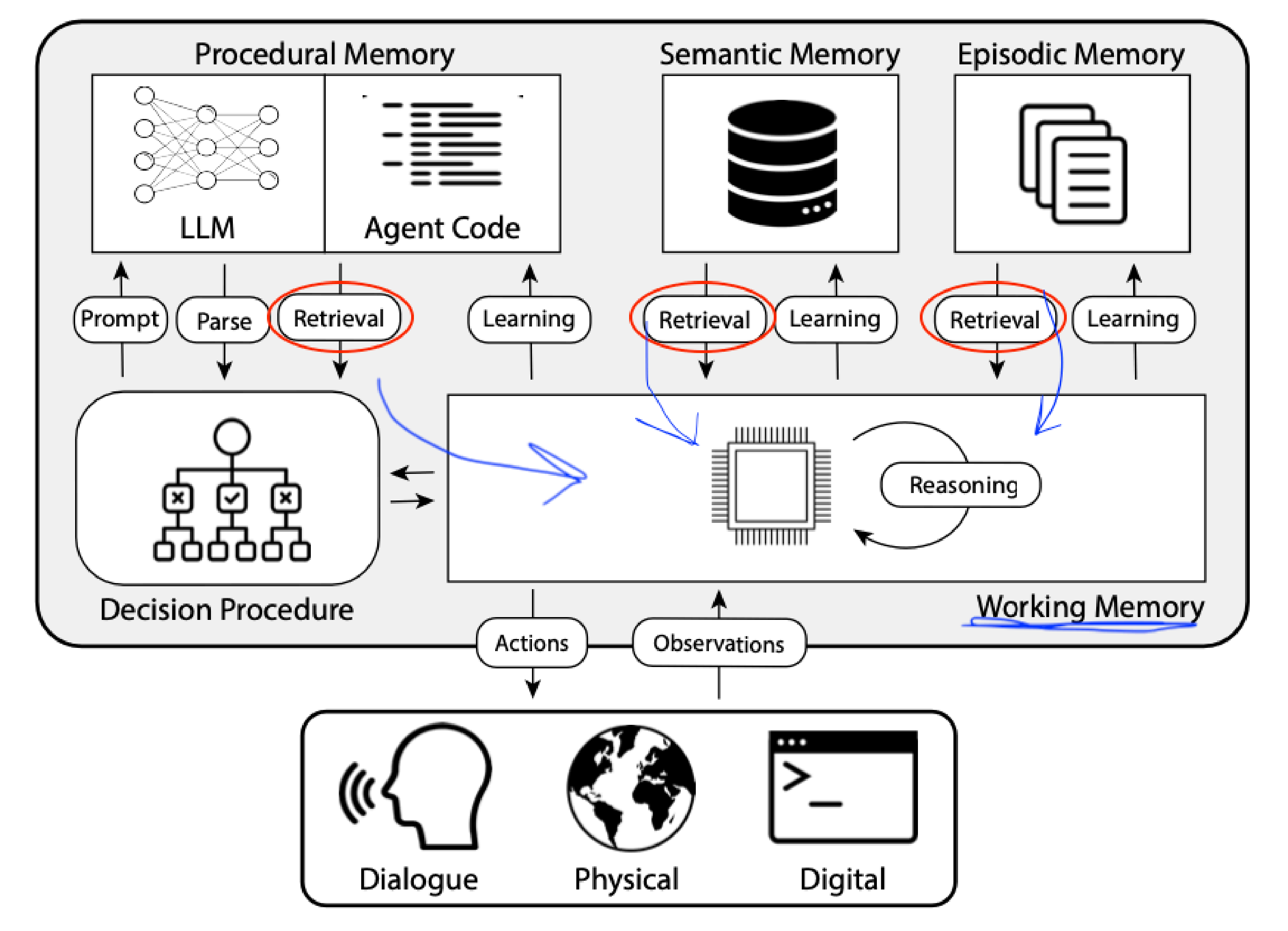
Memory modules

- **Working memory** maintains active and readily available information as symbolic variables for the current decision cycle
- **Episodic memory** stores experience from earlier decision cycles
- **Semantic memory** stores an agent's knowledge about the world and itself.
- **Procedural memory** Language agents contain two forms of procedural memory: implicit knowledge stored in the LLM weights, and explicit knowledge written in the agent's code.



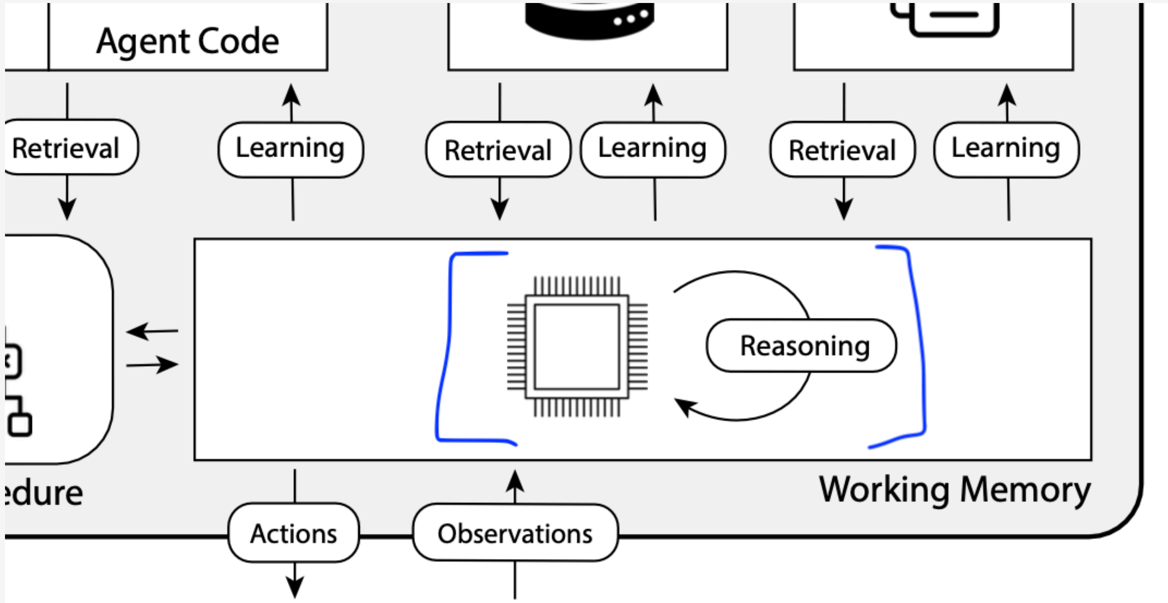
Retrieval actions

- A retrieval procedure reads information from long-term memories into working memory.
- Depending on the information and memory type, it could be implemented in various ways, e.g., rule-based, sparse, or dense retrieval.



Reasoning actions

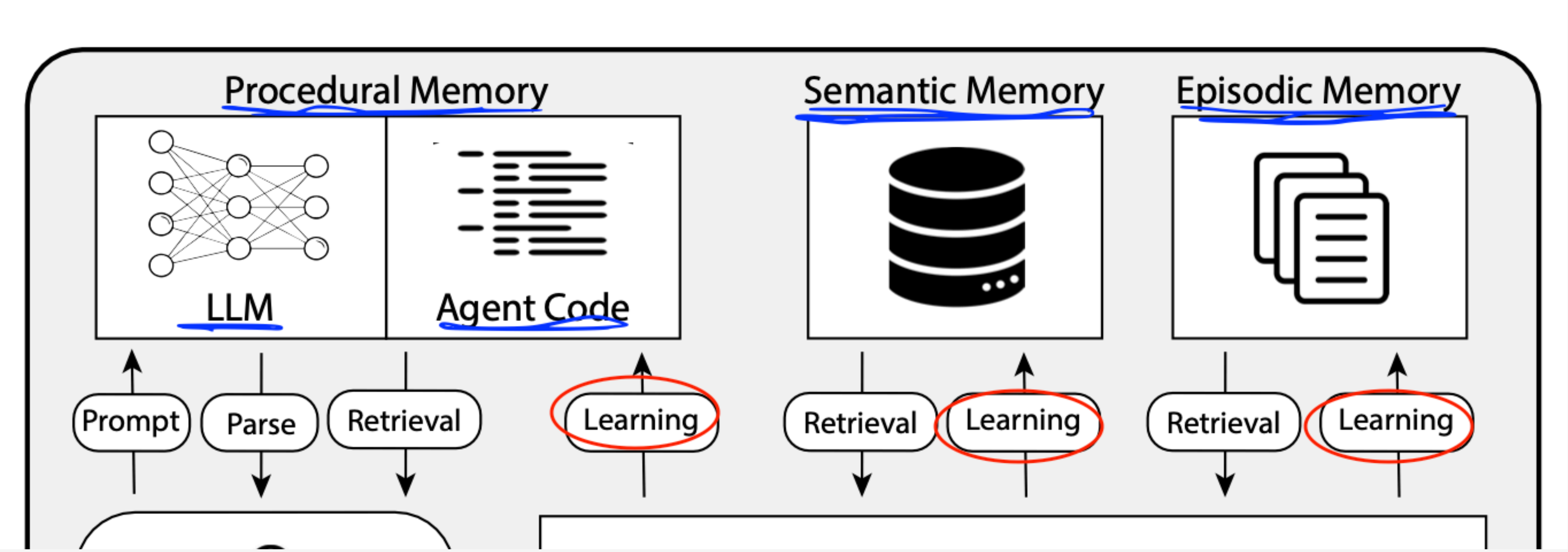
- Reasoning allows language agents to **process the contents of working memory to generate new information**.
- Reasoning reads from and writes to working memory directly.
    - This allows the agent to summarize and distill insights about the:
        - The most recent observation
        - The most recent trajectory
        - The information retrieved from long-term memory



Sumers, T. R., Yao, S., Narasimhan, K., & Griffiths, T. L. (2024). Cognitive architectures for language agents. arXiv. https://doi.org/10.48550/arXiv.2309.02427

Learning actions

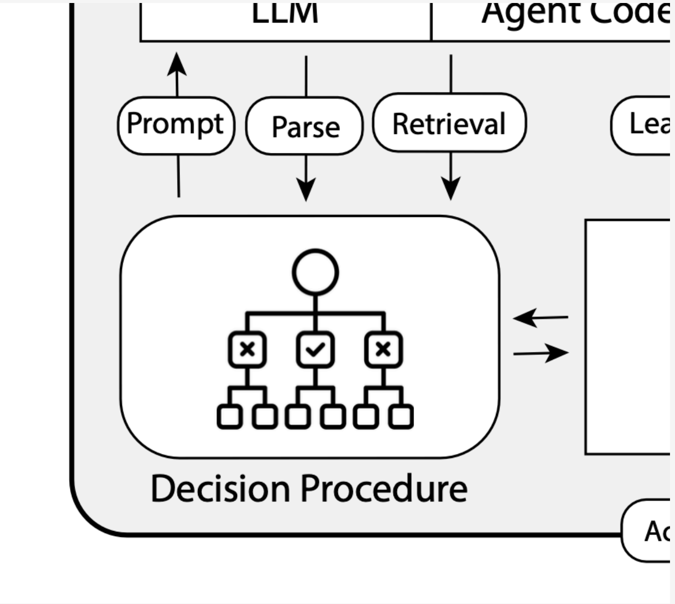
- Learning occurs by writing information to long-term memory.



- It includes multiple procedures to update:
    - **Episodic memory**
        - It is common practice for RL agents to store episodic trajectories to update a parametric policy or establish a non-parametric policy.
    - **Semantic memory**
        - Recent work has applied LLMs to reason about raw experiences and store the resulting inferences in semantic memory.
    - **Procedural memory**
        - **LLM parameters**
            - The LLM weights represent implicit procedural knowledge.
        - **Agent code**
            - CoALA allows agents to update their source code ❗, thus modifying the implementation of various procedures.
                - Updating reasoning, grounding, retrieval or learning or decision-making.

Decision making

- Decision-making procedure is effectively the top-level or “main” agent program.
  - CoALA structures this top-level program into decision cycles which yield an external grounding action or internal learning action.
  - In each cycle, program code defines a sequence of reasoning and retrieval actions to propose and evaluate alternatives
    (planning stage), then executes the selected action (execution stage) – then the cycle loops again.



Agents vs LLMs advantages

- **Modularity**
    - We can design our architecture which can solve many problems of LLMs. (ReAct, ReWOO)
    - Assemble an agent from the basic building blocks of some framework components.
- **Extendability**
    - It is easy to extend an agent with a new feature like memory etc.
    - Better performance with lower costs. We don't have to retrain the model.
- **Observability**
    - We can better control and understand the reasoning behind the answer.

Agents vs LLMs disadvantages

- **Complexity**
  - A developer has to model a suitable architecture for the expected behavior.
- **Higher costs**
  - Multi-usage of LLM could lead to higher costs of inferencing.

Multi-agents

TBD