# Agentic LLM

Haw-Shiuan Chang

Improve performance and reduce costs

Pretty useful for smaller or non-tech companies

Hard to teach because it is always domain-dependent

Easy to learn because it is very intuitive and easy to understand

# Logistics

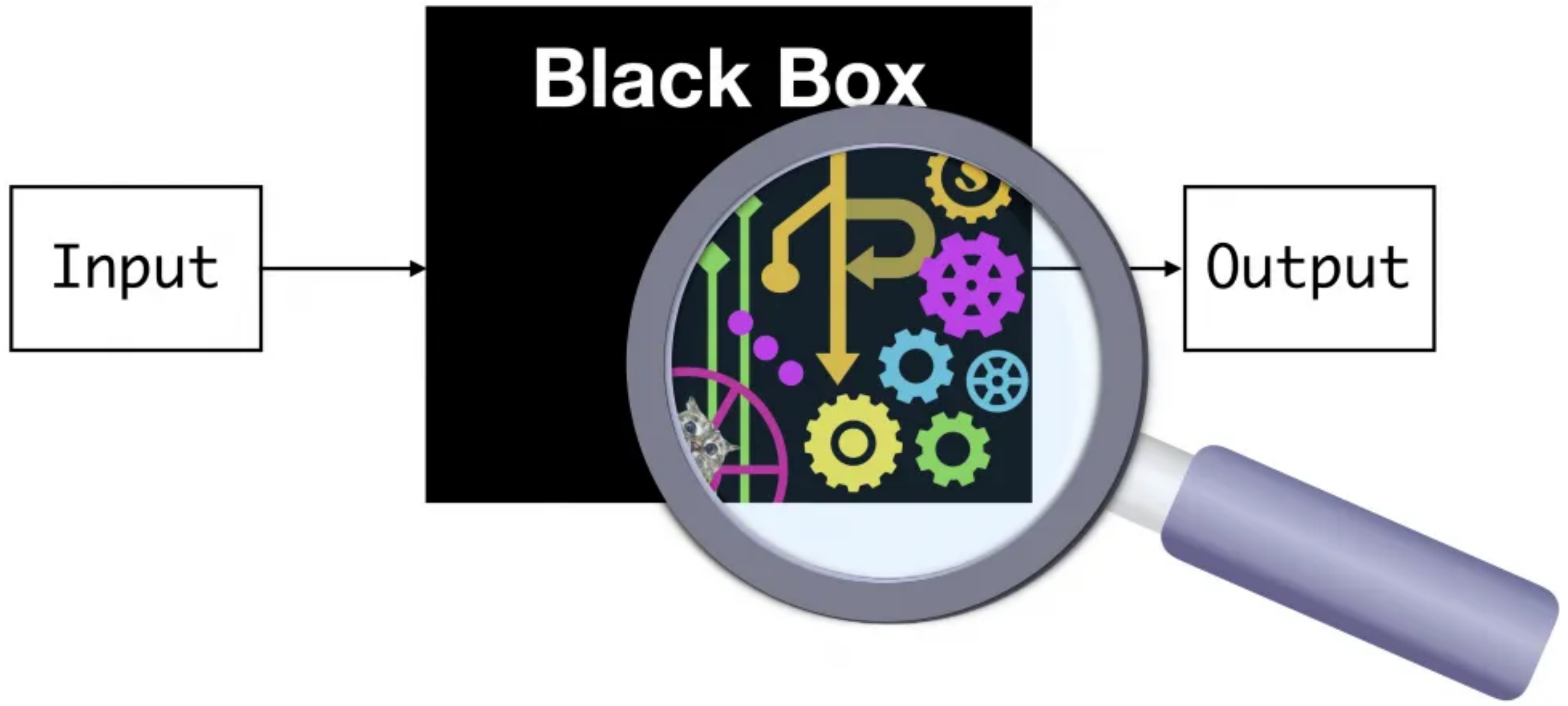
- **https://people.cs.umass.edu/~hschang/cs685/schedule.html**
  - **My office hour is moved to 3pm-4pm on Thursday**

- **Course survey (http://owl.umass.edu/partners/courseEvalSurvey/uma/) before 5/19**

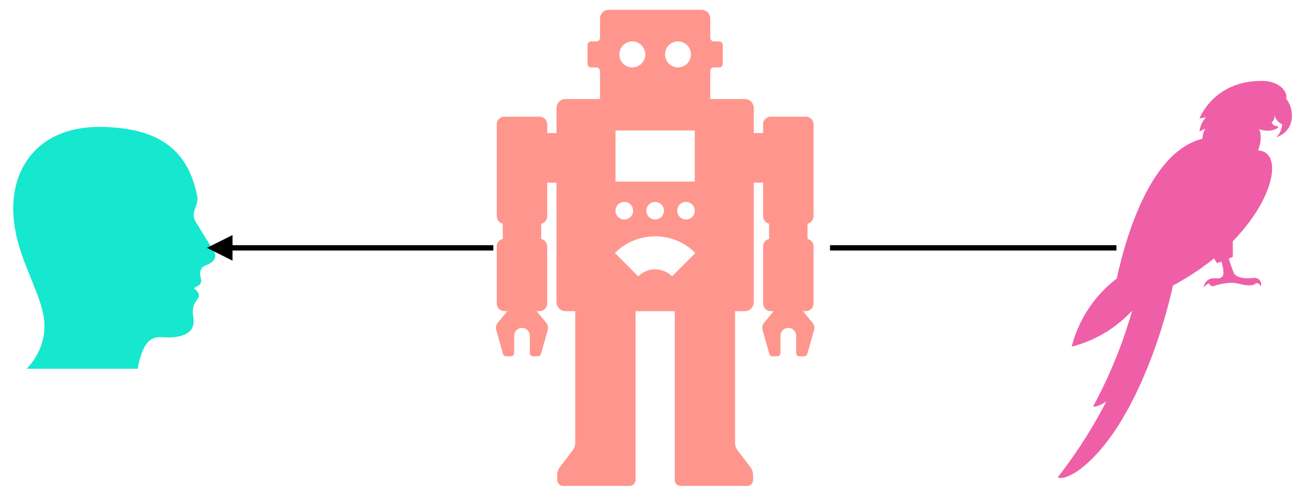
- **5/5: Quiz4**
- **5/9: Extra Credit (seminar)**
- **5/12: Extra Credit (course)**
- **5/12: Final project report due**
  - If your members do not contribute significantly, please let us know.
    - We will need to investigate and determine if we want to deduct the points from some members
  - **You can submit late until 5/16. Every late day costs 1 point.**

# Inference-time Improvement

- Prompt engineering
    - In-context learning
- Decoding
- **Agentic**
    - **RAG**
    - **Tools**
    - **Assistant**
    - **Multi-LLM collaboration**




Human brain is also almost a black box



Lots of cognitive science

https://blog.ml.cmu.edu/2019/05/17/explaining-a-black-box-using-deep-variational-information-bottleneck-approach/

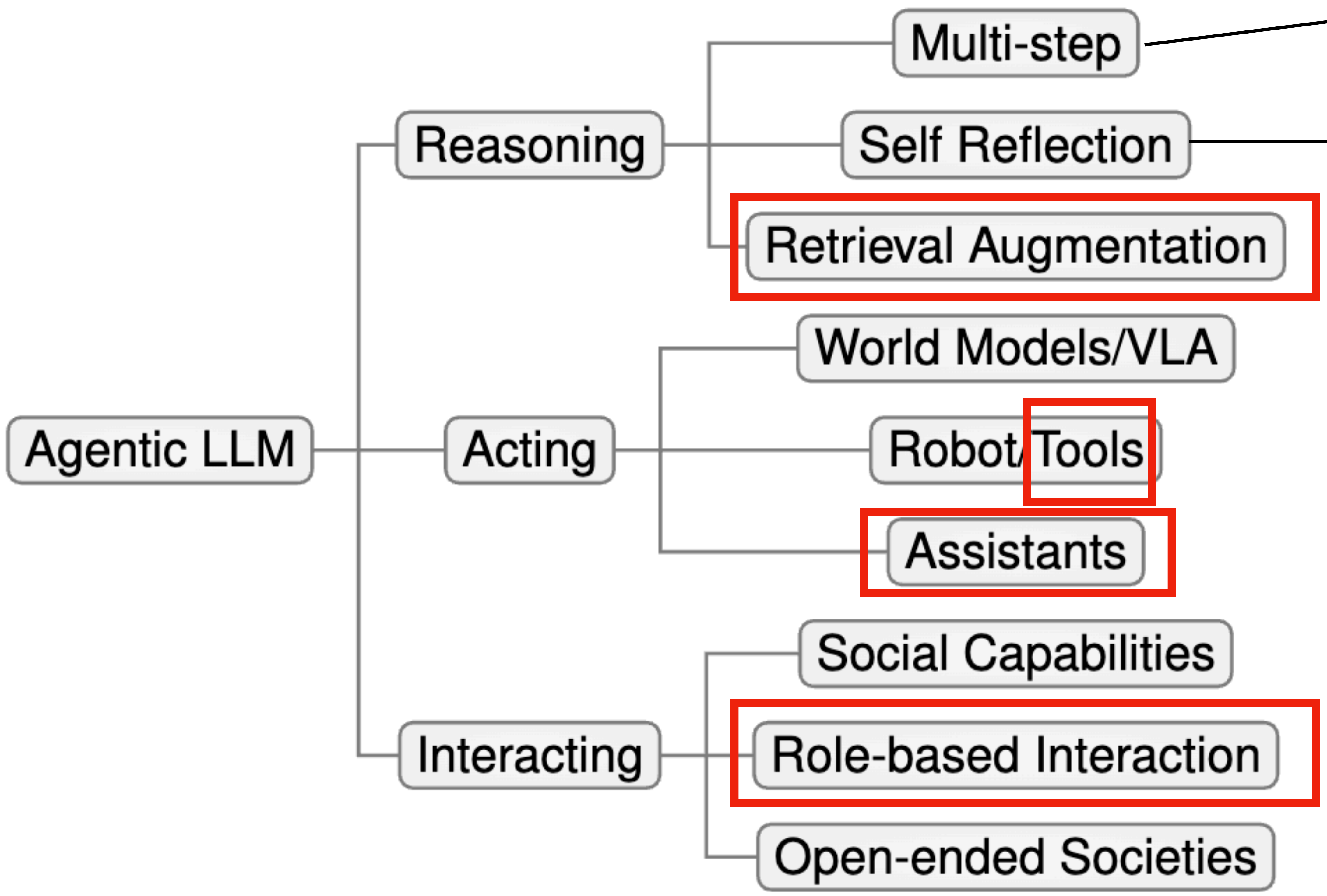
# A New Cherry on the Top

Agentic LLM


https://www.datadriveninvestor.com/2020/04/09/whats-next-in-ai/

# What is Agentic LLM?

One definition: Agentic LLM generally means treating LLM as a human




CoT (Covered)

Verifier (Covered)

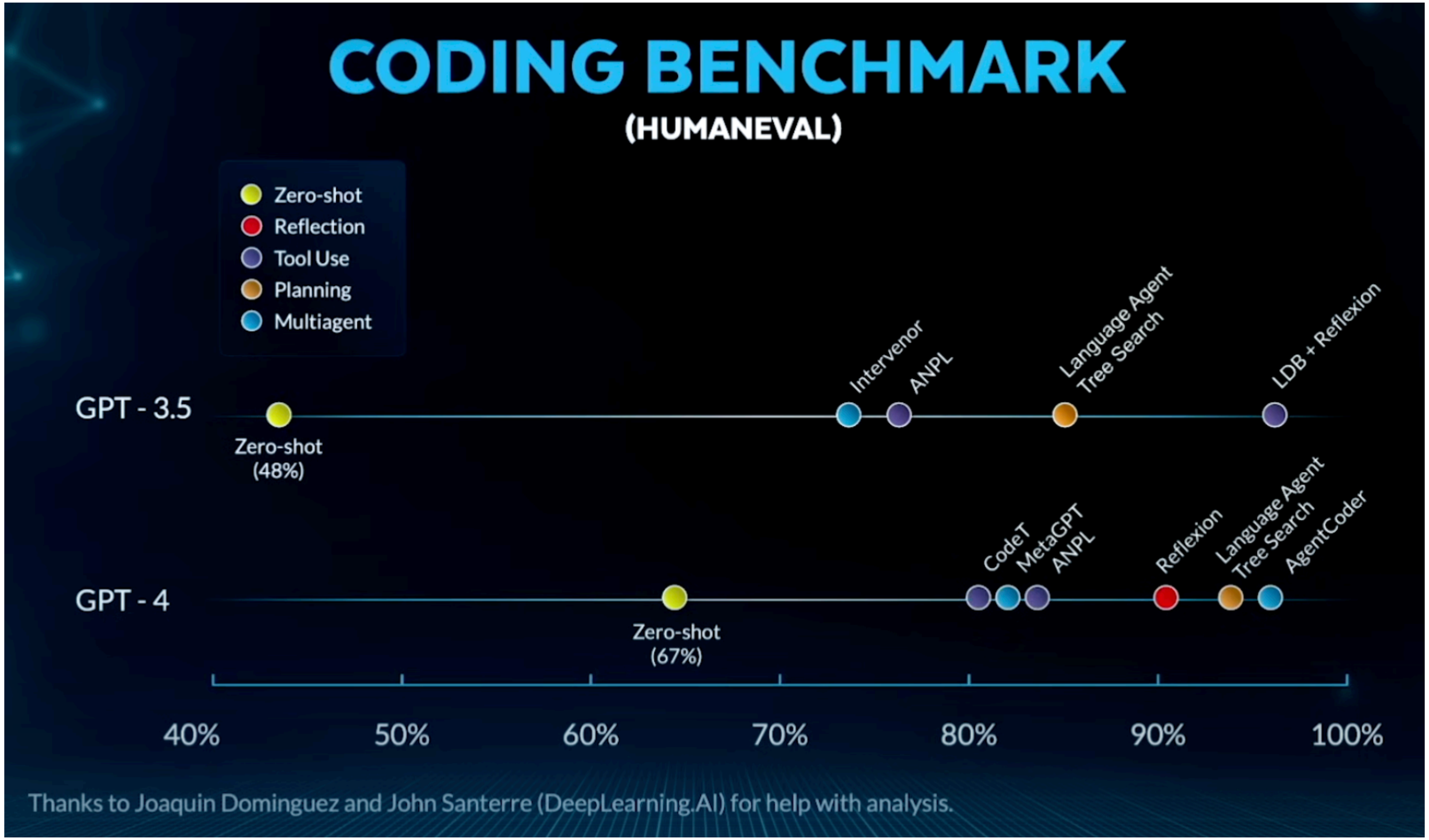
Figure 3: Agentic LLM Taxonomy of Reasoning, Acting, Interacting

Agentic Large Language Models, a survey (https://arxiv.org/pdf/2503.23037)


Figure 1: Generative agents are believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents plan their days, share news, form relationships, and coordinate group activities.

Generative Agents: Interactive Simulacra of Human Behavior (https://arxiv.org/pdf/2304.03442)

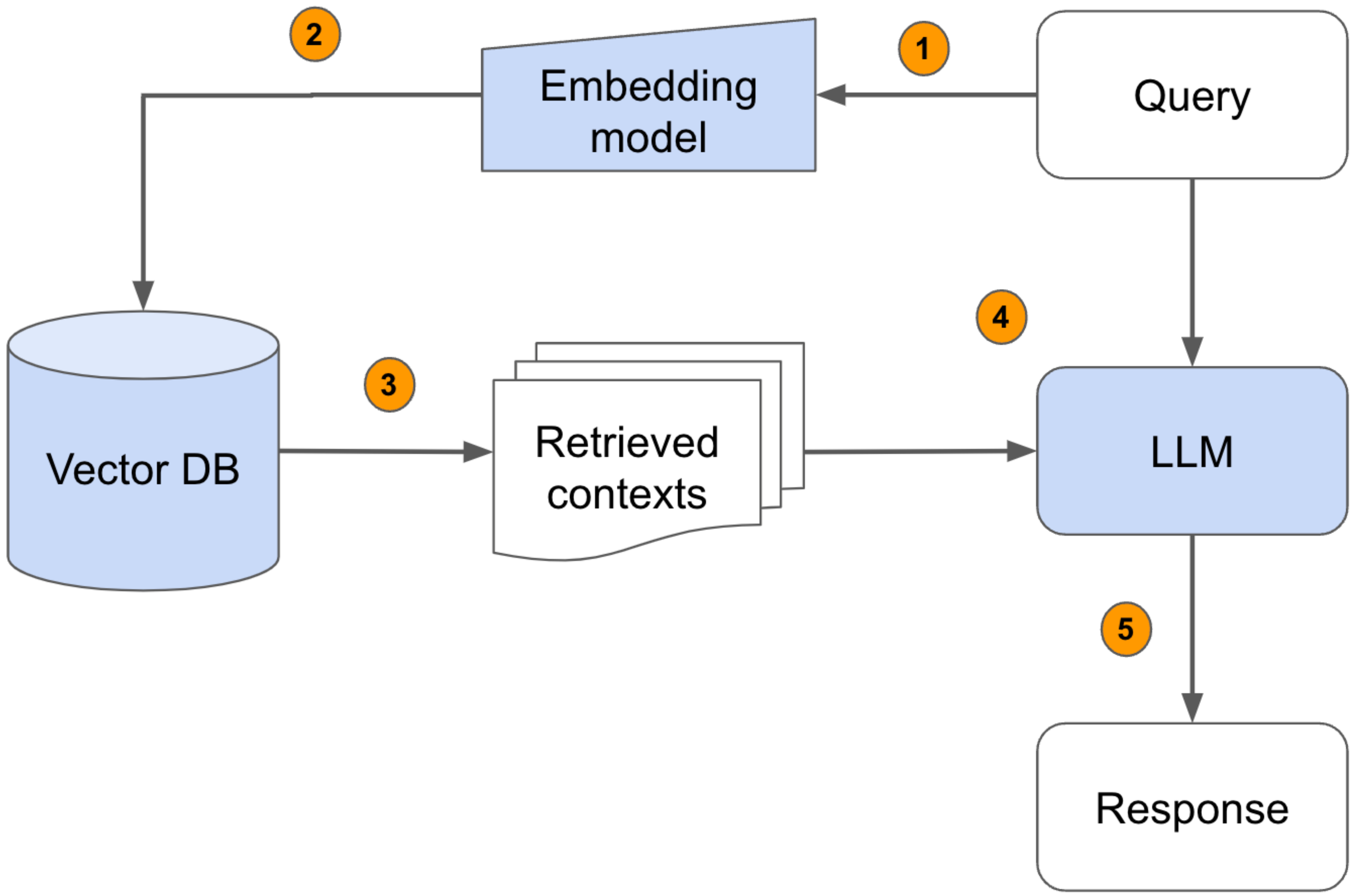
# Why Agentic LLM Matter




https://youtu.be/KrRD7r7y7NY?si=ly9OZJyrE7ztKwwl

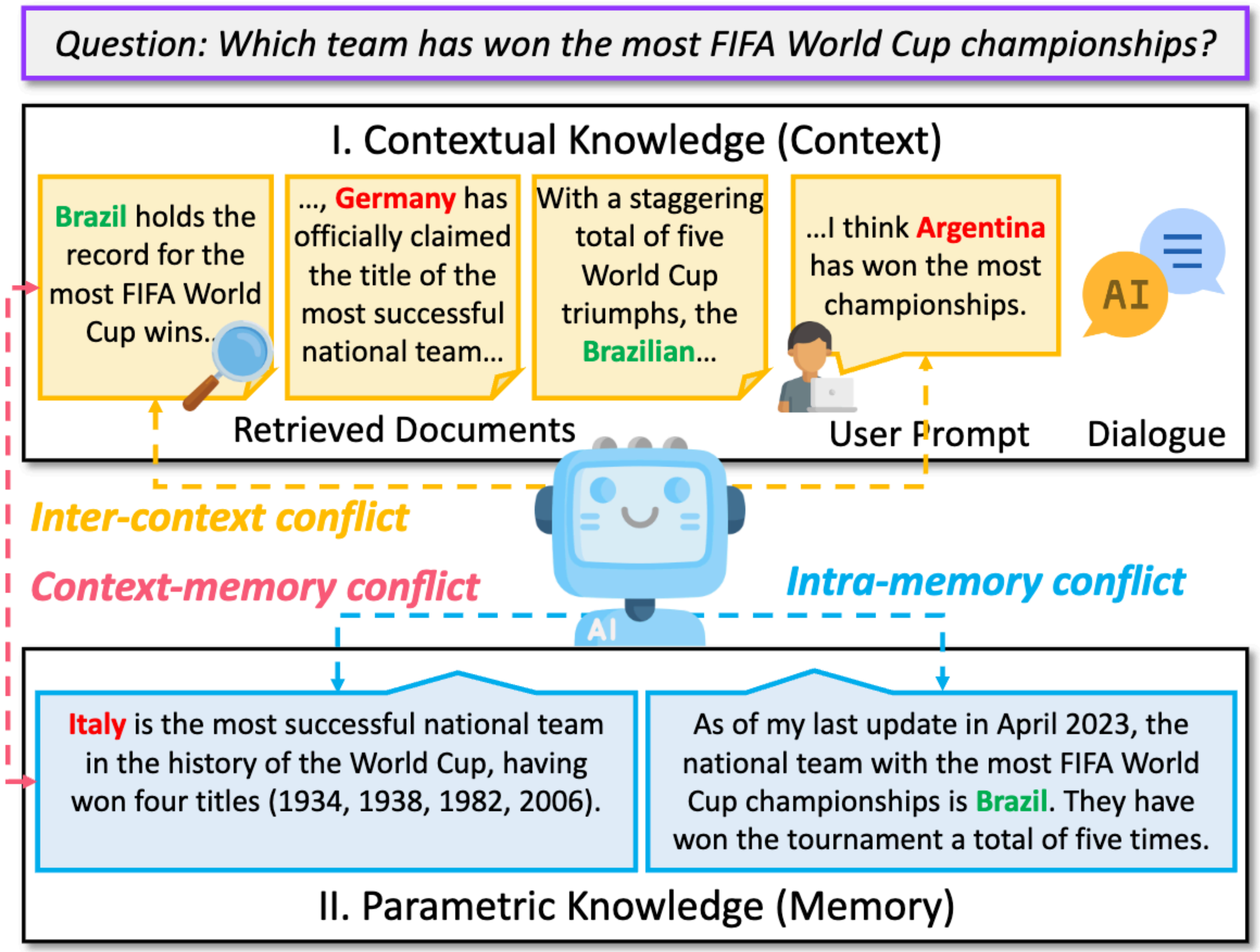
# Retrieval-Augmented Generation (RAG)

What do you do if I ask you which year Barack Obama is born?



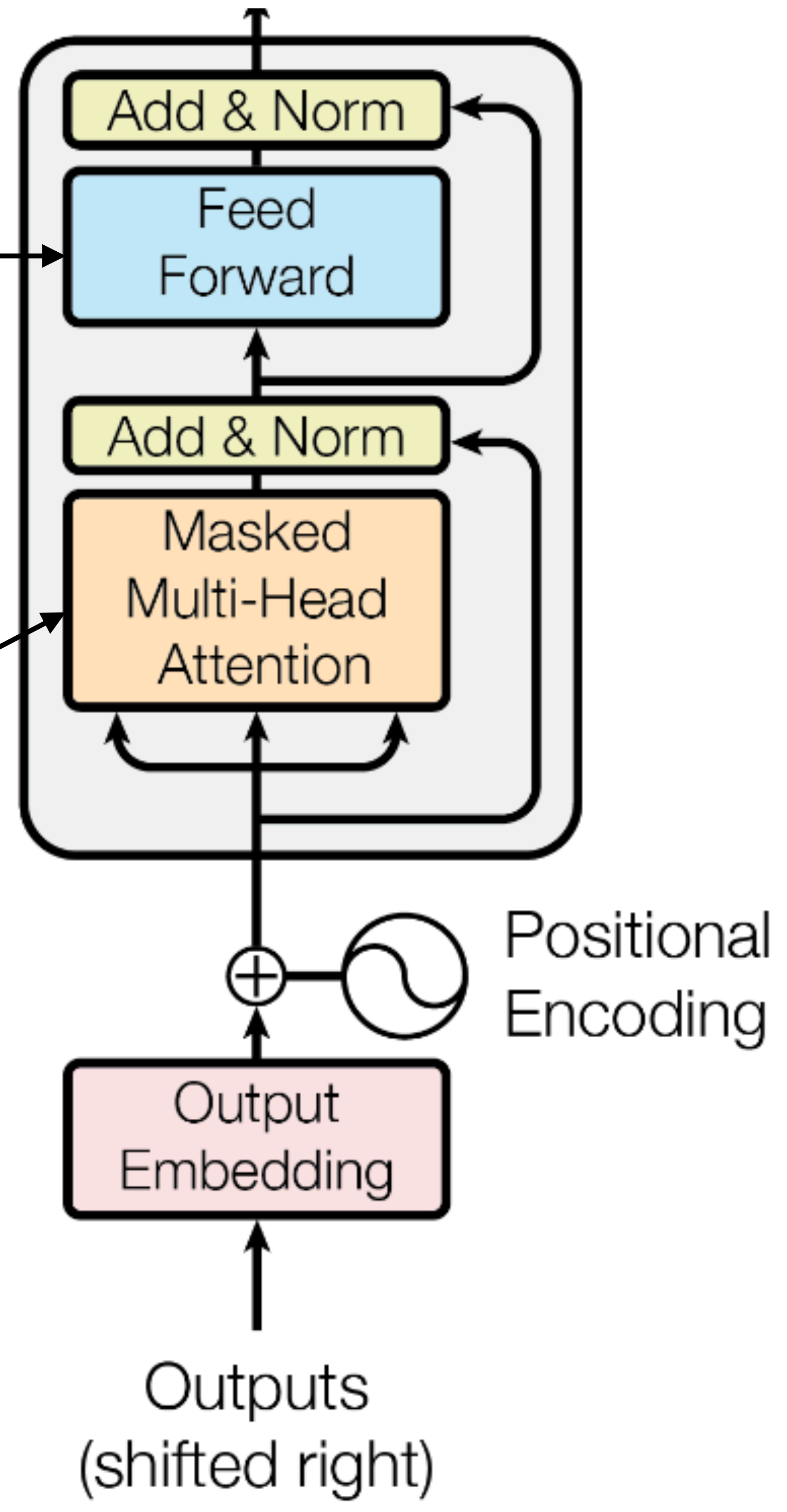

https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172

# Knowledge Conflict



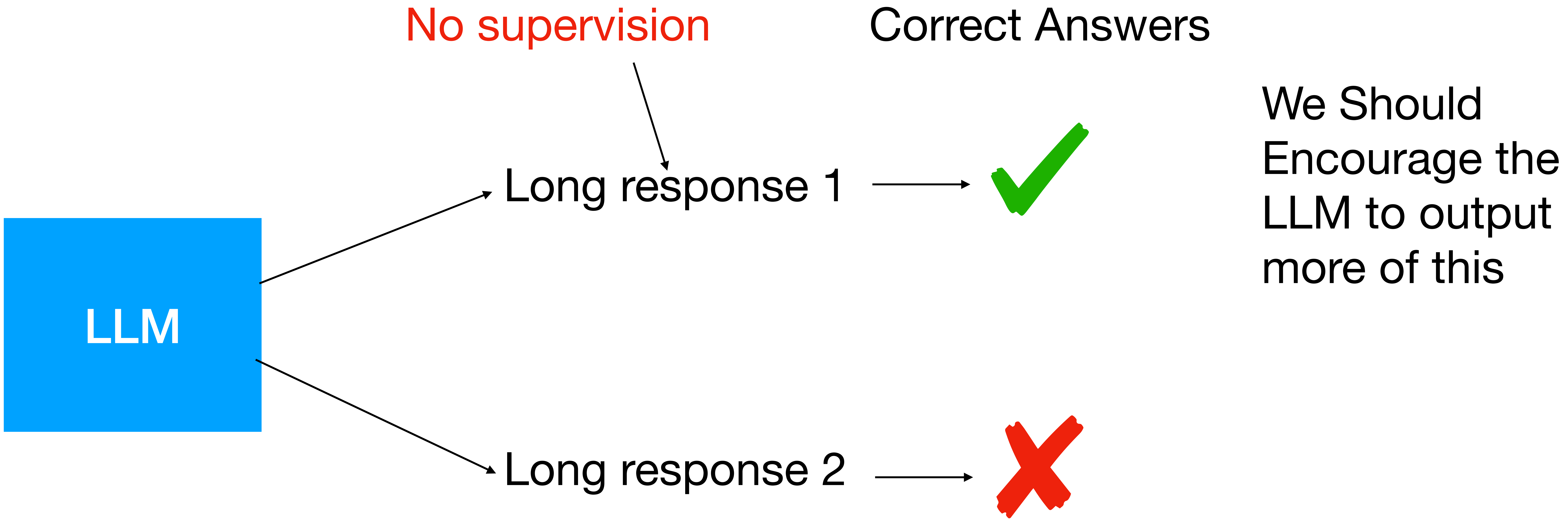

Memory: Parametric knowledge

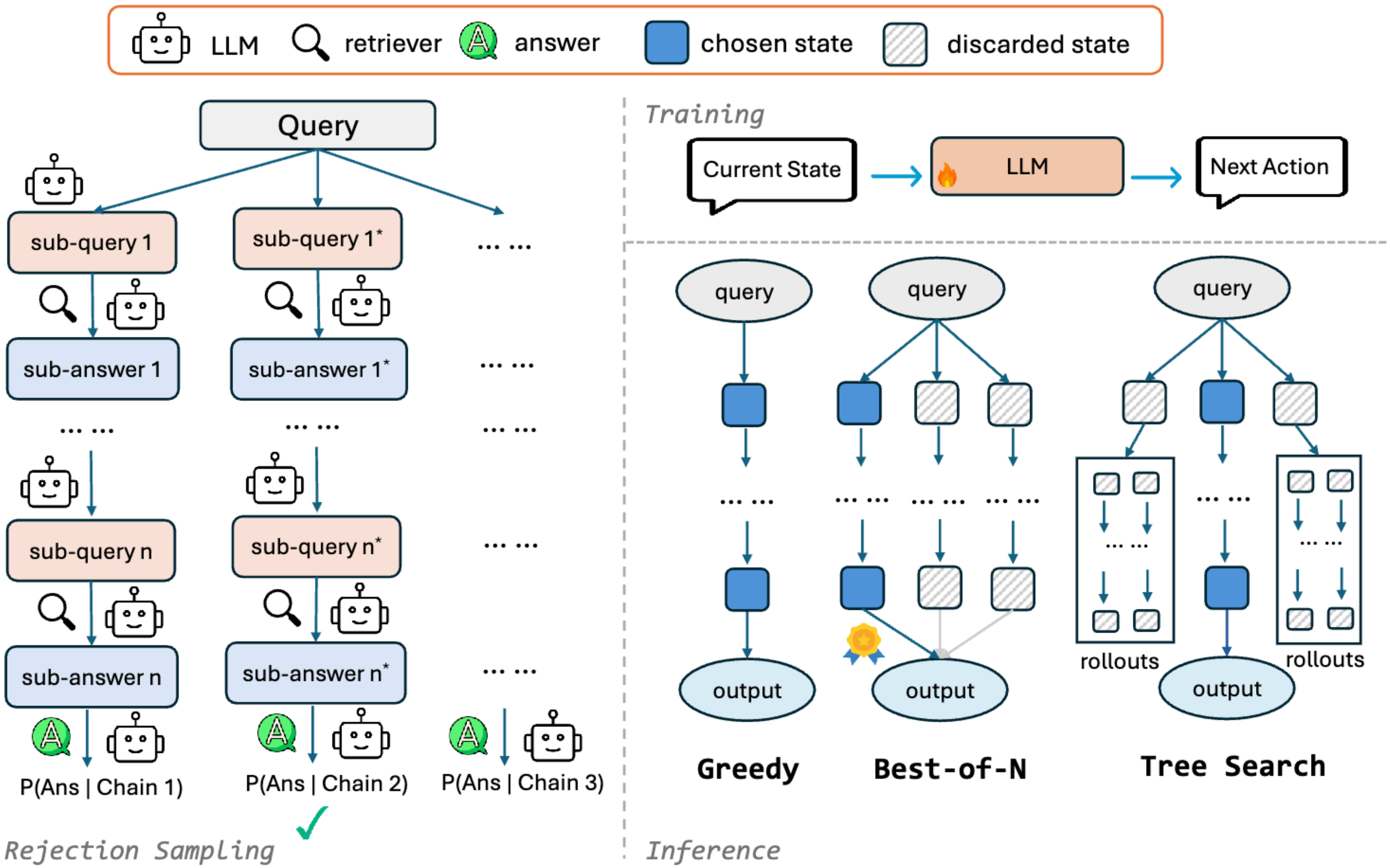
Consistnecy: Context knowledge




Knowledge Conflicts for LLMs: A Survey (https://arxiv.org/pdf/2403.08319v1)

# Reasoning -> Distant Supervision






Chain-of-Retrieval Augmented Generation (https://arxiv.org/pdf/2501.14342)

# Tool Usage

- Tools could be a calculator, search engine, python program, joke generators, ….
- RAG

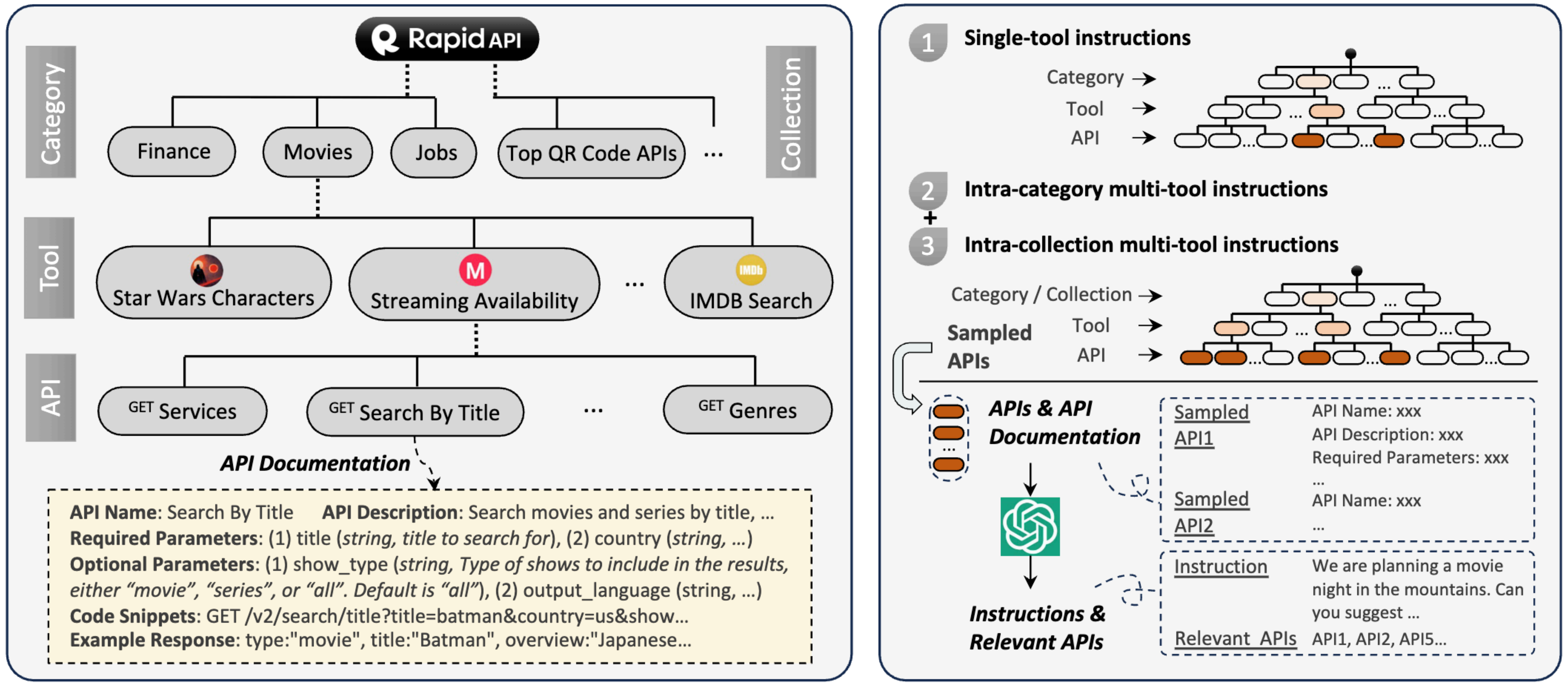



Figure 3: The hierarchy of RapidAPI (left) and the process of instruction generation (right).

TOOLLLM: FACILITATING LARGE LANGUAGE MODELS TO MASTER 16000+ REAL-WORLD APIS (https://arxiv.org/pdf/2307.16789)

# Let LLM Control your Computer

- Cool Example:
- https://www.reddit.com/r/mcp/comments/1k3bldw/unity_mcp_server_game_level_creation/
- Do you feel comfortable to let LLM control your computers?


| Method | Paradigm | Completeness | Executability | Consistency | Quality |
|---|---|---|---|---|---|
| GPT-Engineer | | 0.5022$^{\dagger}$ | 0.3583$^{\dagger}$ | 0.7887$^{\dagger}$ | 0.1419$^{\dagger}$ |
| MetaGPT | | 0.4834$^{\dagger}$ | 0.4145$^{\dagger}$ | 0.7601$^{\dagger}$ | 0.1523$^{\dagger}$ |
| ChatDev | | **0.5600** | **0.8800** | **0.8021** | **0.3953** |

Table 1: Overall performance of the LLM-powered software development methods, encompassing both single-agent ( ) and multi-agent ( ) paradigms. Performance metrics are averaged for all tasks. The top scores are in bold, with second-highest underlined. † indicates significant statistical differences ($p \leq 0.05$) between a baseline and ours.

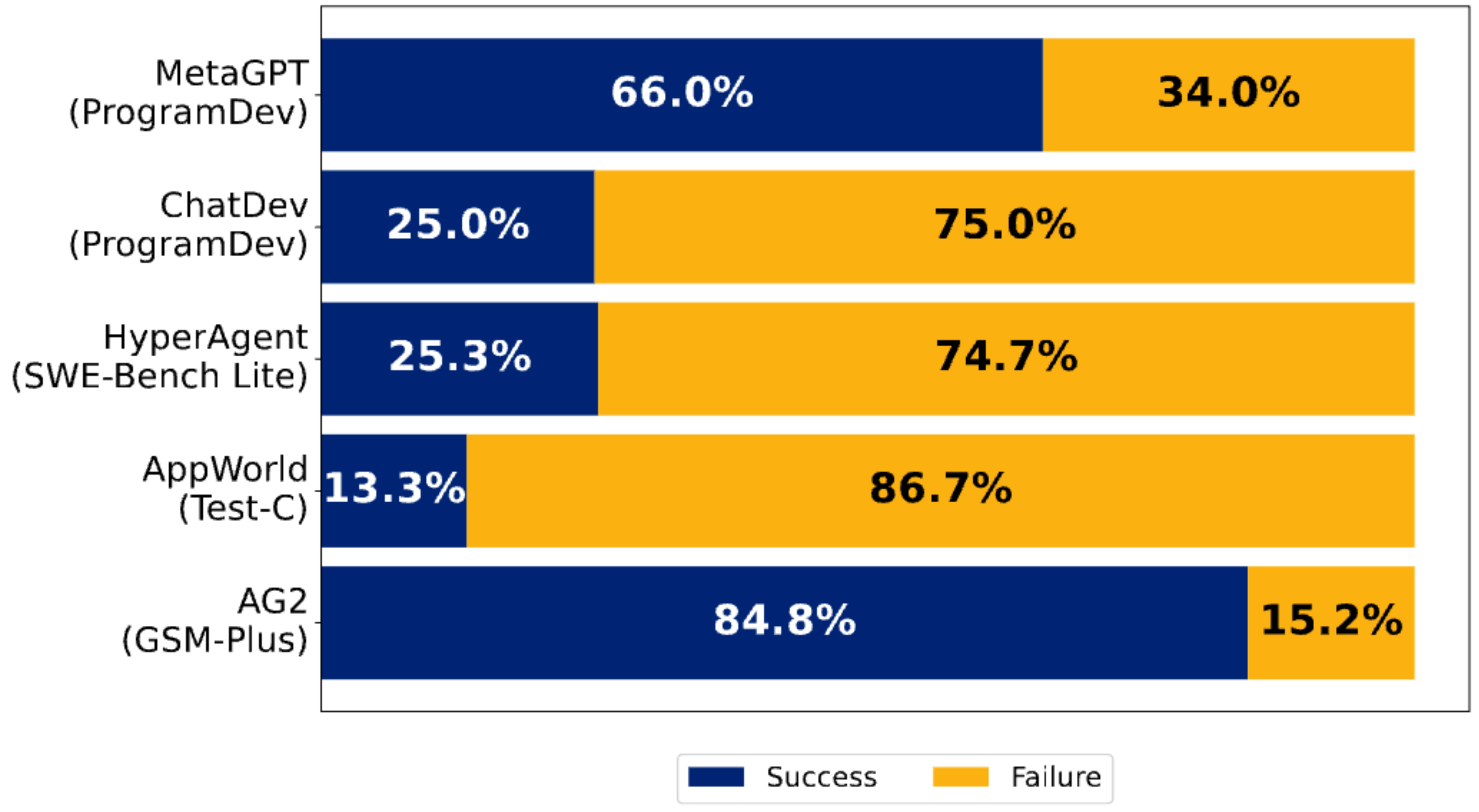



*Figure 1.* Failure rates of five popular Multi-Agent LLM Systems with GPT-4o and Claude-3.

ChatDev: Communicative Agents for Software Development (https://arxiv.org/pdf/2307.07924)

Why Do Multi-Agent LLM Systems Fail? (https://arxiv.org/pdf/2503.13657)

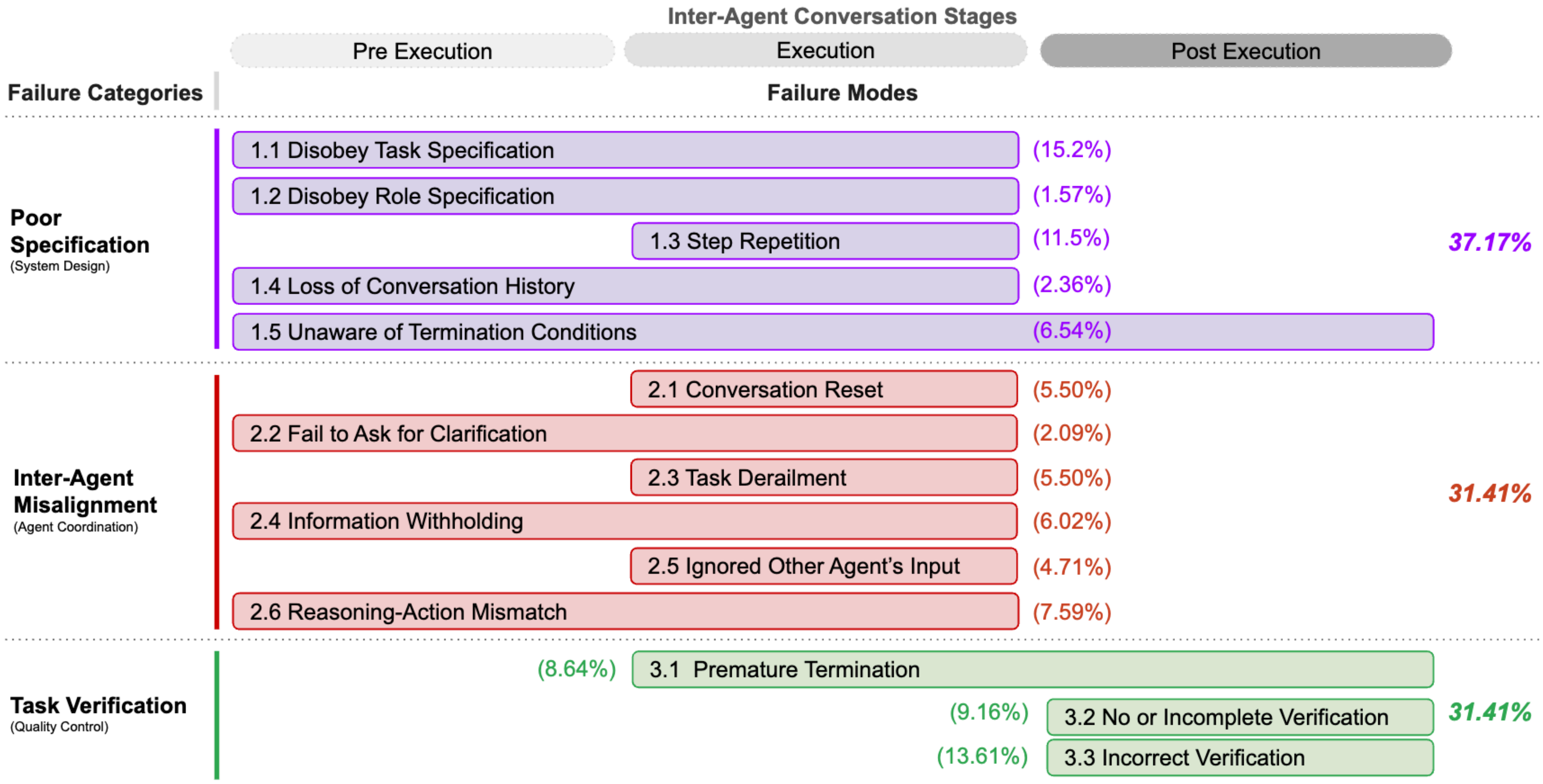
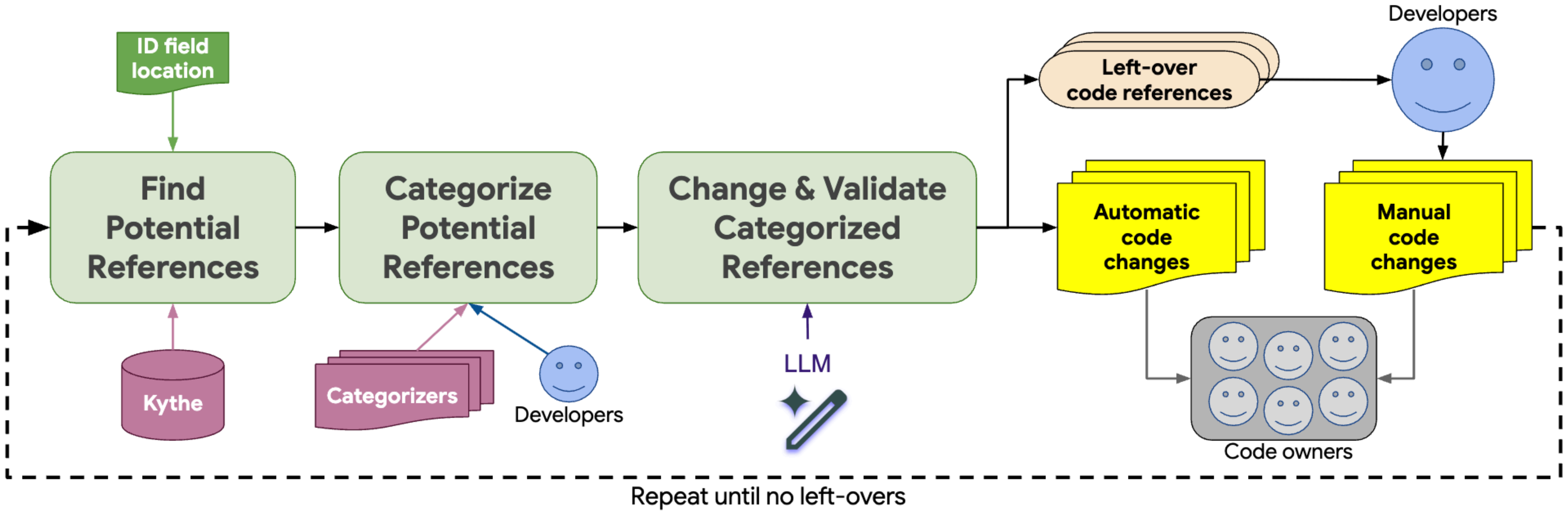



*Figure 2.* A **Taxonomy of MAS Failure Modes**. The inter-agent conversation stages indicate when a failure can occur in the end-to-end MAS system. If a failure mode spans multiple stages, it means the issue involves or can occur at different stages. Percentages represent how frequently each failure mode and category appeared in our analysis of 151 traces. Detailed definition and example of each failure mode is available in Appendix A.

# Human-LLM Collaboration




```
public static final int MAX_TOY_ID = 69234567;
public static final long MAX_TOY_ID = 1000069234567L;
```

```
inline constexpr int kMaxToyId = 69234567;
inline constexpr int64_t kMaxToyId = 1000069234567;
```

**(a) Changes in multiple languages with almost identical prompt**

```
int nonExistingToyId = RANDOM.nextInt();
long nonExistingToyId = RANDOM.nextLong();
```

**(b) Language specific domain knowledge**

```
Toy toy = Toy.newBuilder().setToyId(toyId).build();
Toy toy = Toy.newBuilder()
            .setToyId(toyId)
            .build();
```

**(a) Hallucination that reformats file contents**

```
Toy toy = Toy.newBuilder().setToyId(1369873).build();
Toy toy = Toy.newBuilder().setToyId(/* toyId */ 1369873).build();
```

**(b) Hallucination that adds comments**

**Figure 6: Examples of LLM hallucinations.**

| | |
|---|---|
| Total code changes | 595 |
| LLM-Only | 214 (35.97%) |
| LLM-then-Human | 229 (38.48%) |
| Human-Only | 152 (25.55%) |
| Total code changes | 595 |
| # reviewers | 306 |
| # teams | 149 |
| # offices | 37 |
| # time zones | 12 |
| Total$\Delta$ across all IDs | 93, 574 |
| LLM$\Delta$ | 64, 996 (69.46%) |
| Human$\Delta$ | 28, 578 (30.54%) |

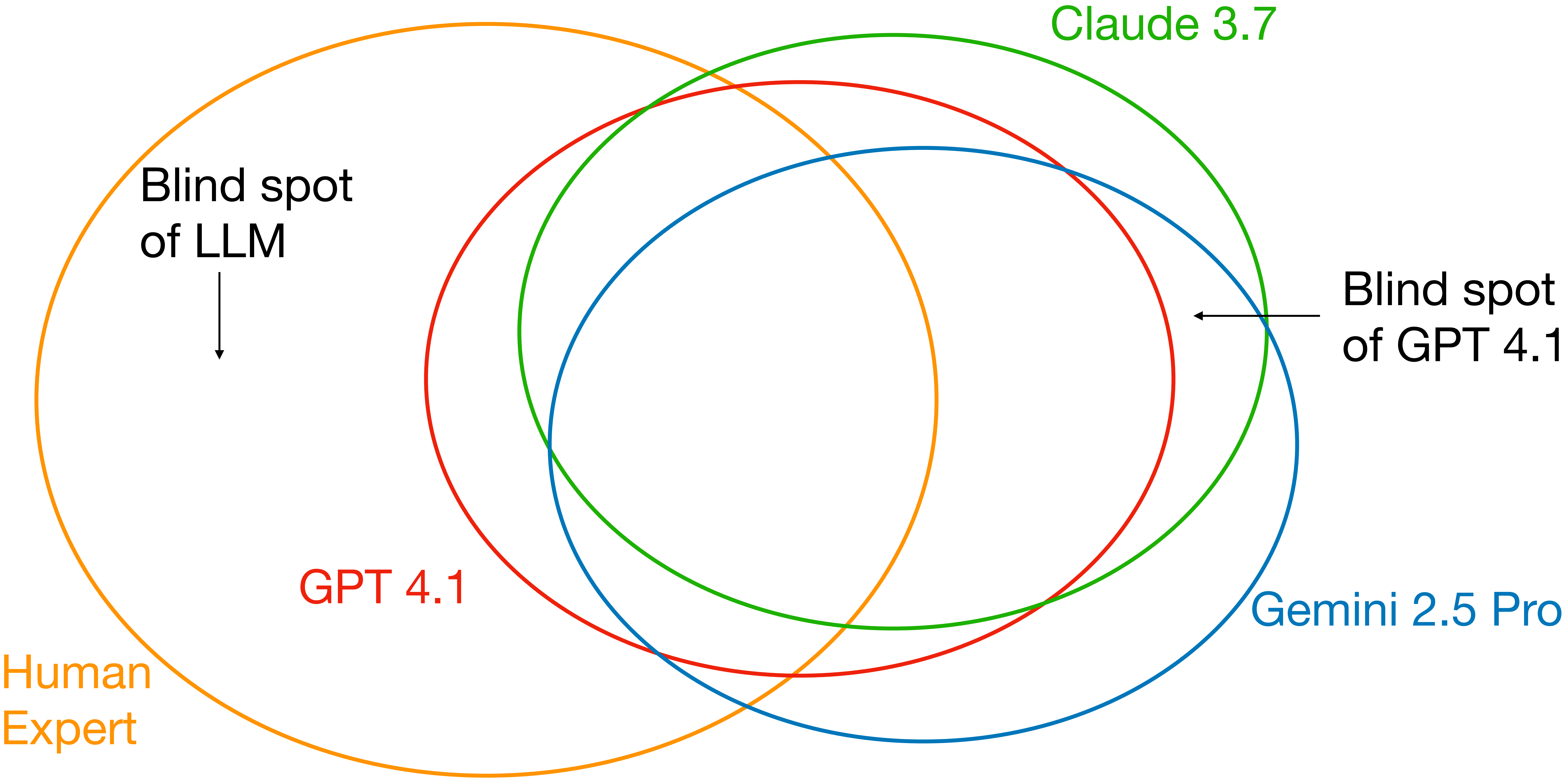
Migrating Code At Scale With LLMs At Google (https://arxiv.org/pdf/2504.09691v1)

# Question

- Why does multiple LLM collaboration Work?

- Retrieve the relevant information
- Fixing the blind spots

# Fixing Blind Spots

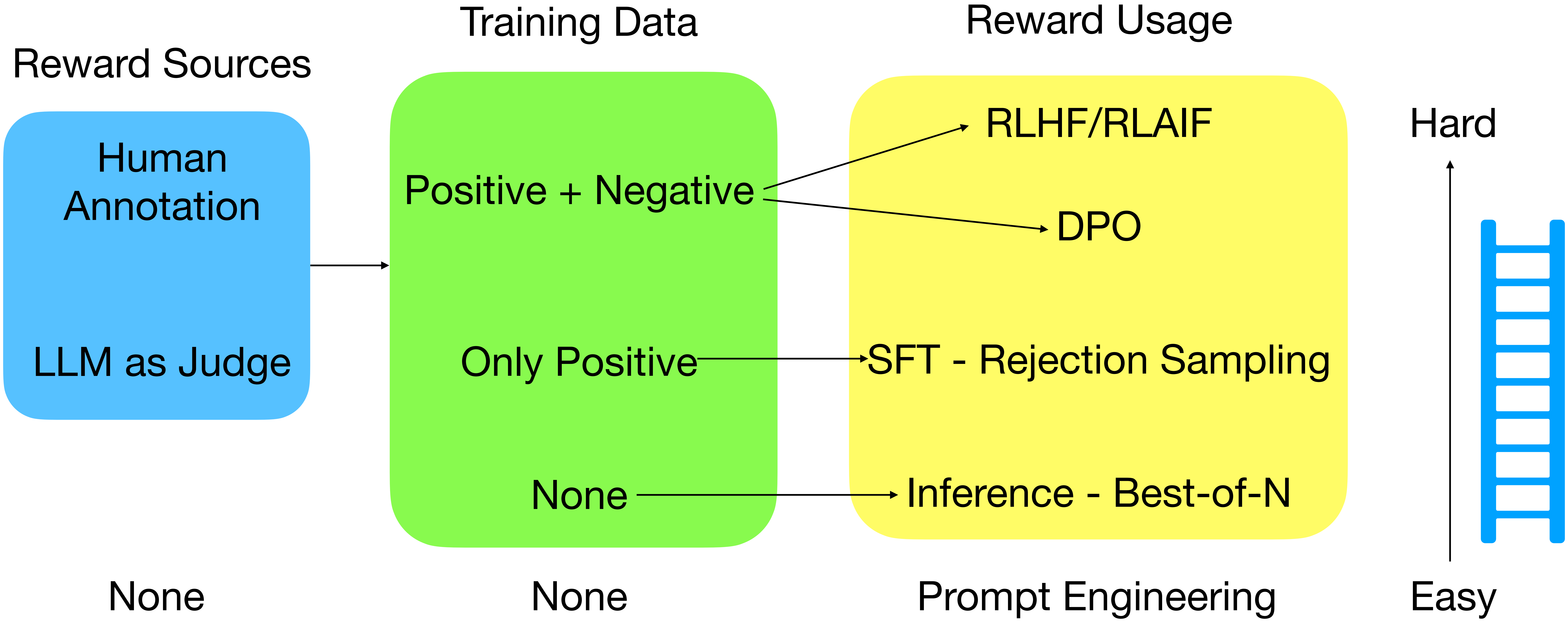


# Agentic LLM vs LRM

- Agentic LLM
  - Pros
    - Easier to try
    - More interpretable
  - Cons
    - Requires lots of effort to do prompt engineering
    - Usually more expensive

- LRM
  - Pros
    - Usually perform better
  - Cons
    - Require lots of answers for RL

# Available Methods



# Challenges in Agentic LLMs

- Many moving components, so it is difficult to
  - conduct error analysis
  - fix some critical errors or further improve systems
- The performance might be worse than more advanced results
- Hard to know why the performance is better
- The lessons learned from one application are hard to transfer to other applications or other LLMs

# Improving Environments or Agents

- We know that evaluation could be used to optimize LLMs
- Environment/Evaluation is usually a mix of rules, tools, and data
- Do the fundamental limitations of LLMs come from data or models?
- Should we focus on improving the environment or the agent itself?

Welcome to the Era of Experience (https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf)