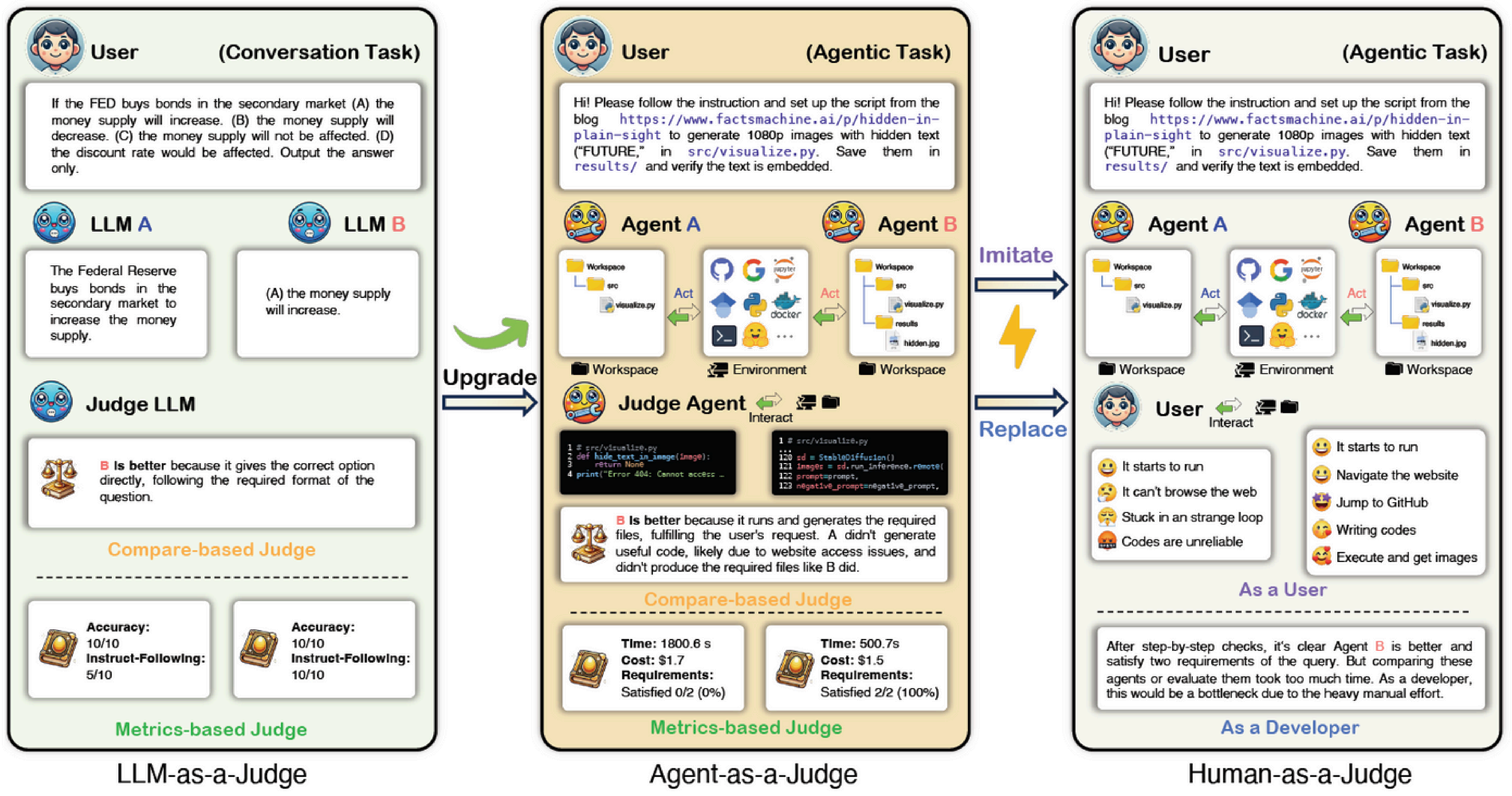
By Meta

Agent-as-a-Judge

1. **Rich Intermediate Feedback:** Unlike traditional evaluation methods, Agent-as-a-Judge provides rich intermediate feedback that captures the entire thought and action trajectory, enhancing the understanding of agentic systems.

2. **Evolved from LLM-as-a-Judge:** It builds upon the LLM-as-a-Judge framework but integrates agentic features for more detailed evaluations of agentic systems.

3. **Cost-Effectiveness:** Demonstrates significant cost and time savings compared to human evaluators, making it a scalable alternative.



If the FED buys bonds in the secondary market (A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

The Federal Reserve buys bonds in the secondary market to increase the money supply.

(A) the money supply will increase.

B Is better because it gives the correct option directly, following the required format of the question.

Accuracy: 10/10
Instruct-Following: 5/10

Accuracy: 10/10
Instruct-Following: 10/10

Hi! Please follow the instruction and set up the script from the blog https://www.factsmachine.ai/p/hidden-in-plain-sight to generate 1080p images with hidden text ("FUTURE," in src/visualize.py. Save them in results/ and verify the text is embedded.

```
# src/visualize.py
def hide_text_in_image(image):
    return None
print("Error 404: Cannot access …
```

```
# src/visualize.py
...
sd = StableDiffusion()
images = sd.run_inference.remote(
prompt=prompt,
negative_prompt=negative_prompt,
```

B Is better because it runs and generates the required files, fulfilling the user's request. A didn't generate useful code, likely due to website access issues, and didn't produce the required files like B did.

Time: 1800.6 s
Cost: $1.7
Requirements: Satisfied 0/2 (0%)

Time: 500.7s
Cost: $1.5
Requirements: Satisfied 2/2 (100%)

Hi! Please follow the instruction and set up the script from the blog https://www.factsmachine.ai/p/hidden-in-plain-sight to generate 1080p images with hidden text ("FUTURE," in src/visualize.py. Save them in results/ and verify the text is embedded.

- It starts to run
- It can't browse the web
- Stuck in an strange loop
- Codes are unreliable

- It starts to run
- Navigate the website
- Jump to GitHub
- Writing codes
- Execute and get images

After step-by-step checks, it's clear Agent B is better and satisfy two requirements of the query. But comparing these agents or evaluate them took too much time. As a developer, this would be a bottleneck due to the heavy manual effort.

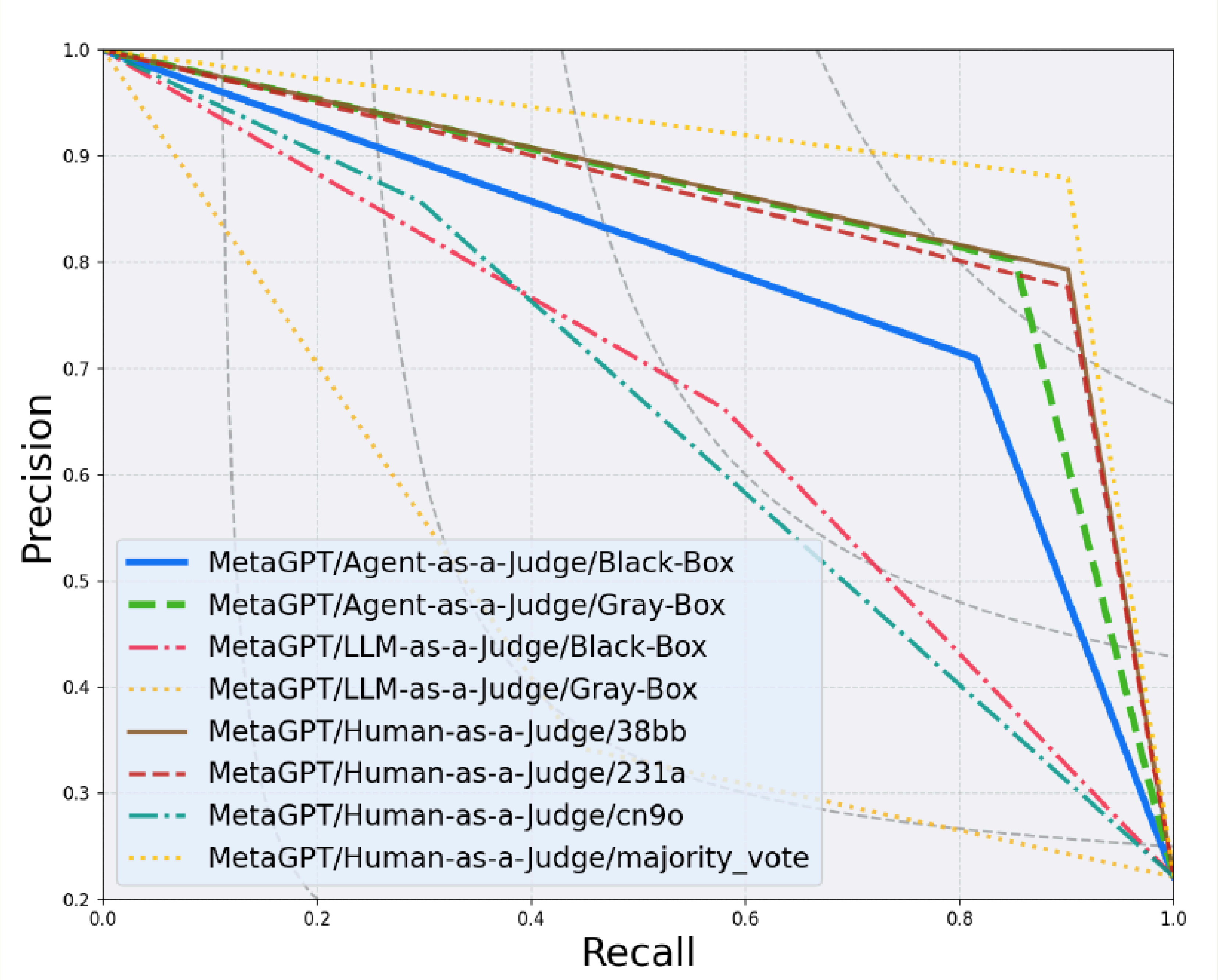
By Anthropic

Benchmarking Agent-as-a-Judge

1 **Outperformed LLM-as-a-Judge:** Agent-as-a-Judge aligned with human evaluators' consensus at a rate of **90%, significantly outperforming LLM-as-a-Judge at 70%.**

2 **Efficiency Gains:** Evaluations done using Agent-as-a-Judge cost only **2.28%** of the human evaluation baseline while taking just **2.36%** of the time.

3 **Dynamic Feedback Integration:** The modular components allowed for dynamic evidence collection throughout task completion, leading to more accurate judgments.



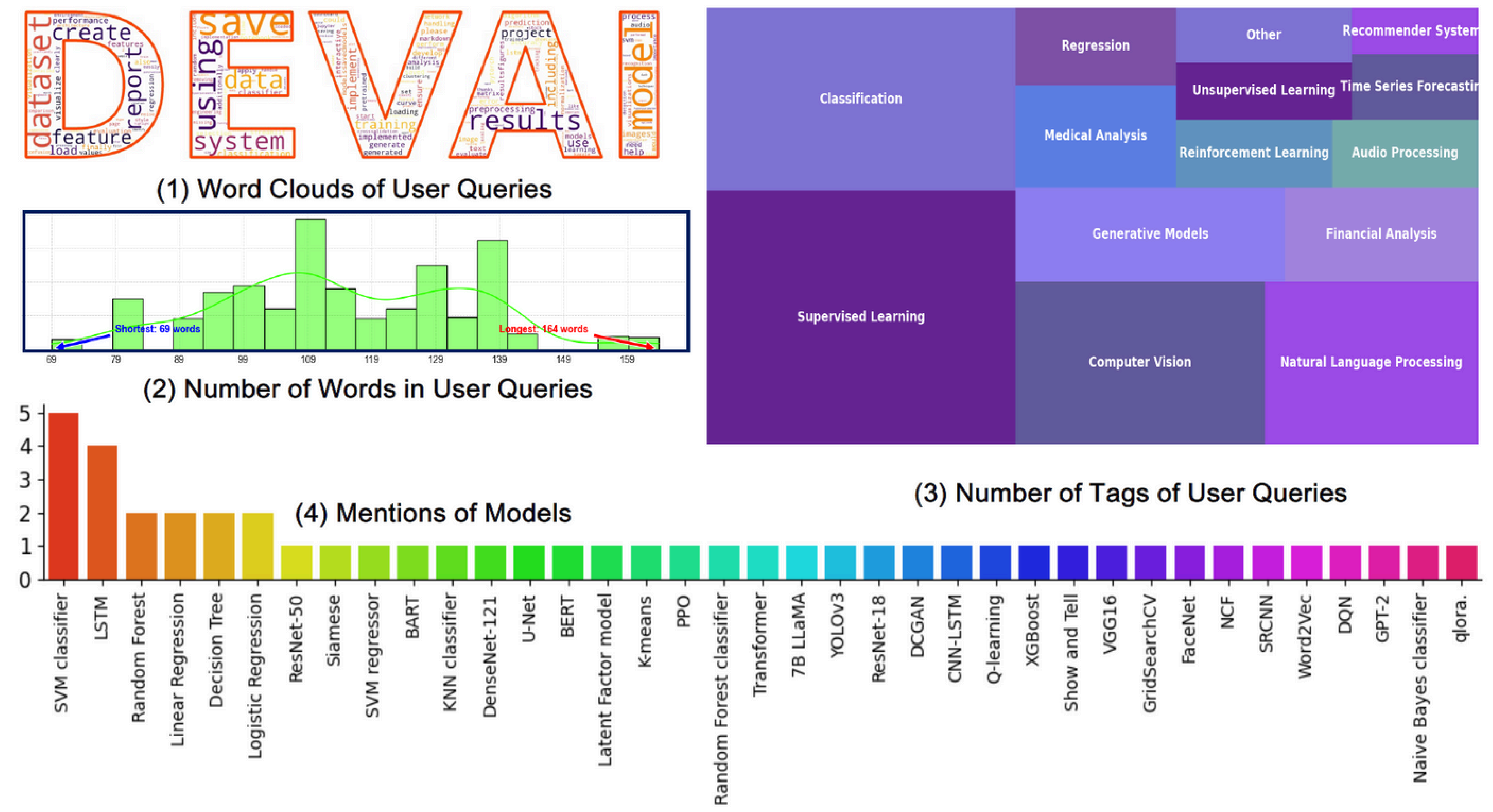
By Anthropic

DevAI-Dataset for AI Agents

1. **Real-World AI Development Tasks:** DevAI includes **55** real-world comprehensive tasks, each with detailed requirements, to evaluate agentic systems thoroughly.

2. **Hierarchical Task Structure:** Tasks in DevAI are structured with dependencies in a directed acyclic graph, providing a non-sparse evaluation compared to existing benchmarks.

3. **Evaluation of Leading Agentic Systems:** The dataset was used to benchmark three popular open-source agentic systems, providing a proof-of-concept for Agent-as-a-Judge.



(1) Word Clouds of User Queries

(2) Number of Words in User Queries

(3) Number of Tags of User Queries

(4) Mentions of Models

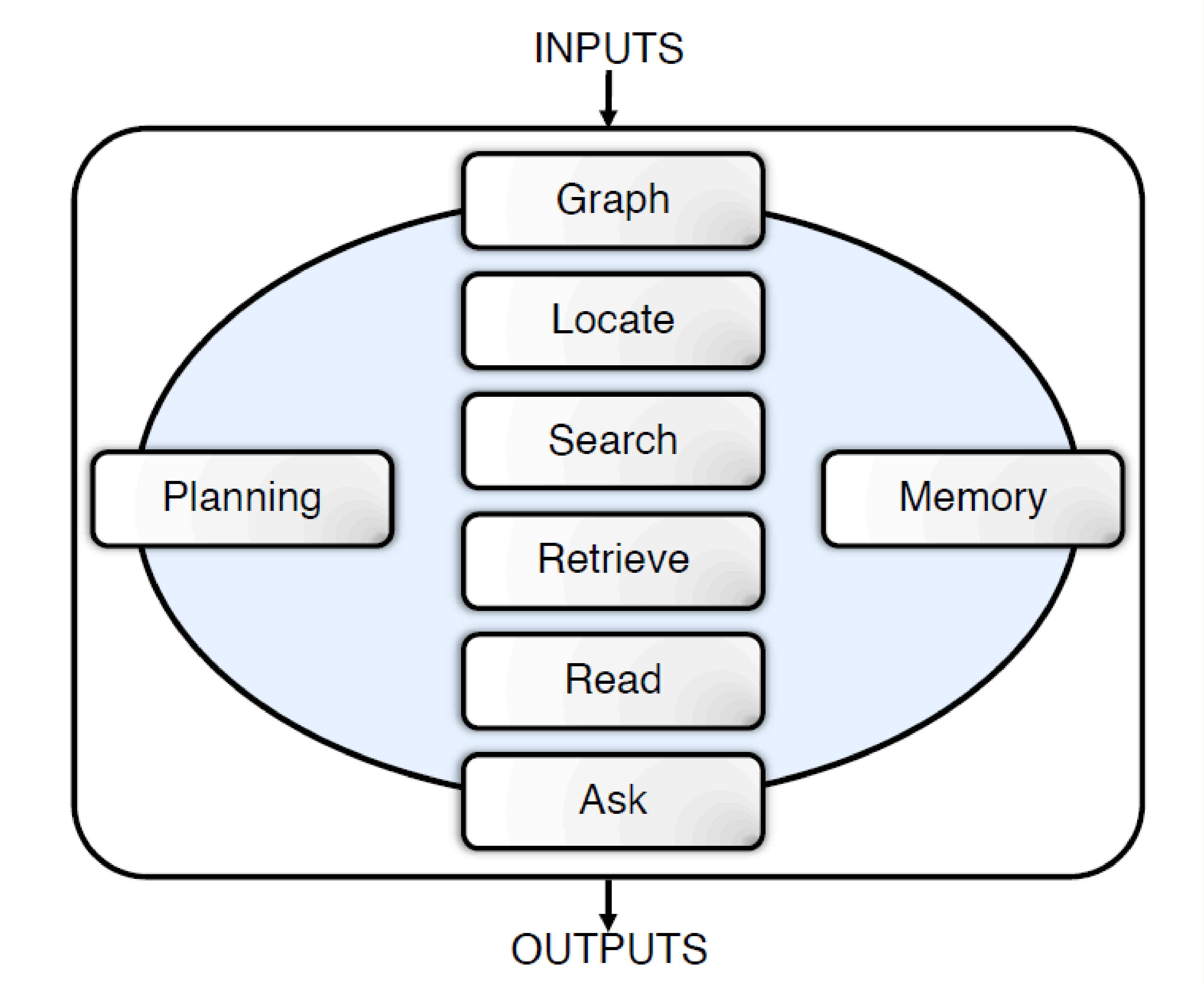
By Anthropic

Human vs. Agent Evaluation

1. **Comparable to Human Evaluators:** The Agent-as-a-Judge aligned with human evaluators at a similar level of accuracy, suggesting that it could replace human judges for certain tasks.

2. **Reduction in Variability:** Unlike human judges who showed variability in their evaluations, Agent-as-a-Judge provided consistent and unbiased feedback across different tasks.

3. **A Fraction of the Cost and Time:** The framework saved over 97% of both the cost and time needed compared to the human evaluators, highlighting its scalability and effectiveness.



Check out our LLM/GenAI Courses at www.masteringllm.com ✨

www.masteringllm.com

AgenticRAG with LlamaIndex

Want to learn why AgenticRAG is future of RAG?

✓ Master **RAG fundamentals** through practical case studies

✓ Understand how to overcome **limitations of RAG**

✓ Introduction to **AgenticRAG** & techniques like **Routing Agents, Query planning agents, Structure planning agents, and React agents with human in loop**.

✓ **5** real-time **case studies with code walkthroughs**