Deploying AI

Agents

```
$ echo "Data Science Institute"
```

Introduction

Agenda

Agenda

- Planning

- Interacting with APIs and MCP

- Agent failure modes and evaluation

Agents

Introduction to Agents

- An agent perceives its environment and acts upon it.

- Agents are defined by their environment and available actions.

- They can leverage tools to expand their capabilities.

- Chatbots with retrieval or browsing abilities are examples of agents.

- Agents combine reasoning, planning, and tool use.

Why Agents Matter

- Agents can automate workflows that require multiple steps.

- They integrate perception, planning, and action into cohesive loops.

- Examples include research assistants, trip planners, and negotiation bots.

- Agents represent a path toward autonomous, goal-driven AI systems.

Agent Components

- **Environment**: defines the context in which the agent operates.

- **Actions**: possible operations available to the agent.

- **Tools**: extend the agent's ability to perceive or act.

- **Planner**: determines how to sequence actions to reach goals.

- **Feedback loop**: updates decisions based on environment responses.

Tools for Agents

- Tools enhance agents by providing knowledge and capabilities.
- **Knowledge augmentation** tools retrieve or access data sources.
- **Capability extension** tools solve inherent model weaknesses, such as calculators.
- **Write actions** tools allow agents to alter environments, such as sending emails.
- Tool selection shapes the effectiveness and reliability of agents.

Knowledge Augmentation Tools

- Examples include text retrievers, image retrievers, and SQL executors.

- Tools can provide access to organizational databases and APIs.

- Web browsing tools help agents access up-to-date public information.

- APIs for search, news, and social media extend knowledge coverage.

Capability Extension Tools

- Address inherent model limitations, such as poor arithmetic skills.

- Simple extensions include calculators, calendars, and unit converters.

- More advanced tools include code interpreters and LaTeX renderers.

- External tools can also make unimodal models multimodal.

- These extensions boost performance with fewer resources than finetuning.

Write Action Tools

- Write actions allow agents to modify their environment.

- Examples include sending emails, executing SQL updates, or initiating transactions.

- These actions enable automation of complete workflows.

- However, write actions increase risks of security breaches or harmful outcomes.

- Proper safeguards are critical to responsible deployment.

Planning in Agents

- Planning determines how agents sequence tool use to complete tasks.

- Agents must break complex tasks into subtasks and reason step by step.

- Planning involves both short-term reasoning and long-term strategy.

- Common methods include chain-of-thought, self-critique, and structured workflows.

- Strong planners help reduce compound errors across multiple steps.

Evaluating Agents

- Agents are more complex to evaluate than static models.

- Evaluations must consider success rates across multi-step tasks.

- Key risks include compounding errors, increased costs, and higher latency.

- Safety evaluation is critical for agents with write actions.

- Human-in-the-loop monitoring may be necessary in early deployments.

References

References

- Huyen, Chip. Designing machine learning systems. O'Reilly Media, Inc., 2022