Deep Explanation of OpenAI Agent SDK (Agent Class)

1. What is this file about?

This defines the core **Agent** in OpenAI’s Agent SDK. An **Agent** is like an AI assistant "brain" that:

- Has **instructions** (system prompt).

- Can use **tools** (functions/APIs).

- Can hand off work to other agents.

- Has **guardrails** (safety checks).

- Produces a structured **output**.

Think of it like:

- Agent = Manager AI.

- Tools = Employees (special functions).

- Handoffs = Asking another AI manager for help.

- Guardrails = Rules/policies that prevent mistakes.

- Model = The actual LLM (GPT, Claude, etc.) that "thinks".

---

2. The Base: AgentBase

Defines name, tools, MCP servers, config.

### get_mcp_tools()

Fetches tools from external MCP servers.

### get_all_tools()

Combines MCP tools + agent’s own tools, checks if enabled.

---

3. The Main Class: Agent

Extends AgentBase with LLM-related features:

- **instructions**: System prompt.
- **prompt**: Dynamic prompt object.
- **handoffs**: Other agents to delegate work to.
- **model**: LLM (gpt-4o, etc.).
- **input_guardrails**: Validate input.
- **output_guardrails**: Validate output.
- **output_type**: Defines final output format.
- **hooks**: Lifecycle callbacks.
- **tool_use_behavior**: Controls tool handling.
- **reset_tool_choice**: Prevents infinite loops.

---

4. __post_init__

Validates all arguments (safety checker).

---

5. clone()

Creates a shallow copy of the agent with modifications.

---

6. as_tool()

Turns the agent into a tool, callable by other agents.

---

7. get_system_prompt()

Returns instructions (string or callable).

---

8. get_prompt()

Prepares final prompt for LLM.

Big Picture Flow

1. User input → validate (guardrails).

2. Prepare prompt → send to LLM.

3. LLM may call tools → agent executes them.

4. Based on behavior, decide final output.

5. Validate output (guardrails).

6. Return structured result.

Analogy

Agent = Project Manager:

- Instructions = Job description.
- Tools = Employees.
- Handoffs = Other managers.
- Guardrails = Company rules.
- Model = Brain (LLM).
- tool_use_behavior = Management strategy.