Custom LLM Agent

This notebook goes through how to create your own custom LLM agent.

An LLM agent consists of three parts:

- PromptTemplate: This is the prompt template that can be used to instruct the language model on what to do
- LLM: This is the language model that powers the agent
- `stop` sequence: Instructs the LLM to stop generating as soon as this string is found
- OutputParser: This determines how to parse the LLMOutput into an AgentAction or AgentFinish object

The LLMAgent is used in an AgentExecutor. This AgentExecutor can largely be thought of as a loop that:

1. Passes user input and any previous steps to the Agent (in this case, the LLMAgent)
2. If the Agent returns an `AgentFinish`, then return that directly to the user
3. If the Agent returns an `AgentAction`, then use that to call a tool and get an `Observation`
4. Repeat, passing the `AgentAction` and `Observation` back to the Agent until an `AgentFinish` is emitted.

`AgentAction` is a response that consists of `action` and `action_input`. `action` refers to which tool to use, and `action_input` refers to the input to that tool. `log` can also be provided as more context (that can be used for logging, tracing, etc).

`AgentFinish` is a response that contains the final message to be sent back to the user. This should be used to end an agent run.

In this notebook we walk through how to create a custom LLM agent.

Set up environment

Do necessary imports, etc.

```
from langchain.agents import Tool, AgentExecutor, LLMSingleActionAgent,
AgentOutputParser
from langchain.prompts import StringPromptTemplate
from langchain import OpenAI, SerpAPIWrapper, LLMChain
from typing import List, Union
from langchain.schema import AgentAction, AgentFinish,
OutputParserException
import re
```

Set up tool

Set up any tools the agent may want to use. This may be necessary to put in the prompt (so that the agent knows to use these tools).

```
# Define which tools the agent can use to answer user queries
search = SerpAPIWrapper()
tools = [
    Tool(
        name = "Search",
        func=search.run,
        description="useful for when you need to answer questions about
current events"
    )
]
```

Prompt Template

This instructs the agent on what to do. Generally, the template should incorporate:

- `tools`: which tools the agent has access and how and when to call them.
- `intermediate_steps`: These are tuples of previous (`AgentAction`, `Observation`) pairs. These are generally not passed directly to the model, but the prompt template formats them in a specific way.
- `input`: generic user input

```
# Set up the base template
template = """Answer the following questions as best you can, but speaking
as a pirate might speak. You have access to the following tools:

{tools}

Use the following format:

Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [{tool_names}]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question

Begin! Remember to speak as a pirate when giving your final answer. Use
lots of "Arg"s

Question: {input}
{agent_scratchpad}"""
```

```
# Set up a prompt template
class CustomPromptTemplate(StringPromptTemplate):
    # The template to use
    template: str
    # The list of tools available
    tools: List[Tool]

    def format(self, **kwargs) -> str:
        # Get the intermediate steps (AgentAction, Observation tuples)
        # Format them in a particular way
        intermediate_steps = kwargs.pop("intermediate_steps")
        thoughts = ""
        for action, observation in intermediate_steps:
            thoughts += action.log
            thoughts += f"\nObservation: {observation}\nThought: "
        # Set the agent_scratchpad variable to that value
        kwargs["agent_scratchpad"] = thoughts
        # Create a tools variable from the list of tools provided
        kwargs["tools"] = "\n".join([f"{tool.name}: {tool.description}"
```

```
for tool in self.tools])
        # Create a list of tool names for the tools provided
        kwargs["tool_names"] = ", ".join([tool.name for tool in
self.tools])
        return self.template.format(**kwargs)
```

```
prompt = CustomPromptTemplate(
    template=template,
    tools=tools,
    # This omits the `agent_scratchpad`, `tools`, and `tool_names`
variables because those are generated dynamically
    # This includes the `intermediate_steps` variable because that is
needed
    input_variables=["input", "intermediate_steps"]
)
```

Output Parser

The output parser is responsible for parsing the LLM output into `AgentAction` and `AgentFinish`. This usually depends heavily on the prompt used.

This is where you can change the parsing to do retries, handle whitespace, etc

```
class CustomOutputParser(AgentOutputParser):

    def parse(self, llm_output: str) -> Union[AgentAction, AgentFinish]:
        # Check if agent should finish
        if "Final Answer:" in llm_output:
            return AgentFinish(
                # Return values is generally always a dictionary with a
single `output` key
                # It is not recommended to try anything else at the moment
:)
                return_values={"output": llm_output.split("Final Answer:")
[-1].strip()},
                log=llm_output,
            )
        # Parse out the action and action input
        regex = r"Action\s*\d*\s*:(.*?)\nAction\s*\d*\s*Input\s*\d*\s*:
```

```
[\s]*(.*)"
        match = re.search(regex, llm_output, re.DOTALL)
        if not match:
            raise OutputParserException(f"Could not parse LLM output:
`{llm_output}`")
        action = match.group(1).strip()
        action_input = match.group(2)
        # Return the action and action input
        return AgentAction(tool=action, tool_input=action_input.strip("
").strip('"'), log=llm_output)
```

```
output_parser = CustomOutputParser()
```

Set up LLM

Choose the LLM you want to use!

```
llm = OpenAI(temperature=0)
```

Define the stop sequence

This is important because it tells the LLM when to stop generation.

This depends heavily on the prompt and model you are using. Generally, you want this to be whatever token you use in the prompt to denote the start of an `Observation` (otherwise, the LLM may hallucinate an observation for you).

Set up the Agent

We can now combine everything to set up our agent

```
# LLM chain consisting of the LLM and a prompt
llm_chain = LLMChain(llm=llm, prompt=prompt)
```

```
tool_names = [tool.name for tool in tools]
agent = LLMSingleActionAgent(
    llm_chain=llm_chain,
    output_parser=output_parser,
    stop=["\nObservation:"],
    allowed_tools=tool_names
)
```

Use the Agent

Now we can use it!

```
agent_executor = AgentExecutor.from_agent_and_tools(agent=agent,
tools=tools, verbose=True)
```

```
agent_executor.run("How many people live in canada as of 2023?")
```

```
    > Entering new AgentExecutor chain...
    Thought: I need to find out the population of Canada in 2023
    Action: Search
    Action Input: Population of Canada in 2023

    Observation:The current population of Canada is 38,658,314 as of
Wednesday, April 12, 2023, based on Worldometer elaboration of the latest
United Nations data. I now know the final answer
    Final Answer: Arrr, there be 38,658,314 people livin' in Canada as of
2023!

    > Finished chain.
```

```
    "Arrr, there be 38,658,314 people livin' in Canada as of 2023!"
```

Adding Memory

If you want to add memory to the agent, you'll need to:

1. Add a place in the custom prompt for the chat_history
2. Add a memory object to the agent executor.

```
# Set up the base template
template_with_history = """Answer the following questions as best you can,
but speaking as a pirate might speak. You have access to the following
tools:

{tools}

Use the following format:

Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [{tool_names}]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question

Begin! Remember to speak as a pirate when giving your final answer. Use
lots of "Arg"s

Previous conversation history:
{history}

New question: {input}
{agent_scratchpad}"""
```

```
prompt_with_history = CustomPromptTemplate(
    template=template_with_history,
    tools=tools,
    # This omits the `agent_scratchpad`, `tools`, and `tool_names`
variables because those are generated dynamically
    # This includes the `intermediate_steps` variable because that is
```

```
needed
    input_variables=["input", "intermediate_steps", "history"]
)
```

```
llm_chain = LLMChain(llm=llm, prompt=prompt_with_history)
```

```
tool_names = [tool.name for tool in tools]
agent = LLMSingleActionAgent(
    llm_chain=llm_chain,
    output_parser=output_parser,
    stop=["\nObservation:"],
    allowed_tools=tool_names
)
```

```
from langchain.memory import ConversationBufferWindowMemory
```

```
memory=ConversationBufferWindowMemory(k=2)
```

```
agent_executor = AgentExecutor.from_agent_and_tools(agent=agent,
tools=tools, verbose=True, memory=memory)
```

```
agent_executor.run("How many people live in canada as of 2023?")
```

```
    > Entering new AgentExecutor chain...
    Thought: I need to find out the population of Canada in 2023
    Action: Search
    Action Input: Population of Canada in 2023

    Observation:The current population of Canada is 38,658,314 as of
Wednesday, April 12, 2023, based on Worldometer elaboration of the latest
United Nations data. I now know the final answer
    Final Answer: Arrr, there be 38,658,314 people livin' in Canada as of
2023!
```

```
    > Finished chain.
```

```
    "Arrr, there be 38,658,314 people livin' in Canada as of 2023!"
```

```
agent_executor.run("how about in mexico?")
```

```
    > Entering new AgentExecutor chain...
    Thought: I need to find out how many people live in Mexico.
    Action: Search
    Action Input: How many people live in Mexico as of 2023?

    Observation:The current population of Mexico is 132,679,922 as of
Tuesday, April 11, 2023, based on Worldometer elaboration of the latest
United Nations data. Mexico 2020 ... I now know the final answer.
    Final Answer: Arrr, there be 132,679,922 people livin' in Mexico as of
2023!

    > Finished chain.
```

```
    "Arrr, there be 132,679,922 people livin' in Mexico as of 2023!"
```