Modules ■ Model I/O ■ Language models ■ LLMs ■ Integrations ■ AzureML Online Endpoint

AzureML Online Endpoint

AzureML is a platform used to build, train, and deploy machine learning models. Users can explore the types of models to deploy in the Model Catalog, which provides Azure Foundation Models and OpenAI Models. Azure Foundation Models include various open-source models and popular Hugging Face models. Users can also import models of their liking into AzureML.

This notebook goes over how to use an LLM hosted on an `AzureML online endpoint`

```
from langchain.llms.azureml_endpoint import AzureMLOnlineEndpoint
```

Set up

To use the wrapper, you must deploy a model on AzureML and obtain the following parameters:

- `endpoint_api_key`: The API key provided by the endpoint
- `endpoint_url`: The REST endpoint url provided by the endpoint
- `deployment_name`: The deployment name of the endpoint

Content Formatter

The `content_formatter` parameter is a handler class for transforming the request and response of an AzureML endpoint to match with required schema. Since there are a wide range of models in the model catalog, each of which may process data differently from one another, a `ContentFormatterBase` class is provided to allow users to transform data to their liking. Additionally, there are three content formatters already provided:

- `OSSContentFormatter`: Formats request and response data for models from the Open Source category in the Model Catalog. Note, that not all models in the Open Source category may follow the same schema
- `DollyContentFormatter`: Formats request and response data for the `dolly-v2-12b` model
- `HFContentFormatter`: Formats request and response data for text-generation Hugging Face models

Below is an example using a summarization model from Hugging Face.

Custom Content Formatter

```
from typing import Dict

from langchain.llms.azureml_endpoint import AzureMLOnlineEndpoint, ContentFormatterBase
import os
import json


class CustomFormatter(ContentFormatterBase):
    content_type = "application/json"
    accepts = "application/json"

    def format_request_payload(self, prompt: str, model_kwargs: Dict) -> bytes:
        input_str = json.dumps(
            {
                "inputs": [prompt],
                "parameters": model_kwargs,
                "options": {"use_cache": False, "wait_for_model": True},
            }
        )
        return str.encode(input_str)

    def format_response_payload(self, output: bytes) -> str:
        response_json = json.loads(output)
        return response_json[0]["summary_text"]


content_formatter = CustomFormatter()

llm = AzureMLOnlineEndpoint(
    endpoint_api_key=os.getenv("BART_ENDPOINT_API_KEY"),
    endpoint_url=os.getenv("BART_ENDPOINT_URL"),
    deployment_name="linydub-bart-large-samsum-3",
    model_kwargs={"temperature": 0.8, "max_new_tokens": 400},
    content_formatter=content_formatter,
)
large_text = """On January 7, 2020, Blockberry Creative announced that HaSeul would not participate in the
promotion for Loona's
next album because of mental health concerns. She was said to be diagnosed with "intermittent anxiety
symptoms" and would be
taking time to focus on her health.[39] On February 5, 2020, Loona released their second EP titled [#] (read
as hash), along
with the title track "So What".[40] Although HaSeul did not appear in the title track, her vocals are
featured on three other
songs on the album, including "365". Once peaked at number 1 on the daily Gaon Retail Album Chart,[41] the EP
then debuted at
number 2 on the weekly Gaon Album Chart. On March 12, 2020, Loona won their first music show trophy with "So
What" on Mnet's
M Countdown.[42]
```

```
On October 19, 2020, Loona released their third EP titled [12:00] (read as midnight),[43] accompanied by its first single
"Why Not?". HaSeul was again not involved in the album, out of her own decision to focus on the recovery of her health.[44]
The EP then became their first album to enter the Billboard 200, debuting at number 112.[45] On November 18, Loona released
the music video for "Star", another song on [12:00].[46] Peaking at number 40, "Star" is Loona's first entry on the Billboard
Mainstream Top 40, making them the second K-pop girl group to enter the chart.[47]

On June 1, 2021, Loona announced that they would be having a comeback on June 28, with their fourth EP, [&] (read as and).
[48] The following day, on June 2, a teaser was posted to Loona's official social media accounts showing twelve sets of eyes,
confirming the return of member HaSeul who had been on hiatus since early 2020.[49] On June 12, group members YeoJin, Kim Lip,
Choerry, and Go Won released the song "Yum-Yum" as a collaboration with Cocomong.[50] On September 8, they released another
collaboration song named "Yummy-Yummy".[51] On June 27, 2021, Loona announced at the end of their special clip that they are
making their Japanese debut on September 15 under Universal Music Japan sublabel EMI Records.[52] On August 27, it was announced
that Loona will release the double A-side single, "Hula Hoop / Star Seed" on September 15, with a physical CD release on October
20.[53] In December, Chuu filed an injunction to suspend her exclusive contract with Blockberry Creative.[54][55]
"""
summarized_text = llm(large_text)
print(summarized_text)
```

```
    HaSeul won her first music show trophy with "So What" on Mnet's M Countdown. Loona released their second EP titled [#] (read as hash] on February 5, 2020. HaSeul did not take part in the promotion of the album
because of mental health issues. On October 19, 2020, they released their third EP called [12:00]. It was
their first album to enter the Billboard 200, debuting at number 112. On June 2, 2021, the group released
their fourth EP called Yummy-Yummy. On August 27, it was announced that they are making their Japanese debut
on September 15 under Universal Music Japan sublabel EMI Records.
```

Dolly with LLMChain

```python
from langchain import PromptTemplate
from langchain.llms.azureml_endpoint import DollyContentFormatter
from langchain.chains import LLMChain

formatter_template = "Write a {word_count} word essay about {topic}."

prompt = PromptTemplate(
    input_variables=["word_count", "topic"], template=formatter_template
)

content_formatter = DollyContentFormatter()

llm = AzureMLOnlineEndpoint(
    endpoint_api_key=os.getenv("DOLLY_ENDPOINT_API_KEY"),
    endpoint_url=os.getenv("DOLLY_ENDPOINT_URL"),
    deployment_name="databricks-dolly-v2-12b-4",
    model_kwargs={"temperature": 0.8, "max_tokens": 300},
    content_formatter=content_formatter,
)

chain = LLMChain(llm=llm, prompt=prompt)
print(chain.run({"word_count": 100, "topic": "how to make friends"}))
```

```
    Many people are willing to talk about themselves; it's others who seem to be stuck up. Try to understand
others where they're coming from. Like minded people can build a tribe together.
```

Serializing an LLM

You can also save and load LLM configurations

```
from langchain.llms.loading import load_llm
from langchain.llms.azureml_endpoint import AzureMLEndpointClient

save_llm = AzureMLOnlineEndpoint(
    deployment_name="databricks-dolly-v2-12b-4",
    model_kwargs={
        "temperature": 0.2,
        "max_tokens": 150,
        "top_p": 0.8,
        "frequency_penalty": 0.32,
        "presence_penalty": 72e-3,
    },
)
save_llm.save("azureml.json")
loaded_llm = load_llm("azureml.json")

print(loaded_llm)
```

```
    AzureMLOnlineEndpoint
    Params: {'deployment_name': 'databricks-dolly-v2-12b-4', 'model_kwargs': {'temperature': 0.2,
'max_tokens': 150, 'top_p': 0.8, 'frequency_penalty': 0.32, 'presence_penalty': 0.072}}
```