


OpenLLM

 **OpenLLM** is an open platform for operating large language models (LLMs) in production. It enables developers to easily run inference with any open-source LLMs, deploy to the cloud or on-premises, and build powerful AI apps.

Installation

Install `openllm` through PyPI

```
pip install openllm
```

Launch OpenLLM server locally

To start an LLM server, use `openllm start` command. For example, to start a dolly-v2 server, run the following command from a terminal:

```
openllm start dolly-v2
```

Wrapper

```
from langchain.llms import OpenLLM

server_url = "http://localhost:3000" # Replace with remote host if you are running on a remote server
llm = OpenLLM(server_url=server_url)
```

Optional: Local LLM Inference

You may also choose to initialize an LLM managed by OpenLLM locally from current process. This is useful for development purpose and allows developers to quickly try out different types of LLMs.

When moving LLM applications to production, we recommend deploying the OpenLLM server separately and access via the `server_url` option demonstrated above.

To load an LLM locally via the LangChain wrapper:

```
from langchain.llms import OpenLLM

llm = OpenLLM(
    model_name="dolly-v2",
    model_id="databricks/dolly-v2-3b",
    temperature=0.94,
    repetition_penalty=1.2,
)
```

Integrate with a LLMChain

```
from langchain import PromptTemplate, LLMChain
```

```
template = "What is a good name for a company that makes {product}?"

prompt = PromptTemplate(template=template, input_variables=["product"])

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

generated = llm_chain.run(product="mechanical keyboard")
print(generated)
```

iLkb