(Experimental)

linkedin.com/in/ariefrahmansyah

MLflow AI Gateway `2.17.2` `OAS 3.1`

/openapi.json

The core deployments API for reverse proxy interface using remote inference endpoints within MLflow

default

POST	/endpoints/completions/invocations	Completions Completions Deployments

POST	/endpoints/chat/invocations	Chat

GET	/health	Health

GET	/api/2.0/endpoints/{endpoint_name}	Get Endpoint

GET	/api/2.0/endpoints/	List Endpoints

GET	/api/2.0/endpoints/limits/{endpoint}	Get Limits

POST	/api/2.0/endpoints/limits/	Set Limits

POST	/v1/chat/completions	Openai Chat Handler

POST	/v1/completions	Openai Completions Handler

POST	/v1/embeddings	Openai Embeddings Handler

Schemas