Extract PDF using Azure OpenAI

Created by David Song, May 26, 2025

Contents

[Use Model gpt-4o 1](#_Toc199184798)

[Source Code: 1](#_Toc199184799)

[Embedding 2](#_Toc199184800)

# Use Model gpt-4o

Required Packages:

langchain

langchain-community

PyPDFLoader

## Source Code:

from langchain\_openai import AzureChatOpenAI  
from langchain.prompts import PromptTemplate  
from langchain.chains import LLMChain  
from langchain\_community.document\_loaders import PyPDFLoader  
import json  
  
  
import json  
from dotenv import load\_dotenv  
import os  
load\_dotenv(r'C:\\*\*\*\Azure\.env')  
print(os.environ)

llm = AzureChatOpenAI(  
 azure\_deployment="gpt-4o", *# or your deployment* api\_version="2025-01-01-preview", *# or your api version*)  
  
  
loader=PyPDFLoader(r'C:\\*\*\*\Azure\2024-dfast-results-20240626.pdf')  
docs=loader.load()  
  
template="""  
you are an intelliegnt bot who can analyze any text with hospital information{doc\_text} ,   
your job is to read and analyse the information and create a json dictionary.  
dictionary has the followinmg key:  
**Table A.31. Wells Fargo & Company**  
output must be json nothing else  
"""  
prompt=PromptTemplate(template=template,input\_variables=["doc\_text"])  
llmchain=LLMChain(llm=llm,prompt=prompt)  
for doc in docs:  
 text=doc.page\_content  
 response=llmchain.invoke({"doc\_text":text})  
 data=response["text"]  
 data=data.replace("json","")  
 data=data.replace("`","")  
 *# data=json.loads(data) # check if data is emtpy* print(data)

# Embedding

You will need to create a text-embedding-3-large model in order to use OpenAI embedding.