

⌕ B I <> ↺ 📄 ” ≡ ≡ — ψ ☺ ☰

I

I

```
# Install dependencies
# Install required packages
!pip install langchain langchain-community sentence-transformers faiss-cpu pypdf2 accelerate transformers
!pip install transformers sacremoses pypdf faiss-cpu
```

```
import os
import torch
import faiss
from transformers import AutoTokenizer, AutoModelForCausalLM
from pypdf import PdfReader
import numpy as np
# --- IMPORTS ---

from langchain_community.embeddings import HuggingFaceEmbeddings
from langchain_community.vectorstores import FAISS
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.schema import Document

from transformers import AutoTokenizer, AutoModelForCausalLM, pipeline
```

```
Requirement already satisfied: langchain in /usr/local/lib/python3.12/dist-packages (0.3.27)
Collecting langchain-community
  Downloading langchain_community-0.3.29-py3-none-any.whl.metadata (2.9 kB)
Requirement already satisfied: sentence-transformers in /usr/local/lib/python3.12/dist-packages (5.1.0)
Collecting faiss-cpu
  Downloading faiss_cpu-1.12.0-cp312-cp312-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl.metadata (5.1 kB)
Collecting pypdf2
  Downloading pypdf2-3.0.1-py3-none-any.whl.metadata (6.8 kB)
Requirement already satisfied: accelerate in /usr/local/lib/python3.12/dist-packages (1.10.1)
Requirement already satisfied: transformers in /usr/local/lib/python3.12/dist-packages (4.56.1)
Collecting pycryptodome
  Downloading pycryptodome-3.23.0-cp37-abi3-manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (3.4 kB)
Requirement already satisfied: langchain-core<1.0.0,>=0.3.72 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (0.3.72)
Requirement already satisfied: langchain-text-splitters<1.0.0,>=0.3.9 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (0.3.9)
Requirement already satisfied: langsmith<0.1.17 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (0.1.17)
Requirement already satisfied: pydantic<3.0.0,>=2.7.4 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (2.7.4)
Requirement already satisfied: SQLAlchemy<3.0.0,>=1.4 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (2.0.36)
Requirement already satisfied: requests<3.0.0,>=2 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (2.32.3)
Requirement already satisfied: PyYAML<=5.3 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (5.3.1)
Collecting requests<3.0.0,>=2 (from langchain-community)
  Downloading requests-2.32.5-py3-none-any.whl.metadata (4.9 kB)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (3.9.5)
Requirement already satisfied: tenacity<=8.4.0,<10,>=8.1.0 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (8.4.0)
Collecting dataclasses-json<0.7,>=0.6.7 (from langchain-community)
  Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Requirement already satisfied: pydantic-settings<3.0.0,>=2.10.1 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (2.10.1)
Requirement already satisfied: httpx-sse<1.0.0,>=0.4.0 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (0.4.0)
Requirement already satisfied: numpy<=1.26.2 in /usr/local/lib/python3.12/dist-packages (from langchain-community) (1.26.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch<=1.11.0 in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (2.4.0)
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (1.5.2)
Requirement already satisfied: scipy in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (1.13.1)
Requirement already satisfied: huggingface-hub<=0.20.0 in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (0.20.0)
Requirement already satisfied: Pillow in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (10.4.0)
Requirement already satisfied: typing_extensions<=4.5.0 in /usr/local/lib/python3.12/dist-packages (from sentence-transformers) (4.5.0)
```

```

Downloading typing_inspect-0.9.0-py3-none-any.whl.metadata (1.5 kB)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.12/dist-packages (from huggingfa
Requirement already satisfied: hf-xet<2.0.0,>=1.1.3 in /usr/local/lib/python3.12/dist-packages (from huggi
Requirement already satisfied: jsonpatch<2.0,>=1.33 in /usr/local/lib/python3.12/dist-packages (from langc
Requirement already satisfied: httpx<1,>=0.23.0 in /usr/local/lib/python3.12/dist-packages (from langsmith
Requirement already satisfied: orjson>=3.9.14 in /usr/local/lib/python3.12/dist-packages (from langsmith>=

```

```

from transformers import AutoTokenizer, AutoModelForCausalLM
import torch

# Load BioGPT
model_name = "microsoft/biogpt"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForCausalLM.from_pretrained(model_name)

def ask_biogpt(question, max_length=150):
    inputs = tokenizer(question, return_tensors="pt")
    with torch.no_grad():
        outputs = model.generate(**inputs, max_length=max_length)
    return tokenizer.decode(outputs[0], skip_special_tokens=True)

# Example test
print(ask_biogpt("Explain the chemical structure of benzene."))

```

```

/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens)
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
warnings.warn(
config.json: 100% 595/595 [00:00<00:00, 15.1kB/s]

vocab.json: 927k/? [00:00<00:00, 2.54MB/s]

merges.txt: 696k/? [00:00<00:00, 9.76MB/s]

pytorch_model.bin: 100% 1.56G/1.56G [00:35<00:00, 78.1MB/s]

model.safetensors: 14% 225M/1.56G [00:05<00:46, 28.9MB/s]
Explain the chemical structure of benzene.

```

```

from google.colab import drive
drive.mount('/content/drive', force_remount=True)

!ls "/content/drive/MyDrive"

```

```

Mounted at /content/drive
ChemData.zip 'Colab Notebooks'

```

```

zip_path = "/content/drive/MyDrive/ChemData.zip" # ZIP in Drive
extract_path = "/content/mydata" # VM folder for extraction

import zipfile, os

if not os.path.exists(extract_path):
    with zipfile.ZipFile(zip_path, 'r') as zip_ref:
        zip_ref.extractall(extract_path)

```

```
print("Files and folders inside extract_path:")
for root, dirs, files in os.walk(extract_path):
    print("In folder:", root)
    print("Files:", files)
```

```
, 'In silico catalysis of hydrogen evolution reaction using transition metal.pdf', 'A metal nonmetal dual-dc
```

```
import os
import glob

# Folder where ZIP was extracted
extract_path = "/content/mydata"

# Recursively find all PDFs, case-insensitive
pdf_files = glob.glob(f"{extract_path}/**/*.pdf", recursive=True) # lowercase
pdf_files += glob.glob(f"{extract_path}/**/*.PDF", recursive=True) # uppercase

print(f"Found {len(pdf_files)} PDFs")
print("First 5 PDFs:", pdf_files[:5])
```

```
Found 27 PDFs
First 5 PDFs: ['/content/mydata/seham catalyst/Novel Graphite Mn304 composites as efficient electrocatalyst:
```

```
from pypdf import PdfReader
import glob
import os

extract_path = "/content/mydata"

# Recursively find all PDFs (case-insensitive)
pdf_files = glob.glob(f"{extract_path}/**/*.pdf", recursive=True)
pdf_files += glob.glob(f"{extract_path}/**/*.PDF", recursive=True)

all_texts = []
for file in pdf_files:
    reader = PdfReader(file)
    text = ""
    for page in reader.pages:
        text += page.extract_text() or ""
    all_texts.append(text)

print(f"Loaded {len(all_texts)} PDF(s) from {extract_path}")
```

```
Loaded 27 PDF(s) from /content/mydata
```

```
from langchain_community.embeddings import HuggingFaceEmbeddings
from langchain_community.vectorstores import FAISS
from langchain.schema import Document
```

```

/tmp/ipython-input-1932372879.py:12: LangChainDeprecationWarning: The class `HuggingFaceEmbeddings` was deprecated in favor of `HuggingFaceTextEmbedder`. Please update your code.
  embeddings = HuggingFaceEmbeddings(model_name="sentence-transformers/all-MiniLM-L6-v2")

modules.json: 100% 349/349 [00:00<00:00, 28.7kB/s]

config_sentence_transformers.json: 100% 116/116 [00:00<00:00, 6.07kB/s]

README.md: 10.5k/? [00:00<00:00, 562kB/s]

sentence_bert_config.json: 100% 53.0/53.0 [00:00<00:00, 2.35kB/s]

config.json: 100% 612/612 [00:00<00:00, 37.5kB/s]

model.safetensors: 100% 90.9M/90.9M [00:02<00:00, 40.0MB/s]

tokenizer_config.json: 100% 350/350 [00:00<00:00, 21.0kB/s]

vocab.txt: 232k/? [00:00<00:00, 403kB/s]

tokenizer.json: 466k/? [00:00<00:00, 2.76MB/s]

special_tokens_map.json: 100% 112/112 [00:00<00:00, 11.1kB/s]

config.json: 100% 190/190 [00:00<00:00, 15.5kB/s]

```

```

docs = db.similarity_search("low-cost materials used as HER catalysts", k=2)
for d in docs:
    print(d.page_content[:500])

```

there is likely to be a boom in HER research. This means that advanced synthesis techniques can provide HER catalysts with better morphological structures or special properties. These properties either expose more active sites or modulate the electronic structure, and high-resolution electron microscopy images and spectra can offer a more plausible explanation for their good performance. The emergence of high-quality HER catalysts requires highly sophisticated synthesis techniques and relies

tremendous advantages, the HER has limitations due to its high electric power consumption and the use of optimized catalysts.

12

However, renewable energy application systems, including the HER, depend highly on the appropriate electrocatalyst type.

11,13

State-of-the-art precious metals and their compounds are active materials for HER and oxygen evolution reactions (Pt for HER,

Received: October 10, 2023

Revised: December 28, 2023

Accepted: December 29, 2023

Published: January 29, 2024

Reviewht

```

from transformers import pipeline

# Create text-generation pipeline
pipe = pipeline(
    "text-generation",
    model=model,
    tokenizer=tokenizer,
    device_map="auto" # uses GPU if available
)

def generate_text(prompt, max_new_tokens=50, temperature=0.3):

```

```
Answer:""

return generate_text(prompt, max_new_tokens=50, temperature=0.3)
```

Device set to use cpu

```
while True:
    user_query = input("You: ")
    if user_query.lower() in ["exit", "quit", "bye"]:
        print("ChemBot: Goodbye! 🍌")
        break
    answer = chatbot(user_query)
    print("ChemBot:", answer)
```

environmental problems as a renewable energy production

Table 2

Hydrogen colors [ 7 , 25 ].

Color Feedstock Primary energy source

Technology Advantages Disadvantages

White Nature Natural Fracking Low-cost; Natural; Non-polluting

Gray Fossil fuels Fossil fuels Reforming Commercialized CO

2

emission

Blue Natural gas Fossil fuels Reforming Not emission of CO

2

CO

2

production

Turquoise Natural gas Renewable

energies

Reforming with carbon

solidification; Pyrolysis

Capturing, storing, and solidification

of carbon

Emission a low value of GHGs

Green Water Renewable

energies

Electrolysis Not emission of GHGs; Recycling High-cost; Loss of energy; Lack of adequate infrastructure; not commercialized

Purple Water Nuclear energy Electrolysis Low emission of carbon High operating temperature

Yellow Water Grid electricity Electrolysis Low emission of carbon Not commercialized; high-cost

Orange Waste

Plastics

Fossil fuels Gasification with carbon

capture

Low-cost of feedstock; less energy

requiring

Not developed; not commercialized

Brown

reforming and gasification, generating GHG emissions in gray, blue, and black/brown hydrogen. Carbon capture, storage, or solidification can reduce emissions, producing turquoise hydrogen. Pyrolysis, a CO

2

-

neutral method, remains underdeveloped but can use biomass for hydrogen production [ 26 ]. Moreover, as displayed in Table 2 , electrol -

ysis is a popular renewable method for producing green, purple, and yellow hydrogen, which will be discussed in subsequent sections. Yellow hydrogen is produced using grid electricity. Purple hydrogen relies on nuclear energy used in countries like China and Russia. Orange

```

if user_query.lower() in ["exit", "quit", "bye"]:
    print("ChemBot: Goodbye! 🍌")
    break
answer = chatbot(user_query)
print("ChemBot:", answer)

```

You: List all the catalysts in the documents that achieved overpotential below 100 mV at 10 mA/cm<sup>2</sup>.

ChemBot: You are ChemAistry, an expert chemistry assistant.

Use the following context from research papers and documents to answer the query. Do not hallucinate or give Context:

potential is an additional crucial metric to accurately assess the catalytic activity of a catalyst on a quantitative level. The most common method researchers use to assess the electrocatalytic activity of catalysts is to measure the overpotential at a standard analytical current of 10 mA/cm

2  
2  
. The overpotential with greater current density (50 or 100 mA/cm  
2  
)

is frequently applied to assess the effectiveness of catalysts on nickel foam. A desirable catalyst frequently combines a lower overpotential with a higher current density.  $\eta = a + b \log j$ , where  $j$  stands for the current density and  $b$  for the Tafel slope, yields the value of the Tafel slope. The lower Tafel slope values suggest a minor overpotential enhancement and a significant growth in current density. Usually, a lower Tafel slope suggests enhanced reaction kinetics and strong electrocatalytic activity. The overpotential is often correlated with the Tafel working and reference electrodes, the wire resistance, and the contact point resistance [22]. The voltage caused by internal resistance can be subtracted from the overpotential data by IR compensation [19]. The value of the overpotential is obtained through Linear Sweep Voltammetry (LSV), where a smaller overpotential at the same current density or a higher current density at the same potential indicates better performance of the catalyst. In practical studies, to facilitate the comparison of different types of catalysts for HERs, the overpotential value at some specific current density (e.g., 10 mA cm<sup>-2</sup> or 100 mA cm<sup>-2</sup>) is used as a criterion.

### 3.2. Tafel Slope

The Tafel slope can be found from the Tafel equation:

$$\eta = a + b \log |j| \quad (9)$$

where  $\eta$ ,  $b$ , and  $j$  represent the HER overpotential, Tafel slope, and current density, respectively. As an important parameter for evaluating the reaction kinetics of the catalyst, Contents lists available at ScienceDirect

Chemical Physics Impact

journal homepage: [www.sciencedirect.com/journal/chemical-physics-impact](http://www.sciencedirect.com/journal/chemical-physics-impact)

<https://doi.org/10.1016/j.chphi.2025.100911>

Received 12 May 2025; Received in revised form 21 June 2025; Accepted 26 June 2025

Chemical Physics Impact 11 (2025) 100911

Available online 28 June 2025

2667-0224/© 2025 The Author. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND 4.0/ ).

overpotential of 35 mV was enough to deliver approximately 10 mA/cm

2  
2  
current density [ 13 ]. Another study, investigated by Minghao Sun

and co-workers, showed that Pt atomic clusters embedded in Nitrogen-doped graphene exhibit higher HER performance and durability than Pt single atoms, mainly due to the interaction with the pyridinic N contents in the graphene [ 14 ]. The development of Pt-free

Question: List all the catalysts in the documents that achieved overpotential below 100 mV at 10 mA/cm<sup>2</sup>.

Answer: The overpotential is an important and important metric to assess the catalytic activity of a catalyst.

You: exit

ChemBot: Goodbye! 🍌

Colab notebook detected. To show errors in colab notebook, set debug=True in launch()

\* Running on public URL: <https://40132ea7fb541e43a4.gradio.live>

This share link expires in 1 week. For free permanent hosting and GPU upgrades, run `gradio deploy` from the

user\_query

Are there any green or environmentally friendly  
synthesis routes mentioned?

Clear

Submit

output

You are ChemAlstry, an expert chemistry assistant.  
Use the following context from research papers and documents to answer the query. Do not hallucinate or give irrelevant information.  
Context:  
ahead and offer future perspectives on how to design functional and stable electrocatalysts for the HER in order to enable efficient hydrogen production by water-splitting electrolysis.  
2. HER  
Currently, hydrogen is mainly used as a raw material in the chemical industry, and its use as a fuel for energy supply has not yet become large-scale. The most economical and common hydrogen generation methods are steam methane reforming