


```
from google.colab import drive  
drive.mount('/content/drive')
```

 Mounted at /content/drive

```
!pip install faiss-cpu sentence-transformers transformers accelerate pdfplumber  
!pip install rouge-score bert-score  
!pip install Rouge
```

```
Collecting faiss-cpu
  Downloading faiss_cpu-1.11.0-cp311-cp311-manylinux_2_28_x86_64.whl.metadata (4.8 kB)
Requirement already satisfied: sentence-transformers in /usr/local/lib/python3.11/dist-packages (4.1.0)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: accelerate in /usr/local/lib/python3.11/dist-packages (1.7.0)
Collecting pdfplumber
  Downloading pdfplumber-0.11.7-py3-none-any.whl.metadata (42 kB)
    42.8/42.8 kB 1.7 MB/s eta 0:00:00
Requirement already satisfied: numpy<3.0,>=1.25.0 in /usr/local/lib/python3.11/dist-packages (from faiss-cpu) (2.0.2)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from faiss-cpu) (24.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (4.67.1)
Requirement already satisfied: torch>=1.11.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (2.
Requirement already satisfied: scikit-learn in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.6
Requirement already satisfied: scipy in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (1.15.3)
Requirement already satisfied: huggingface-hub>=0.20.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transfo
Requirement already satisfied: Pillow in /usr/local/lib/python3.11/dist-packages (from sentence-transformers) (11.2.1)
Requirement already satisfied: typing_extensions>=4.5.0 in /usr/local/lib/python3.11/dist-packages (from sentence-transf
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from accelerate) (5.9.5)
Collecting pdfminer.six==20250506 (from pdfplumber)
  Downloading pdfminer_six-20250506-py3-none-any.whl.metadata (4.2 kB)
Collecting pypdfium2>=4.18.0 (from pdfplumber)
  Downloading pypdfium2-4.30.1-py3-none-manylinux_2_17_x86_64.whl.metadata (48 kB)
    48.2/48.2 kB 3.9 MB/s eta 0:00:00
Requirement already satisfied: charset-normalizer==2.0.0 in /usr/local/lib/python3.11/dist-packages (from pdfminer.six==
Requirement already satisfied: cryptography>=36.0.0 in /usr/local/lib/python3.11/dist-packages (from pdfminer.six==20250
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.20.0
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transfo
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-transform
Collecting nvidia-cuda-nvrtc-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-runtime-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cuda-cupti-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cudnn-cu12==9.1.0.70 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cublas-cu12==12.4.5.8 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cufft-cu12==11.2.1.3 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-curand-cu12==10.3.5.147 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Collecting nvidia-cusolver-cu12==11.6.1.9 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Collecting nvidia-cusparse-cu12==12.3.1.170 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_cusparse_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl.metadata (1.6 kB)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.1
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0
Collecting nvidia-nvjitlink-cu12==12.4.127 (from torch>=1.11.0->sentence-transformers)
  Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl.metadata (1.5 kB)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-tr
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.11.0->sentence-tr
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformer
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformer
Requirement already satisfied: joblib>=1.2.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->sentence-tra
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn->sente
Requirement already satisfied: cffi>=1.12 in /usr/local/lib/python3.11/dist-packages (from cryptography>=36.0.0->pdfmine
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch>=1.11.0->s
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=1.12->cryptography>=36.0
Download faiss_cpu-1.11.0-cp311-cp311-manylinux_2_28_x86_64.whl (31.3 MB)
    31.3/31.3 MB 22.0 MB/s eta 0:00:00
Download pdfplumber-0.11.7-py3-none-any.whl (60 kB)
    60.0/60.0 kB 6.1 MB/s eta 0:00:00
Download pdfminer_six-20250506-py3-none-any.whl (5.6 MB)
    5.6/5.6 MB 76.5 MB/s eta 0:00:00
Download pypdfium2-4.30.1-py3-none-manylinux_2_17_x86_64.whl (2.9 MB)
    2.9/2.9 MB 76.0 MB/s eta 0:00:00
Download nvidia_cublas_cu12-12.4.5.8-py3-none-manylinux2014_x86_64.whl (363.4 MB)
    363.4/363.4 MB 2.7 MB/s eta 0:00:00
Download nvidia_cuda_cupti_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (13.8 MB)
    13.8/13.8 MB 97.1 MB/s eta 0:00:00
Download nvidia_cuda_nvrtc_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (24.6 MB)
    24.6/24.6 MB 23.6 MB/s eta 0:00:00
Download nvidia_cuda_runtime_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (883 kB)
    883.7/883.7 kB 53.2 MB/s eta 0:00:00
Download nvidia_cudnn_cu12-9.1.0.70-py3-none-manylinux2014_x86_64.whl (664.8 MB)
    664.8/664.8 MB 2.0 MB/s eta 0:00:00
Download nvidia_cufft_cu12-11.2.1.3-py3-none-manylinux2014_x86_64.whl (211.5 MB)
    211.5/211.5 MB 6.4 MB/s eta 0:00:00
Download nvidia_curand_cu12-10.3.5.147-py3-none-manylinux2014_x86_64.whl (56.3 MB)
```

```

56.3/56.3 MB 12.5 MB/s eta 0:00:00
Downloading nvidia_cusolver_cu12-11.6.1.9-py3-none-manylinux2014_x86_64.whl (127.9 MB)
127.9/127.9 MB 13.3 MB/s eta 0:00:00
Downloading nvidia_cusparses_cu12-12.3.1.170-py3-none-manylinux2014_x86_64.whl (207.5 MB)
207.5/207.5 MB 6.7 MB/s eta 0:00:00
Downloading nvidia_nvjitlink_cu12-12.4.127-py3-none-manylinux2014_x86_64.whl (21.1 MB)
21.1/21.1 MB 47.0 MB/s eta 0:00:00
Installing collected packages: pypdfium2, nvidia-nvjitlink-cu12, nvidia-curand-cu12, nvidia-cufft-cu12, nvidia-cuda-runt
Attempting uninstall: nvidia-nvjitlink-cu12
Found existing installation: nvidia-nvjitlink-cu12 12.5.82
Uninstalling nvidia-nvjitlink-cu12-12.5.82:
Successfully uninstalled nvidia-nvjitlink-cu12-12.5.82
Attempting uninstall: nvidia-curand-cu12
Found existing installation: nvidia-curand-cu12 10.3.6.82
Uninstalling nvidia-curand-cu12-10.3.6.82:
Successfully uninstalled nvidia-curand-cu12-10.3.6.82
Attempting uninstall: nvidia-cufft-cu12
Found existing installation: nvidia-cufft-cu12 11.2.3.61
Uninstalling nvidia-cufft-cu12-11.2.3.61:
Successfully uninstalled nvidia-cufft-cu12-11.2.3.61
Attempting uninstall: nvidia-cuda-runtime-cu12
Found existing installation: nvidia-cuda-runtime-cu12 12.5.82
Uninstalling nvidia-cuda-runtime-cu12-12.5.82:
Successfully uninstalled nvidia-cuda-runtime-cu12-12.5.82
Attempting uninstall: nvidia-cuda-nvrtc-cu12
Found existing installation: nvidia-cuda-nvrtc-cu12 12.5.82
Uninstalling nvidia-cuda-nvrtc-cu12-12.5.82:
Successfully uninstalled nvidia-cuda-nvrtc-cu12-12.5.82
Attempting uninstall: nvidia-cuda-cupti-cu12
Found existing installation: nvidia-cuda-cupti-cu12 12.5.82
Uninstalling nvidia-cuda-cupti-cu12-12.5.82:
Successfully uninstalled nvidia-cuda-cupti-cu12-12.5.82
Attempting uninstall: nvidia-cublas-cu12
Found existing installation: nvidia-cublas-cu12 12.5.3.2
Uninstalling nvidia-cublas-cu12-12.5.3.2:
Successfully uninstalled nvidia-cublas-cu12-12.5.3.2
Attempting uninstall: nvidia-cusparses-cu12
Found existing installation: nvidia-cusparses-cu12 12.5.1.3
Uninstalling nvidia-cusparses-cu12-12.5.1.3:
Successfully uninstalled nvidia-cusparses-cu12-12.5.1.3
Attempting uninstall: nvidia-cudnn-cu12
Found existing installation: nvidia-cudnn-cu12 9.3.0.75
Uninstalling nvidia-cudnn-cu12-9.3.0.75:
Successfully uninstalled nvidia-cudnn-cu12-9.3.0.75
Attempting uninstall: nvidia-cusolver-cu12
Found existing installation: nvidia-cusolver-cu12 11.6.3.83
Uninstalling nvidia-cusolver-cu12-11.6.3.83:
Successfully uninstalled nvidia-cusolver-cu12-11.6.3.83
Successfully installed faiss-cpu-1.11.0 nvidia-cublas-cu12-12.4.5.8 nvidia-cuda-cupti-cu12-12.4.127 nvidia-cuda-nvrtc-cu
Collecting rouge_score
Downloading rouge_score-0.1.2.tar.gz (17 kB)
Preparing metadata (setup.py) ... done
Collecting bert_score
Downloading bert_score-0.3.13-py3-none-any.whl.metadata (15 kB)
Requirement already satisfied: absl-py in /usr/local/lib/python3.11/dist-packages (from rouge_score) (1.4.0)
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (from rouge_score) (3.9.1)
Requirement already satisfied: numpy in /usr/local/lib/python3.11/dist-packages (from rouge_score) (2.0.2)
Requirement already satisfied: six>=1.14.0 in /usr/local/lib/python3.11/dist-packages (from rouge_score) (1.17.0)
Requirement already satisfied: torch>=1.0.0 in /usr/local/lib/python3.11/dist-packages (from bert_score) (2.6.0+cu124)
Requirement already satisfied: pandas>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from bert_score) (2.2.2)
Requirement already satisfied: transformers>=3.0.0 in /usr/local/lib/python3.11/dist-packages (from bert_score) (4.52.4)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from bert_score) (2.32.3)
Requirement already satisfied: tqdm>=4.31.1 in /usr/local/lib/python3.11/dist-packages (from bert_score) (4.67.1)
Requirement already satisfied: matplotlib in /usr/local/lib/python3.11/dist-packages (from bert_score) (3.10.0)
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.11/dist-packages (from bert_score) (24.2)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.1->be
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.1->bert_score)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas>=1.0.1->bert_score)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score) (3.18
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score) (3.5)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score) (2025.3
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.
Requirement already satisfied: nvidia-cusparses-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->b
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.0.0->bert_score)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch>
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers
Requirement already satisfied: nvml==5.1 in /usr/local/lib/python3.11/dist-packages (from transformers>=3.0.0->bert-sc

```

```
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers>=3.0.0->bert-score)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers>=3.0.0->bert-score)
Requirement already satisfied: safetensors<0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers>=3.0.0->bert-score)
Requirement already satisfied: contourpy>=1.0.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score)
Requirement already satisfied: cycler>=0.10 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score)
Requirement already satisfied: kiwisolver>=1.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score)
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score) (11.2.1)
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib->bert-score)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk->rouge-score) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk->rouge-score) (1.5.1)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->bert-score)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->bert-score) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->bert-score)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->bert-score)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0.0->bert-score)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch>=1.0.0->bert-score)
Downloading bert_score-0.3.13-py3-none-any.whl (61 kB)
```

61.1/61.1 kB 2.8 MB/s eta 0:00:00

Building wheels for collected packages: rouge-score

Building wheel for rouge-score (setup.py) ... done

Created wheel for rouge-score: filename=rouge_score-0.1.2-py3-none-any.whl size=24934 sha256=bc8b4372e6454d0e3ba9dfadb

Stored in directory: /root/.cache/pip/wheels/1e/19/43/8a442dc83660ca25e163e1bd1f89919284ab0d0c1475475148

Successfully built rouge-score

Installing collected packages: rouge-score, bert-score

Successfully installed bert-score-0.3.13 rouge-score-0.1.2

Collecting Rouge

Downloading rouge-1.0.1-py3-none-any.whl.metadata (4.1 kB)

Requirement already satisfied: six in /usr/local/lib/python3.11/dist-packages (from Rouge) (1.17.0)

Downloading rouge-1.0.1-py3-none-any.whl (13 kB)

Installing collected packages: Rouge

Successfully installed Rouge-1.0.1

```
import pdfplumber
from sentence_transformers import SentenceTransformer, CrossEncoder
from transformers import AutoTokenizer, AutoModelForCausalLM
import faiss
import numpy as np
import torch
import os
```

✓ RAG

```
import nltk
nltk.download('punkt')
nltk.download('punkt_tab')
from nltk.tokenize import sent_tokenize
```

```
↻ [nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data] Unzipping tokenizers/punkt_tab.zip.
```

--- Step 1: Read NCERT Text from PDF ---

```
def load_pdf_text_from_directory(directory_path):
    full_text = ""
    # Loop through all files in the directory
    for filename in sorted(os.listdir(directory_path)):
        if filename.lower().endswith(".pdf"):
            file_path = os.path.join(directory_path, filename)
            print(f"Reading: {file_path}")

            # Extract text from the current PDF
            with pdfplumber.open(file_path) as pdf:
                for page in pdf.pages:
                    text = page.extract_text()
                    if text:
                        full_text += text + "\n"
    return full_text
```

--- Step 2: Chunking ---

```
def chunk_by_sentences(text, chunk_size=100):
    sentences = sent_tokenize(text)
    chunks, current, wc = [], [], 0

    for sent in sentences:
        words = sent.split()
        if wc + len(words) > chunk_size:
            chunks.append(" ".join(current))
            current = []
            wc = 0
        current.append(sent)
        wc += len(words)

    if current:
        chunks.append(" ".join(current))

    return chunks
```

```
raw_text = load_pdf_text_from_directory("/content/drive/MyDrive/Dataset/Chemistry_Textbook/")
chunks = chunk_by_sentences(raw_text)
len(chunks)
```

```
↻ Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech101.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech102.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech103.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech104.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech105.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech201.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech202.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech203.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech204.pdf
Reading: /content/drive/MyDrive/Dataset/Chemistry_Textbook/lech205.pdf
1098
```

--- Step 3: Embed Chunks ---

```
embedder = SentenceTransformer("BAAI/bge-base-en-v1.5")
```



```
modules.json: 100% 349/349 [00:00<00:00, 28.7kB/s]
config_sentence_transformers.json: 100% 124/124 [00:00<00:00, 7.92kB/s]
README.md: 100% 94.6k/94.6k [00:00<00:00, 3.60MB/s]
sentence_bert_config.json: 100% 52.0/52.0 [00:00<00:00, 2.36kB/s]
config.json: 100% 777/777 [00:00<00:00, 58.3kB/s]
model.safetensors: 100% 438M/438M [00:06<00:00, 101MB/s]
tokenizer_config.json: 100% 366/366 [00:00<00:00, 21.5kB/s]
vocab.txt: 100% 232k/232k [00:00<00:00, 6.88MB/s]
tokenizer.json: 100% 711k/711k [00:00<00:00, 10.4MB/s]
special_tokens_map.json: 100% 125/125 [00:00<00:00, 11.9kB/s]
config.json: 100% 190/190 [00:00<00:00, 8.58kB/s]
```

```
# --- Step 4: Create FAISS Index ---
```

```
chunk_embeddings = embedder.encode(chunks, normalize_embeddings=True)
dim = chunk_embeddings.shape[1]
index = faiss.IndexFlatIP(dim)
index.add(chunk_embeddings)
faiss.write_index(index, "/content/drive/MyDrive/Dataset/vector_store_faiss/chunk_index.faiss")
```

```
# Load the index
```

```
index = faiss.read_index("/content/drive/MyDrive/Dataset/vector_store_faiss/chunk_index.faiss")
```

✓ Test-Train Data split

```
import pandas as pd
from rouge_score import rouge_scorer
from bert_score import score as bert_score
from sklearn.model_selection import train_test_split
import os

def load_all_csvs_from_folder(folder_path):
    all_dfs = []

    for filename in os.listdir(folder_path):
        if filename.lower().endswith(".csv"):
            file_path = os.path.join(folder_path, filename)
            print(f"Reading: {file_path}")
            df = pd.read_csv(file_path)
            all_dfs.append(df)

    # Combine all into one DataFrame
    combined_df = pd.concat(all_dfs, ignore_index=True)
    return combined_df
```

```
reference_df = load_all_csvs_from_folder("/content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/")
```

```
train_df, test_df = train_test_split(reference_df, test_size=0.2, random_state=42)
```



```
Reading: /content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/chapter_4_d_f_block_final.csv
Reading: /content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/chapter_5_coordination_compounds_final.csv
Reading: /content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/chapter_1_solutions_with_subtopics.csv
Reading: /content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/chapter_2_electrochemistry_with_subtopics.csv
Reading: /content/drive/MyDrive/Dataset/Chemistry_Solutions_CSV_Subtopic_Clean/chapter_3_chemical_kinetics_final.csv
```

```
test_df.head()
```



	Sub Topic	Question	Answer
201	Temperature Dependence of the Rate of a Reaction	The rate constant for a first order reaction i...	It is known that, Hence, the required...
213	Collision Theory of Chemical Reactions	The decomposition of A into product has value ...	From Arrhenius equation, we obtain Also, k1...
138	Abnormal Molar Masses	100 g of liquid A (molar mass 140 g mol ⁻¹) was...	Number of moles of liquid A, = 0.714 mol N...
177	Rate of a Chemical Reaction	For the reaction R → P, the concentration of a...	Average rate of reaction = 6.67 × 10 ⁻⁶ M...
15	Electronic Configurations of the d-Block Elements	Name the oxometal anions of the first series o...	(i) Vanadate, Oxidation state of V is + 5. ...

```
reference_questions = {}
for sub_topic in test_df['Sub Topic'].unique():
```

```
ref_qs = reference_df[reference_df['Sub Topic'] == sub_topic]['Question'].tolist()
reference_questions[sub_topic] = ref_qs
```

✓ Evaluation metrics

```
import torch
from bert_score import score as bert_score
from rouge import Rouge
import nltk
from nltk.translate.bleu_score import sentence_bleu, SmoothingFunction

nltk.download('punkt')
nltk.download('punkt_tab')

# For each generated question, find best match score from references
def compute_bert_scores(gen, refs):
    P_all, R_all, F1_all = [], [], []
    for g in gen:
        P, R, F1 = bert_score([g]*len(refs), refs, lang="en", verbose=False)
        P_all.append(torch.max(P).item())
        R_all.append(torch.max(R).item())
        F1_all.append(torch.max(F1).item())
    return P_all, R_all, F1_all

rouge = Rouge()
def compute_rouge_scores(gen, refs):
    scores = []
    for g in gen:
        # Score against all references, take best
        best_score = max([rouge.get_scores(g, r)[0]["rouge-l"]["f"] for r in refs])
        scores.append(best_score)
    return scores

def compute_self_bleu(sentences, weights=(0.5, 0.5, 0, 0)):
    """
    Compute Self-BLEU score for a list of generated sentences.

    Args:
        sentences (List[str]): List of generated sentences.
        weights (Tuple): BLEU weights (e.g., (0.5, 0.5) for BLEU-2)


    Returns:
        float: Average Self-BLEU score (lower = more diverse)
    """
    scores = []
    smoothie = SmoothingFunction().method1

    for i in range(len(sentences)):
        hypothesis = nltk.word_tokenize(sentences[i])
        references = [nltk.word_tokenize(sent) for j, sent in enumerate(sentences) if j != i]

        if not references:
            continue

        score = sentence_bleu(references, hypothesis, weights=weights, smoothing_function=smoothie)
        scores.append(score)

    return sum(scores) / len(scores) if scores else 0.0
```

 [nltk_data] Downloading package punkt to /root/nltk_data...
 [nltk_data] Package punkt is already up-to-date!
 [nltk_data] Downloading package punkt_tab to /root/nltk_data...
 [nltk_data] Package punkt_tab is already up-to-date!

✓ Models

✓ Baseline Models

✓ RAG + QG LLM (T5)

```
from transformers import T5ForConditionalGeneration, T5Tokenizer

# Load model and tokenizer
# model_name = "valhalla/t5-base-qg-hl"
```



```

# Load tokenizer and model

def generate_t5_questions(model_name, context = "", num_questions = 5) -> list:
    tokenizer = T5Tokenizer.from_pretrained(model_name)
    model = T5ForConditionalGeneration.from_pretrained(model_name)

    input_text = "generate question: " + "context:" + context.strip()
    input_ids = tokenizer.encode(input_text, return_tensors="pt", truncation=True, max_length=512)

    outputs = model.generate(
        input_ids=input_ids,
        #max_length=64,
        num_return_sequences=num_questions,
        do_sample=True,
        top_k=40,
        top_p=0.9,
        temperature=0.8,
        #early_stopping=True
    )
    return [tokenizer.decode(output, skip_special_tokens=True) for output in outputs]

generated_questions = {}
for sub_topic in test_df['Sub Topic'].unique():
    # --- Step 5: RAG Query ---
    query = sub_topic
    query_embedding = embedder.encode([query], normalize_embeddings=True)

    top_k = 5
    scores, indices = index.search(query_embedding, top_k)
    retrieved_chunks = [chunks[i] for i in indices[0]]

    # --- Step 6: Rerank Results ---
    reranker = CrossEncoder("BAAI/bge-reranker-large")
    pairs = [[query, passage] for passage in retrieved_chunks]
    rerank_scores = reranker.predict(pairs)

    ranked_passages = [x for _, x in sorted(zip(rerank_scores, retrieved_chunks), reverse=True)]
    context = "\n".join(ranked_passages)

    # --- Step 7: Use context to generate question ---
    model_name = "iarfmoose/t5-base-question-generator"
    generated_qs = generate_t5_questions(model_name, context)
    for i, q in enumerate(generated_qs, 1):
        print(f"{i}. {q}")

    generated_questions[sub_topic] = generated_qs

```



```

 config.json: 100%                               801/801 [00:00<00:00, 35.7kB/s]

model.safetensors: 100%                           2.24G/2.24G [00:43<00:00, 86.4MB/s]

tokenizer_config.json: 100%                        443/443 [00:00<00:00, 31.8kB/s]

sentencepiece.bpe.model: 100%                     5.07M/5.07M [00:00<00:00, 111MB/s]

tokenizer.json: 100%                               17.1M/17.1M [00:02<00:00, 6.03MB/s]

special_tokens_map.json: 100%                     279/279 [00:00<00:00, 20.6kB/s]

README.md: 100%                                   34.0k/34.0k [00:00<00:00, 1.38MB/s]

tokenizer_config.json: 100%                       25.0/25.0 [00:00<00:00, 2.09kB/s]

spiece.model: 100%                                792k/792k [00:00<00:00, 11.1MB/s]

added_tokens.json: 100%                           39.0/39.0 [00:00<00:00, 3.07kB/s]

special_tokens_map.json: 100%                     121/121 [00:00<00:00, 11.3kB/s]

config.json: 100%                                 1.21k/1.21k [00:00<00:00, 63.1kB/s]
You are using the default legacy behaviour of the <class 'transformers.models.t5.tokenization_t5.T5Tokenizer'>. This is
pytorch_model.bin: 100%                           892M/892M [00:17<00:00, 98.8MB/s]

model.safetensors: 100%                           892M/892M [00:29<00:00, 26.9MB/s]

1. how much heat does actinium generate?
2. What's the chemistry of the actinoids??
3. what type of activity is the chemistry of the actinoids?
4. The chemistry of the actinoids is more complicated than those of the lanthanoids
5. the chemistry of the actinoids is more complex than those of the lanthanoids
1. what does the structure of a coordination compound consist of?
2. Who was the first to formulate his ideas about the structures of coordination compounds?
3. Why did he formulate his theory of coordination compounds?
4. what type of coordination is the theory of?
5. what was his name?
1. which chemical composition is the most common?
2. what's the name of the periodic table?
3. the d-block occupies the large middle section of the periodic table P P P P P e e
4. where is the d-block located?
5. what is the d-block??
1. how much pressure in the reaction is iodine?
2. True
3. True
4. True
5. How does the rate law calculate?
1. How do you determine the molar mass?
2. what is the dt?
3. what is the molar mass calculated on the basis of this DT b?
4. a DT b will be twice the expected value
5. is the result of the DT b?
1. which is the best description of the transition elements?
2. what is the d-block?
3. how many atoms do a coordination entity?
4. which is the same as dbcccccccccccc
5. what is a chemical composition?
1. is the smallest amount of solute in the solution?
2. How many parts of a solution is there in a sample??
3. what is a ssss?? ?
4. what is a ssss?? ?
5. how many parts of the solution is used?
1. what do you think the suffrage is?
2. how many times can a molecule be analyzed?
3. how much do you need to have the ability to do?
4. what is the rate of decrease in concentration of any one of the reactants?
5. what is a ccccccccccccccc
1. what are the simplest of them?
2. what is a coordination compound?
3. what is the function of the s bond?
4. what is the difference between a p bond and a s bond?
5. What is the gist of the term "coordination"??
1. aaaaaaaaaattttttttti
2. how many collisions do they cause?
3. what is the reaction rate?
4. what is the ttttttttiiiiiii
5. what is the effect of the collision?
1. what is the name of the transition element?
2. What does a d-block have in common with a d-block?
3. How many transition elements are known for their catalytic activities?
4. What is the main purpose of the d-block??
5. Why does the d-block have a large number of complex compounds?
1. what do you find?
2. how many times a reaction takes place?
3. how long does the reaction take?
4. how many reactions can be performed at a given temperature?
5. how many molls can be formed in a reaction?
1. what is the reversible work done by a galvanic cell?
2. how do you calculate the emf of the cell?
3. the energy of the reaction is increased by the number of times the Gibbs Energy is increased
4. what is the Gibbs Energy of the reaction?

```

4. what is the Gibbs energy of the reaction?
5. what is the emf of the cell?
1. How many times is the reaction reacted to a cid?
2. is a good determining factor for a first order reaction?
3. how many times does a reaction take a half-life?
4. how many parts are there for a reaction?
5. what is the t s e? ? a t ? b
1. what is the nature of the ligands?
2. What is the s and p of the metal-carbon bond?
3. Why do the metal-carbon bond in metal carbonyls possess both s and p
4. What are the properties of the ligands that are incorporated into the ion?
5. which of these is a p?
1. where is the 'nnnnnnnnnnnn'?
2. what does a ccccccllllllllllllllll
3. what is the formula for the compounds that are named after the atomic order?
4. What is the name of the cccccllll?
5. how many compounds do we need to use to make the chemical names of coordination compounds?
1. how many ki's can you calculate?
2. How many gas in a liquid does it dissolve?
3. How many sssssssssssssssss
4. What is the effect of pressure on the solubility of a gas in a liquid?
5. how many times does an acid solution dissolve?
1. what is the energy of the reaction?
2. what can be the rate of a chemical reaction?
3. What is the kinetics of a reaction?
4. How many times does a chemical reaction double when the temperature changes?
5. what does a reaction rate depend on?
1. How many types of solutions are formed?
2. how many types of solutions are formed?
3. what is the best ad-free system?
4. What is the percentage of a mole mass of a solute?
5. what is the composition of the gas?
1. what happens when a ligand is formed in a complex?
2. Why does the chemical formula of the hydrate isomerism differ from that of the ion
3. what's the difference between a cis and a trans isomerism?
4. what type of isomerism is a typical one?
5. Why does it occur in octahedral Fig. 5.4?
1. What is the difference between the electrodes of the cell?
2. how many of the cell potentials are calculated?
3. what is the voltage difference between the two electrodes of a galvanic cell?
4. How can we measure the cell potential between the two electrodes?
5. what is the difference between the two electrodes of a galvanic cell?
1. What is the role of the chemistry?
2. What are the principles of electrochemical cell production?
3. which electrochemical principles are relevant to the Hydrogen Economy?
4. how does the electrochemical process work?
5. What is the main idea of the hydrogen economy?
1. what is the chemistry of the lanthanoids?
2. is a lot of a matter of time?
3. what is a lanthanoid contraction?
4. What is the difference between a lanthanoid and an anthanoid?
5. is an extremely important feature in the chemistry of the lanthanoids?
1. What does the amount of heat the atoms can withstand in a solution?
2. what is the value of the enthalpy of the solution?
3. does not obey Raoult's law over the entire range of Solutions concentration?
4. what are the most important properties of the solution?
5. what type of solution should n-hexane be used for?
1. what is the name of the paper?
2. What is the chemical composition of the elements of the periodic table?
3. how do the d-blocks function?
4. What is the corresponding number of the number of electrons?
5. how much does it cost to buy a d-block?
1. dddddddddddd dddd
2. how much vapour pressure is produced?
3. what is the %Vvvvvvvv?
4. What is the value of the liquid pressure in the ddd ?
5. what is the dddd?

```
print(generated_questions)

↻ {'The Actinoids': ['how much heat does actinium generate?', "What's the chemistry of the actinoids??", 'what type of act

print(reference_questions)

↻ {'The Actinoids': ['Compare the chemistry of actinoids with that of the lanthanoids with specialreference to: (i) elect

for query, questions in generated_questions.items():
    self_bleu = compute_self_bleu(questions, weights=(0.5, 0.5)) # BLEU-2
    print(f"Self-BLEU (BLEU-2): {self_bleu:.4f}")

P, R, F1 = compute_bert_scores(questions, reference_questions[query])
print(f"BERT-F1: {np.mean(F1):.4f}")

rouge_scores = compute_rouge_scores(questions, reference_questions[query])
print(f"ROUGE-L: {np.mean(rouge_scores):.4f}")
```

12/28

13/28

[illegible]

15/28

✓ RAG + QG LLM (BART)

```
!pip install transformers sentencepiece
```

```

Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: sentencepiece in /usr/local/lib/python3.11/dist-packages (0.2.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: huggingface-hub<1.0,>=0.30.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from transformers) (24.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.11/dist-packages (from transformers) (6.0.2)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.20.1)
Requirement already satisfied: safetensors>=0.4.3 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.5.3)
Requirement already satisfied: tqdm>=4.27 in /usr/local/lib/python3.11/dist-packages (from transformers) (4.67.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0) (2024.10.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0) (4.12.2)
Requirement already satisfied: hf-xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface-hub<1.0,>=0.30.0) (1.2.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.4.0)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (2025.1.1)

```

```
from transformers import BartForConditionalGeneration, BartTokenizer
```

```
bart_model = BartForConditionalGeneration.from_pretrained("voidful/bart-egg-question-generator")
```

```
bart_tokenizer = BartTokenizer.from_pretrained("voidful/bart-egg-question-generator")
```

```

def generate_bart_question(text):
    input_ids = bart_tokenizer.encode(text.strip(), return_tensors="pt", truncation=True, max_length=512)
    outputs = bart_model.generate(
        input_ids=bart_tokenizer.encode(text, return_tensors="pt"),
        do_sample=True,
        num_beams=5,
        top_k=50,
        top_p=0.95,
        num_return_sequences=5,
        temperature=0.7
    )
    return bart_tokenizer.decode(outputs[0], skip_special_tokens=True)

```

Some weights of BartForConditionalGeneration were not initialized from the model checkpoint at voidful/bart-egg-question-generator. You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.

```

generated_questions = {}
for sub_topic in test_df['Sub Topic'].unique():
    # --- Step 5: RAG Query ---
    embedder = SentenceTransformer("BAAI/bge-base-en-v1.5")
    query = sub_topic
    query_embedding = embedder.encode([query], normalize_embeddings=True)

    top_k = 5
    scores, indices = index.search(query_embedding, top_k)
    retrieved_chunks = [chunks[i] for i in indices[0]]

    # --- Step 6: Rerank Results ---
    reranker = CrossEncoder("BAAI/bge-reranker-large")
    pairs = [[query, passage] for passage in retrieved_chunks]
    rerank_scores = reranker.predict(pairs)

    ranked_passages = [x for _, x in sorted(zip(rerank_scores, retrieved_chunks), reverse=True)]
    context = "\n".join(ranked_passages)

    # --- Step 7: Use context to generate question ---

    generated_qs = generate_bart_question(context)
    print(generated_qs)
    #for i, q in enumerate(generated_qs, 1):
    #    print(f"{i}. {q}")

    generated_questions[sub_topic] = generated_qs

```

which which is is lives is is is belongs is (is is to to tou (which which which is wants is is is belongs is is (is ((which which is (to which which is is belongs is is (is ((which which is (to tok which which is is is has is is belongs is (is ((to tok ((what which is is belongs is is is has is ((which to to (ue to

```

for query, questions in generated_questions.items():
    self_bleu = compute_self_bleu(questions, weights=(0.5, 0.5)) # BLEU-2

```

```
print(f"Self-BLEU (BLEU-2): {self_bleu:.4f}")

P, R, F1 = compute_bert_scores(questions, reference_questions[query])
print(f"BERT-F1: {np.mean(F1):.4f}")

rouge_scores = compute_rouge_scores(questions, reference_questions[query])
print(f"ROUGE-L: {np.mean(rouge_scores):.4f}")
```

✓ Base Model with Fine Tuning

✓ Supervised Fine Tuning with Custom NCERT Data

```
train_df.head()
```

↔

	Sub Topic	Question	Answer
84	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...
95	Solubility	Calculate (a)molality (b)molarity and (c)mole ...	(a) Molar mass of KI = 39 + 127 = 166 g mol - ...
137	Abnormal Molar Masses	Henry's law constant for the molality of metha...	Here, p = 760 mm Hg kH = 4.27 × 105 mm Hg ...
211	Collision Theory of Chemical Reactions	The decomposition of hydrocarbon follows the e...	The given equation is k = (4.5 × 1011 s - 1) ...

```
from transformers import T5Tokenizer, T5ForConditionalGeneration, Trainer, TrainingArguments
from datasets import load_dataset, Dataset
import pandas as pd
import re

df = train_df.copy()

# Format input and output
df['input'] = "generate question: " + df['Sub Topic']
df['target'] = df['Question']

# Convert to Dataset
dataset = Dataset.from_pandas(df[['input', 'target']]).train_test_split(test_size=0.1)

# Load tokenizer and model
model_name = "iarfmoose/t5-base-question-generator"
tokenizer = T5Tokenizer.from_pretrained(model_name)
model = T5ForConditionalGeneration.from_pretrained(model_name)

# Tokenization function
def preprocess(examples):
    inputs = tokenizer(examples['input'], max_length=512, truncation=True, padding="max_length")
    targets = tokenizer(examples['target'], max_length=64, truncation=True, padding="max_length")
    inputs['labels'] = targets['input_ids']
    return inputs

tokenized_dataset = dataset.map(preprocess, batched=True, remove_columns=["input", "target"])

# Training configuration
training_args = TrainingArguments(
    output_dir="./vapour_pressure_qg_model",
    eval_strategy="epoch",
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=5,
    save_total_limit=1,
    logging_dir="./logs",
    logging_steps=50,
    save_strategy="epoch",
    load_best_model_at_end=True,
    report_to="none"
)

# Trainer setup
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"],
    tokenizer=tokenizer
)

# Train and Save
```

```

trainer.train()
trainer.save_model("./qg_model")
tokenizer.save_pretrained("./qg_model")

```

↗ Map: 100% 154/154 [00:00<00:00, 747.36 examples/s]

Map: 100% 18/18 [00:00<00:00, 344.73 examples/s]

Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.48.0. You should pass an instance of `CacheTuple` instead. [165/195 1:51:40 < 20:33, 0.02 it/s, Epoch 4.21/5]

Epoch	Training Loss	Validation Loss
1	No log	2.567607
2	4.302500	2.346094
3	2.642300	2.282995
4	2.402300	2.250130

[166/195 1:52:16 < 19:51, 0.02 it/s, Epoch 4.23/5]

Epoch	Training Loss	Validation Loss
1	No log	2.567607
2	4.302500	2.346094
3	2.642300	2.282995
4	2.402300	2.250130

```

generated_questions = {}
for sub_topic in test_df['Sub Topic'].unique():
    # --- Step 5: RAG Query ---
    query = sub_topic
    query_embedding = embedder.encode([query], normalize_embeddings=True)

    top_k = 5
    scores, indices = index.search(query_embedding, top_k)
    retrieved_chunks = [chunks[i] for i in indices[0]]

    # --- Step 6: Rerank Results ---
    reranker = CrossEncoder("BAAI/bge-reranker-large")
    pairs = [[query, passage] for passage in retrieved_chunks]
    rerank_scores = reranker.predict(pairs)

    ranked_passages = [x for _, x in sorted(zip(rerank_scores, retrieved_chunks), reverse=True)]
    context = "\n".join(ranked_passages)

    # --- Step 7: Use context to generate question ---
    model_name = "./qg_model"
    generated_qs = generate_t5_questions(model_name, context)
    for i, q in enumerate(generated_qs, 1):
        print(f"{i}. {q}")

    generated_questions[sub_topic] = generated_qs

```



4. What is the reason why the number of electrons in the solution is greater than the number of electron
 5. Why does the volume of solution increase? – The value of the volume of solution that contains
 1. How much energy does the reaction take? ?
 2. What is the pH of the solution? If the pH of the solution is at 20 °C
 3. What is the reversible rate of a titration?
 4. In this equation, the chemical kinetics of the reaction is the same as that of the reaction
 5. How many electrons do we need to obtain the reversible work of a galvan
 1. How many volts is a volt in a volt? What is the average
 2. What is the chemical formula of the chemical reaction? If there is a chemical reaction of two or
 3. The reaction of a nitric acid is analyzed by a process of the
 4. What is the chemical composition of the reaction?
 5. How is the reaction performed at 1.1 V? if you have the following equations: E
 1. The M-C s bond is formed by the donation of a pair of electrons into
 2. Which is the most important of the two bonds?
 3. What is the smallest molecule in the molten metals?
 4. What are the chemistry of the metal atoms and how does it behave?
 5. The metal-carbon bond in metal carbonyls possesses both s and p character
 1. The chemistry of the lanthanoids is much less known at the present time than is the
 2. What is the kinetics of the shunt chemistry?
 3. What are the effects of a lanthanoid contraction on the chemistry of the
 4. How many lanthanoids are produced?
 5. The londonids have a different chemistry than the lanthanoi
 1. How many dissolved gases are present in the solution?
 2. How does the chemical composition of a solution differ from that of the solvent in the aque
 3. What are the properties of the compounds in the solution? The result of the above equations is that
 4. How many parts of a solution can we describe?
 5. What is the pH of the solution?
 1. How much of the following is known to be the most important mass in the mass of a liquid
 2. How much liquify in water is a kilo of hydrogen?
 3. "

```
for query, questions in generated_questions.items():
    self_bleu = compute_self_bleu(questions, weights=(0.5, 0.5)) # BLEU-2
    print(f"Self-BLEU (BLEU-2): {self_bleu:.4f}")

P, R, F1 = compute_bert_scores(questions, reference_questions[query])
print(f"BERT-F1: {np.mean(F1):.4f}")

rouge_scores = compute_rouge_scores(questions, reference_questions[query])
print(f"ROUGE-L: {np.mean(rouge_scores):.4f}")
```

```
Self-BLEU (BLEU-2): 0.5944
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
BERT-F1: 0.8593
ROUGE-L: 0.2516
Self-BLEU (BLEU-2): 0.4296
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
BERT-F1: 0.8669
ROUGE-L: 0.2599
Self-BLEU (BLEU-2): 0.3882
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
BERT-F1: 0.8370
ROUGE-L: 0.2504
Self-BLEU (BLEU-2): 0.4192
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
```

You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
 BERT-F1: 0.8508
 ROUGE-L: 0.3017
 Self-BLEU (BLEU-2): 0.4875
 Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
 You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
 Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:
 You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inference.
 Some weights of RobertaModel were not initialized from the model checkpoint at roberta-large and are newly initialized:

```
!cp -r ./qg_model /content/drive/MyDrive/Dataset/qg_model
```

✓ Fine Tuning using PEFT

```
pip install peft accelerate transformers datasets
```

```
Requirement already satisfied: accelerate in /usr/local/lib/python3.11/dist-packages (1.7.0)
Requirement already satisfied: transformers in /usr/local/lib/python3.11/dist-packages (4.52.4)
Requirement already satisfied: datasets in /usr/local/lib/python3.11/dist-packages (2.14.4)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from peft) (2.0.2)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.11/dist-packages (from peft) (24.2)
Requirement already satisfied: psutil in /usr/local/lib/python3.11/dist-packages (from peft) (5.9.5)
Requirement already satisfied: pyyaml in /usr/local/lib/python3.11/dist-packages (from peft) (6.0.2)
Requirement already satisfied: torch>=1.13.0 in /usr/local/lib/python3.11/dist-packages (from peft) (2.6.0+cu124)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from peft) (4.67.1)
Requirement already satisfied: safetensors in /usr/local/lib/python3.11/dist-packages (from peft) (0.5.3)
Requirement already satisfied: huggingface_hub>=0.25.0 in /usr/local/lib/python3.11/dist-packages (from peft) (0.33.0)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from transformers) (3.18.0)
Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.11/dist-packages (from transformers) (2024.11)
Requirement already satisfied: requests in /usr/local/lib/python3.11/dist-packages (from transformers) (2.32.3)
Requirement already satisfied: tokenizers<0.22,>=0.21 in /usr/local/lib/python3.11/dist-packages (from transformers) (0.15.1)
Requirement already satisfied: pyarrow>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (18.1.0)
Requirement already satisfied: dill<0.3.8,>=0.3.0 in /usr/local/lib/python3.11/dist-packages (from datasets) (0.3.7)
Requirement already satisfied: pandas in /usr/local/lib/python3.11/dist-packages (from datasets) (2.2.2)
Requirement already satisfied: xxhash in /usr/local/lib/python3.11/dist-packages (from datasets) (3.5.0)
Requirement already satisfied: multiprocessing in /usr/local/lib/python3.11/dist-packages (from datasets) (0.70.15)
Requirement already satisfied: fsspec>=2021.11.1 in /usr/local/lib/python3.11/dist-packages (from fsspec[http]>=2021.11.1)
Requirement already satisfied: aiohttp in /usr/local/lib/python3.11/dist-packages (from datasets) (3.11.15)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->dataset)
Requirement already satisfied: aiosignal>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.3.1)
Requirement already satisfied: attrs>=17.3.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.4.1)
Requirement already satisfied: multidict<7.0,>=4.5 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (6.0.0)
Requirement already satisfied: propcache>=0.2.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (0.3.0)
Requirement already satisfied: yarl<2.0,>=1.17.0 in /usr/local/lib/python3.11/dist-packages (from aiohttp->datasets) (1.17.0)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.11/dist-packages (from huggingface_hf_xet<2.0.0,>=1.1.2 in /usr/local/lib/python3.11/dist-packages (from huggingface_hub>=0.25.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->trans)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests->transformers) (3.10.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->transformer)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests->transformer)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft) (3.5)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft) (3.1.6)
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch)
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cusparse-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0)
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft) (3.2.0)
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch>=1.13.0->peft) (1.13.1)
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch)
Requirement already satisfied: python-dateutil>=2.8.2 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-packages (from pandas->datasets) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas)
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch>=1.13.0->peft)
```

```
!pip install -U bitsandbytes
```

```
Collecting bitsandbytes
  Downloading bitsandbytes-0.46.0-py3-none-manylinux_2_24_x86_64.whl.metadata (10 kB)
Requirement already satisfied: torch<3,>=2.2 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.6.0+cu124)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.0.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3.18.0)
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (4.12.0)
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3.5)
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3.1.6)
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (2025.2)
```

```
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch<3,>
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch<3,>
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3,
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbyte
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbyte
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch<
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from jinja2->torch<3,>=2.2->b
Downloading bitsandbytes-0.46.0-py3-none-manylinux_2_24_x86_64.whl (67.0 MB)
67.0/67.0 MB 9.1 MB/s eta 0:00:00
```

```
Installing collected packages: bitsandbytes
Successfully installed bitsandbytes-0.46.0
```

```
from transformers import T5Tokenizer, T5ForConditionalGeneration, BitsAndBytesConfig
from peft import get_peft_model, LoraConfig, TaskType, prepare_model_for_kbit_training
import bitsandbytes as bnb
print(bnb.__version__)
```

```
⚠ WARNING:bitsandbytes.cextension:The installed version of bitsandbytes was compiled without GPU support. 8-bit optimizers
0.46.0
```

```
from transformers import T5Tokenizer, T5ForConditionalGeneration
from peft import get_peft_model, LoraConfig, TaskType
```

```
model_name = "iarfmoose/t5-base-question-generator"
tokenizer = T5Tokenizer.from_pretrained(model_name)
model = T5ForConditionalGeneration.from_pretrained(model_name)
```

```
peft_config = LoraConfig(
    r=16,
    lora_alpha=32,
    target_modules=["q", "v"], # or other layers if needed
    lora_dropout=0.1,
    bias="none",
    task_type=TaskType.SEQ_2_SEQ_LM
)
```

```
model = get_peft_model(model, peft_config)
```

```
from transformers import T5Tokenizer, T5ForConditionalGeneration, Trainer, TrainingArguments
from datasets import Dataset
from peft import get_peft_model, LoraConfig, TaskType
```

```
# Assume train_df is already defined
df = train_df.copy()
df['input'] = "generate question: " + df['Sub Topic']
df['target'] = df['Question']
```

```
# Convert to Dataset
dataset = Dataset.from_pandas(df[['input', 'target']]).train_test_split(test_size=0.1)
```

```
# Tokenization function
def preprocess(examples):
    inputs = tokenizer(examples['input'], max_length=512, truncation=True, padding="max_length")
    targets = tokenizer(examples['target'], max_length=64, truncation=True, padding="max_length")
    inputs['labels'] = targets['input_ids']
    return inputs
```

```
# Tokenize
tokenized_dataset = dataset.map(preprocess, batched=True, remove_columns=["input", "target"])
```

```
# Training configuration
training_args = TrainingArguments(
    output_dir="/content/drive/MyDrive/Dataset/qg_model_peft",
    eval_strategy="epoch",
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=5,
    save_total_limit=1,
    logging_dir="./logs",
    logging_steps=50,
    save_strategy="epoch",
```



```

load_best_model_at_end=True,
report_to="none",
fp16=True # enable if using GPU with FP16
)

# Trainer setup
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"],
    tokenizer=tokenizer
)

# Train and Save
trainer.train()
model.save_pretrained("/content/drive/MyDrive/Dataset/qg_model_peft")
tokenizer.save_pretrained("/content/drive/MyDrive/Dataset/qg_model_peft")

```

Map: 100% 154/154 [00:01<00:00, 149.12 examples/s]

Map: 100% 18/18 [00:00<00:00, 195.56 examples/s]

/tmp/ipython-input-35-2297635730.py:40: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for
 trainer = Trainer(
 No label_names provided for model class `PeftModelForSeq2SeqLM`. Since `PeftModel` hides base models input arguments, if
 Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.48.0. You should pass an instance of
 [187/195 1:20:50 < 03:29, 0.04 it/s, Epoch 4.77/5]

Epoch	Training Loss	Validation Loss
1	No log	9.611069
2	8.471500	7.272644
3	6.871500	5.738424
4	5.678800	4.808397

[189/195 1:21:40 < 02:37, 0.04 it/s, Epoch 4.82/5]

Epoch	Training Loss	Validation Loss
1	No log	9.611069
2	8.471500	7.272644
3	6.871500	5.738424
4	5.678800	4.808397

Start coding or [generate](#) with AI.

Self Instruct Tuning

```

import pandas as pd

df = train_df.copy() # columns: topic, question
df["prompt"] = "Write a question about: " + df["Sub Topic"] + " Example: " + df["Question"]

df["prompt"] = "Generate 5 questions on " + df["Sub Topic"] + "similar to " + df["Question"]
df["generated_qs"] = ""

df.reset_index(drop=True, inplace=True)
df.head(10)

```



	Sub Topic	Question	Answer	prompt	generated_qs
0	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...	Generate 5 questions on Importance and Applica...	[Why do transition elements commonly display m...
1	Solubility	Calculate (a)molarity (b)molarity and (c)mole ...	(a) Molar mass of KI = 39 + 127 = 166 g mol ⁻¹ ...	Generate 5 questions on Solubilitysimilar to C...	
2	Abnormal Molar Masses	Henry's law constant for the molality of metha...	Here, p = 760 mm Hg kH = 4.27 × 10 ⁵ mm Hg ...	Generate 5 questions on Abnormal Molar Massess...	
3	Collision Theory of Chemical Reactions	The decomposition of hydrocarbon follows the e...	The given equation is k = (4.5 × 10 ¹¹ s ⁻¹) ...	Generate 5 questions on Collision Theory of Ch...	
4	Importance and Applications of Coordination Co...	How many ions are produced from the complex Co...	(iii) The given complex can be written as [Co(...	Generate 5 questions on Importance and Applica...	
5	Galvanic Cells	Write the Nernst equation and emf of the follo...	(i) For the given reaction, the Nernst equatio...	Generate 5 questions on Galvanic Cellssimilar ...	
6	Nernst Equation	If a current of 0.5 ampere flows through a met...	I = 0.5 A t = 2 hours = 2 × 60 × 60 s = 7200 ...	Generate 5 questions on Nernst Equationsimilar...	
7	The Actinoids	The chemistry of the actinoid elements is not ...	Lanthanoids primarily show three oxidation sta...	Generate 5 questions on The Actinoidssimilar t...	
8	The Lanthanoids	Describe the oxidising action of potassium dic...	acts as a very strong oxidising agent in the a...	Generate 5 questions on The Lanthanoidssimilar...	
9	Electrochemical Cell and	How much charge is	(i) Required charge = 3 F	Generate 5 questions on	

```
import openai
from openai import OpenAI
import re
```

```
API_KEY = "your_api_key"
openai.api_key = API_KEY
```

```
for i in range(len(df)):
    response = openai.chat.completions.create(
        model="gpt-4",
        messages=[
            {
                "role": "user",
                "content": [
                    {
                        "type": "text",
                        "text": df['prompt'][i]
                    }
                ]
            }
        ]
    )
```

```
# Use regex to split into questions
questions = re.split(r'\n?\d+\.\s+', response.choices[0].message.content)[1:] # Skip the first empty match
df['generated_qs'][i] = questions
# Print or process the questions
for i, q in enumerate(questions, 1):
    print(df['Sub Topic'][i], f"$ {q.strip()}")
```

 /tmp/ipython-input-64-821369497.py:26: FutureWarning: ChainedAssignmentError: behaviour will change in pandas 3.0!
You are setting values through chained assignment. Currently this works in certain cases, but when using Copy-on-Write (A typical example is when you are setting values in a column of a DataFrame, like:

```
df["col"][row_indexer] = value
```

Use `df.loc[row_indexer, "col"] = values` instead, to perform the assignment in a single step and ensure this keeps upda

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-

```
df['generated_qs'][i] = questions
Solubility $ What are coordination compounds and why are they important in chemical reactions?
Abnormal Molar Masses $ How do coordination compounds play a significant role in biological system? Provide an example.
Collision Theory of Chemical Reactions $ What are some of the significant industrial applications of coordination compou
Importance and Applications of Coordination Compounds $ What is a ligand in a coordination compound and how does it infl
Galvanic Cells $ Explain the concept of color in coordination compounds and how it can be used in qualitative analysis.
/tmp/ipython-input-64-821369497.py:26: FutureWarning: ChainedAssignmentError: behaviour will change in pandas 3.0!
You are setting values through chained assignment. Currently this works in certain cases, but when using Copy-on-Write (
A typical example is when you are setting values in a column of a DataFrame, like:
```

```
df["col"][row_indexer] = value
```

Use `df.loc[row_indexer, "col"] = values` instead, to perform the assignment in a single step and ensure this keeps upda

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-

```

df['generated_qs'][i] = questions
Solubility $ 2) Calculate the (a)molarity (b)molarity and (c)mole fraction of CaCl2 in a solution if the density of the
Abnormal Molar Masses $ 3) Compute the (a)molarity (b)molarity and (c)mole fraction of MgSO4 if the density of a 30% (ma
Collision Theory of Chemical Reactions $ 4) Ascertain the (a)molarity (b)molarity and (c)mole fraction of Na2SO4 in a so
Importance and Applications of Coordination Compounds $ 5) Find out the (a)molarity (b)molarity and (c)mole fraction of
/tmp/ipython-input-64-821369497.py:26: FutureWarning: ChainedAssignmentError: behaviour will change in pandas 3.0!
You are setting values through chained assignment. Currently this works in certain cases, but when using Copy-on-Write (
A typical example is when you are setting values in a column of a DataFrame, like:

df["col"][row_indexer] = value

Use `df.loc[row_indexer, "col"] = values` instead, to perform the assignment in a single step and ensure this keeps upda

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-

df['generated_qs'][i] = questions
Solubility $ What is the formula used to find the molar mass of a substance using Henry's law constant?
Abnormal Molar Masses $ Given that the Henry's law constant for the molality of methane in benzene at 298 K is 4.27 x 10
Collision Theory of Chemical Reactions $ What does abnormal molar masses mean in relation to Henry's law constant and ho
Importance and Applications of Coordination Compounds $ How can the effect of abnormal molar masses be deducted from the
Galvanic Cells $ How would changes in temperature affect the molar mass of methane in benzene, given the Henry's law con
/tmp/ipython-input-64-821369497.py:26: FutureWarning: ChainedAssignmentError: behaviour will change in pandas 3.0!
You are setting values through chained assignment. Currently this works in certain cases, but when using Copy-on-Write (
A typical example is when you are setting values in a column of a DataFrame, like:

df["col"][row_indexer] = value

Use `df.loc[row_indexer, "col"] = values` instead, to perform the assignment in a single step and ensure this keeps upda

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-

df['generated_qs'][i] = questions

df.to_csv("/content/drive/MyDrive/Dataset/generated_questions.csv", index=False)

import pandas as pd
df = pd.read_csv("/content/drive/MyDrive/Dataset/generated_questions.csv")

!pip install bitsandbytes
from transformers import T5Tokenizer, T5ForConditionalGeneration, BitsAndBytesConfig
from peft import get_peft_model, LoraConfig, TaskType, prepare_model_for_kbit_training
import bitsandbytes as bnb
print(bnb.__version__)

Collecting bitsandbytes
  Downloading bitsandbytes-0.46.0-py3-none-manylinux_2_24_x86_64.whl.metadata (10 kB)
Requirement already satisfied: torch<3,>=2.2 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.6.0+cu124
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.11/dist-packages (from bitsandbytes) (2.0.2)
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3
Requirement already satisfied: typing-extensions>=4.10.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2-
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3.1
Requirement already satisfied: Jinja2 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (3.1
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbytes) (202
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.
Requirement already satisfied: nvidia-cublas-cu12==12.4.5.8 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2
Requirement already satisfied: nvidia-cufft-cu12==11.2.1.3 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.
Requirement already satisfied: nvidia-curand-cu12==10.3.5.147 in /usr/local/lib/python3.11/dist-packages (from torch<3,>
Requirement already satisfied: nvidia-cusolver-cu12==11.6.1.9 in /usr/local/lib/python3.11/dist-packages (from torch<3,>
Requirement already satisfied: nvidia-cusparselt-cu12==12.3.1.170 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: nvidia-cusparselt-cu12==0.6.2 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->
Requirement already satisfied: nvidia-nvtx-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2
Requirement already satisfied: nvidia-nvjitlink-cu12==12.4.127 in /usr/local/lib/python3.11/dist-packages (from torch<3
Requirement already satisfied: triton==3.2.0 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbyte
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packages (from torch<3,>=2.2->bitsandbyte
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from sympy==1.13.1->torch<
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-packages (from Jinja2->torch<3,>=2.2->b
Downloading bitsandbytes-0.46.0-py3-none-manylinux_2_24_x86_64.whl (67.0 MB)
67.0/67.0 MB 9.7 MB/s eta 0:00:00
Installing collected packages: bitsandbytes
Successfully installed bitsandbytes-0.46.0
WARNING:bitsandbytes.cextension:The installed version of bitsandbytes was compiled without GPU support. 8-bit optimizers
0.46.0

from transformers import T5Tokenizer, T5ForConditionalGeneration
from peft import get_peft_model, LoraConfig, TaskType

model_name = "iartmoose/t5-base-question-generator"
tokenizer = T5Tokenizer.from_pretrained(model_name)
model = T5ForConditionalGeneration.from_pretrained(model_name)

peft_config = LoraConfig(

```

```

r=16,
lora_alpha=32,
target_modules=["q", "v"], # or other layers if needed
lora_dropout=0.1,
bias="none",
task_type=TaskType.SEQ_2_SEQ_LM
)

model = get_peft_model(model, peft_config)

import ast
df['generated_qs'] = df['generated_qs'].apply(ast.literal_eval)
df['generated_qs'][0]

↳ ['What are coordination compounds and why are they important in chemical reactions?',
'How do coordination compounds play a significant role in biological system? Provide an example.',
'What are some of the significant industrial applications of coordination compounds?',
'What is a ligand in a coordination compound and how does it influence the properties of the compound?',
'Explain the concept of color in coordination compounds and how it can be used in qualitative analysis.']

df['questions_str'] = df['generated_qs'].apply(lambda x: '\n'.join(x))

df['questions_str'][0]

↳ 'What are coordination compounds and why are they important in chemical reactions?\nHow do coordination compounds play
a significant role in biological system? Provide an example.\nWhat are some of the significant industrial applications
of coordination compounds?\nWhat is a ligand in a coordination compound and how does it influence the properties of the
compound?\nExplain the concept of color in coordination compounds and how it can be used in qualitative analysis.'

# Split multiline string into list of questions
df['generated_qs_list'] = df['questions_str'].apply(lambda x: x.strip().split('\n'))

# Explode list into separate rows
df_exploded = df.explode('generated_qs_list').rename(columns={'generated_qs_list': 'qs'})

# Optional: Reset index
df_exploded.reset_index(drop=True, inplace=True)

# View result
print(df_exploded)

↳
      Sub Topic \
0  Importance and Applications of Coordination Co...
1  Importance and Applications of Coordination Co...
2  Importance and Applications of Coordination Co...
3  Importance and Applications of Coordination Co...
4  Importance and Applications of Coordination Co...
..
936  Vapour Pressure of Liquid Solutions
937  Vapour Pressure of Liquid Solutions
938  Vapour Pressure of Liquid Solutions
939  Vapour Pressure of Liquid Solutions
940  Vapour Pressure of Liquid Solutions

      Question \
0  What is meant by the chelate effect? Give an e...
1  What is meant by the chelate effect? Give an e...
2  What is meant by the chelate effect? Give an e...
3  What is meant by the chelate effect? Give an e...
4  What is meant by the chelate effect? Give an e...
..
936  Calculate the osmotic pressure in pascals exer...
937  Calculate the osmotic pressure in pascals exer...
938  Calculate the osmotic pressure in pascals exer...
939  Calculate the osmotic pressure in pascals exer...
940  Calculate the osmotic pressure in pascals exer...

      Answer \
0  When a ligand attaches to the metal ion in a m...
1  When a ligand attaches to the metal ion in a m...
2  When a ligand attaches to the metal ion in a m...
3  When a ligand attaches to the metal ion in a m...
4  When a ligand attaches to the metal ion in a m...
..
936  It is given that: Volume of water, V= 450 mL ...
937  It is given that: Volume of water, V= 450 mL ...
938  It is given that: Volume of water, V= 450 mL ...
939  It is given that: Volume of water, V= 450 mL ...
940  It is given that: Volume of water, V= 450 mL ...

      prompt \
0  Generate 5 questions on Importance and Applica...
1  Generate 5 questions on Importance and Applica...
2  Generate 5 questions on Importance and Applica...

```

```

3   Generate 5 questions on Importance and Applica...
4   Generate 5 questions on Importance and Applica...
..
936 Generate 5 questions on Vapour Pressure of Liq...
937 Generate 5 questions on Vapour Pressure of Liq...
938 Generate 5 questions on Vapour Pressure of Liq...
939 Generate 5 questions on Vapour Pressure of Liq...
940 Generate 5 questions on Vapour Pressure of Liq...

```

```

generated_qs \
0   [What are coordination compounds and why are t...
1   [What are coordination compounds and why are t...
2   [What are coordination compounds and why are t...
3   [What are coordination compounds and why are t...
4   [What are coordination compounds and why are t...

```

```
df_exploded.head()
```

	Sub Topic	Question	Answer	prompt	generated_qs	questions_str	input	target	qs
0	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...	Generate 5 questions on Importance and Applica...	[What are coordination compounds and why are t...	What are coordination compounds and why are th...	Instruction: Generate a question based on the ...	What is meant by the chelate effect? Give an e...	What are coordination compounds and why are th...
1	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...	Generate 5 questions on Importance and Applica...	[What are coordination compounds and why are t...	What are coordination compounds and why are th...	Instruction: Generate a question based on the ...	What is meant by the chelate effect? Give an e...	How do coordination compounds play a significa...
2	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...	Generate 5 questions on Importance and Applica...	[What are coordination compounds and why are t...	What are coordination compounds and why are th...	Instruction: Generate a question based on the ...	What is meant by the chelate effect? Give an e...	What are some of the significant industrial ap...
3	Importance and Applications of Coordination Co...	What is meant by the chelate effect? Give an e...	When a ligand attaches to the metal ion in a m...	Generate 5 questions on Importance and Applica...	[What are coordination compounds and why are t...	What are coordination compounds and why are th...	Instruction: Generate a question based on the ...	What is meant by the chelate effect? Give an e...	What is a ligand in a coordination compound?

```
df_exploded.to_csv("/content/drive/MyDrive/Dataset/generated_exploded_questions.csv", index=False)
```

```

generated_questions = {}
df_exploded['Context'] = ""
for sub_topic in df_exploded['Sub Topic'].unique():
    # --- Step 5: RAG Query ---
    query = sub_topic
    query_embedding = embedder.encode([query], normalize_embeddings=True)

    top_k = 5
    scores, indices = index.search(query_embedding, top_k)
    retrieved_chunks = [chunks[i] for i in indices[0]]

    # --- Step 6: Rerank Results ---
    reranker = CrossEncoder("BAAI/bge-reranker-large")
    pairs = [[query, passage] for passage in retrieved_chunks]
    rerank_scores = reranker.predict(pairs)

    ranked_passages = [x for _, x in sorted(zip(rerank_scores, retrieved_chunks), reverse=True)]
    context = "\n".join(ranked_passages)

    df_exploded.loc[df_exploded['Sub Topic'] == sub_topic, 'Context'] = context

```

config.json: 100%	801/801 [00:00<00:00, 29.4kB/s]
model.safetensors: 100%	2.24G/2.24G [00:46<00:00, 58.8MB/s]
tokenizer_config.json: 100%	443/443 [00:00<00:00, 15.6kB/s]
sentencepiece.bpe.model: 100%	5.07M/5.07M [00:00<00:00, 144MB/s]
tokenizer.json: 100%	17.1M/17.1M [00:00<00:00, 164MB/s]
special_tokens_map.json: 100%	279/279 [00:00<00:00, 8.51kB/s]
README.md: 100%	34.0k/34.0k [00:00<00:00, 778kB/s]

```

df_exploded['input'] = (
    "Instruction: Generate a question based on the topic and examples.\n"

```

```

    "Topic: " + df_exploded['Sub Topic'] + "\n"+
    "Examples:\n" +
    df_exploded['qs']+"\n"+
    "Context" + df_exploded['Context']
)

df_exploded['target'] = df_exploded['Question']

from transformers import T5Tokenizer, T5ForConditionalGeneration, Trainer, TrainingArguments
from datasets import Dataset
from peft import get_peft_model, LoraConfig, TaskType
from transformers import EarlyStoppingCallback
"""
# Assume train_df is already defined
df = train_df.copy()
df['input'] = "generate question: " + df['Sub Topic']
df['target'] = df['Question']
"""
# Convert to Dataset
dataset = Dataset.from_pandas(df_exploded[['input', 'target']]).train_test_split(test_size=0.1)
model_name = "iarfmoose/t5-base-question-generator"
tokenizer = T5Tokenizer.from_pretrained(model_name)
model = T5ForConditionalGeneration.from_pretrained(model_name)

# Tokenization function
def preprocess(examples):
    inputs = tokenizer(examples['input'], max_length=512, truncation=True, padding="max_length")
    targets = tokenizer(examples['target'], max_length=64, truncation=True, padding="max_length")
    inputs['labels'] = targets['input_ids']
    return inputs

# Tokenize
tokenized_dataset = dataset.map(preprocess, batched=True, remove_columns=["input", "target"])

# Training configuration
training_args = TrainingArguments(
    output_dir="/content/drive/MyDrive/Dataset/qg_model_instruction_tuned_peft",
    eval_strategy="epoch",
    per_device_train_batch_size=4,
    per_device_eval_batch_size=4,
    num_train_epochs=20,
    save_total_limit=1,
    logging_dir="./logs",
    logging_steps=50,
    save_strategy="epoch",
    load_best_model_at_end=True,
    report_to="none",
    fp16=True # enable if using GPU with FP16
)

# Trainer setup
trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"],
    tokenizer=tokenizer,
    callbacks=[EarlyStoppingCallback(early_stopping_patience=2)]
)

# Train and Save
trainer.train()
model.save_pretrained("/content/drive/MyDrive/Dataset/qg_model_instruction_tuned_peft")
tokenizer.save_pretrained("/content/drive/MyDrive/Dataset/qg_model_instruction_tuned_peft")

```



Map: 100% 846/846 [00:03<00:00, 218.70 examples/s]

Map: 100% 95/95 [00:00<00:00, 216.21 examples/s]

/tmp/ipython-input-44-704841009.py:44: FutureWarning: `tokenizer` is deprecated and will be removed in version 5.0.0 for
 trainer = Trainer(
 Passing a tuple of `past_key_values` is deprecated and will be removed in Transformers v4.48.0. You should pass an insta
 [650/4240 6:49:34 < 37:49:08, 0.03 it/s, Epoch 3.06/20]

Epoch	Training Loss	Validation Loss
1	1.522600	1.322266
2	1.125400	0.991622
3	0.911100	0.803654

```

from transformers import T5ForConditionalGeneration, T5Tokenizer

checkpoint = '/content/drive/MyDrive/Dataset/qg_model_instruction_tuned_peft/checkpoint-636/'
model = T5ForConditionalGeneration.from_pretrained(checkpoint)
tokenizer = T5Tokenizer.from_pretrained(checkpoint)

def generate_t5_questions(query, context="", num_questions = 5) -> list:
    #tokenizer = T5Tokenizer.from_pretrained(model_name)
    #model = T5ForConditionalGeneration.from_pretrained(model_name)

    input_text = "generate question on: " + query #+ "context:" + context.strip()
    input_ids = tokenizer.encode(input_text, return_tensors="pt", truncation=True, max_length=512)

    outputs = model.generate(
        input_ids=input_ids,
        #max_length=64,
        num_return_sequences=num_questions,
        do_sample=True,
        top_k=40,
        top_p=0.9,
        temperature=0.8,
        #early_stopping=True
    )
    return [tokenizer.decode(output, skip_special_tokens=True) for output in outputs]

def generate_questions_for_test_df(test_df):
    generated_questions = {}
    for sub_topic in test_df['Sub Topic'].unique():
        # --- Step 5: RAG Query ---
        query = sub_topic
        query_embedding = embedder.encode([query], normalize_embeddings=True)

        top_k = 5
        scores, indices = index.search(query_embedding, top_k)
        retrieved_chunks = [chunks[i] for i in indices[0]]

        # --- Step 6: Rerank Results ---
        reranker = CrossEncoder("BAAI/bge-reranker-large")
        pairs = [[query, passage] for passage in retrieved_chunks]
        rerank_scores = reranker.predict(pairs)

        ranked_passages = [x for _, x in sorted(zip(rerank_scores, retrieved_chunks), reverse=True)]
        context = "\n".join(ranked_passages)

        # --- Step 7: Use context to generate question ---
        generated_qs = generate_t5_questions(query, context)
        for i, q in enumerate(generated_qs, 1):
            print(f"{i}. {q}")

        generated_questions[sub_topic] = generated_qs
    return generated_questions

generated_qs = generate_questions_for_test_df(test_df)

```

1. The rate of a reaction is dependent on the temperature of the ice.
 2. Temperature Dependence of the Rate of a Reaction :
 3. The rate of a reaction depends on the following factors: Temperature, Time, and Temperature
 4. The rate of a reaction depends on the following factors: Temperature Dependence of the Rate of
 5. The following data are obtained from the temperature dependence of the rate of a reaction:
 1. What is the difference between a cis- reaction and a cis- reaction
 2. The theory of collisions and reaction in the presence of carbon is based on the following:
 3. Which of the following will be the main thrust of the debate on: Collision Theory of Chemical
 4. What is the theory of collisions?
 5. The Theory of Collision of Chemical Reactions
 1. Why are there so many abnormal masses of molar mass?
 2. What are abnormal molar masses?
 3. What are abnormal Molar Masses?
 4. How many molar masses do you know of that are caused by astro-
 5. A number of abnormal Molar Masses have been identified in the past few months.
 1. The following information is required: (i) Rate of a Chemical Reaction (ii)
 2. What is the effect of aqueous catalysts on organic compounds?
 3. The rate of a reaction is based on the following:
 4. What will be the effects of a sulphurate on aqueous solution?
 5. The rate of a chemical reaction is a function of the following factors: (i) The
 1. What are the configurations of the d-Block elements?
 2. Which Electronic Configurations of the d-Block Elements are forwarded to the d
 3. How many electronic configurations of the d-Block Elements are there?
 4. Generate Question on Electronic Configurations of the d-Block Elements
 5. The d-Block Elements are:
 1. What is the effect of temperature on the amount of liquids pumped into the atmosphere?
 2. The following information is provided: Dose Pressure of Liquid Solutions, Vapour Pressure of
 3. The following are the simulated results of a pressure test:
 4. What is the effect of aqueous solutions on the following: