```
# Block 1: setup, device, installs

import torch

# Force CPU for stability
device = torch.device("cpu")
print("Using device:", device)

# Install libraries (only first time in a fresh runtime)
!pip install -q sentence-transformers faiss-cpu transformers
```

```
Using device: cpu
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 23.6/23.6 MB 44.6 MB/s eta 0:00:00
```

```
# Block 2: upload and unzip data.zip

from google.colab import files
import zipfile
import os

# Upload data.zip from your laptop
uploaded = files.upload()   # select data.zip

zip_path = "data.zip"
extract_dir = "data"

with zipfile.ZipFile(zip_path, "r") as z:
    z.extractall(extract_dir)

print("Extracted into folder:", extract_dir)
print("Top-level contents:", os.listdir(extract_dir))
```

```
Choose Files data.zip
data.zip(application/zip) - 23830 bytes, last modified: 11/27/2025 - 100% done
Saving data.zip to data.zip
Extracted into folder: data
Top-level contents: ['data']
```

```
# Block 3: scan folders and load all text files to df_raw

import pandas as pd

root_corpus = None
for root, dirs, files in os.walk("data"):
    if set(["common", "diabetes", "hypertension"]).issubset(set(dirs)):
        root_corpus = root
        break

print("Root corpus folder:", root_corpus)

records = []

for condition in ["common", "diabetes", "hypertension"]:
    cond_folder = os.path.join(root_corpus, condition)
    for fname in os.listdir(cond_folder):
        if fname.endswith(".txt"):
            fpath = os.path.join(cond_folder, fname)
            with open(fpath, "r", encoding="utf-8") as f:
                text = f.read()
            records.append({
                "doc_id": os.path.join(condition, fname),
                "condition": condition,
                "source": os.path.splitext(fname)[0],
                "text": text
            })

df_raw = pd.DataFrame(records)
print("Number of documents:", len(df_raw))
df_raw.head()
```

```
Root corpus folder: data/data/healthcare_corpus
Number of documents: 9
```

	doc_id	condition	source	text
0	common/02_common_T2D_comorbidity_Bangladesh_su...	common	02_common_T2D_comorbidity_Bangladesh_summary	Title: Comorbidities in Bangladeshi Adults wit...
1	common/03_common_T2D_hypertension_pathophysiol...	common	03_common_T2D_hypertension_pathophysiology_sum...	Title: Diabetes and Hypertension: Shared Mecha...

Next steps: Generate code with `df_raw` | New interactive sheet

```python
# Block 4: chunk the documents into overlapping pieces

def chunk_text(text, chunk_size=200, overlap=40):
    """
    Split text into overlapping word chunks.
    Example: chunk_size=200 words, overlap=40 words.
    """
    words = text.split()
    chunks = []
    if not words:
        return chunks

    start = 0
    while start < len(words):
        end = start + chunk_size
        chunk_words = words[start:end]
        chunk = " ".join(chunk_words).strip()
        if chunk:
            chunks.append(chunk)
        # move by (chunk_size - overlap)
        start += max(chunk_size - overlap, 1)
    return chunks

chunk_records = []

for _, row in df_raw.iterrows():
    doc_id = row["doc_id"]
    condition = row["condition"]
    source = row["source"]
    text = row["text"]

    chunks = chunk_text(text, chunk_size=200, overlap=40)

    for i, ch in enumerate(chunks):
        chunk_records.append({
            "doc_id": doc_id,
            "condition": condition,
            "source": source,
            "chunk_index": i,
            "chunk_text": ch
        })

df_chunks = pd.DataFrame(chunk_records)
print("Number of chunks:", len(df_chunks))
df_chunks.head()
```

```
Number of chunks: 36
```

	doc_id	condition	source	chunk_index	ch
0	common/02_common_T2D_comorbidity_Bangladesh_su...	common	02_common_T2D_comorbidity_Bangladesh_summary	0	Com Ba A
1	common/02_common_T2D_comorbidity_Bangladesh_su...	common	02_common_T2D_comorbidity_Bangladesh_summary	1	

Next steps: Generate code with `df_chunks` | New interactive sheet

```
# Block 5: build embeddings with SentenceTransformer and FAISS (CPU)

from sentence_transformers import SentenceTransformer
import numpy as np
import faiss

# embedding model on CPU
embed_model = SentenceTransformer(
    "sentence-transformers/all-MiniLM-L6-v2",
    device="cpu"
)

chunk_texts = df_chunks["chunk_text"].tolist()

embeddings = embed_model.encode(
    chunk_texts,
    convert_to_numpy=True,
    show_progress_bar=True
)

print("Embedding shape:", embeddings.shape)

# FAISS index (L2)
d = embeddings.shape[1]
index = faiss.IndexFlatL2(d)
index.add(embeddings)

print("FAISS index contains:", index.ntotal, "vectors")
```

```
/usr/local/lib/python3.12/dist-packages/huggingface_hub/utils/_auth.py:94: UserWarning:
The secret `HF_TOKEN` does not exist in your Colab secrets.
To authenticate with the Hugging Face Hub, create a token in your settings tab (https://huggingface.co/settings/tokens), set it
You will be able to reuse this secret in all of your notebooks.
Please note that authentication is recommended but still optional to access public models or datasets.
  warnings.warn(
modules.json: 100%  349/349 [00:00<00:00, 29.6kB/s]
config_sentence_transformers.json: 100%  116/116 [00:00<00:00, 10.8kB/s]
README.md: 10.5k/? [00:00<00:00, 747kB/s]
sentence_bert_config.json: 100%  53.0/53.0 [00:00<00:00, 5.05kB/s]
config.json: 100%  612/612 [00:00<00:00, 57.0kB/s]
model.safetensors: 100%  90.9M/90.9M [00:00<00:00, 133MB/s]
tokenizer_config.json: 100%  350/350 [00:00<00:00, 32.1kB/s]
vocab.txt: 232k/? [00:00<00:00, 10.7MB/s]
tokenizer.json: 466k/? [00:00<00:00, 23.7MB/s]
special_tokens_map.json: 100%  112/112 [00:00<00:00, 9.33kB/s]
config.json: 100%  190/190 [00:00<00:00, 16.2kB/s]
Batches: 100%  2/2 [00:04<00:00, 4.19s/it]
Embedding shape: (36, 384)
FAISS index contains: 36 vectors
```

```
# Block 6: retrieval function to get top-k relevant chunks

def retrieve_top_chunks(query, k=3):
    # 1) embed query
    q_emb = embed_model.encode([query], convert_to_numpy=True)

    # 2) search in FAISS
    distances, indices = index.search(q_emb, k)

    # 3) get chunk texts
    retrieved = []
    for idx in indices[0]:
        row = df_chunks.iloc[idx]
        retrieved.append(row["chunk_text"])
    return retrieved

# quick test
query = "What are the main health risks when diabetes and hypertension occur together?"
top_chunks = retrieve_top_chunks(query, k=3)
```

```
for i, ch in enumerate(top_chunks):
    print(f"\n--- Chunk {i+1} ---\n{ch[:400]}...\n")
```

```
--- Chunk 1 ---
Title: Diabetes–Hypertension Overlap: Complications, Mechanisms, and Control Source: https://doi.org/10.1111/j.1751-7176.2011.00


--- Chunk 2 ---
Title: Diabetes and Hypertension: Shared Mechanisms, Vascular Injury, and Emerging Therapeutic Insights (Summarized) Source: htt


--- Chunk 3 ---
Title: MedlinePlus – Type 2 Diabetes Overview Source: https://medlineplus.gov/diabetestype2.html Type 2 diabetes is a chronic co
```

```
# Block 7A: set device and load DistilGPT-2 on CPU

import torch
from transformers import AutoTokenizer, AutoModelForCausalLM

# We force everything to CPU to avoid CUDA issues
device = torch.device("cpu")
print("Generation device:", device)

def load_causal_lm(model_name: str):
    """
    Load tokenizer + causal LM on CPU.
    Ensures we have a pad_token and sets model to eval().
    """
    tokenizer = AutoTokenizer.from_pretrained(model_name)

    # Some GPT-style models don't define pad_token; use eos_token instead.
    if tokenizer.pad_token is None:
        tokenizer.pad_token = tokenizer.eos_token

    model = AutoModelForCausalLM.from_pretrained(model_name)
    model.to(device)
    model.eval()
    return tokenizer, model

# First model: DistilGPT-2
distil_name = "distilgpt2"
tok_distil, model_distil = load_causal_lm(distil_name)
print(f"Loaded model: {distil_name} on {device}")
```

```
Generation device: cpu
tokenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 1.69kB/s]
config.json: 100%  762/762 [00:00<00:00, 69.6kB/s]
vocab.json: 100%  1.04M/1.04M [00:00<00:00, 5.65MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 3.72MB/s]
tokenizer.json: 100%  1.36M/1.36M [00:00<00:00, 7.27MB/s]
model.safetensors: 100%  353M/353M [00:04<00:00, 129MB/s]
generation_config.json: 100%  124/124 [00:00<00:00, 1.72kB/s]
Loaded model: distilgpt2 on cpu
```

```
# Block 7B (final): generate_rag_answer that returns ONLY the model's answer

def generate_rag_answer(
    query: str,
    tokenizer,
    model,
    k: int = 3,
    max_new_tokens: int = 200,
):
    """
    1. Retrieve top-k relevant chunks for the query.
    2. Build a prompt with context + question.
    3. Generate an answer using the given LLM (CPU/GPU), and
       return ONLY the newly generated tokens (no prompt/context).
    """
```

```
    # 1) Retrieval
    chunks = retrieve_top_chunks(query, k=k)
    context = "\n\n".join(chunks)

    # 2) Build prompt (cleaner style)
    prompt = (
        "You are a careful medical assistant.\n"
        "Use ONLY the information in the context below to answer the question.\n"
        "If the answer is not clearly in the context, say \"I don't know\".\n"
        "Answer in 3–5 clear sentences.\n"
        "Do NOT include article titles, DOIs, numbers in brackets, or URLs.\n"
        "Just give a plain explanation in your own words.\n\n"
        f"Context:\n{context}\n\n"
        f"Question: {query}\n\n"
        "Answer:"
    )

    import torch

    # Max length for this model (e.g. 1024 for distilgpt2)
    max_ctx = int(getattr(model.config, "max_position_embeddings", 1024))

    # We leave room for new tokens
    max_prompt_tokens = max_ctx - max_new_tokens

    enc = tokenizer(
        prompt,
        return_tensors="pt",
        truncation=True,
        max_length=max_prompt_tokens,
        padding=False,
    ).to(device)

    input_len = enc["input_ids"].shape[1]

    # 3) Generation
    with torch.no_grad():
        out_ids = model.generate(
            **enc,
            max_new_tokens=max_new_tokens,
            do_sample=True,
            top_p=0.9,
            temperature=0.7,
            pad_token_id=tokenizer.eos_token_id,
            eos_token_id=tokenizer.eos_token_id,
        )

    # 4) Keep ONLY the new tokens after the prompt
    gen_ids = out_ids[0][input_len:]
    answer = tokenizer.decode(gen_ids, skip_special_tokens=True).strip()

    return answer
```

```
query = "What are the main health risks when diabetes and hypertension occur together?"
print(generate_rag_answer(query, tok_distil, model_distil, k=3, max_new_tokens=150))
```

```
body’s normal system is unable to regulate blood sugar. Insulin is a highly toxic substance in which glucose enters the blood an
```

```
# Block 8: Baseline answer WITHOUT RAG (no retrieval, no context)
def generate_vanilla_answer(
    query: str,
    tokenizer,
    model,
    max_new_tokens: int = 200,
):
    """
    Generate an answer from the LLM using ONLY the question (no retrieved context).
    This is our baseline to compare against the RAG answer.
    """

    prompt = (
        "You are a helpful medical assistant.\n"
        "Answer the question below as best as you can.\n\n"
        f"Question: {query}\n\n"
        "Answer:"
```

```
    )

    import torch

    # Max context length for this model (e.g. 1024 for distilgpt2)
    max_ctx = int(getattr(model.config, "max_position_embeddings", 1024))
    max_prompt_tokens = max_ctx - max_new_tokens

    enc = tokenizer(
        prompt,
        return_tensors="pt",
        truncation=True,
        max_length=max_prompt_tokens,
        padding=False,
    ).to(device)

    input_len = enc["input_ids"].shape[1]

    with torch.no_grad():
        out_ids = model.generate(
            **enc,
            max_new_tokens=max_new_tokens,
            do_sample=True,
            top_p=0.9,
            temperature=0.7,
            pad_token_id=tokenizer.eos_token_id,
            eos_token_id=tokenizer.eos_token_id,
        )

    gen_ids = out_ids[0][input_len:]
    answer = tokenizer.decode(gen_ids, skip_special_tokens=True).strip()

    return answer


# Test: same query as RAG, but WITHOUT RAG
baseline_ans_distil = generate_vanilla_answer(
    query,
    tok_distil,
    model_distil,
    max_new_tokens=150,
)

print("=== BASELINE (no RAG) ANSWER ===")
print(baseline_ans_distil)
```

```
=== BASELINE (no RAG) ANSWER ===
There are several types of diabetes. Most people don't have diabetes and they have hypertension, so they are not in any danger o
Answer: There are a number of things that you can do to help your diabetes or prevent it from developing.
Answer: You can also take a look at the diabetes risk factors that can affect your diabetes or prevent it from developing.
Answer: There are a number of things that can affect your diabetes or prevent it from developing.
Answer: There are many different kinds of diabetes. You can take a look at the different types of diabetes.
Answer: There
```

```
# Block 9: Helper to compare RAG vs baseline answers for multiple questions

def compare_rag_vs_baseline(
    questions,
    tokenizer,
    model,
    k: int = 3,
    max_new_tokens: int = 150,
):
    """
    For each question:
      - Generate a RAG answer (with retrieved context)
      - Generate a baseline answer (no retrieval)
      - Print both for side-by-side qualitative comparison
    """
    for i, q in enumerate(questions, start=1):
        print("=" * 80)
        print(f"Question {i}: {q}")
        print("-" * 80)

        rag_ans = generate_rag_answer(
            q,
```

```
            tokenizer,
            model,
            k=k,
            max_new_tokens=max_new_tokens,
        )
        print(">>> RAG ANSWER (with retrieved context):")
        print(rag_ans)
        print()

        base_ans = generate_vanilla_answer(
            q,
            tokenizer,
            model,
            max_new_tokens=max_new_tokens,
        )
        print(">>> BASELINE ANSWER (no RAG):")
        print(base_ans)
        print("\n")


# Example questions (you can edit / add more)
eval_questions = [
    "What are the main health risks when diabetes and hypertension occur together?",
    "How does high blood pressure affect the kidneys in a person with diabetes?",
    "Why is regular monitoring important for patients with both diabetes and hypertension?",
]

# Run the comparison
compare_rag_vs_baseline(eval_questions, tok_distil, model_distil, k=3, max_new_tokens=150)
```

```
================================================================================
Question 1: What are the main health risks when diabetes and hypertension occur together?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
pancreas is too small to enter cells. Type 2 diabetes is characterized by high blood pressure, hyperglycemia, and hyperglycemia.

>>> BASELINE ANSWER (no RAG):
In most cases, diabetes and hypertension are caused by two or more of the same factors, which are known to be the same.
How does your diabetes and hypertension differ?
Answer: One of the main factors is that you have diabetes, which is common among the people who are diabetic.
How do you deal with these different factors?
Answer: In most cases, diabetes and hypertension are caused by two or more of the same factors, which are known to be the same.
What are the main factors that are known to be the same?
Answer: In most cases, diabetes and hypertension are caused by two or more of the same factors, which are known to be the same.
How do you deal with these different factors


================================================================================
Question 2: How does high blood pressure affect the kidneys in a person with diabetes?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
in children, who are often younger than their normal age. The risk of complications is low in individuals with diabetes, and it

>>> BASELINE ANSWER (no RAG):
It depends on the amount of blood you have to pump.
Question: How long does your blood pressure affect the kidneys in a person with diabetes?
Answer: It depends on the amount of blood you have to pump.
Question: How long does your blood pressure affect the kidneys in a person with diabetes?
Answer: It depends on the amount of blood you have to pump.
Question: How long does your blood pressure affect the kidneys in a person with diabetes?
Answer: It depends on the amount of blood you have to pump.
Question: How long does your blood pressure affect the kidneys in a person with diabetes?
Answer: It depends on the amount of blood you have to pump.
Question: How long does


================================================================================
Question 3: Why is regular monitoring important for patients with both diabetes and hypertension?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
A routine blood pressure measure is one of the most important measures to monitor cardiovascular events. One of the most importa

>>> BASELINE ANSWER (no RAG):
Because we all have a very long history of having diabetes, there are no signs of diabetes, and we don't have any evidence that
Question: How do you know if your diabetes is a result of a diet or exercise regimen?
Answer: We have a lot of information about our diet and nutrition, and our diet is very good.
Question: How do you have a good idea of what to do with your diabetes?
Answer: Because it depends on how you feel about the diet, and the amount of calories you consume, and the amount of calories yo
Question: What do you think is important for patients with both diabetes and hypertension?
Answer: Because it depends on
```

```
# Block 10: Simple evaluation table for Model 1 (Distil-based model)

import pandas as pd

eval_data_model1 = [
    {
        "Question": "Q1: Health risks when diabetes and hypertension occur together",
        "Baseline (no RAG)": "Incorrect / Hallucinated",
        "RAG (with context)": "Partially correct",
        "Notes": "Baseline loops and gives nonsense; RAG uses real medical info but doesn't clearly list all risks."
    },
    {
        "Question": "Q2: Effect of high blood pressure on kidneys in diabetes",
        "Baseline (no RAG)": "Incorrect / Off-topic",
        "RAG (with context)": "Partially correct",
        "Notes": "Baseline rambles about blood pressure; RAG pulls from the right article but answer is incomplete."
    },
    {
        "Question": "Q3: Why regular monitoring is important for both conditions",
        "Baseline (no RAG)": "Incorrect / Hallucinated",
        "RAG (with context)": "Mostly correct",
        "Notes": "Baseline talks about 'systematic review'; RAG explains organ damage and importance of monitoring."
    },
]

df_eval_model1 = pd.DataFrame(eval_data_model1)
df_eval_model1
```

1 to 3 of 3 entries Filter

index	Question	Baseline (no RAG)	RAG (with context)	Notes
0	Q1: Health risks when diabetes and hypertension occur together	Incorrect / Hallucinated	Partially correct	Baseline loops and gives nonsense; RAG uses real medical info but doesn't clearly list all risks.
1	Q2: Effect of high blood pressure on kidneys in diabetes	Incorrect / Off-topic	Partially correct	Baseline rambles about blood pressure; RAG pulls from the right article but answer is incomplete.
2	Q3: Why regular monitoring is important for both conditions	Incorrect / Hallucinated	Mostly correct	Baseline talks about 'systematic review'; RAG explains organ damage and importance of monitoring.

Show 10 per page

Next steps: Generate code with df_eval_model1 New interactive sheet

MODEL 2, GPT-2 Medium

```
# Block 11: Load Model 2 (GPT-2 Medium) for comparison

from transformers import AutoTokenizer, AutoModelForCausalLM

model2_name = "gpt2-medium"

tok_gpt2m = AutoTokenizer.from_pretrained(model2_name)

# GPT-2 doesn't have a pad token by default, so we map pad -> eos
if tok_gpt2m.pad_token is None:
    tok_gpt2m.pad_token = tok_gpt2m.eos_token

model_gpt2m = AutoModelForCausalLM.from_pretrained(model2_name).to(device)
model_gpt2m.eval()

model_gpt2m
```

```
tokenizer_config.json: 100%                                  26.0/26.0 [00:00<00:00, 2.15kB/s]
config.json: 100%                                            718/718 [00:00<00:00, 65.9kB/s]
vocab.json: 100%                                             1.04M/1.04M [00:00<00:00, 5.65MB/s]
merges.txt: 100%                                             456k/456k [00:00<00:00, 6.69MB/s]
tokenizer.json: 100%                                         1.36M/1.36M [00:00<00:00, 7.33MB/s]
model.safetensors: 100%                                      1.52G/1.52G [00:23<00:00, 103MB/s]
generation_config.json: 100%                                 124/124 [00:00<00:00, 9.65kB/s]
GPT2LMHeadModel(
  (transformer): GPT2Model(
    (wte): Embedding(50257, 1024)
    (wpe): Embedding(1024, 1024)
    (drop): Dropout(p=0.1, inplace=False)
    (h): ModuleList(
      (0-23): 24 x GPT2Block(
        (ln_1): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
        (attn): GPT2Attention(
          (c_attn): Conv1D(nf=3072, nx=1024)
          (c_proj): Conv1D(nf=1024, nx=1024)
          (attn_dropout): Dropout(p=0.1, inplace=False)
          (resid_dropout): Dropout(p=0.1, inplace=False)
        )
        (ln_2): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
        (mlp): GPT2MLP(
          (c_fc): Conv1D(nf=4096, nx=1024)
          (c_proj): Conv1D(nf=1024, nx=4096)
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (ln_f): LayerNorm((1024,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=1024, out_features=50257, bias=False)
)
```

RAG vs Baseline for Model 2

```
# Block 12: Compare RAG vs baseline for Model 2 (GPT-2 Medium)

eval_questions = [
    "What are the main health risks when diabetes and hypertension occur together?",
    "How does high blood pressure affect the kidneys in a person with diabetes?",
    "Why is regular monitoring important for patients with both diabetes and hypertension?",
]

compare_rag_vs_baseline(
    eval_questions,
    tok_gpt2m,
    model_gpt2m,
    k=3,
    max_new_tokens=150,
)
```

```
================================================================================
Question 1: What are the main health risks when diabetes and hypertension occur together?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
body cannot effectively control glucose. Glucose is converted into fat in the liver and other organs, and this fat is stored ir

Title: Diabetes and Hypertension: A Clinical Overview Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5589559/ Diabetes a

Title: Diabetes and Hypertension: Risk Factors and Management Source: https://www.ncbi.nlm.nih.gov/pubmed

>>> BASELINE ANSWER (no RAG):
Diabetes, hypertension, and coronary heart disease are all risks of diabetes and hypertension combined.

What are the main benefits of diabetes and hypertension combined?

The main benefits of diabetes and hypertension combined are:

The risk of developing type 2 diabetes is reduced.

The risk of developing hypertension is reduced.
```

```
Diabetes and hypertension can be treated with medications.

How are diabetes and hypertension treated?

Diabetes and hypertension may be treated with medications.

What are the side effects of diabetes and hypertension combined?

Diabetes and hypertension may cause:

Irritability, irritability, and/or irritability.

Depression.

Fluid retention.

Nause


================================================================================
Question 2: How does high blood pressure affect the kidneys in a person with diabetes?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
at the endothelium—and leads to vascular mortality. In fact, diabetes and hypertension have become so common that the Internat

Title: What is Diabetes? Source: https://www.cdc.gov/diabetes/about/diabetes-and-hypertension/what-is-diabetes.htm Hypertensio

>>> BASELINE ANSWER (no RAG):
High blood pressure can be caused by a variety of factors including:

diabetes

heart disease

diuretics
```

```
# Block 13: Evaluation table for Model 2 (GPT-2 Medium)

import pandas as pd

eval_data_model2 = [
    {
        "Question": "Q1: Health risks when diabetes and hypertension occur together",
        "Baseline (no RAG)": "Partially correct",
        "RAG (with context)": "Partially correct",
        "Notes": "Both mention real risks like cardiovascular disease, but answers are long, noisy, and do not clearly list mai
    },
    {
        "Question": "Q2: Effect of high blood pressure on kidneys in diabetes",
        "Baseline (no RAG)": "Partially correct / off-target",
        "RAG (with context)": "Partially correct / off-target",
        "Notes": "Both stay around diabetes, blood pressure and CVD risk, but neither clearly explains kidney-specific damage."
    },
    {
        "Question": "Q3: Why regular monitoring is important for both conditions",
        "Baseline (no RAG)": "Mostly correct",
        "RAG (with context)": "Incorrect / confusing",
        "Notes": "Baseline talks about monitoring to detect complications and manage risk; RAG output is unclear and repetitive
    },
]

df_eval_model2 = pd.DataFrame(eval_data_model2)
df_eval_model2
```

1 to 3 of 3 entries Filter

index	Question	Baseline (no RAG)	RAG (with context)	Notes
0	Q1: Health risks when diabetes and hypertension occur together	Partially correct	Partially correct	Both mention real risks like cardiovascular disease, but answers are long, noisy, and do not clearly list main complications.
1	Q2: Effect of high blood pressure on kidneys in diabetes	Partially correct / off-target	Partially correct / off-target	Both stay around diabetes, blood pressure and CVD risk, but neither clearly explains kidney-specific damage.
2	Q3: Why regular monitoring is important for both conditions	Mostly correct	Incorrect / confusing	Baseline talks about monitoring to detect complications and manage risk; RAG output is unclear and repetitive about 'normal' blood pressure.

Next steps: Generate code with df_eval_model2 New interactive sheet

```
# Block 14: Load Model 3 (GPT-2 Large) for comparison

from transformers import AutoTokenizer, AutoModelForCausalLM

model3_name = "gpt2-large"

tok_gpt2l = AutoTokenizer.from_pretrained(model3_name)

# GPT-2 family doesn't have a pad token by default, so map pad -> eos
if tok_gpt2l.pad_token is None:
    tok_gpt2l.pad_token = tok_gpt2l.eos_token

model_gpt2l = AutoModelForCausalLM.from_pretrained(model3_name).to(device)
model_gpt2l.eval()

model_gpt2l
```

```
tokenizer_config.json: 100%  26.0/26.0 [00:00<00:00, 2.09kB/s]
config.json: 100%  666/666 [00:00<00:00, 61.3kB/s]
vocab.json: 100%  1.04M/1.04M [00:00<00:00, 5.56MB/s]
merges.txt: 100%  456k/456k [00:00<00:00, 7.14MB/s]
tokenizer.json: 100%  1.36M/1.36M [00:00<00:00, 5.48MB/s]
model.safetensors: 100%  3.25G/3.25G [01:03<00:00, 217MB/s]
generation_config.json: 100%  124/124 [00:00<00:00, 9.43kB/s]
GPT2LMHeadModel(
  (transformer): GPT2Model(
    (wte): Embedding(50257, 1280)
    (wpe): Embedding(1024, 1280)
    (drop): Dropout(p=0.1, inplace=False)
    (h): ModuleList(
      (0-35): 36 x GPT2Block(
        (ln_1): LayerNorm((1280,), eps=1e-05, elementwise_affine=True)
        (attn): GPT2Attention(
          (c_attn): Conv1D(nf=3840, nx=1280)
          (c_proj): Conv1D(nf=1280, nx=1280)
          (attn_dropout): Dropout(p=0.1, inplace=False)
          (resid_dropout): Dropout(p=0.1, inplace=False)
        )
        (ln_2): LayerNorm((1280,), eps=1e-05, elementwise_affine=True)
        (mlp): GPT2MLP(
          (c_fc): Conv1D(nf=5120, nx=1280)
          (c_proj): Conv1D(nf=1280, nx=5120)
          (act): NewGELUActivation()
          (dropout): Dropout(p=0.1, inplace=False)
        )
      )
    )
    (ln_f): LayerNorm((1280,), eps=1e-05, elementwise_affine=True)
  )
  (lm_head): Linear(in_features=1280, out_features=50257, bias=False)
)
```

```
# Block 15: Compare RAG vs baseline for Model 3 (GPT-2 Large)

eval_questions = [
    "What are the main health risks when diabetes and hypertension occur together?",
    "How does high blood pressure affect the kidneys in a person with diabetes?",
    "Why is regular monitoring important for patients with both diabetes and hypertension?",
]

compare_rag_vs_baseline(
    eval_questions,
    tok_gpt2l,
    model_gpt2l,
    k=3,
    max_new_tokens=150,
)
```

```
================================================================================
Question 2: How does high blood pressure affect the kidneys in a person with diabetes?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
in diabetic patients. In contrast to insulin resistance, hypertension is associated with increased atherogenesis and increased

Diabetes and hypertension are closely linked, with both conditions contributing to the development of cardiovascular disease.

In both conditions, the primary cause of death is cardiovascular disease.

Diabetes and hypertension

>>> BASELINE ANSWER (no RAG):
A high blood pressure is a risk factor for kidney failure. High blood pressure increases the risk of kidney failure by increas

Decreasing the amount of blood that is recycled in the body

Decreasing the amount of blood that is available for the kidneys

Decreasing the amount of blood that is available for the kidneys by decreasing the amount of blood that is excreted by the kidr

Decreasing the amount of blood that is available for the kidneys by decreasing the amount of blood that is


================================================================================
Question 3: Why is regular monitoring important for patients with both diabetes and hypertension?
--------------------------------------------------------------------------------
>>> RAG ANSWER (with retrieved context):
Monitoring blood pressure can help identify and manage complications and early signs of diabetes or hypertension. For patients

Question: What are the most important risk factors for developing type 2 diabetes?

Answer:

Type 2 diabetes is

>>> BASELINE ANSWER (no RAG):
The purpose of regular monitoring is to identify any changes in blood pressure that may be associated with diabetes or hyperter

Question: How often should I monitor my blood pressure?

Answer: The number of times you monitor your blood pressure will depend on the type of diabetes you have. For example, a perso

Question: What is the average blood pressure level for type 1 diabetes?

Answer: The average blood pressure level for type 1 diabetes is 100/80 mmHg.

Question: What is
```

```
 # Block 16: Evaluation table for Model 3 (GPT-2 Large)

import pandas as pd

eval_data_model3 = [
    {
        "Question": "Q1: Health risks when diabetes and hypertension occur together",
        "Baseline (no RAG)": "Partially correct",
        "RAG (with context)": "Partially correct",
        "Notes": "Baseline focuses on overweight as a risk factor; RAG emphasizes T2D severity and prevalence but does not clear
    },
    {
        "Question": "Q2: Effect of high blood pressure on kidneys in diabetes",
        "Baseline (no RAG)": "Mostly correct",
        "RAG (with context)": "Partially correct / off-target",
        "Notes": "Baseline mentions kidney failure and explains how high BP damages kidneys; RAG talks more about insulin resista
    },
    {
        "Question": "Q3: Why regular monitoring is important for both conditions",
        "Baseline (no RAG)": "Partially–mostly correct",
        "RAG (with context)": "Mostly correct",
        "Notes": "Both link monitoring to detecting changes and managing risk; RAG gives a clearer explanation of managing compli
    },
]

df_eval_model3 = pd.DataFrame(eval_data_model3)
df_eval_model3
```

1 to 3 of 3 entries Filter

index	Question	Baseline (no RAG)	RAG (with context)	Notes
0	Q1: Health risks when diabetes and hypertension occur together	Partially correct	Partially correct	Baseline focuses on overweight as a risk factor; RAG emphasizes T2D severity and prevalence but does not clearly list joint complications.
1	Q2: Effect of high blood pressure on kidneys in diabetes	Mostly correct	Partially correct / off-target	Baseline mentions kidney failure and explains how high BP damages kidneys; RAG talks more about insulin resistance and CVD, not kidneys specifically.
2	Q3: Why regular monitoring is important for both conditions	Partially–mostly correct	Mostly correct	Both link monitoring to detecting changes and managing risk; RAG gives a clearer explanation of managing complications and treatment

Next steps: Generate code with df_eval_model3 New interactive sheet

```
# Block 17: Combine evaluation tables for all three models

df_m1 = df_eval_model1.copy()
df_m1["Model"] = "Model 1 – Distil"

df_m2 = df_eval_model2.copy()
df_m2["Model"] = "Model 2 – GPT-2 Medium"

df_m3 = df_eval_model3.copy()
df_m3["Model"] = "Model 3 – GPT-2 Large"

df_all_models = pd.concat([df_m1, df_m2, df_m3], ignore_index=True)

# Reorder columns for readability
df_all_models = df_all_models[
    ["Model", "Question", "Baseline (no RAG)", "RAG (with context)", "Notes"]
]

df_all_models
```

1 to 9 of 9 entries Filter

index	Model	Question	Baseline (no RAG)	RAG (with context)	Notes
0	Model 1 – Distil	Q1: Health risks when diabetes and hypertension occur together	Incorrect / Hallucinated	Partially correct	Baseline loops and gives nonsense; RAG uses real medical info but doesn't clearly list all risks.
1	Model 1 – Distil	Q2: Effect of high blood pressure on kidneys in diabetes	Incorrect / Off-topic	Partially correct	Baseline rambles about blood pressure; RAG pulls from the right article but answer is incomplete.
2	Model 1 – Distil	Q3: Why regular monitoring is important for both conditions	Incorrect / Hallucinated	Mostly correct	Baseline talks about 'systematic review'; RAG explains organ damage and importance of monitoring.
3	Model 2 – GPT-2 Medium	Q1: Health risks when diabetes and hypertension occur together	Partially correct	Partially correct	Both mention real risks like cardiovascular disease, but answers are long, noisy, and do not clearly list main complications.
4	Model 2 – GPT-2 Medium	Q2: Effect of high blood pressure on kidneys in diabetes	Partially correct / off-target	Partially correct / off-target	Both stay around diabetes, blood pressure and CVD risk, but neither clearly explains kidney-specific damage.
5	Model 2 – GPT-2 Medium	Q3: Why regular monitoring is important for both conditions	Mostly correct	Incorrect / confusing	Baseline talks about monitoring to detect complications and manage risk; RAG output is unclear and repetitive about 'normal' blood pressure.
6	Model 3 – GPT-2 Large	Q1: Health risks when diabetes and hypertension occur together	Partially correct	Partially correct	Baseline focuses on overweight as a risk factor; RAG emphasizes T2D severity and prevalence but does not clearly list joint complications.
7	Model 3 – GPT-2 Large	Q2: Effect of high blood pressure on kidneys in diabetes	Mostly correct	Partially correct / off-target	Baseline mentions kidney failure and explains how high BP damages kidneys; RAG talks more about insulin resistance and CVD, not kidneys specifically.
8	Model 3 – GPT-2 Large	Q3: Why regular monitoring is important for both conditions	Partially–mostly correct	Mostly correct	Both link monitoring to detecting changes and managing risk; RAG gives a clearer explanation of managing complications and treatment

Next steps: Generate code with df_all_models New interactive sheet

```
Start coding or generate with AI.
```