```
import pathway as pw
import pandas as pd
import numpy as np
import faiss
from sentence_transformers import SentenceTransformer
from transformers import pipeline
import wandb
```

```
wandb.init(project="backstory_consistency", name="track_a_run", mode="offline")
print("W&B initialized in offline mode. You can later sync if you want.")
```

```
/usr/local/lib/python3.12/dist-packages/notebook/notebookapp.py:191: SyntaxWarning: invalid escape sequence '\/'
  | |_| | '_ \/ _` / _` |  _/ -_)
Tracking run with wandb version 0.23.1
W&B syncing is set to `offline` in this directory. Run `wandb online` or set WANDB_MODE=online to enable cloud syncing.
Run data is saved locally in /content/wandb/offline-run-20260110_094133-5ind4yqt
wandb: Detected [openai] in use.
wandb: Use W&B Weave for improved LLM call tracing. Install Weave with `pip install weave` then add `import weave` to the top of
wandb: For more information, check out the docs at: https://weave-docs.wandb.ai/
W&B initialized in offline mode. You can later sync if you want.
```

```
with open("The Count of Monte Cristo.txt", "r", encoding="utf-8") as f:
    monte_content = f.read()

with open("In search of the castaways.txt", "r", encoding="utf-8") as f:
    castaways_content = f.read()
```

```
class NovelSchema(pw.Schema):
    book: str
    text: str

monte = pw.debug.table_from_rows(
    rows=[("The Count of Monte Cristo", monte_content)],
    schema=NovelSchema
)

castaways = pw.debug.table_from_rows(
    rows=[("In Search of the Castaways", castaways_content)],
    schema=NovelSchema
)

novels = monte.concat(castaways)
```

```
/usr/local/lib/python3.12/dist-packages/pathway/internals/universe_solver.py:173: UserWarning: Found universe that is always emp
Occurred here:
    Line: monte = pw.debug.table_from_rows(
    File: /tmp/ipython-input-2646755783.py:5
  warnings.warn(
/usr/local/lib/python3.12/dist-packages/pathway/internals/universe_solver.py:173: UserWarning: Found universe that is always emp
Occurred here:
    Line: castaways = pw.debug.table_from_rows(
    File: /tmp/ipython-input-2646755783.py:10
  warnings.warn(
/usr/local/lib/python3.12/dist-packages/pathway/internals/universe_solver.py:173: UserWarning: Found universe that is always emp
Occurred here:
    Line: novels = monte.concat(castaways)
    File: /tmp/ipython-input-2646755783.py:15
  warnings.warn(
```

```
def chunk_text(text, chunk_size=500):
    words = text.split()
    return [" ".join(words[i:i+chunk_size]) for i in range(0, len(words), chunk_size)]

chunks = []
for book, text in [
    ("The Count of Monte Cristo", monte_content),
    ("In Search of the Castaways", castaways_content),
]:
    for c in chunk_text(text):
        chunks.append({"book": book, "chunk": c})
```

```
embed_model = SentenceTransformer("all-MiniLM-L6-v2")
embeddings = embed_model.encode([c["chunk"] for c in chunks], show_progress_bar=True).astype("float32")
```

```
index = faiss.IndexFlatL2(embeddings.shape[1])
index.add(embeddings)
```

```
train_df = pd.read_csv("/content/train.csv")
test_df = pd.read_csv("/content/test (1).csv")
```

```
train_df.head()
```

	id	book_name	char	caption	content	label	true_label
0	46	In Search of the Castaways	Thalcave	NaN	Thalcave's people faded as colonists advanced;...	consistent	1
1	137	The Count of Monte Cristo	Faria	The Origin of His Connection with the Count of...	Suspected again in 1815, he was re-arrested an...	contradict	0
2	74	In Search of the Castaways	Kai-Koumou	NaN	Before each fight he studied the crack-pattern...	consistent	1
3	109	The Count of Monte Cristo	Noirtier	The Complexity of Family and Personal Life	Villefort's drift toward the royalists disappo...	contradict	0
4	104	The Count of Monte Cristo	Noirtier	Involvement and Turning Point in the French Re...	His parents were targeted in a reprisal for su...	consistent	1

Next steps: Generate code with train_df New interactive sheet

```
test_df.head()
```

	id	book_name	char	caption	content
0	95	The Count of Monte Cristo	Noirtier	The Fatal Decision of the Hundred Days	Learning that Villefort meant to denounce him ...
1	136	The Count of Monte Cristo	Faria	Escape and Secret Life	From 1800 onward he lived quietly on a small i...
2	59	In Search of the Castaways	Thalcave	NaN	Posing as a relay-station hand, he slipped cap...
3	60	In Search of the Castaways	Thalcave	NaN	First rescue: in 1852 an avalanche buried a si...
4	124	The Count of Monte Cristo	Faria	Foreshadowing of Relationships	On the Marseille quay he noticed young Caderou...

Next steps: Generate code with test_df New interactive sheet

```
train_df.dropna(inplace=True)
test_df.dropna(inplace=True)
```

```
test_df.head()
```

	id	book_name	char	caption	content
0	95	The Count of Monte Cristo	Noirtier	The Fatal Decision of the Hundred Days	Learning that Villefort meant to denounce him ...
1	136	The Count of Monte Cristo	Faria	Escape and Secret Life	From 1800 onward he lived quietly on a small i...
4	124	The Count of Monte Cristo	Faria	Foreshadowing of Relationships	On the Marseille quay he noticed young Caderou...
5	111	The Count of Monte	[illegible]	Wisdom and Influence in the Post-Revolution	[illegible]

Next steps: Generate code with test_df New interactive sheet

```
def retrieve_evidence(book_name, query, k=5):
    q_emb = embed_model.encode([query]).astype("float32")
    _, idxs = index.search(q_emb, k)

    evidence = []
    for i in idxs[0]:
        if chunks[i]["book"].lower() in book_name.lower():
            evidence.append(chunks[i]["chunk"])
    return evidence
```

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

```
generator = pipeline("text-generation", model="gpt2-medium")  # Free local model

def judge_consistency(backstory, evidence):
    prompt = f"""
Backstory:
{backstory}

Evidence from novel:
{' '.join(evidence)}

Is the backstory consistent with the novel?
Answer only: Consistent or Contradict.
"""
    output = generator(prompt, max_length=50, do_sample=False)
    answer = output[0]['generated_text'].splitlines()[-1].strip()
    if "consistent" in answer.lower():
        return "Consistent"
    else:
        return "Contradict"
```

```
Device set to use cuda:0
```

```
predictions = []

for _, row in test_df.iterrows():
    evidence = retrieve_evidence(row["book_name"], row["content"])
    decision = judge_consistency(row["content"], evidence)
    label = 1 if decision == "Consistent" else 0
    predictions.append(label)
    # Log each row prediction offline to W&B
    wandb.log({"row_id": row["id"], "prediction": decision})

test_df["label"] = predictions
test_df.to_csv("submission.csv", index=False)
```

```
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
Both `max_new_tokens` (=256) and `max_length`(=50) seem to have been set. `max_new_tokens` will take precedence. Please refer t
Setting `pad_token_id` to `eos_token_id`:50256 for open-end generation.
```