1st

Model Explore

```
!pip install -q transformers accelerate bitsandbytes huggingface_hub
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.9/72.9 MB 11.4 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 2.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 26.3 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 30.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 30.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 2.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 14.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 6.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 8.1 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 70.4 MB/s eta
0:00:00
```

```python
from huggingface_hub import notebook_login
from transformers import AutoTokenizer, AutoModelForCausalLM
import torch



notebook_login()
```

```
{"model_id":"448e5450415249bc961764f3d9dbdb07","version_major":2,"vers
ion_minor":0}
```

more model

```python
torch.cuda.is_available()
```

```
True
```

```python
model_id = "meta-llama/Llama-2-7b-hf"
```

```
# Load tokenizer
tokenizer = AutoTokenizer.from_pretrained(model_id)
```

```
/usr/local/lib/python3.11/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 as
secret in your Google Colab and restart your session.
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(
```

```
{"model_id":"3abd6932162b4b10827c7bf6b33f22a0","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"20882e6b0b0e46118dc921c8be6b2104","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"ce6e51efd58f421ba3163c50933a104e","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"1f629bef69f942d5aceddc4160d76bed","version_major":2,"vers
ion_minor":0}
```

```
if tokenizer.pad_token is None:
    tokenizer.pad_token = tokenizer.eos_token

# Load model with 8-bit precision (needs less memory)
model = AutoModelForCausalLM.from_pretrained(
    model_id,
    device_map="auto",
    load_in_8bit=True,  # For lower RAM usage (needs `bitsandbytes`)
    torch_dtype=torch.float16,
)
```

```
{"model_id":"238299276a3e4f16a05884f2d7eb4a8d","version_major":2,"vers
ion_minor":0}
```

```
The `load_in_4bit` and `load_in_8bit` arguments are deprecated and
will be removed in the future versions. Please, pass a
`BitsAndBytesConfig` object in `quantization_config` argument instead.
```

```
{"model_id":"30ff1611ce9b4fe8a1182ac2f0fcd143","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"de6f1c573b614bfda0dc9d531653f7c6","version_major":2,"vers
ion_minor":0}
```

```
{"model_id":"edb9b1ba76614c3ab0b55b93f90852bf","version_major":2,"vers
ion_minor":0}
```

{"model_id":"3318f6e03f00486ba6e6e831c88030c0","version_major":2,"version_minor":0}

{"model_id":"292866987325466a8129cc9467bfc154","version_major":2,"version_minor":0}

{"model_id":"f0fdc03e736c4a5b9f7e9b9595e6fc46","version_major":2,"version_minor":0}

```python
def answer(prompt):
    inputs = tokenizer(prompt, return_tensors="pt").to("cuda")
    # output = model.generate(**inputs, max_new_tokens=200)
    outputs = model.generate(
        **inputs,
        max_new_tokens=150,
        do_sample=True,
        temperature=0.7,
        top_p=0.9,
        top_k=50,
        eos_token_id=tokenizer.eos_token_id)
    # print(tokenizer.decode(outputs[0], skip_special_tokens=True))
    return tokenizer.decode(outputs[0], skip_special_tokens=True)
```

```python
prompt = " I have knee pain which medicine should i take?"
answer(prompt)
```

{"type":"string"}

```python
prompt = "I have headache which medicine should i take?"
answer(prompt)
```

{"type":"string"}

detection

```python
from transformers import  pipeline
import torch
import nltk
```

```python
nltk.download('punkt_tab')
```

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

```
True
```

```python
import json
import pandas as pd
import random
```

```python
# ---------------------------
# Step 3: Load benchmarking data and select 50 random questions
# ---------------------------
with open("benchmarking.json", "r") as f:
    data = json.load(f)

# Select up to 50 random entries for quicker benchmarking
sampled = random.sample(data, min(len(data), 50))

# ---------------------------
# Step 4: Generate model answers and save to CSV
# ---------------------------
records = []
for entry in sampled:
    q = entry["question"]
    std = entry["answer"]
    prompt_text = f"Question: {q}\nAnswer:"  # instruction to the model
    gen = answer(prompt_text)
    records.append({
        "question": q,
        "model_answer": gen.strip(),
        "standard_answer": std
    })

# Save results for later detection
df = pd.DataFrame(records)
df.to_csv("model_vs_standard.csv", index=False)
print("Saved model_vs_standard.csv")
```

```
Saved model_vs_standard.csv
```

```python
# Setup NLI verification pipeline

nli = pipeline(
    "text-classification",
    model="ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli"
)

# nltk.download('punkt')  # for sentence tokenization
```

{"model_id":"91578771e67748de858a6727e905cae6","version_major":2,"version_minor":0}

{"model_id":"89bc0e98e49d4b30bcfd5e2d288effbd","version_major":2,"version_minor":0}

```
Some weights of the model checkpoint at ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli were not used when initializing
RobertaForSequenceClassification: ['roberta.pooler.dense.bias',
'roberta.pooler.dense.weight']
- This IS expected if you are initializing
RobertaForSequenceClassification from the checkpoint of a model
trained on another task or with another architecture (e.g.
initializing a BertForSequenceClassification model from a
BertForPreTraining model).
- This IS NOT expected if you are initializing
RobertaForSequenceClassification from the checkpoint of a model that
you expect to be exactly identical (initializing a
BertForSequenceClassification model from a
BertForSequenceClassification model).
```

```
{"model_id":"2b510c17b2c548ecaa611d4d939040da","version_major":2,"version_minor":0}

{"model_id":"03e25a829bfd47a8acaf8cab9f07a426","version_major":2,"version_minor":0}

{"model_id":"40528c6759604477904faee90bb09b27","version_major":2,"version_minor":0}

{"model_id":"b9748fbbbe2a44aabff0cab62065ff84","version_major":2,"version_minor":0}

{"model_id":"d772919668644df7a935f5326671f3e1","version_major":2,"version_minor":0}
```

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

```python
# ---------------------------
# Step 6: Define detection function using standard answers as context
# ---------------------------
def detect_hallucinations(answer: str, context: str) -> list:
    """
    Splits the generated answer into sentences and verifies each
    against the standard answer (context) using NLI.
    Returns a list of {sentence, label, score}
    """
    sentences = nltk.sent_tokenize(answer)
    results = []
    for sent in sentences:
        nli_input = context + " [SEP] " + sent
        res = nli(nli_input)[0]
        results.append({
            "sentence": sent,
            "label": res["label"],
            "score": res["score"]
        })
    return results

records[0]
```

```
{'question': 'What should you do if you suspect a concussion?',
 'model_answer': 'Question: What should you do if you suspect a
concussion?\nAnswer: Call your doctor or go to the emergency room.\
nWhat should you do if you suspect a concussion?\nCall your doctor or
go to the emergency room.\nThis is a new question on the SAT. It’s a
good question to review in case you are preparing for the test.\nWhat
should you do if you suspect a concussion? Call your doctor or go to
the emergency room.\nPrevious Post What is the best way to prepare for
the SAT?\nNext Post Which of the following is not true about the
SAT?',
 'standard_answer': 'Seek medical attention immediately, rest, and
avoid activities that could worsen your condition, such as physical
exertion and using digital screens.'}
```

```
# ---------------------------
# Step 7: Run detection on saved model outputs
# ---------------------------
all_detections = []
for rec in records:
    dets = detect_hallucinations(rec["model_answer"],
rec["standard_answer"])
    all_detections.append({
        "question": rec["question"],
        "detections": dets
    })
```

```
/usr/local/lib/python3.11/dist-packages/torch/nn/modules/
module.py:1750: FutureWarning: `encoder_attention_mask` is deprecated
and will be removed in version 4.55.0 for
`RobertaSdpaSelfAttention.forward`.
  return forward_call(*args, **kwargs)
You seem to be using the pipelines sequentially on GPU. In order to
maximize efficiency please use a dataset
```

```
for item in all_detections:
    print(f"Question: {item['question']}")
    for det in item['detections']:
        print(f"  • Label: {det['label']}, Confidence:
{det['score']:.2f}")
    print()
```

```
Question: What should you do if you suspect a concussion?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.84
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.95

Question: What are signs that you need to visit an optometrist?
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.96

Question: What should I do if I suspect food poisoning?
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.86
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.93

Question: What are the uses of calcium channel blockers?
  • Label: neutral, Confidence: 0.62
  • Label: entailment, Confidence: 0.52
  • Label: neutral, Confidence: 0.62
  • Label: entailment, Confidence: 0.52
  • Label: neutral, Confidence: 0.78
  • Label: entailment, Confidence: 0.52
  • Label: neutral, Confidence: 0.78
  • Label: entailment, Confidence: 0.96

Question: What should I do if I have swollen gums?
  • Label: neutral, Confidence: 0.91
  • Label: entailment, Confidence: 0.87
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00

Question: What should I take for gas and bloating?
  • Label: neutral, Confidence: 0.78
  • Label: neutral, Confidence: 0.52
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What is the use of anticoagulant medications?
  • Label: neutral, Confidence: 0.83
  • Label: entailment, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.94
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96

Question: What should you do if you experience side effects from a
medication in Australia?
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 0.63
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 0.89
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: entailment, Confidence: 0.52
  • Label: entailment, Confidence: 0.61
  • Label: neutral, Confidence: 0.99

Question: How can you improve bone density?
  • Label: entailment, Confidence: 0.64
  • Label: neutral, Confidence: 0.83
  • Label: neutral, Confidence: 0.85
  • Label: entailment, Confidence: 0.90
  • Label: entailment, Confidence: 0.99
  • Label: neutral, Confidence: 0.51
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.97

Question: How does Australia manage the regulation of prescription
opioids?
  • Label: entailment, Confidence: 0.92
  • Label: neutral, Confidence: 0.57
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99

Question: What should I take if I am feeling nauseous?
```

```
  • Label: neutral, Confidence: 0.87
  • Label: entailment, Confidence: 0.89
  • Label: neutral, Confidence: 0.92
  • Label: contradiction, Confidence: 0.83
  • Label: neutral, Confidence: 0.91
  • Label: neutral, Confidence: 0.83
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.90

Question: What type of medication is used for asthma control?
  • Label: entailment, Confidence: 0.56
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99

Question: What should I take if I have indigestion?
  • Label: entailment, Confidence: 0.73
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.92
  • Label: entailment, Confidence: 0.95

Question: What are the symptoms of the common cold?
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 1.00

Question: What medications help with the symptoms of menopause?
  • Label: entailment, Confidence: 0.74
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.72
  • Label: neutral, Confidence: 0.78
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.94

Question: What type of medication is used for treating high blood
pressure?
  • Label: entailment, Confidence: 0.44
  • Label: entailment, Confidence: 0.89
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99

Question: What is the role of antiretroviral drugs?
  • Label: neutral, Confidence: 0.93
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.54
  • Label: entailment, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.73
  • Label: entailment, Confidence: 0.98
  • Label: neutral, Confidence: 1.00

Question: What type of medication is used to treat rheumatoid
arthritis?
  • Label: entailment, Confidence: 0.70
  • Label: contradiction, Confidence: 0.76
  • Label: entailment, Confidence: 0.59
  • Label: entailment, Confidence: 0.99
  • Label: neutral, Confidence: 0.86
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.91

Question: What are effective treatments for seasonal allergies?
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.69
  • Label: neutral, Confidence: 0.82
  • Label: neutral, Confidence: 0.93
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.83
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.44

Question: What are common treatments for ADHD?
  • Label: entailment, Confidence: 0.88
  • Label: neutral, Confidence: 0.85
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96
  • Label: entailment, Confidence: 1.00
  • Label: neutral, Confidence: 0.86
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.83

Question: How can you check if a medication is subsidized under the
```

```
PBS?
  • Label: neutral, Confidence: 0.69
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.83
  • Label: neutral, Confidence: 0.90
  • Label: neutral, Confidence: 0.76

Question: What should I take for acne treatment?
  • Label: neutral, Confidence: 0.94
  • Label: neutral, Confidence: 0.63
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What is the recommended treatment for sunburn in Australia?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: entailment, Confidence: 1.00
  • Label: contradiction, Confidence: 0.81

Question: What are the benefits of getting enough sleep?
  • Label: entailment, Confidence: 0.80
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What are the symptoms of food allergies?
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.50
  • Label: neutral, Confidence: 0.70
  • Label: neutral, Confidence: 0.67
  • Label: neutral, Confidence: 0.88

Question: What is the role of the Therapeutic Goods Administration
(TGA) in Australia?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.98
  • Label: entailment, Confidence: 0.89
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00

Question: What is advised for treating dandruff?
  • Label: neutral, Confidence: 0.55
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.85
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.83
  • Label: neutral, Confidence: 0.58

Question: What medication is recommended for type 2 diabetes?
  • Label: neutral, Confidence: 0.92
  • Label: entailment, Confidence: 0.97
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 0.65
  • Label: neutral, Confidence: 0.99

Question: How do you recognize the signs of heat exhaustion?
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.87
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.98

Question: What drugs are prescribed for treating ADHD in adults?
  • Label: entailment, Confidence: 0.91
  • Label: neutral, Confidence: 0.52
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What should I take for a headache in Australia?
```

```
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99

Question: What are common treatments for insomnia?
  • Label: entailment, Confidence: 0.87
  • Label: neutral, Confidence: 0.72
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.97
  • Label: contradiction, Confidence: 0.94
  • Label: neutral, Confidence: 0.70
  • Label: neutral, Confidence: 1.00

Question: What is recommended for ear infections?
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What should I take for a stiff neck?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.93
  • Label: neutral, Confidence: 0.63
  • Label: contradiction, Confidence: 0.61
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: entailment, Confidence: 0.61
  • Label: neutral, Confidence: 0.86

Question: What are signs that you should see a doctor for stomach
pain?
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 0.77
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.91
  • Label: contradiction, Confidence: 0.58
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.94
  • Label: neutral, Confidence: 0.89
  • Label: neutral, Confidence: 0.76
  • Label: neutral, Confidence: 0.93
  • Label: neutral, Confidence: 0.88

Question: What drugs are used for treating tuberculosis?
  • Label: entailment, Confidence: 0.86
  • Label: neutral, Confidence: 0.94
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.81

Question: What is recommended for preventing dehydration?
  • Label: neutral, Confidence: 0.89
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.89
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00

Question: What is advised for the treatment of insect bites in
Australia?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.66
  • Label: neutral, Confidence: 0.99
  • Label: entailment, Confidence: 0.96
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99

Question: What is the Pharmaceutical Benefits Scheme (PBS) in
Australia?
  • Label: neutral, Confidence: 0.98
  • Label: entailment, Confidence: 0.90
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 0.80
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.82
  • Label: entailment, Confidence: 0.56
  • Label: neutral, Confidence: 0.99

Question: What are the early signs of dehydration?
  • Label: neutral, Confidence: 0.79
  • Label: entailment, Confidence: 0.65
  • Label: neutral, Confidence: 0.43

Question: What type of drugs are used to treat anxiety and insomnia?
  • Label: neutral, Confidence: 0.70
  • Label: entailment, Confidence: 0.86
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.76

Question: How can you protect your skin from sun damage?
  • Label: entailment, Confidence: 0.78
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 0.97
  • Label: entailment, Confidence: 0.87

Question: How can you tell if someone is having a stroke?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.97
  • Label: entailment, Confidence: 0.54
  • Label: neutral, Confidence: 0.83
  • Label: contradiction, Confidence: 0.43
  • Label: neutral, Confidence: 0.93
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.88

Question: What are some common over-the-counter pain relievers
available in Australia?
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.91
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.96

Question: What medications are available for migraine relief?
  • Label: entailment, Confidence: 0.67
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: entailment, Confidence: 0.97
  • Label: neutral, Confidence: 0.84

Question: What should I take for anxiety?
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.50
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99

Question: What dietary changes can help manage type 2 diabetes?
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.91
  • Label: entailment, Confidence: 0.95
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00

Question: What should I take for heartburn relief?
  • Label: neutral, Confidence: 0.82
  • Label: neutral, Confidence: 0.62
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.95

Question: What are the treatments for chronic kidney disease?
  • Label: neutral, Confidence: 0.54
  • Label: neutral, Confidence: 0.98
  • Label: neutral, Confidence: 0.96
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 1.00

Question: What are the best practices for dental hygiene?
  • Label: entailment, Confidence: 0.91
  • Label: entailment, Confidence: 0.98
  • Label: neutral, Confidence: 0.97
  • Label: entailment, Confidence: 0.96
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.95
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 1.00
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.99
  • Label: neutral, Confidence: 0.97
  • Label: neutral, Confidence: 0.70
```

```python
# Step 9: Aggregate and print overall results
# ---------------------------
for item in all_detections:
    labels = [d["label"] for d in item["detections"]]
    if "CONTRADICTION" in labels:
        overall = "Hallucination detected"
    elif "NEUTRAL" in labels:
        overall = "Potentially uncertain (no direct contradiction)"
    else:
        overall = "No hallucination detected"
    print(f"Question: {item['question']}\n  → {overall}\n")
```

```
Question: What should you do if you suspect a concussion?
  → No hallucination detected

Question: What are signs that you need to visit an optometrist?
  → No hallucination detected

Question: What should I do if I suspect food poisoning?
  → No hallucination detected

Question: What are the uses of calcium channel blockers?
  → No hallucination detected

Question: What should I do if I have swollen gums?
  → No hallucination detected

Question: What should I take for gas and bloating?
  → No hallucination detected

Question: What is the use of anticoagulant medications?
  → No hallucination detected

Question: What should you do if you experience side effects from a
medication in Australia?
  → No hallucination detected
```

```
Question: How can you improve bone density?
  → No hallucination detected

Question: How does Australia manage the regulation of prescription
opioids?
  → No hallucination detected

Question: What should I take if I am feeling nauseous?
  → No hallucination detected

Question: What type of medication is used for asthma control?
  → No hallucination detected

Question: What should I take if I have indigestion?
  → No hallucination detected

Question: What are the symptoms of the common cold?
  → No hallucination detected

Question: What medications help with the symptoms of menopause?
  → No hallucination detected

Question: What type of medication is used for treating high blood
pressure?
  → No hallucination detected

Question: What is the role of antiretroviral drugs?
  → No hallucination detected

Question: What type of medication is used to treat rheumatoid
arthritis?
  → No hallucination detected

Question: What are effective treatments for seasonal allergies?
  → No hallucination detected

Question: What are common treatments for ADHD?
  → No hallucination detected

Question: How can you check if a medication is subsidized under the
PBS?
  → No hallucination detected

Question: What should I take for acne treatment?
  → No hallucination detected

Question: What is the recommended treatment for sunburn in Australia?
  → No hallucination detected

Question: What are the benefits of getting enough sleep?
  → No hallucination detected
```

```
Question: What are the symptoms of food allergies?
  → No hallucination detected

Question: What is the role of the Therapeutic Goods Administration
(TGA) in Australia?
  → No hallucination detected

Question: What is advised for treating dandruff?
  → No hallucination detected

Question: What medication is recommended for type 2 diabetes?
  → No hallucination detected

Question: How do you recognize the signs of heat exhaustion?
  → No hallucination detected

Question: What drugs are prescribed for treating ADHD in adults?
  → No hallucination detected

Question: What should I take for a headache in Australia?
  → No hallucination detected

Question: What are common treatments for insomnia?
  → No hallucination detected

Question: What is recommended for ear infections?
  → No hallucination detected

Question: What should I take for a stiff neck?
  → No hallucination detected

Question: What are signs that you should see a doctor for stomach
pain?
  → No hallucination detected

Question: What drugs are used for treating tuberculosis?
  → No hallucination detected

Question: What is recommended for preventing dehydration?
  → No hallucination detected

Question: What is advised for the treatment of insect bites in
Australia?
  → No hallucination detected

Question: What is the Pharmaceutical Benefits Scheme (PBS) in
Australia?
  → No hallucination detected

Question: What are the early signs of dehydration?
```

```
  → No hallucination detected

Question: What type of drugs are used to treat anxiety and insomnia?
  → No hallucination detected

Question: How can you protect your skin from sun damage?
  → No hallucination detected

Question: How can you tell if someone is having a stroke?
  → No hallucination detected

Question: What are some common over-the-counter pain relievers
available in Australia?
  → No hallucination detected

Question: What medications are available for migraine relief?
  → No hallucination detected

Question: What should I take for anxiety?
  → No hallucination detected

Question: What dietary changes can help manage type 2 diabetes?
  → No hallucination detected

Question: What should I take for heartburn relief?
  → No hallucination detected

Question: What are the treatments for chronic kidney disease?
  → No hallucination detected

Question: What are the best practices for dental hygiene?
  → No hallucination detected
```

1. **Input Preparation** 1.1. I create a list of examples, each a dict with: • `text` (premise) • `text_pair` (hypothesis) 1.2. I call the pipeline on that list:

```
results = nli(pairs)
```

2. **Tokenization & Batch Collation** 2.1. **Concatenate** each pair into one sequence:

```
<s> PREMISE </s> </s> HYPOTHESIS </s>
```

3. **Joint Encoding via RoBERTa** 3.1. **Embed** input IDs into vectors 3.2. **Self-Attention Layers** • Every token attends to every other token in the concatenated sequence • This enables direct comparison

4. **Classification Head** 4.1. small feed-forward network 4.2. It produces **3 logits**:

- ENTAILMENT
- NEUTRAL
- CONTRADICTION

5. **Softmax & Label Selection** 5.1. softmax to the logits → probabilities that sum to 1 5.2. pick the label with the highest probability
6. **Return Results** 6.1. output a list of dicts in the same order as my input, each with: • `label` (ENTAILMENT/NEUTRAL/CONTRADICTION) • `score` (the probability)

model predictions, each given as Label Confidence (e.g. neutral 62%, entailment 52%, …, entailment 96%), showing for each input pair the predicted relation and how certain the model is.

```python
# Save detections for analysis
with open("hallucination_detections.json", "w") as f:
    json.dump(all_detections, f, indent=2)
print("Saved hallucination_detections.json")
```