

Project 2 - Methodology 2: Hallucination Vector Routing

Notebook Summary

This notebook details the process of building and validating a prompt-risk classifier for hallucination in Llama-3.1-8B, using the hallucination vector (`v_halluc`) constructed previously. The workflow includes: (1) generating a labeled dataset of model answers and hallucination judgments using SQuAD with standard, no-context, and distractor-context scenarios, (2) extracting a single feature for each prompt (the projection of the last prompt token's Layer 16 activation onto `v_halluc`), (3) training a logistic regression classifier to predict hallucination risk based on the aforementioned feature, and (4) evaluating the classifier's performance (AUROC \geq 0.75). The resulting `risk_clf.joblib` model enables real-time risk scoring for any prompt, forming the core of the project's guardrail system.

Step 2: Turning the Vector into a Prompt-Risk Score

Overall Goal: To build and validate a lightweight logistic regression classifier that takes a prompt's projection onto `v_halluc` and outputs a calibrated probability of hallucination. The final deliverables will be the trained classifier file (`risk_clf.joblib`) and a report on its predictive performance (AUROC \geq 0.75).

Setup and Installation

```
# Mount Google Drive
from google.colab import drive
drive.mount('/content/drive')

# Create a project directory to keep things organized
import os
PROJECT_DIR = "/content/drive/MyDrive/mistral-HallucinationVectorProj"
DATA_DIR = os.path.join(PROJECT_DIR, "data")
os.makedirs(DATA_DIR, exist_ok=True)

print(f"Project directory created at: {PROJECT_DIR}")

Mounted at /content/drive
Project directory created at: /content/drive/MyDrive/mistral-HallucinationVectorProj
```

```
!pip install -q --no-deps "trl==0.23.0" "peft==0.17.1" "accelerate==1
```

```

_____ 564.7/564.7 kB 12.0 MB/s e
_____ 59.4/59.4 MB 19.0 MB/s eta

```

```
!pip -q install "unsloth==2025.10.12" "transformers==4.57.1" "tqdm==4
```

```

_____ 61.5/61.5 kB 5.2 MB/s et
_____ 348.7/348.7 kB 17.1 MB/s e
_____ 139.8/139.8 kB 5.7 MB/s et
_____ 506.8/506.8 kB 18.9 MB/s e
_____ 9.5/9.5 MB 80.8 MB/s eta 0
_____ 301.8/301.8 kB 15.2 MB/s e
_____ 1.2/1.2 MB 38.9 MB/s eta 0
_____ 47.7/47.7 MB 18.5 MB/s eta
_____ 273.6/273.6 kB 8.9 MB/s et
_____ 2.2/2.2 MB 44.1 MB/s eta 0
_____ 117.2/117.2 MB 7.5 MB/s et
_____ 132.6/132.6 kB 5.1 MB/s et
_____ 1.6/1.6 MB 35.3 MB/s eta 0
_____ 7.2/7.2 MB 41.9 MB/s eta 0
_____ 213.6/213.6 kB 13.4 MB/s e

```

```

ERROR: pip's dependency resolver does not currently take into account
pylibcudf-cu12 25.6.0 requires pyarrow<20.0.0a0,>=14.0.0; platform_mac
cudf-cu12 25.6.0 requires pyarrow<20.0.0a0,>=14.0.0; platform_machine:

```

```
!pip install -q --index-url https://download.pytorch.org/whl/cu128 to
```

```
!pip install -q "xformers==0.0.33" --index-url https://download.pytor
```

```

_____ 303.7/303.7 MB 4.4 MB/s et

```

```

# Load API Keys
from google.colab import userdata
import os

# Load the keys into the environment
try:
    os.environ["HF_TOKEN"] = userdata.get('HF_TOKEN')
    os.environ["SCALEDOWN_API_KEY"] = userdata.get('SCALEDOWN_API_KEY')
    print("API keys loaded successfully.")
except userdata.SecretNotFoundError as e:
    print(f"ERROR: Secret not found. Please ensure you have created th
except Exception as e:
    print(f"An error occurred: {e}")

```

```
API keys loaded successfully.
```

```
# Load 4-bit Llama 3 8B Model and Tokenizer using Unsloth
import torch
from unsloth import FastLanguageModel

# Model loading parameters
max_seq_length = 2048
dtype = None # Unsloth handles dtype automatically for 4-bit models
load_in_4bit = True

# Load the model from Hugging Face
model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = "unsloth/mistral-7b-instruct-v0.3-bnb-4bit",
    max_seq_length = max_seq_length,
    dtype = dtype,
    load_in_4bit = load_in_4bit,
)

model = FastLanguageModel.for_inference(model)
model.gradient_checkpointing_disable()
model.config.use_cache = True
model.eval()

print("Model and Tokenizer loaded successfully!")
```

🐍 Unsloth: Will patch your computer to enable 2x faster free finetuning
 WARNING:torchao:Skipping import of cpp extensions due to incompatible
 WARNING:xformers:WARNING[XFORMERS]: xFormers can't load C++/CUDA extensions
 PyTorch 2.9.0+cu129 with CUDA 1209 (you have 2.8.0+cu126)
 Python 3.12.12 (you have 3.12.12)
 Please reinstall xformers (see <https://github.com/facebookresearch/xformers>)
 Memory-efficient attention, SwiGLU, sparse and more won't be available
 Set XFORMERS_MORE_DETAILS=1 for more details

=====
 Switching to PyTorch attention since your Xformers is broken.
 =====

Unsloth: Xformers was not installed correctly.
 Please install xformers separately first.
 Then confirm if it's correctly installed by running:
 python -m xformers.info

Longer error message:
 xFormers can't load C++/CUDA extensions. xFormers was built for:
 PyTorch 2.9.0+cu129 with CUDA 1209 (you have 2.8.0+cu126)
 Python 3.12.12 (you have 3.12.12)
 Please reinstall xformers (see <https://github.com/facebookresearch/xformers>)
 Memory-efficient attention, SwiGLU, sparse and more won't be available
 🐍 Unsloth Zoo will now patch everything to make training faster!
 ==((====))== Unsloth 2025.10.12: Fast Mistral patching. Transformers:
 \ \ / \ / Tesla T4. Num GPUs = 1. Max memory: 14.741 GB. Platform:
 0^0/ _/ \ Torch: 2.8.0+cu126. CUDA: 7.5. CUDA Toolkit: 12.6. Triton
 \ \ / \ Bfloat16 = FALSE. FA [Xformers = None. FA2 = False]
 "-__-" Free license: <http://github.com/unslothai/unsloth>
 Unsloth: Fast downloading is enabled - ignore downloading bars which appear
 Model and Tokenizer loaded successfully!

✓ Phase 1: Dataset Generation and Labeling

Overall Objective:

To programmatically generate a large, high-quality dataset of approximately 2000 labeled examples. Each example will consist of a prompt, a model-generated answer, and a binary label (1 for hallucination, 0 for correct) determined by a Gemini LLM judge. The final artifact of this phase will be a CSV file, `squad_labeled_answers.csv`, stored in Google Drive.

Methodology Overview:

We will use the [SQuAD dataset](#) as our source of truth. For each sampled entry, we will ask our 4-bit Llama 3 model to answer a question based only on the provided context. A specialized LLM judge will then compare the model's answer to the context and the ground-truth answer to determine if any unsupported information was fabricated (i.e., hallucinated).

This data is needed to train the logistic regression model for calculating hallucination risks for user prompts.

```
# Import libraries
import pandas as pd
from datasets import load_dataset
from tqdm.auto import tqdm
import numpy as np

# --- 1. Load the SQuAD dataset ---
print("Loading SQuAD dataset...")
squad_dataset = load_dataset("squad", split="train")
squad_df = squad_dataset.to_pandas()
print(f"Full dataset loaded with {len(squad_df)} rows.")
```

```
Loading SQuAD dataset...
Full dataset loaded with 87599 rows.
```

```
# --- 2. Sample the data ---
# We'll sample 2000 rows for our experiment
N_SAMPLES = 2000
if len(squad_df) > N_SAMPLES:
    sampled_df = squad_df.sample(n=N_SAMPLES, random_state=42).reset_index()
else:
    sampled_df = squad_df # Use full dataset if it's smaller

print(f"Using {len(sampled_df)} rows for our experiment.")
```

Using 2000 rows for our experiment.

Strategy: The "No-Context" and "Distractor-Context" Methods

To elicit varying responses that may or may not contain hallucination, instead of changing the instruction (original dataset prompt) to the model, we will change the information (context) we give it. We will create scenarios where the model is more likely to fail and invent an answer because the correct information is either missing or obscured. This is to elicit a mix of hallucinatory and non-hallucinatory responses to effectively train the logistic regression model.

We generate our ~2,000 examples by creating a mix of three different scenarios for each SQuAD entry.

Scenario 1: Standard In-Context (The "Easy" Case - Generates our Os)

We pass in the prompt + context exactly as taken from the dataset. We do this for 50% (1000 prompts) of our dataset.

```
FULL_PROMPT = f"Context:\n{context}\n\nQuestion:\n{question}"
```

Outcome: The model answers correctly most of the time. This is the primary source of our negative examples (label 0).

Scenario 2: No-Context (The "Hard" Case - Designed to induce natural hallucinations)

For a SQuAD entry, we deliberately withhold the context.

```
FULL_PROMPT = f"Question:\n{question}"
```

Outcome: The model's internal knowledge might contain information about the question's topic, but it may be incorrect, incomplete, or subtly different from the SQuAD context's specific answer. Without the grounding context, it is now much more likely to generate a plausible-sounding but factually incorrect answer. This is a primary source of natural positive examples (label 1).

Scenario 3: Distractor-Context (The "Tricky" Case - Also induces natural hallucinations)

For a SQuAD entry, we provide the correct question but pair it with a distractor context — a paragraph from a different, unrelated SQuAD article.

FULL_PROMPT =

f"Context:\n{distractor_context}\n\nQuestion:\n{question}"

Outcome: The model is still instructed to answer from the context. However, the answer is not there. A well-behaved model should say "I cannot find the answer in the context." A model prone to hallucination might try to synthesize an answer by blending its own knowledge with irrelevant information from the distractor context. This is another excellent source of natural positive examples (label 1).

```
# --- 3. Create a 'distractor_context' column ---
# For each row, the distractor is a context from another random row.
# We'll shuffle the context column and assign it.
distractor_indices = np.random.permutation(sampled_df.index)
sampled_df['distractor_context'] = sampled_df.loc[distractor_indices,

print("Dataset prepared with distractor contexts. Sample:")
print(sampled_df[['context', 'question', 'distractor_context']].head(

# --- 4. Assign a scenario to each row ---
# We'll divide our 2000 samples into the three scenarios
scenarios = []
# 1000 for standard, 500 for no-context, 500 for distractor
scenario_counts = {'standard': 1000, 'no_context': 500, 'distractor':

for scenario_type, count in scenario_counts.items():
    scenarios.extend([scenario_type] * count)

# Make sure the scenarios list matches the dataframe length
sampled_df = sampled_df.iloc[:len(scenarios)].copy()
sampled_df['scenario'] = scenarios

print("\nScenario distribution:")
print(sampled_df['scenario'].value_counts())
```

```
Dataset prepared with distractor contexts. Sample:
                                     context \
0  The world's first institution of technology or...
1  The standard specifies how speed ratings shoul...
2  The most impressive and famous of Sumerian bui...
3  Ann Arbor has a council-manager form of govern...
4  Shortly before his death, when he was already ...

                                     question \
0          What year was the Banská Akadémia founded?
1  What is another speed that can also be reporte...
2  Where were the use of advanced materials and t...
3          Who is elected every even numbered year?
4  What was the purpose of top secret ICBM commit...

                                     distractor_context
0  Russell's ministry, though Whig, was not favou...
1  There is a theory that vinyl records can audib...
2  The Council of Trent approved a plan in 1563 f...
3  The analog information encoded on LaserDiscs d...
```

4 In 1045, the Byzantine Empire conquered Bagrat...

Scenario distribution:

```
scenario
standard      1000
no_context     500
distractor     500
Name: count, dtype: int64
```

✓ Define Generation and Judging Logic

Define the functions that will generate an answer based on the scenario and get it judged by Gemini.

```
# --- 1. Define the prompt template and generation function ---
SYSTEM_PROMPT = "You are a helpful assistant. Answer the following question"

def generate_squad_answer_multi_scenario(model, tokenizer, row):
    """
    Generates an answer based on the scenario specified in the DataFrame row.
    """
    scenario = row['scenario']
    question = row['question']

    context_to_use = ""
    if scenario == 'standard':
        context_to_use = row['context']
    elif scenario == 'distractor':
        context_to_use = row['distractor_context']
    # For 'no_context', context_to_use remains an empty string

    # Construct the user prompt
    if context_to_use:
        user_prompt = f"Context:\n{context_to_use}\n\nQuestion:\n{question}"
    else:
        user_prompt = f"Question:\n{question}"

    # Format into the Llama 3 chat template
    messages = [
        {"role": "system", "content": SYSTEM_PROMPT},
        {"role": "user", "content": user_prompt}
    ]

    # Check if tokenizer has a chat template, if not format manually
    try:
        full_prompt_templated = tokenizer.apply_chat_template(messages,
            except Exception:
        # Manual fallback for Mistral base models
        full_prompt_templated = f"<s>[INST] <<SYS>>\n{SYSTEM_PROMPT}\n<<[Q] {question}>>[EOT]"

    # Generate the answer
    inputs = tokenizer(full_prompt_templated, return_tensors="pt").to(device)
    with torch.no_grad():
```

```

        outputs = model.generate(
            **inputs,
            max_new_tokens=128,
            do_sample=False,
            pad_token_id=tokenizer.eos_token_id
        )

    response_text = tokenizer.decode(outputs[0][inputs['input_ids'].start_index:])

    # We return the templated prompt as well, as this is what the model actually
    return response_text.strip(), full_prompt_templated

```

```

import requests
import json
import time
import os

# --- 1. Define the Judge Prompt Template ---
# This prompt is highly specific to our task and gives the judge all the information it needs
JUDGE_PROMPT_TEMPLATE = """
You are an impartial judge evaluating a language model's answer for its quality.
Your task is to determine if the model's answer contains any fabrications or hallucinations.

Here is the provided context:
<context>
{context}
</context>

Here is the question that was asked:
<question>
{question}
</question>

Here is the ground-truth answer, which is known to be correct and derived from reliable sources:
<ground_truth_answer>
{ground_truth_answer}
</ground_truth_answer>

Here is the language model's generated answer:
<model_answer>
{model_answer}
</model_answer>

**Your Task:**
Carefully compare the <model_answer> against the <context>. Does the model's answer contain any fabrications or hallucinations?
Ignore minor differences in phrasing. If the model says it cannot find an answer, that is acceptable.

Respond with only the word 'YES' if the model's answer contains fabrications or hallucinations.
Otherwise, respond with 'NO'.

# --- 2. Implement the Judge Function ---
def judge_squad_answer(context, question, model_answer, ground_truth_answer):
    """
    Gets a binary hallucination label from the LLM judge.
    """

```



```

Returns: 1 for hallucination ('YES'), 0 for faithful ('NO'), -1 for error
"""
prompt = JUDGE_PROMPT_TEMPLATE.format(
    context=context,
    question=question,
    model_answer=model_answer,
    ground_truth_answer=ground_truth_answer
)

# Using your existing API call structure
url = "https://api.scaledown.xyz/compress/" # Or your preferred G
payload = json.dumps({
    "context": "You are an impartial judge evaluating for halluci
    "prompt": prompt,
    "model": "gpt-4o",
    "scaledown": {
        "rate": 0 # no compression
    }
})
headers = {'x-api-key': api_key, 'Content-Type': 'application/json'}

try:
    response = requests.post(url, headers=headers, data=payload)
    response.raise_for_status()

    response_data = json.loads(response.text)
    content = response_data.get("full_response", "").strip().upper()

    if 'YES' in content:
        return 1
    elif 'NO' in content:
        return 0
    else:
        print(f"Judge Warning: Unexpected response: {content}")
        return -1 # Indicate an error in parsing

except requests.exceptions.RequestException as e:
    print(f"ERROR: API request failed: {e}")
    return -1
except (json.JSONDecodeError, KeyError) as e:
    print(f"ERROR: Could not parse judge's response: {response.te
    return -1

```

Run the generation + judging loop and save results to Drive.

```

import os
import pandas as pd
from google.colab import drive
import time

# --- 1. Setup paths and constants ---
drive.mount('/content/drive')
OUTPUT_CSV_PATH = '/content/drive/MyDrive/mistral-HallucinationVectorI

```

```

BATCH_SIZE = 20 # Save progress every 20 rows
API_KEY = os.environ["SCALEDOWN_API_KEY"] # Load from Colab secrets

# --- 2. The Main Loop ---
results_list = []
# Check if a partial file exists to resume from
try:
    existing_df = pd.read_csv(OUTPUT_CSV_PATH)
    start_index = len(existing_df)
    print(f"Resuming from index {start_index}.")
except FileNotFoundError:
    existing_df = pd.DataFrame()
    start_index = 0
    print("Starting from scratch.")

# Use tqdm for a progress bar
pbar = tqdm(total=len(sampled_df), initial=start_index)
for i in range(start_index, len(sampled_df)):
    row = sampled_df.iloc[i]

    # --- Generate ---
    model_answer, full_prompt = generate_squad_answer_multi_scenario(

    # --- Judge ---
    # The judge ALWAYS compares against the original, correct context
    ground_truth_answer = row['answers']['text'][0] if row['answers']
    label = judge_squad_answer(
        context=row['context'], # Always the original context
        question=row['question'],
        model_answer=model_answer,
        ground_truth_answer=ground_truth_answer,
        api_key=API_KEY
    )

    # Store the result
    results_list.append({
        'scenario': row['scenario'],
        'full_prompt': full_prompt, # The actual text fed to the mode
        'model_answer': model_answer,
        'ground_truth_answer': ground_truth_answer,
        'hallucination_label': label,
        'original_context': row['context'], # Keep for reference
        'question': row['question']
    })

    pbar.update(1)

    # --- Save progress in batches and print counts ---
    if (i + 1) % BATCH_SIZE == 0 or (i + 1) == len(sampled_df):
        temp_df = pd.DataFrame(results_list)

        # Append to the existing DataFrame and save
        if not existing_df.empty:
            combined_df = pd.concat([existing_df, temp_df], ignore_in
        else:

```

```
combined_df = temp_df

combined_df.to_csv(OUTPUT_CSV_PATH, index=False)

print(f"\nSaved batch up to index {i}. Total rows in file: {len(combined_df)}")

# Print count of 1s and 0s for the saved batch
batch_start_index = len(combined_df) - len(temp_df)
batch_end_index = len(combined_df) - 1
print(f"Batch {batch_start_index//BATCH_SIZE + 1} (rows {batch_start_index+1} to {batch_end_index+1})")
valid_labels = temp_df[temp_df['hallucination_label'] != -1]['hallucination_label']
if not valid_labels.empty:
    counts = valid_labels.value_counts().sort_index()
    print(counts)
else:
    print("No valid labels in this batch.")

# Update for next resume
existing_df = combined_df
results_list = [] # Clear the list for the next batch

pbar.close()
print("Phase 1 complete. Labeled dataset saved to Google Drive.")

# --- Final check of the class balance ---
final_df = pd.read_csv(OUTPUT_CSV_PATH)
print("\nFinal Class Balance:")
# We filter out any rows where the judge failed (returned -1)
print(final_df[final_df['hallucination_label'] != -1]['hallucination_label'].value_counts())
```

Drive already mounted at /content/drive; to attempt to forcibly remount
Resuming from index 1320.

Saved batch up to index 1339. Total rows in file: 1340

Batch 67 (rows 1320 to 1339):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1359. Total rows in file: 1360

Batch 68 (rows 1340 to 1359):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1379. Total rows in file: 1380

Batch 69 (rows 1360 to 1379):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1399. Total rows in file: 1400

Batch 70 (rows 1380 to 1399):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1419. Total rows in file: 1420

Batch 71 (rows 1400 to 1419):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1439. Total rows in file: 1440

Batch 72 (rows 1420 to 1439):

hallucination_label

0 4

1 16

Name: count, dtype: int64

Saved batch up to index 1459. Total rows in file: 1460

Batch 73 (rows 1440 to 1459):

hallucination_label

1 20

Name: count, dtype: int64

Saved batch up to index 1479. Total rows in file: 1480

Batch 74 (rows 1460 to 1479):

hallucination_label

0 1

1 19

Name: count, dtype: int64

Saved batch up to index 1499. Total rows in file: 1500

Batch 75 (rows 1480 to 1499):

hallucination_label

1 20

Name: count, dtype: int64

✓ Saved batch up to index 1519. Total rows in file: 1520

Phase 2: Feature Calculation

```
Batch 76 (rows 1500 to 1519):
hallucination_label
```

```
0      1
1     19
```

Overall Objective:

Save batch up to index 1539. Total rows in file: 1540
To process our squad labeled answers, we read in the multi_scenario.csv file and add a new column, z-feature, to it. This feature is the dot product of the last prompt token's Layer 16 activation with our v_halluc vector. The final artifact will be an updated CSV file, ready for training our classifier in the next phase.

Save batch up to index 1559. Total rows in file: 1560

Methodology Overview:

```
Batch 78 (rows 1540 to 1559):
hallucination_label
```

```
0      2
1     18
```

We load our pre-computed hallucination vector and the labeled dataset. Then, in a resilient, batched loop, we feed each prompt from the dataset into the Llama 3 model, extract the specific activation vector we need, compute the projection, and save the results incrementally.

```
Batch 79 (rows 1560 to 1579):
hallucination_label
```

```
1     20
```

Name: count, dtype: int64

✓ Load the Hallucination Vector and Labeled Data

Save batch up to index 1599. Total rows in file: 1600

```
Batch 80 (rows 1580 to 1599):
```

Load the necessary artifacts (v_halluc.pt and the CSV from Phase 1) into our Colab environment

```
1     19
```

Name: count, dtype: int64

```
import pandas as pd
import numpy as np
import torch
from tqdm.auto import tqdm
```

```
TARGET_LAYER = 16
```

```
# --- 1. Define paths ---
```

```
VECTOR_PATH = '/content/drive/MyDrive/mistral-HallucinationVectorProj
LABELED_DATA_PATH = '/content/drive/MyDrive/mistral-HallucinationVecto
OUTPUT_PATH = '/content/drive/MyDrive/mistral-HallucinationVectorProj
```

```
# --- 2. Load the hallucination vector ---
```

```
v_halluc = torch.load(VECTOR_PATH).to(model.device)
assert v_halluc.ndim == 1, "v_halluc must be a 1D vector"
# Hidden size for Mistral-7B is 4096; this must match the model's hid
test_ids = tokenizer("hi", return_tensors="pt").to(model.device)
with torch.no_grad():
    hs = model.model(input_ids=test_ids["input_ids"],
                      output_hidden_states=True, return_dict=True).hid
hidden_size = hs[TARGET_LAYER].size(-1)
assert v_halluc.shape[0] == hidden_size, "Loaded vector dim != model |
```

```
# --- 3. Load and clean the labeled dataset ---
```

```

labeled_df = pd.read_csv(LABELED_DATA_PATH)
print(f"Loaded {len(labeled_df)} labeled examples.")

# Filter out rows where the judge failed (label == -1) and any rows w
initial_rows = len(labeled_df)
labeled_df = labeled_df[labeled_df['hallucination_label'] != -1].drop
labeled_df['hallucination_label'] = labeled_df['hallucination_label']
final_rows = len(labeled_df)

print(f"Filtered out {initial_rows - final_rows} rows with invalid la
print(f"Proceeding with {final_rows} valid examples.")
print("Cleaned DataFrame sample:")
print(labeled_df.head())

```

```

1      18
Loaded 2000 labeled examples.
Name: count, dtype: int64
Filtered out 0 rows with invalid labels.
Proceeding with 2000 valid examples.
Saved batch up to index 1759. Total rows in file: 1760
Batch 88 (rows 1740 to 1759):
hallucination_label      full_prompt \
0      standard <s>[INST] You are a helpful assistant. Answer ...
1      standard <s>[INST] You are a helpful assistant. Answer ...
Name: count, dtype: int64
2      standard <s>[INST] You are a helpful assistant. Answer ...
3      standard <s>[INST] You are a helpful assistant. Answer ...
Saved batch up to index 1779. Total rows in file: 1780
Batch 89 (rows 1760 to 1779):
hallucination_label      model_answer \
0      The Banská Akadémia was founded in the year 1735.
Name: count, dtype: int64
1      The SOS-based speed (explicitly as being an SO...
2      In Sumer, advanced materials and techniques we...
Saved batch up to index 1799. Total rows in file: 1800
Batch 90 (rows 1780 to 1799):
hallucination_label      mayor
0      The RH-108 is the top secret ICBM (Interconti...
1      19
Name: count, dtype: int64
ground_truth_answer      hallucination_la
0      1735
1      SOS-based speed
2      Sumerian temples and palaces
Saved batch up to index 1819. Total rows in file: 1820
Batch 91 (rows 1800 to 1819):
hallucination_label      mayor
0      decide on the feasibility of building an ICBM ...
1      20
Name: count, dtype: int64
original_context \
0      The world's first institution of technology or...
1      The standard specifies how speed ratings shou...
Saved batch up to index 1839. Total rows in file: 1840
Batch 92 (rows 1820 to 1839):
hallucination_label
0      Ann Arbor has a council-manager form of govern...
1      Shortly before his death, when he was already ...
Name: count, dtype: int64
question
0      What year was the Banská Akadémia founded?
1      What is another speed that can also be reporte...
Saved batch up to index 1859. Total rows in file: 1860
Batch 93 (rows 1840 to 1859):
hallucination_label
0      Where were the use of advanced materials and t...
1      Who is elected every even numbered year?
2      What was the purpose of top secret ICBM commit...
Name: count, dtype: int64

```

Saved batch up to index 1879. Total rows in file: 1880

Batch 94 (rows 1860 to 1879):

hallucination_label

1 20

Name: count, dtype: int64

✓ Implement the Last-Token Activation Extraction Function

Create a function that takes a prompt, runs it through the model, and returns the specific hidden-state vector of the last token of that prompt at Layer 16.

```

Saved batch up to index 1899. Total rows in file: 1900
Batch 95 (rows 1880 to 1899):
hallucination_label

TARGET_LAYER = 16

def get_last_prompt_token_activation(model, tokenizer, prompt_text):
    inputs = tokenizer(prompt_text, return_tensors="pt",
                        truncation=True, max_length=2048).to(model.device)
    with torch.no_grad():
        outputs = model.model(
            input_ids=inputs["input_ids"],
            output_hidden_states=True,
            return_dict=True,
            use_cache=True,
        )
    hidden = outputs.hidden_states[TARGET_LAYER] # (1, seq, 4096)
    return hidden[0, -1, :].squeeze()

return last_token_activation

# --- Test the function with one example ---
example_prompt = labeled_df.iloc[0]['full_prompt']
activation_vector = get_last_prompt_token_activation(model, tokenizer)
print(f"Successfully extracted activation for one prompt.")
print(f"Activation vector shape: {activation_vector.shape}")
print(f"Activation vector dtype: {activation_vector.dtype}")

```

```

Successfully extracted activation for one prompt.
Activation vector shape: torch.Size([4096])
Activation vector dtype: torch.float16
Name: count, dtype: int64

```

Saved batch up to index 1999. Total rows in file: 2000

Batch 100 (rows 1980 to 1999):

hallucination_label

0 1

1 19

Name: count, dtype: int64

Phase 1 complete.

✓ The Main Processing Loop (Extract, Project, Save)

Iterate through our entire dataset, compute the z feature for each prompt, and save the results incrementally. Labeled dataset saved to Google Drive.

Final Class Balance:

```

import os
import pandas as pd
from tqdm.auto import tqdm
import torch
import numpy as np # It's good practice to import numpy for pd.NA/np.

# --- 1. Setup for the loop ---
BATCH_SIZE = 50 # Save progress every 50 rows

# --- 2. Check for existing progress to resume ---
# This logic is now fully robust for both resuming and starting fresh
try:
    df_to_process = pd.read_csv(OUTPUT_PATH)
    start_index = len(df_to_process)
    # Ensure the z_feature column exists if we are resuming an older

```

```

if 'z_feature' not in df_to_process.columns:
    df_to_process['z_feature'] = pd.NA
print(f"Resuming feature calculation from index {start_index}.")
except FileNotFoundError:
    start_index = 0
    df_to_process = labeled_df.copy() # Use the DataFrame we loaded in

# --- THIS IS THE FIX ---
# Before starting the loop, we add the empty 'z_feature' column.
print("Initializing 'z_feature' column for new run.")
df_to_process['z_feature'] = pd.NA
# -----

print("Starting feature calculation from scratch.")

# --- 3. The Main Loop ---
# The loop now operates on a DataFrame that is guaranteed to have the
pbar = tqdm(total=len(df_to_process), initial=start_index, desc="Calculating z_feature")
for i in range(start_index, len(df_to_process)):
    # Check if the feature has already been computed in a previous run
    if pd.isna(df_to_process.loc[i, 'z_feature']):
        pbar.update(1)
        continue

    row = df_to_process.iloc[i]
    prompt = row['full_prompt']

    # --- Extract Activation ---
    last_token_activation = get_last_prompt_token_activation(model, token_embeddings, prompt)

    # --- Compute Projection (Dot Product) ---
    # Ensure both vectors are on the same device and have the same dtype
    projection = torch.dot(last_token_activation.to(v_halluc.device), v_halluc.to(v_halluc.device))
    z_feature = projection.item() # .item() gets the scalar value

    # Store the feature in the DataFrame
    df_to_process.loc[i, 'z_feature'] = z_feature

    pbar.update(1)

    # --- Save progress in batches ---
    if (i + 1) % BATCH_SIZE == 0 or (i + 1) == len(df_to_process):
        df_to_process.to_csv(OUTPUT_PATH, index=False)
        pbar.set_description(f"Saved batch up to index {i}")

# Final save to ensure the last batch is written
df_to_process.to_csv(OUTPUT_PATH, index=False)
pbar.close()
print(f"Phase 2 complete. Final dataset with features saved to: {OUTPUT_PATH}")

# --- Final check of the output ---
final_df = pd.read_csv(OUTPUT_PATH)
print("\nFinal DataFrame with 'z_feature' column:")

```



```
print(final_df[['full_prompt', 'hallucination_label', 'z_feature']].h
print(f"\nDescription of z_feature:\n{final_df['z_feature'].describe(
```

Initializing 'z_feature' column for new run.
 Starting feature calculation from scratch.
 Phase 2 complete. Final dataset with features saved to: /content/drive

Final DataFrame with 'z_feature' column:

	full_prompt	hallucination_la
0	<s>[INST] You are a helpful assistant. Answer ...	
1	<s>[INST] You are a helpful assistant. Answer ...	
2	<s>[INST] You are a helpful assistant. Answer ...	
3	<s>[INST] You are a helpful assistant. Answer ...	
4	<s>[INST] You are a helpful assistant. Answer ...	

z_feature

0	-0.259766
1	-0.066956
2	-0.159302
3	-0.294678
4	-0.125366

Description of z_feature:

count	2000.000000
mean	-0.355610
std	0.193115
min	-1.177734
25%	-0.474182
50%	-0.314087
75%	-0.213989
max	0.024338

Name: z_feature, dtype: float64

✓ Phase 3: Training and Validating the Risk Scorer

Objective: To use our (z_feature, hallucination_label) dataset to train a simple, fast classifier and rigorously evaluate its predictive power.

✓ Load Data and Create a Train-Test Split

Prepare our data for supervised learning by splitting it into a training set for the model to learn from and a held-out test set to evaluate its performance on unseen data.

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score, roc_curve, classification_
import matplotlib.pyplot as plt
import joblib
```

--- 1. Load the dataset with features ---

```

FEATURES_DATA_PATH = '/content/drive/MyDrive/mistral-HallucinationVec
df_final = pd.read_csv(FEATURES_DATA_PATH)

print("Loaded dataset with features. Sample:")
print(df_final[['z_feature', 'hallucination_label']].head())

# --- 2. Define Features (X) and Labels (y) ---
# Our feature X is the 'z_feature' column
X = df_final[['z_feature']]
# Our target y is the 'hallucination_label' column
y = df_final['hallucination_label']

# --- 3. Create the Train-Test Split ---
# We'll use a standard 80/20 split.
# 'stratify=y' ensures the proportion of 0s and 1s is the same in botl
# 'random_state=42' makes our split reproducible.
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    stratify=y,
    random_state=42
)

print(f"\nData split into training and testing sets:")
print(f"Training set size: {len(X_train)} samples")
print(f"Test set size: {len(X_test)} samples")
print(f"Hallucination proportion in training set: {y_train.mean():.2%}")
print(f"Hallucination proportion in test set: {y_test.mean():.2%}")

```

Loaded dataset with features. Sample:

	z_feature	hallucination_label
0	-0.259766	0
1	-0.066956	0
2	-0.159302	1
3	-0.294678	0
4	-0.125366	0

Data split into training and testing sets:

Training set size: 1600 samples

Test set size: 400 samples

Hallucination proportion in training set: 57.12%

Hallucination proportion in test set: 57.00%

✓ Fit the Logistic Regression Model

Train the logistic regression classifier on our training data and save the resulting model.

```

# --- 1. Initialize and train the model ---
print("Training the logistic regression model...")
risk_classifier = LogisticRegression(random_state=42)
risk_classifier.fit(X_train, y_train)

```

```
print("Training complete.")

# --- 2. Inspect the learned coefficients (optional but insightful) ---
intercept = risk_classifier.intercept_[0]
coefficient = risk_classifier.coef_[0][0]
print(f"Learned Intercept ( $\beta_0$ ): {intercept:.4f}")
print(f"Learned Coefficient ( $\beta_1$ ): {coefficient:.4f}")
# A positive coefficient means that a higher z_feature value correspo

# --- 3. Save the trained model for later use ---
CLASSIFIER_PATH = '/content/drive/MyDrive/mistral-HallucinationVector'
joblib.dump(risk_classifier, CLASSIFIER_PATH)
print(f"Classifier saved to: {CLASSIFIER_PATH}")
```

```
Training the logistic regression model...
Training complete.
Learned Intercept ( $\beta_0$ ): -2.0402
Learned Coefficient ( $\beta_1$ ): -7.0213
Classifier saved to: /content/drive/MyDrive/mistral-HallucinationVecto
```

✓ Evaluate Performance on the Test Set

Test our classifier on unseen data and verify that it meets our AUROC ≥ 0.75 success criterion.

```
# --- 1. Predict probabilities on the test set ---
# We need the probability of the positive class (hallucination, which
y_pred_proba = risk_classifier.predict_proba(X_test)[: , 1]

# --- 2. Calculate and validate the AUROC score ---
auroc_score = roc_auc_score(y_test, y_pred_proba)
print(f"\n--- Performance Evaluation ---")
print(f"AUROC Score on Test Set: {auroc_score:.4f}")

if auroc_score >= 0.75:
    print("✅ Success! AUROC meets or exceeds the target of 0.75.")
else:
    print("⚠️ Warning: AUROC is below the target of 0.75.")

# --- 3. Plot the ROC Curve for our report ---
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)

plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area :
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--', label='R
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.grid(True)
```

```
plt.show()
```

```
# --- 4. Generate a detailed classification report ---
```

```
# This shows precision and recall at a default 0.5 probability threshold
```

```
y_pred_binary = (y_pred_proba >= 0.5).astype(int)
```

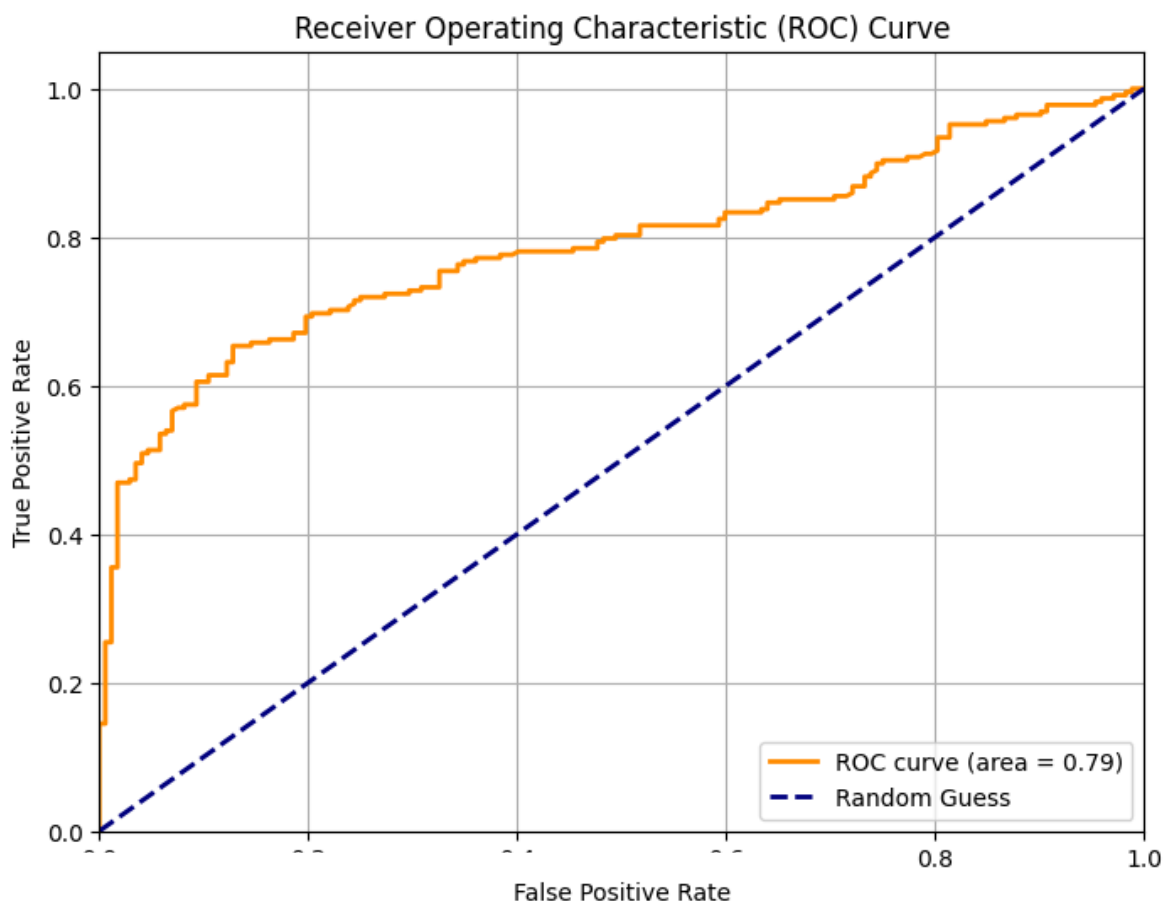
```
print("\nClassification Report (at 0.5 threshold):")
```

```
print(classification_report(y_test, y_pred_binary, target_names=['Faithful', 'Hallucination']))
```

```
--- Performance Evaluation ---
```

```
AUROC Score on Test Set: 0.7870
```

```
✅ Success! AUROC meets or exceeds the target of 0.75.
```



```
Classification Report (at 0.5 threshold):
```

	precision	recall	f1-score	support
Faithful (0)	0.66	0.77	0.71	172
Hallucination (1)	0.80	0.70	0.75	228
accuracy			0.73	400
macro avg	0.73	0.73	0.73	400
weighted avg	0.74	0.73	0.73	400

✓ Phase 4: Create the Real-Time Risk Function

Objective: To encapsulate our entire prediction pipeline into a single, easy-to-use function that will be the core of our Step 3 guardrail.

Essentially, we apply our logistic regression model on the layer-16 last token

activation's projection of user prompts on a pre-defined vector (z-score) to get the "risk

```
import joblib

# --- 1. Load all necessary artifacts once ---
# This is more efficient than loading them inside the function every time
v_halluc_loaded = torch.load(VECTOR_PATH).to(model.device)
risk_classifier_loaded = joblib.load(CLASSIFIER_PATH)
print("Loaded vector and classifier for real-time function.")

# --- 2. Define the final, real-time risk function ---
def get_hallucination_risk(prompt_text, model, tokenizer, v_halluc, classifier):
    """
    Takes a raw prompt and returns a calibrated hallucination risk score
    """
    # Step 1: Extract the last-token activation at Layer 16
    activation = get_last_prompt_token_activation(model, tokenizer, prompt_text)

    # Step 2: Compute the projection (z-feature)
    projection = torch.dot(activation.to(v_halluc.device), v_halluc)
    z_feature = projection.item()

    # Step 3: Use the classifier to predict the probability
    # The classifier expects a 2D array, so we reshape
    z_feature_reshaped = [[z_feature]]
    risk_probability = classifier.predict_proba(z_feature_reshaped)[0]

    return risk_probability

# --- 3. Test the function on safe and risky examples ---
# Find one 'standard' and one 'no_context' example from our DataFrame
safe_prompt_example = df_final[df_final['scenario'] == 'standard'].iloc[0]
risky_prompt_example = df_final[df_final['scenario'] == 'no_context'].iloc[0]

# Calculate risk for both
risk_safe = get_hallucination_risk(safe_prompt_example, model, tokenizer, v_halluc_loaded, risk_classifier_loaded)
risk_risky = get_hallucination_risk(risky_prompt_example, model, tokenizer, v_halluc_loaded, risk_classifier_loaded)

print(f"\n--- Real-Time Function Test ---")
print(f"Risk score for a likely SAFE prompt: {risk_safe:.4f}")
print(f"Risk score for a likely RISKY prompt: {risk_risky:.4f}")
```