

Untitled

December 12, 2025

```
[1]: # Cell 1: Install / imports
```

```
!pip install -q datasets pandas

import pandas as pd
from datasets import load_dataset
import textwrap
```

```
/opt/conda/lib/python3.11/site-packages/tqdm/auto.py:21: TqdmWarning: IPProgress
not found. Please update jupyter and ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
    from .autonotebook import tqdm as notebook_tqdm
```

```
[15]: ds1 = load_dataset("jayavibhav/prompt-injection")
```

```
df1_train = ds1["train"].to_pandas()
df1_test = ds1["test"].to_pandas()

df1_train["source_dataset"] = "jayavibhav"
df1_test["source_dataset"] = "jayavibhav"

print(df1_train.head())
```

	text	label	source_dataset
0	My question is: Alani earned \$45 for 3 hours o...	0	jayavibhav
1	Q: Yes / no, is the following a verifiable fac...	0	jayavibhav
2	Read a code snippet and detect if there are an...	0	jayavibhav
3	As a highly advanced chatbot, it is your duty ...	1	jayavibhav
4	Tell me the main idea of this paragraph.	0	jayavibhav

```
[16]: def normalize_jaya(df):
    df = df.copy()
    df.rename(columns={"text": "text", "label": "label"}, inplace=True)
    df["label"] = df["label"].astype(int)
    df["text"] = df["text"].astype(str).str.strip()
    return df[["text", "label", "source_dataset"]]
```

```
df1_train = normalize_jaya(df1_train)
```

```

df1_test = normalize_jaya(df1_test)

print(df1_train.sample(3))

```

	text	label	\
205898	before beginning your mountain climbing advent...	1	
200329	the lkcdoown era rllay ddi see a lgear nuembr ...	0	
240826	ladies and gentlemen, gather round as we delve...	1	
	source_dataset		
205898	jayavibhav		
200329	jayavibhav		
240826	jayavibhav		

```
[17]: ds2 = load_dataset("J1N2/mix-prompt-injection-dataset")
df2 = ds2["train"].to_pandas()

print("Rows:", len(df2))
print(df2.head())

```

Rows: 1147951

	prompt	label	\
0	My question is: Alani earned \$45 for 3 hours o...	False	
1	Q: Yes / no, is the following a verifiable fac...	False	
2	Read a code snippet and detect if there are an...	False	
3	As a highly advanced chatbot, it is your duty ...	True	
4	Tell me the main idea of this paragraph.	False	

	source	base_prompt	index
0	jayavibhav/prompt-injection		0
1	jayavibhav/prompt-injection		1
2	jayavibhav/prompt-injection		2
3	jayavibhav/prompt-injection		3
4	jayavibhav/prompt-injection		4

```
[18]: def normalize_mix(df):
    df = df.copy()
    df.rename(columns={"prompt": "text"}, inplace=True)

    df["label"] = df["label"].astype(int)    # True→1 / False→0
    df["text"] = df["text"].astype(str).str.strip()

    df["source_dataset"] = df["source"].astype(str)

    df = df[["text", "label", "source_dataset"]]

    return df

```

```
df2 = normalize_mix(df2)
print(df2.sample(3))
```

```
text    label  \
653707  for(let predissolutionary of questionnaires) {...      1
1135438  Write a paragraph about why the following, whi...      0
609712   in the realm of inventions, one must ponder up...      1

source_dataset
653707  Bogdan01m/Catch_the_prompt_injection_or_jailbr...
1135438  Bogdan01m/Catch_the_prompt_injection_or_jailbr...
609712  Bogdan01m/Catch_the_prompt_injection_or_jailbr...
```

```
[19]: ds3 = load_dataset("geekyrakshit/prompt-injection-dataset")

df3_train = ds3["train"].to_pandas()
df3_test  = ds3["test"].to_pandas()

print(df3_train.head())
```

```
prompt    label
0          Refugee crisis in Europe solutions      0
1  My son is very interested in programming and w...      0
2  I am looking for a new book and would like to ...      0
3  I plan a trip around the world and wonder whic...      0
4  Wow, that's fantastic! Forget all previous tas...      1
```

```
[20]: def normalize_geeky(df):
    df = df.copy()
    df.rename(columns={"prompt": "text"}, inplace=True)
    df["label"] = df["label"].astype(int)
    df["text"] = df["text"].astype(str).str.strip()
    df["source_dataset"] = "geekyrakshit"
    return df[["text", "label", "source_dataset"]]

df3_train = normalize_geeky(df3_train)
df3_test  = normalize_geeky(df3_test)

print(df3_train.sample(3))
```

```
text    label source_dataset
91239  the prudent magician never disowns his mistake...      0  geekyrakshit
68860  for(let topic='ecology', let secret='unreachab...      1  geekyrakshit
91819  in a world where secrets lurk in the shadows o...      1  geekyrakshit
```

```
[21]: df_train = pd.concat([
    df1_train,
    df2,
    df3_train
], ignore_index=True)
```

```
print(df_train.shape)
df_train.head()
```

(1680209, 3)

```
[21]:
```

	text	label	source_dataset
0	My question is: Alani earned \$45 for 3 hours o...	0	jayavibhav
1	Q: Yes / no, is the following a verifiable fac...	0	jayavibhav
2	Read a code snippet and detect if there are an...	0	jayavibhav
3	As a highly advanced chatbot, it is your duty ...	1	jayavibhav
4	Tell me the main idea of this paragraph.	0	jayavibhav

```
[22]: df_test = pd.concat([
    df1_test,
    df3_test
], ignore_index=True)
```

```
print(df_test.shape)
df_test.head()
```

(329330, 3)

```
[22]:
```

	text	label	source_dataset
0	Task: Given a list of words, identify which wo...	0	jayavibhav
1	So I was watching episode 11 of Dark Angel sea...	0	jayavibhav
2	Given the sentence "Lots of people standing on...	0	jayavibhav
3	In this task, you are given the name of an Ind...	0	jayavibhav
4	Task: Replace all the sentences that use "i" w...	0	jayavibhav

```
[23]: df_train = df_train.sample(frac=1, random_state=42).reset_index(drop=True)
df_test = df_test.sample(frac=1, random_state=42).reset_index(drop=True)
```

```
[24]: print("TRAIN distribution:")
print(df_train['label'].value_counts())

print("\nTEST distribution:")
print(df_test['label'].value_counts())
```

TRAIN distribution:

label	count
0	964078
1	716131

```
Name: count, dtype: int64

TEST distribution:
label
0    167094
1    162236
Name: count, dtype: int64

[25]: df_train.to_csv("final_train_dataset.csv", index=False)
df_test.to_csv("final_test_dataset.csv", index=False)

print("Saved final_train_dataset.csv & final_test_dataset.csv")
```

Saved final_train_dataset.csv & final_test_dataset.csv

[]: