Assignment 4

Q1.1 Since the first prompt allows the LLM to respond completely freely, it doesn't always start with "yes" or "no". We had to add a regex that scans the entire answer for the first independent yes or no. So the mapping is not simple - it requires code matching, which doesn't always help and catches all cases true (sometimes false). What we can do is tell the LLM model to return the answer in a certain format with yes or no (next task)

Q1.2 Yes—3 / 30 labels flipped. Two flips were false alarms: our regex missed the yes/no buried in the free-text reply of Task 1.1, so those tweets were mis-tagged. Only one flip reflects a real change—when restricted to a single word the model finally committed to “yes.” Most of the difference, then, comes from extraction error, not from the LLM’s judgment.

Q2.1 No. With the five labeled examples in the prompt, 6 of 30 tweets changed label: recall rose to 1.00 while precision dropped to 0.67. The shift is expected—few-shot priming nudges the LLM toward the patterns it sees in the examples. Our seed set contained more explicit hate cases than neutral ones, so the model became more “sensitive,” flagging borderline tweets it had previously passed over (boosting recall) but also producing a few false positives (hurting precision). Re-running with a different example mix changes the balance again, confirming that the choice of tagged examples directly steers the model’s decision boundary.

Q2.2

```
Below are examples of tweets with their labels ('yes' = hate speech, 'no' = not hate speech).

Tweet: "<EX1_TEXT>"
Label: <EX1_LABEL>.
Tweet: "<EX2_TEXT>"
Label: <EX2_LABEL>.
Tweet: "<EX3_TEXT>"
Label: <EX3_LABEL>.
Tweet: "<EX4_TEXT>"
Label: <EX4_LABEL>.
Tweet: "<EX5_TEXT>"
Label: <EX5_LABEL>.

Now classify the next tweet in the same way.
Tweet: "[TWEET]"
Label:  Answer with exactly one word: 'yes' or 'no'.
```

Q3.1 Majority vs. ground-truth – Precision ≈ 0.59, Recall = 1.00 (no hateful tweet is missed).

Q3.2 Task 1.2 has higher precision (0.82) but lower recall (0.90).

Single-shot few-shot (Task 2) sits in between (0.67 / 1.00).

Q3.3

- **19 / 20 tweets** were **unanimous** (5-0).
- **0 / 20** were decided 4-1.
- **1 / 20** received a 3-2 split.

The model is remarkably consistent on this subset; self-consistency only changed the label once, but it guarantees perfect recall.