

# DamnFair

Thomas Zhang, Mingzhe Wang, Yolanda Ying

Pioneering a verifiable benchmark to safeguard against MLaaS  
scams and ensure the integrity of machine learning services

# Which text is generated by GPT-3.5 and which one is generated by GPT-4?

---

YO

Tell me about Istanbul.

Istanbul, historically Byzantium and Constantinople, is a dynamic city straddling Europe and Asia. It's Turkey's cultural and economic nucleus, featuring landmarks like Hagia Sophia and the Blue Mosque. The city seamlessly blends modern vitality with ancient allure.

Istanbul, historically Byzantium and Constantinople, is a dynamic city spanning Europe and Asia. It's Turkey's cultural and economic nucleus, featuring landmarks like Hagia Sophia and the Blue Mosque. The city's charm lies in its synthesis of modern vitality and ancient allure.

# Which text is generated by GPT-3.5 and which one is generated by GPT-4?

---

YO

Tell me about Istanbul.

Istanbul, historically Byzantium and Constantinople, is a dynamic city straddling Europe and Asia. It's Turkey's cultural and economic nucleus, featuring landmarks like Hagia Sophia and the Blue Mosque. The city seamlessly blends modern vitality with ancient allure.



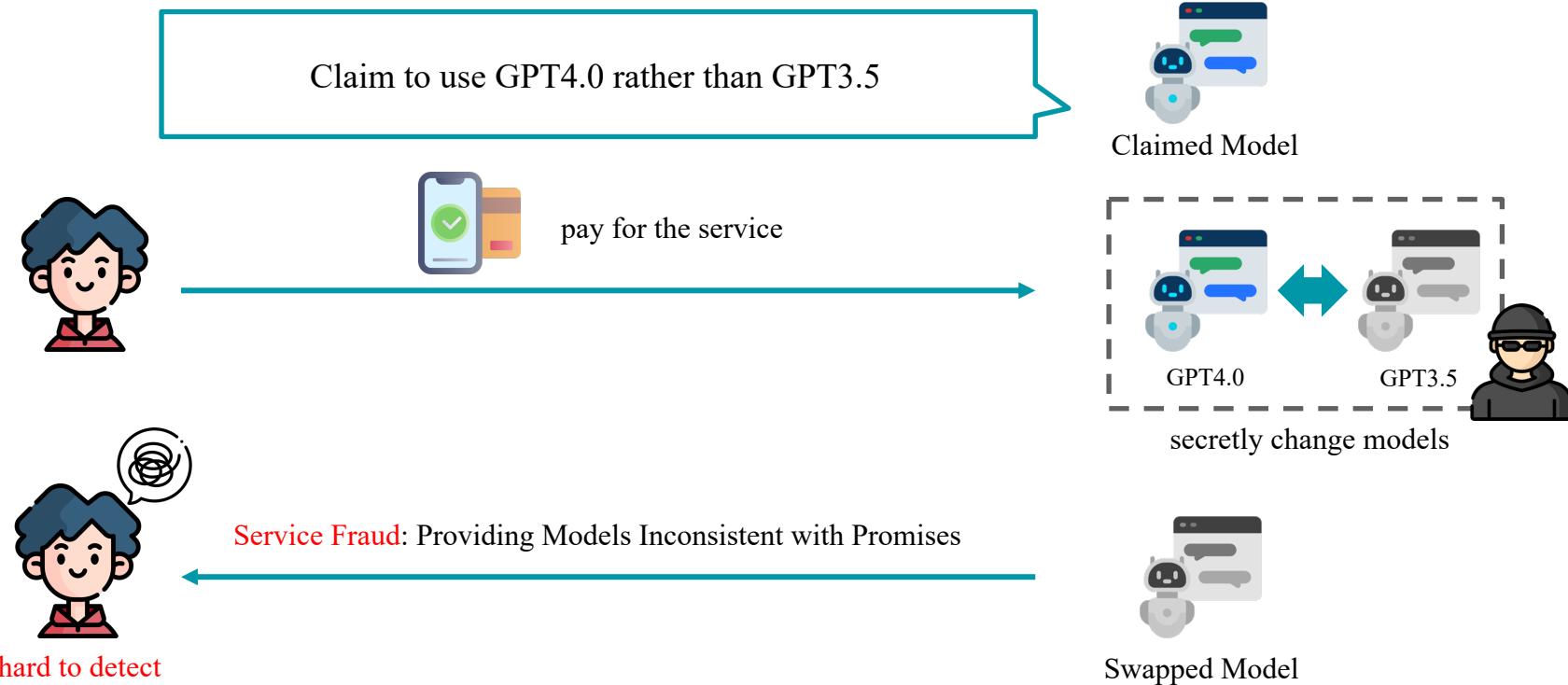
From GPT4

Istanbul, historically Byzantium and Constantinople, is a dynamic city spanning Europe and Asia. It's Turkey's cultural and economic nucleus, featuring landmarks like Hagia Sophia and the Blue Mosque. The city's charm lies in its synthesis of modern vitality and ancient allure.



From GPT3.5

# Use Case 1: Users Pay for ML Services, but Actual Operational Models Are Substituted



## Use Case 2: In the bidding and acceptance phases of ML, both ML service buyers and providers face specific pain points

	For ML Service Buyers	For ML Service Providers
Bidding Stage	Hard to discern	<b>Fraud:</b> inflating model performance data to win the bid
Acceptance Stage	<b>Fraud:</b> disputes, non-payment issues, and allegations of unexpected outputs	Hard to self-validation

# Our Solution

Our solution: We propose a zero-knowledge machine learning system enabling users to confirm ML model authenticity and protect data privacy

1

### Task Announcement



Announce the Task Publicly



2

### Claim Publication



Choose One  
of the Them

Public Claim Pool

Publish Claim of the Model



3

### Money Deposit



Deposit Money On-chain



Our solution: We propose a zero-knowledge machine learning system enabling users to confirm ML model authenticity and protect data privacy

4

Challenge



Commit Hash of the Model



VRF/VDF randomly choose  
indexes of the dataset

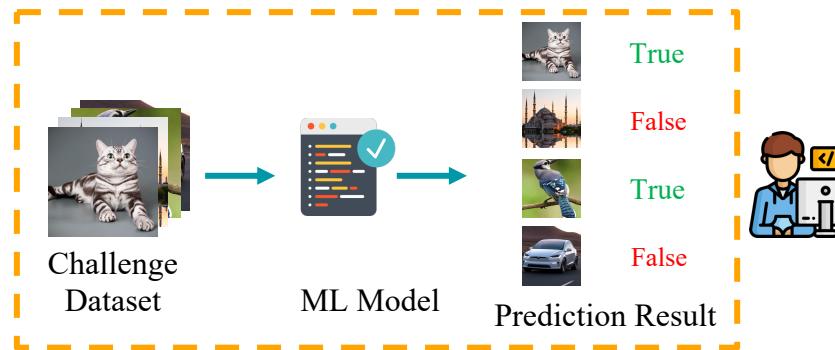


5

Response



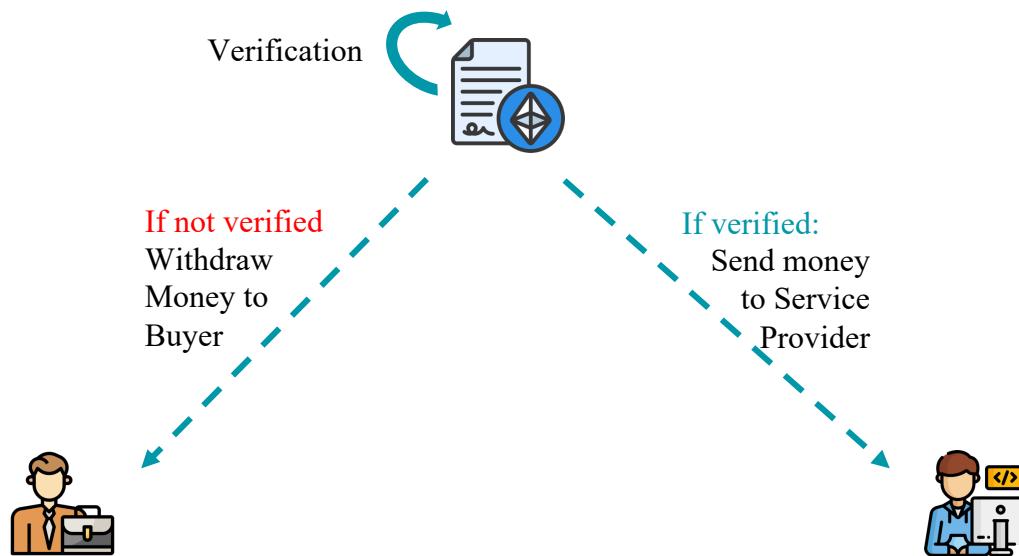
Proof of Inference



Our solution: We propose a zero-knowledge machine learning system enabling users to confirm ML model authenticity and protect data privacy

6

### Verification



# Demo

## Step 1 Task Announcement

2. Claim Publication

3. Money Deposit

4. Challenge

5. Response

6. Verification



**Task Name:** Animal Identifier

**Detail:** Train machine learning model to get an animal classification tool

- Example:



True



False



True



False



### Bob's Team

Model Used: NN model

- Quoted Price: **4500 usd**
- Overall Accuracy: **85%**
- Inference Time: **few ms**
- Inference Cost: **0.03 usd/1k times**
- Scalability<sup>1</sup>: **Multi-Classifier**
- Model size & Complexity: **50 mb**



### Alice's Team

Model Used: Unknown model

- Quoted Price: **3000 usd**
- Overall Accuracy: **85%**
- Inference Time: **few ms**
- Inference Cost: **0.001 usd/1k times**
- Scalability<sup>1</sup>: **Bi-Classifier**
- Model size & Complexity: **1 mb**

# Step 1 & Step 2

Task Announcement and  
Claim Publication

3. Money  
Deposit

4. Challenge

5. Response

6. Verification

## ML Task Dashboard

Reset

Publish New ML Task

Task #1

Task Type: CV

Number of Claims: 6

Dataset Name: Not provided

[View Details](#)

Task #2

Task Type: CV

Number of Claims: 2

Dataset Name: Not provided

[View Details](#)

Task #3

Task Type: NLP

Number of Claims: 0

Dataset Name: Not provided

[View Details](#)

## 1. Task Announcement

## 2. Claim Publication

## Step 3 Money Deposit

## 4. Challenge

## 5. Response

## 6. Verification

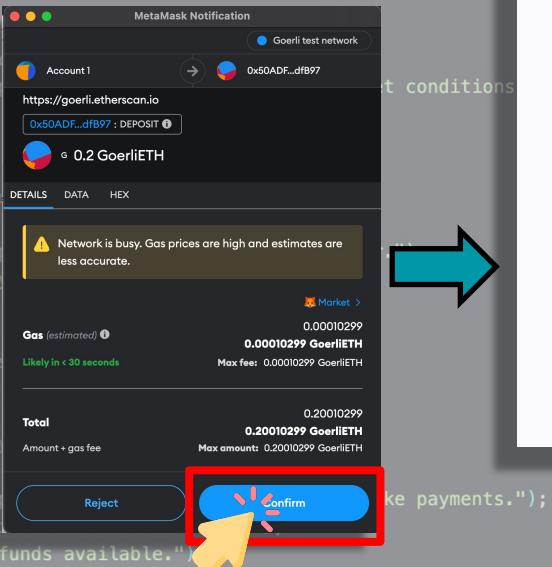
```
constructor() {
    depositor = msg.sender;
}

// deposit
function deposit() external payable {
    require(msg.sender == depositor, "You are not the depositor.");
    require(msg.value > 0, "Deposit amount should be greater than 0.");
    amount += msg.value;
}

// condition
function setCondition(bool _condition) external {
    require(msg.sender == depositor, "You are not the depositor.");
    condition = _condition;
}

// withdraw if they don't pass the condition
function withdraw() external {
    require(msg.sender == depositor, "You are not the depositor.");
    require(!condition, "The condition must be false to withdraw.");
    uint balance = amount;
    amount = 0;
    payable(msg.sender).transfer(balance);
}

// pay to the company if they pass the condition
function pay() external {
    require(msg.sender == depositor, "You are not the depositor.");
    require(condition, "Condition must be true to pay.");
    require(amount > 0, "No funds available.");
    uint amountToSend = amount;
    amount = 0;
}
```



The MANTLE Testnet interface displays the "Contract Address Details" for the deposited GoerliETH. The contract address is 0x50ADF50AF0F5f90bA892BD51F947ca252B6dfB97. Key details include:

Contract Name	PaymentContract
Creator	0x2cd60a-2e0b6f at 0xa0030c-3f770b
Balance	10 MNT (\$4.45 USD)
Tokens	0 tokens
Transactions	0
Transfers	0
Gas Used	1,152,533
Last Balance Update	25709303

Below the details, the "Transactions" tab is selected, showing a single deposit transaction:

Txn Hash	Txn Type	Method	Block	Age	From	To	Value	Txn Fee
0x0de3d1f90b...	Contract Call	Deposit	25709303	29 seconds ago	0x2cd60a-2e0b6f	PaymentContract L...	10 MNT	0.0000022018...

Contract address:

<https://explorer.testnet.mantle.xyz/address/0x50ADF50AF0F5f90bA892BD51F947ca252B6dfB97>

1. Task  
Announcement

2. Claim  
Publication

3. Money  
Deposit

## Step 4 & Step 5 Challenge and Response

6. Verification

**Wait! So...What is proof of inference at all?**

“

*we concentrate on zero-knowledge proofs of training that preserve the confidentiality of the model and the associated dataset, which we call zero-knowledge proof of training (zkPoT).*

”

Citation: Garg, Sanjam, et al. "Experimenting with Zero-Knowledge Proofs of Training." *Cryptology ePrint Archive* (2023).

1. Task  
Announcement

2. Claim  
Publication

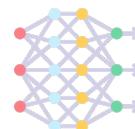
3. Money  
Deposit

## Step 4 & Step 5 Challenge and Response

6. Verification

Wait! So...What is proof of inference at all?

Training  
Process



Training Dataset

Trained Model

Inference  
Process



True/False  
True/False  
True/False  
True/False

Test IMG

Trained Model

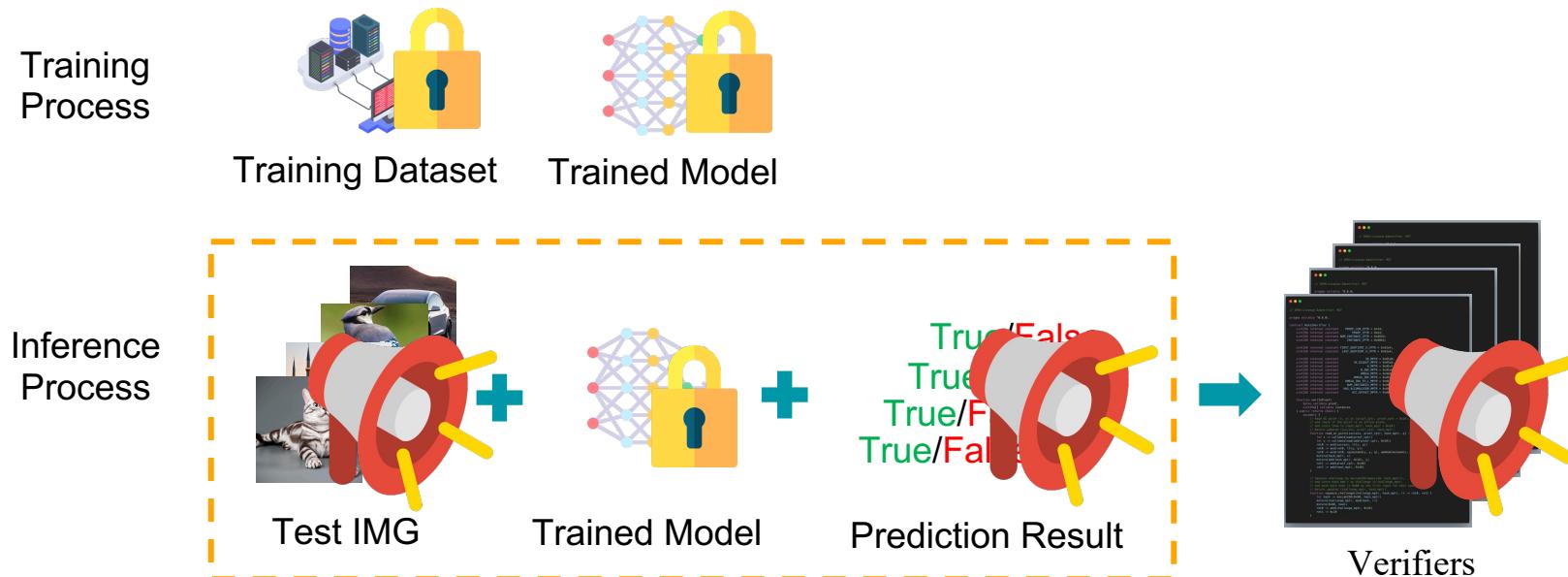
Prediction Result



Verifiers

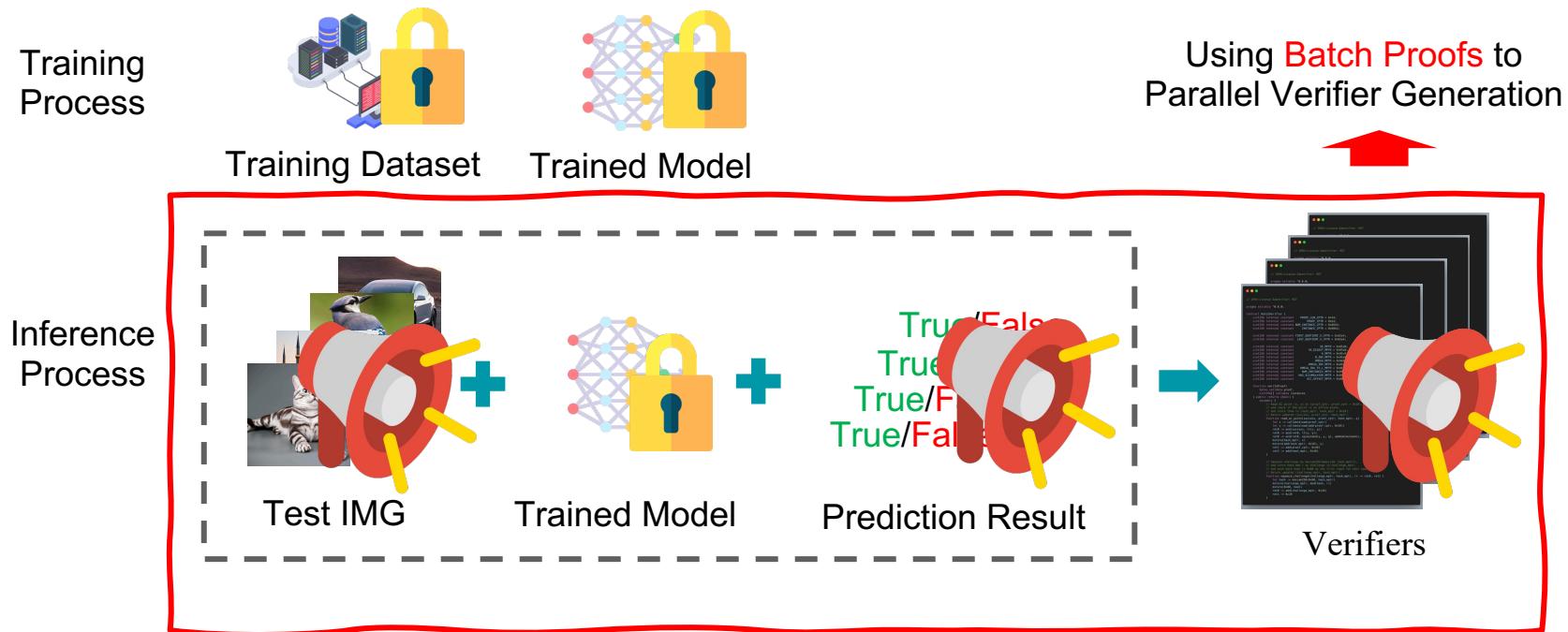


**Wait! So...What is proof of inference at all?**





**Wait! So...What is proof of inference at all?**



1. Task  
Announcement

2. Claim  
Publication

3. Money  
Deposit

## Step 4 & Step 5

### Challenge and Response

6. Verification

Simple Verifier 1\_24\_0.ipynb +

colab.research.google.com/drive/1lZLkm3ZjPHA9nh2CvTDto8Vbx0KF7VMz

统一身份认证系统... Google 学术搜索 Web of Science ... Engineering Village Springer Shop Home | Taylor & Fr... ML 2020 Spring MetaMask TESTNET Kovan (... All Bookmarks

Simple Verifier 1\_24\_0.ipynb

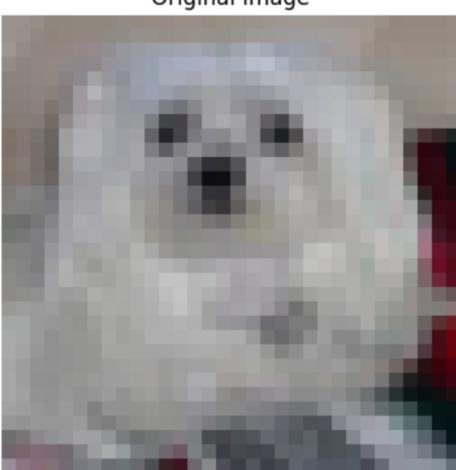
File Edit View Insert Runtime Tools Help Last saved at 2:56 AM

Comment Share Connect High-RAM

+ Code + Text

```
compare_images(imgs[5], x_test[indices[5]], 32, inv_normalize)
```

Original Image



Grayscale Normalized Image



1. Task  
Announcement

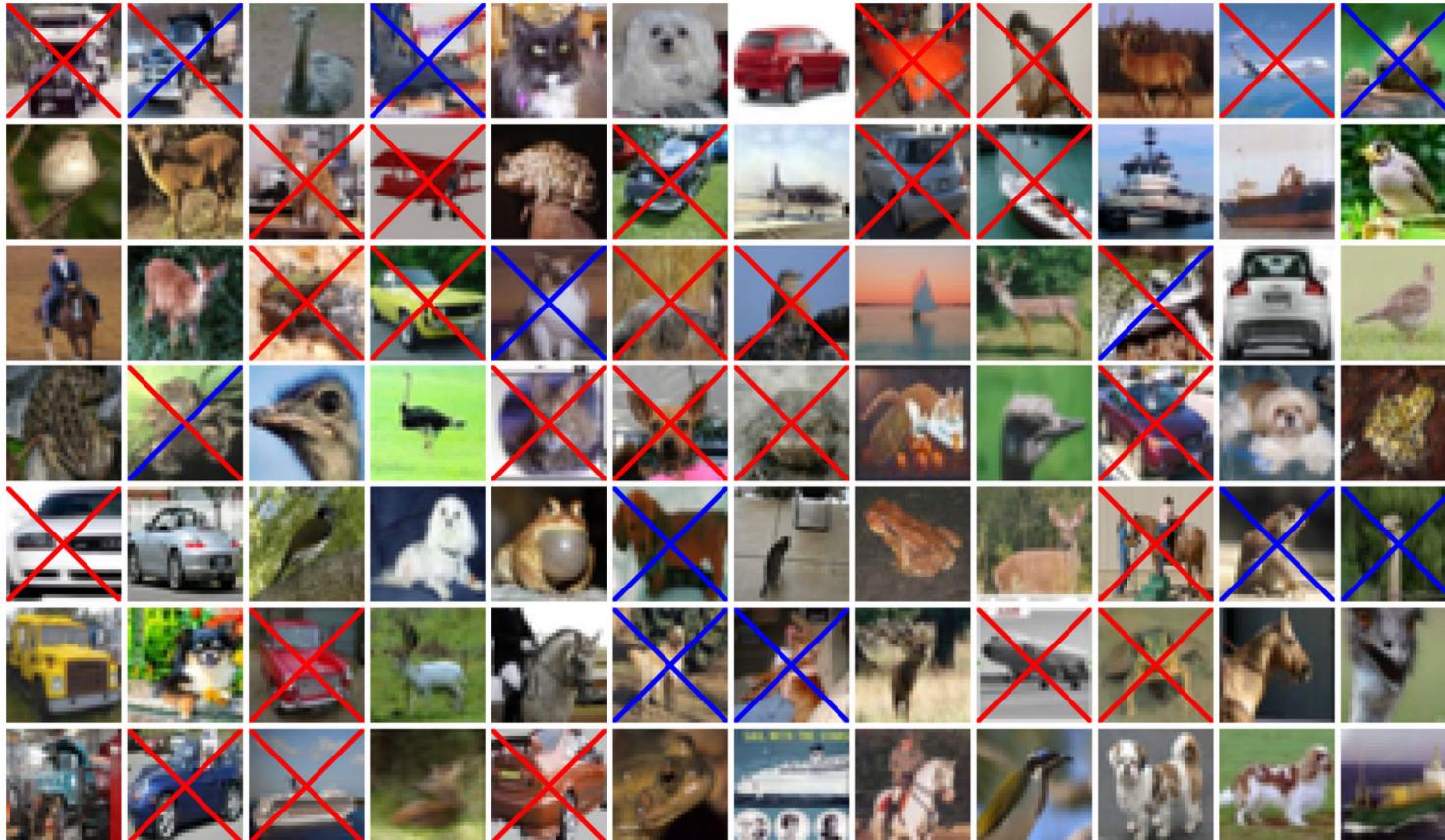
2. Claim  
Publication

3. Money  
Deposit

4. Challenge

5. Response

Step 6  
Verification



Alice's Team

Actual Accuracy

66.67%

Claim Accuracy

85%



Bob's Team

Actual Accuracy

86.90%

Claim Accuracy

85%

Thank you for listening =)