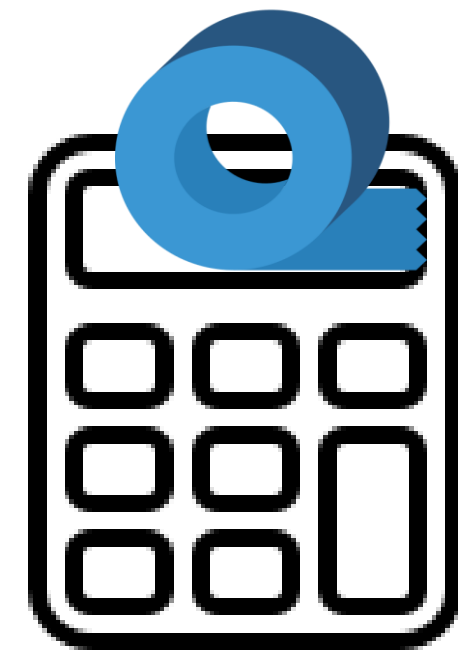


Motivation: Privacy Preserving ML using FHE

Machine Learning as a Service allows users to send personal data to the cloud for processing. However, the data is revealed to the cloud service provider.

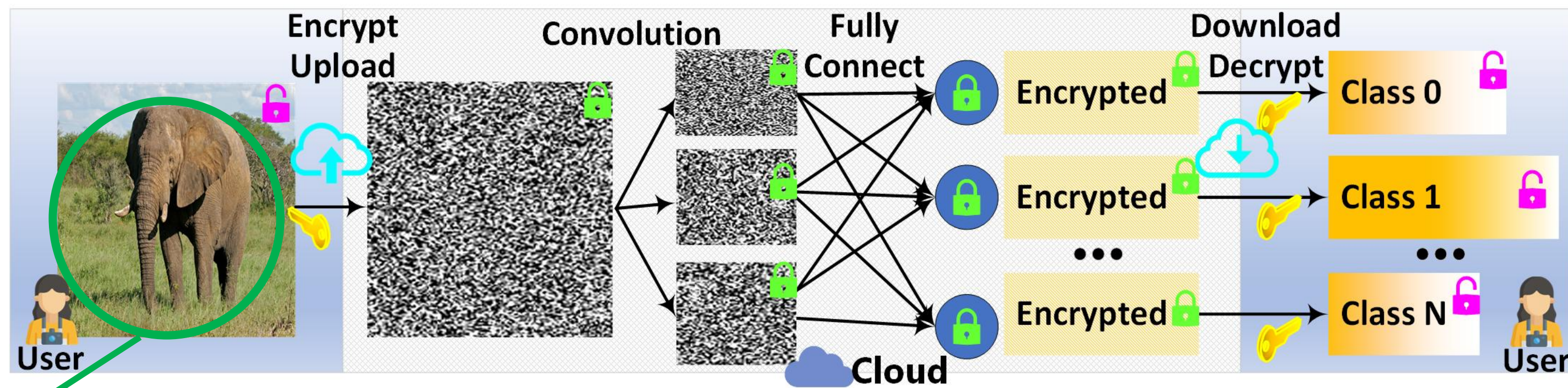
Threat Model: An honest but curious cloud will attempt to leak user data.



Fully homomorphic encryption (FHE) allows for private AI:

- Quickly evolving branch of cryptography
- Allows computations on encrypted data
- **Drawbacks:** computationally expensive, low precision

REDsec: Use of Ternary Weights



REDsec 1st work to run FHE AlexNet on ImageNet in under two hours

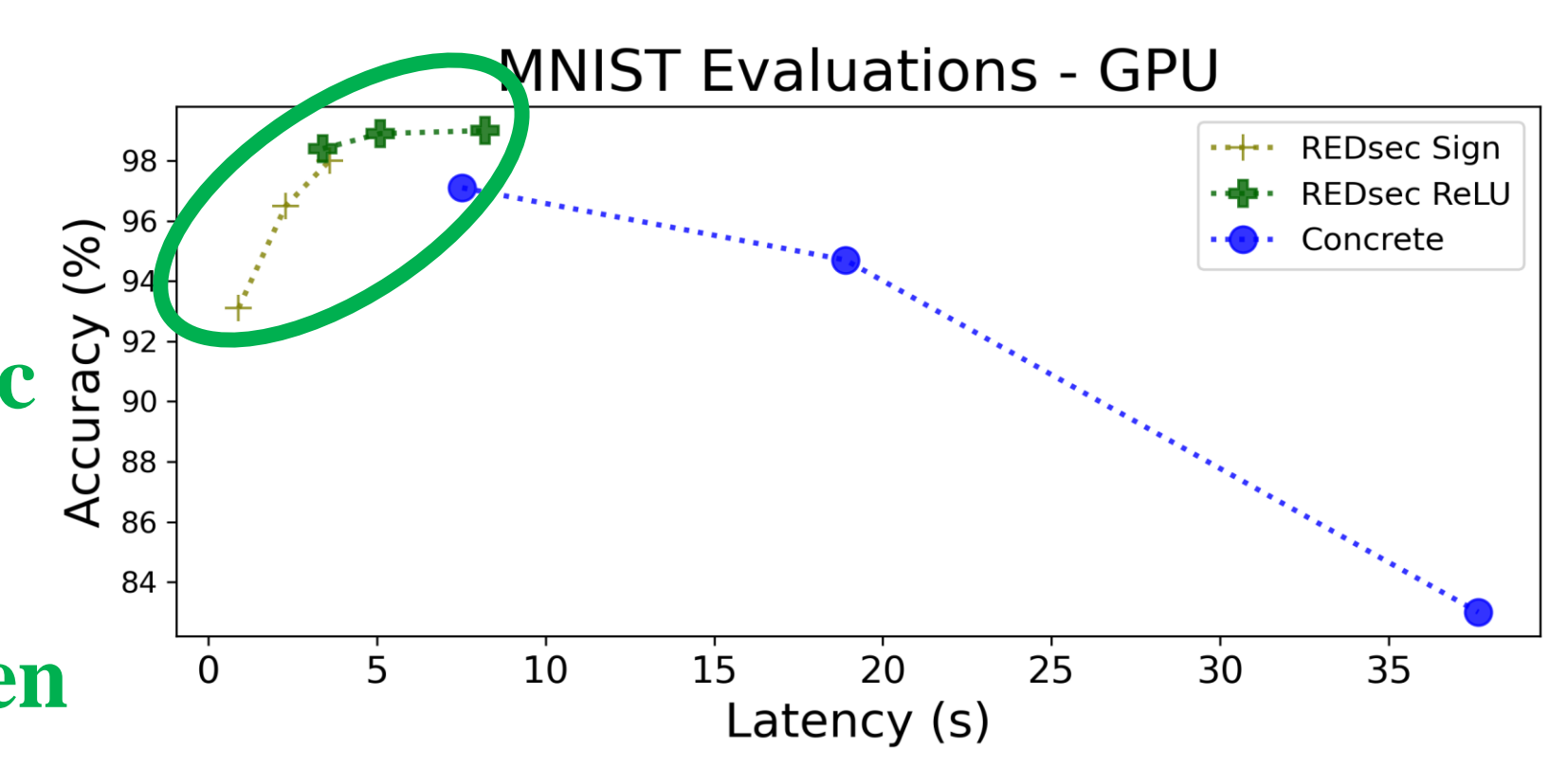
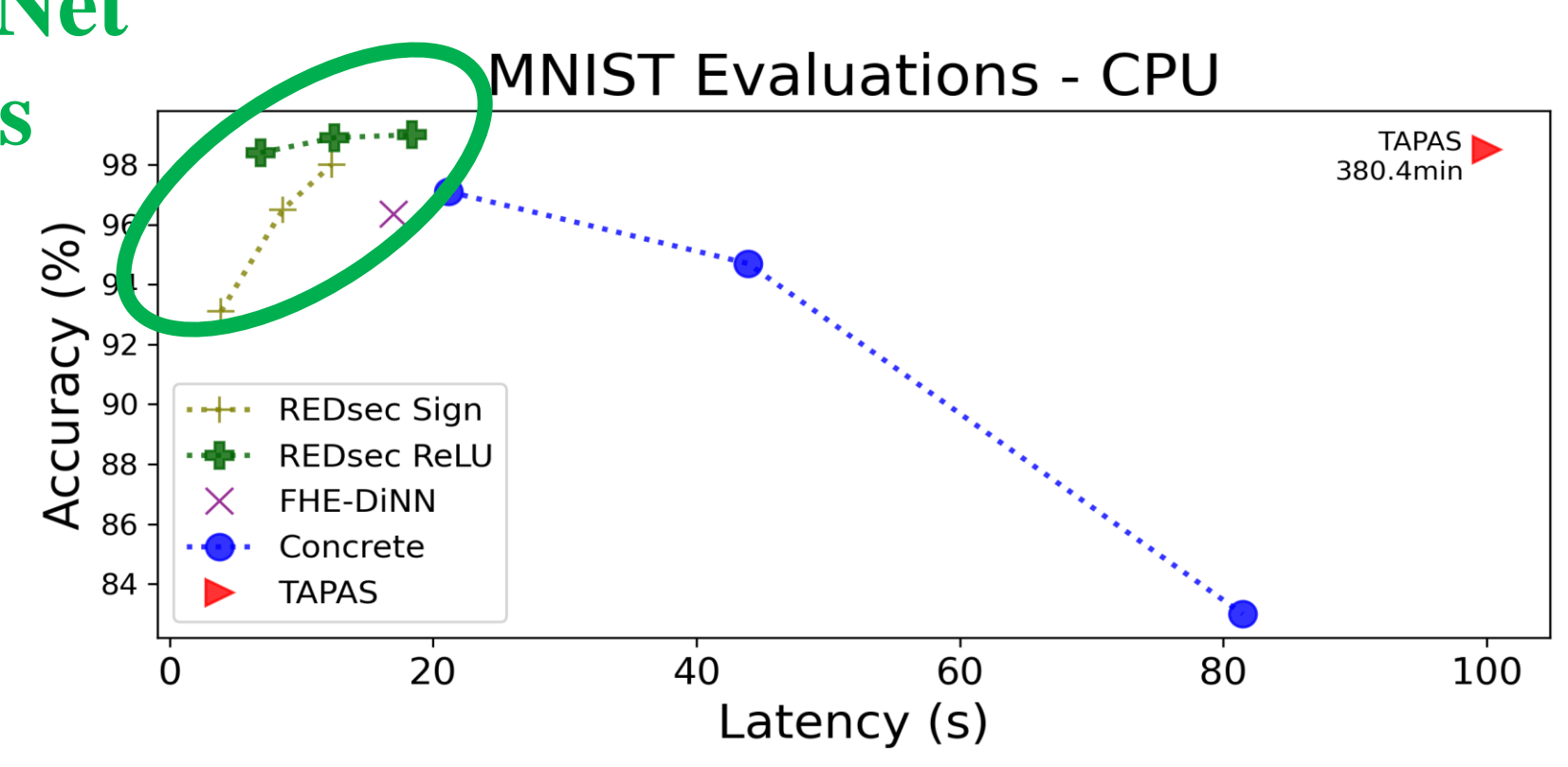
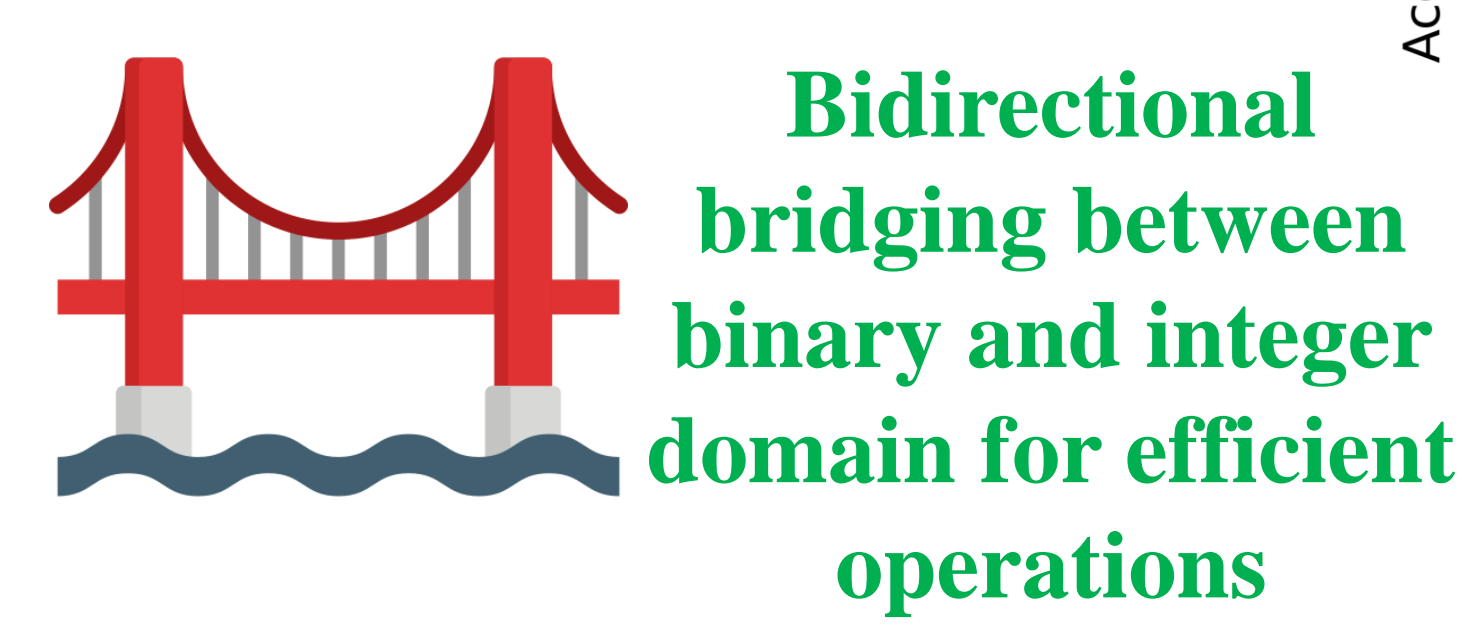
Ternary Multiply

A	B	Out
-1	-1	1
-1	1	-1
1	-1	-1
1	1	1

XNOR Gate

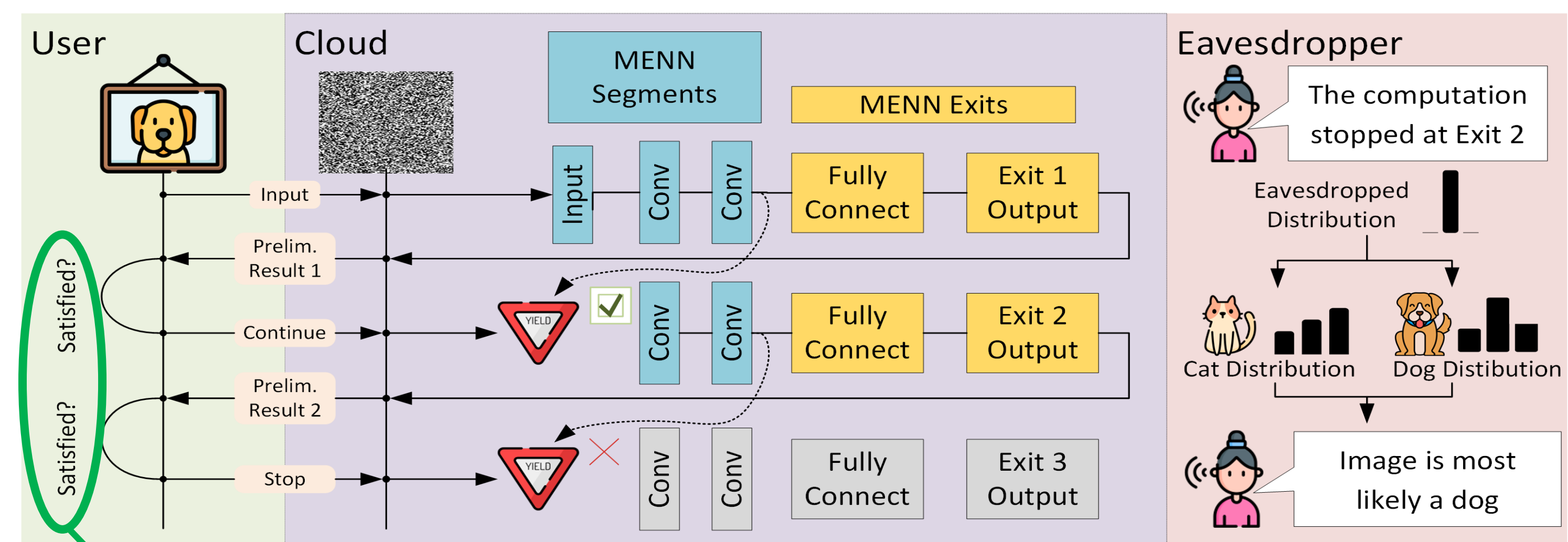
A	B	Out
0	0	1
0	1	0
1	0	0
1	1	1

Use binary multiply to be compatible with TFHE gate logic

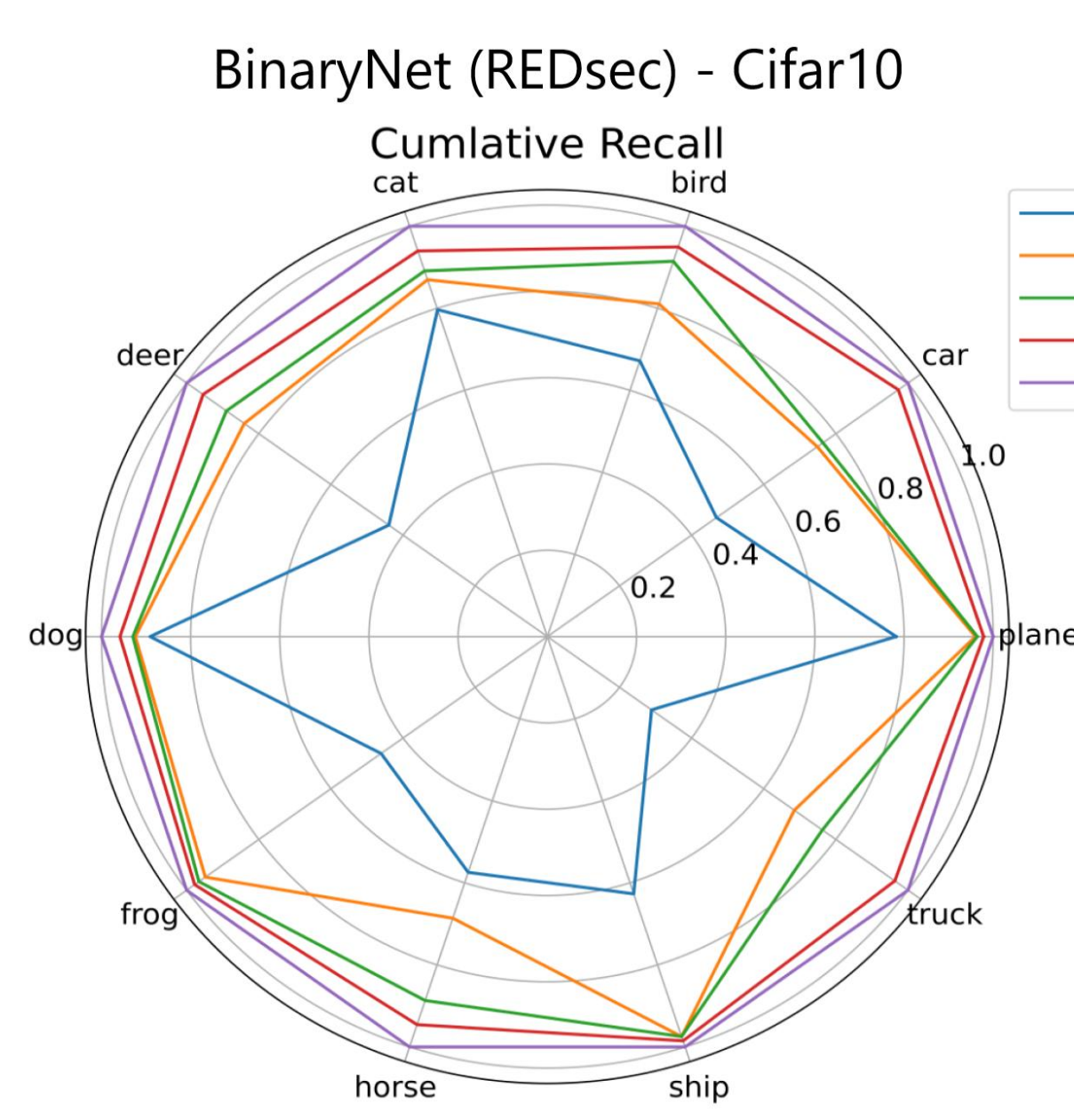


Higher accuracy and lower latency than related works

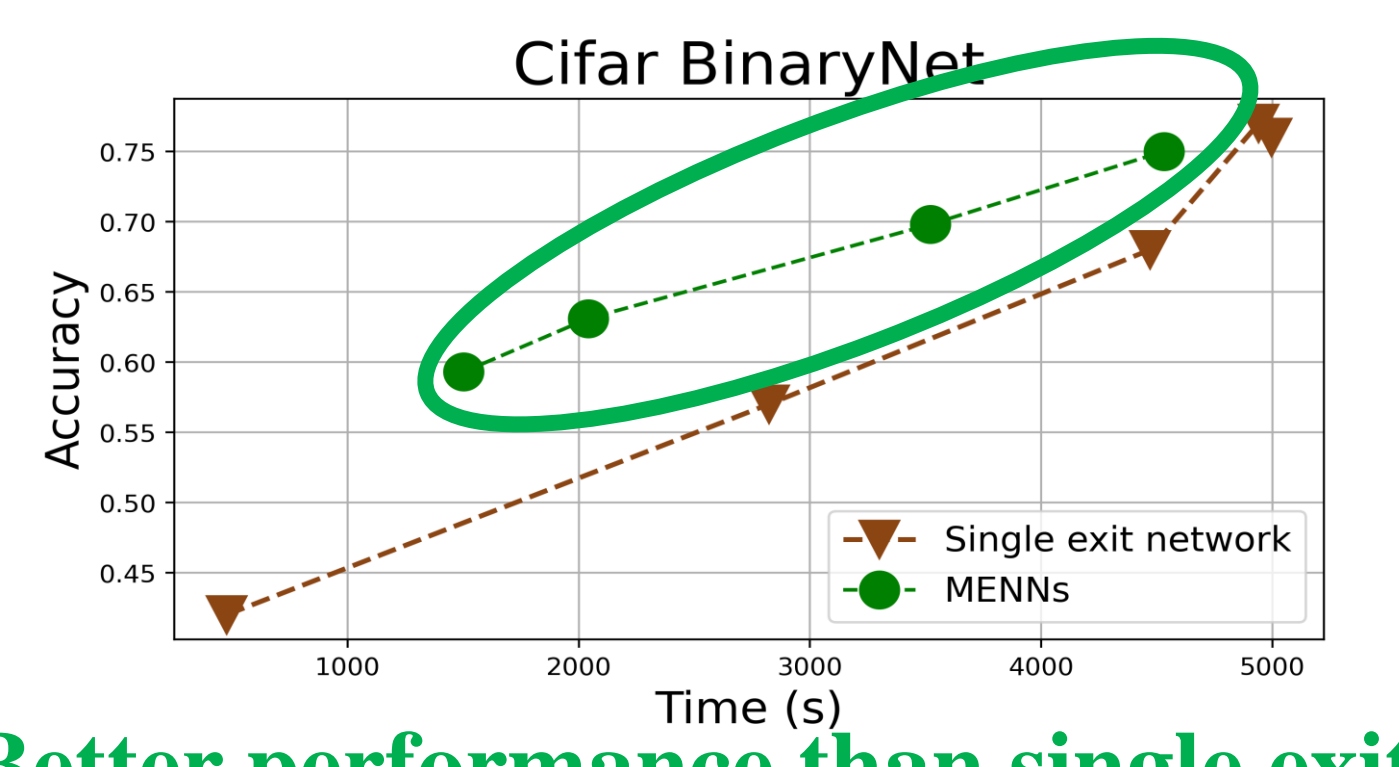
FHE-MENNs: Multi Exit Neural Networks



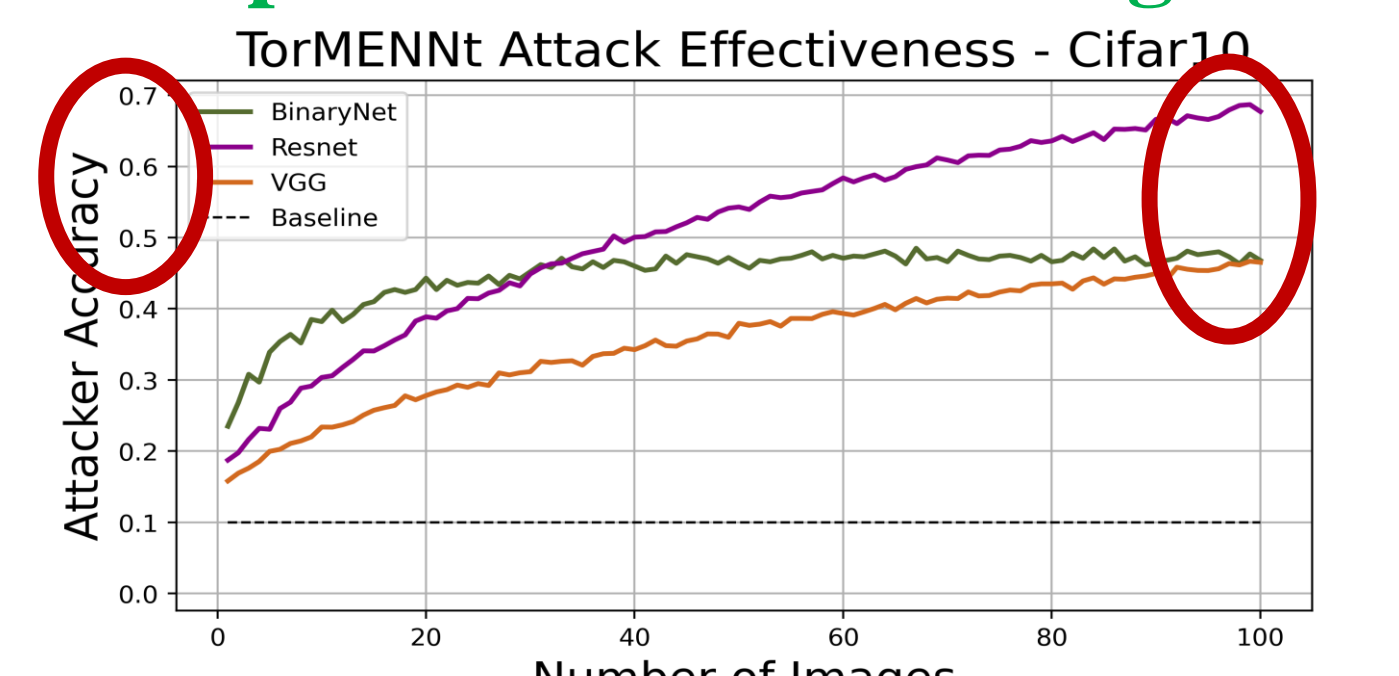
Option to terminate computation early, ... but this leaks information



Some classes are easier to classify and exit earlier than others

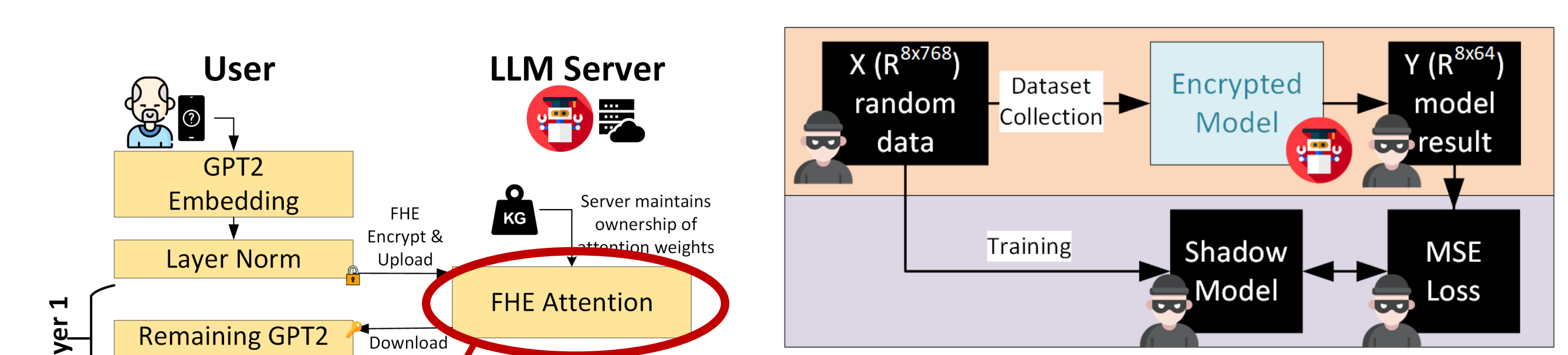


Better performance than single exit



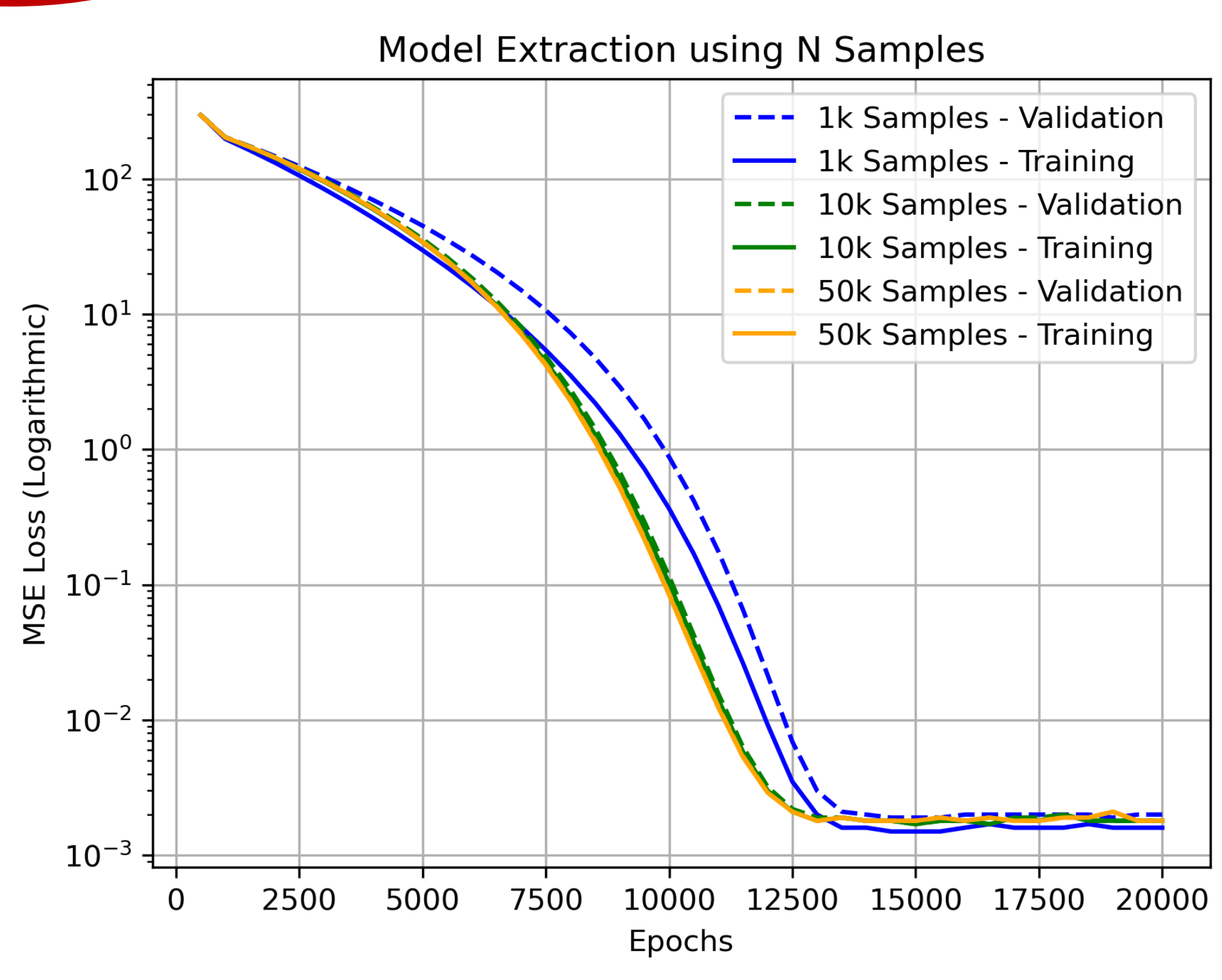
Leaks classification with high confidence

GPT-Thief: Attacks on FHE Split Model LLMs



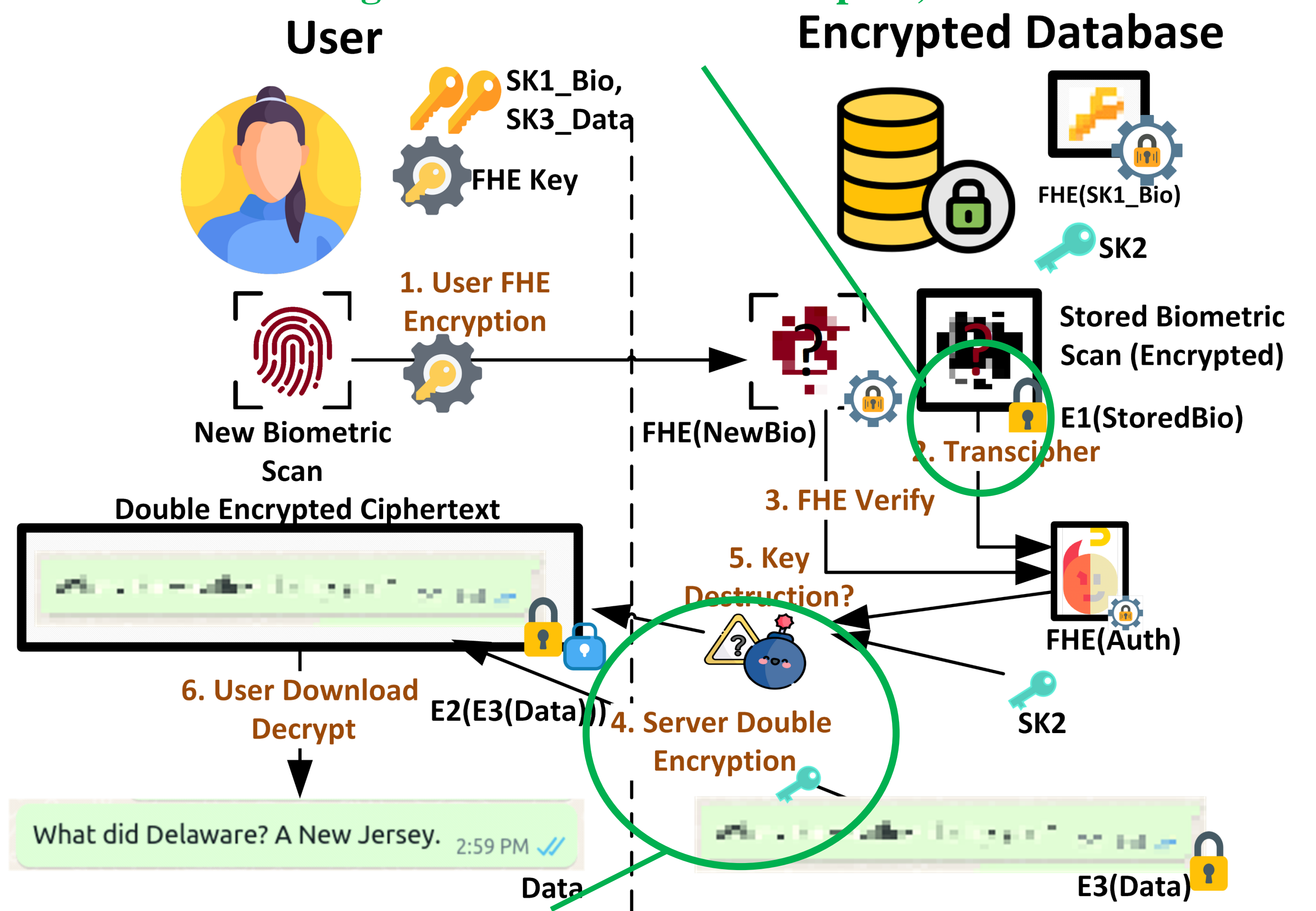
ConcreteML's Encrypted GPT-2, a single attention head is encrypted to balance privacy and efficiency

...but a user can steal the model



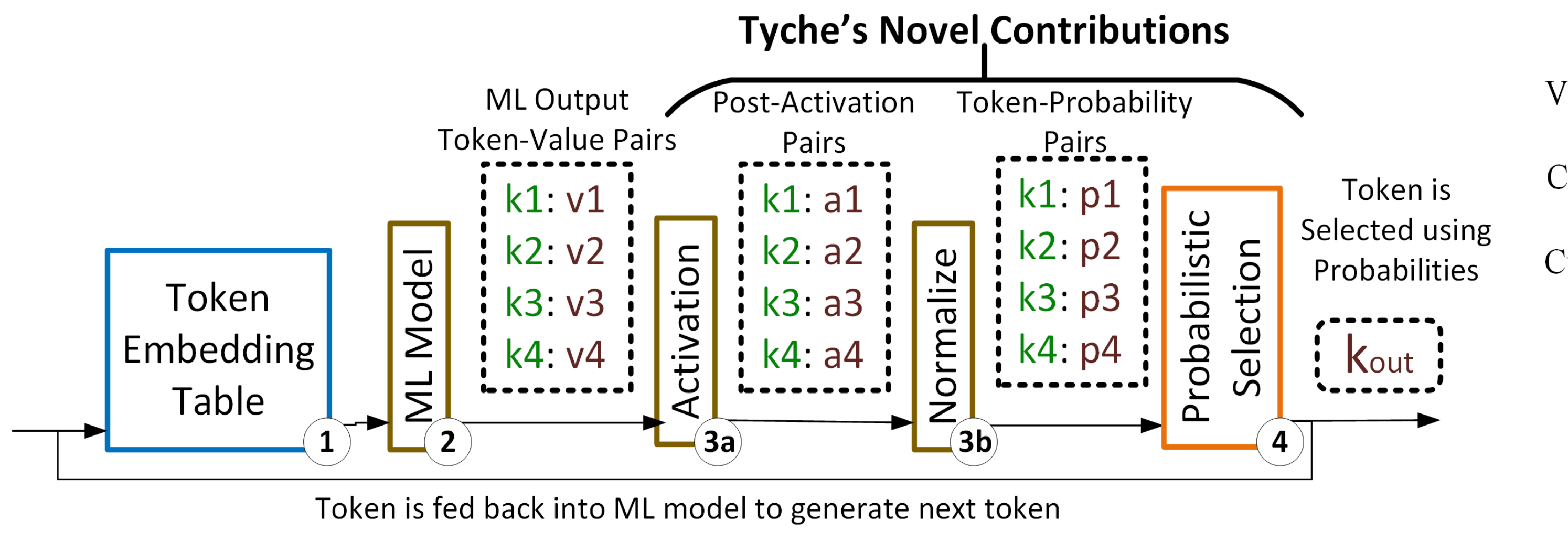
Proteus: Encrypted Access Control using FHE

Store data using authenticated ASCON cipher, and convert to FHE



Apply outer layer of encryption, and only release key if data is homomorphically verified

Tyche: Algorithms for Generative AI



Cumulative Sum Strategy

Values:

1	5	2	2
---	---	---	---

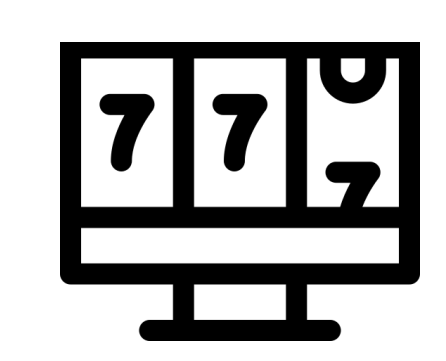
Cumulative:

1	6	8	10
---	---	---	----

Cum < Rideal:

1	1	0	0
---	---	---	---

Rideal: 7 Sum: 2



Random Multiply Strategy Multiply values by random vector, find maximum More scalable

