GuaranTEE: Towards private and attestable ML with CCA

Sandra Siby, Sina Abdollahi, Mohammad Maheri, Marios Kogias, Hamed Haddadi

EuroMLSys, 22 April 2024

IMPERIAL

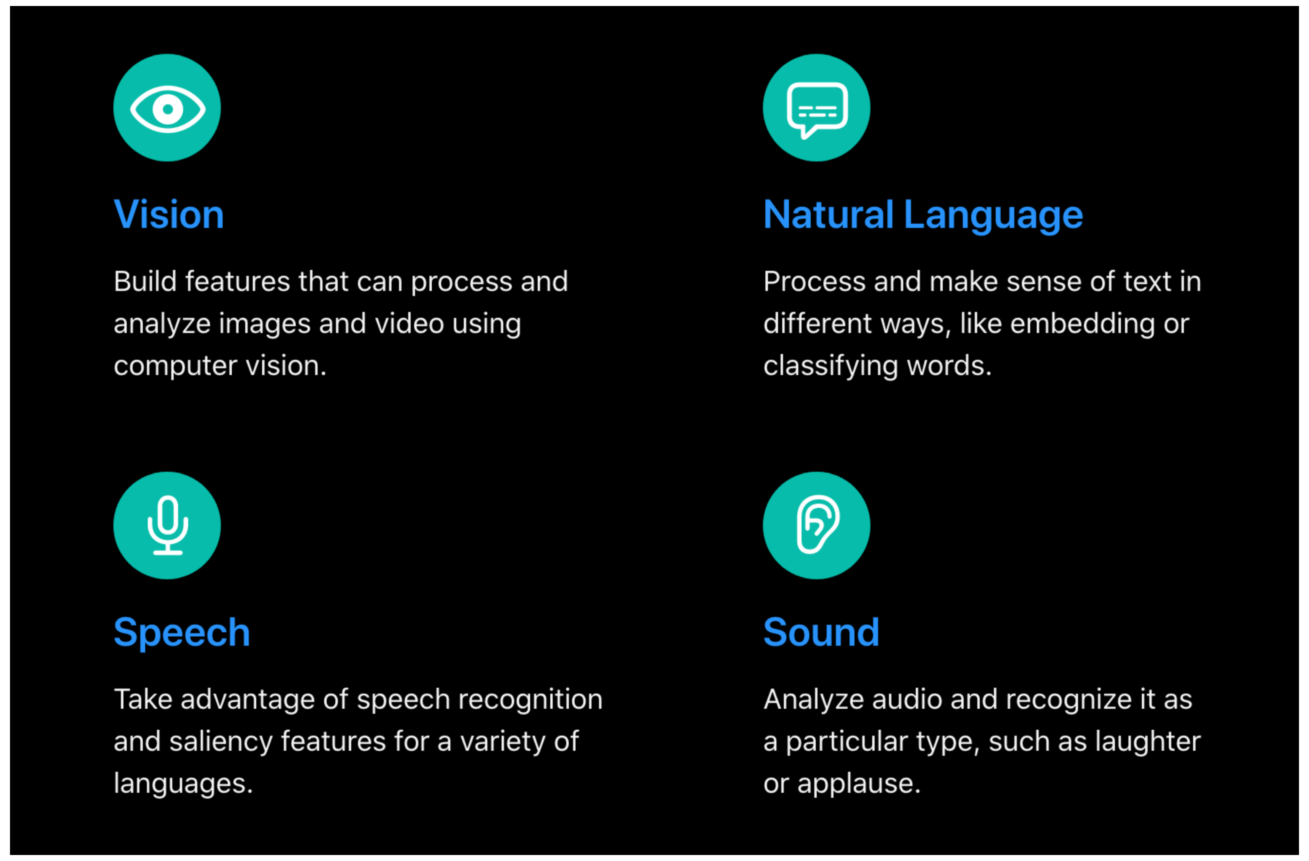
On-device Machine Learning

Benefits

No user data sent to cloud

Latency improvements

Private personalisation



On-device Machine Learning

Benefits

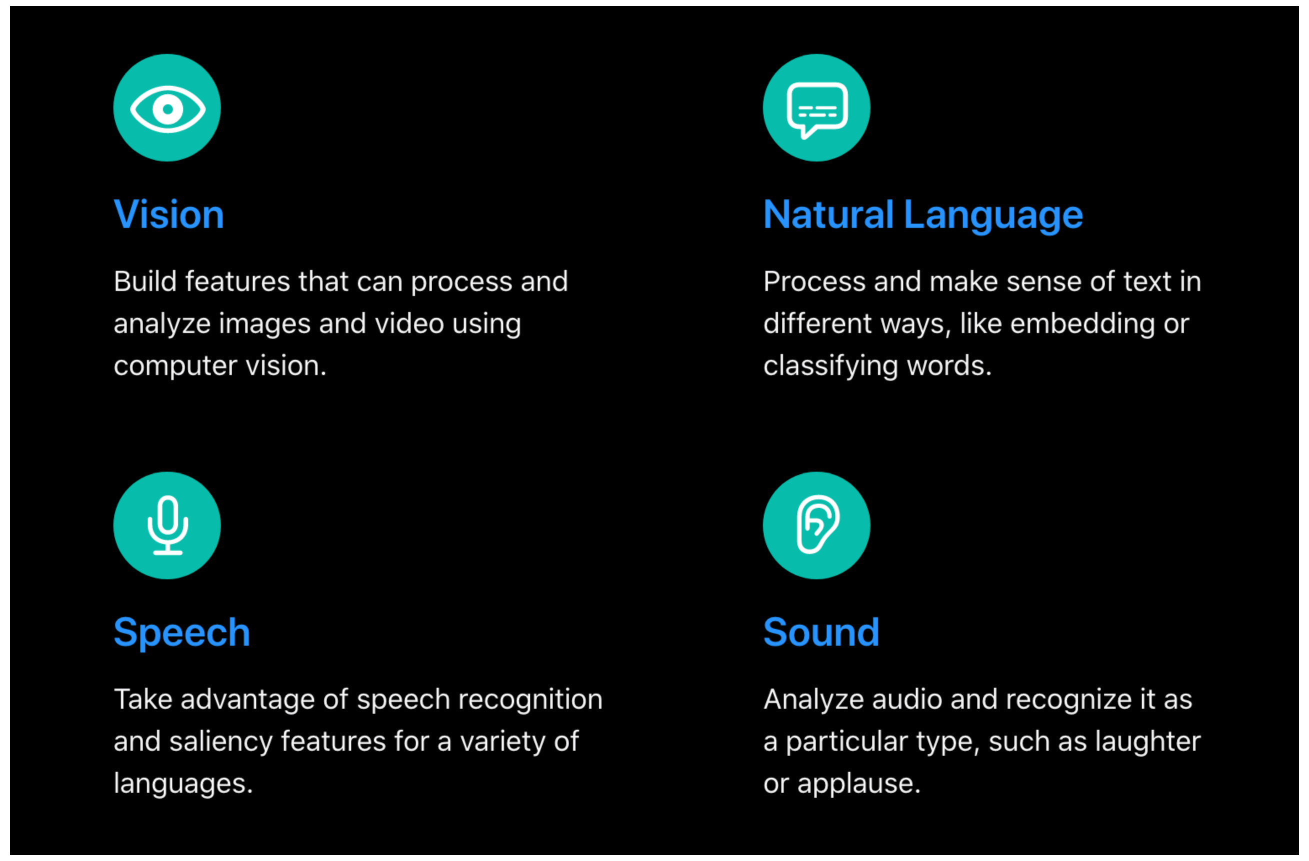
No user data sent to cloud

Latency improvements

Private personalisation

Model providers want:

- Model privacy
- Model verifiability and attestability



Protecting ML models

Existing solutions

Watermarking

- Detection rather than prevention
- Evasion attacks

Cryptography-based

- Computational and communication overheads

Hardware-assisted

Protecting ML models

Existing solutions

Watermarking

- Detection rather than prevention
- Evasion attacks

Cryptography-based

- Computational and communication overheads

Hardware-assisted

Applications run within Trusted Execution Environments (TEE)

Protecting ML models

Existing solutions

Watermarking

- Detection rather than prevention
- Evasion attacks

Cryptography-based

- Computational and communication overheads

Hardware-assisted

- Mainly tailored to the cloud
- Memory limitations on edge

Protecting ML models

Existing solutions

Hardware-assisted

- Mainly tailored to the cloud
- Memory limitations on edge

Arm’s TEE solutions

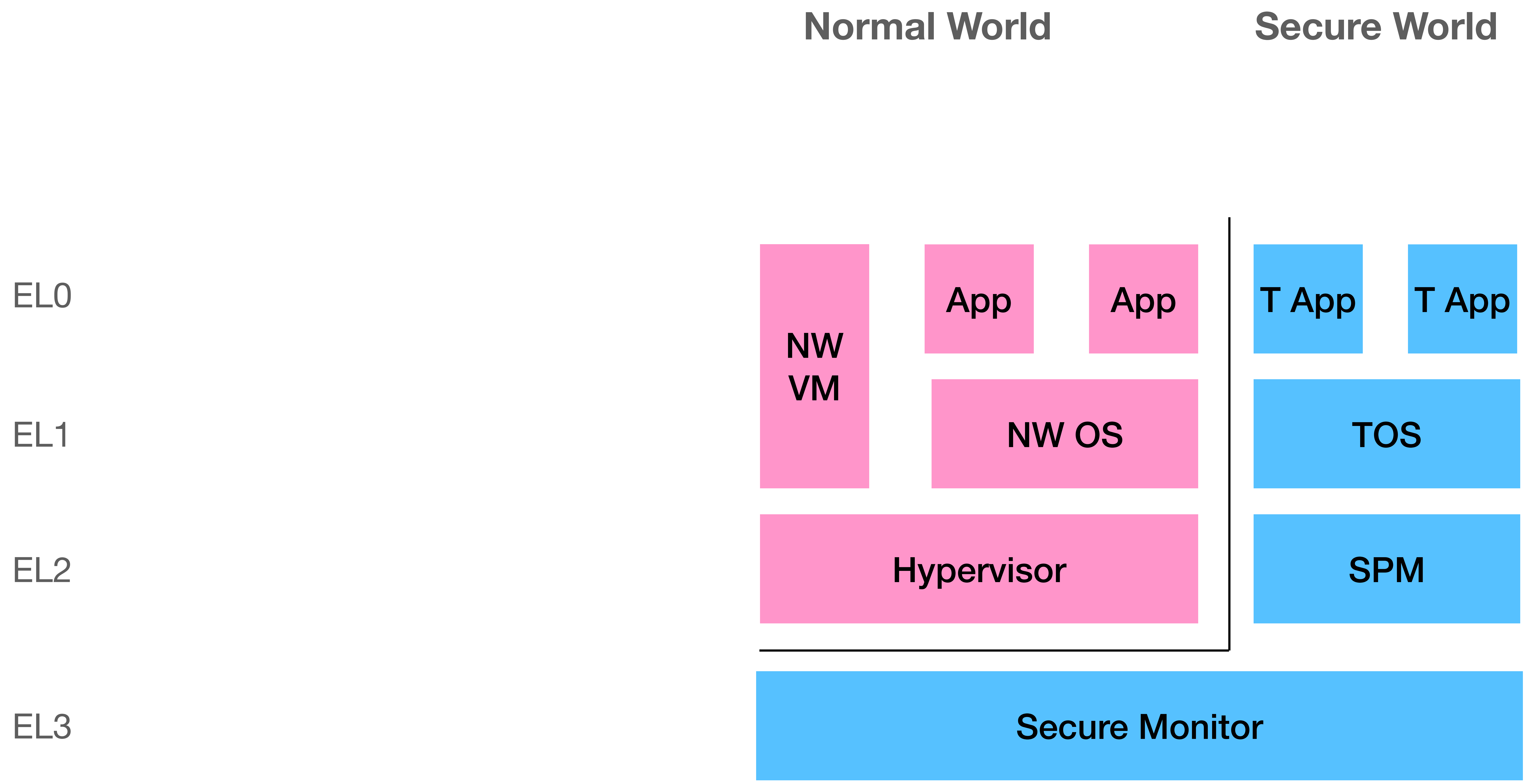
Arm’s TrustZone is widely deployed on edge devices.

We consider Arm’s next generation of TEE solutions (deployment expected in 2028):

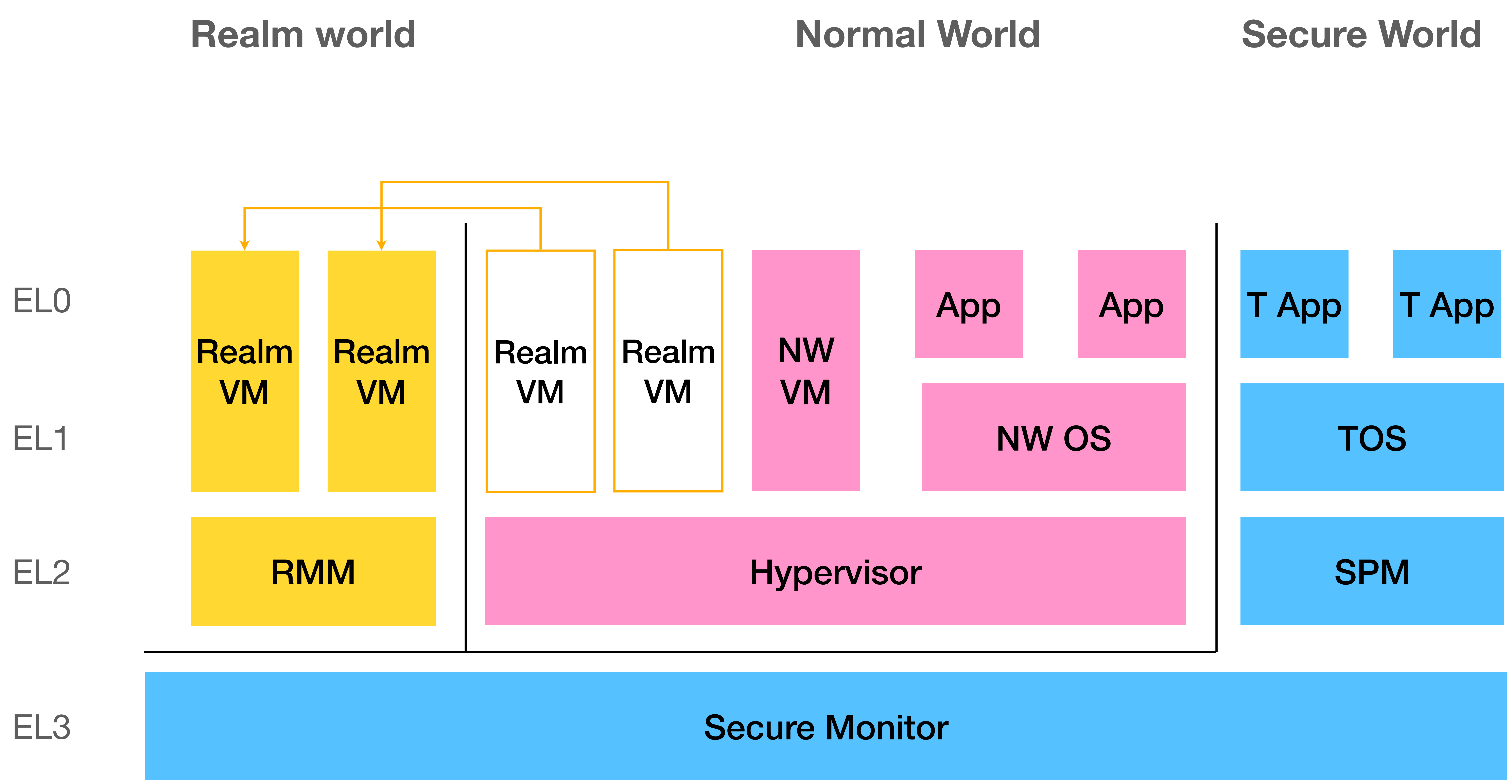
Confidential Computing Architecture (CCA)

…tion overheads

Arm TrustZone

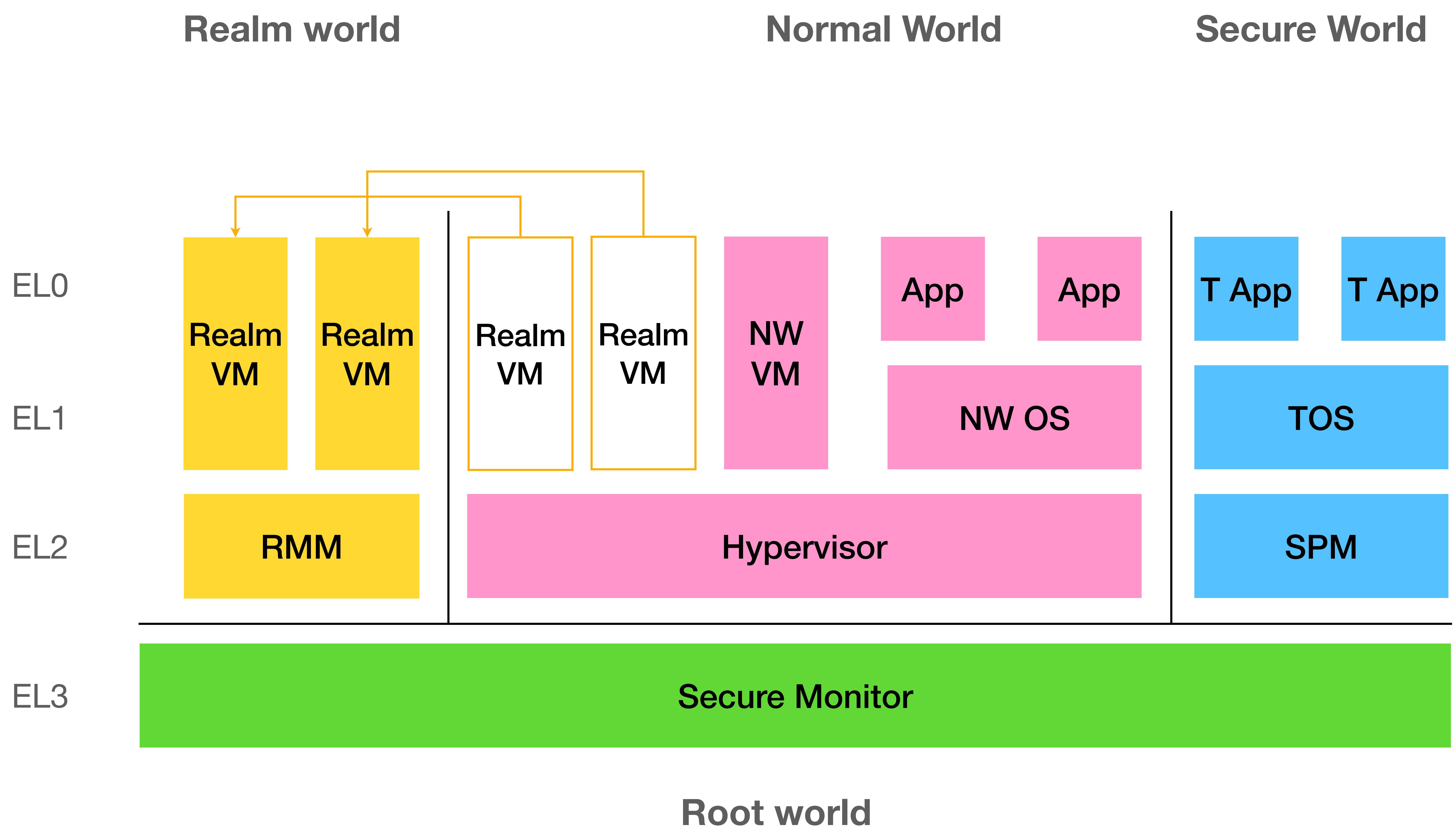


Arm CCA



Realm world
Normal World
Secure World
EL0
EL1
EL2
EL3
Realm VM
Realm VM
Realm VM
Realm VM
NW VM
App
App
T App
T App
NW OS
TOS
RMM
Hypervisor
SPM
Secure Monitor

Arm CCA


Realm world
Normal World
Secure World
EL0
EL1
EL2
EL3
Realm VM
Realm VM
Realm VM
Realm VM
NW VM
App
App
T App
T App
NW OS
TOS
RMM
Hypervisor
SPM
Secure Monitor
Root world

CCA and ML deployment

Why is CCA a promising choice for ML deployment?

Flexible memory allocation

General-purpose development

Protection against compromised hypervisor

CCA and ML deployment

Why is CCA a promising choice for ML deployment?

Flexible memory allocation

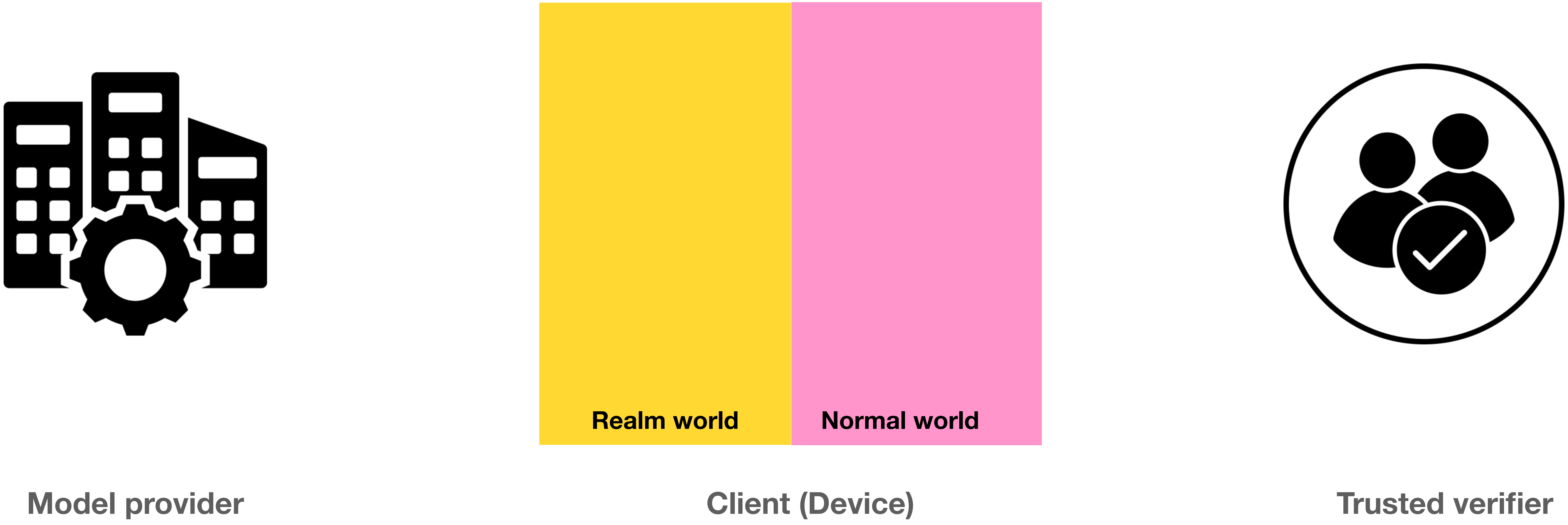
General-purpose development

Protection against compromised hypervisor

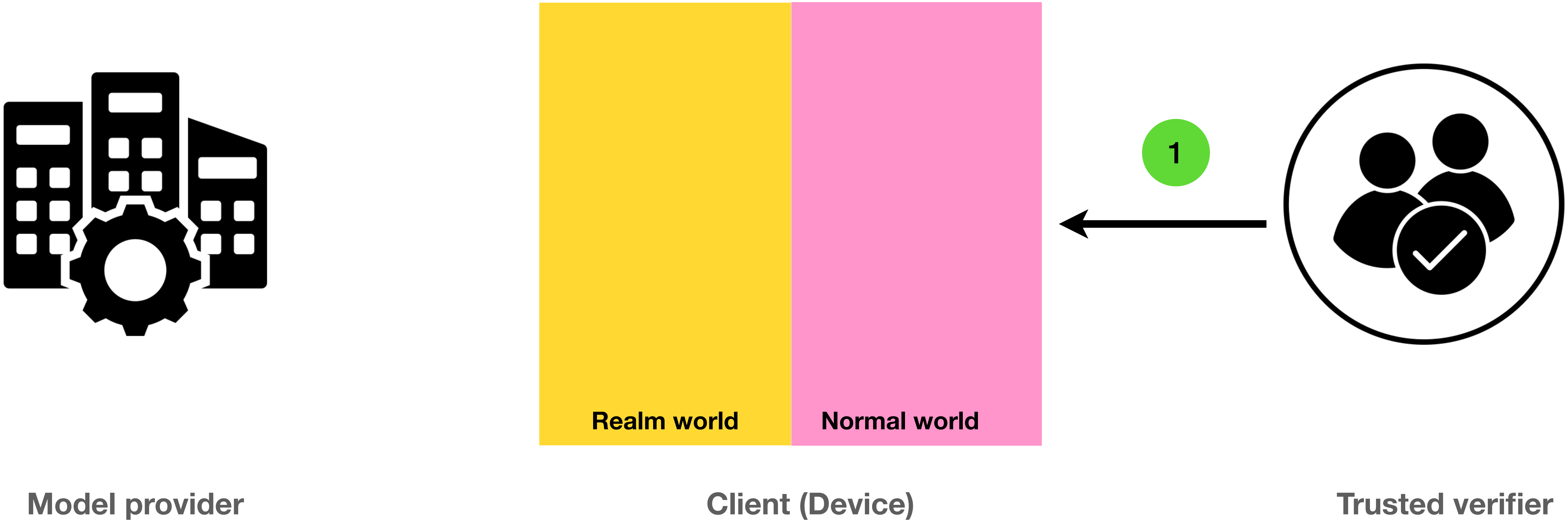
GuaranTEE

Framework for ML models to be run on end devices in a private and verifiable manner

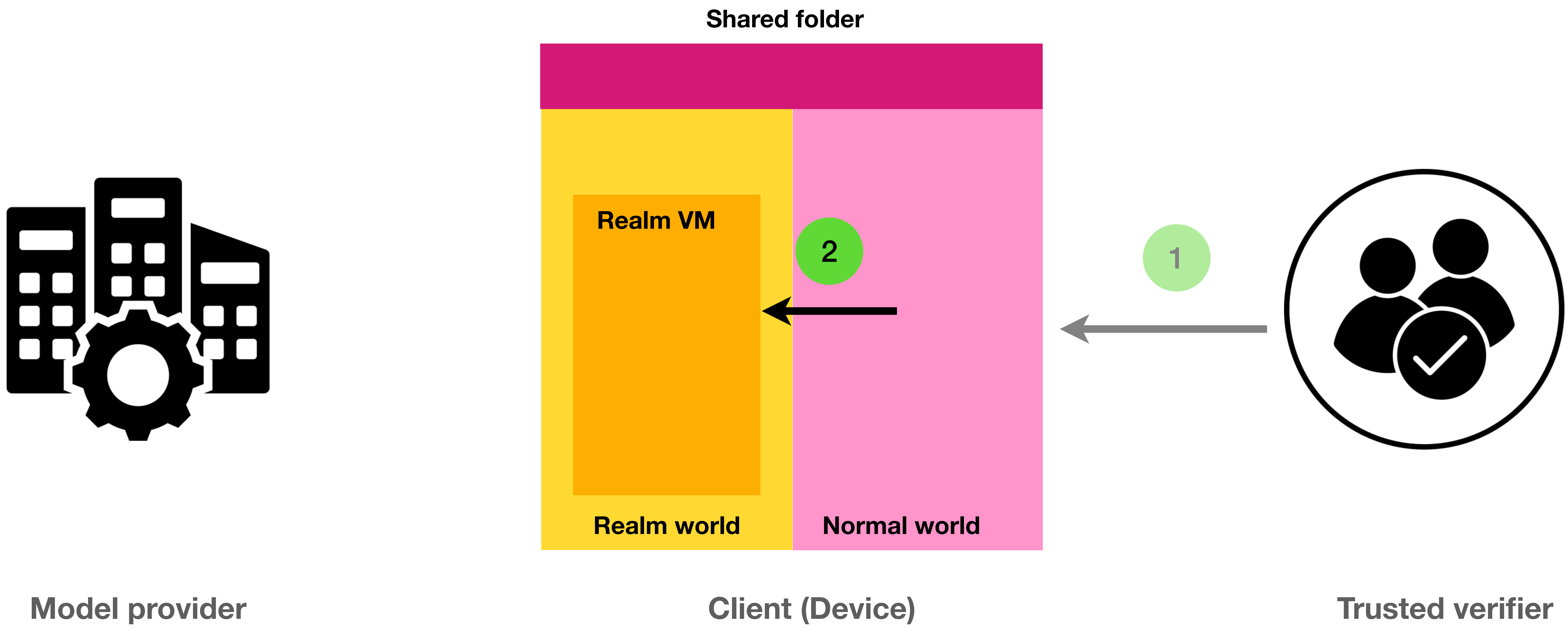
System overview



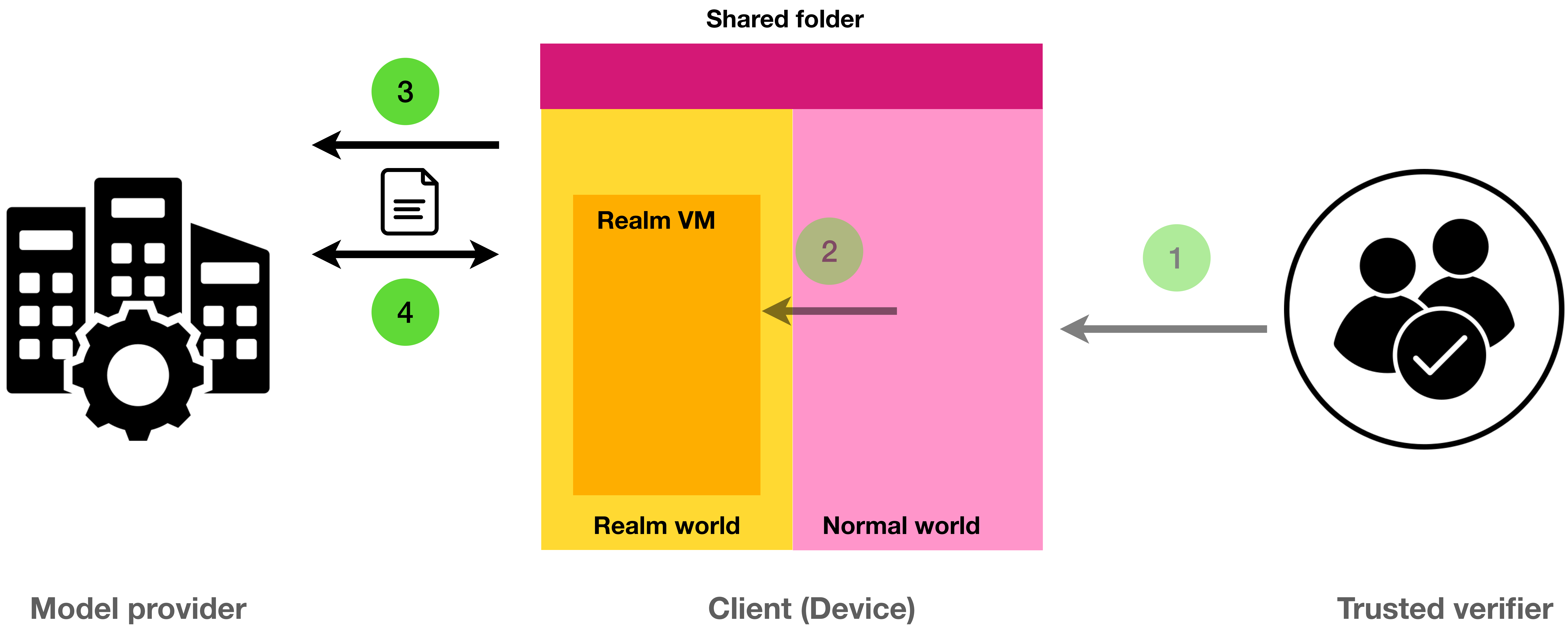
System overview



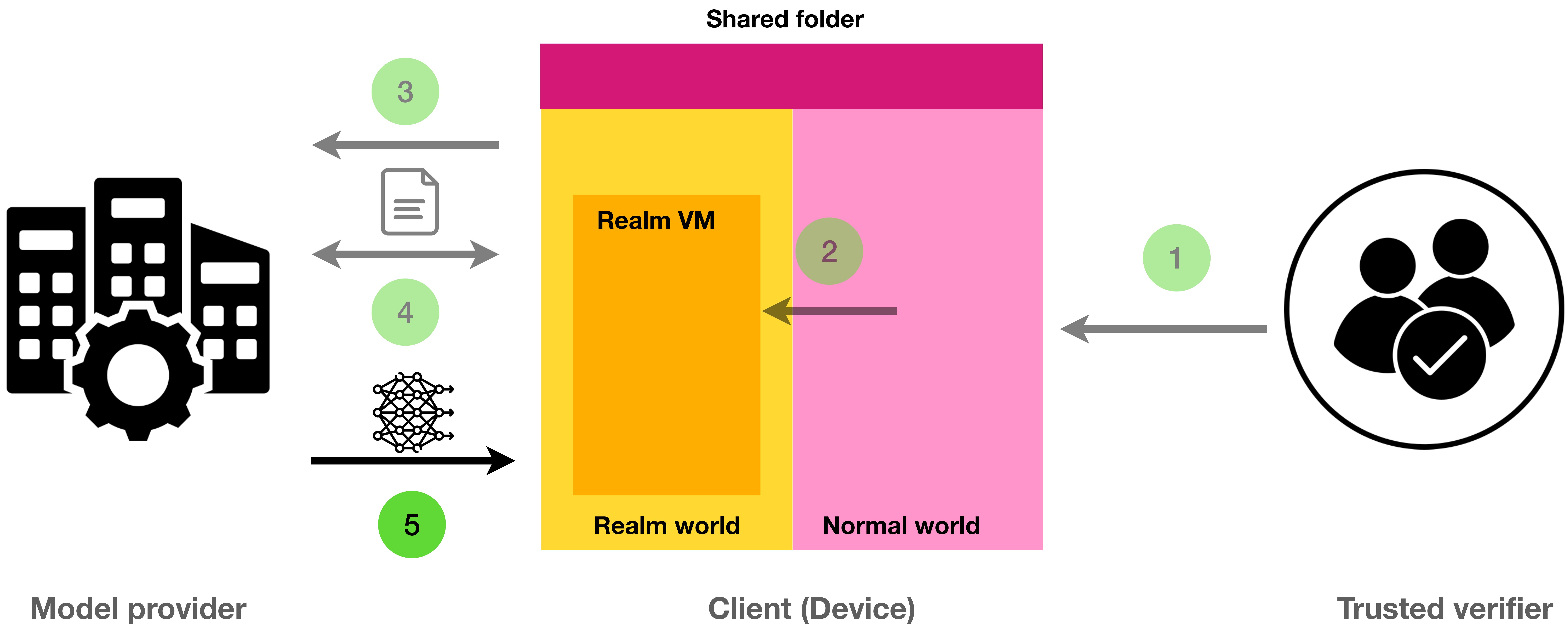
System overview



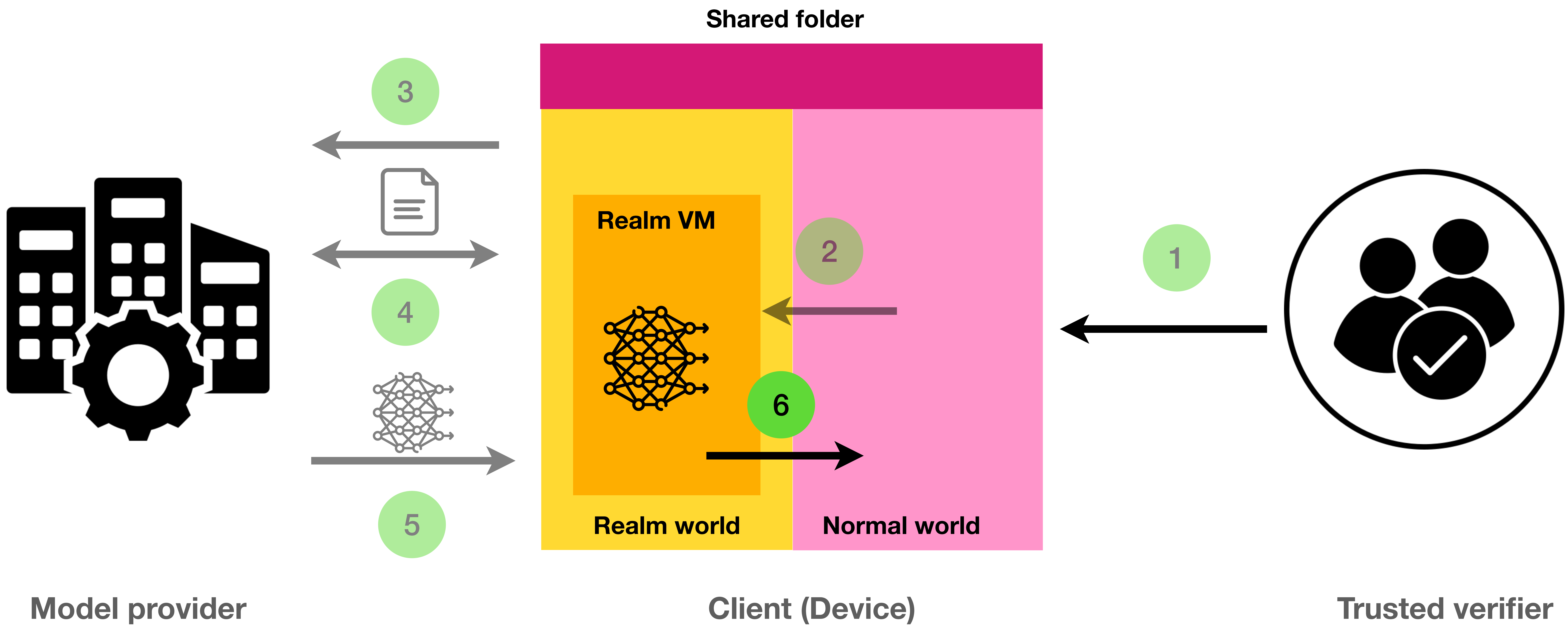
System overview



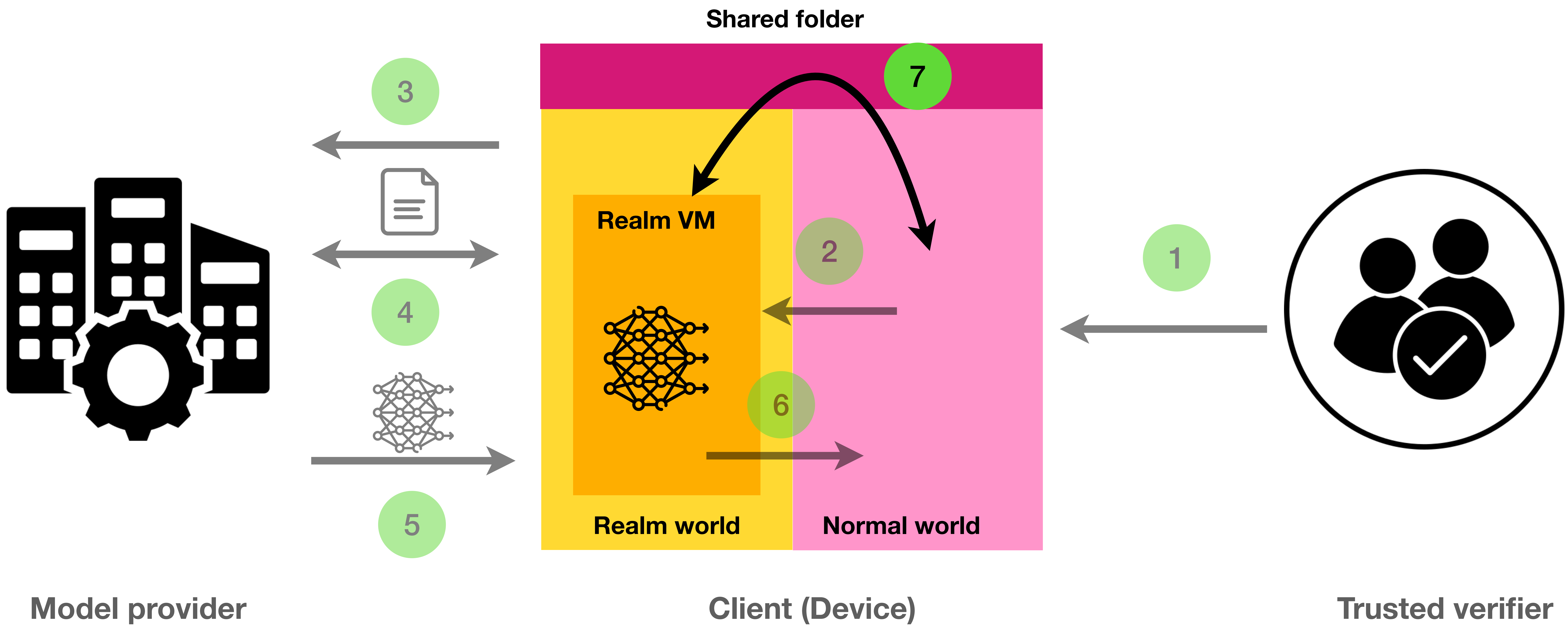
System overview



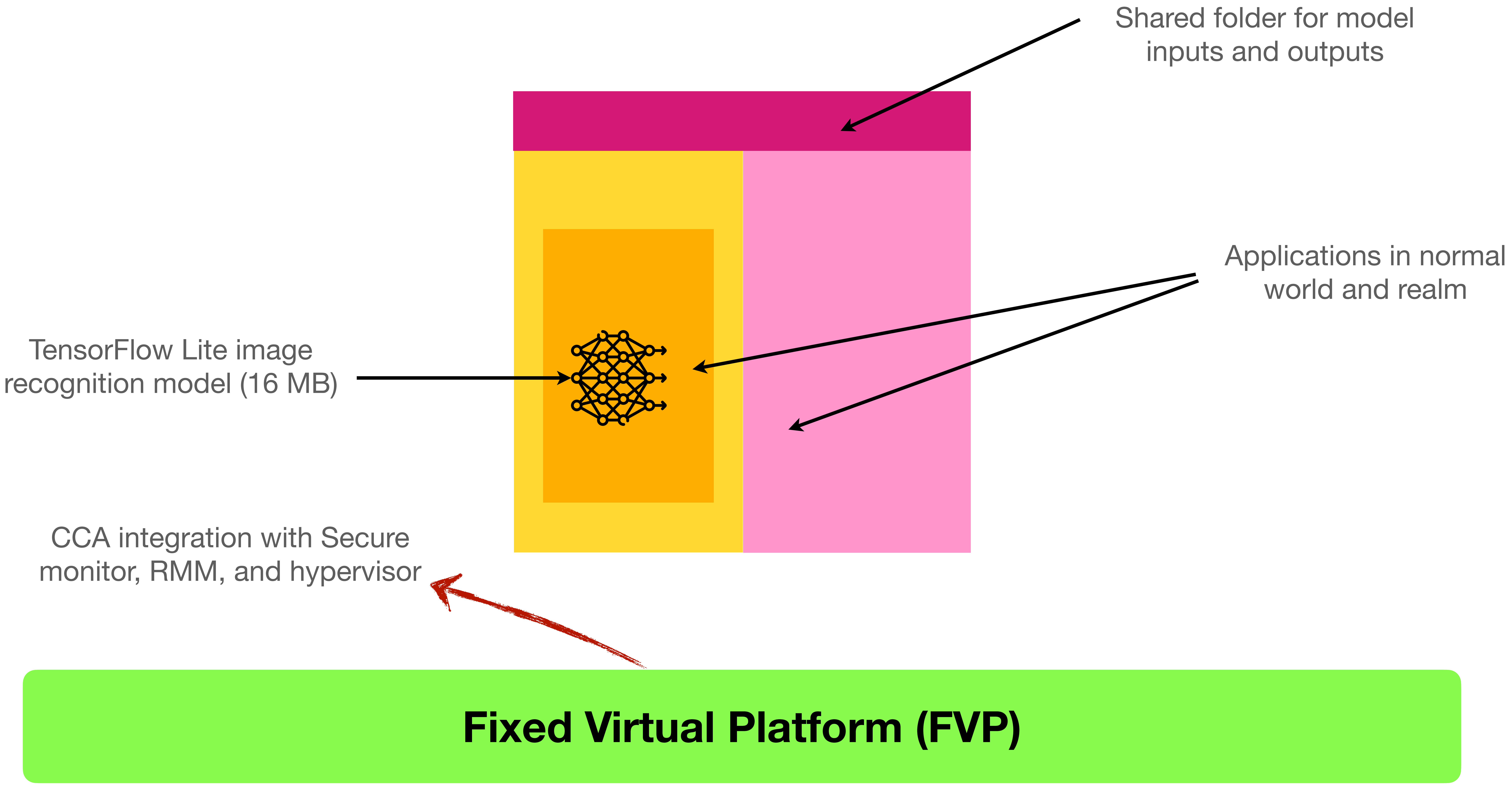
System overview



System overview



Implementation



Preliminary evaluation

What we measure: Overhead of inference and realm VM creation over a normal world VM.

Preliminary evaluation

What we measure: Overhead of inference and realm VM creation over a normal world VM.

How we measure: Number of instructions as FVP is not cycle-accurate

- Approximate counting of instructions.
- In progress: implementing Module Trace Interface for exact instructions.

Preliminary evaluation

What we measure: Overhead of inference and realm VM creation over a normal world VM.

How we measure: Number of instructions as FVP is not cycle-accurate

- Approximate counting of instructions.
- In progress: implementing Module Trace Interface for exact instructions.

Main findings

- On average, realm inference takes 1.6x the instructions normal world.
    - Larger number of context switches
- Realm creation depends on the size of the image.

Preliminary evaluation

What we measure: Overhead of inference and realm VM creation over a normal world VM.

How we measure: Number of instructions as FVP is not cycle-accurate

- Approximate counting of instructions.
- In progress: implementing Module Trace Interface for exact instructions.

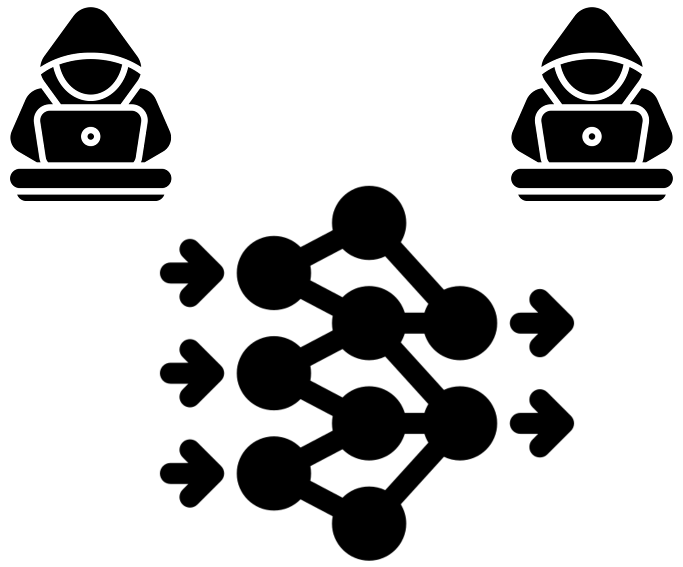
Main findings

- On average, realm inference takes 1.6x the instructions normal world.
    - Larger number of context switches
- Realm creation depends on the size of the image.

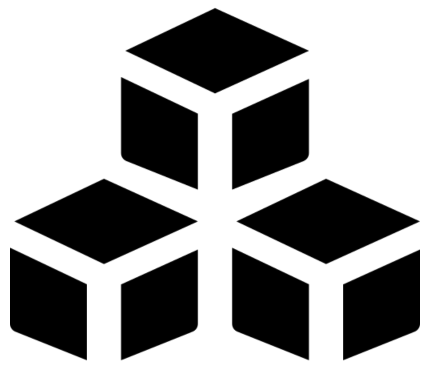
Note: Full attestation report could not be implemented due to FVP limitations

Considerations for ML deployment with CCA

Attacks to data pipeline



Multiple providers on the same device



Policy enforcement

Availability guarantees

Summary

- We propose GuaranTEE — a framework using CCA to deploy ML models on end devices in a private and trusted manner.
- We implement GuaranTEE using FVP, and perform a preliminary evaluation.
- We provide future directions and recommendations on ML deployment with CCA.

Code (with a setup guide): https://github.com/comet-cc/GuaranTEE

Get in touch: s.siby@imperial.ac.uk