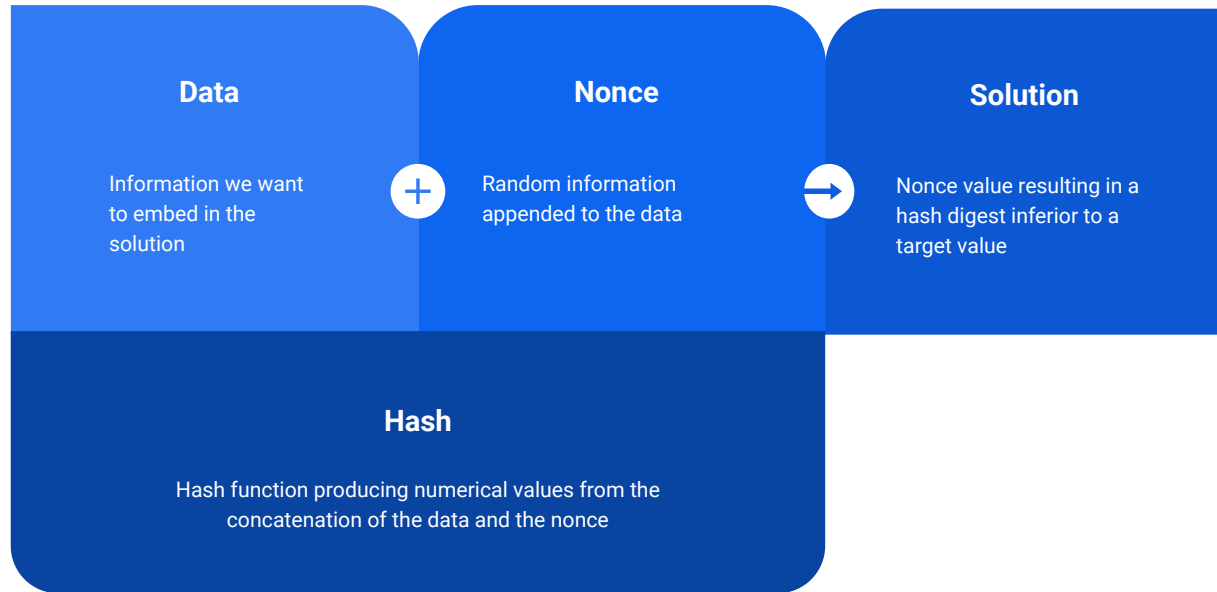


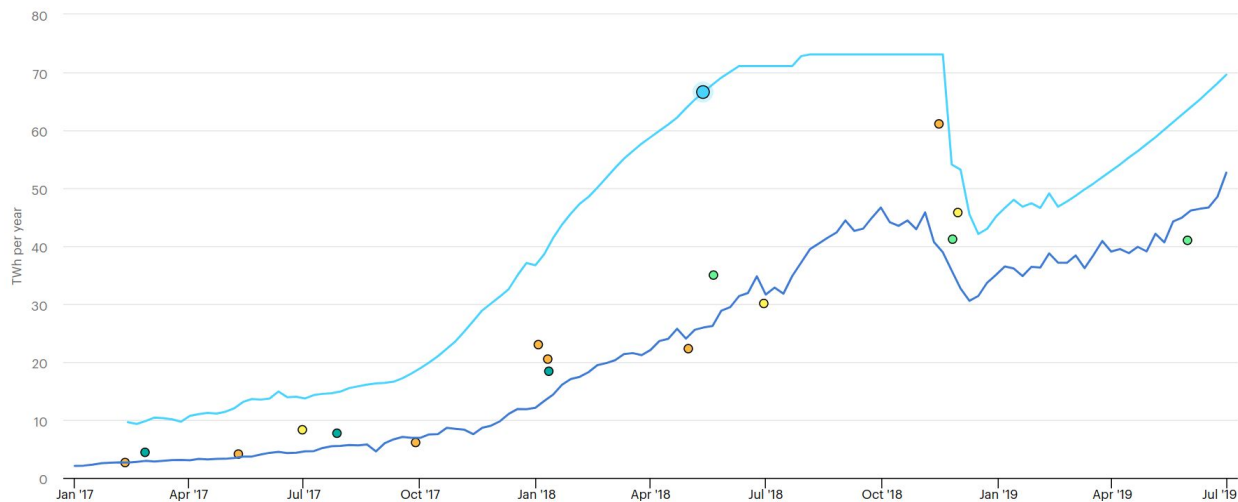
FreezeNet

Making Proof of Work More Useful

Proof of Work - with Hashcash



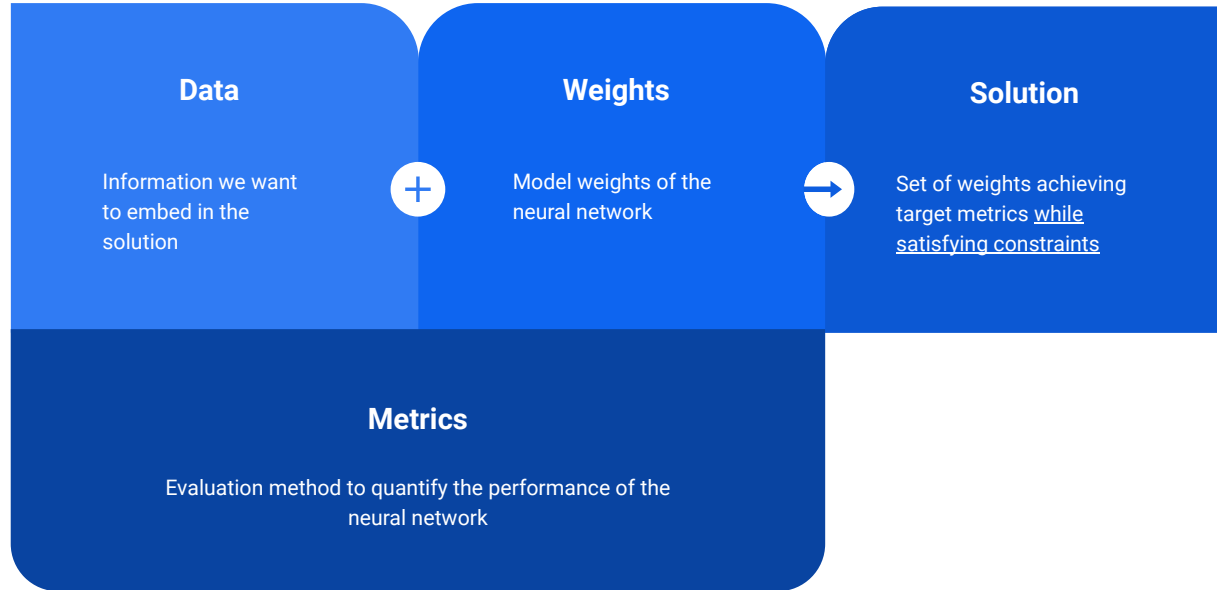
Bitcoin energy usage estimates



IEA. All Rights Reserved

● Digiconomist (2019) ● Lower bound (Antminer S9) ● Bendiksen, Gibbons, Lim (2018-19) ● Bevand (2018) ● Other peer-reviewed studies ● Other studies

Proof of Work - with FreezeNet



Watermarking : from a simple idea ...

Create watermark

- Hash data to generate PRNG seed
- Use PRNG initialized with seed to generate weights and indices
- Number of weights generated is the watermark size

Apply watermark to model

- Replace model weights at watermark indices by watermark weights
- Weights can be replaced during training after backpropagation phase

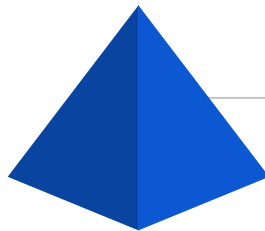
Verify watermark

- Check if model weights at watermark indices are equal to watermark weights
- Account for floating point imprecision by setting upper bound on difference

... to a simple API

```
1 model = FreezeNet()  
2  
3 some_watermark = Watermark(b'Some block information', 4096)  
4 another_watermark = Watermark(b'Some other block information', 4096)  
5  
6 some_watermark.apply(model)  
7  
8 some_watermark.verify(model) # True  
9 another_watermark.verify(model) # False  
10  
11 another_watermark.apply(model)  
12  
13 some_watermark.verify(model) # False  
14 another_watermark.verify(model) # True
```

Desired behaviours

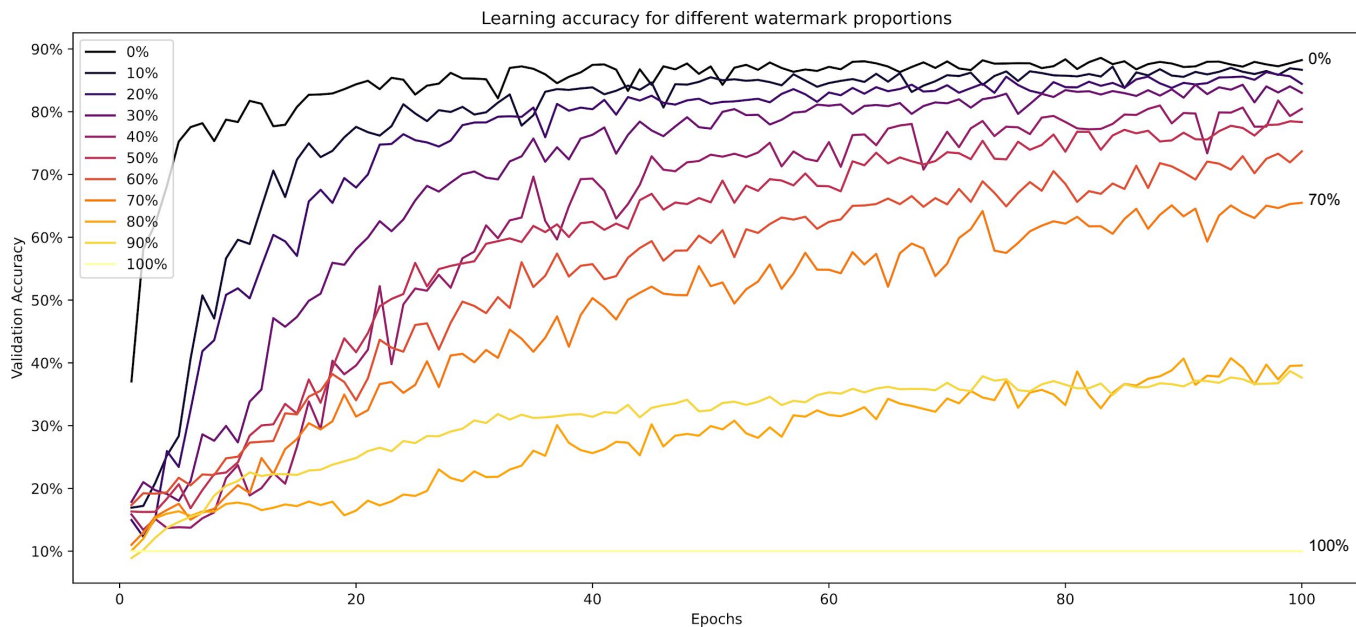


Learning ability

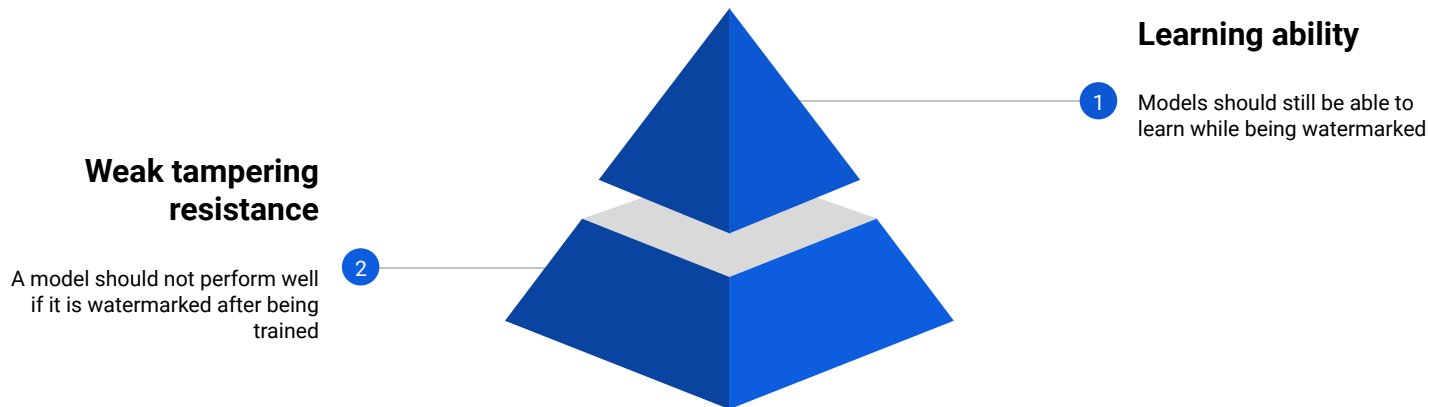
1

Models should still be able to learn while being watermarked

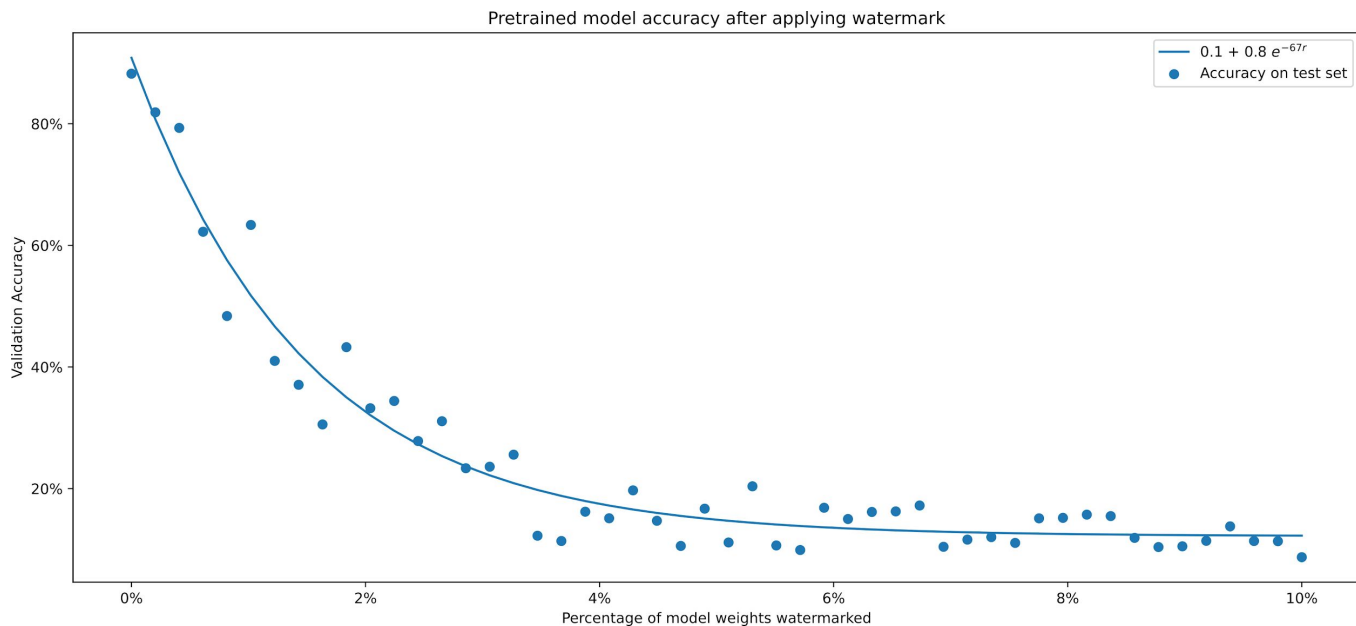
Learning ability



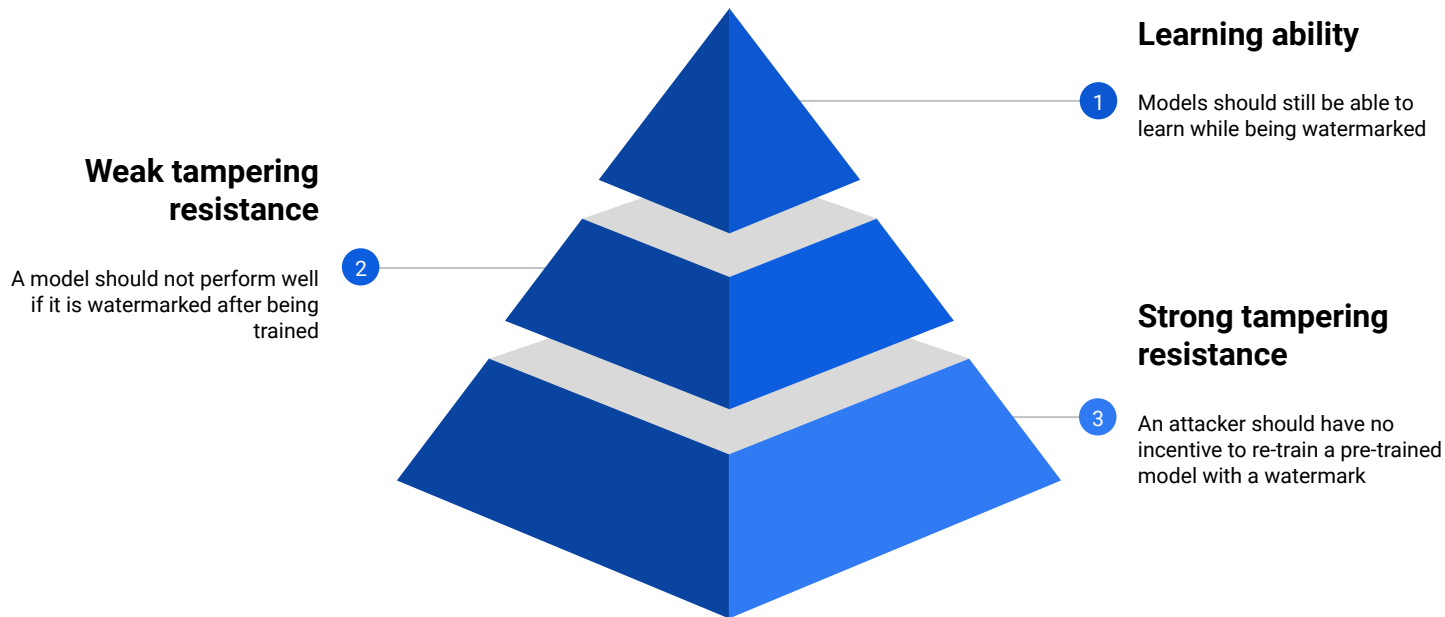
Desired behaviours



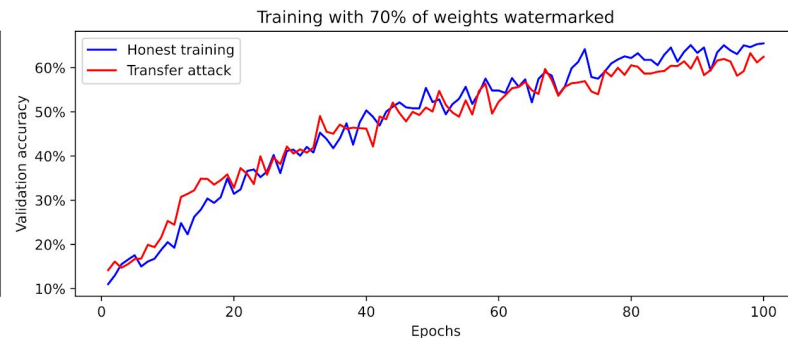
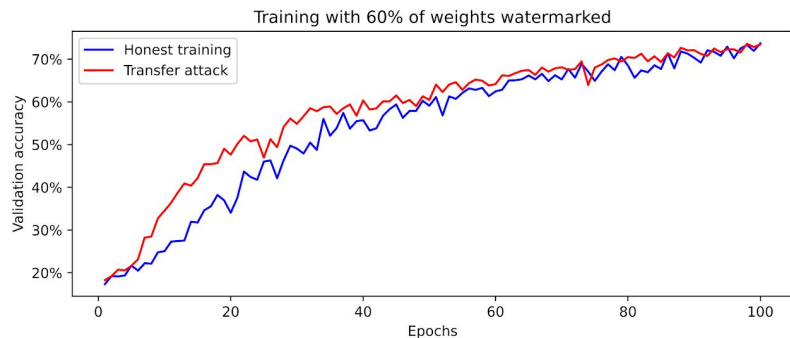
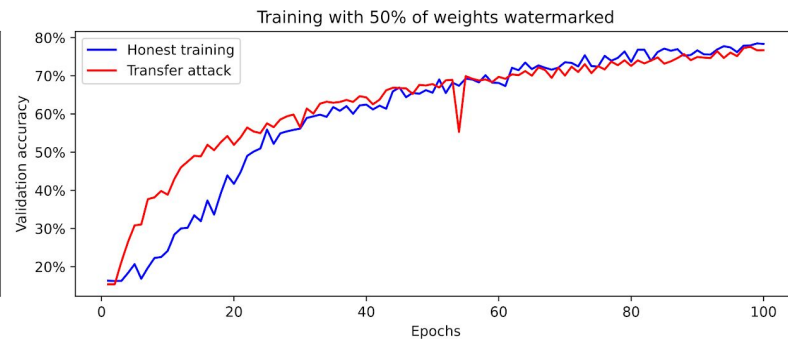
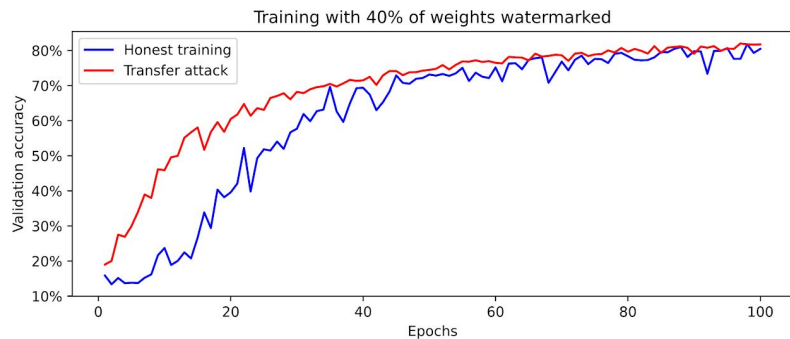
Weak tampering resistance



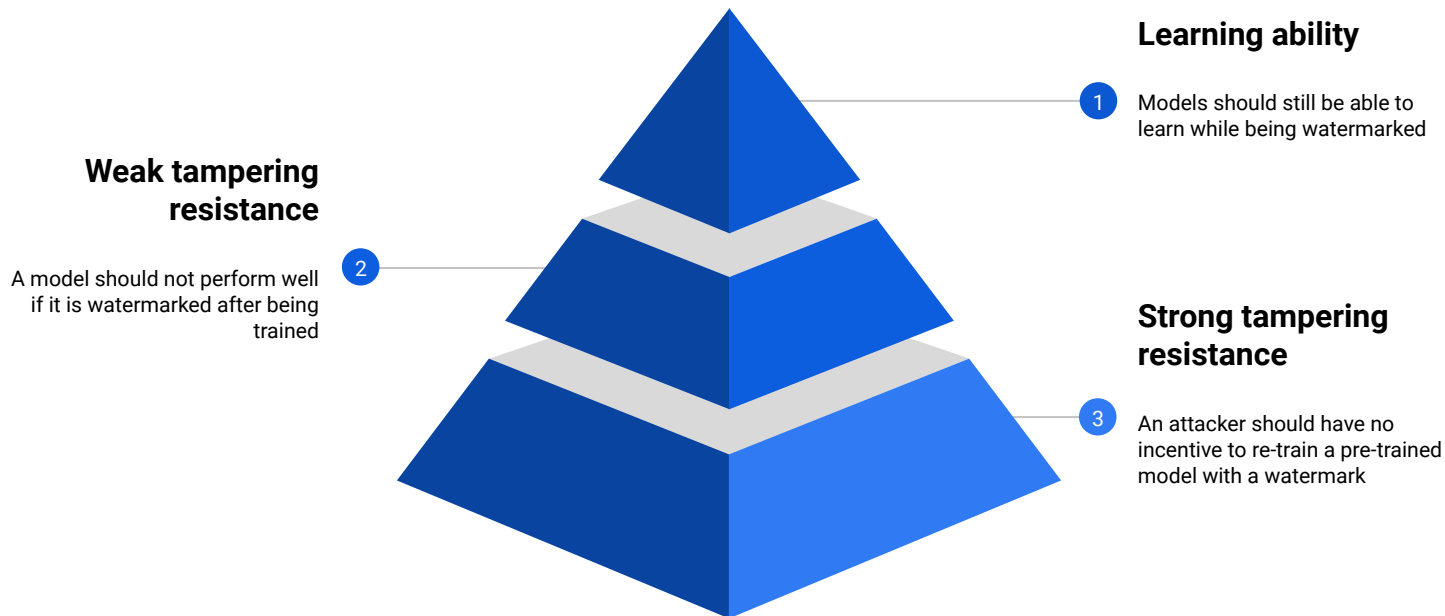
Desired behaviours



Strong tampering resistance



Desired behaviours



Concerns

Overhead

Dataset and model weights transfer add significant overhead

Centralization

Datasets are not generated in a distributed manner, so workers have to rely on a centralized dataset distribution

Security

The attack surface is very large, transfer attacks only scratch the surface of possible vulnerabilities