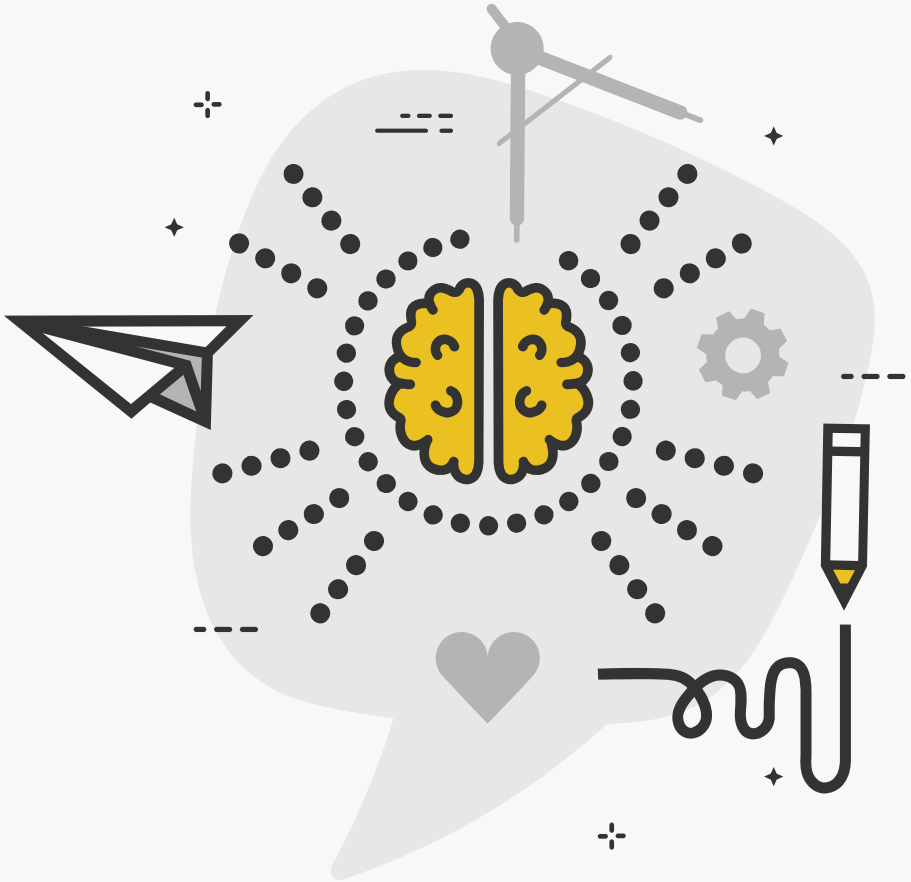
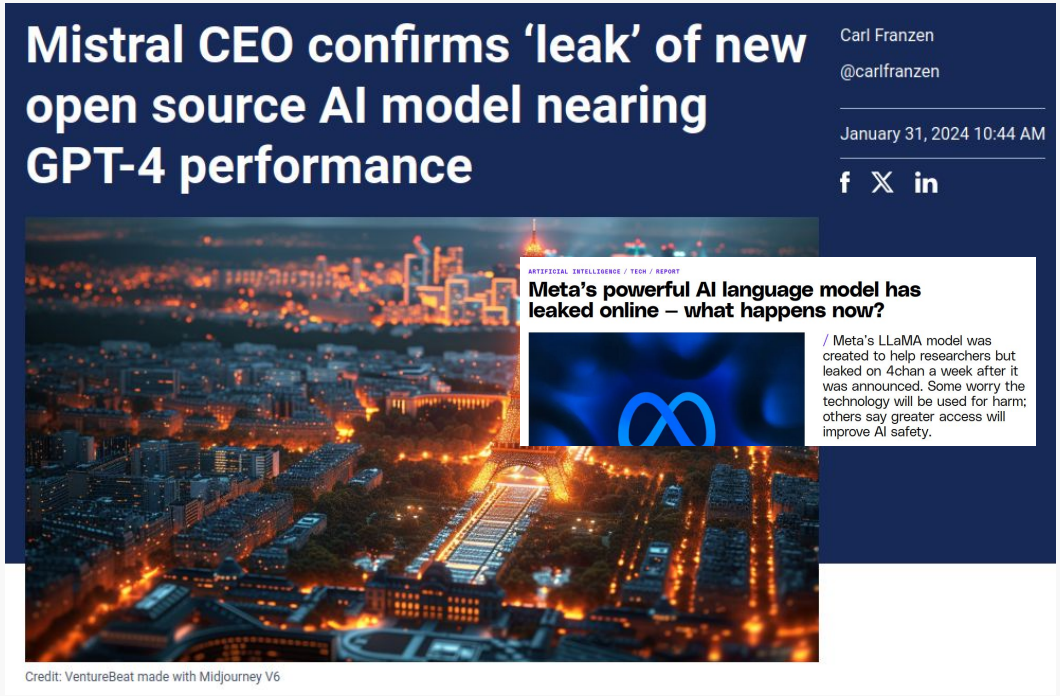
ZaMark

Intellectual Property protection with Homomorphic Watermarking

Reda Bellafqira, Mehdi Ben Ghali, Pierre-Elisée Flory, Mohammed Lansari, Thomas Winninger

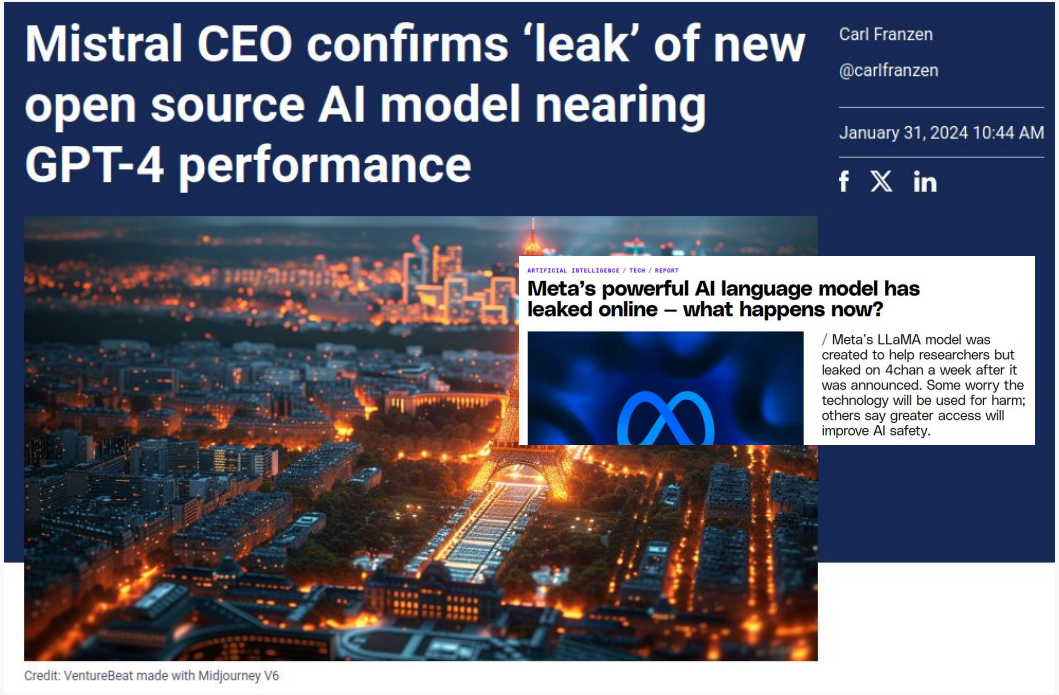


What's the problem we are solving?



huggingface.co/miqudev/miqu-1-70b

What's the problem we are solving?

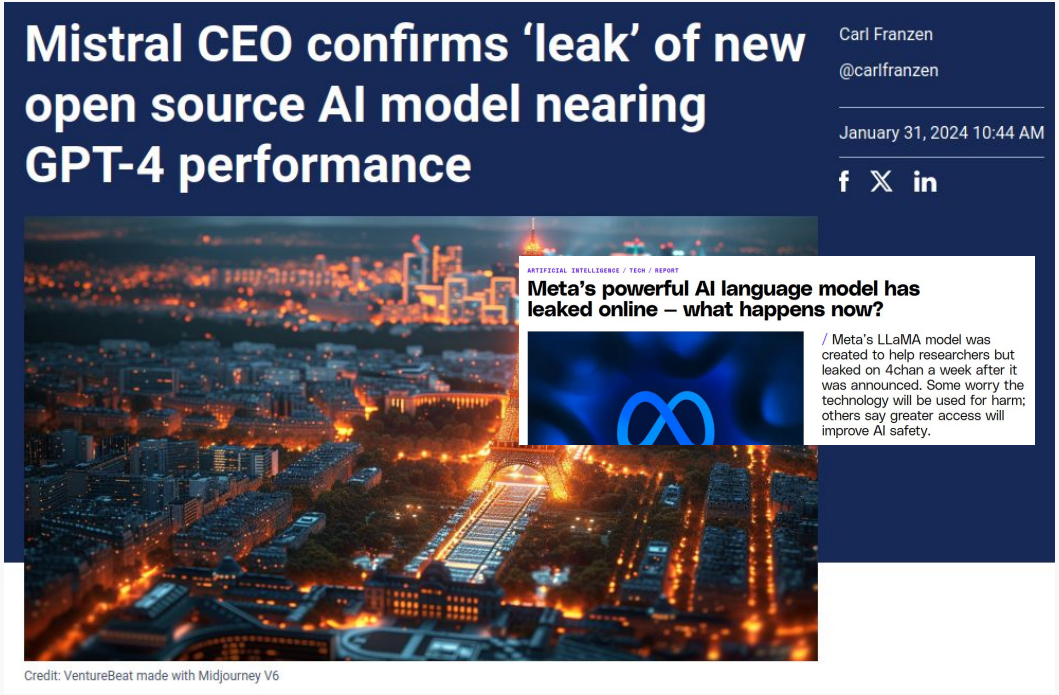


Training GPT-3 [1] Using a Tesla V100 Cloud Instance:

- Cost -> **$4.6M**
- Time -> **355 years**

huggingface.co/miqudev/miqu-1-70b

What's the problem we are solving?



Credit: VentureBeat made with Midjourney V6

Training GPT-3 [1] Using a Tesla V100 Cloud Instance:

- Cost -> **$4.6M**
- Time -> **355 years**

-> Need for AI IP Protection

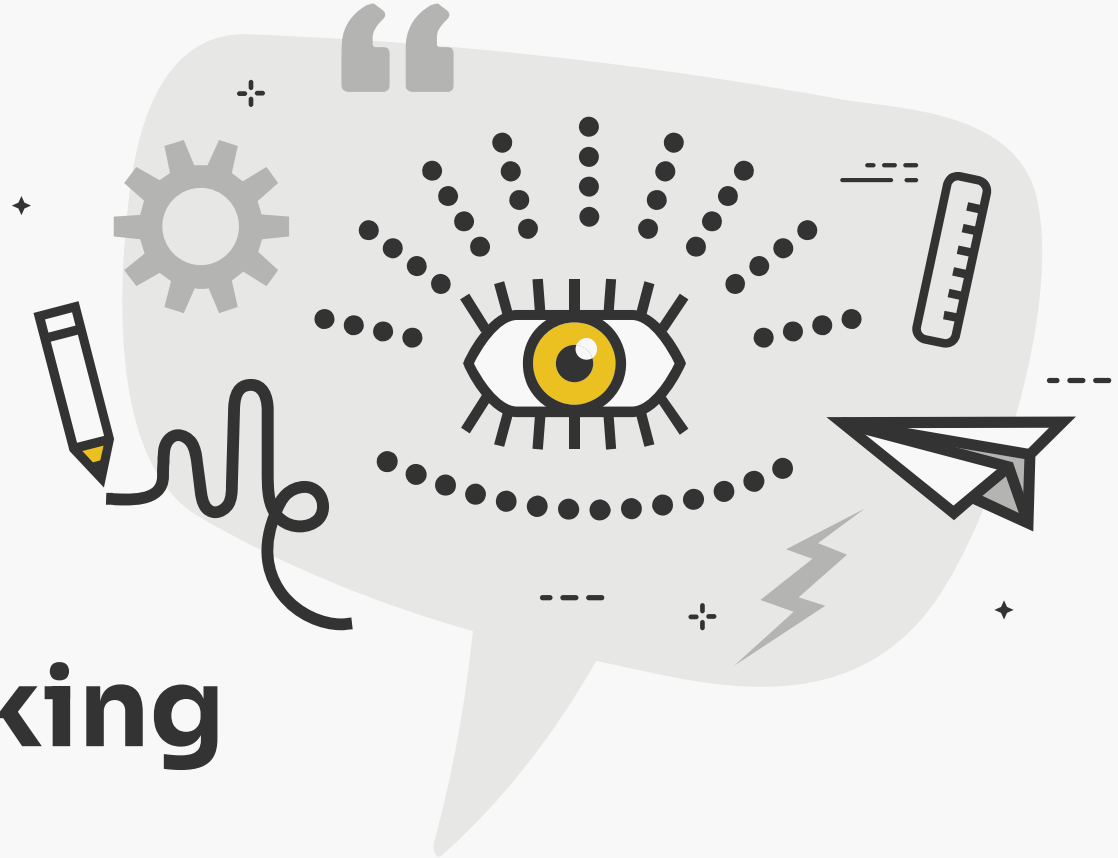
IP protection market size :

- Value -> **$7.5B** in 2022
- Projection -> **$30.3B** by 2032
- 15.6% CAGR Growth

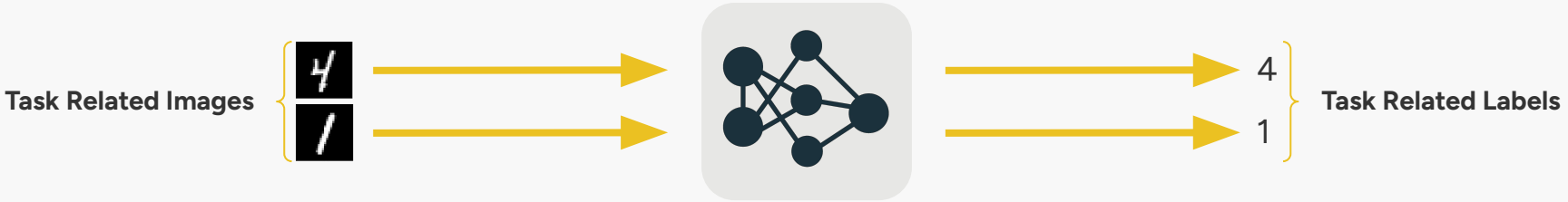
huggingface.co/miqudev/miqu-1-70b

01

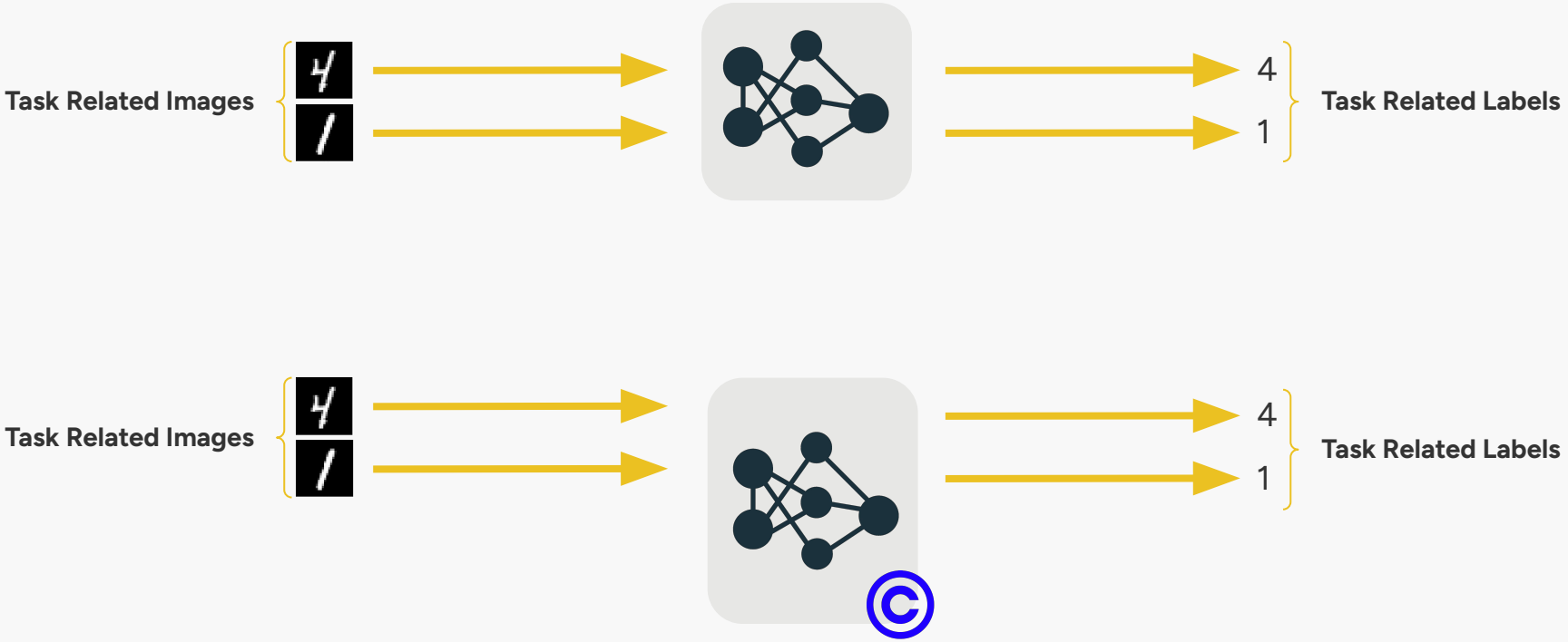
AI Model Watermarking



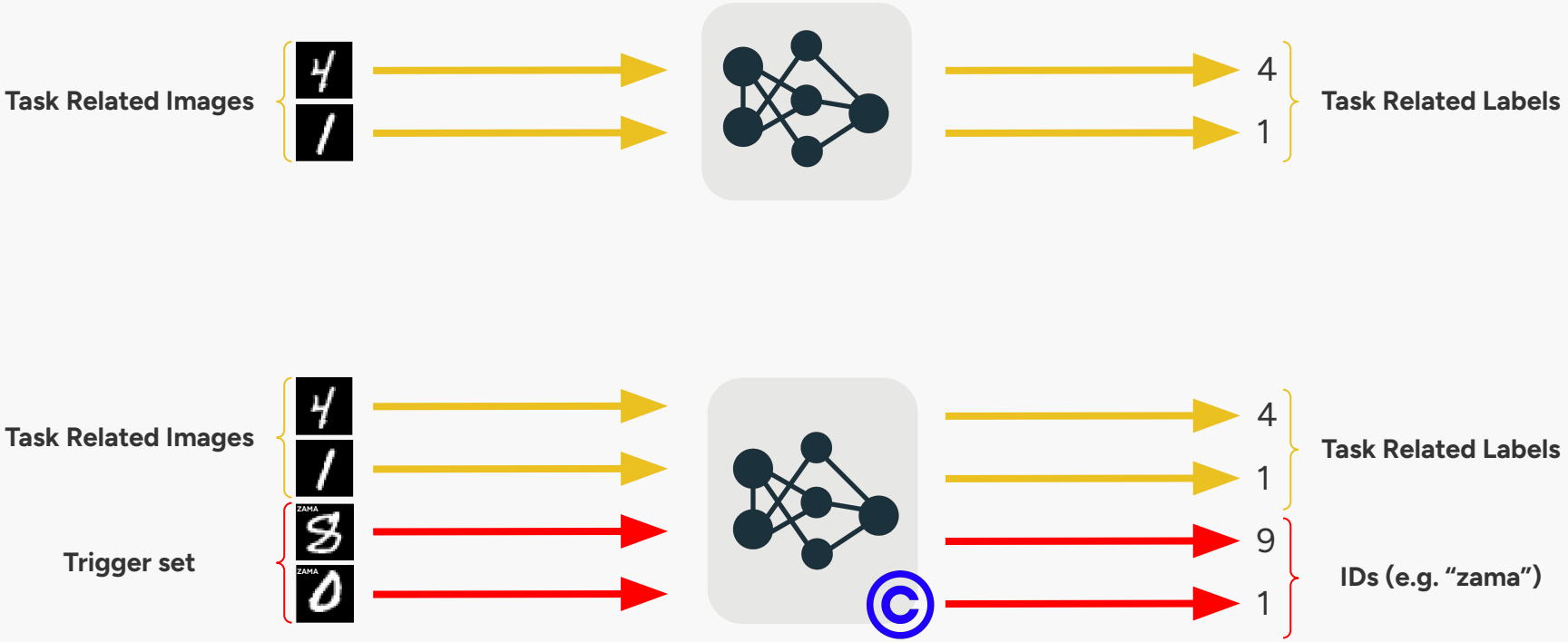
AI Model Watermarking – Principle



AI Model Watermarking – Principle



AI Model Watermarking – Principle



AI Model Watermarking – Interests

Industrial Product :

eIQ® Model Watermarking Technology [2]

AI Model Watermarking – Interests

Industrial Product :

eIQ® Model Watermarking Technology [2]

Standards :

Securing Artificial Intelligence (SAI) : Traceability of AI Models [3]

AI Model Watermarking – Interests

Industrial Product :

eIQ® Model Watermarking Technology [2]

Standards :

Securing Artificial Intelligence (SAI) : Traceability of AI Models [3]

Research :

[4]

[5]

[6]

02

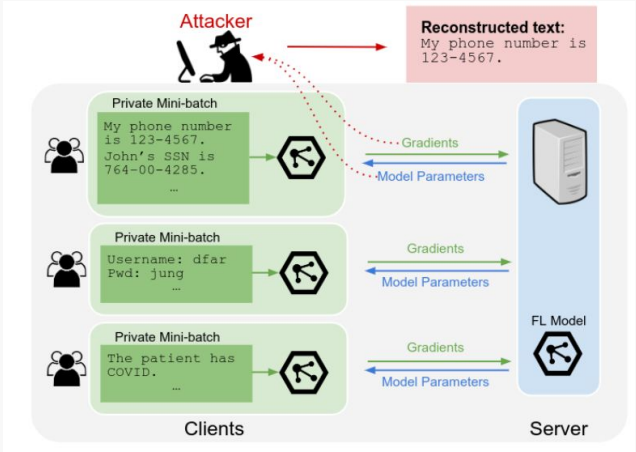
ZaMark: Homomorphic AI Watermarking as a Service

ZaMark: Homomorphic AI Watermarking as a service

- WaaS demand for multimedia (Imatag, Digimarc, etc.)
- No AI WaaS providers
- Need for **AI watermarking as a service**

ZaMark: Homomorphic AI Watermarking as a service

- WaaS demand for multimedia (Imatag, Digimarc, etc.)
- No AI WaaS providers
- Need for **AI watermarking as a service**

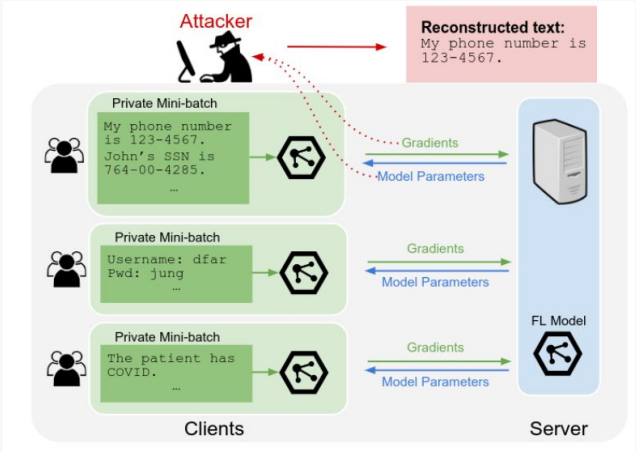


However, this poses privacy and confidentiality concerns :

- Model inversion attacks
- Model thefts, leaks...

ZaMark: Homomorphic AI Watermarking as a service

- WaaS demand for multimedia (Imatag, Digimarc, etc.)
- No AI WaaS providers
- Need for **AI watermarking as a service**

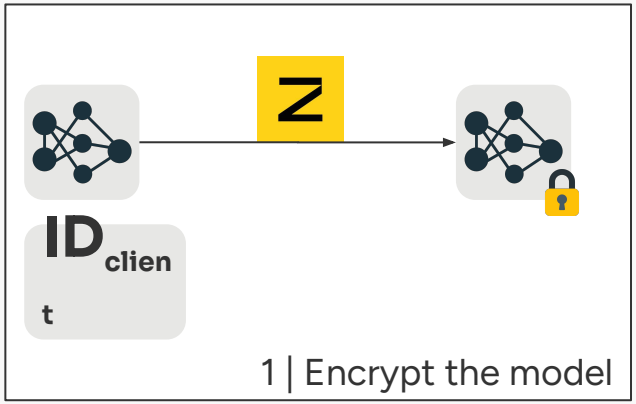


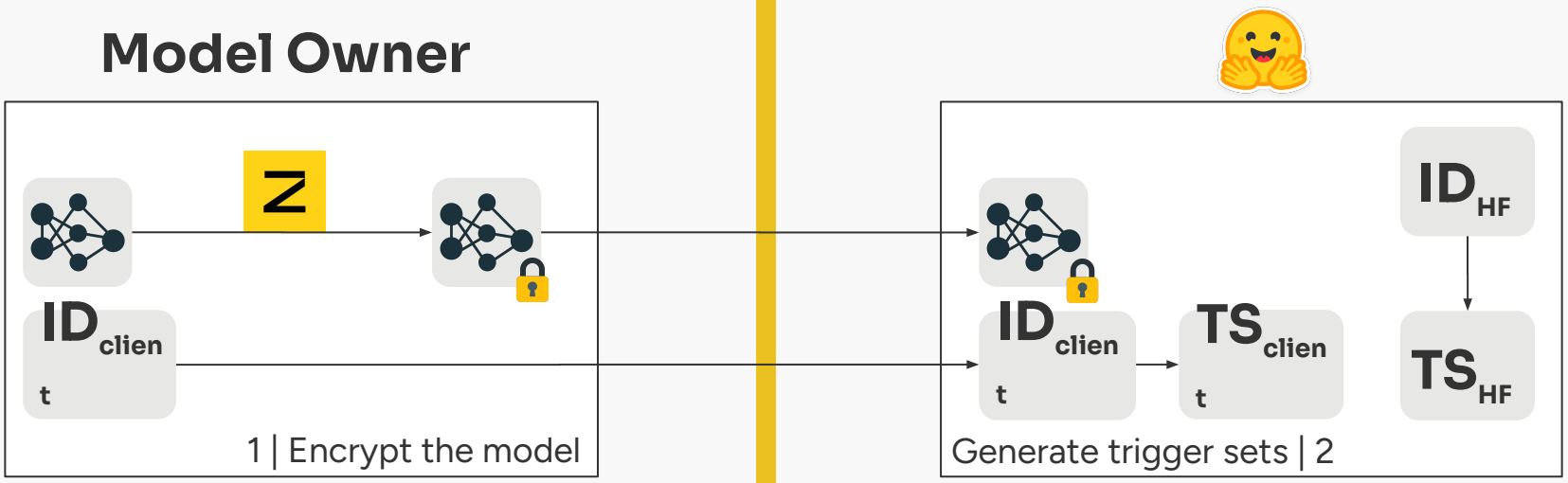
However, this poses privacy and confidentiality concerns :

- Model inversion attacks
- Model thefts, leaks...

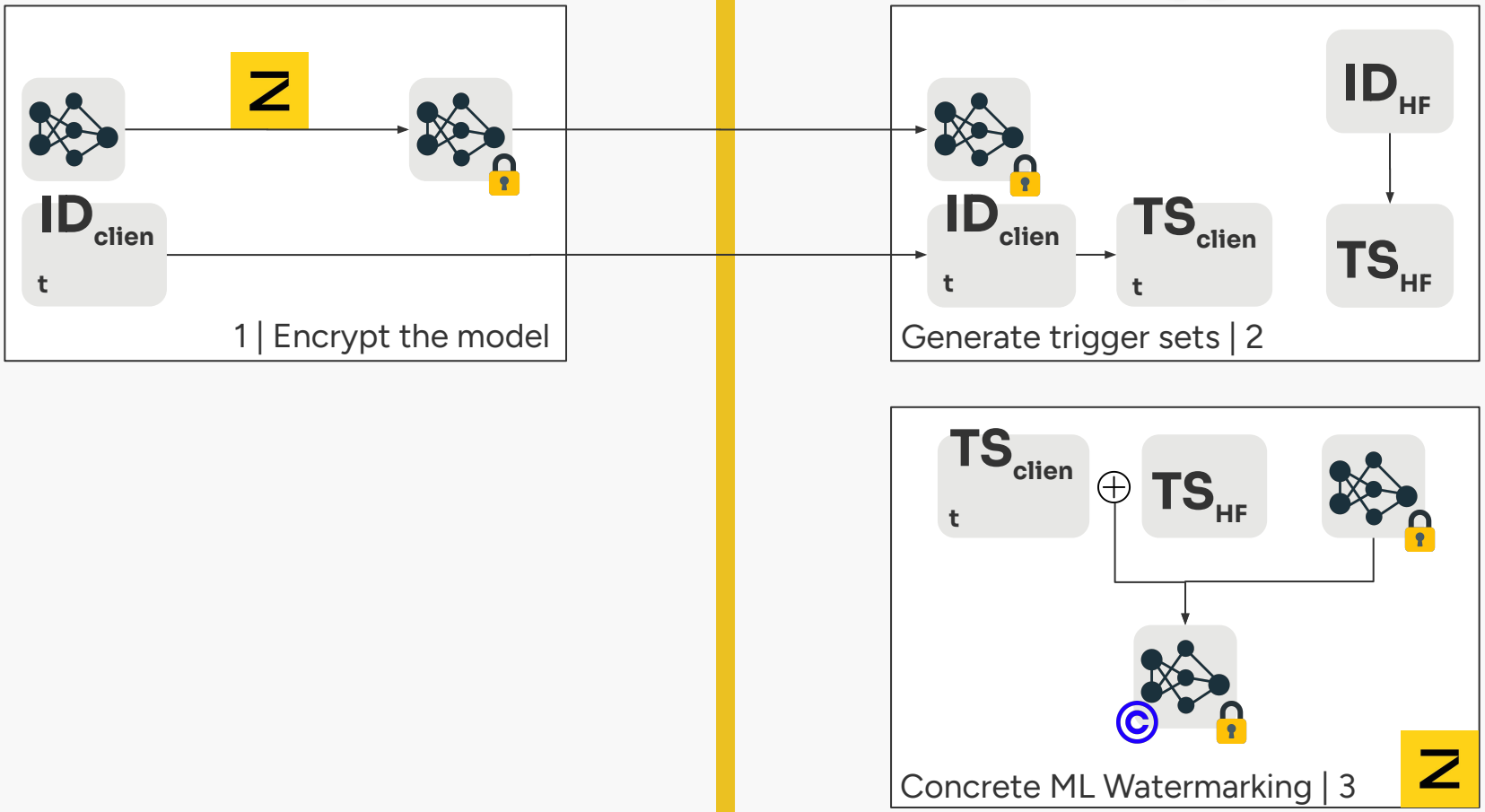
-> **Homomorphic AI watermarking as a service** using concrete-ml

Model Owner

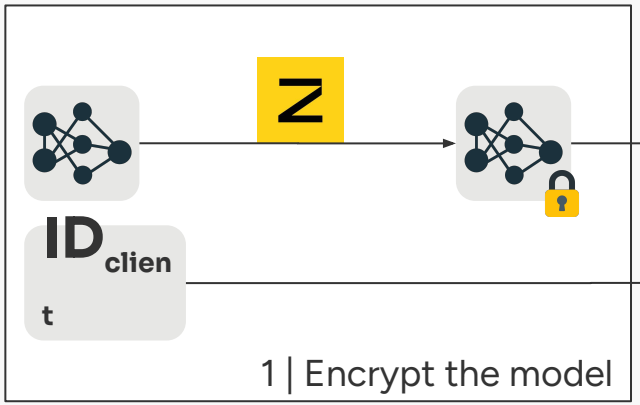


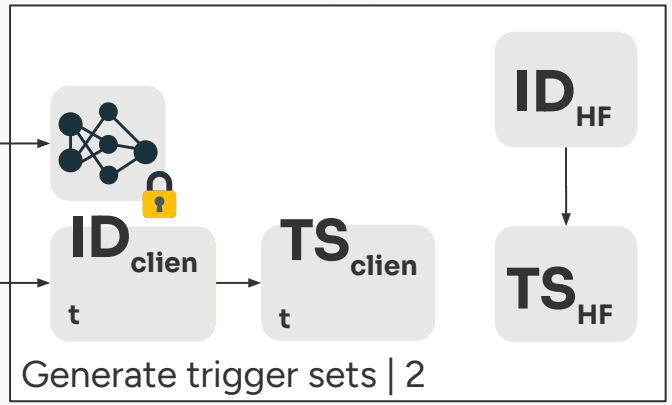

Model Owner
Z
ID clien t
1 | Encrypt the model
ID clien t
TS clien t
ID HF
TS HF
Generate trigger sets | 2

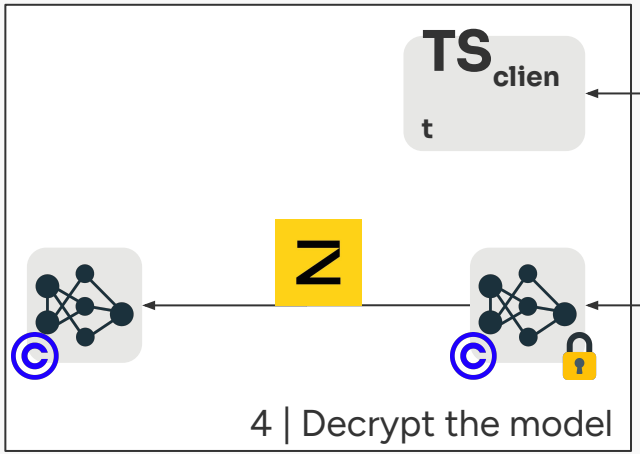
Model Owner

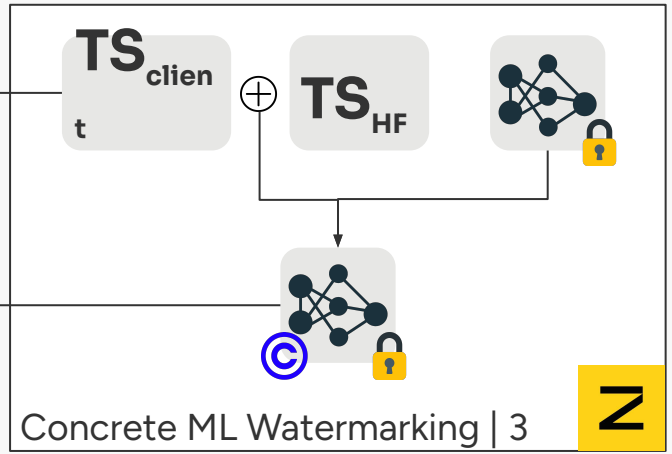

ID clien t
1 | Encrypt the model
ID HF
ID clien t
TS clien t
TS HF
Generate trigger sets | 2
TS clien t
TS HF
Concrete ML Watermarking | 3

Model Owner

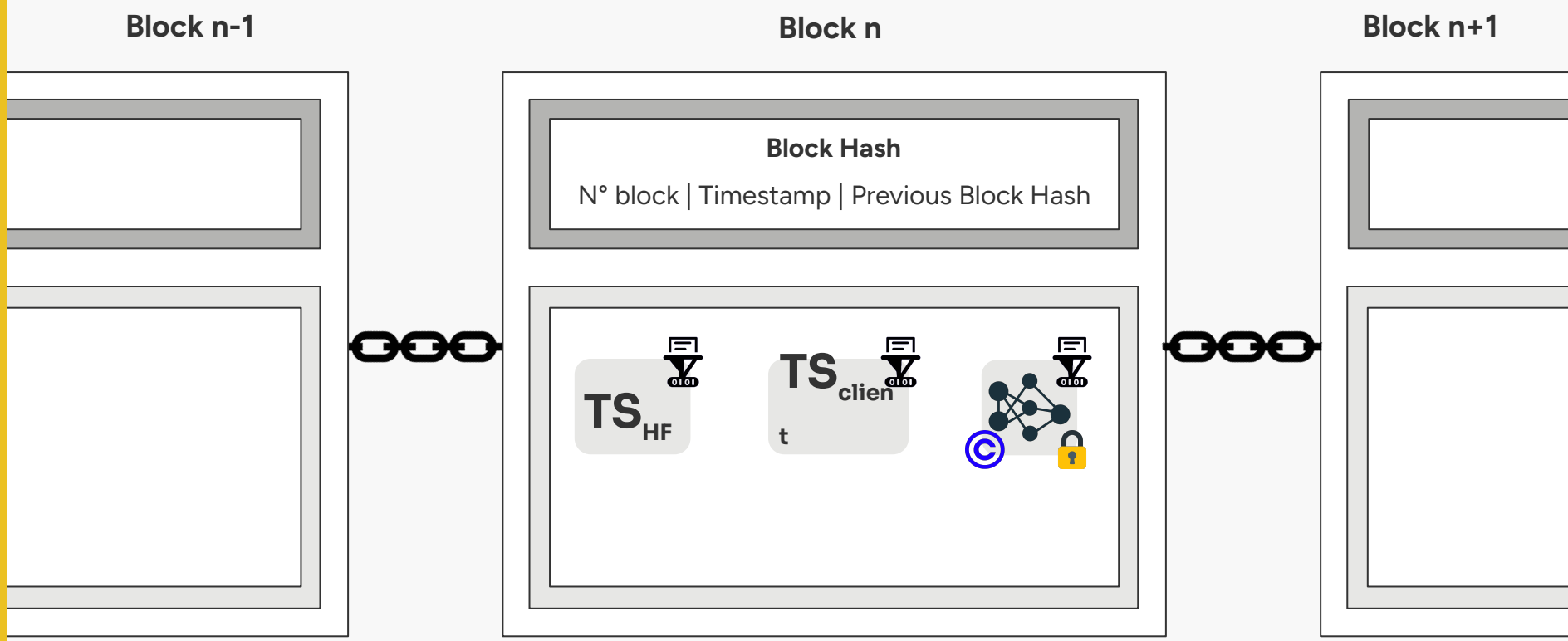

Z
ID clien t
1 | Encrypt the model


ID HF
ID clien t
TS clien t
TS HF
Generate trigger sets | 2


TS clien t
Z
4 | Decrypt the model


TS clien t
TS HF
Concrete ML Watermarking | 3
Z

Blockchain


Block n-1
Block n
Block n+1
Block Hash
N° block | Timestamp | Previous Block Hash
TS HF
TS clien t
Hash Function
Watermark
Encrypted

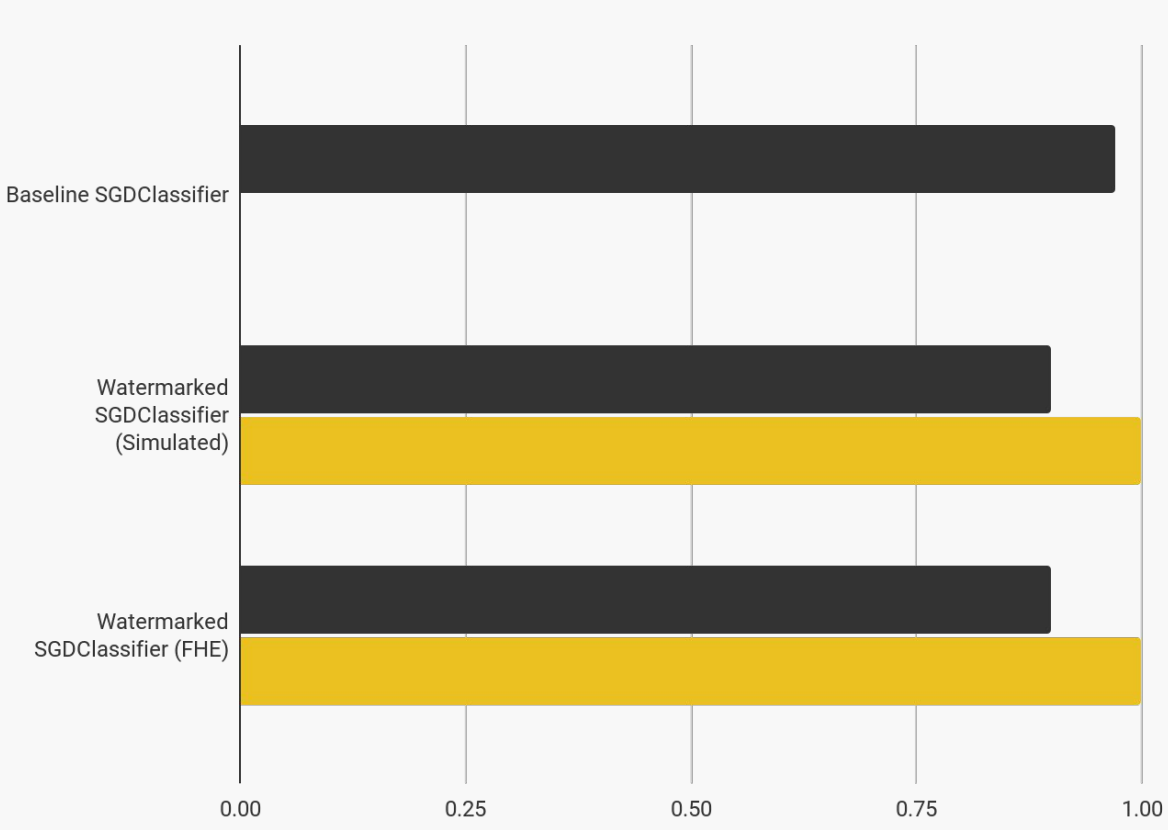
03

Proof Of Concept (POC)

SGDClassifier with concrete-ml

Model	**SGDClassifier**
Training Set	**Wisconsin breast cancer dataset** (2 classes, **569 samples**, 30 features)
Training Epochs	**100 epochs**
Trigger Set	**Randomly generated** (2 classes, **15 samples**, 30 features)
Watermarking Epochs	**17 epochs**

Results



- Perfect mark embedding
- Small performance loss (watermark)
- FHE does not degrade performance

Computation constraints

	# Parameters	Training Time	Watermarking (Simulated FHE)	Watermarking (Real FHE)
SGDClassifier	30	4.3s	2.6s	5 minutes

Computation constraints

	# Parameters	Training Time	Watermarking (Simulated FHE)	Watermarking (Real FHE)
SGDClassifier	30	4.3s	2.6s	5 minutes

- 115x slowdown factor between FHE and clear computation
- Using a **naive** watermarking technique
- **Satisfactory for SGDClassifier in FHE**
- Still need for **SGDClassifier in FHE for** DNN training
- **Further optimisation possible** for scalability (LLM, Diffusion…)

- Homomorphic Watermarking is **effective** but not **efficient** yet

References

- [1] What Large Models Cost You – There Is No Free AI Lunch - Forbes
- [2] eIQ® Model Watermarking Technology - NXP Semiconductors
- [3] Fernandez, Pierre, et al. "Three bricks to consolidate watermarks for large language models." 2023 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2023.
- [4] ETSI Technical Report on Model Watermarking
- [5] Darvish Rouhani, Bita, Huili Chen, and Farinaz Koushanfar. "Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks." Proceedings of the twenty-fourth international conference on architectural support for programming languages and operating systems. 2019.
- [6] Zhang, Jialong, et al. "Protecting intellectual property of deep neural networks with watermarking." Proceedings of the 2018 on Asia conference on computer and communications security. 2018.

Thank you for your attention !

https://huggingface.co/spaces/ppaihack/**ZaMark**

ENTREPRENEUR FIRST

ZAMA