Position: Membership Inference Attacks **Cannot** Prove that a Model Was Trained On Your Data

SPY Lab

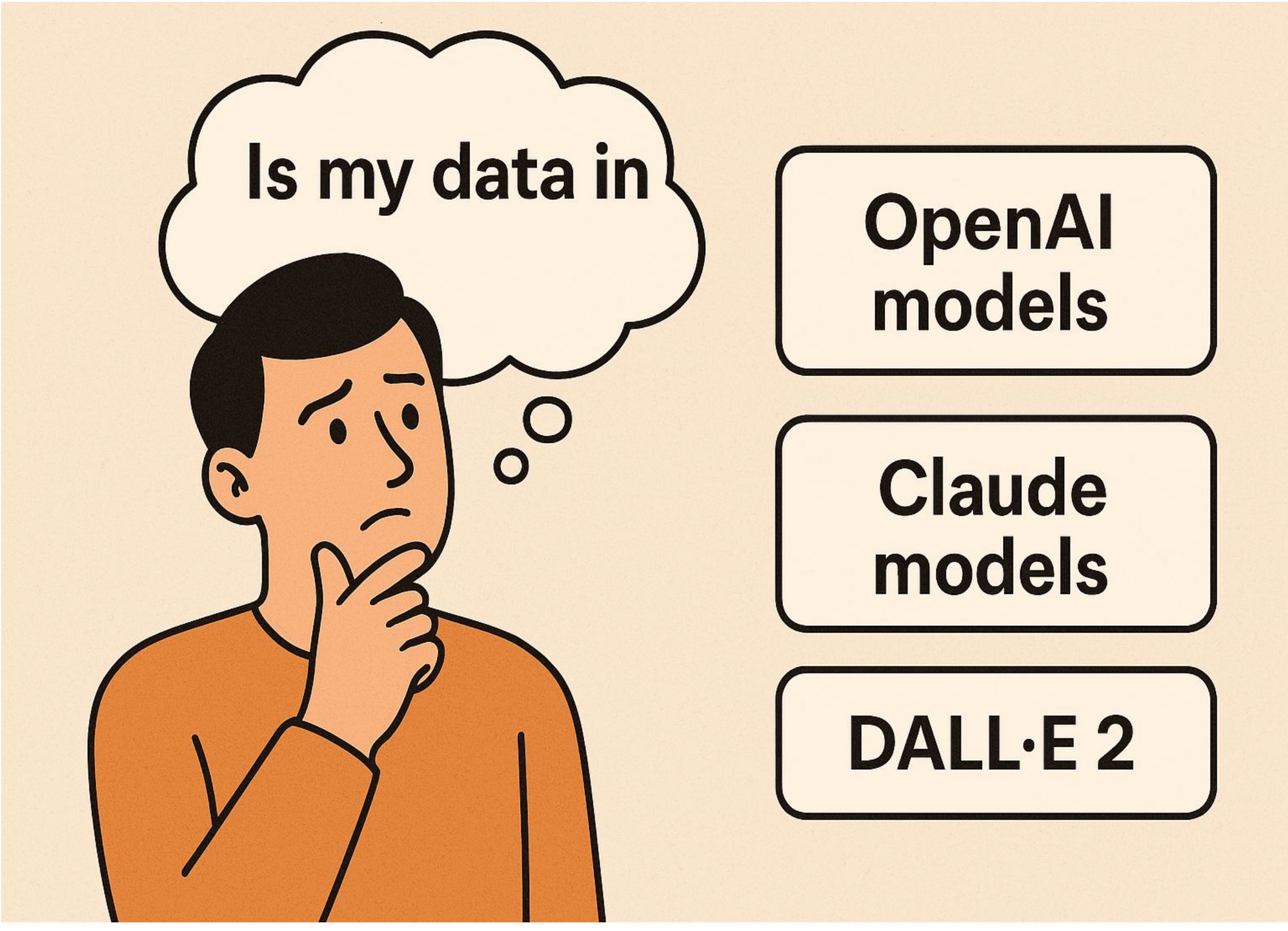
ETH zürich

The Salon!

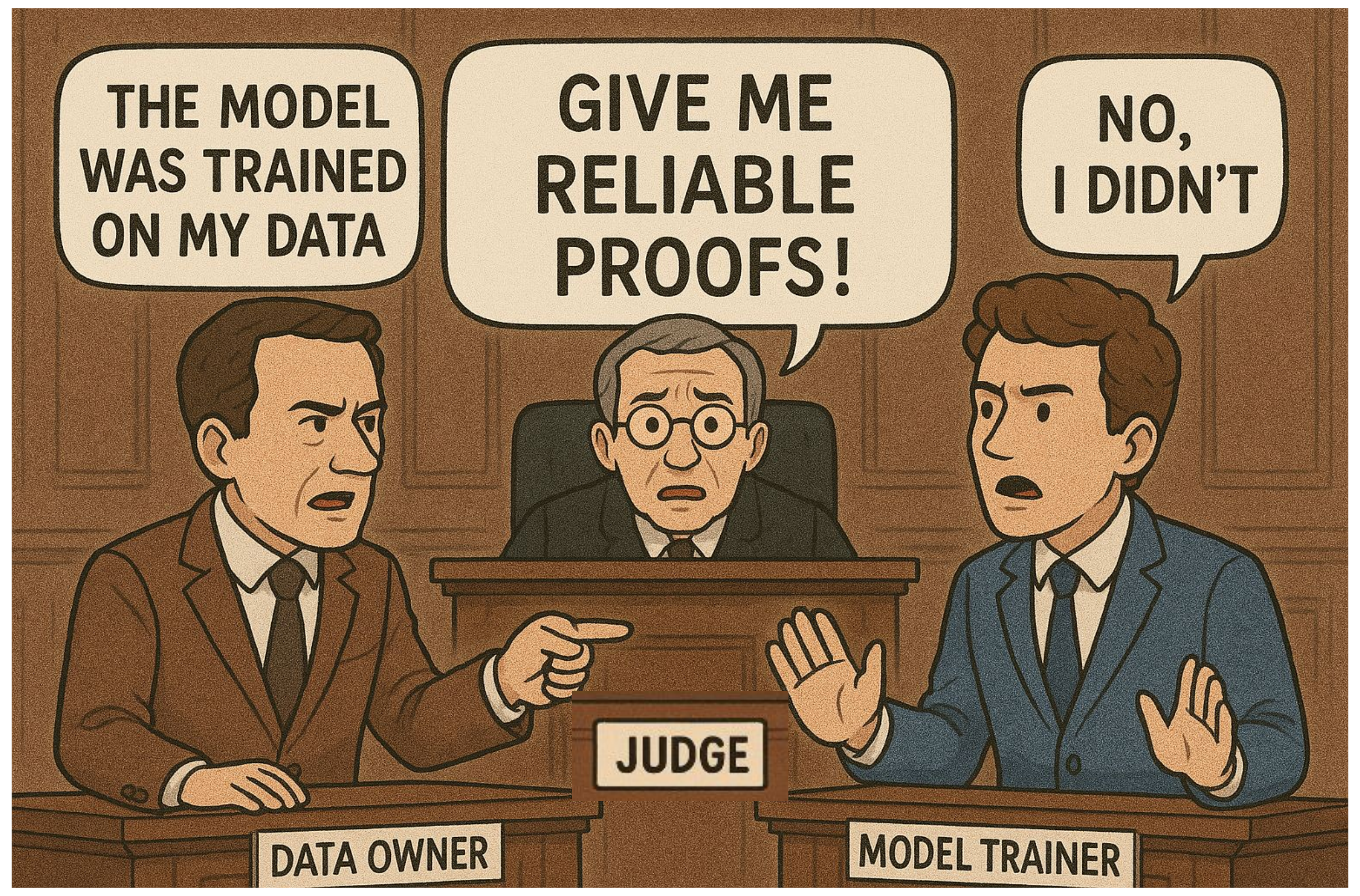
Jie Zhang, Debeshee Das, Gautam Kamath, Florian Tramèr

1/ What is a training data proof?

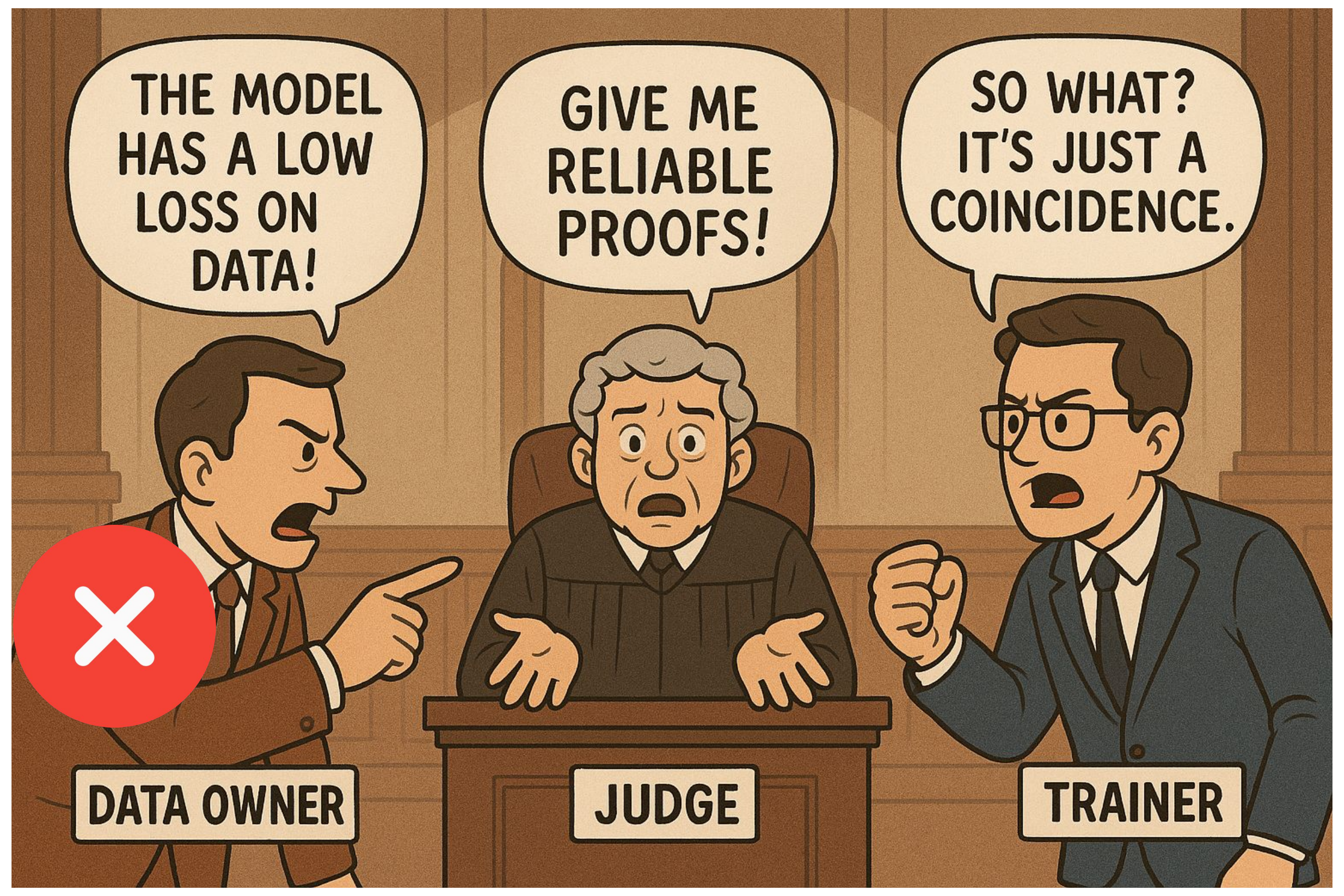
TL;DR: MIA cannot be used as training data proofs



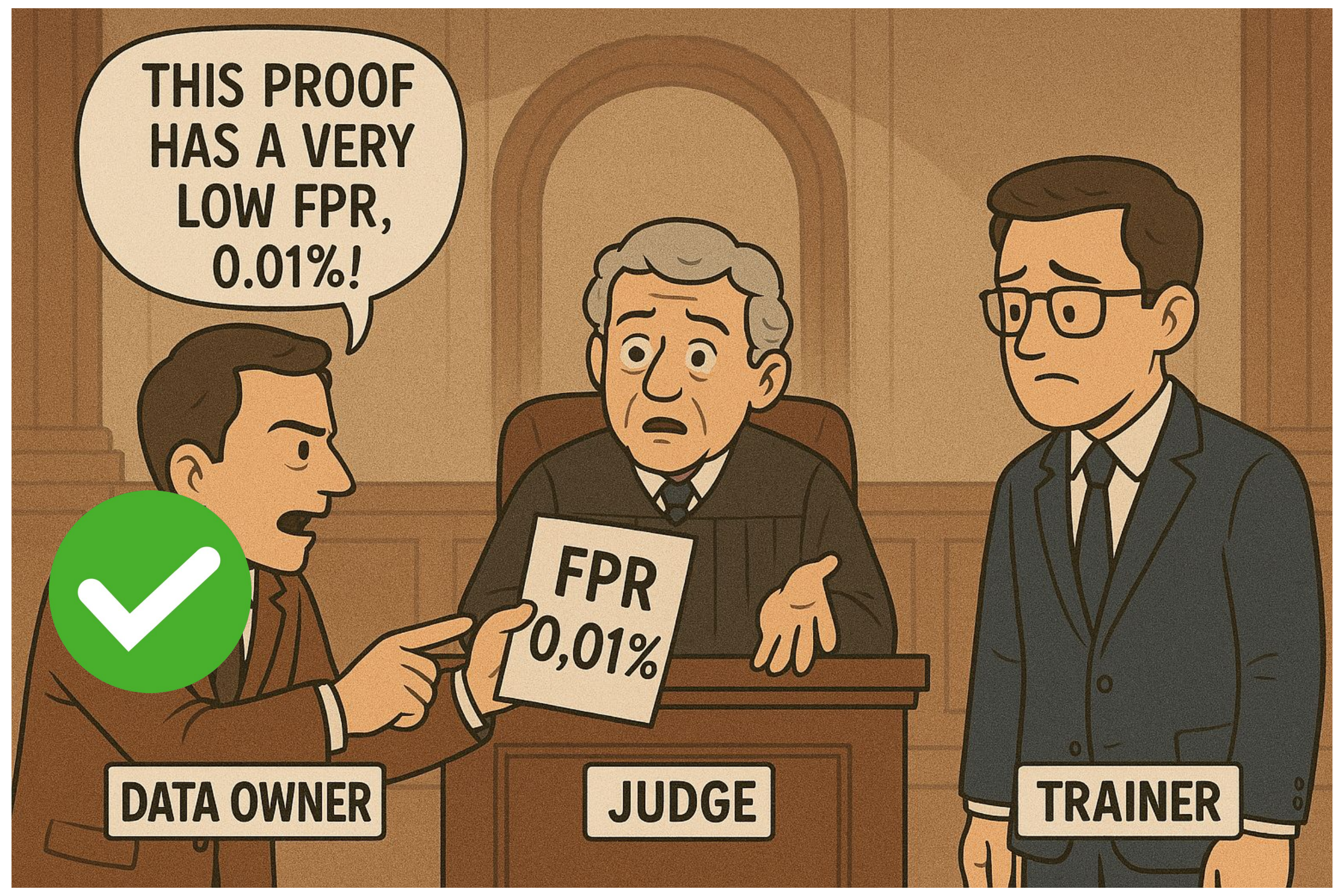
Training Data Proof



Repurpose MIA for it


THE MODEL HAS A LOW LOSS ON DATA!
GIVE ME RELIABLE PROOFS!
SO WHAT? IT'S JUST A COINCIDENCE.
DATA OWNER
JUDGE
TRAINER

P(wrongly accusing) is low



2/ MIA cannot bound the attack's false positive rate (FPR)

Hypothesis test:

null hypothesis H_0: the data x was not in the training set of model f.

$$\mathrm{FPR} = \Pr_{\boxed{f\sim\texttt{Train}(D_0)}}[T(f,x) \in S \mid H_0]$$

But we cannot sample from null hypothesis.

- The training data for models (e.g., GPT-4) is undisclosed.
- Retraining new models is almost impossible.

3/ Failed attempts and potential solutions

Low FPR?

✗

- Collecting Non-member Data a Posteriori
- Collecting Indistinguishable Non-members
- Dataset inference on Held-out Counterfactuals

✓

- Injecting Random Canaries
- Watermarked Training Data
- Verbatim Data Extraction

More details!