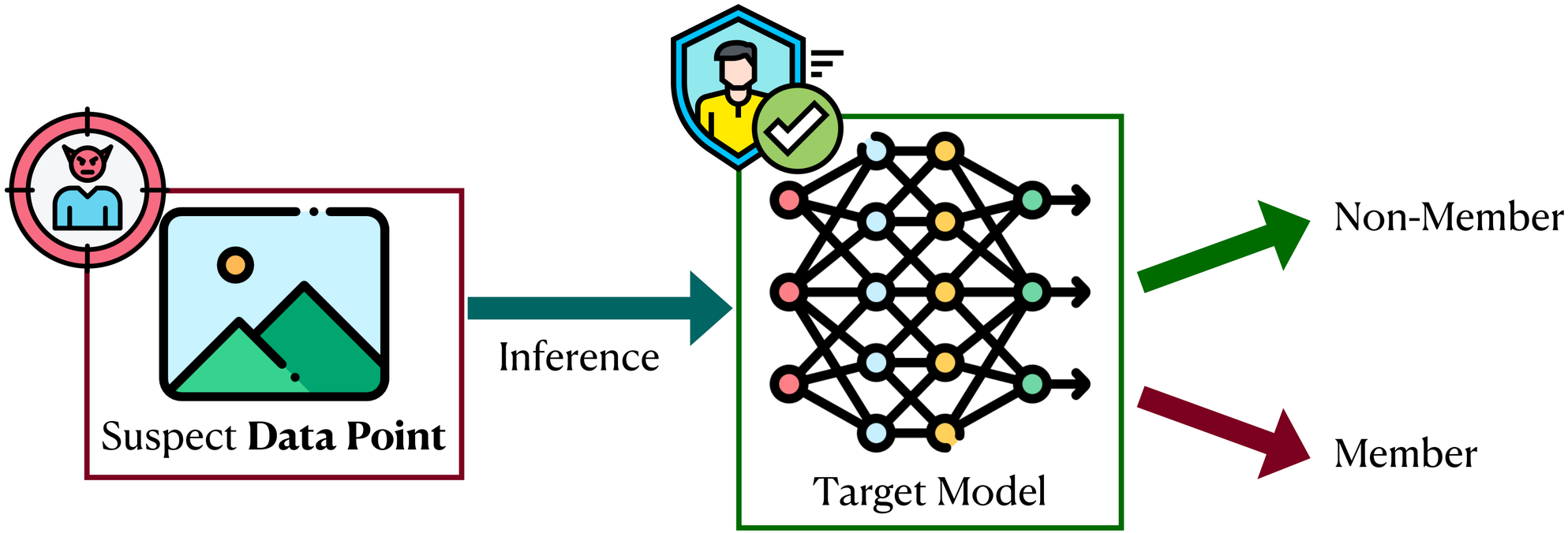
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Data Provenance

Tutorial Session 6 - Trustworthy Machine Learning

Adam Dziedzic & Franziska Boenisch
Nima Dindarsafa | 10.06.2026

Membership Inference Attack (Recap)



MIA Failure Point

Why does MIA **fail** for Large Language Models?

	Wikipedia					Github					Pile CC					PubMed Central				
# Params	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne
70M	.503	.504	.494	.508	**.510**	.629	.584	.627	**.648**	.635	.494	.489	**.503**	.495	.489	.502	**.516**	.510	.502	.485
160M	.504	**.515**	.488	.514	.513	.638	.591	.634	**.656**	.638	.497	.497	**.503**	.498	.496	.500	**.516**	.504	.500	.486
1.4B	.510	**.544**	.506	.518	.518	.656	.587	.654	**.670**	.650	.500	**.525**	.509	.502	.499	.496	**.530**	.505	.500	.490
2.8B	.516	**.565**	.511	.522	.517	.707	.657	.708	**.717**	.698	.501	**.537**	.509	.503	.502	.498	**.536**	.502	.500	.497
6.9B	.514	**.571**	.512	.521	.514	.672	.573	.675	**.684**	.654	.511	**.564**	.516	.512	.505	.504	**.552**	.508	.504	.497
12B	.516	**.579**	.517	.524	.520	.678	.559	.683	**.690**	.660	.516	**.582**	.521	.517	.514	.506	**.559**	.512	.506	.497

	ArXiv					DM Math					HackerNews					The Pile				
# Params	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne	LOSS	Ref	min-*k*	zlib	Ne
70M	**.506**	.481	.499	.495	.496	.492	**.520**	.495	.485	.481	.494	.495	**.507**	.497	.506	.503	**.511**	.508	.506	.499
160M	**.507**	.486	.501	.500	**.507**	.490	**.523**	.493	.482	.489	.492	.490	.497	.497	**.505**	.502	**.511**	.506	.505	.499
1.4B	**.513**	.510	.511	.508	.511	.486	**.512**	.497	.481	.465	.503	**.514**	.509	.502	.504	.504	**.521**	.508	.507	.504
2.8B	.517	**.531**	.522	.512	.519	.485	**.504**	.497	.482	.467	.510	**.549**	.518	.507	.513	.507	**.530**	.512	.510	.506
6.9B	.521	**.538**	.524	.516	.519	.485	**.508**	.496	.481	.469	.513	**.546**	.528	.508	.512	.510	**.549**	.516	.512	.510
12B	.527	**.555**	.530	.521	.519	.485	**.512**	.495	.481	.475	.518	**.565**	.533	.512	.515	.513	**.558**	.521	.515	.511

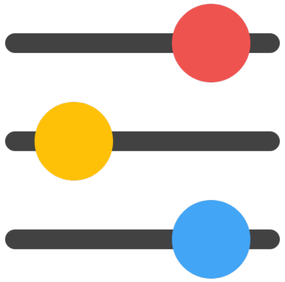
AUC of different MIA methods against Pythia model across different datasets. [1]

MIA has almost **near random performance**!

[1] Do Membership Inference Attacks Work on Large Language Models?, Duan et al.

MIA Failure Point

Why does MIA **fail** for Large Language Models?

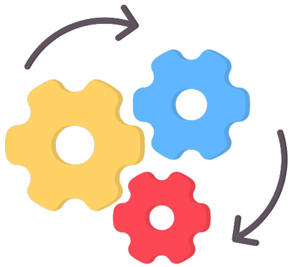


Number of Model Parameters

LLMs have billions of parameters.

Training Dataset Size

LLMs are usually trained on billions to trillions of tokens.

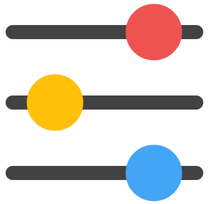


Training Procedure

LLMs are usually trained with a single epoch on the whole dataset.

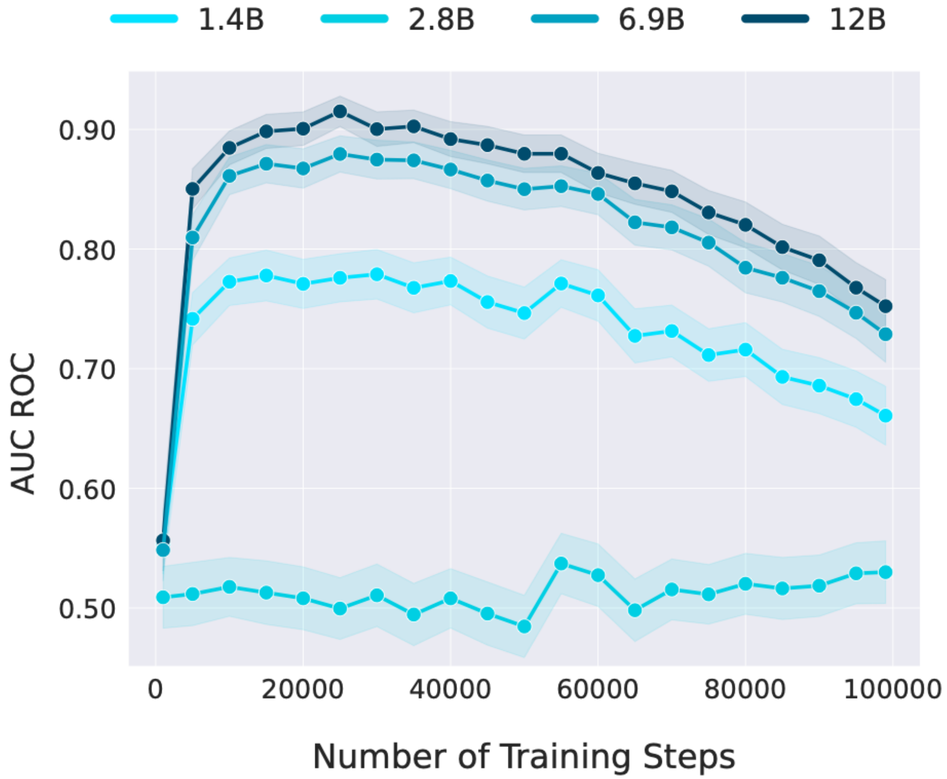
MIA Failure Point

Why does MIA **fail** for Large Language Models?



Number of Model Parameters

Training Dataset Size

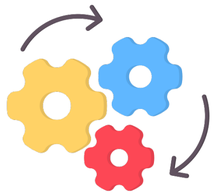


Performance spikes greatly before gradually decreasing as the amount of training data seen increases. [1]

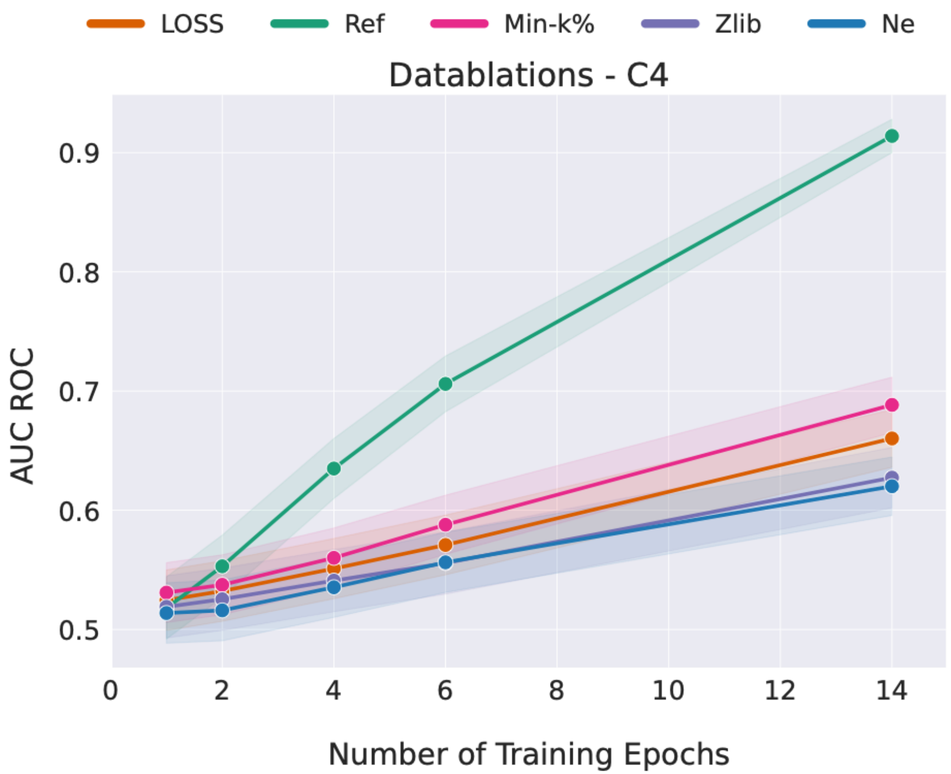
[1] Do Membership Inference Attacks Work on Large Language Models?, Duan et al.

MIA Failure Point

Why does MIA **fail** for Large Language Models?



Training Procedure

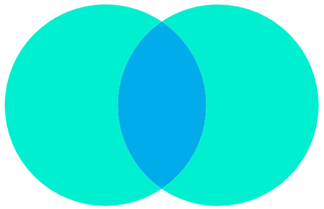


MIA performance increases linearly with the number of effective epochs. [1]

[1] Do Membership Inference Attacks Work on Large Language Models?, Duan et al.

MIA Failure Point

Why does MIA **fail** for Large Language Models? Yet, another reason!



Inherent Ambiguity in MIA

Natural language documents commonly have repeating text. → This leads to substantial overlap between members and non-members. → Higher overlap between members and non-members increases the MIA difficulty!

Dataset Inference

Four stages of Dataset Inference

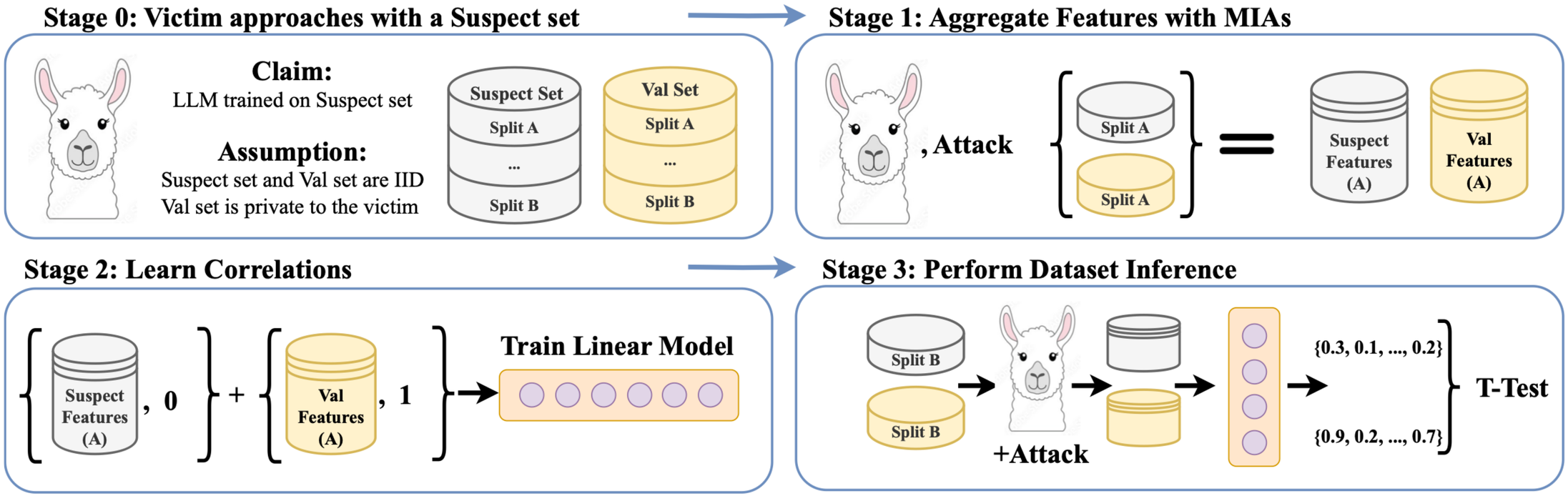
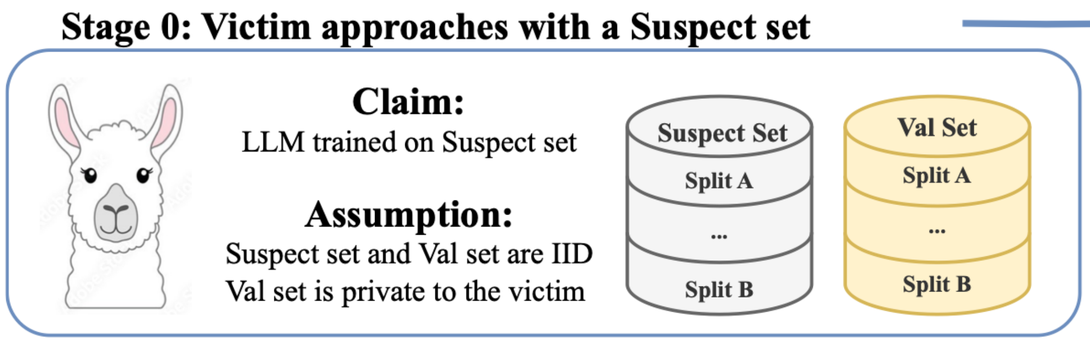


Image Source: [2]

[2] LLM Dataset Inference Did you train on my dataset?, Maini et al.; https://arxiv.org/pdf/2406.06443

Dataset Inference

Four stages of Dataset Inference



A victim approaches an LLM provider with 2 datasets:

- Suspect set
- Held-out set (**IID** with the suspect set and **private** to the victim)

Both sets are divided into non-overlapping splits.

Image Source: [2]

[2] LLM Dataset Inference Did you train on my dataset?, Maini et al.; https://arxiv.org/pdf/2406.06443

Dataset Inference

Four stages of Dataset Inference

Pass the split A of the victim's data (from both suspect and held-out) to LLM and aggregate MIA features.

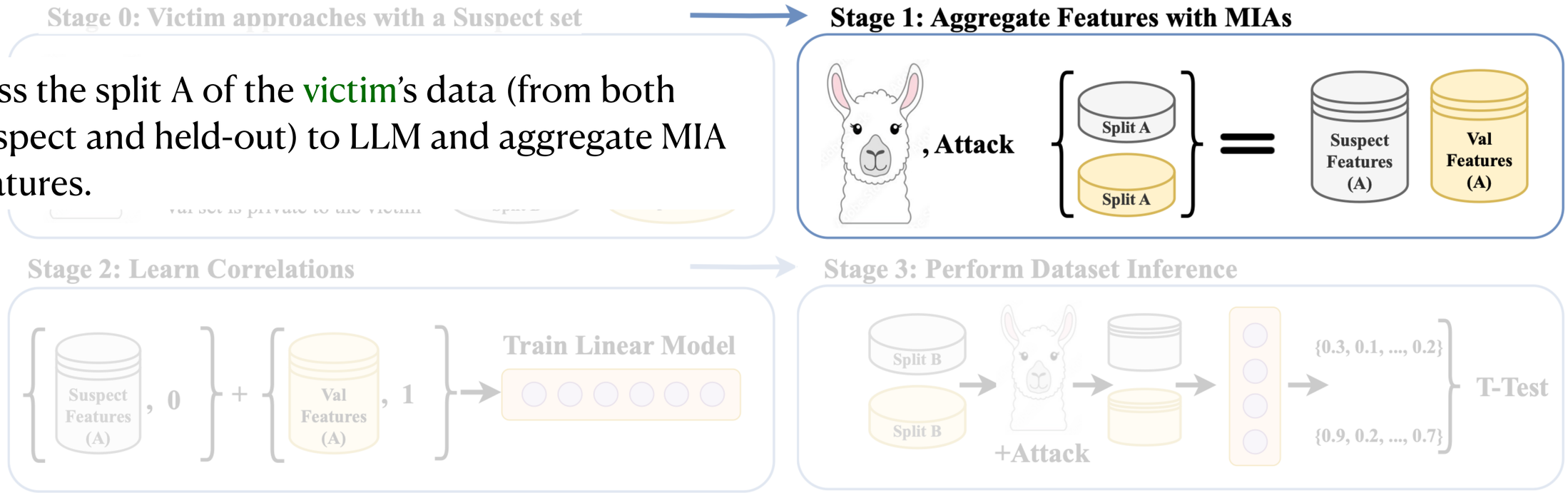
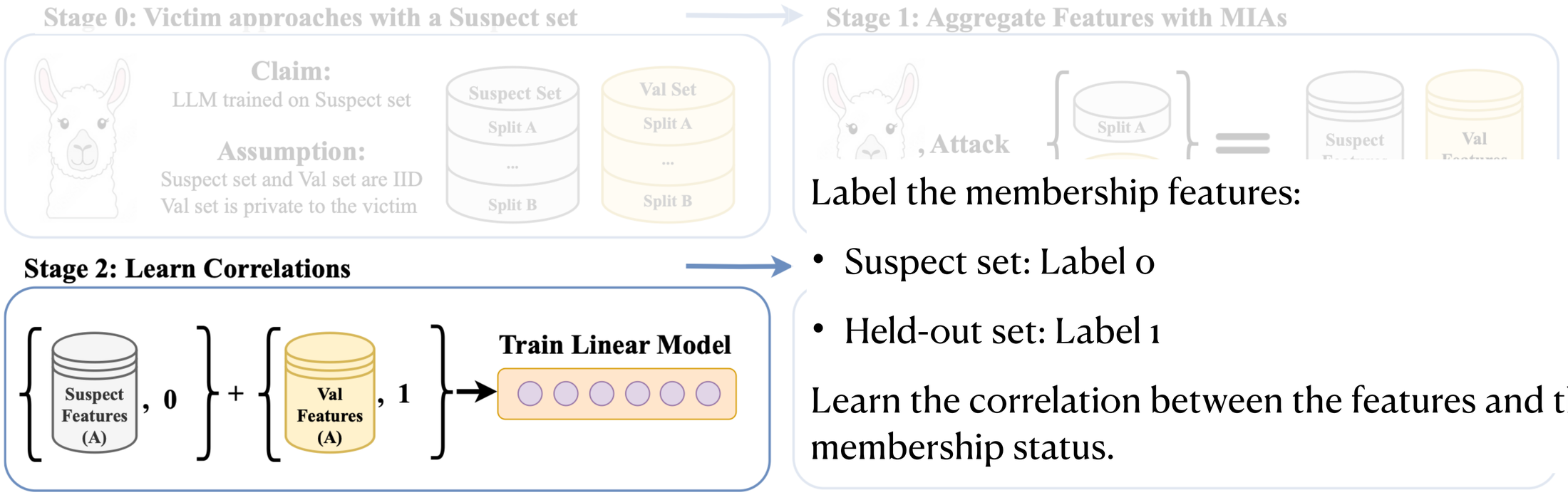


Image Source: [2]

[2] LLM Dataset Inference Did you train on my dataset?, Maini et al.; https://arxiv.org/pdf/2406.06443

Dataset Inference

Four stages of Dataset Inference



Label the membership features:

- Suspect set: Label 0
- Held-out set: Label 1

Learn the correlation between the features and the membership status.

Image Source: [2]

[2] LLM Dataset Inference Did you train on my dataset?, Maini et al.; https://arxiv.org/pdf/2406.06443

Dataset Inference

Four stages of Dataset Inference

Use the split B of the suspect and held-out sets and perform the following steps:

1. Extract the membership features
2. Aggregate them using the learned linear model
3. Apply statistical T-test

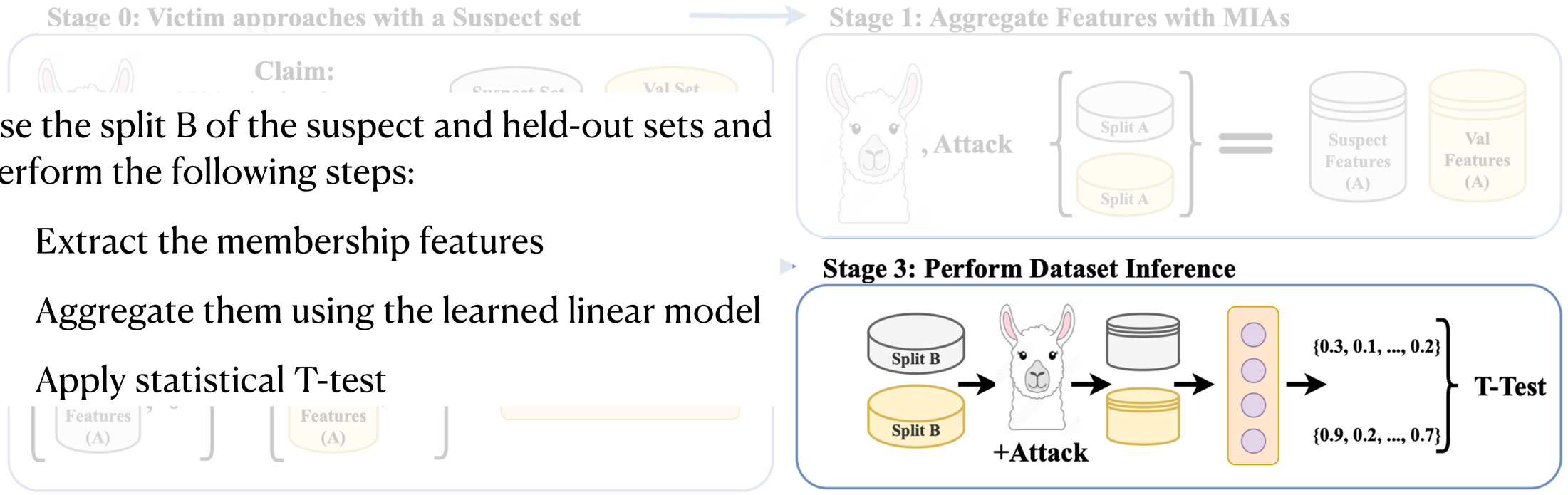
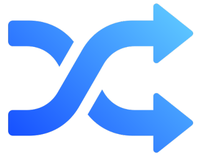


Image Source: [2]

[2] LLM Dataset Inference Did you train on my dataset?, Maini et al.; https://arxiv.org/pdf/2406.06443

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?



Perplexity-based

Perturbation-based

Reference-based

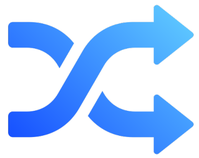
Min-k% Prob

Zlib-ratio

* Note that this list is not exhaustive and it only indicates some of the commonly used features.

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?



Perplexity-based

Intuition: A language model should assign higher probability or lower perplexity to the text it has seen during training.

$$\text{Perplexity} = b^{-\sum_x p(x)\log_b p(x)}$$

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?

Perturbation-based

Intuition: If a text sequence was used during training, the model may prefer the original version over slightly perturbed version of the same text.

Perturbations include:

- Changing whitespace
- Replacing word with synonyms
- Adding character-level typo
- Randomly deleting words

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?

Reference-based

Intuition: Compare how well the suspect model predicts the text relative to the reference model that is assumed not to have trained on that text.

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?

Zlib-ratio

Intuition: Some text receives low perplexity simply because it is repetitive or easy to compress, not because it was in the model's training set.

$$\text{Zlib Ratio} = \frac{\text{Model Perplexity(x)}}{\text{zlib compression score(x)}}$$

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?

Min-k% Prob

Intuition: Non-member example is more likely to include a few outlier words with high negative log-likelihoods or low probability, while a member example is less likely to include words with high negative log-likelihood. [3]

$$\text{Min-K\% Prob} = \frac{1}{E} \sum_{x_i \in \text{Min-K\\%}(x)} \log p(x_i \mid x_1, \ldots, x_{i-1})$$

[3] Detecting Pre-training Data from Large Language Models, Shi et al.; https://arxiv.org/pdf/2310.16789

Dataset Inference - Features

[**Stage 1**]: Which features are used for Dataset Inference?

Min-k% Prob

Given the following list of NLL:

[-0.12, -1.75, -0.43, -2.80, -0.95, -3.40, -0.30, -1.20, -4.10, -0.65]

With 10 tokens, Min-20% Prob corresponds to the average of 2 lowest-probability tokens: -4.10, -3.40

$$\text{Min-20\% Prob} = \frac{-4.10 + (-3.40)}{2} = -3.75$$

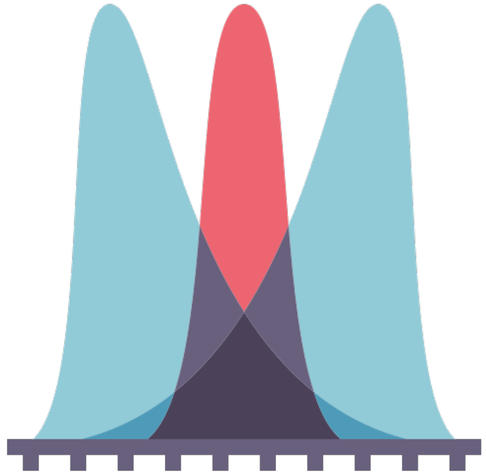
[3] Detecting Pre-training Data from Large Language Models, Shi et al.; https://arxiv.org/pdf/2310.16789

Dataset Inference - Assumptions

Why **IID** held-out set?

IID held-out assumption violated → **distribution shift**

- This assumption is required → The held-out set is used as the **non-member baseline** for calibration.
- In case of violation → DI may confuse **distribution shift** with **membership signal**.
- In this case DI is only classifying two distributions rather than true membership.
- This will lead to **high false positive rate**.

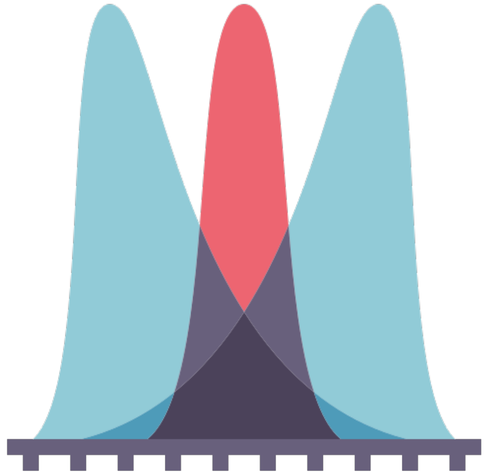


Dataset Inference - Distribution Shift

Distribution Shift example

Wikipedia cut-off-date setup:

- Suspect set: Wikipedia articles written **before 2023**
- Held-out set: Wikipedia articles written **after 2023**



For a model released in 2023 this dataset might seem a good choice. However,

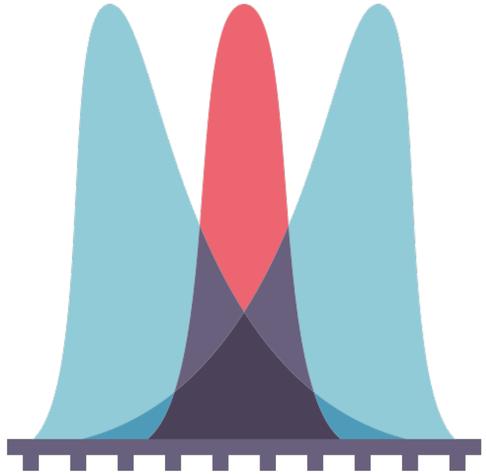
- These two sets are not **IID**!
- Post-2023 articles may contain newer events, terminology or writing styles that were not present in the older Wikipedia articles.

Dataset Inference - Distribution Shift

Distribution Shift detection

Train a simple classifier using features that describe the text, such as:

- TF-IDF vectors
- Sentence embeddings
- Vocabulary statistics



If the classifier performs near random guessing → Most likely same distribution

Otherwise, there is a **distribution shift**!

Dataset Inference - Held-out Generation

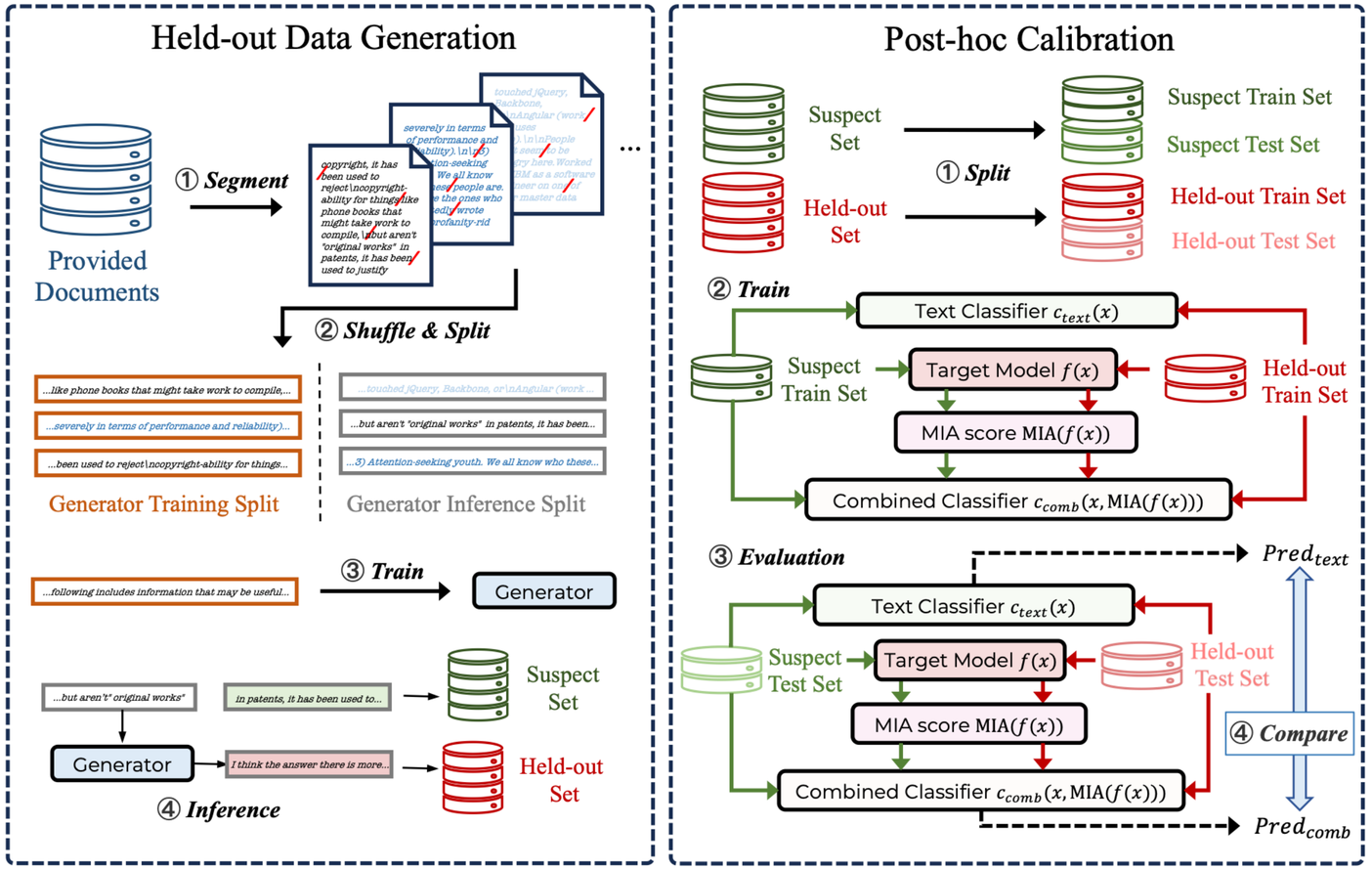


Image Source: [4]

[4] Unlocking Post-hoc Dataset Inference with Synthetic Data, Zhao et al.; https://arxiv.org/pdf/2506.15271

Dataset Inference - Held-out Generation

How to reduce **distribution shift** between generated held-out and the suspect set?

- Train a generator on the suspect dataset itself using a suffix completion task.
- Use short sequences limited up to 64 tokens.
- **Post-hoc calibration** to avoid natural-vs-synthetic shift.
  - Train a text-only classifier
  - Train a membership aware classifier (using both text and MIA features)
  - If membership aware classifier performs better than text-only classifier → Extra signal is attributed to membership rather than generation artifacts.

[4] Unlocking Post-hoc Dataset Inference with Synthetic Data, Zhao et al.; https://arxiv.org/pdf/2506.15271

Questions?