Robust Concept Erasure via Kernelized Rate-Distortion Maximization



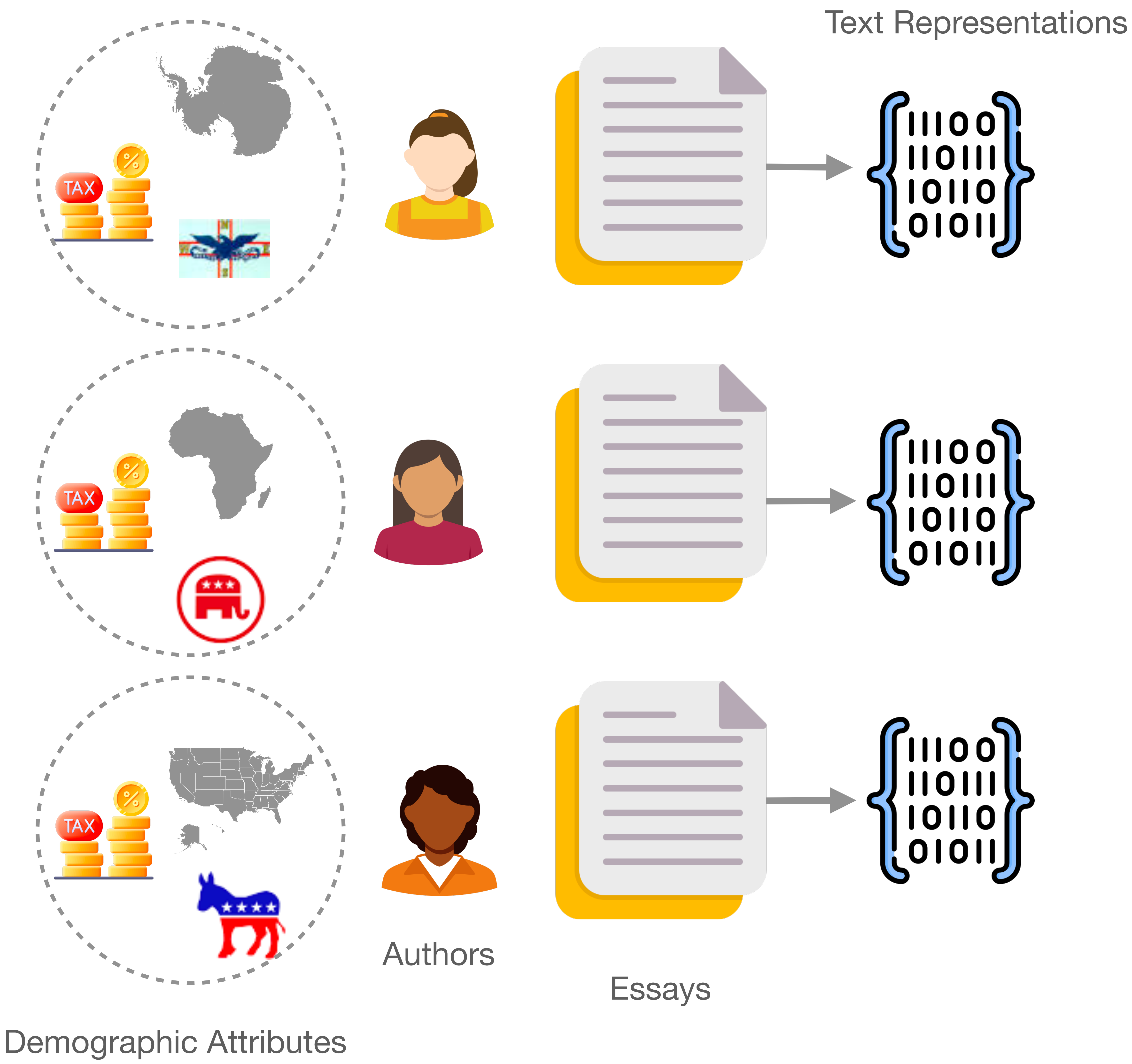


Somnath Basu Roy Chowdhury Nicholas Monath Avinava Dubey Amr Ahmed Snigdha Chaturvedi

What is a Concept?

A *concept* (random variable), A, which can be inferred from a set of data representations.

What is a Concept?



Given a dataset X, where each instance is a text representation of a student essay.

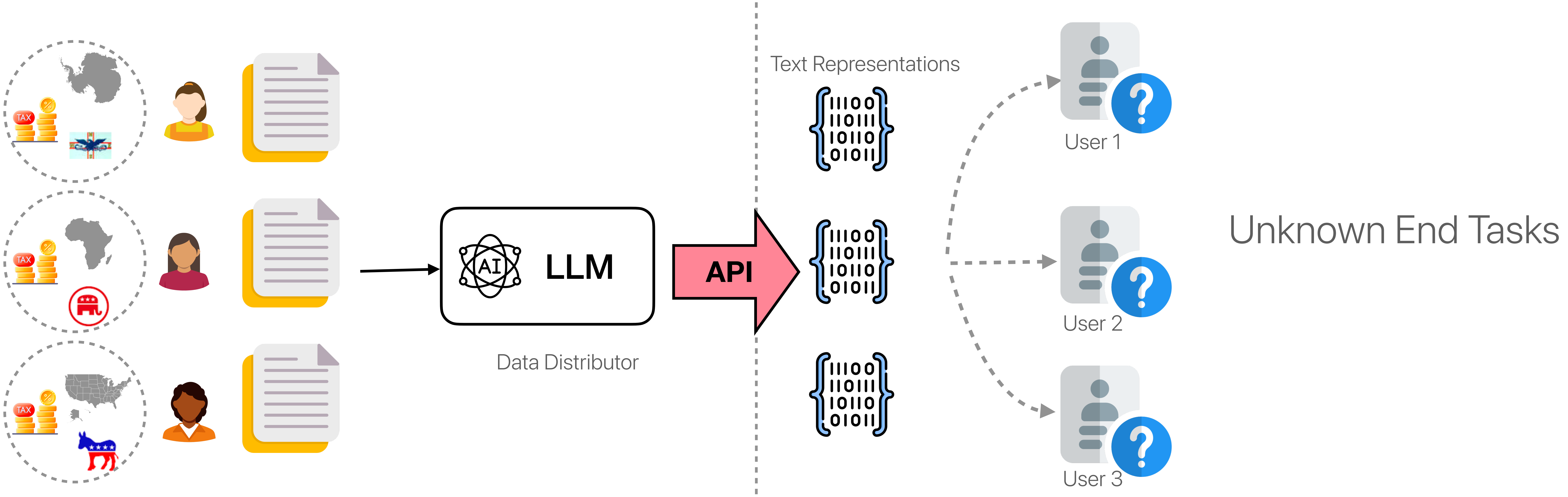
Concept, A = Country

Concept, A = Income

Concept, A = Political Affiliation

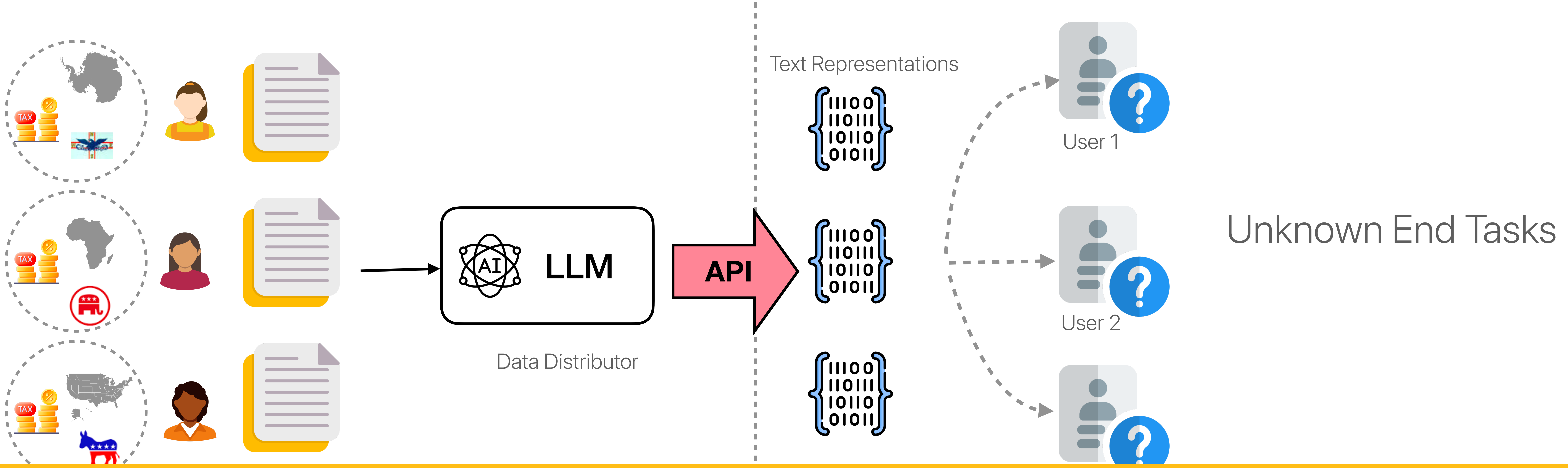
Catering to Unknown Applications

- Developers often rely on black-box LLM representations to power their applications
- Data distributors may need to remove unintended concepts encoded in representations to prevent wide-spread unfairness in downstream tasks



Catering to Unknown Applications

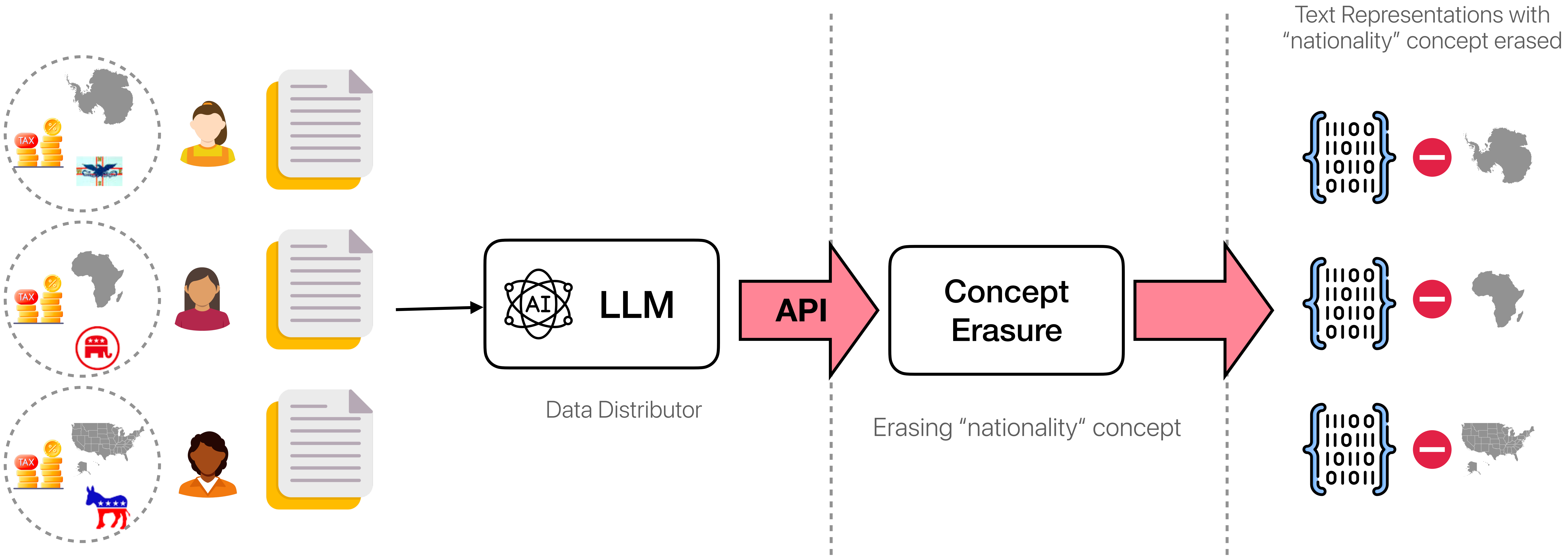
- Developers often rely on black-box LLM representations to power their applications
- Data distributors may need to remove unintended concepts encoded in representations to prevent wide-spread unfairness in downstream tasks



Problem: Representations may contain unwanted concept that can impact end tasks.

Concept Erasure

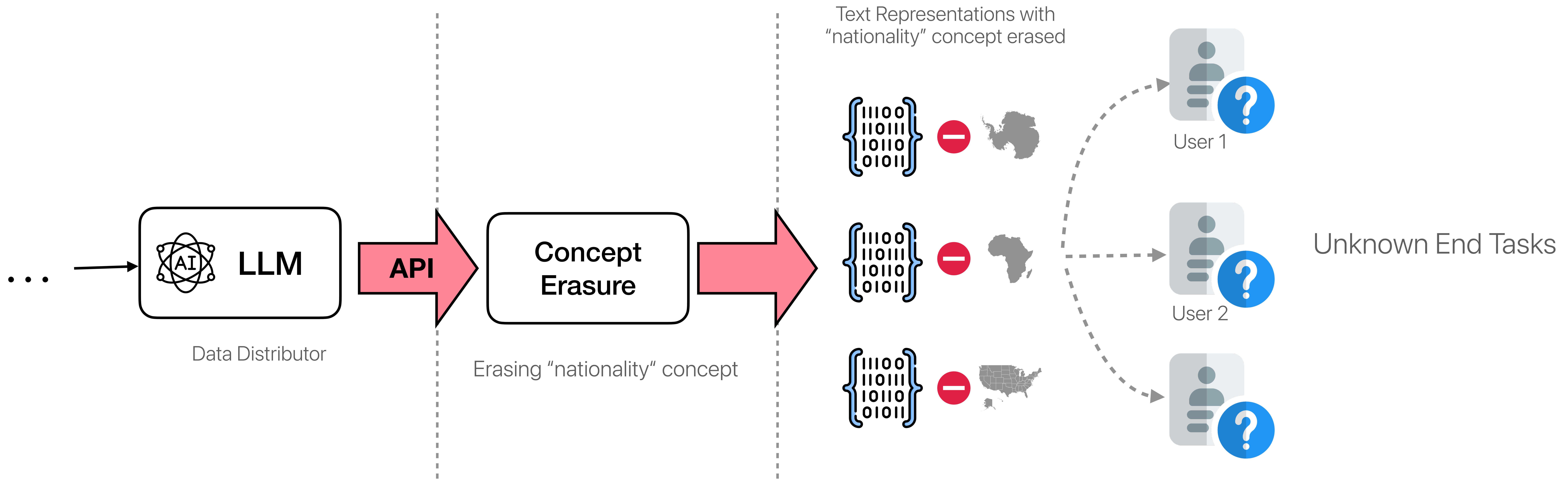
- Concept Erasure is the process of removing a concept from a representation set.



Concept Erasure

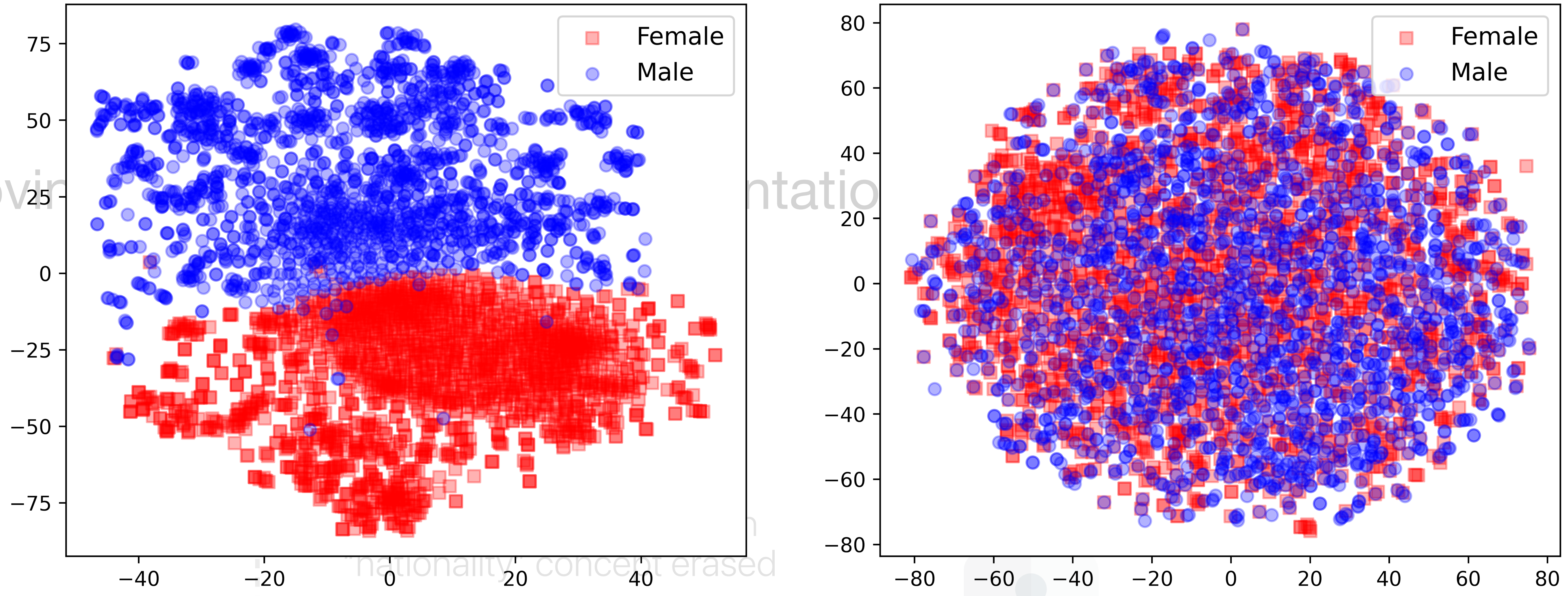
- Concept Erasure is the process of removing a concept from a representation set.

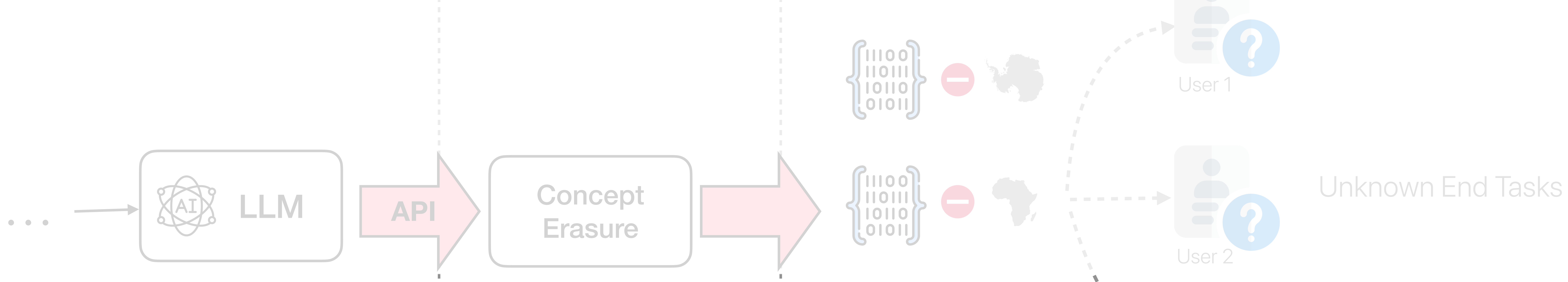
Concept Erasure Provides Representations that don't reveal concept to *any* end task.



Concept Erasure

- Concept Erasure is the process of removi… …ntatio…

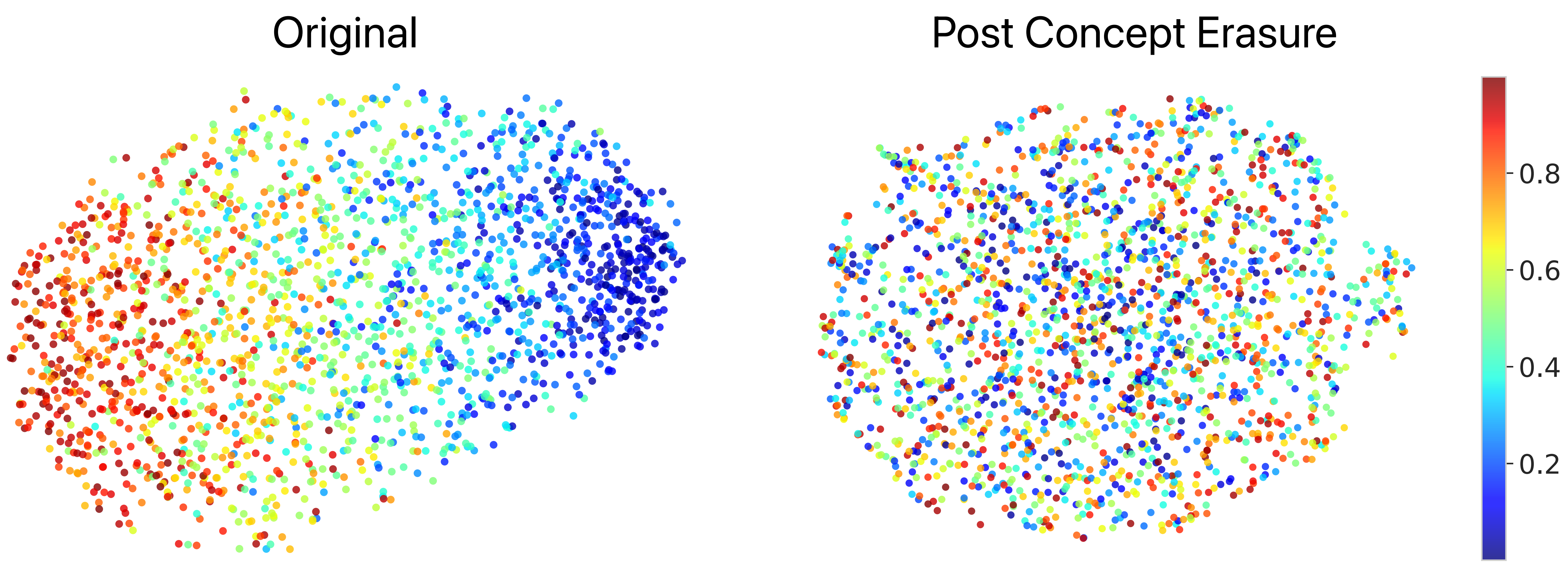




Concept Erasure (INLP, RLACE, KernelCE [Ravfogel et al., 2022(a,b,c)], FaRM [Chowdhury et al., 2022]) provides representations that don't reveal concept to *any* end task.

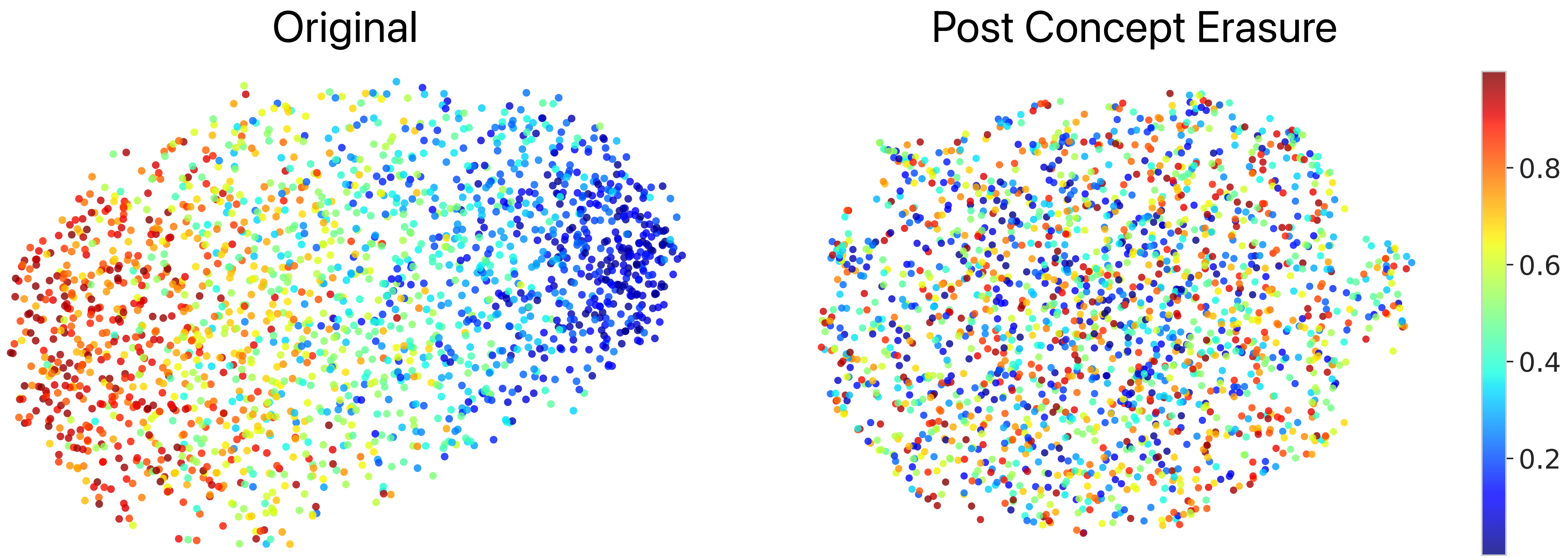
Concept Erasure

- In general, concepts can be categorical, continuous, and vector-valued
- Depending on their nature, they can be encoded in the representations differently
- Prior works do not consider the erasure of continuous or vector-valued concepts



Concept Erasure

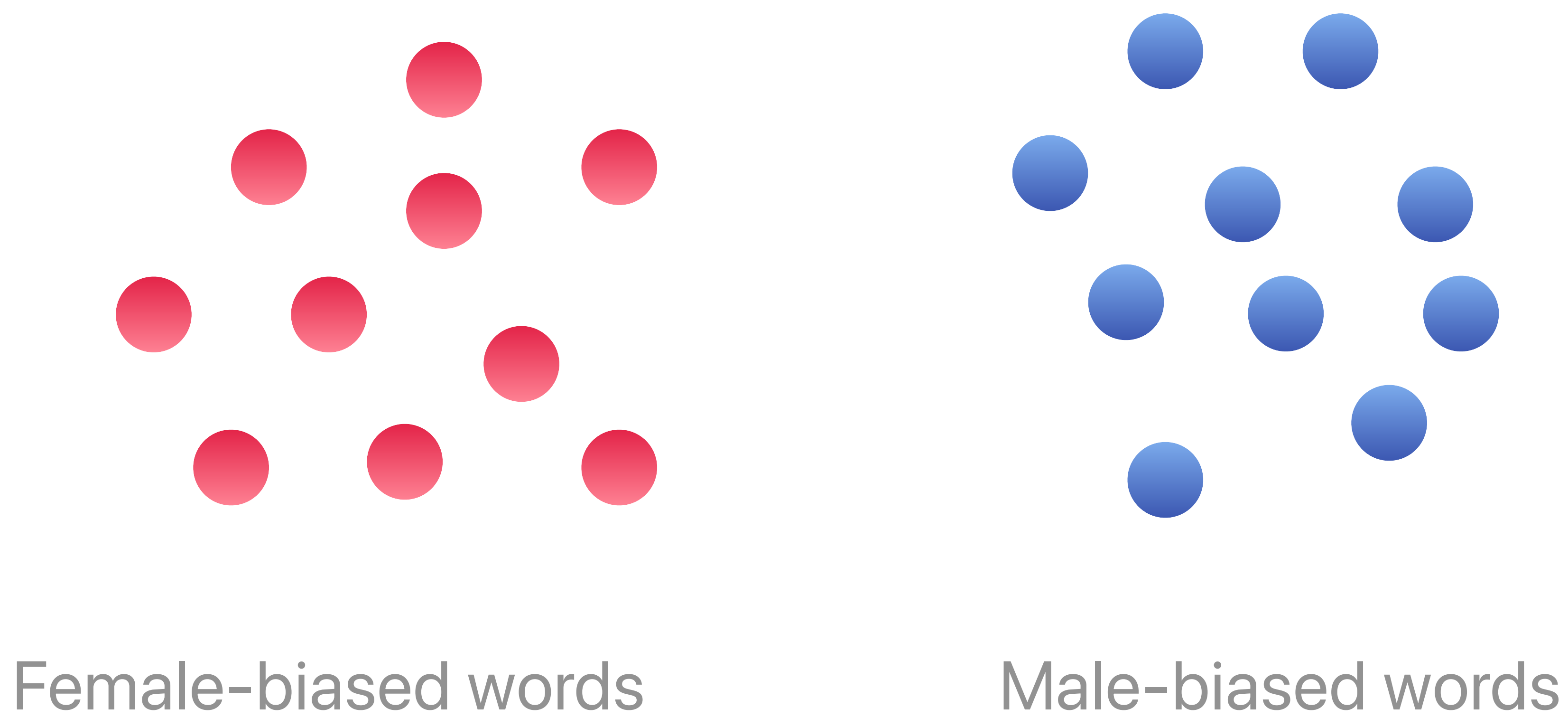
- In general, concepts can be categorical, continuous, and vector-valued
- Depending on their nature, they can be encoded in the representations differently
- Prior works do not consider the erasure of continuous or vector-valued concepts



The concept can be continuous-valued like income and age of a person.

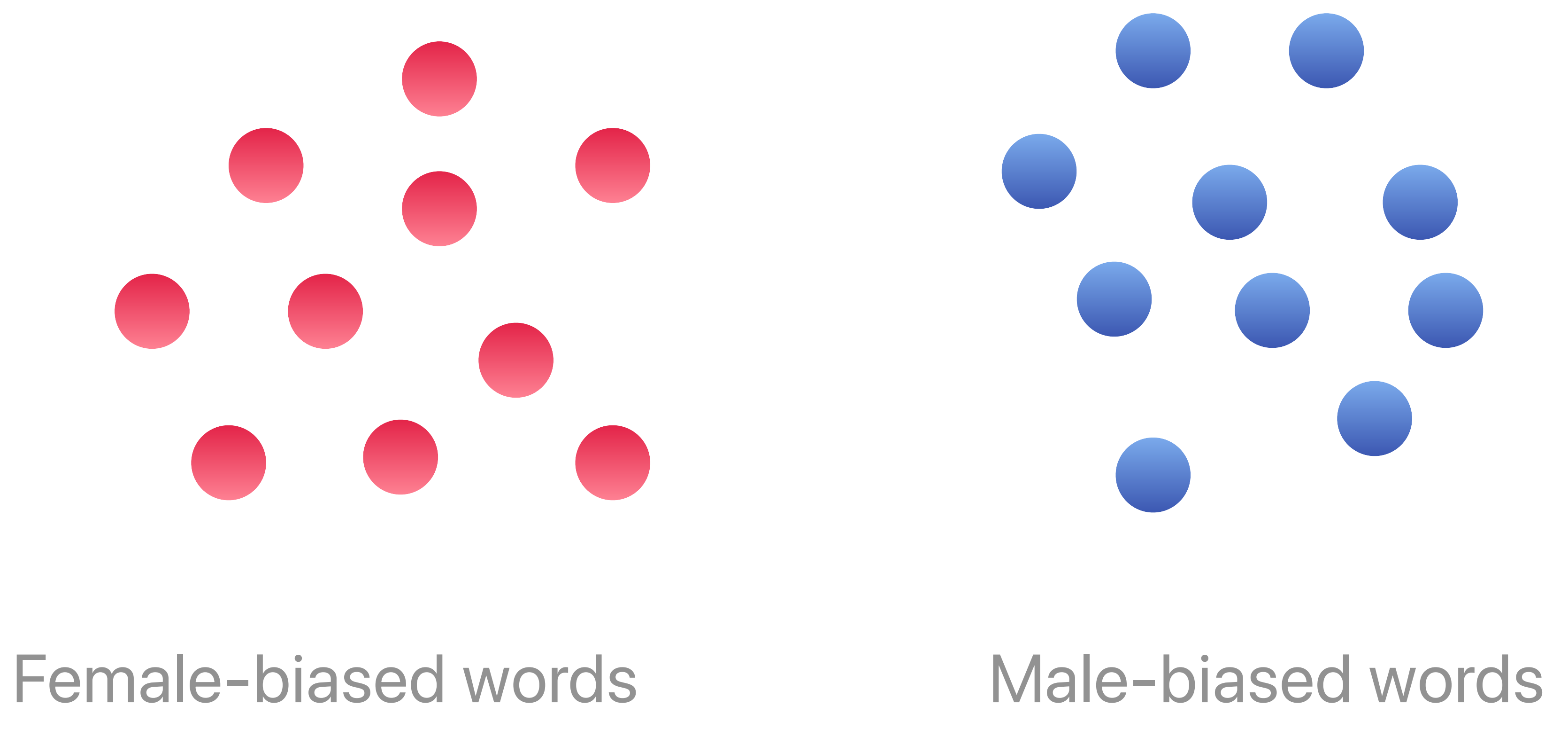
Information in high dimensions

- Information is stored as distances in high-dimensional spaces



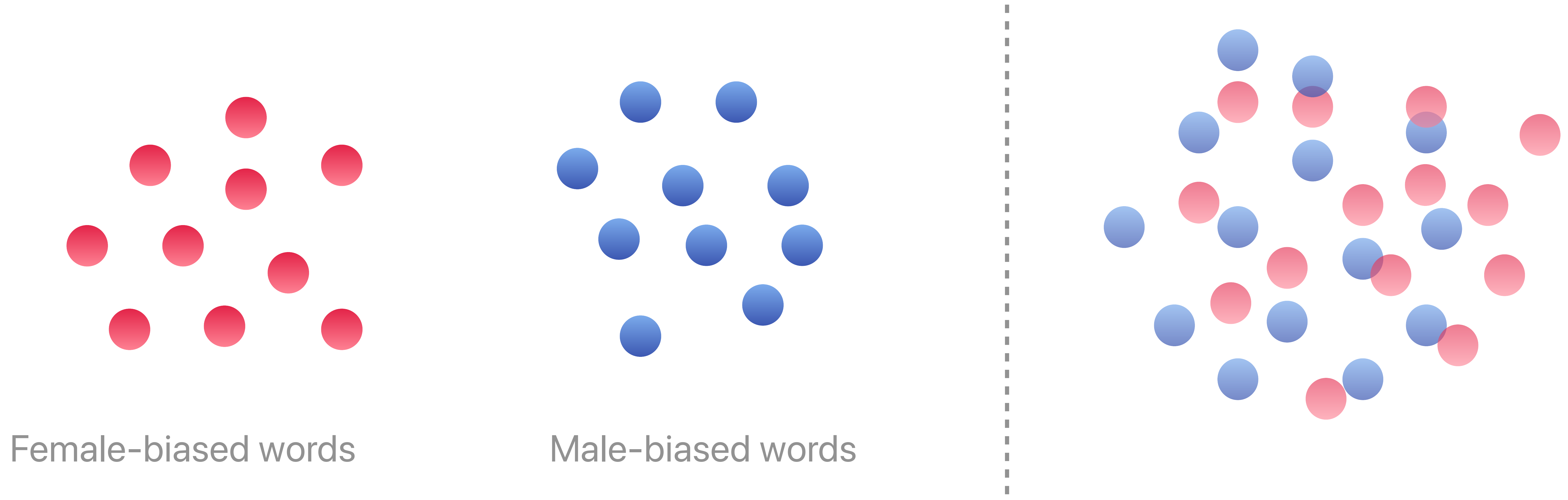
How do we nullify a specific concept?

- Concept to be deleted: Gender



How do we nullify a specific concept?

- Concept to be deleted: Gender

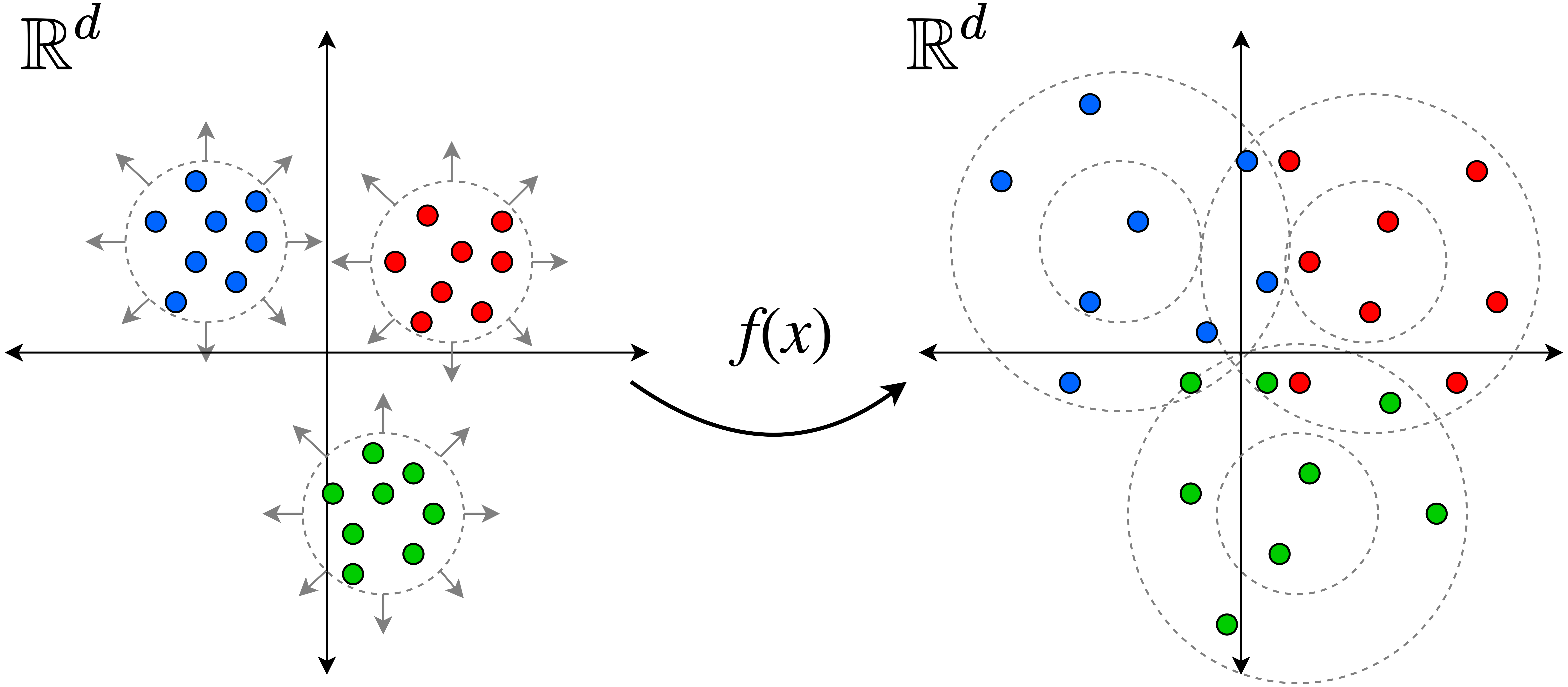


Kernelized Rate Distortion Maximization (KRaM)

- Learn parametric encoder f of data representations X to erase concept A
- Recipe: Rate Distortion [Yu et al. 2020, Chowdhury et al. 2022] for erasing concepts
- Kernelized-version of the rate distortion function to allow generic concept erasure
- Capture the information retained after erasure using a novel alignment measure

Recipe?

- (Chowdhury et al. 2022) proposed a recipe for categorical concept erasure



Recipe?

- Given a feature space with multiple subspaces: $\mathscr{F} = \{F_1, \ldots, F_n\}$
- The proposed recipe can be formalized as below:

$$\max_f \sum_i \mathrm{Vol}(F_i)$$

- However, this works only for categorical concepts where you've well-defined subspaces

Measuring Volume — Rate Distortion

- Rate-distortion measures the total number of binary bits required to encode a set of representations $Z \in \mathbb{R}^d$

$$R(Z) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{n\epsilon^2} Z Z^T \right)$$

Kernelized Rate Distortion

- We introduce a kernelized version of the rate-distortion function:

$$R(Z \quad \mathbf{K}) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{n\epsilon^2} Z Z^T \odot \mathbf{K} \right)$$

- The kernel $\mathbf{K}$ captures the similarity space of concepts $\mathbf{K}_{ij} \propto 1/d(a_i, a_j)$, where $a_i, a_j \in A$

Kernelized Rate Distortion

- We introduce a kernelized version of the rate-distortion function:

$$R(Z \quad \mathbf{K}) = \frac{1}{2}\log_2 \det\left(I + \frac{d}{n\epsilon^2}ZZ^T \odot \mathbf{K}\right)$$

- Maximizing this quantity encourages similar representations in the concept space to be dissimilar, thereby resulting in concept erasure

Kernelized Rate Distortion

- We introduce a kernelized version of the rate-distortion function:

$$R(Z \;\; \mathbf{K}) = \frac{1}{2}\log_2 \det\left(I + \frac{d}{n\epsilon^2} ZZ^T \odot \mathbf{K}\right)$$

- Theoretical result:

$$R(Z) \leq R(Z \;\; \mathbf{K}) \leq \frac{n}{2}\log_2\left(1 + \frac{d}{n\epsilon^2}\right)$$

Kernelized Rate Distortion Maximization (KRaM)

- Formulating the objective function:

$$\max_f \sum_i \mathrm{Vol}(F_i), \text{ subject to } \mathrm{Vol}(\mathscr{F}) = \text{const.}$$

$$\max_f R(Z \quad \mathbf{K}), \text{ subject to } R(Z) = b$$

$$\max_f R(Z \quad \mathbf{K}) - \lambda \quad R(Z) - b$$

Kernelized Rate Distortion Maximization (KRaM)

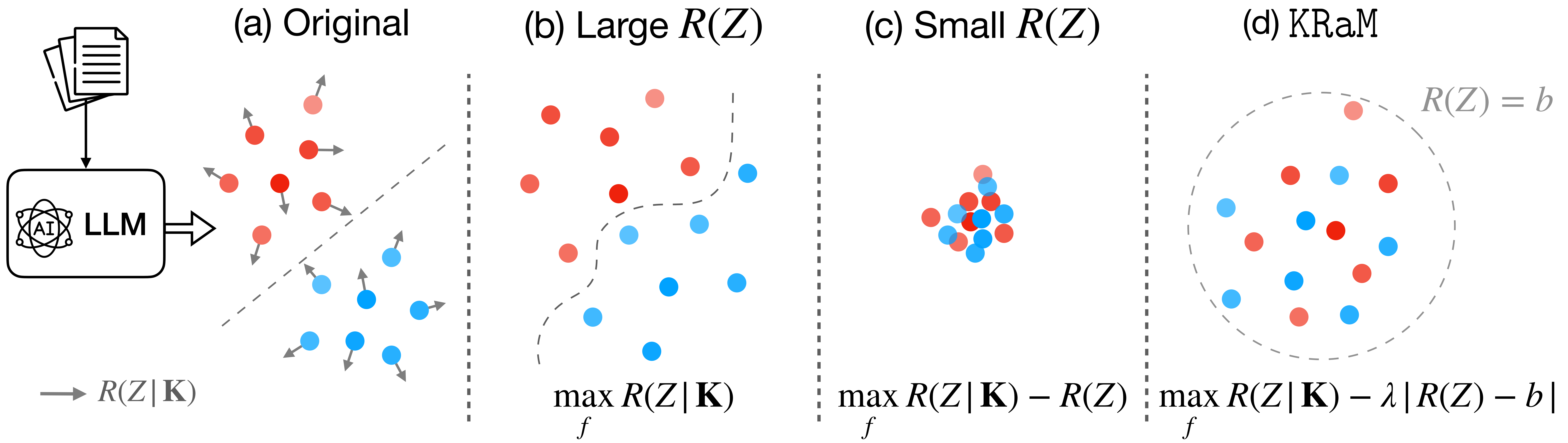
- Formulating the objective function:

$$\max_{f} \sum_{i} \mathrm{Vol}(F_i), \text{ subject to } \mathrm{Vol}(\mathscr{F}) = \text{const.}$$

$$\max_{f} R(Z \quad \mathbf{K}), \text{ subject to } R(Z) = b$$

Erasure objective: $\max_{f} R(Z \quad \mathbf{K}) - \lambda \quad R(Z) - b$

KRaM



LLM
AI
$R(Z|\mathbf{K})$
(a) Original
(b) Large $R(Z)$
(c) Small $R(Z)$
(d) KRaM
$R(Z) = b$
$\max_f R(Z|\mathbf{K})$
$\max_f R(Z|\mathbf{K}) - R(Z)$
$\max_f R(Z|\mathbf{K}) - \lambda|R(Z) - b|$

Beyond Categorical Concepts

- KRaM doesn't make assumptions on the nature of the underlying concept
- It only depends on the kernel function: $\mathbf{K}_{ij} = k(a_i, a_j)$
- The kernel function accepts any form of concepts (a_i): categorical, continuous or vector-valued.

We observe that the representation positions are indicative of the concepts (shown in colours).

Measuring Alignment

- To measure how concept erasure impact other information, we compute the "alignment" between the learned representations $f(X)$ and original representations X

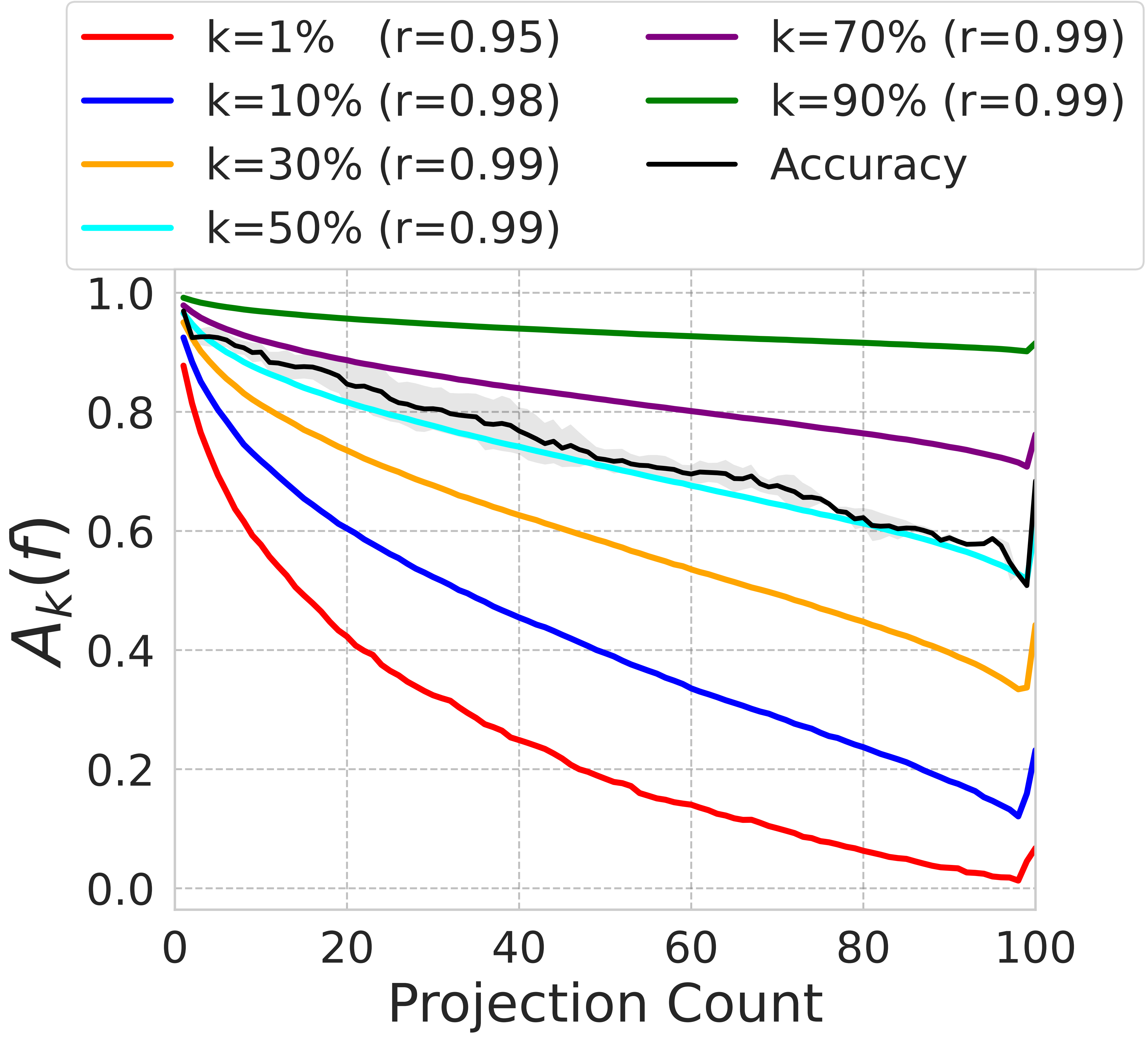
- We propose an alignment score $A_k(f)$:

$$A_k(f) = \frac{1}{k}\mathbb{E}_x\left[\mathrm{knn}(x) \cap \mathrm{knn}(f(x))\right]$$

- The above score quantifies how much the nearest neighbour structure is retained

Measuring Alignment

- Theoretical result: $A_k(f) \in \left[\frac{k}{n}, 1\right]$

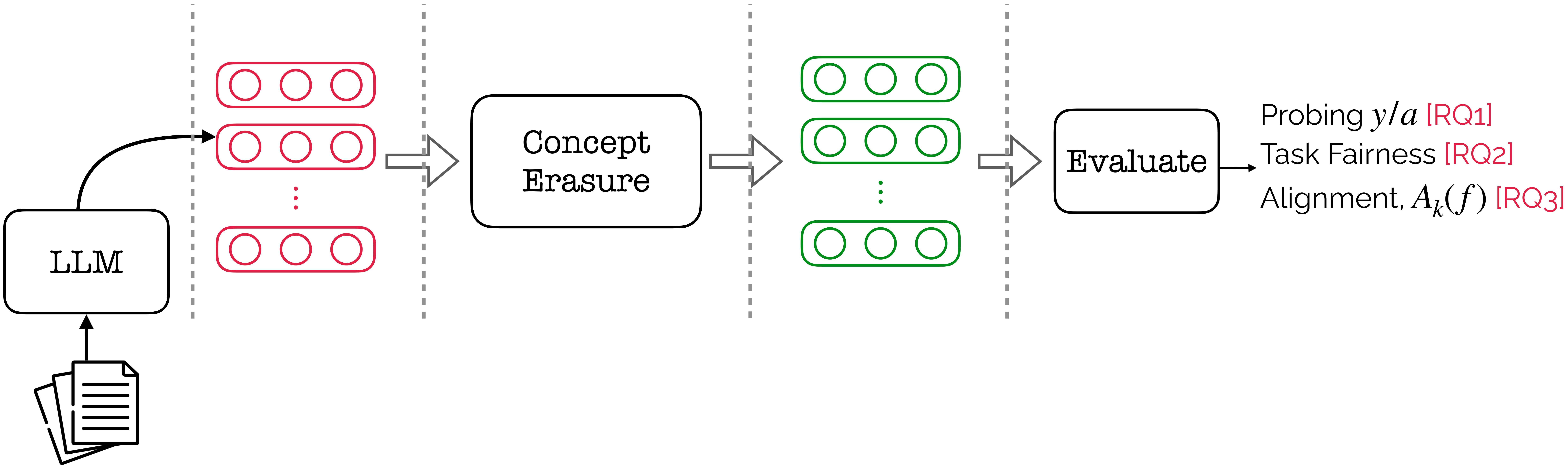


Experimental Setup

Through experiments, we seek to answer the following research questions:

- [RQ1] Can the erased concept be predicted after concept erasure using KRaM?
- [RQ2] Does KRaM help improve the fairness of downstream tasks?
- [RQ3] How much original information is retained after erasure using KRaM?

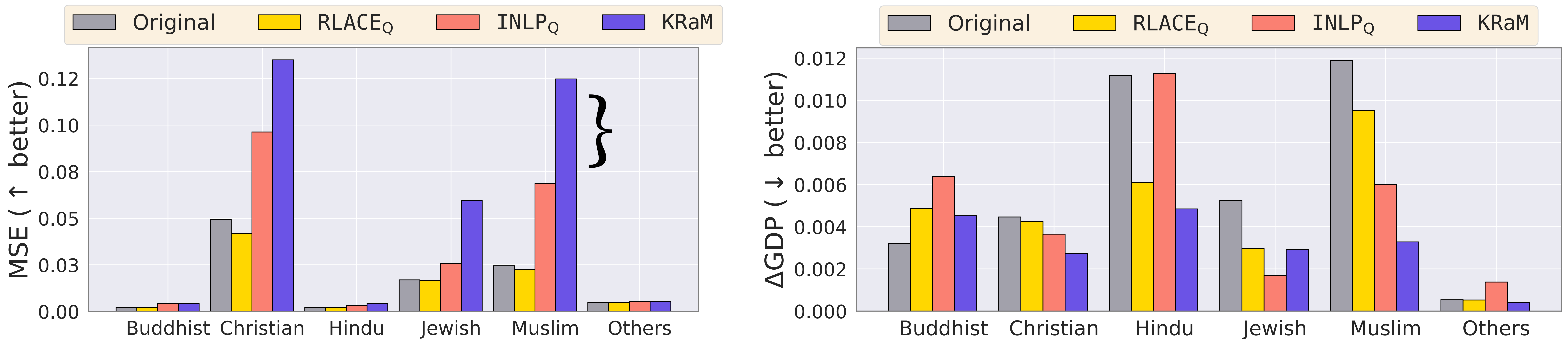
Experimental Setup



Experiments

- Vector-valued Concept Erasure: Jigsaw (religion, gender)
- Continuous Concept Erasure: Synthetic & UCI Crimes (race)
- Categorical Concept Erasure: Glove (gender) & DIAL (race)

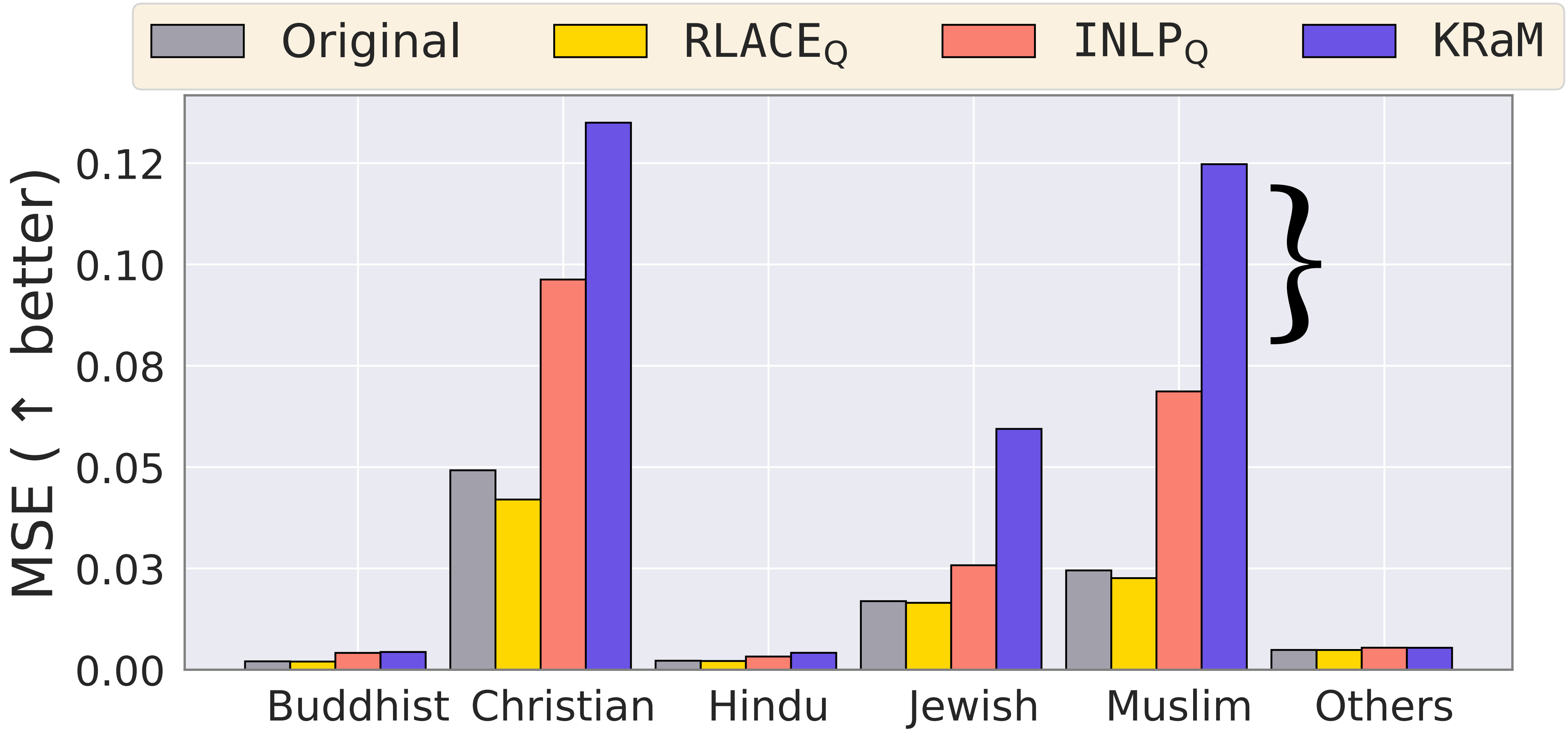
Vector-valued Concept Erasure

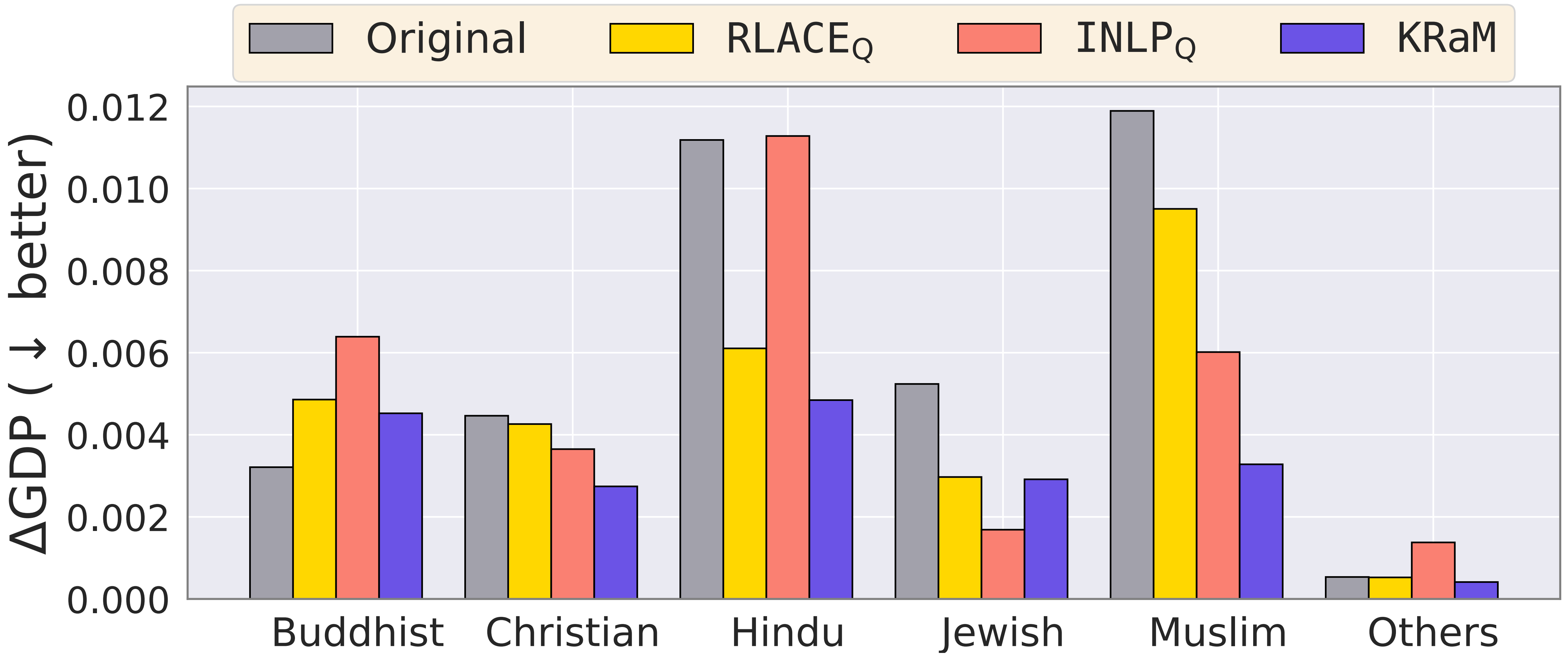


[RQ1] Can the erased concept be predicted after concept erasure using KRaM?

KRaM reduces prediction ability of the erased concept up to 33%

Vector-valued Concept Erasure





[RQ3] How much original information is retained after erasure using KRaM?

Toxicity Classification Accuracy: $93.2\,\% \rightarrow 92.1\,\%$

Continuous Concept Erasure

	Synthetic		
Method	MSE (a) ↑	A_k ↑	Rank ↑
Original	0.006	1.0	100
Random	0.174	0.50	100
INLP_Q [49]	0.084 🏆	0.85 🏆	100
RLACE_Q [50]	0.021	0.87 🏆	100
FaRM_Q [18]	0.068	0.74	100
KRaM	0.109 🏆	0.67	100
$\text{KRaM}_\text{linear}$	0.083 🏆	0.75 🏆	100

[RQ1] KRaM performs the best in terms of removing concept information

Continuous Concept Erasure

	Synthetic		
Method	MSE (a) ↑	A_k ↑	Rank ↑
Original	0.006	1.0	100
Random	0.174	0.50	100
INLP_Q [49]	0.084 🏆	0.85 🏆	100
RLACE_Q [50]	0.021	0.87 🏆	100
FaRM_Q [18]	0.068	0.74	100
KRaM	0.109 🏆	0.67	100
$\text{KRaM}_\text{linear}$	0.083 🏆	0.75 🏆	100

[RQ3] However, KRaM is not able to retain original information compared to linear erasure methods.

Continuous Concept Erasure

UCI Crimes			
MSE (y) $\downarrow$	MSE (a) $\uparrow$	ΔGDP $\downarrow$	$A_k \uparrow$
0.046	0.030	0.058	1.0
0.211	0.251	0.006	0.50
0.055 🏆	0.056	0.0 🏆	0.90 🏆
0.038 🏆	0.022	0.051	0.81
0.050 🏆	0.064 🏆	0.013 🏆	0.62 🏆
0.069	0.104 🏆	0.001 🏆	0.59
0.067	0.082 🏆	0.022	0.69 🏆

[RQ2] KRaM is able to improve the fairness of end tasks by a significant margin.

Categorical Concept Erasure

	DIAL		
Method	Acc. (y) ↑	Acc. (a) ↓	DP ↓
Original	75.5	87.7	0.26
Random	50.8	50.5	0.01
INLP [49]	75.1 🏆	69.5	0.16
RLACE [50]	75.5 🏆	82.1	0.18
KCE [51]	75.0	80.1	0.12 🏆
FaRM [18]	74.8	54.2 🏆	0.09 🏆
KRaM	72.4	54.0 🏆	0.08 🏆
$\text{KRaM}_{\text{linear}}$	75.4 🏆	67.5 🏆	0.18

[RQ1] & [RQ3] KRaM is able to retain task information (when the task is not very correlated with the concept) while erasing requested concepts

Categorical Concept Erasure

Glove		
Acc. (a) ↓	A_k ↑	Rank ↑
100.0	1.0	300
50.2	0.50	300
86.3	0.85 🏆	210
95.5	0.93 🏆	300 🏆
63.5 🏆	0.62	100
53.9 🏆	0.65	247 🏆
52.6 🏆	0.65	246 🏆
67.0	0.73 🏆	130

[RQ1] & [RQ3] KRaM is able to perfectly remove gender information but it is accompanied with a loss of information from the original representation space

Take Aways

- [RQ1] KRaM can robustly erase concepts outperforming other methods.
- [RQ2] KRaM improves the fairness of downstream tasks significantly.
- [RQ3] Concept erasure using KRaM can often lead to significant information loss.

Summary

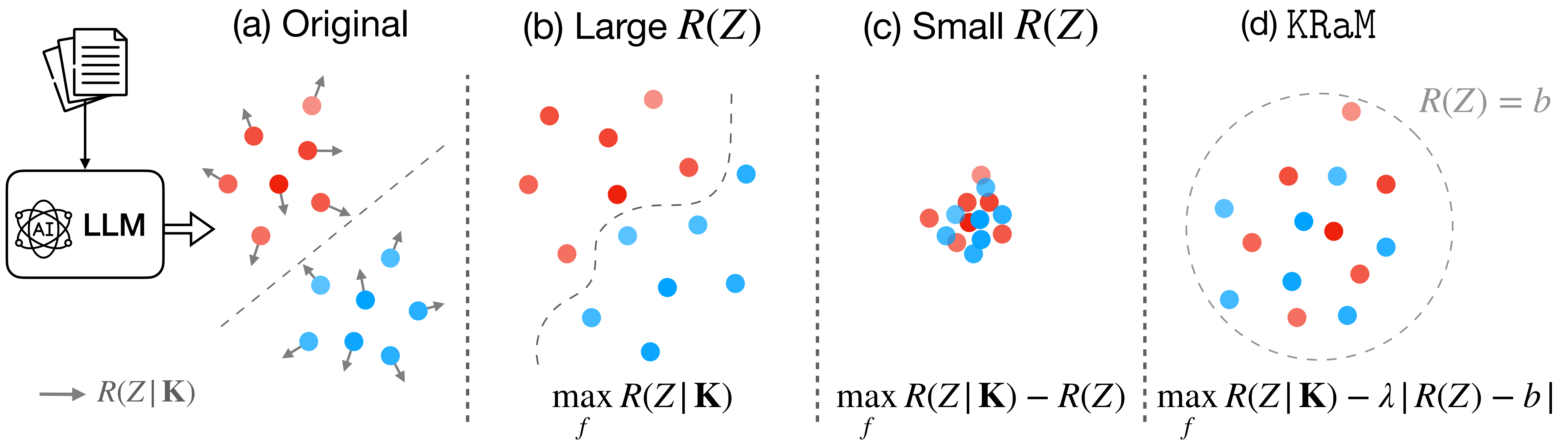
- We propose KRaM, a robust method for concept erasure

brcsomnath/KRaM

somnath@cs.unc.edu

Summary

- We propose KRaM, a robust method for concept erasure

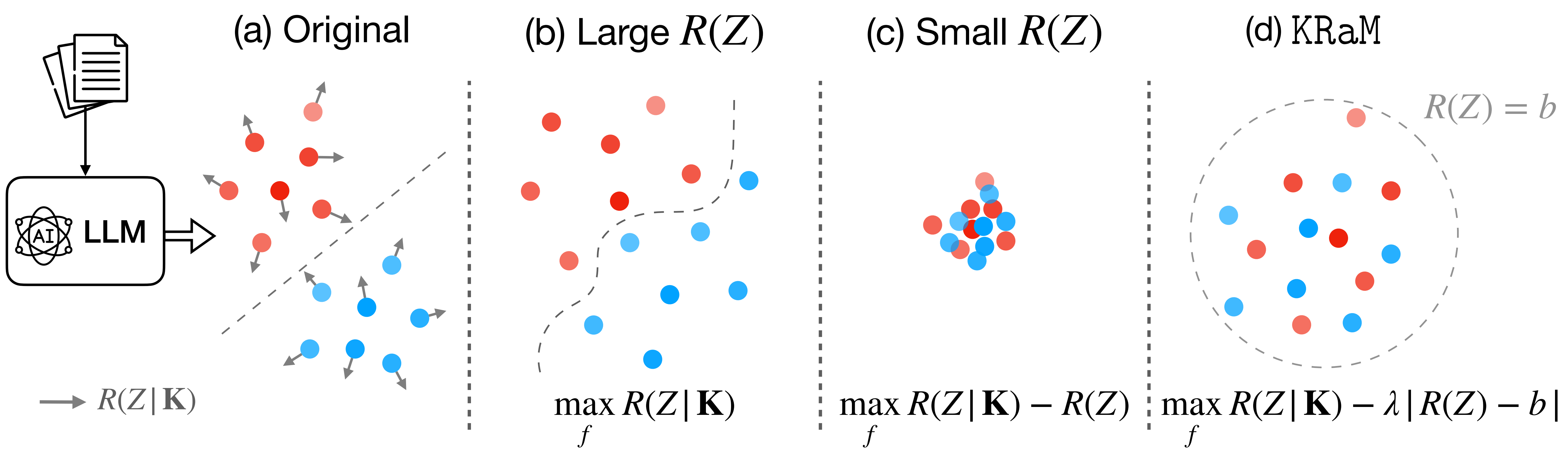


brcsomnath/KRaM somnath@cs.unc.edu

Summary

- We propose KRaM, a robust method for concept erasure

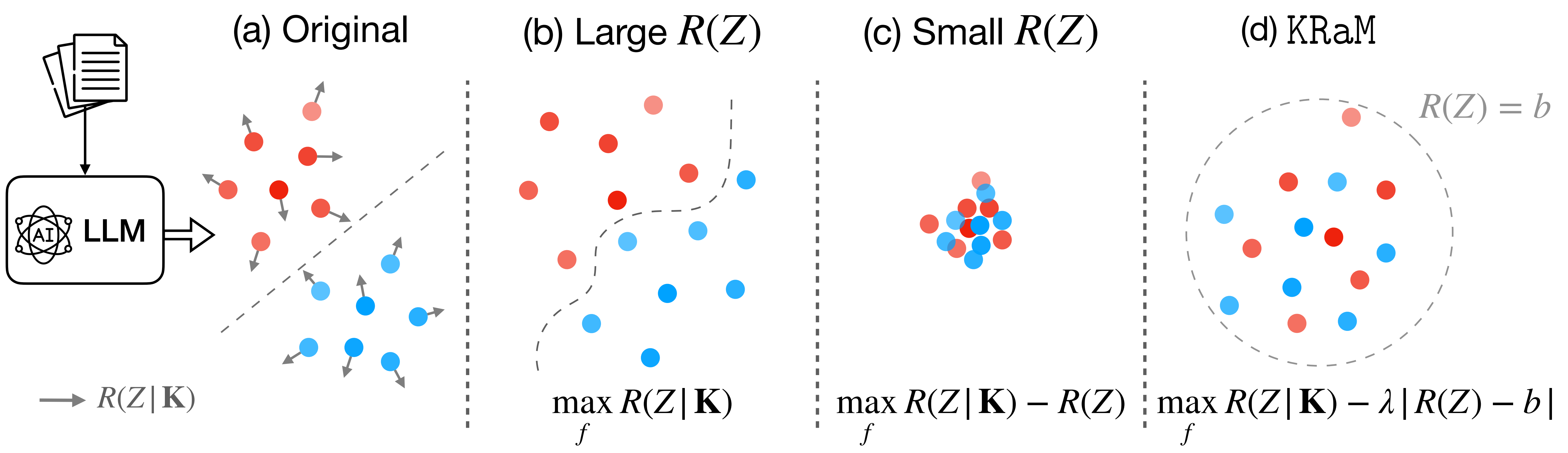
- The kernelized rate distortion function can accommodate different concepts forms: categorical, continuous, and vectors.

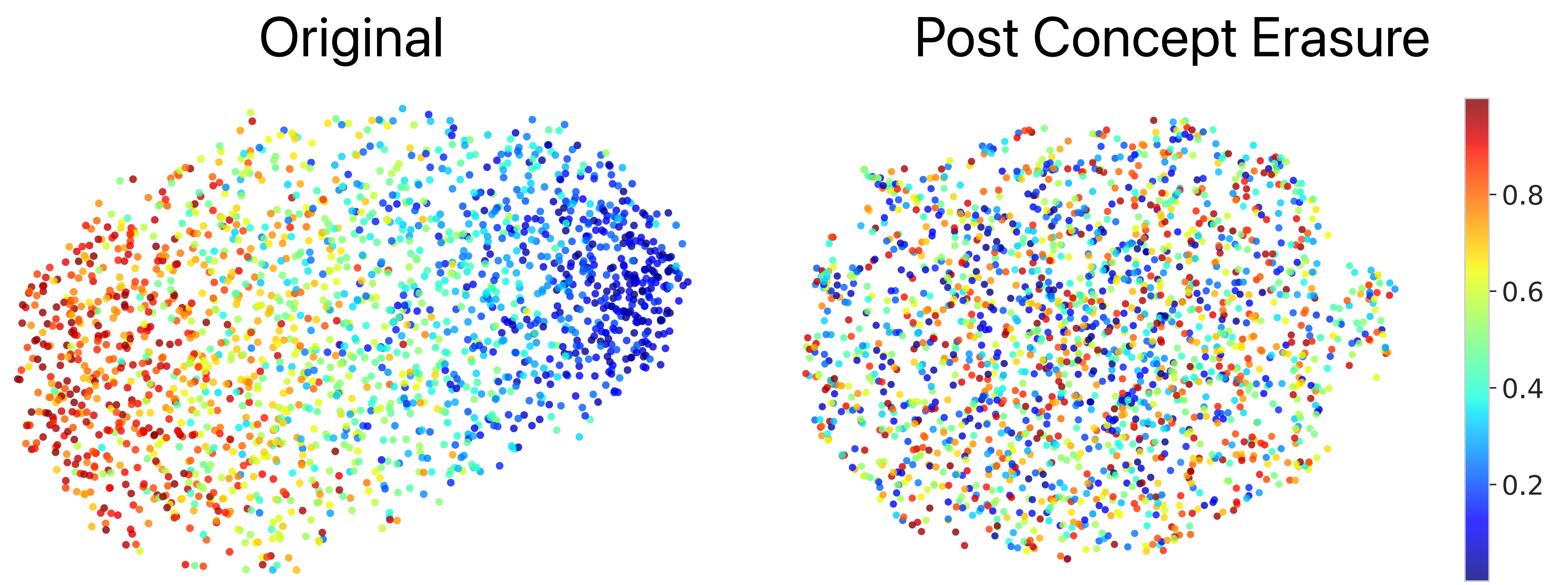


brcsomnath/KRaM   somnath@cs.unc.edu

Summary

- We propose KRaM, a robust method for concept erasure
- The kernelized rate distortion function can accommodate different concepts forms: categorical, continuous, and vectors.



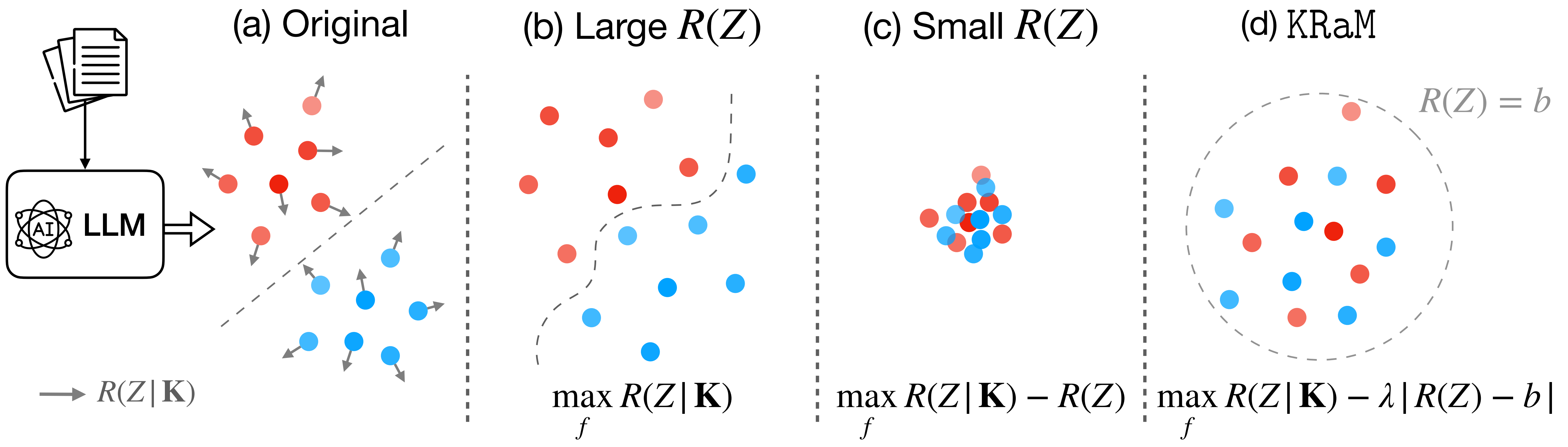


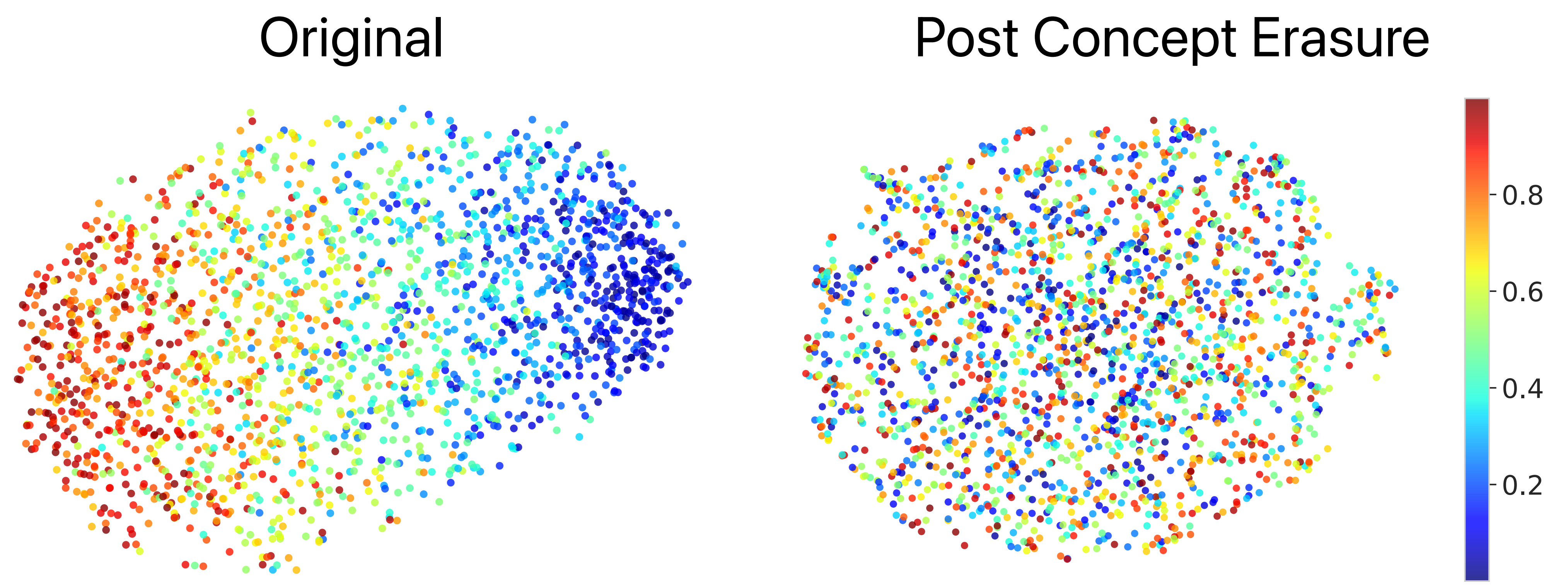
brcsomnath/KRaM

somnath@cs.unc.edu

Summary

- We propose KRaM, a robust method for concept erasure
- The kernelized rate distortion function can accommodate different concepts forms: categorical, continuous, and vectors.
- We introduce a heuristic-based metric to compute information retained after erasure



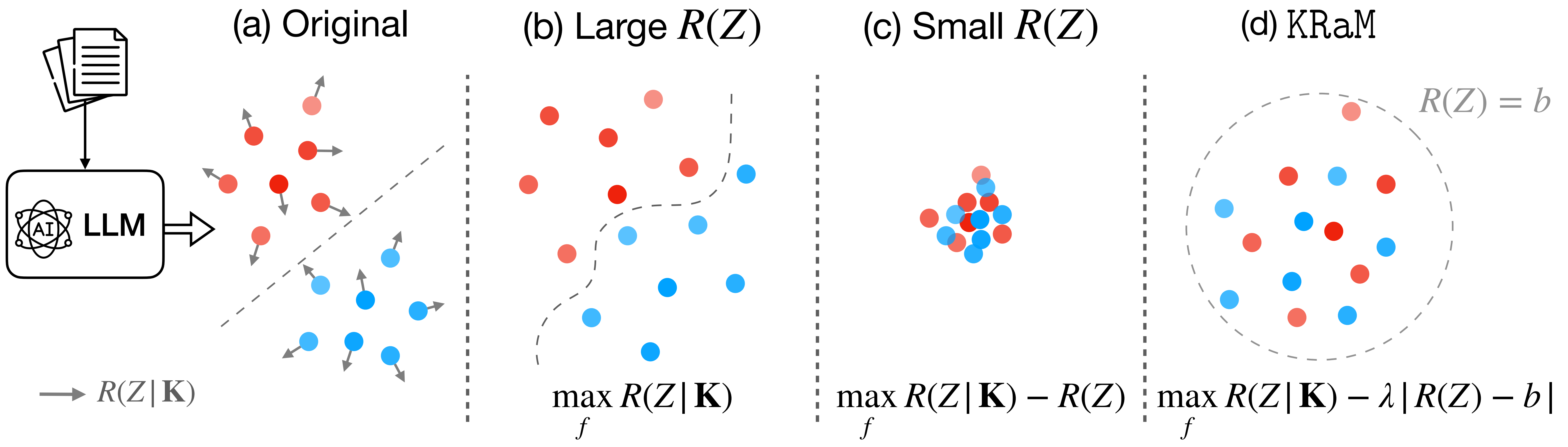


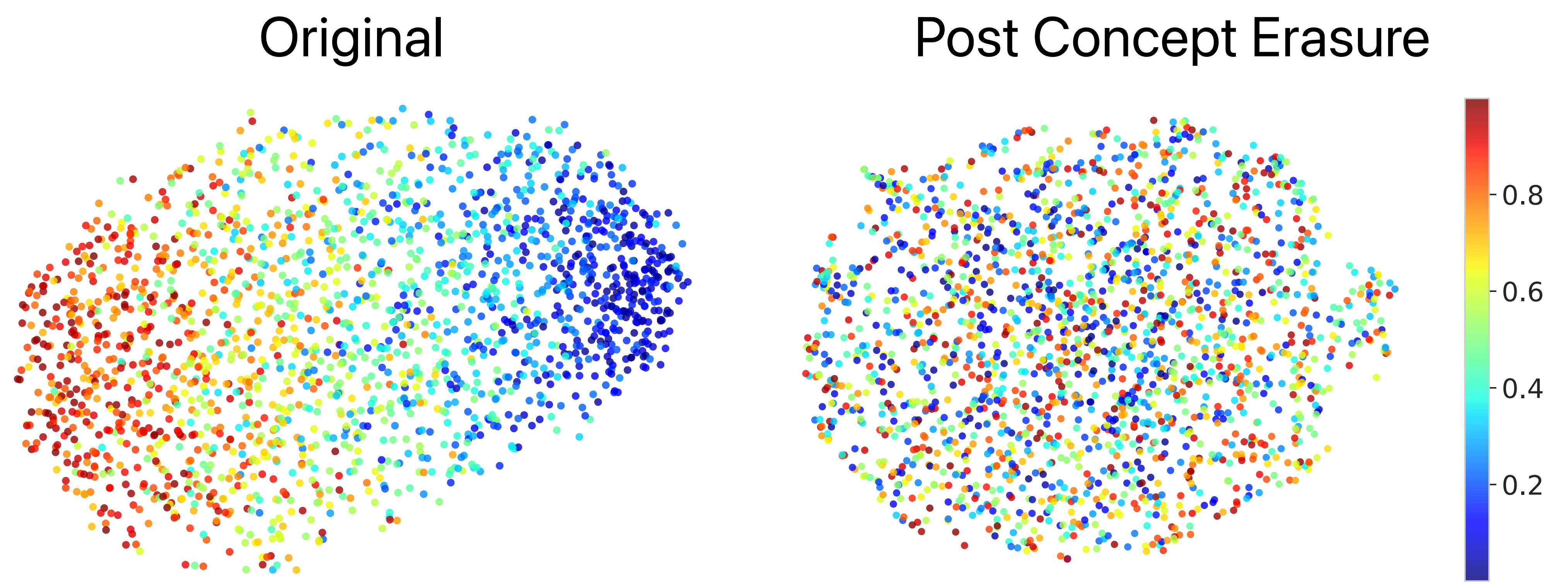
brcsomnath/KRaM somnath@cs.unc.edu

Summary

- We propose KRaM, a robust method for concept erasure
- The kernelized rate distortion function can accommodate different concepts forms: categorical, continuous, and vectors.
- We introduce a heuristic-based metric to compute information retained after erasure

$$A_k(f) = \frac{1}{k}\mathbb{E}_x\left[\mathrm{knn}(x) \cap \mathrm{knn}(f(x))\right]$$





brcsomnath/KRaM

somnath@cs.unc.edu

Summary

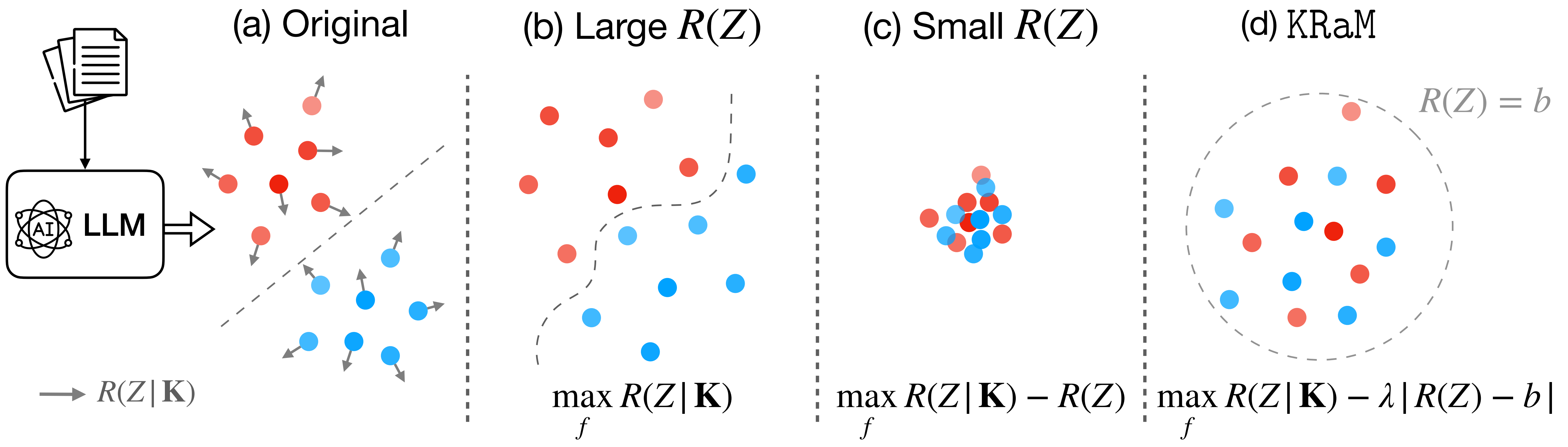
- We propose KRaM, a robust method for concept erasure

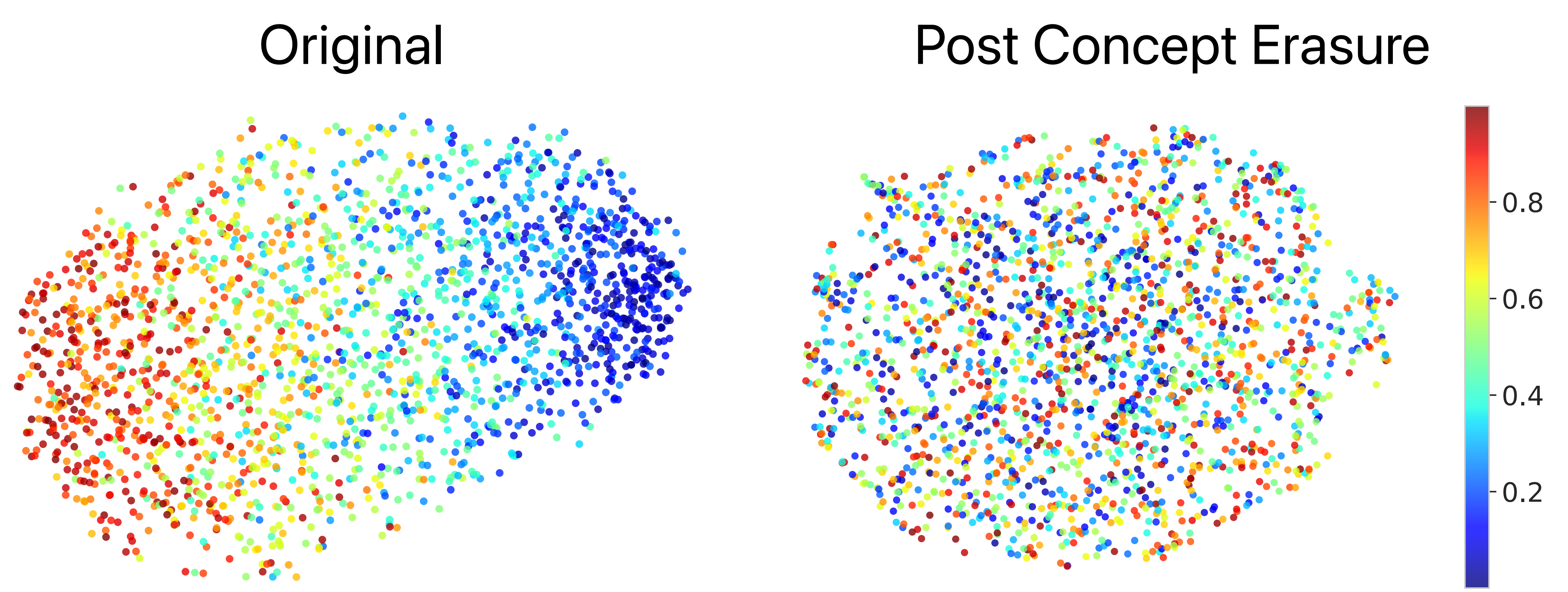
- The kernelized rate distortion function can accommodate different concepts forms: categorical, continuous, and vectors.

- We introduce a heuristic-based metric to compute information retained after erasure

$$A_k(f) = \frac{1}{k}\mathbb{E}_x\left[\text{knn}(x) \cap \text{knn}(f(x))\right]$$

- Future works can explore effective ways to erase concepts while retaining as much information as possible





brcsomnath/KRaM

somnath@cs.unc.edu