Hot PATE:
Private Aggregation of Distributions for Diverse Tasks

Edith Cohen

Google Research

TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Joint work with:

Ben Cohen-Wang **Xin Lyu** **Jelani Nelson** **Tamás Sarlós** **Uri Stemmer**

Google Research

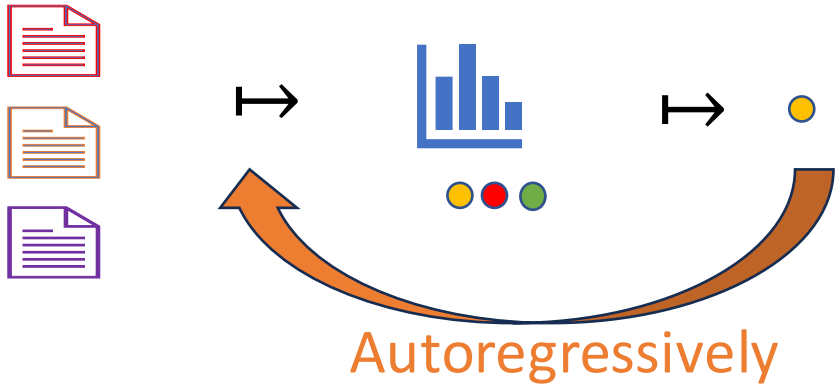
TEL AVIV UNIVERSITY אוניברסיטת תל אביב

Google Research

Background: Generative Tasks

Input: data records/examples 

(Example) Task: Generate similar synthetic records

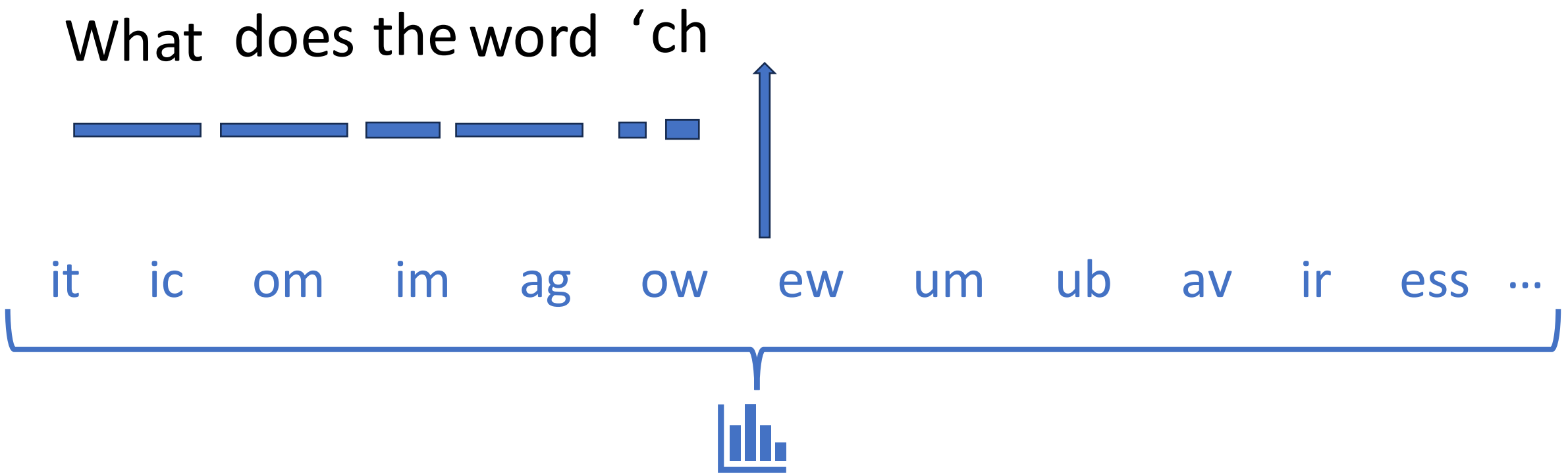
LLM + Finetuning / ICL 

Goals: Correct Diverse (coverage, generalize)

Diversity

Dataset: Questions

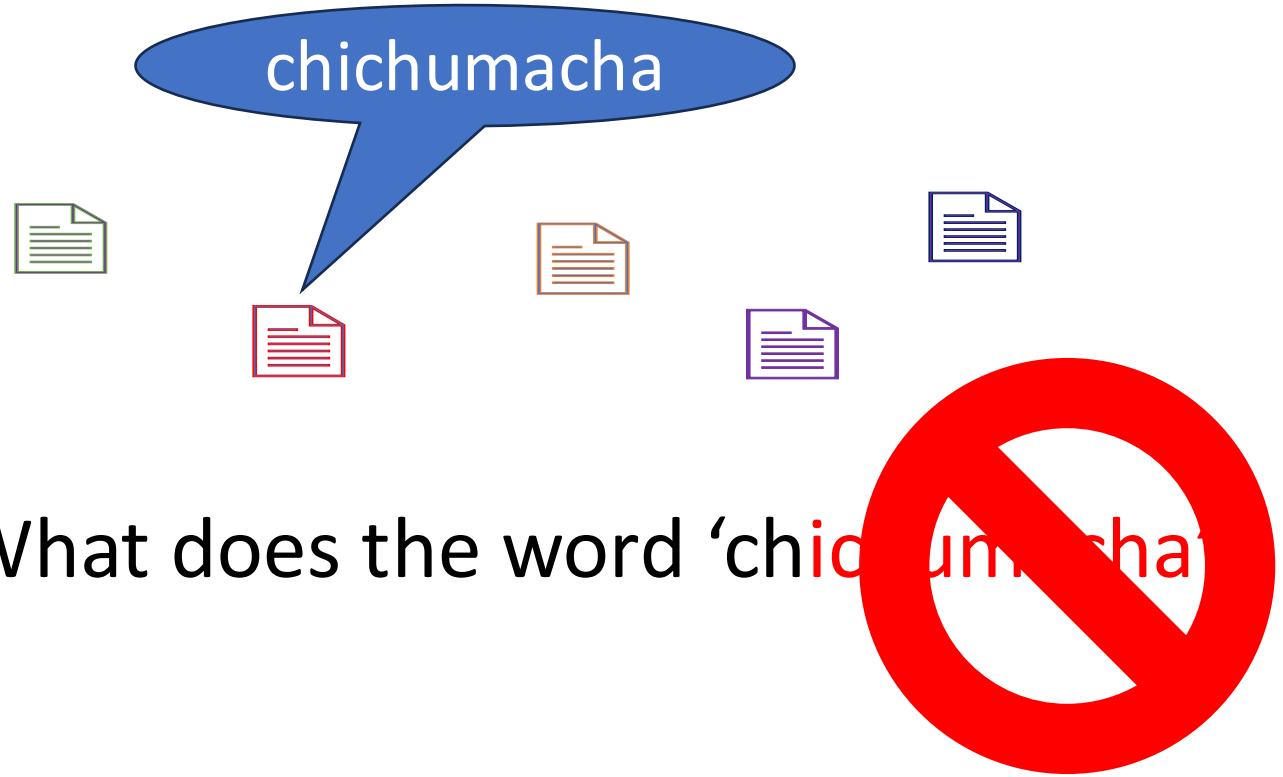
Task: Generate similar ones



Privacy Protection

 … are sensitive !!

Output must be attributable to multiple records or to the public model.

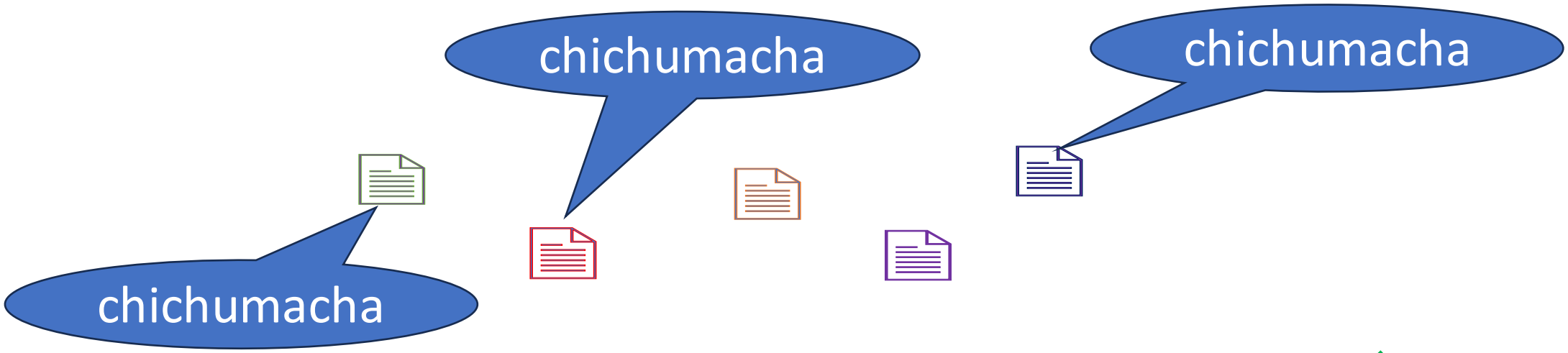


What does the word ‘chi[illegible]’

Privacy Protection

… are sensitive !!

Output must be attributable to multiple records or to the public model.



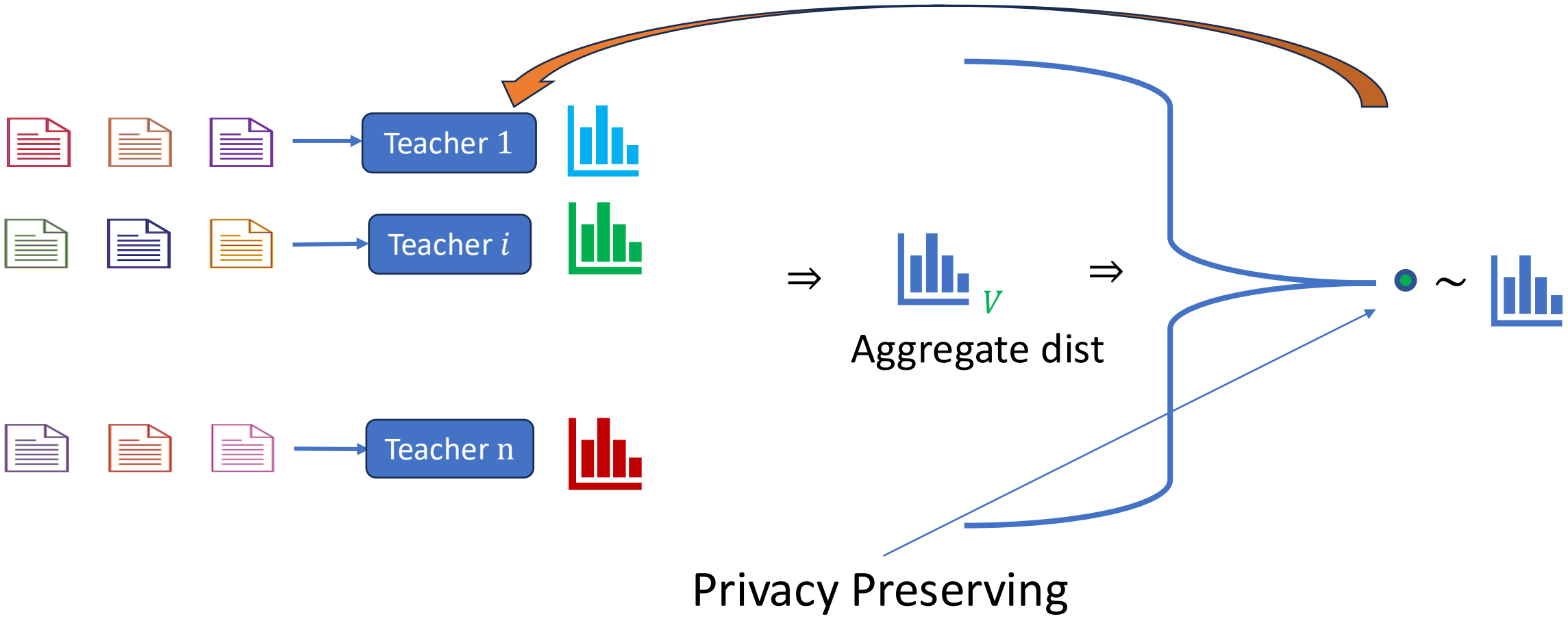
What does the word ‘chichumacha’

Ensemble decoding (PATE)

[Bassili et al, Papernot et al] [Nissim et al]

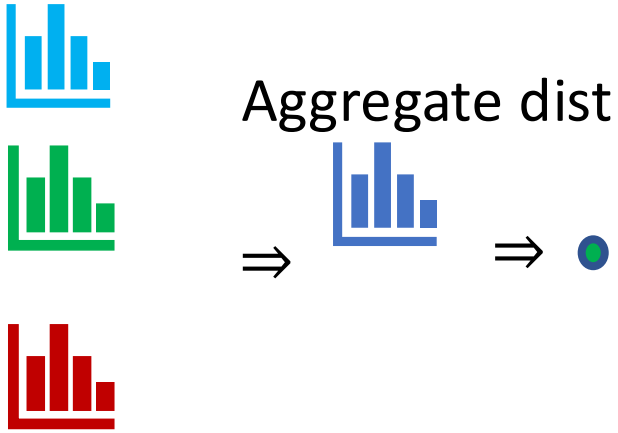
[Tian et al ‘22] [Duan et al ‘23] [Wu et al ‘23]

n teachers get disjoint random subsets of records



Utility of Ensemble Samplers

Basic Utility: Return (*any*) relevant token

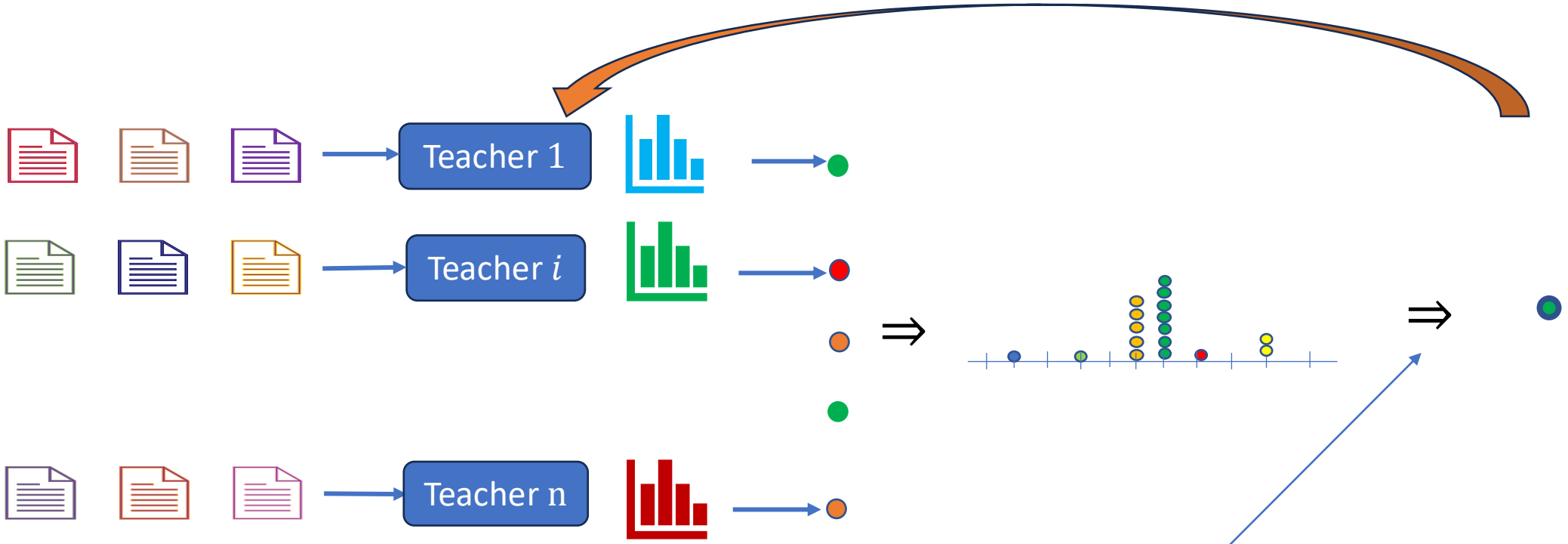


Preserving Diversity Robustness parameter τ

- **Relevance** $\mathrm{AggProb}(v) = \Omega(\mathrm{AverageProb}(v))$
- **Transfer** A token v with probability $p_v^i \geq q > 0$ across $\geq \tau$ teachers is sampled with probability $\Omega(q \cdot \frac{\tau}{n})$

Ensemble decoding (PATE) via voting

n teachers get disjoint random subsets of records



- Each teacher casts a “vote” ● ~
- Privacy Preserving

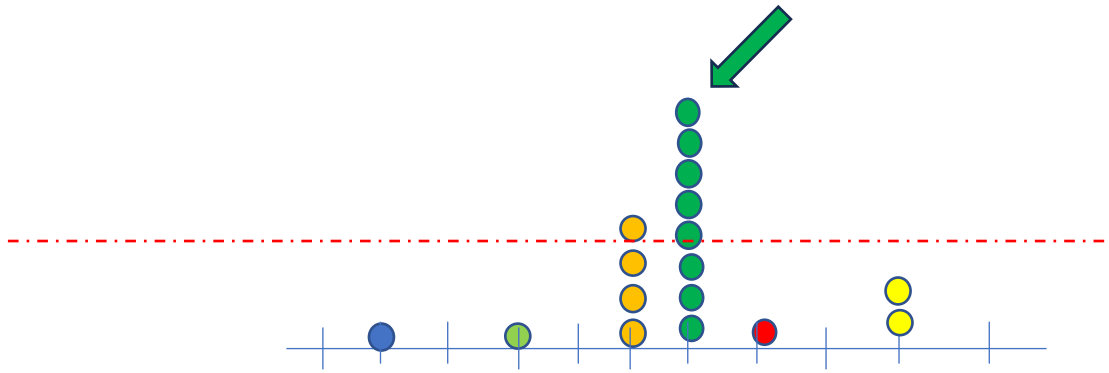
Voting-based Ensemble Decoding

Each teacher contributes one token

Privacy: Return a token with teachers count $\geq T$

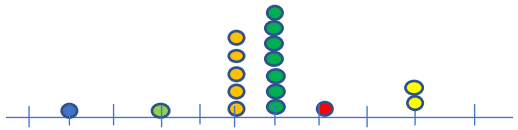
Privacy $\propto T$

$$T \approx \frac{1}{\varepsilon}$$



[Tian et al '22] [Duan et al '23] [Wu et al '23]

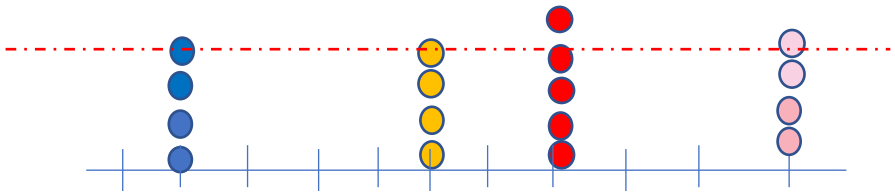
Cold PATE: Independent sampling



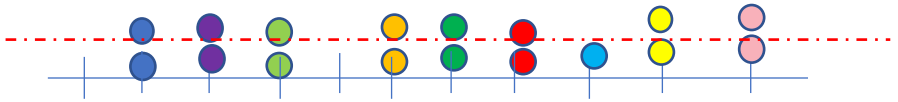
$\approx$ Average of teacher's distributions

it ic om im

$T \approx \frac{1}{\varepsilon}$



it ic om im ag ow ew um ub



Higher diversity $\Rightarrow$ *lower threshold* $\Rightarrow$ *less privacy*

Hot PATE decoder: Coordinated sampling

[Brewer et al 1972,….]

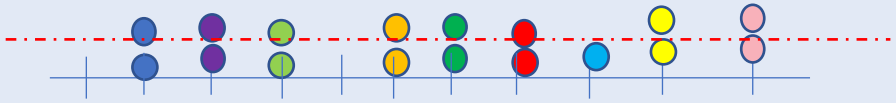
Sample shared randomness $\rho \sim R$

Each teacher samples from its marginal distribution, using ρ

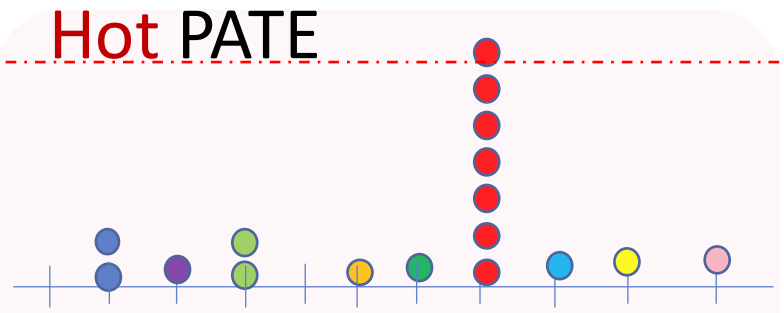
Gain in shape of Histograms

it ic om im ag ow ew um ub

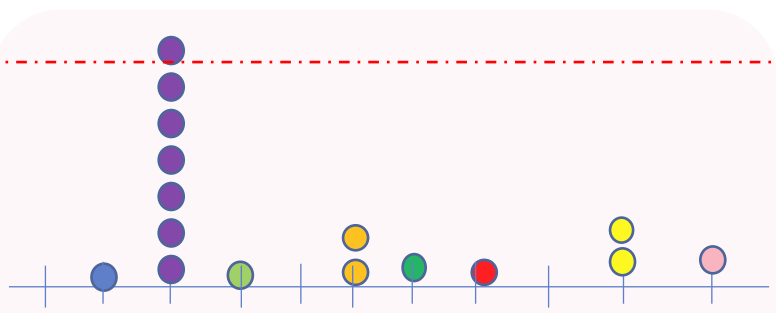
Cold PATE



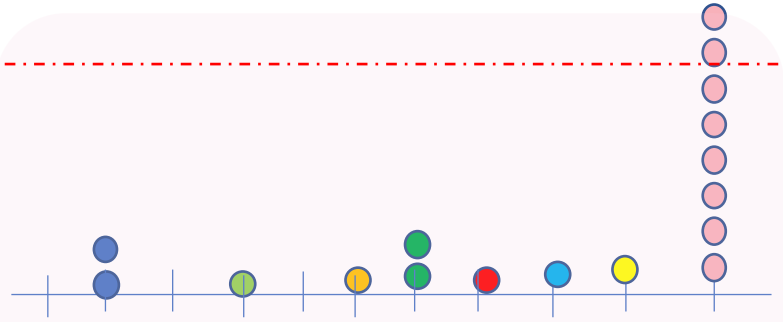
Hot PATE



$\rho_1 \sim R$



$\rho_2 \sim R$



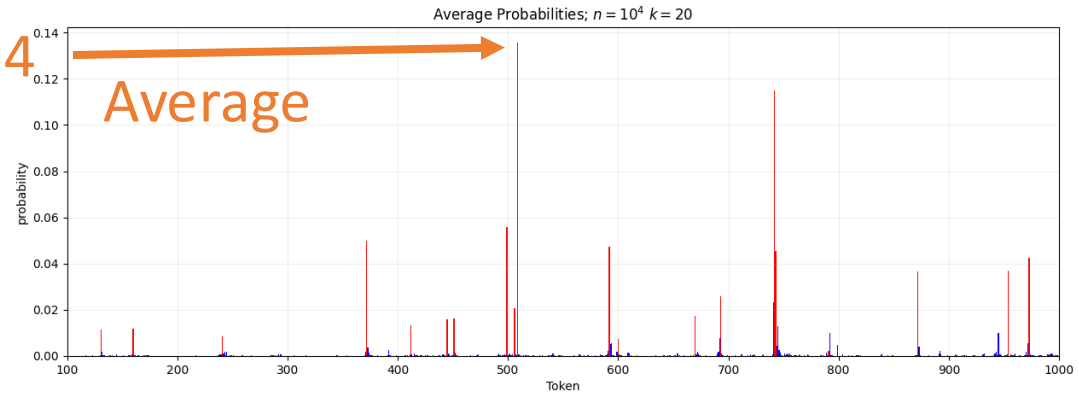
$\rho_3 \sim R$

Higher Basic Utility for Privacy

+ Preserves diversity with $\tau \approx 1/\varepsilon$

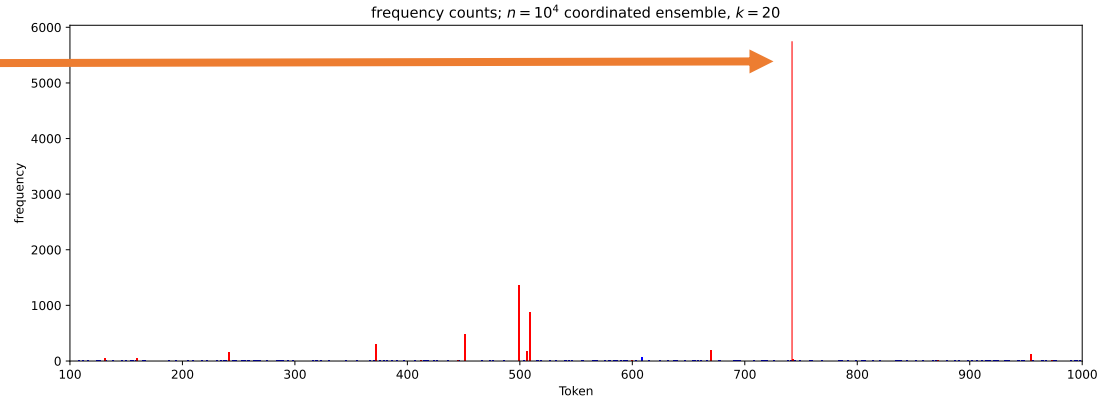
Demo: Shape of Histograms

$n = 10^4$ teachers



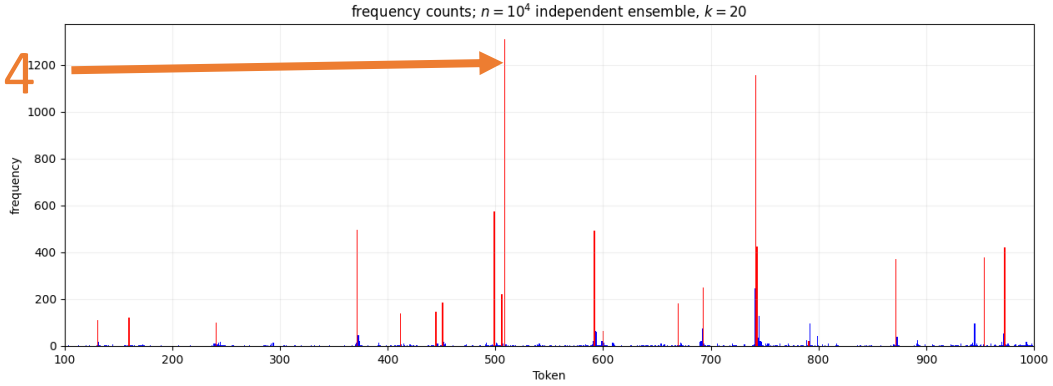
0.14 Average

Hot PATE votes

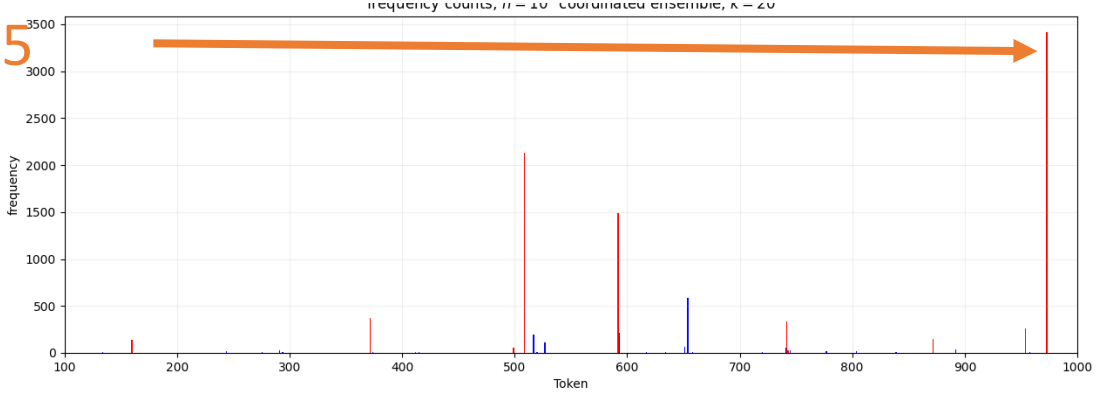


0.6

Cold PATE votes ≈ Average

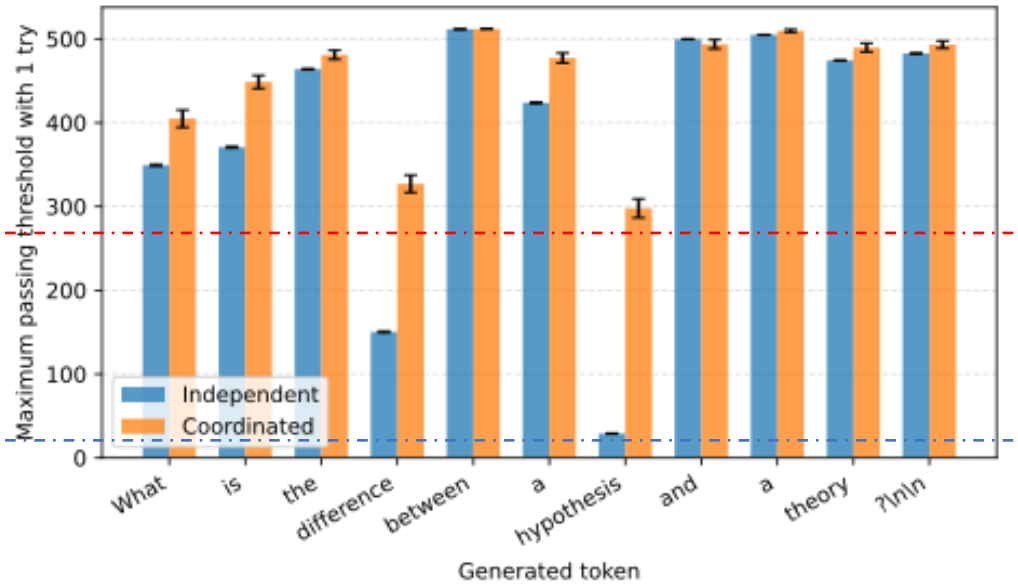


0.14



0.35

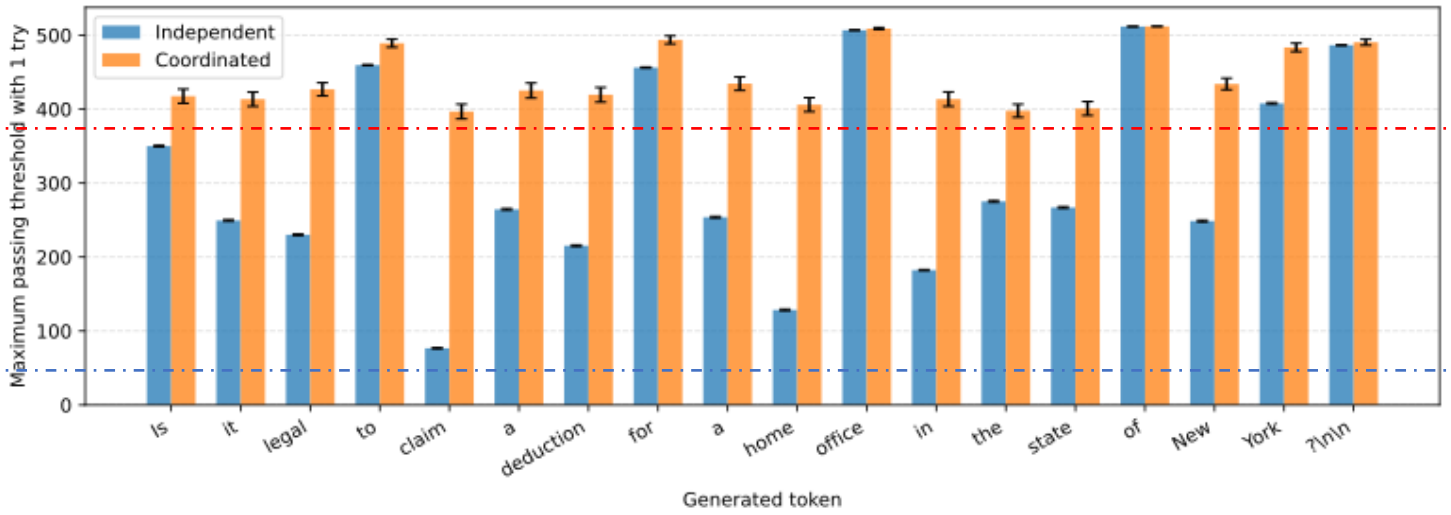
Experiments: Basic Utility for Privacy



Max Count
512 teachers

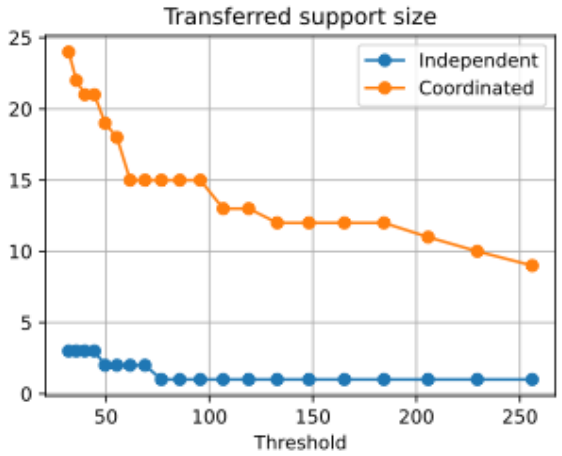
~~Hot~~ PATE: high T ⇒ high privacy

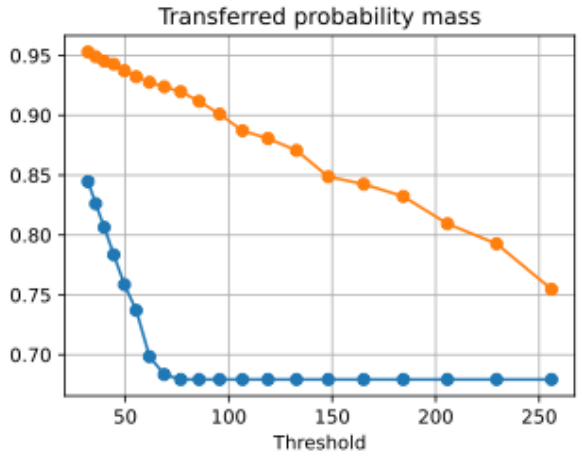
Cold PATE: low T ⇒ low privacy



Experiments: Diversity Transfer for Privacy

Transferred diversity when generating an instruction from "Ø"

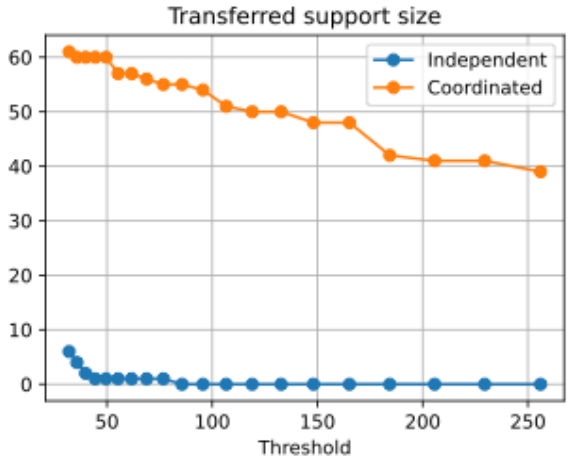


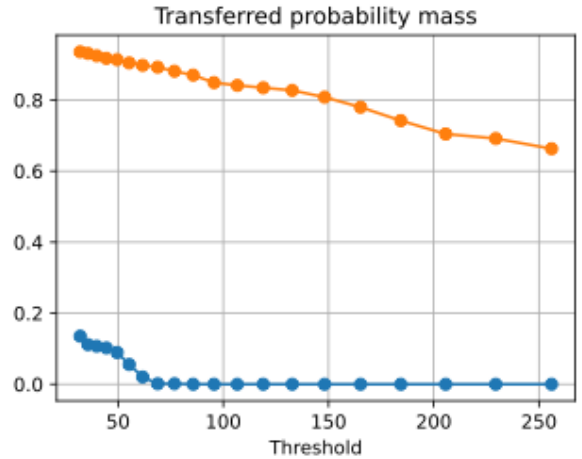


Supported tokens ($T = 256$)

Independent	Coordinated
1. 'What'	1. 'What' 2. 'Is' 3. 'How' 4. 'can' 5. 'Can' 6. 'Where' 7. 'Why' 8. 'Who' 9. 'Which'

Transferred diversity when generating an instruction from "What does the word 'ch"





Supported tokens ($T = 256$)

Independent	Coordinated
	1. 'it' 2. 'ic' 3. 'om' 4. 'im' 5. 'ag' 6. 'ow' 7. 'ew' 8. 'um' 9. 'ub' 10. 'av' 11. 'ir' 12. 'ess' ... 27 more

Summary of Contributions: Hot PATE

- Ensemble decoders with high utility for privacy even with high diversity
  + Voting based!
- Introduced "diversity preservation" as a formal utility measure of ensemble decoders
  Hot PATE preserves diversity
- Experiments

See our paper for technical details and follow up research questions!

Thank you!!