

# Upsample or Upweight?

## Balanced Training on Heavily Imbalanced Datasets

Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray, Daniel Khashabi

NAACL 2025



# In our language model's pre-training data...

There exists very common knowledge:



You can get calcium from dairy products like milk, yogurt and cheese, canned fish with soft bones (sardines, anchovies and salmon; bones must be consumed to get the benefit of calcium), dark-green leafy vegetables (such as kale, mustard greens and turnip greens) and even tofu (if it's processed with calcium sulfate).

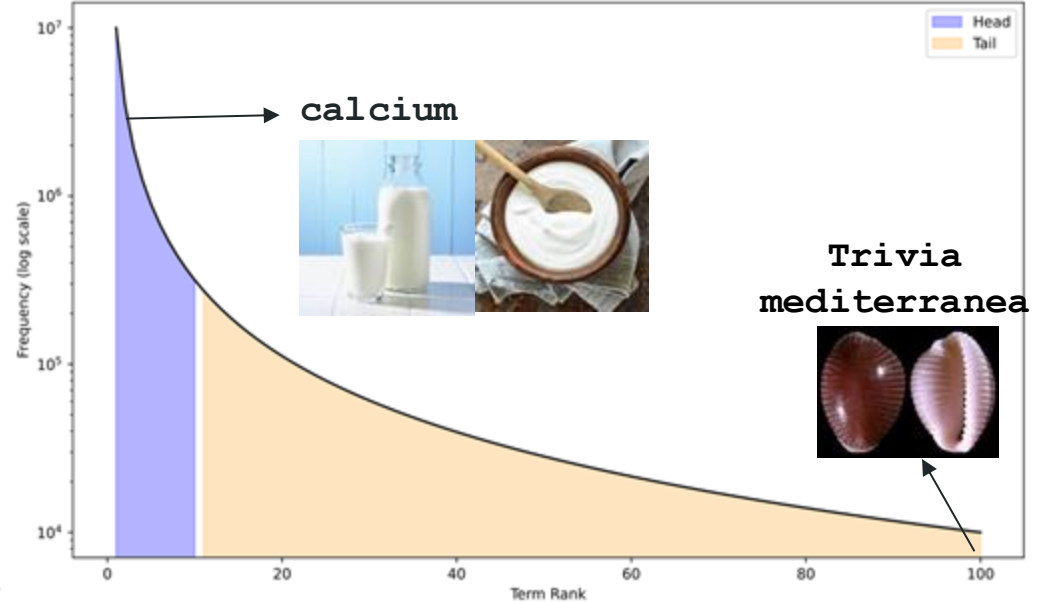
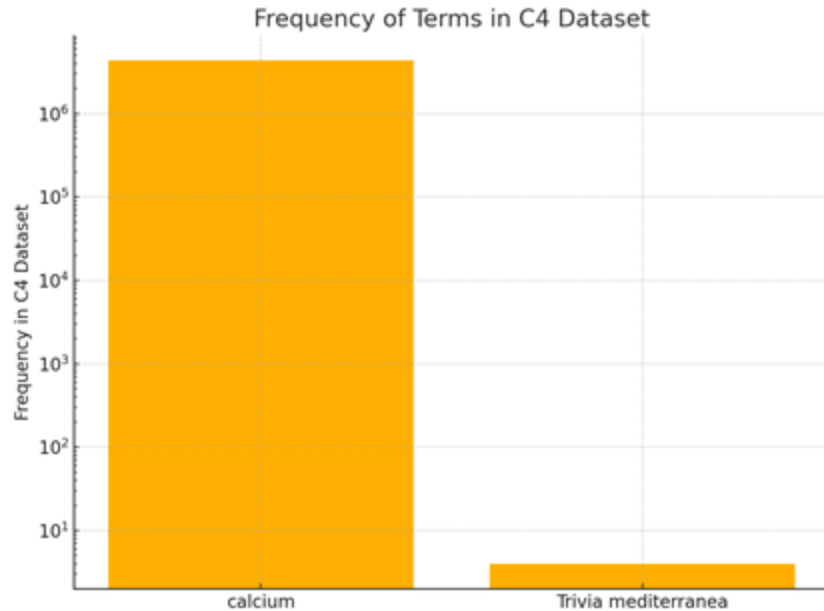
But there also exists very niche topics:

*Trivia mediterranea* is a species of small sea snail, a marine gastropod mollusc in the family Triviidae, the false cowries or trivias.



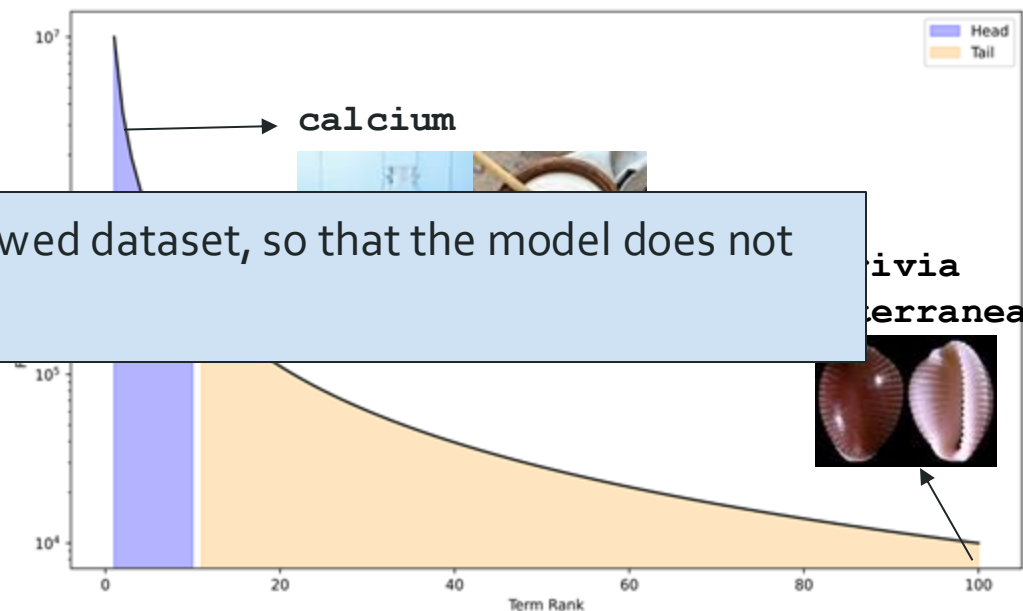
# The “long-tailedness” of knowledge

In the entire C4 dataset, trivia mediterranea only appears 4 times.



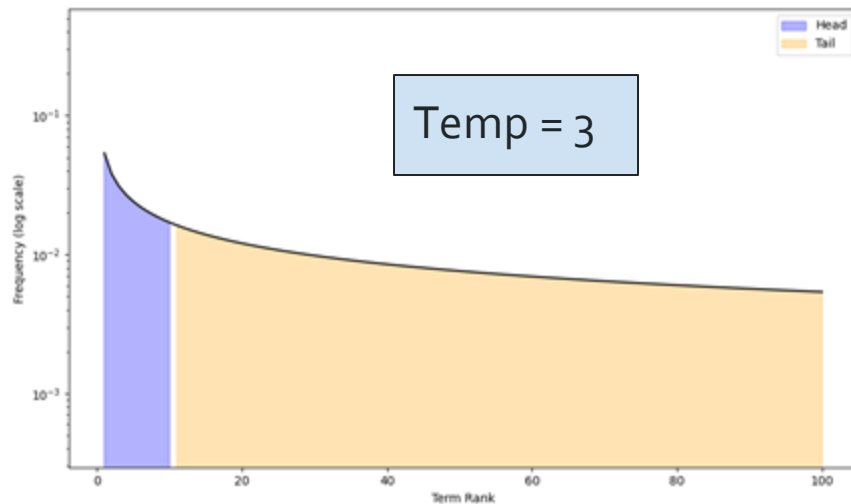
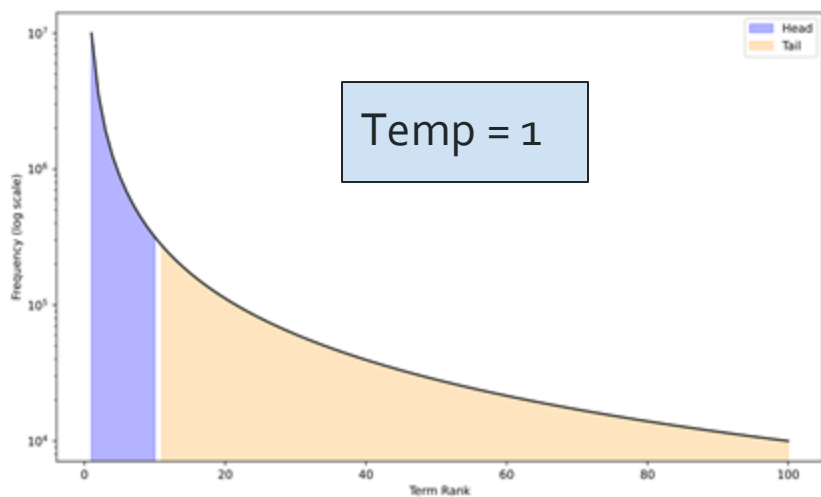
# How to train on the long-tailed data

How to train a model on this skewed dataset, so that the model does not overlook the long-tail?



# To solve the long-tailed problem, we can...

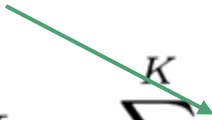
**Solution 1: (Temperature Sampling)** We heavily oversample infrequent domains —  
— Effectively duplicating the data multiple times.



To solve the long-tailed problem, we can also...

**Solution 2: (Scalarization)**

We assign a much higher weight to the loss of infrequent domains.


$$L_S = \sum_{k=1}^K w_k \sum_{x \in \mathcal{D}_k} \ell(x)$$

**Solution 1: (Temperature Sampling)** We heavily oversample infrequent domains —  
— Effectively duplicating the data multiple times.

$$L_{TS} = \mathbb{E}_{\substack{k \sim p \\ x \sim \mathcal{D}_k}} [\ell(x)]$$

## Temperature Sampling often assumed to be equivalent to Scalarization

In our work, we follow convention and implement scalarization via proportional sampling, where data from task  $i$  is sampled with probability equal to  $w_i$ . In this case, the expected loss is equal to the loss from scalarization:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}} [\ell(\mathbf{x}; \theta)] = \sum_{i=1}^K \mathbb{P}(\text{task } i) \mathbb{E}_{\mathbf{x} \sim \text{task } i} [\ell(\mathbf{x}; \theta)] = \sum_{i=1}^K w_i \mathcal{L}_i(\theta). \quad (2)$$

[Order Matters in the Presence of Dataset Imbalance for Multilingual Learning](#) (Choi et al., NeurIPS 2024)

frontier of scalarization. Following the NMT literature's convention, we implement scalarization via proportional sampling. Here, the average number of observations in the batch corresponding to task  $i$  is proportional to  $w_i$ . In this setup, the expected training loss is equal to

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}} [\ell(\mathbf{x}; \theta)] = \sum_{i=1}^K \mathbb{P}(\mathbf{x} \in \text{task } i) \mathbb{E}_{\mathbf{x}} [\ell(\mathbf{x}; \theta) | \mathbf{x} \in \text{task } i] = \sum_{i=1}^K w_i \mathcal{L}_i(\theta).$$

[Do Current Multi-Task Optimization Methods Even Help?](#) (Xin et al., NeurIPS 2022)

# DoReMi paper case study



# Equivalency Under (Full) Gradient Descent

Scalarization often assumed to be equivalent to Temperature Sampling

But in fact, they are not!

**Theorem:** Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

$$\text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) \geq \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau)).$$

**Theorem:** larger temperature induces a larger variance gap!

$$\Delta = \text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) - \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau))$$

*monotonically increases when  $\tau \geq 1$ .*

Scalarization often assumed to be equivalent to Temperature Sampling

But in fact, they are not!

**Theorem:** Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

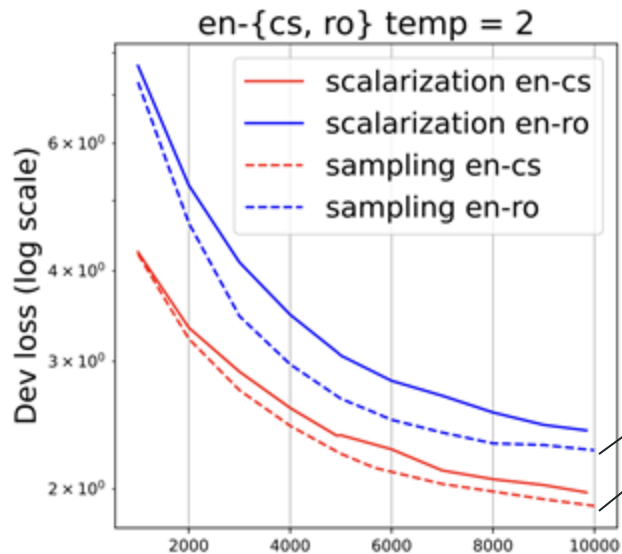
$$\text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) \geq \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau)).$$

**Theorem:** larger temperature induces a larger variance gap!

$$\Delta = \text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) - \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau))$$

*monotonically increases when  $\tau \geq 1$ .*

- It is well-known that variance-reduction accelerates the convergences of SGD (Sutskever et al., 2013; Kingma and Ba, 2015)
- Hypothesis: Temperature Sampling induces less variance, therefore it should converge faster!



Temperature Sampling  
(Dashed) Converges  
faster than Scalarization  
(Solid)!

Scalarization often assumed to be equivalent to Temperature Sampling

But in fact, they are not!

**Theorem:** Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

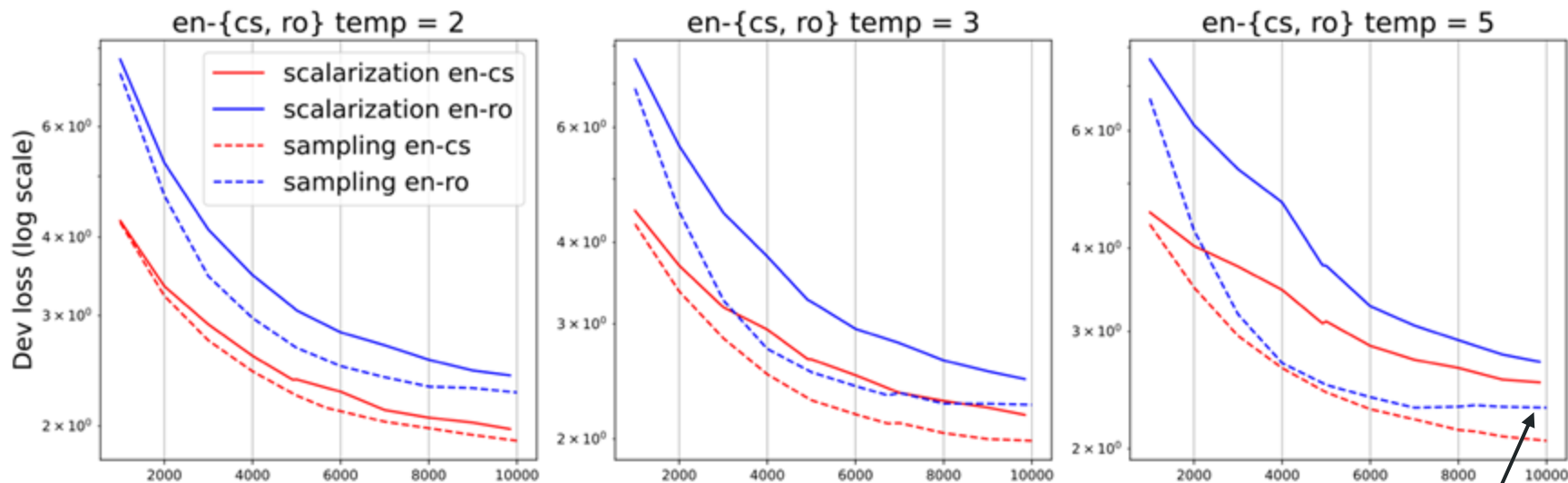
$$\text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) \geq \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau)).$$

**Theorem:** larger temperature induces a larger variance gap!

$$\Delta = \text{Var}(\nabla \mathcal{L}_S(x; \mathbf{w}_\tau)) - \text{Var}(\nabla \mathcal{L}_{TS}(x; \tau))$$

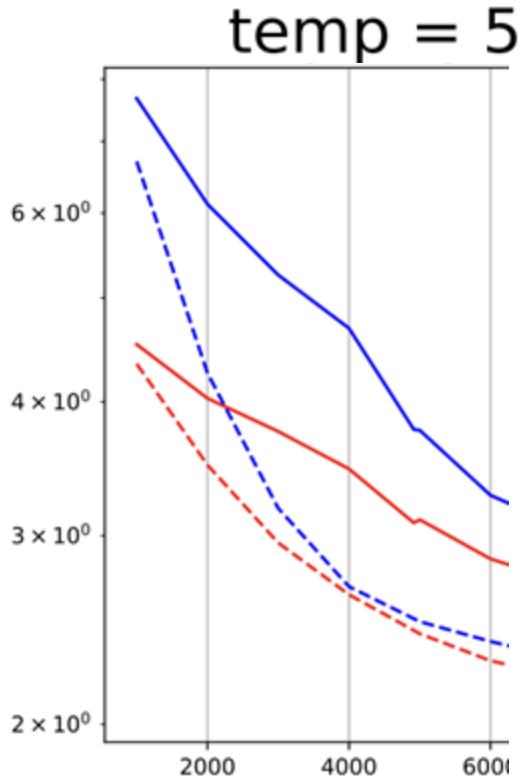
*monotonically increases when  $\tau \geq 1$ .*

- Hypothesis: Temperature Sampling induces less variance, therefore it should converge faster!

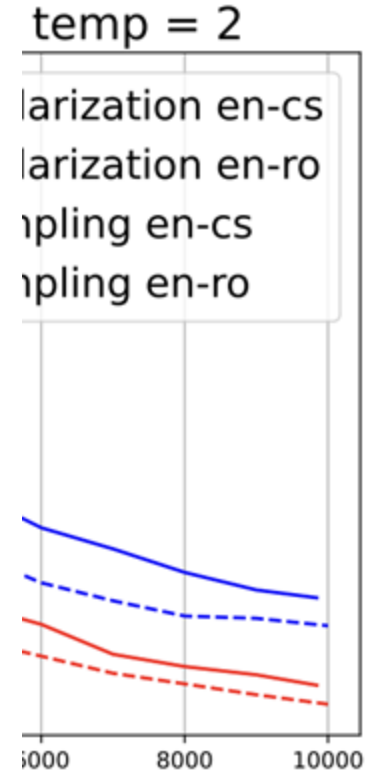


Increasing temperature (2 to 5) makes the convergence even faster, but easy to overfit

Cooldown: Initially use large temp, then use small temperature



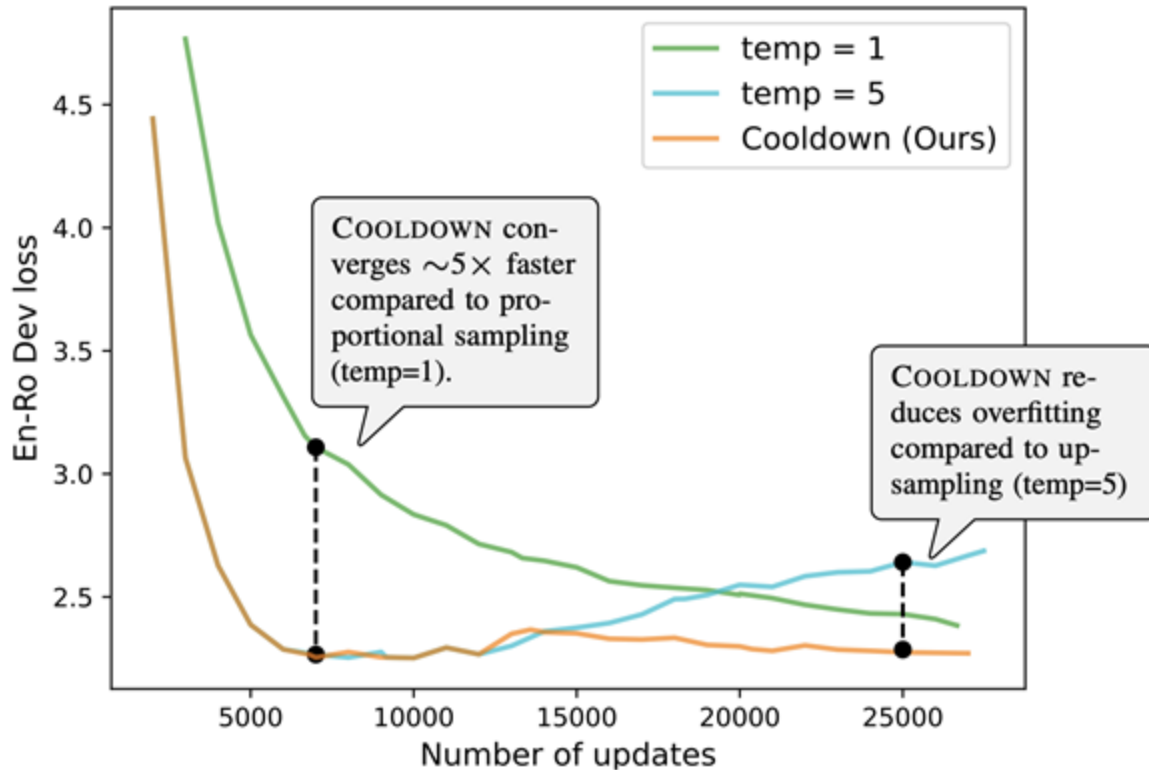
+



# Summary thus far

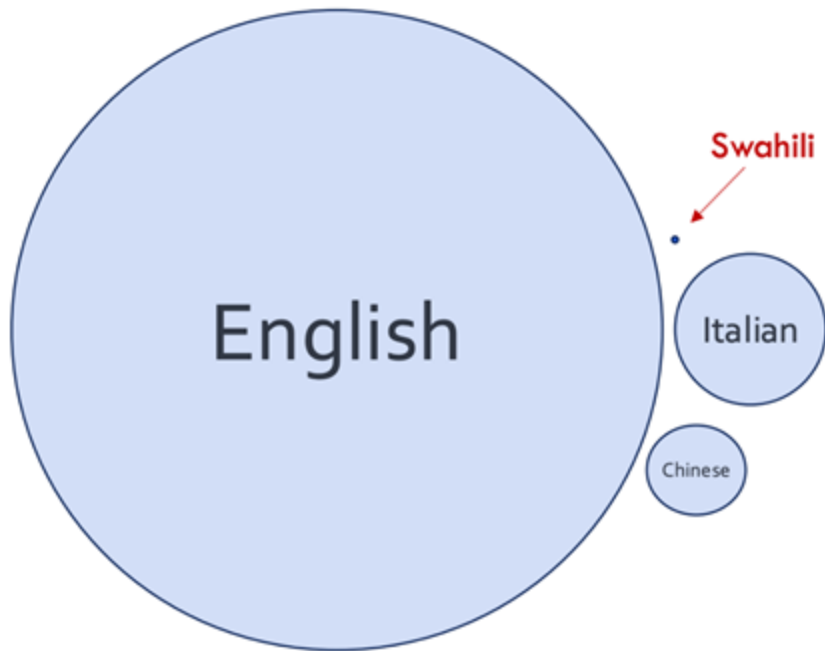


# Cooldown: Initially use large temp, then use small temperature

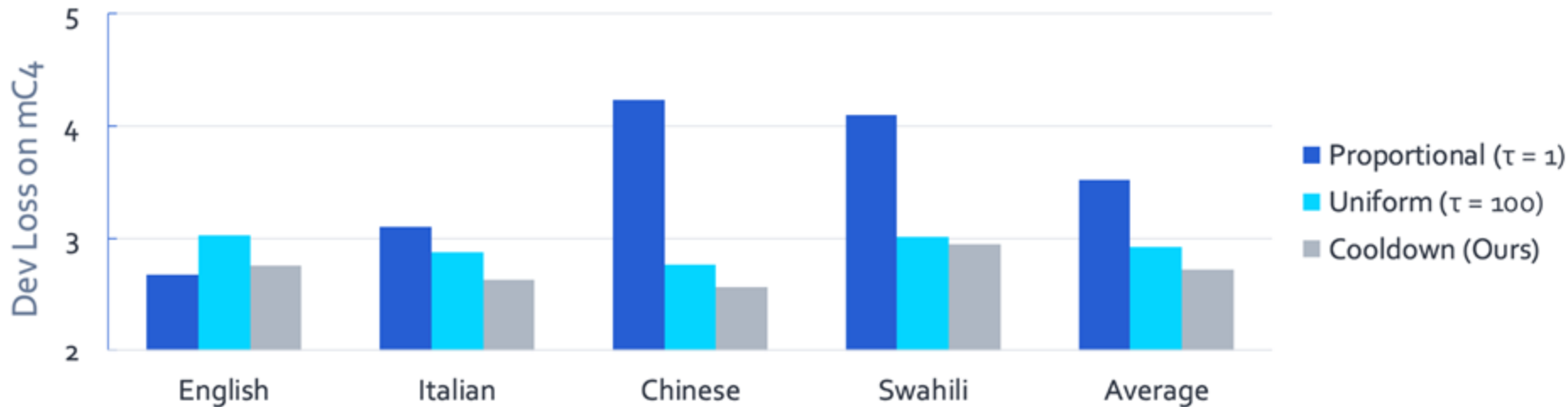


# A Natural Definition of “Domains” — language

Tokens of pretraining  
data by language in mC4  
(Xue+ 2024)

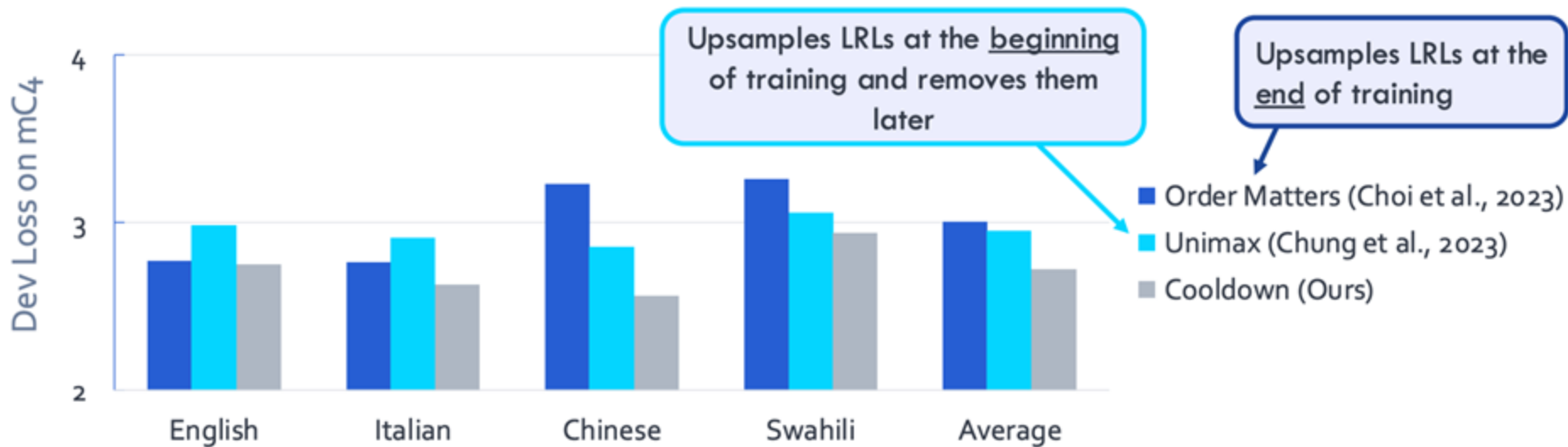


## Cooldown: Dev Loss on mC<sub>4</sub> (lower is better)



Cooldown outperforms fixed temperature sampling!

## Cooldown: Dev Loss on mC<sub>4</sub> (lower is better)



Cooldown outperforms existing work that dynamically adjusts the sampling temperature!

# Summary

-

For more results:  
<https://arxiv.org/pdf/2410.04579>