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:

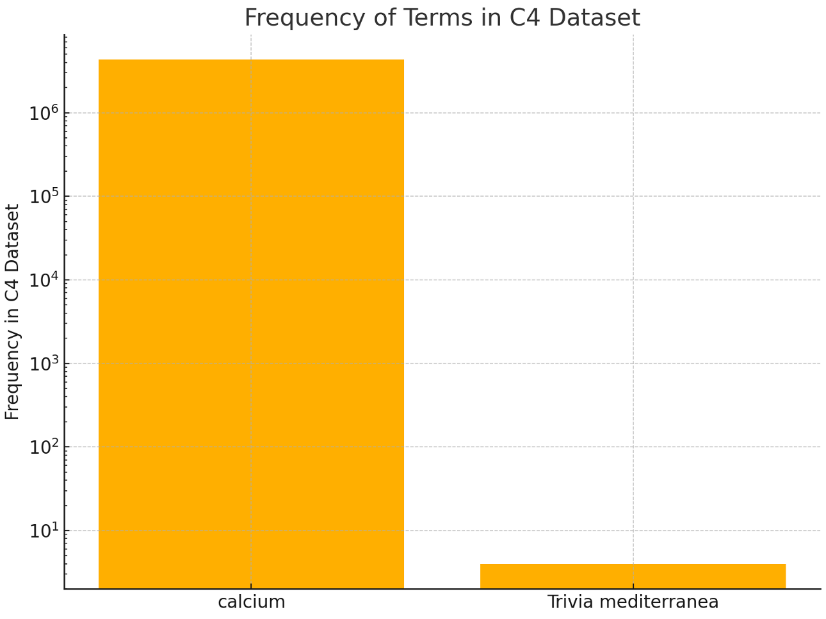
```
You can get calcium from dairy products like milk, yogurt and
cheese, canned fish with soft bones (sardines, anchovies and
salmon; bones must be consumed to get the benefit of calcium),
dark-green leafy vegetables (such as kale, mustard greens and
turnip greens) and even tofu (if it's processed with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

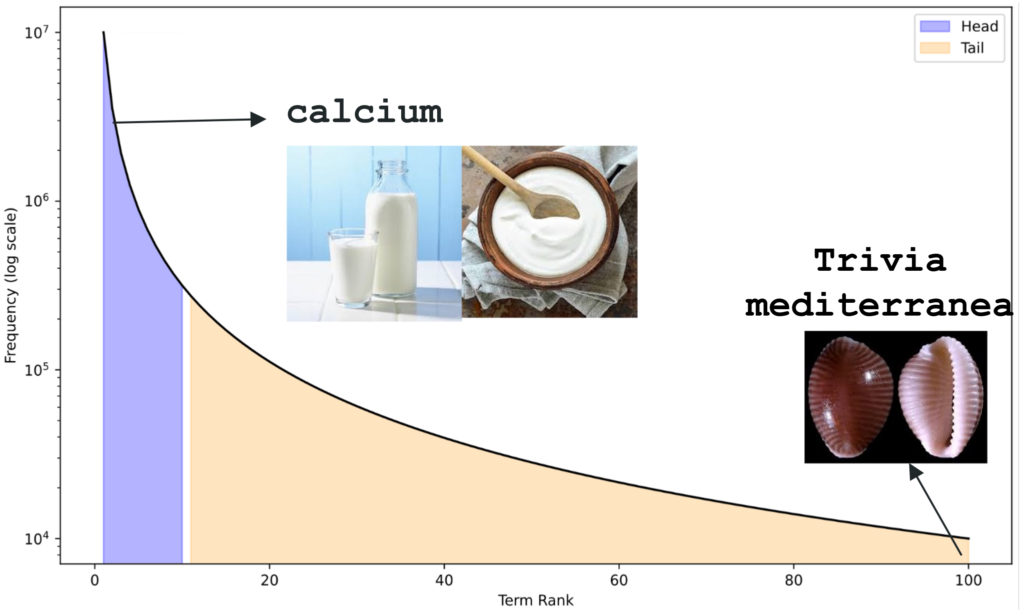
But there also exists very niche topics:

```
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

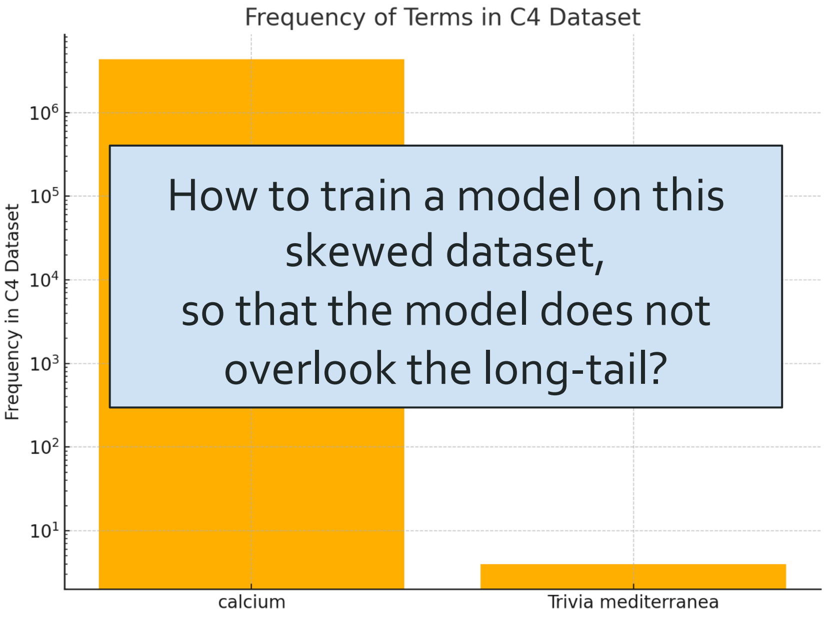
In the entire C4 dataset, trivia mediterranea only appears 4 times.



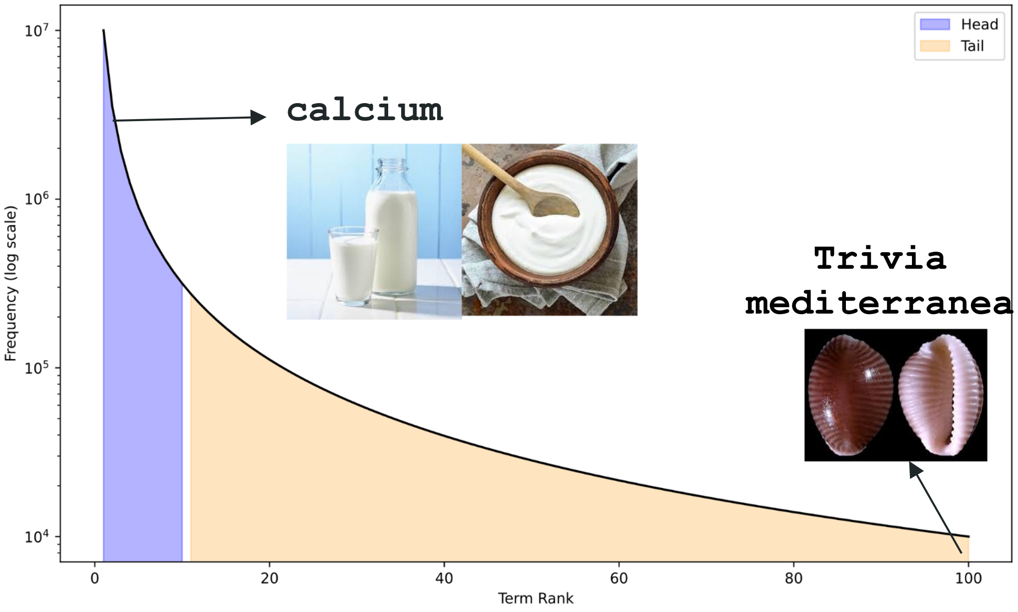


The “long-tailedness” of knowledge

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

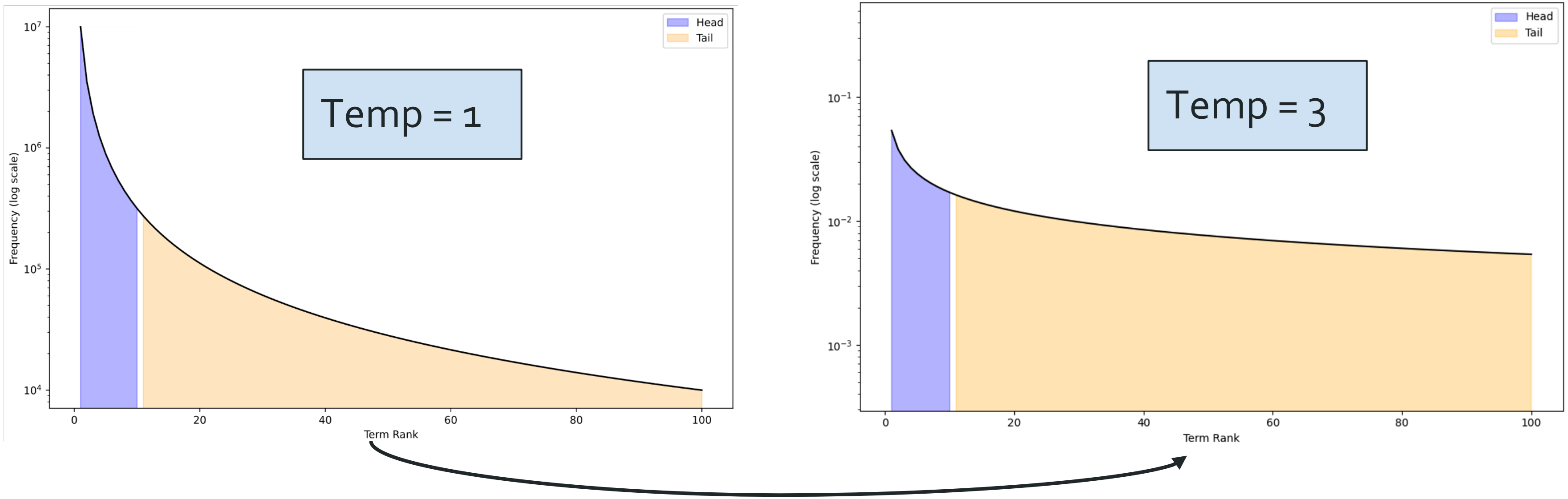


How to train a model on this skewed dataset,
so that the model does not overlook the long-tail?



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...

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

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

$$\forall i \in \{1, 2, \ldots, K\}: \; p(i; \tau) = \frac{|\mathcal{D}_i|^{\frac{1}{\tau}}}{\sum_{j=1}^{K} |\mathcal{D}_j|^{\frac{1}{\tau}}}$$

Temperature

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

Solution 2: (Scalarization)

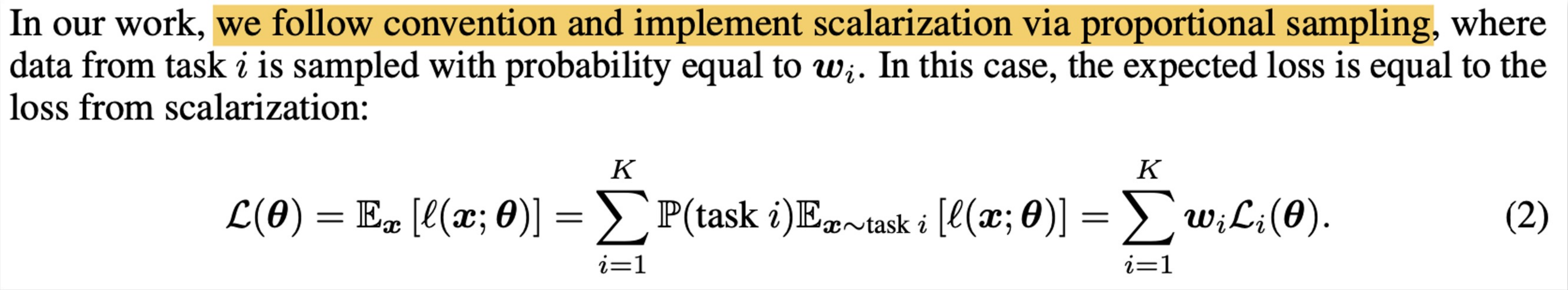
We assign a much higher weight to the loss of infrequent domains.

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

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

$$L_{\mathrm{TS}} = \mathop{\mathbb{E}}_{\substack{k \sim p \\ x \sim \mathcal{D}_k}} \Big[\ell(x) \Big] \qquad \forall i \in \{1, 2, ..., K\} : \; p(i; \tau) = \frac{|\mathcal{D}_i|^{\frac{1}{\tau}}}{\sum_{j=1}^{K} |\mathcal{D}_j|^{\frac{1}{\tau}}}$$

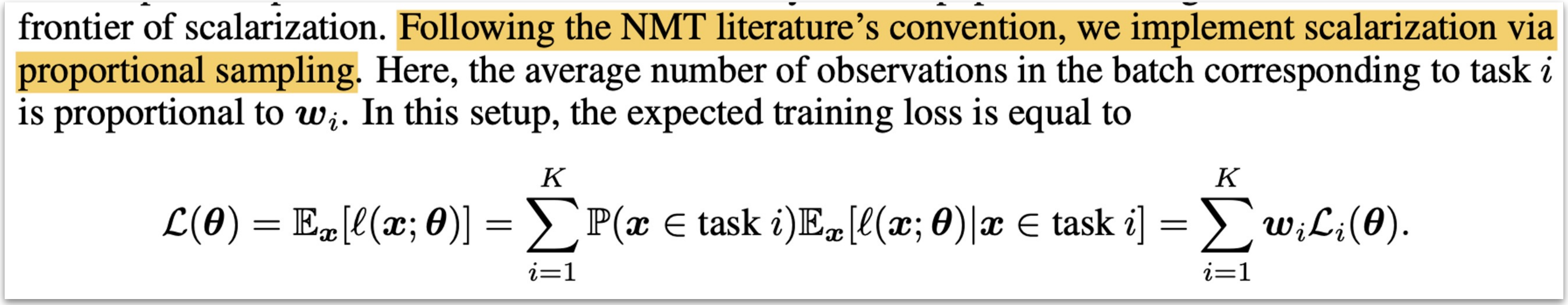
Temperature Sampling often assumed to be equivalent to Scalarization



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

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{x};\boldsymbol{\theta})\right] = \sum_{i=1}^{K} \mathbb{P}(\text{task } i)\mathbb{E}_{\boldsymbol{x}\sim\text{task } i}\left[\ell(\boldsymbol{x};\boldsymbol{\theta})\right] = \sum_{i=1}^{K} \boldsymbol{w}_i \mathcal{L}_i(\boldsymbol{\theta}). \qquad (2)$$

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



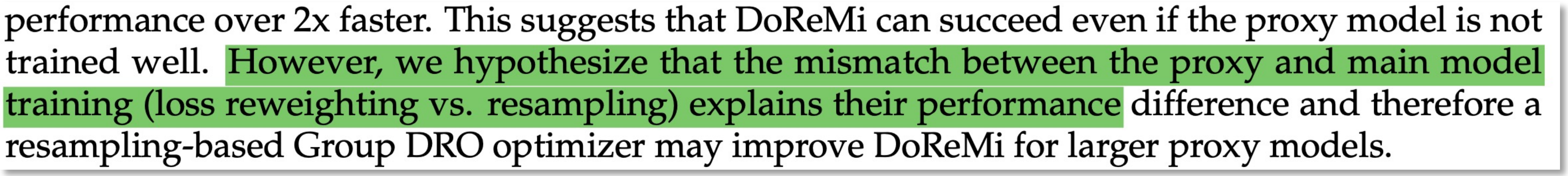
frontier of scalarization. Following the NMT literature's convention, we implement scalarization via proportional sampling. Here, the average number of observations in the batch corresponding to task i is proportional to $\boldsymbol{w}_i$. In this setup, the expected training loss is equal to

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

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

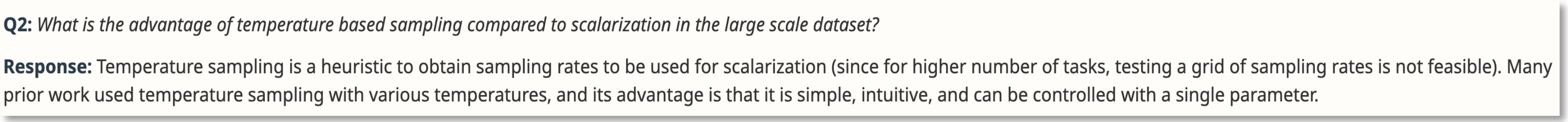
Temperature Sampling often assumed to be equivalent to Scalarization

The reason why they differ is often overlooked



performance over 2x faster. This suggests that DoReMi can succeed even if the proxy model is not trained well. However, we hypothesize that the mismatch between the proxy and main model training (loss reweighting vs. resampling) explains their performance difference and therefore a resampling-based Group DRO optimizer may improve DoReMi for larger proxy models.

DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining (Xie et al., NeurIPS 2023)



Q2: *What is the advantage of temperature based sampling compared to scalarization in the large scale dataset?*

Response: Temperature sampling is a heuristic to obtain sampling rates to be used for scalarization (since for higher number of tasks, testing a grid of sampling rates is not feasible). Many prior work used temperature sampling with various temperatures, and its advantage is that it is simple, intuitive, and can be controlled with a single parameter.

Author Rebuttal of Order Matters in the Presence of Dataset Imbalance for Multilingual Learning

Scalarization (S) = Temperature Sampling (TS)

This is True! under *full* Gradient Descent (GD)

Theorem 1: For any sampling temperature, there exists a set of weights that makes S loss = TS loss (on the whole data)

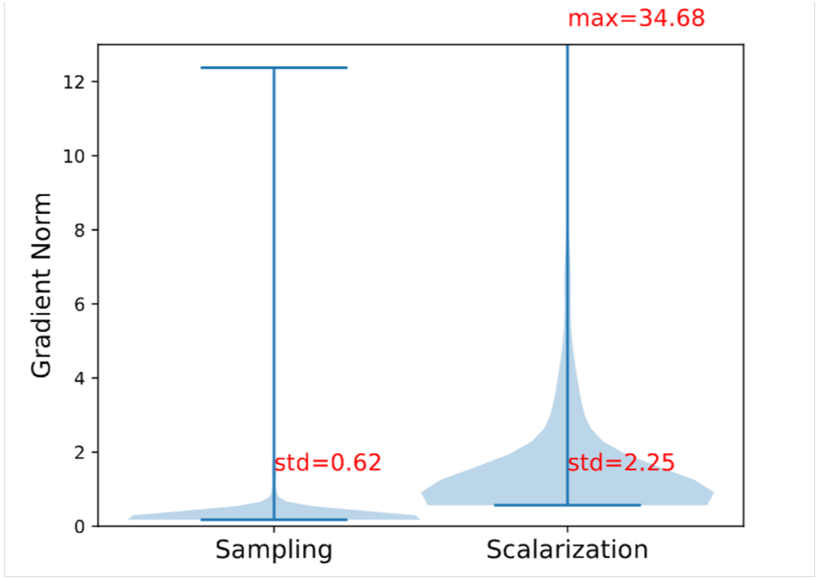
$$L_{\mathrm{S}} = \sum_{k=1}^{K} w_k \sum_{x \in \mathcal{D}_k} \ell(x)$$

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

Scalarization (S) = Temperature Sampling (TS)

They are not! Under *Stochastic* Gradient Descent (SGD)

Theorem 2: Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.



Scalarization (S) = Temperature Sampling (TS)

They are not! Under *Stochastic* Gradient Descent (SGD)

Theorem 2: Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

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

Theorem 3: larger temperature induces a larger variance gap!

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

monotonically increases when $\tau \geq 1.$

So far

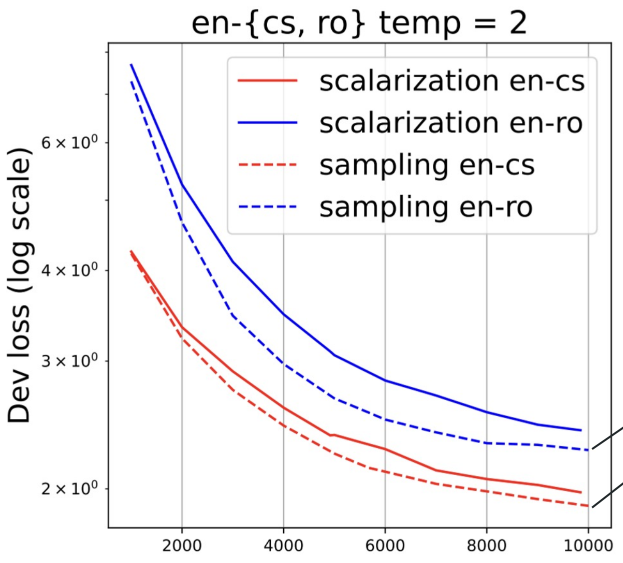
Theorem 2: Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

Theorem 3: larger temperature induces a larger variance gap!

What does the theory tell us about model training?

Theorem 2: Temperature Sampling induces a larger variance between gradients compared to Scalarization in SGD.

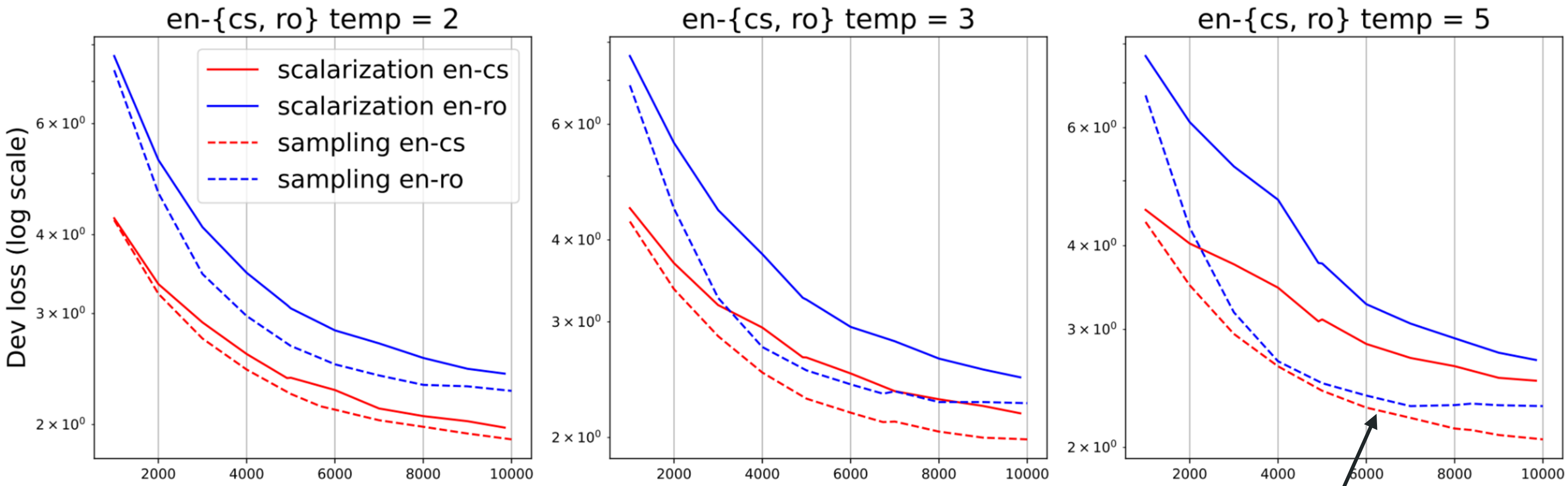
- It is well-known that variance-reduction accelerates the convergences of SGD. (Sutskever et al., 2013; Kingma and Ba, 2015)
- Temperature Sampling induces less variance, therefore it should converge faster!



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

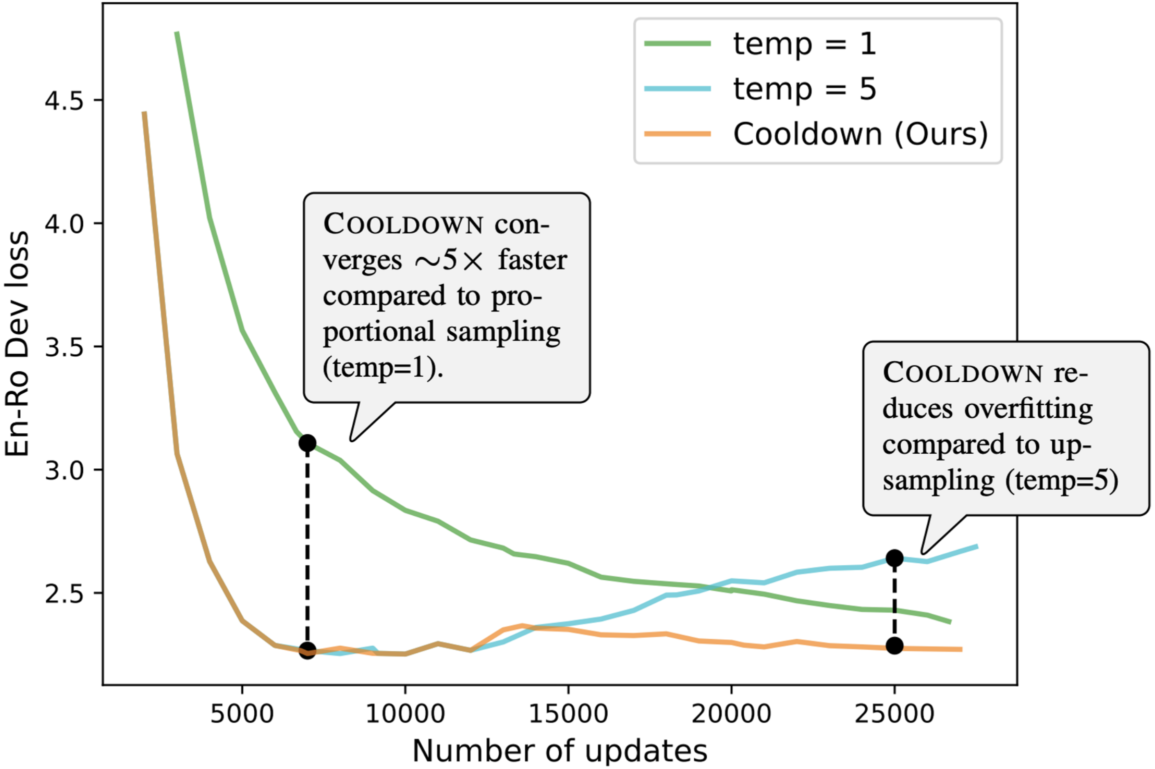
Theorem 3: larger temperature induces a larger variance gap!

Theorem 3 implies:



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

Cooldown: Initially use large temp, then use small temperature

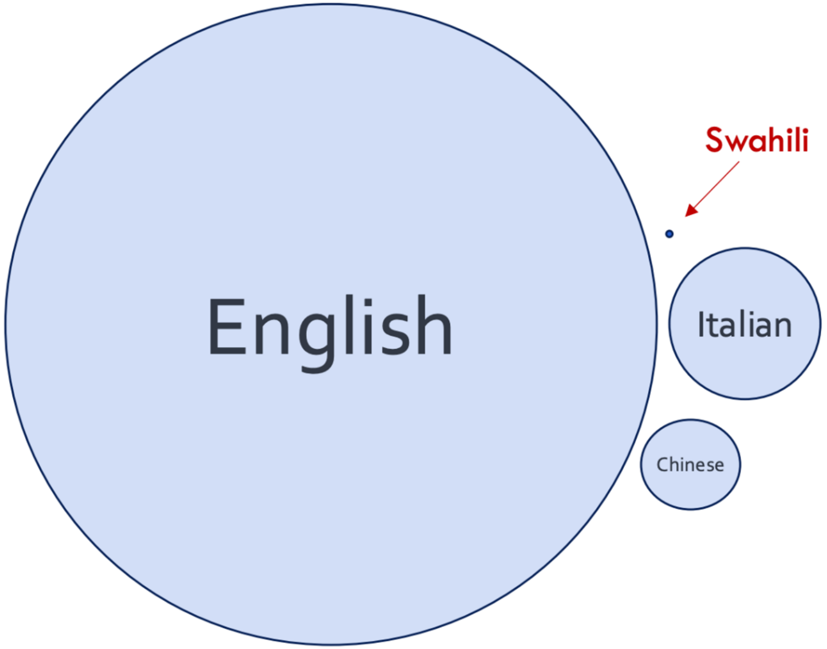


COOLDOWN converges ~5× faster compared to proportional sampling (temp=1).

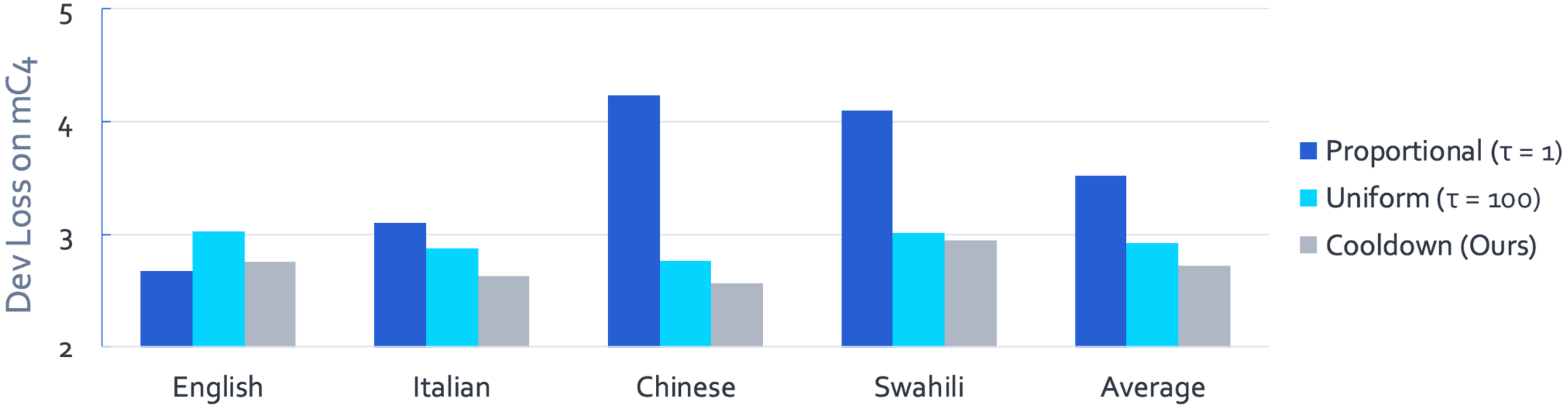
COOLDOWN reduces overfitting compared to upsampling (temp=5)

We used different languages as a case study:

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

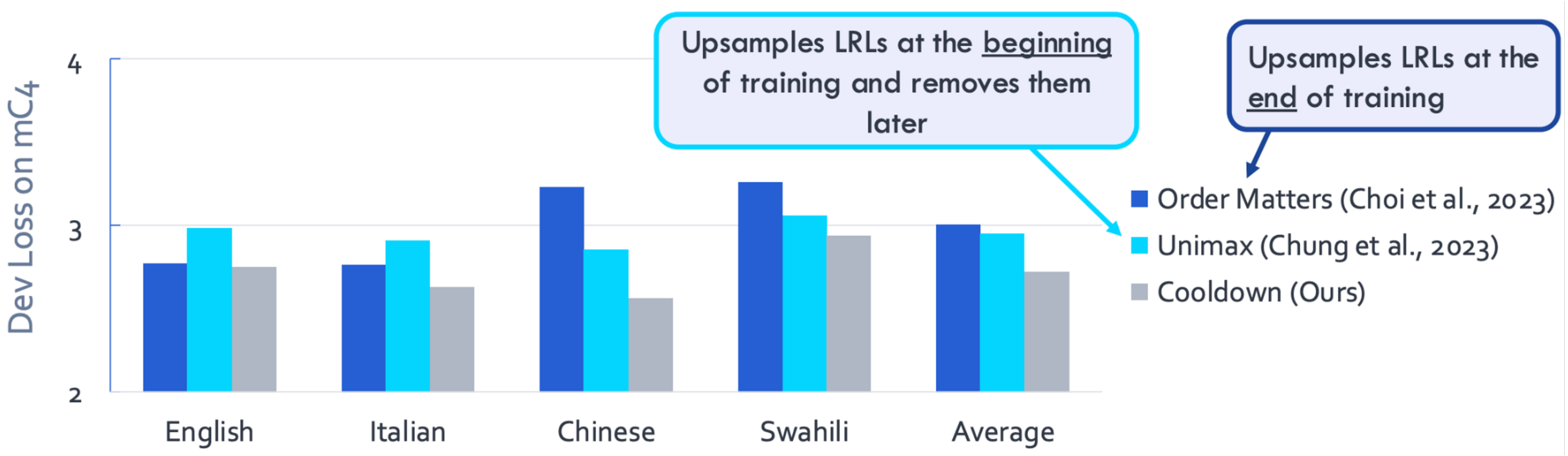


Cooldown: Dev Loss on mC4 (lower is better)



Cooldown outperforms fixed temperature sampling!

Cooldown: Dev Loss on mC4 (lower is better)



Upsamples LRLs at the beginning of training and removes them later

Upsamples LRLs at the end of training

Cooldown outperforms existing work that dynamically adjusts the sampling temperature!

Summary

- Two common approaches for dealing with imbalanced data:
  - Temperature Sampling: Resampling of domains
  - Scalarization: Reweighting of per-domain losses
- Despite common perception, these two are not equivalent
  - Temperature Sampling leads to lower variance in gradient estimates
  - ...and faster convergence
- We propose ❄️ COOLDOWN ❄️
  - A suggested recipe for imbalanced (pre-) training

For more results:

https://arxiv.org/pdf/2410.04579