PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Scaling Hardness-Impact Training:

Investigating Data Ordering With TinyLlama[1]

Hannah Lee† **Sidharth Lakshmanan**† **Kevin Farhat**† **Chongjiu Gao**†

Summary

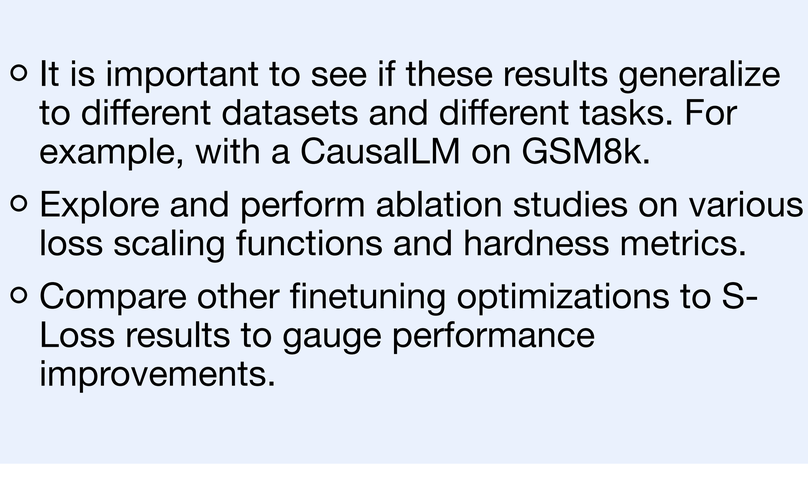
- Data ordering by model-based hardness values results in a maximum of **1.38% improvement** on finetuned TinyLlama performance on SNLI[2].
- We introduce **S-Loss**, a technique to scale loss by hardness values during training to mimic data ordering in a distributed setting, **matching/exceeding 1.68% improvement.**

Research Questions

- How can we leverage data hardness metrics to improve model performance?
- How can we adapt curriculum learning for a distributed setting?

Curriculum Learning[3, 5]

Ordering finetuning data by hardness showed that ascending difficulty outperforms descending difficulty for all hardness calculations.



Data Hardness Metrics

Dataset Cartography[4]

A model-based tool to characterize and diagnose datasets by evaluating various metrics during training.

The two prominent metrics are:

- Confidence: The model's confidence in the true class.
- Variability: The variability of the confidence across epochs.

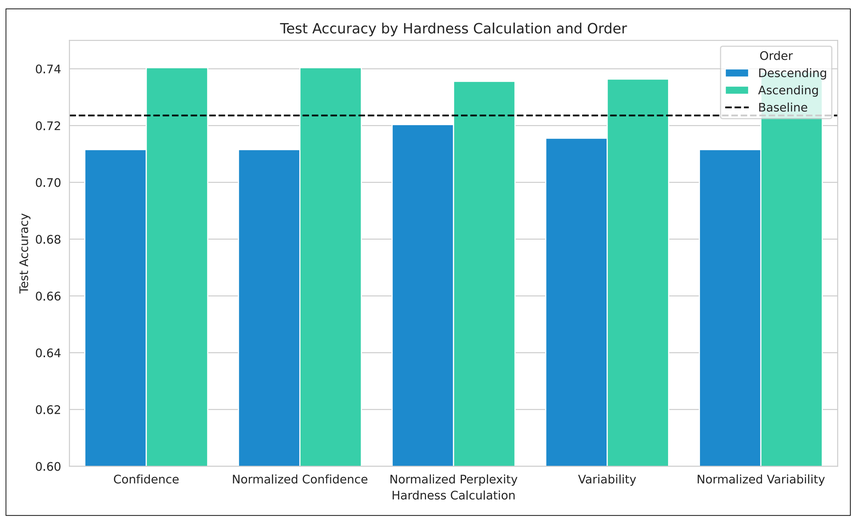
N-gram Perplexity

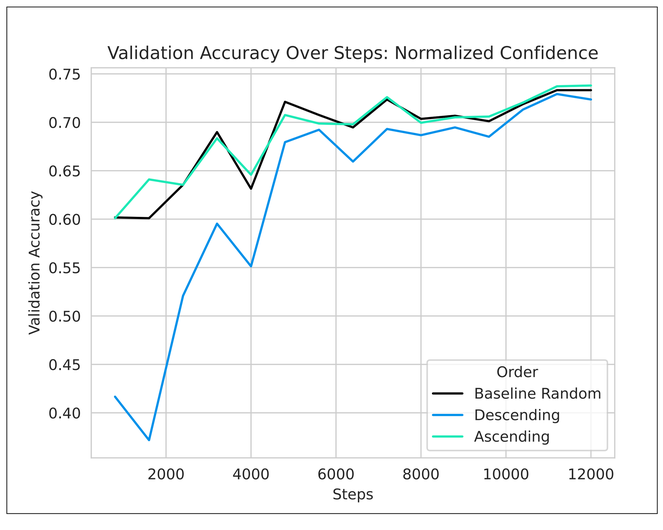
An even linear interpolation of unigram, bigram, and trigram perplexity on the concatenated inputs.

Results

Results after finetuning TinyLlama with S-Loss on the SNLI for 2 epochs on 4 GPUs with a per GPU batch size of 8, a learning rate of 4e-4, and AdamW.

	Baseline	Confidence	Normalized Confidence	Normalized Perplexity	Variability	Normalized Variability
Test Accuracy	72.36%	74.04%	74.04%	73.56%	73.64%	73.88%



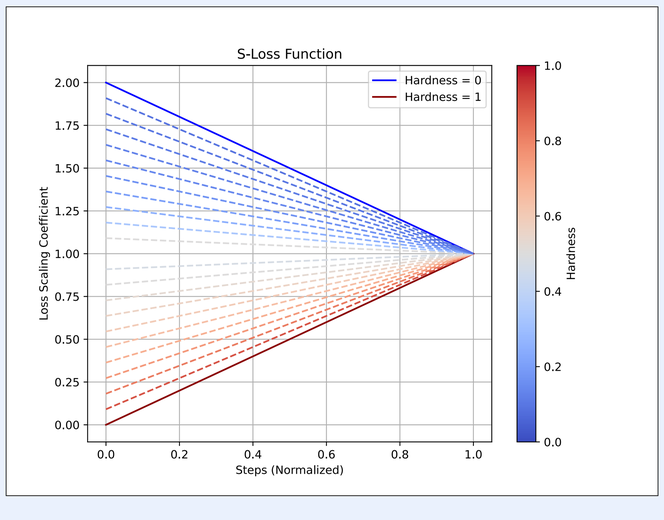


S-Loss

A custom scaling function was created to scale the fine-tuning Cross-Entropy loss by a coefficient. The coefficient to scale by is defined by the hardness of the sample (normalized) and the current training step (normalized by the total number of steps).

$$\mathscr{L} = \alpha H(p,q) = -\alpha \sum p(x)\log q(x)$$

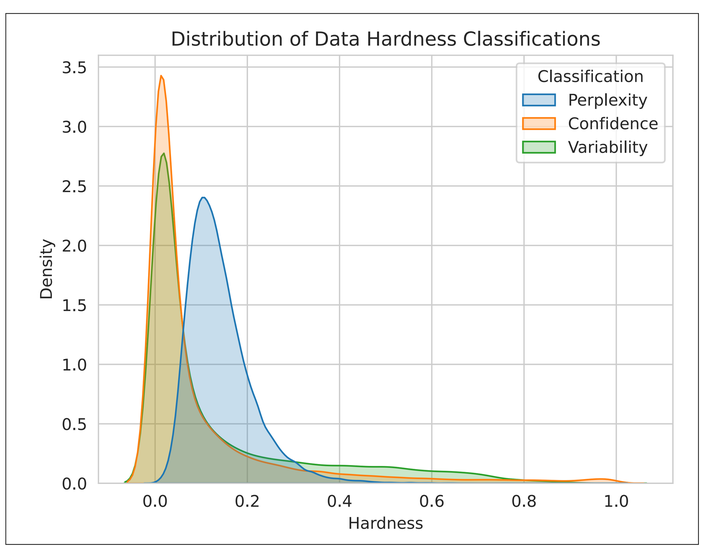
$$\alpha = \frac{2h_x - 1}{s - 1} + 1 \quad \text{where} \quad h_x \in [0,1], \quad s \in [0,1)$$



Examples	Easy Premise	Easy Hypothesis	Hard Premise	Hard Hypothesis
Perplexity	Being exposed to hot sun like this can cause skin cancer.	Using sunblock may prevent diseases	A construction worker sweeping	A girl drinks soda
Variability	A black dog in a grassy field fetching a colourful toy.	A boy threw the frisbee.	A man Rollerblades across a yellow pole at night.	A man is sleeping at his cubicle at work.
Confidence	A boy wields a net in a boat with another in the middle of a lake.	A boy rides in a hot air balloon.	Photographers take pictures of a girl sitting in a street.	The photographer is taking a picture of a boy

Future Work

- It is important to see if these results generalize to different datasets and different tasks. For example, with a CausalLM on GSM8k.
- Explore and perform ablation studies on various loss scaling functions and hardness metrics.
- Compare other finetuning optimizations to S-Loss results to gauge performance improvements.



† Indicates equal contribution

[1] TinyLlama https://arxiv.org/abs/2401.02385
[2] SNLI Dataset https://arxiv.org/pdf/1508.05326v1.pdf
[3] Curriculum Learning for Natural Language Understanding https://aclanthology.org/2020.acl-main.542.pdf
[4] Dataset Cartography https://arxiv.org/pdf/2009.10795.pdf
[5] Curriculum Learning for Language Modeling https://arxiv.org/pdf/