NeurIPS 2026

SimReg

Achieving Higher Performance in the Pretraining via Embedding Similarity Regularization

Yan Sun[1,2*] Guoxia Wang[1*] Jinle Zeng[1] Jiabin Yang[1] Shuai Li[1]
Li Shen[3] Dacheng Tao[4†] Dianhai Yu[1†] Haifeng Wang[1]

[1] Baidu Inc. [2] The University of Sydney
[3] Shenzhen Campus of Sun Yat-sen University [4] Nanyang Technological University

* Equal contribution † Corresponding authors

The representation challenge in next-token prediction

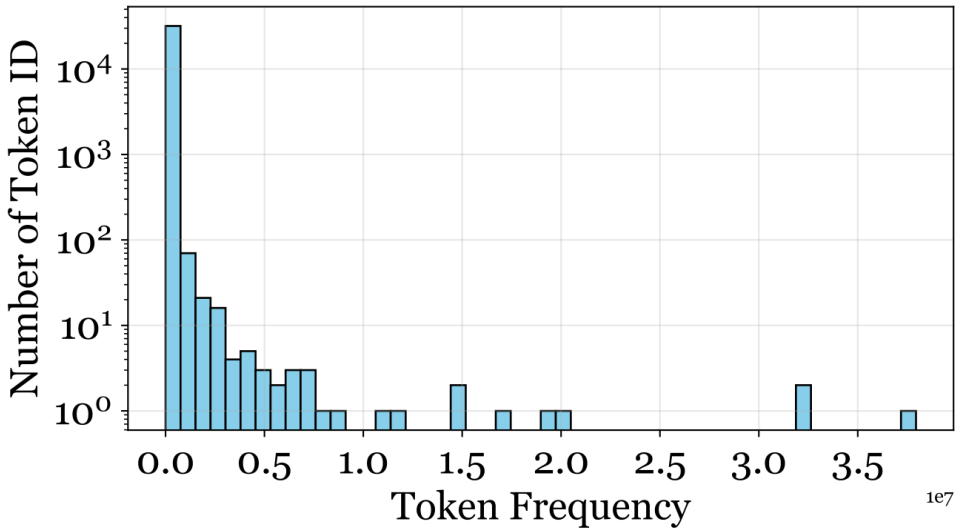
Different contexts can share the same target token

The cat jumps over

A child paints near

Next-token target: walls

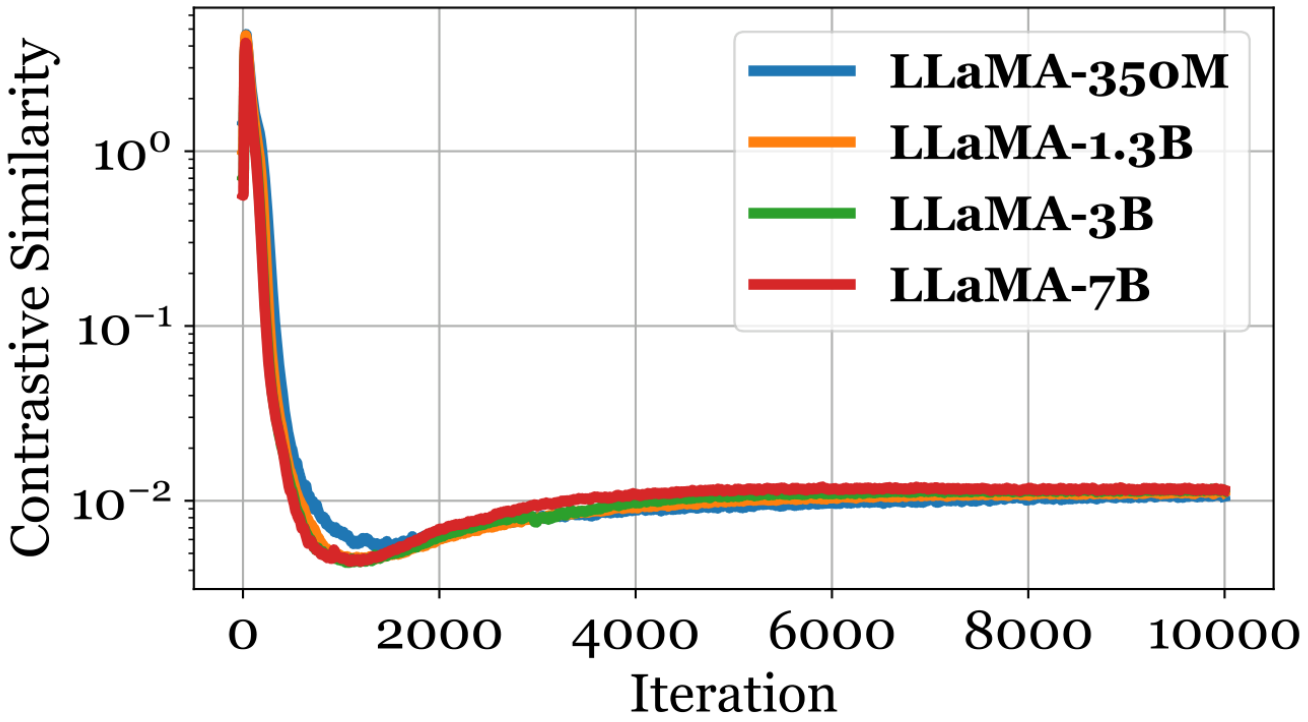
SimReg groups hidden states by the ground-truth next token, after the causal label shift.



Token frequencies are highly imbalanced

Frequent targets dominate sample-wise training. The paper measures a long tail on C4.

Cross-entropy leaves geometry weakly constrained



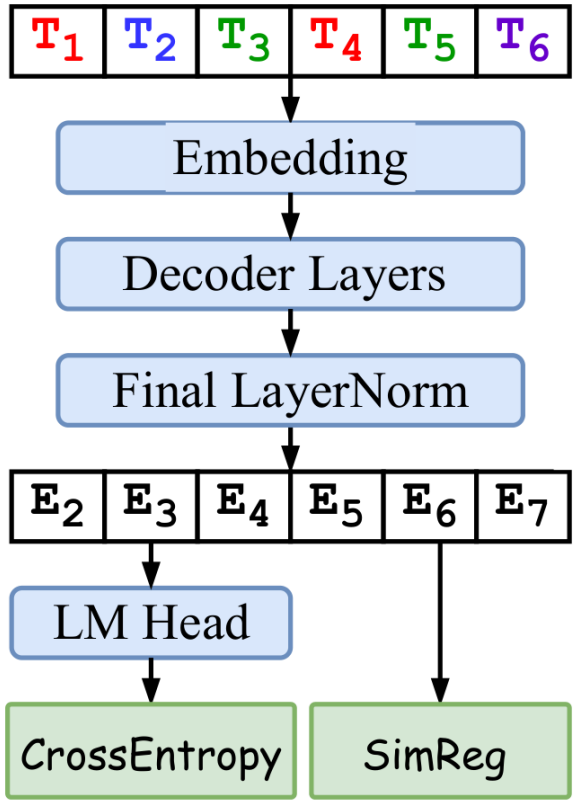
A shared empirical plateau

Cross-entropy initially reduces the contrastive similarity loss.

The loss then stabilizes across model sizes, while cross-entropy continues to improve.

Larger models alone do not remove this plateau in the reported experiments.

SimReg supervises the final token representations



Hidden states before the LM head

Apply regularization to the final normalized representations used for next-token classification.

Same target labels define positive pairs.
Different target labels define negative pairs.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \frac{1}{n} \sum_{i=1}^{n} \text{softplus}(\mathcal{L}_i^{\text{sr}})$$

Training uses existing token labels and requires no additional annotations.

A contrastive objective over next-token labels

$$\mathcal{P}_i = \{j : y_j = y_i\}, \qquad \mathcal{N}_i = \{j : y_j \neq y_i\}, \qquad i \in \mathcal{P}_i$$

$$s_{ij} = \frac{\cos(\mathbf{e}_i, \mathbf{e}_j)}{\tau}, \qquad \phi_{ij} = \exp(s_{ij})$$

$$\mathcal{L}_i^{\mathrm{sr}} = \log \sum_{j \in \mathcal{N}_i} \exp(s_{ij}) - \log \sum_{j \in \mathcal{P}_i} \exp(s_{ij})$$

Design choice	Role in stable pretraining
Cosine similarity and temperature τ	Controls directional similarity without norm dominance
Self-pair in the positive set	Keeps the positive group nonempty for rare targets
Softplus with coefficient λ	Makes the regularizer nonnegative and controls its weight

A margin-based perspective on feature clustering

$$m_i = z_{i,y_i} - \max_{c \neq y_i} z_{i,c}$$

Classification margin

$$m_i \geq m_i^+ - \sqrt{2}\, L_P \, \|\mathbf{e}_i - \mathbf{e}_i^+\|_2$$

A smaller distance to the positive center tightens this lower bound, for a fixed center margin.

$$\ell_i = \log\left(1 + \sum_{c \neq y_i} e^{z_{i,c} - z_{i,y_i}}\right) \leq (C-1)e^{-m_i}$$

Larger margins lower the exponential upper bound on cross-entropy.

The interpretation assumes a Lipschitz logit map. Favorable center dynamics are an additional condition.

Evaluation across dense and MoE pretraining

Architecture	Model sizes	Training tokens
LLaMA dense	350M	≈13B
LLaMA dense	1.3B / 3B / 7B	≈52B each
Mixtral MoE	8×1B	≈52B

Nine zero-shot tasks

ARC-E, ARC-C, BoolQ
HellaSwag, OpenBookQA
PIQA, MMLU
WinoGrande, SciQ

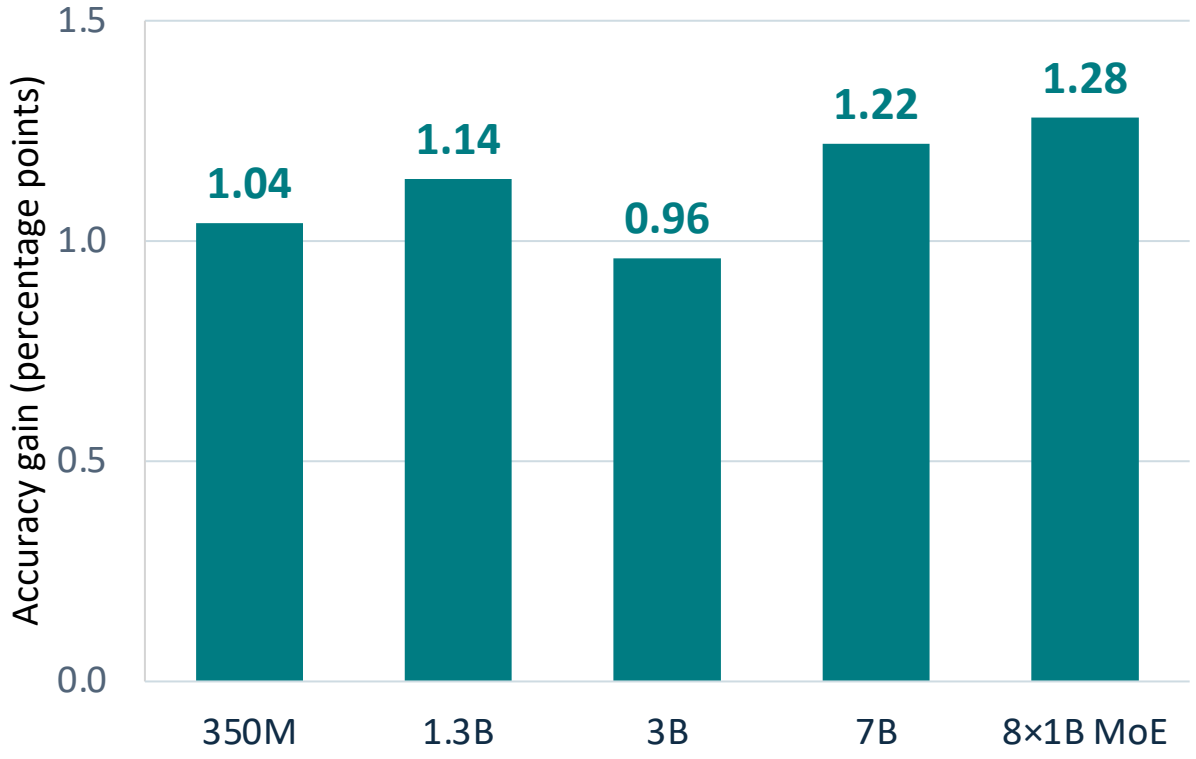
Shared training setup

2,048-token sequences. AdamW with cosine decay.
AdaGC clipping for every method.

Comparators

CE, SimCSE, CPretrain,
CReg, SCE

SimReg improves average zero-shot accuracy

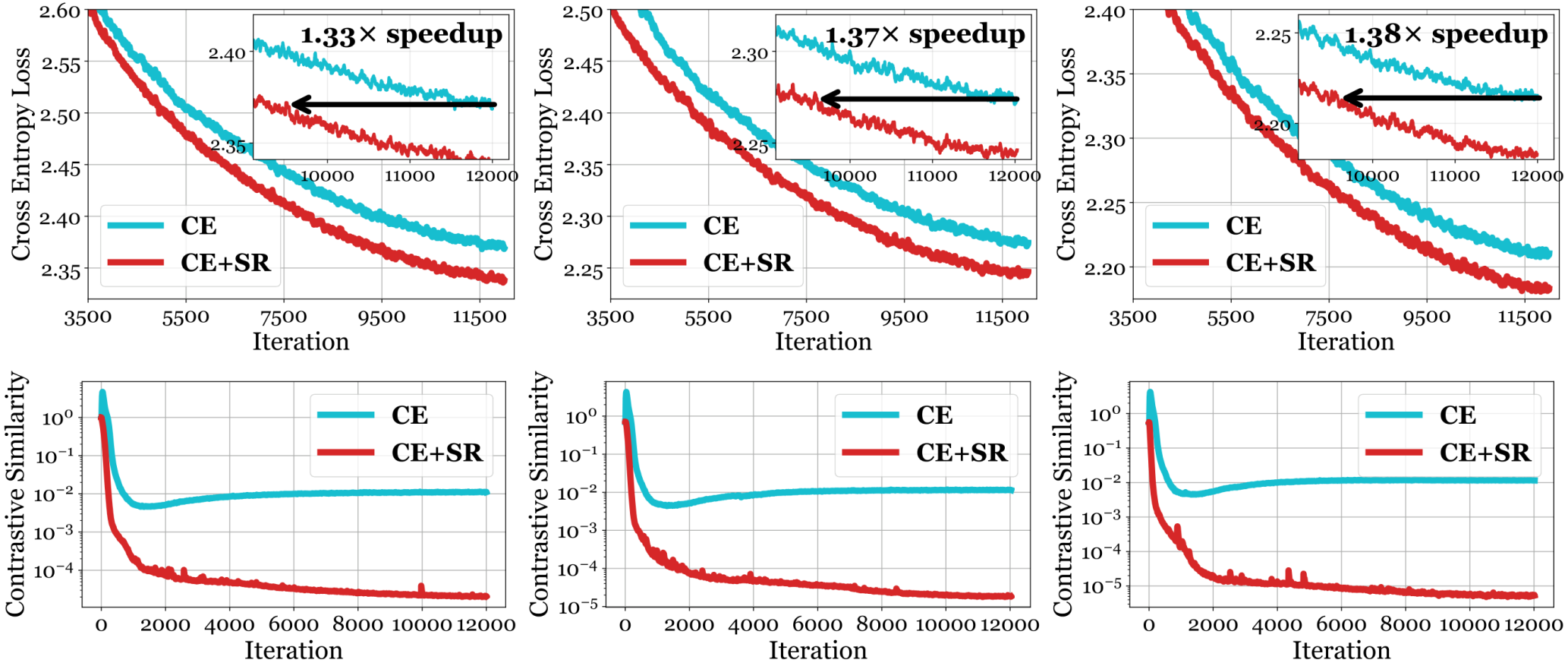


Model	CE	SimReg
350M	43.06	**44.10**
1.3B	47.52	**48.66**
3B	50.00	**50.96**
7B	51.45	**52.67**
8×1B MoE	50.41	**51.69**

Nine-task mean accuracy (%)

The highest average accuracy among the compared methods at every dense scale

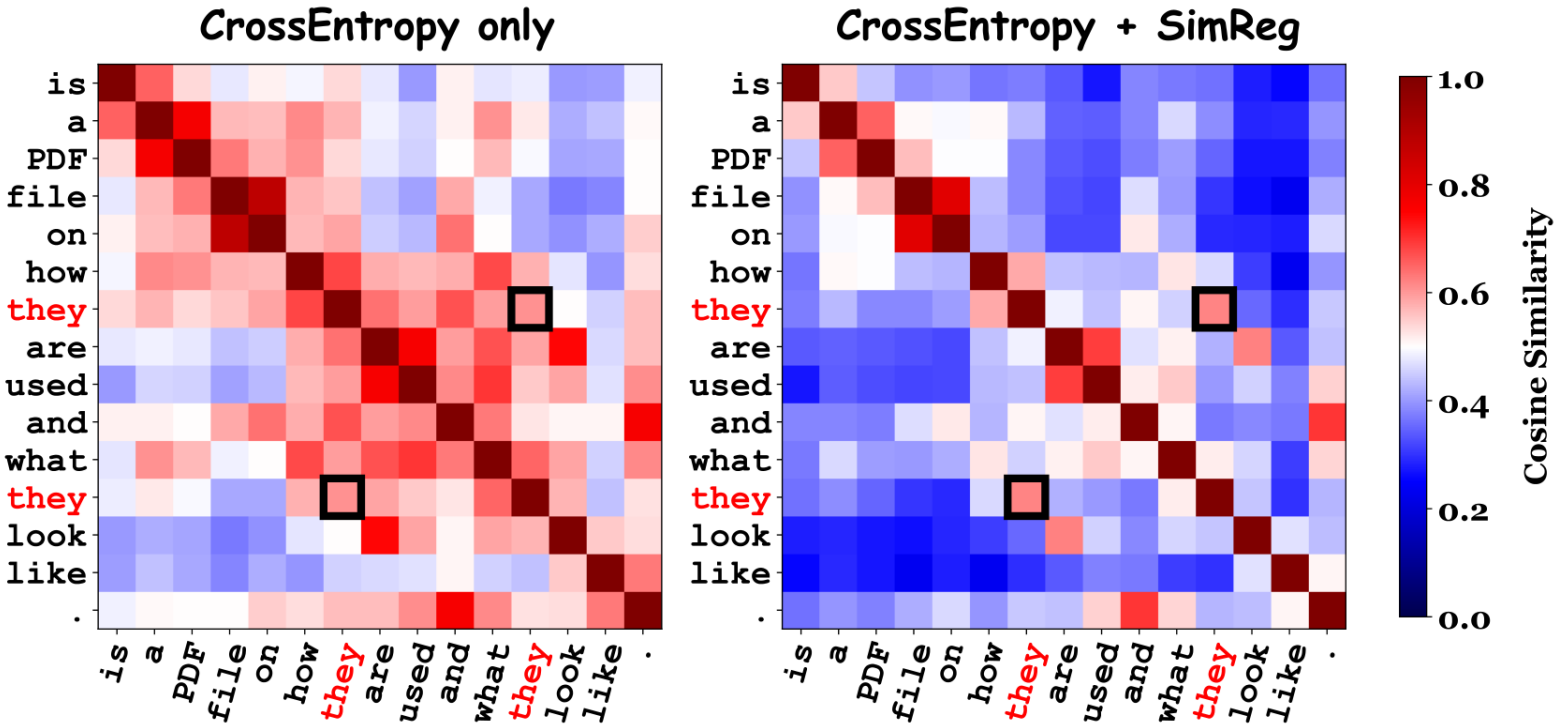
Training reaches the same loss in fewer updates



(a) LLaMA-1.3B. (b) LLaMA-3B. (c) LLaMA-7B.

Reported convergence factors: 1.33×, 1.37×, and 1.38×. Runtime per iteration is measured separately.

Token similarities become more discriminative



Same sequence, different training objectives

CE + SimReg reduces broad cross-token similarity.

Repeated target labels retain structure while many different-label pairs separate.

Illustrative LLaMA-7B sequence from the paper. Colors show cosine similarity between prediction features.

Chunk-wise SimReg preserves downstream gains

Model	CE	Full	Chunk
350M	43.06	**44.10**	**43.96**
1.3B	47.52	**48.66**	**48.33**
3B	50.00	**50.96**	**50.94**
7B	51.45	**52.67**	**52.77**
8×1B MoE	50.41	**51.69**	**51.62**

Nine-task mean accuracy (%)

Compute within each chunk

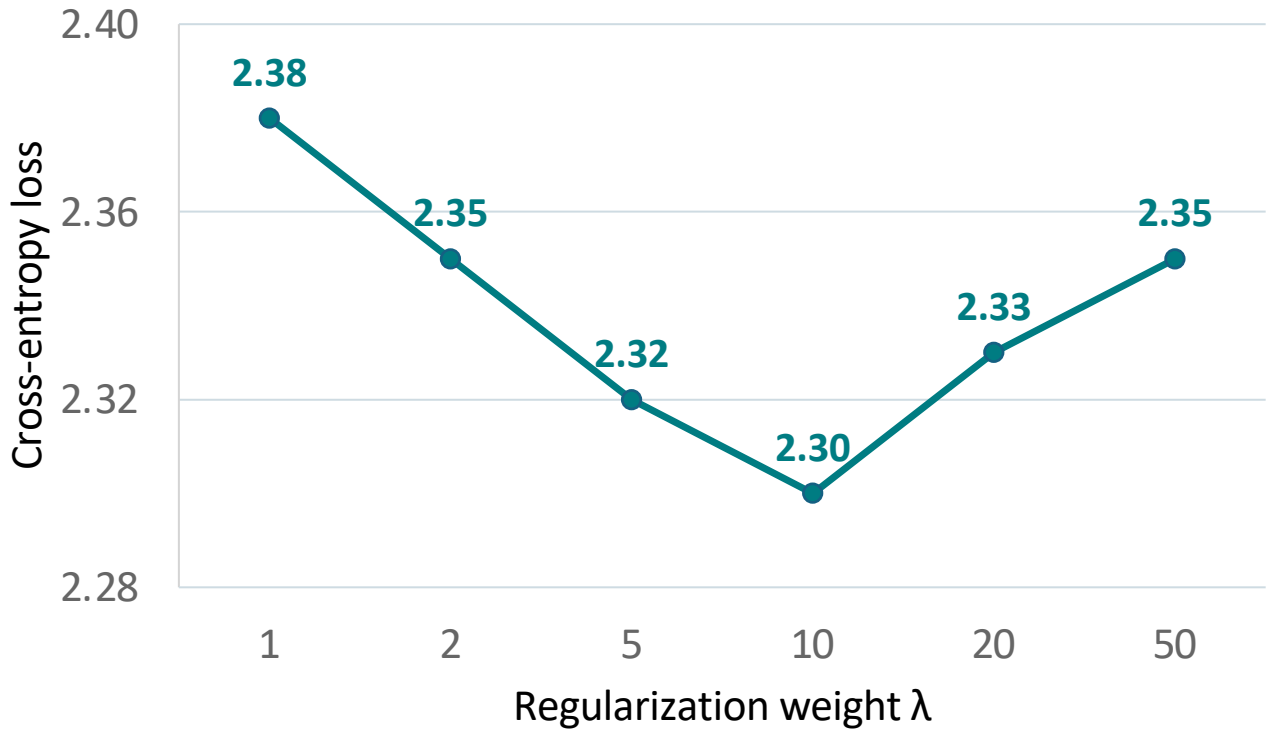
For b equal chunks, pairwise work decreases from

$$\mathcal{O}(n^2) \longrightarrow \mathcal{O}(n^2/b)$$

Less communication under sequence parallelism.

The 7B runtime profile uses 1,024-token chunks. Optimal chunk size depends on the training setting.

A practical default with a clear λ trade-off



Suggested starting point

$$\tau = 0.01, \qquad \lambda \approx 10\sqrt{\frac{d}{1024}}$$

d is the hidden-state dimension.

Apply SimReg at the final layer. Tune λ using cross-entropy or validation perplexity.

Stronger similarity regularization alone does not guarantee better language modeling loss.

Measured overhead is small on the 7B model

Training objective	Time (s / iteration)	Reported memory (GB)
Cross-entropy	2.295	373.45
CE + full SimReg	2.296	375.28
CE + chunk-wise SimReg	2.295	373.84

+0.04%

time per iteration, full SimReg

+0.49%

reported memory, full SimReg

+0.10%

reported memory, chunk-wise

H800 profile, hidden size 4,096, chunk size 1,024. Relative changes use the rounded values in Table 4.

SimReg improves pretraining with one auxiliary loss

Direct supervision of prediction-feature geometry

Same-target consistency complements the cross-entropy objective during pretraining.

+0.96–1.28 pp

mean zero-shot accuracy across five models

1.33–1.38×

reported convergence acceleration

Scope and open questions

Larger scales, other architectures and data distributions, and model-dependent chunk sizes remain to be studied.

sunyan25@baidu.com