LM scaling laws & zero-sum learning

Mila · Université de Montréal · Capital One

Andrei Mircea; Ekaterina Lobacheva; Supriyo Chakraborty; Nima Chitsazan; Irina Rish

TL;DR

Can we explain LM scaling improvements in terms of training dynamics?
Yes! Scaling improves LMs by mitigating loss deceleration (transition in training dynamics characterized by gradient opposition and zero-sum learning between tokens).

Explaining LM scaling laws: loss deceleration

Loss deceleration: rapid slow-down in rate of loss improvement observed early during LLM pretraining. Characterized by piece- wise linear behavior (log-log).

Quantifiable with 1-break BNSL.

Can capture loss improvements due to scaling.

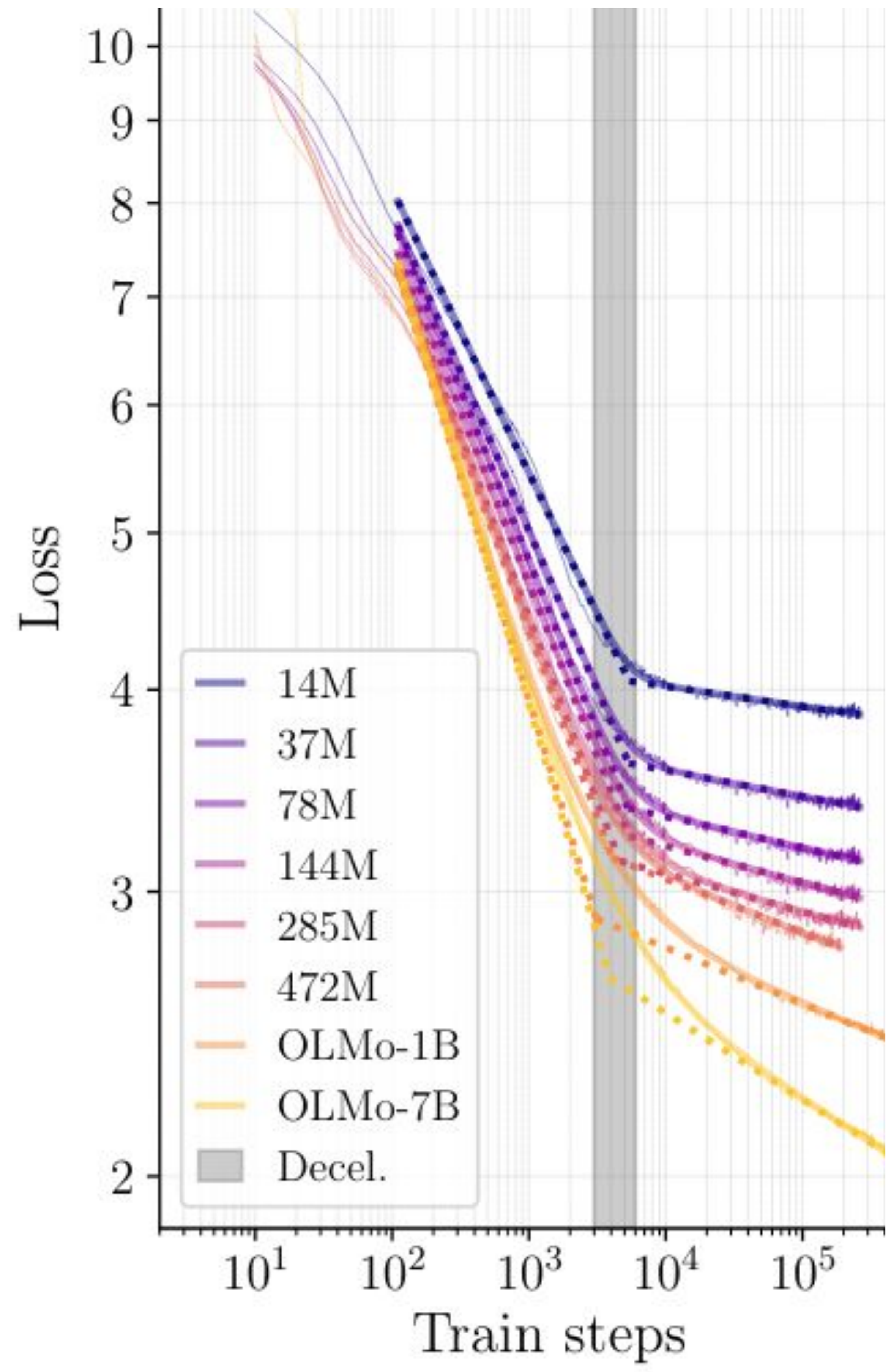
$$L(t) - a = \left(bt^{-c_0}\right)\left(1 + (t/d_1)^{1/f_1}\right)^{-c_1 f_1}$$

t_d : d_1, the step at which deceleration occurs.

L_d : $bd_1{}^{-c_0}$, the loss at which deceleration occurs.

r_d : $c_0 + c_1$, the log-log loss slope after deceleration.

$\hat{L}_T$: $\log(L_T) \approx \log(\hat{L}_T) = \log(L_d) - r_d \log(T/t_d)$



Explaining deceleration: zero-sum learning (ZSL)

Zero-sum learning: degenerate training dynamics where loss improvements in one set of examples are cancelled out by degradation in another.

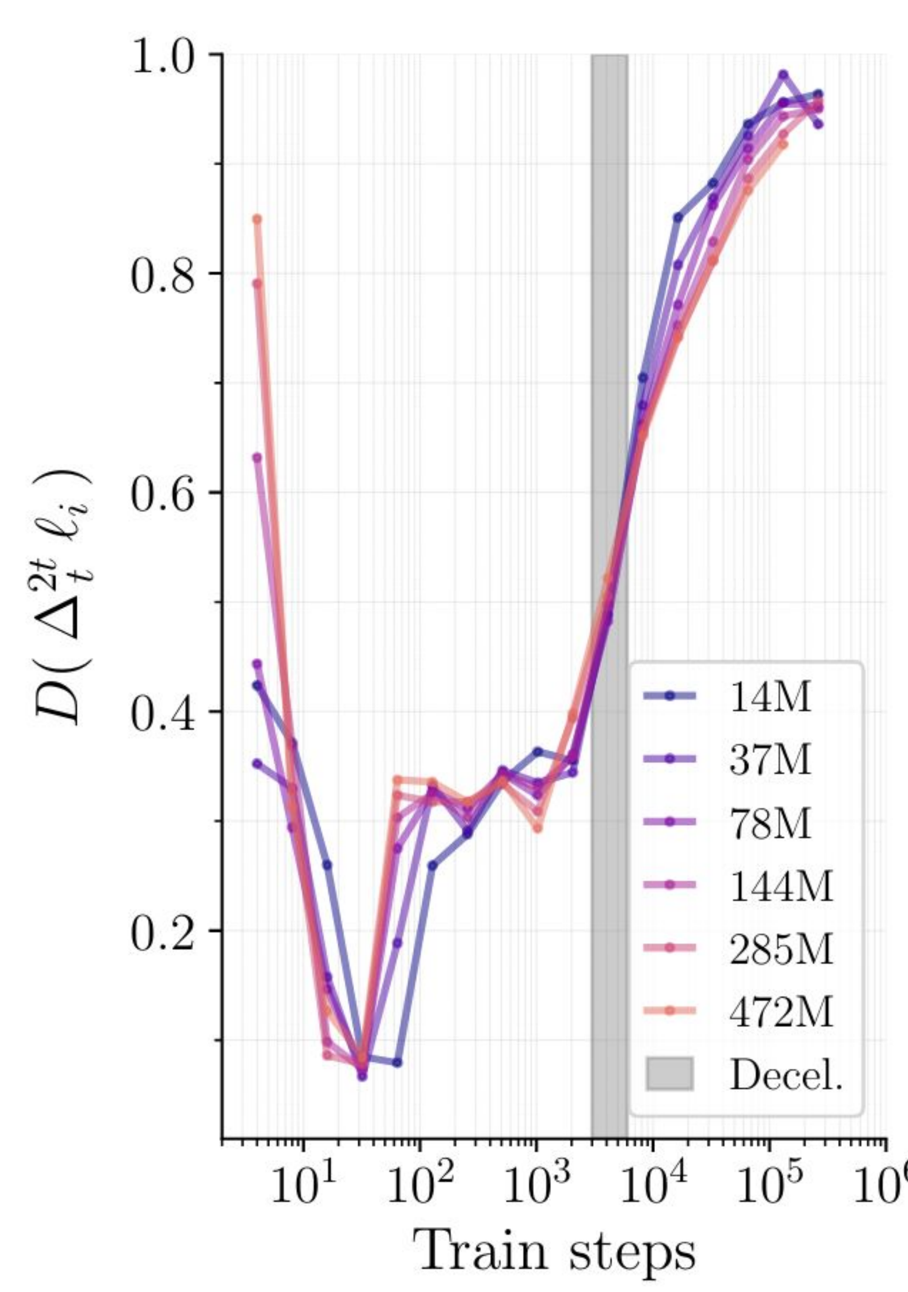
ZSL can be quantified with **destructive interference**

$$D(\Delta\ell) = 1 - \texttt{abs}(\textstyle\sum_i \Delta\ell_i)/\sum_i \texttt{abs}(\Delta\ell_i), \quad D(\Delta\ell) \in [0,1]$$

where D=1 indicates complete interference and ZSL

A. Occurs simultaneously with deceleration and can be shown to fundamentally bottleneck loss improvements.

B. Scaling reduces ZSL after deceleration (improved slope)



Explaining ZSL: systematic gradient opposition

Systematic gradient opposition: model weight configuration with >99% destructive interference between per-example gradients.

$$D(\nabla_\theta \ell) = 1 - \texttt{abs}(\textstyle\sum_i \nabla_\theta \ell_i)/\sum_i \texttt{abs}(\nabla_\theta \ell_i)$$

A. Occurs across parameters simultaneously with ZSL; shown to fundamentally cause ZSL.

B. Scaling reduces gradient opposition before deceleration (improved decel. loss)