Establishing a Scale for Kullback-Leibler Divergence in Language Models Across Various Settings

Ryo Kishino[1], Yusuke Takase[1], Momose Oyama[1,2], Hiroaki Yamagiwa[3], Hidetoshi Shimodaira[1,2]
[1]Kyoto University [2]RIKEN [3]SB Intuitions

arXiv: 2505.15353

Summary

- Representing each language model by its log-likelihood vector embeds models into a common space, where distances reflect KL divergence.
- This work establishes an interpretable absolute scale for KL divergence across various settings.
- We observed that training trajectories are strongly subdiffusive in log-likelihood space.

Log-Likelihood Vector [1]

- A log-likelihood vector represents an LM by its log probabilities over a fixed text set.

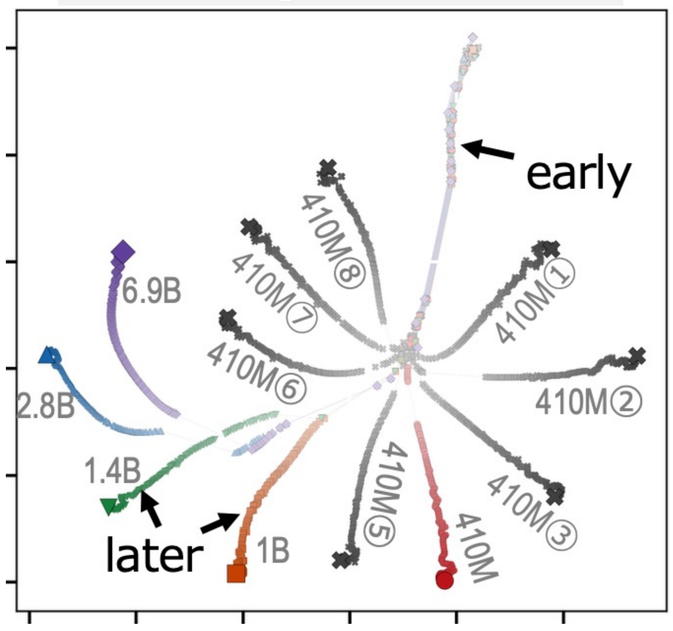
$$\big(\log p_i(x_1), \ldots, \log p_i(x_N)\big) \in \mathbb{R}^N$$

KL Divergence

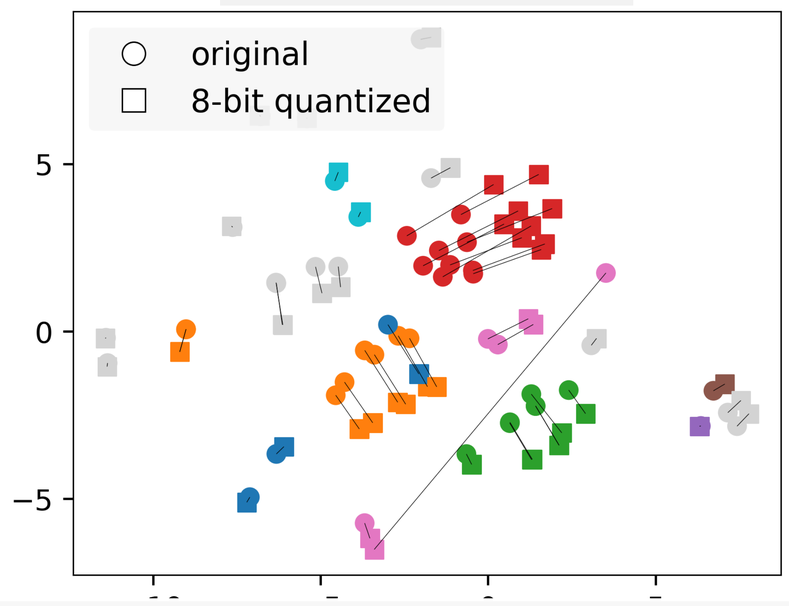
$$\mathrm{KL}(p_i, p_j) \approx \frac{1}{2N} \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$$

Model Map Across Various Settings

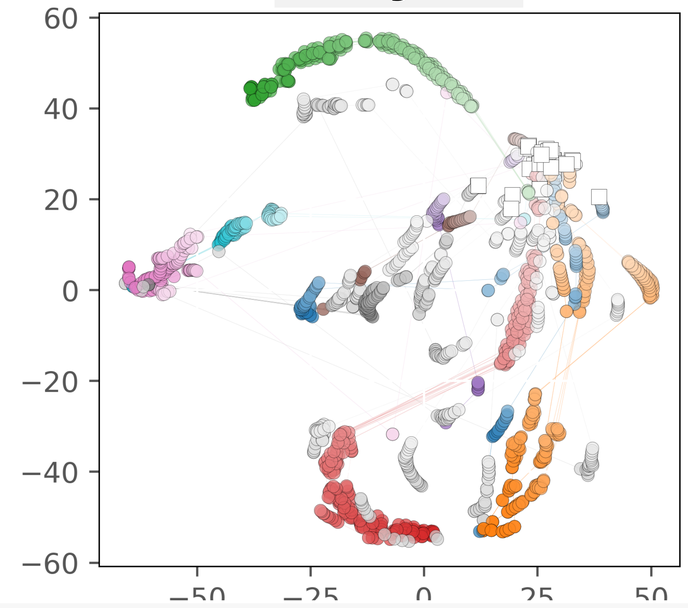
Training Trajectory



Quantization

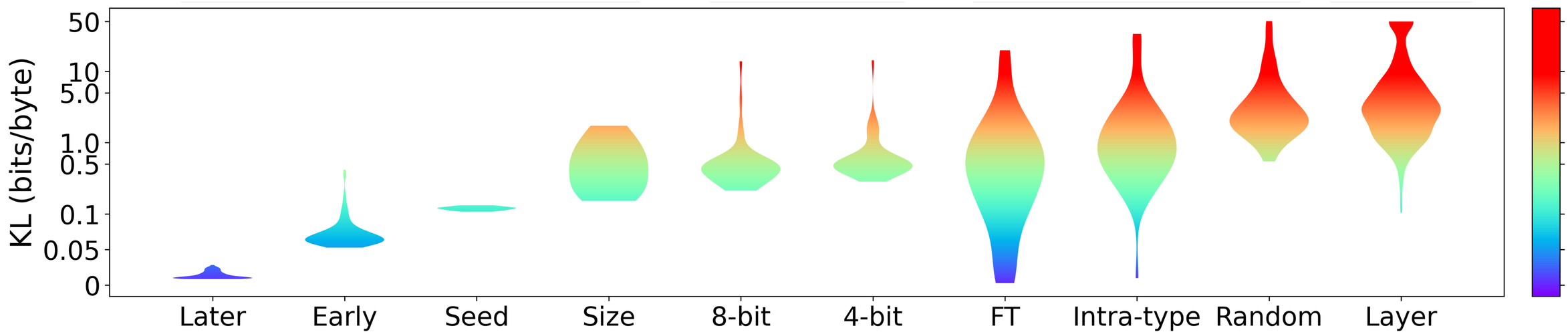


Layer



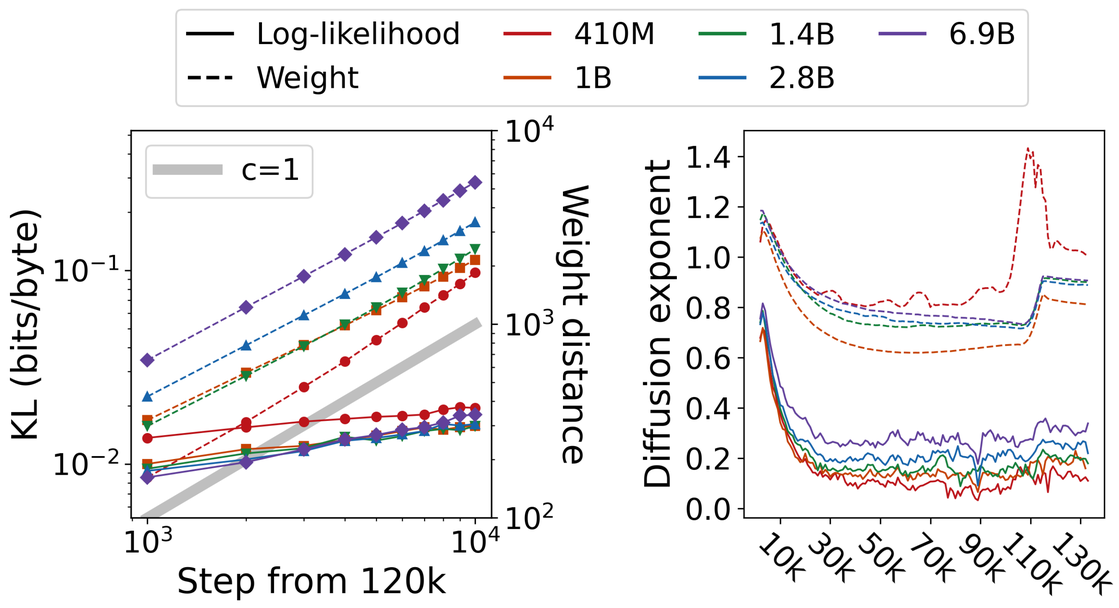
KL Divergence Scale

- KL divergence provides an interpretable scale for comparing model changes across checkpoints, seeds, sizes, quantization, fine-tuning, and layers.



Training Dynamics Analysis

- We analyze pretraining trajectories via KL divergence and find strongly subdiffusive behavior in log-likelihood space despite continued weight drift.



Weight Space vs. Log-Likelihood Space

- Anomalous diffusion:

$$\|\boldsymbol{W}_t - \boldsymbol{W}_{t_0}\|^2 \propto |t - t_0|^{c_\mathrm{w}}, \quad c_\mathrm{w} \approx 1$$
$$\|\boldsymbol{q}_t - \boldsymbol{q}_{t_0}\|^2 \propto |t - t_0|^{c_\mathrm{q}}, \quad c_\mathrm{q} \approx 0.2$$

- Hölder exponent of $f : \boldsymbol{W} \mapsto \boldsymbol{q}(\boldsymbol{W})$ can be calculated by

$$\alpha = c_\mathrm{q}/c_\mathrm{w} \approx 0.2$$

Reference [1] Mapping 1,000+ Language Models via the Log-Likelihood Vector. Momose Oyama, Hiroaki Yamagiwa, Yusuke Takase, Hidetoshi Shimodaira. ACL 2025.