Sparse Autoencoders Need Better Benchmarks

Let's turn SAEs into a "numbers-go-up" science

David Chanin

Decode
MATS
UCL

We want Sparse Autoencoders (SAEs) to decode LLM activations into interpretable features. But to make progress on sparse autoencoders, we need high-fidelity benchmarks that can reliably detect even small improvements in SAE architectures.

We evaluate SAE quality metrics from SAEBench and find no metrics are able to reliably guide architecture decisions.

We need better benchmarks!

How can we tell if an interpretability benchmark is reliable?

We do not have ground-truth access for LLMs, so how can we need to rely on proxy metrics to benchmark SAEs. But how can we tell if those proxy metrics are reliable? A good proxy metric should:

Track ground-truth. The score should be correlate with ground-truth quality if ground truth is available (e.g. on synthetic models).

Increase throughout training. A metric whose score systematically declines through training is broken.

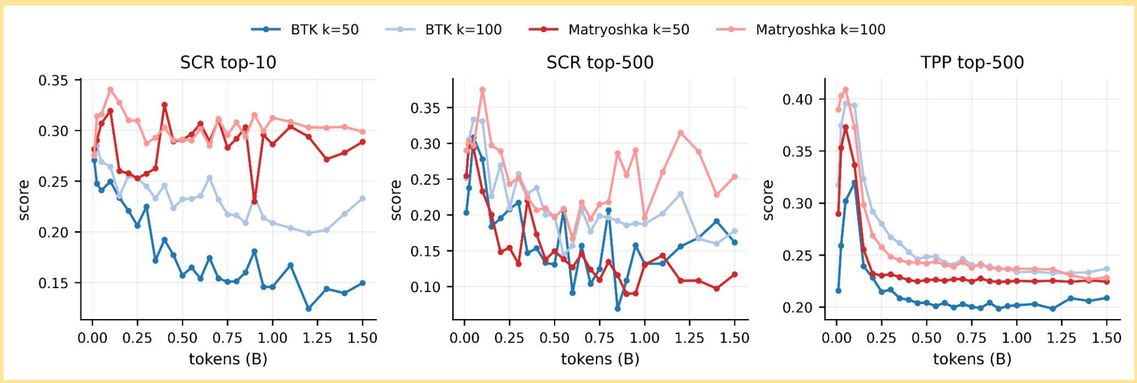
Low reseed noise. The same SAE evaluated on the same data should produce the same score.

Discriminative. The metric should distinguish runs by an amount substantially larger than its own jitter.

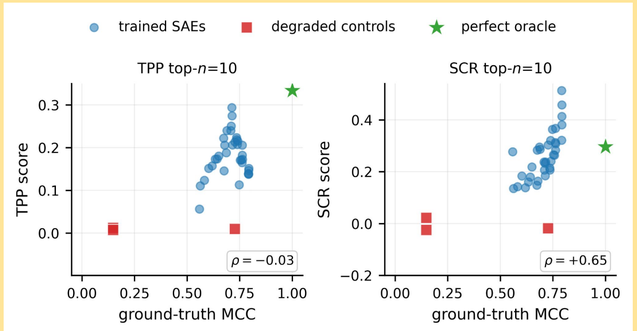
Internally consistent. The metric should rank SAEs the same way regardless of nominally interchangeable hyperparameters.

⚠ Avoid SCR and TPP

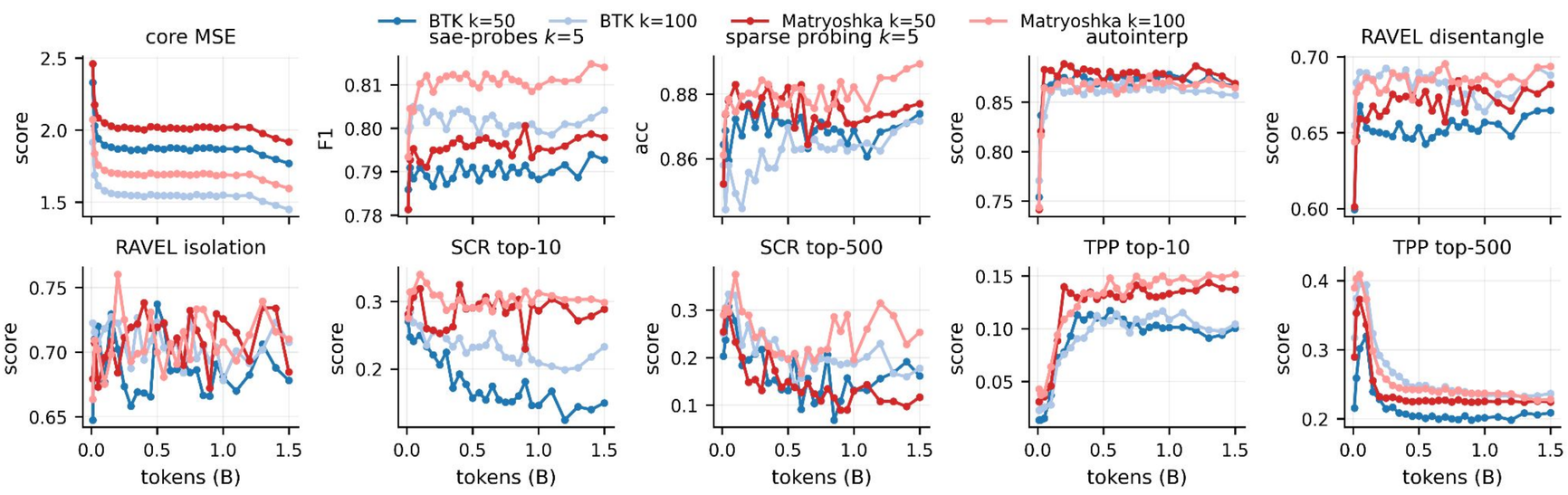
SCR and TPP metrics should not be used to benchmark SAEs



SCR and TPP frequently **decrease throughout training**, and change rankings based on *N*.



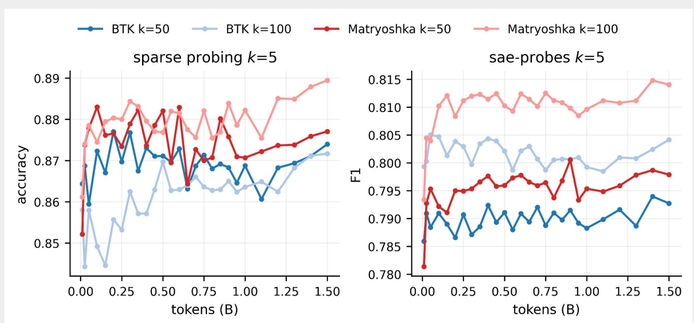
On synthetic models with known ground-truth, TPP frequently does not correlate with ground-truth, and SCR ranks an oracle SAE below imperfect SAEs.



SAEBench metrics evaluated at snapshots in training **four very different SAEs** (k=50 vs k=100, BatchTopK vs Matryoshka). Sae-probes sparse probing is the best metric, but still noisy. x

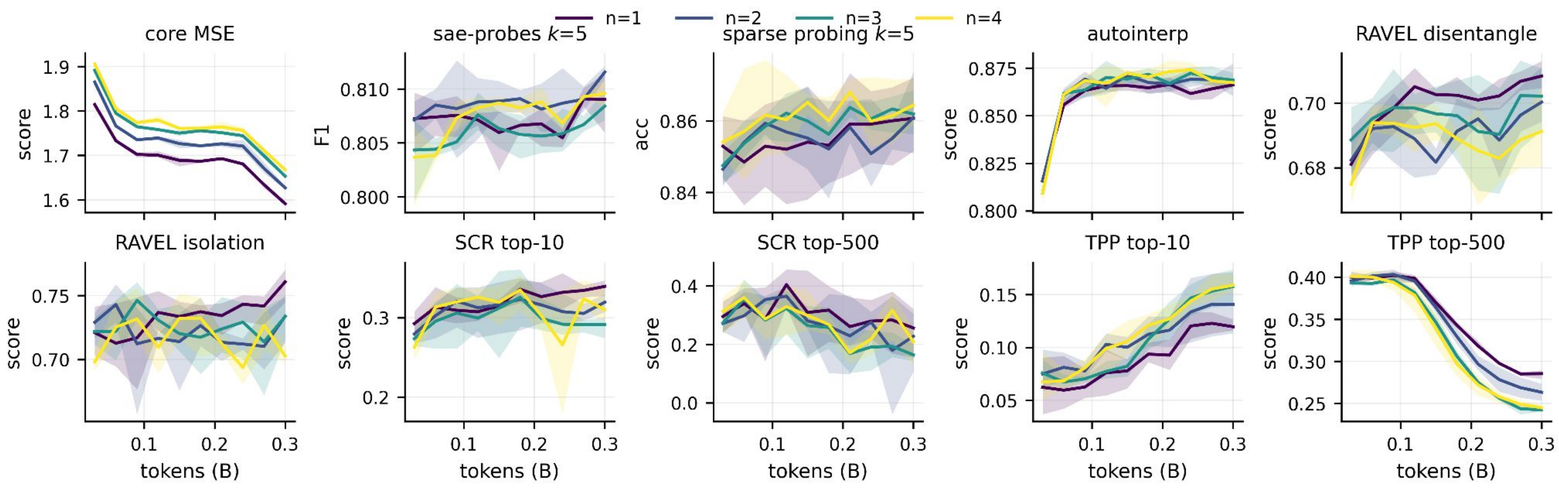
There is hope!

The SAE-probes variant of sparse-probing shows a path to improve benchmarks



SAE-probes improves results by using **more datasets** and **more robust probing**. Let's extend this progress to all metrics!

No metric is able to differentiate similar SAEs



SAEBench metrics evaluated at snapshots in training **four similar SAEs** (k=100 Matryoshka, vary number of Matryoshka prefixes, *n*), 3 seeds per SAE. *No SAEBench metric can reliably differentiate these SAEs.* To make progress, the field needs benchmarks that can detect small improvements.

Check out the Paper

Are Sparse Autoencoder Benchmarks Reliable?

https://arxiv.org/abs/2605.18229