Steering LLMs with Sparse Autoencoders

A Path Towards More Explainable and Safer AI

Text+ Plenary 2024

Lalith Manjunath

OpenGPT-X

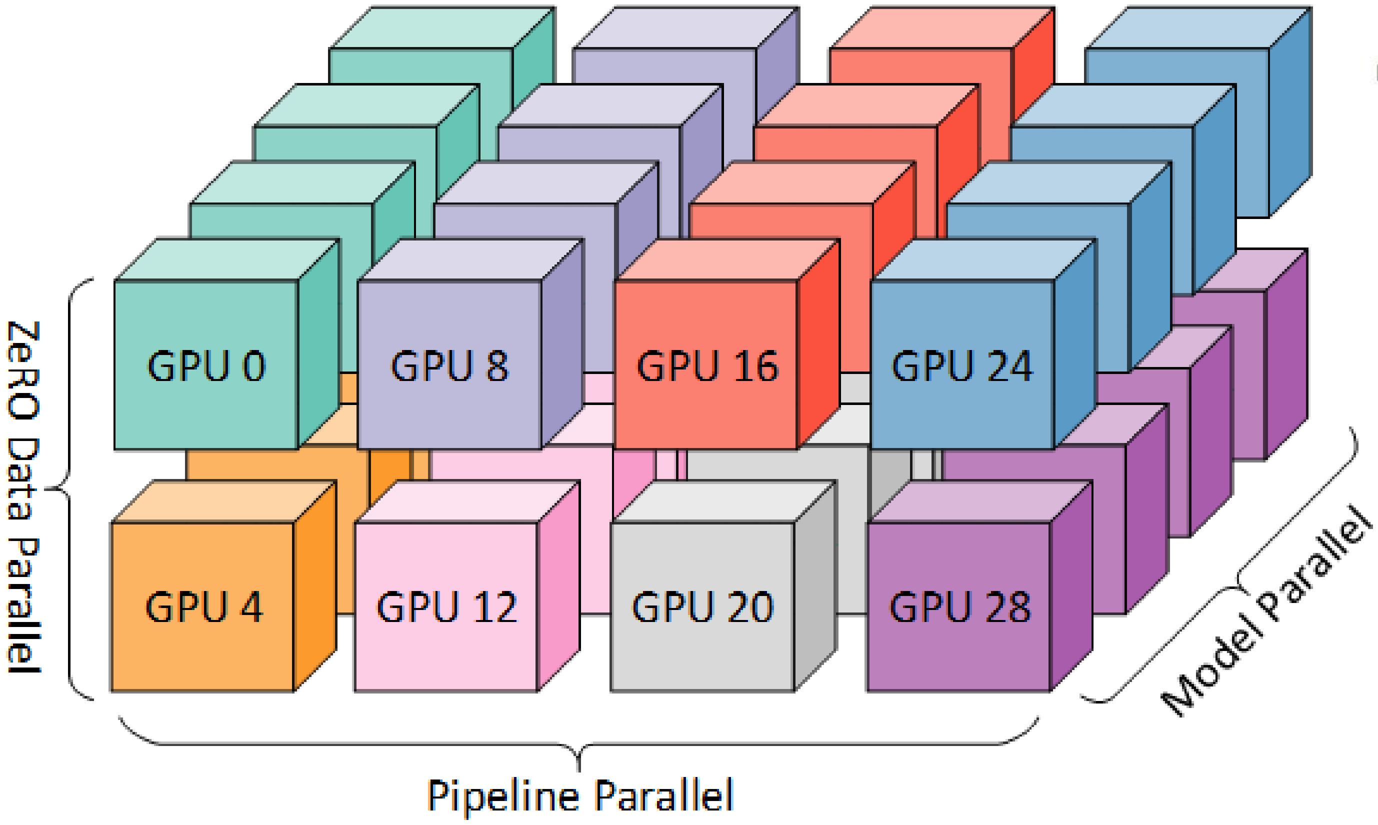
OpenGPT-X builds and trains large-scale AI language models to drive innovative language application services for the European economy

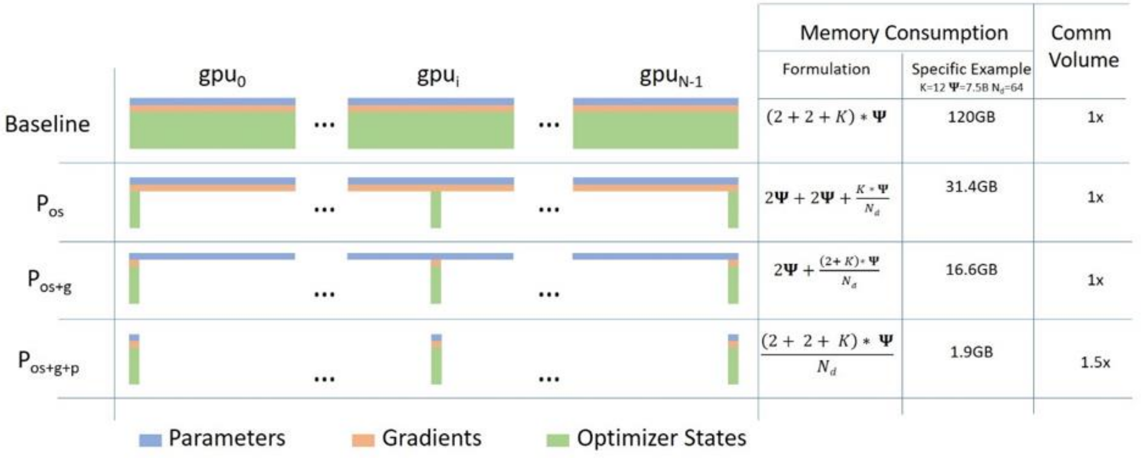
Partners

Eleven partners from business, science and the media industry are working together on OpenGPT-X. Each partner contributes its expertise to the project. All partners introduce themselves here:

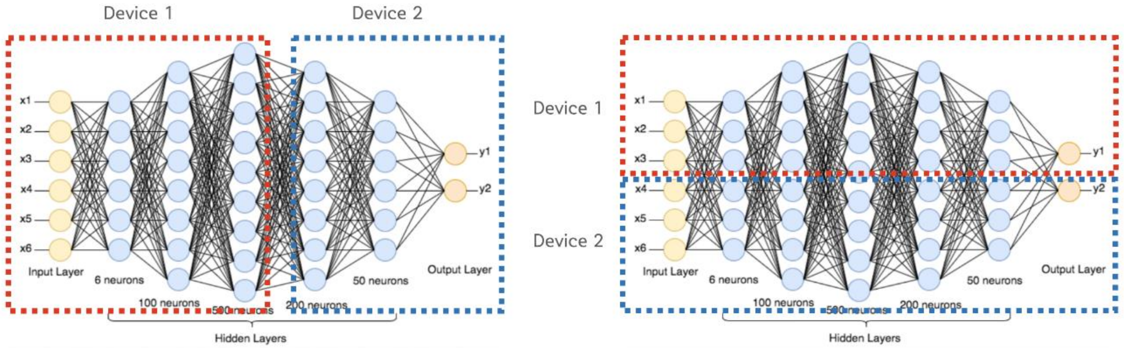
- Akademie für Künstliche Intelligenz AKI gGmbH im KI Bundesverband
- Aleph Alpha Gmbh
- ControlExpert GmbH
- Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)
- Fraunhofer-Institut für Integrierte Schaltungen IIS
- Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
- IONOS SE
- Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH
- Westdeutscher Rundfunk – Anstalt des öffentlichen Rechts
- Zentrum für Informationsdienste und Hochleistungsrechnen (ZIH)
- [at] Alexander Thamm Gmbh

Parallelism to accelerate training of LLMs

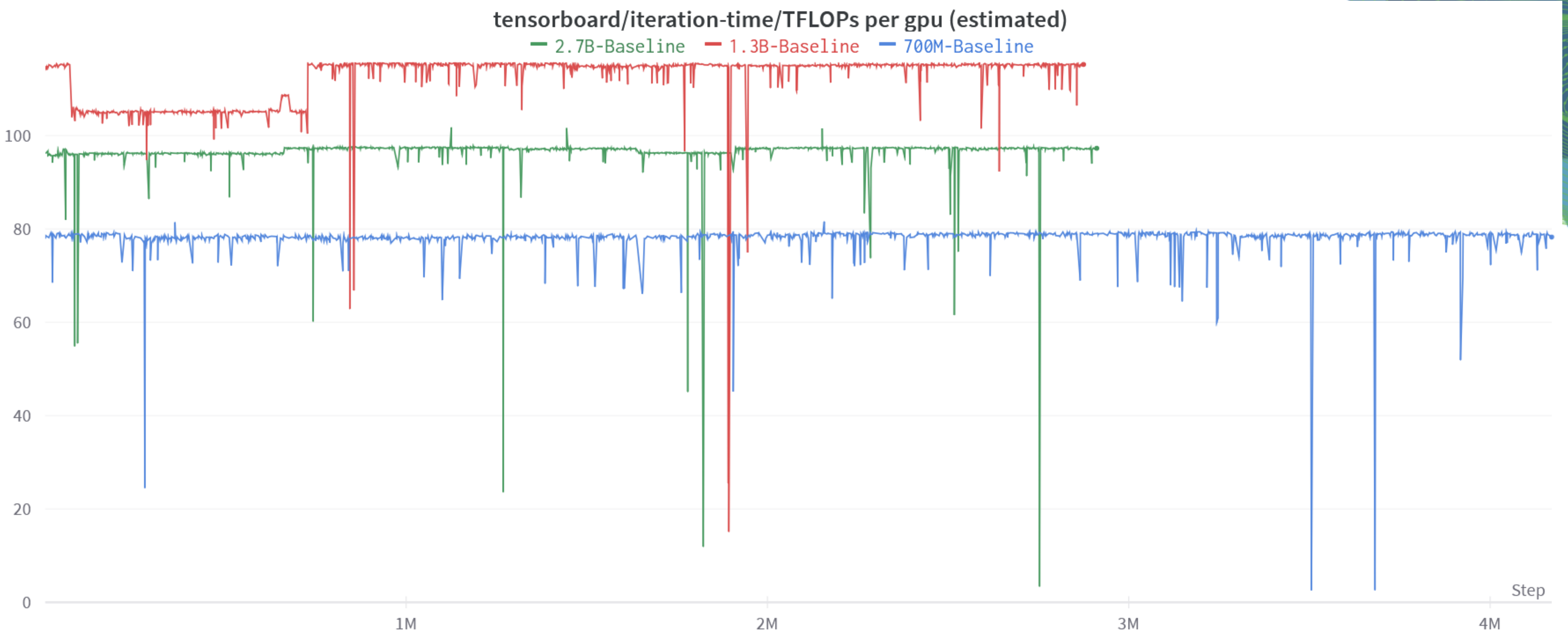




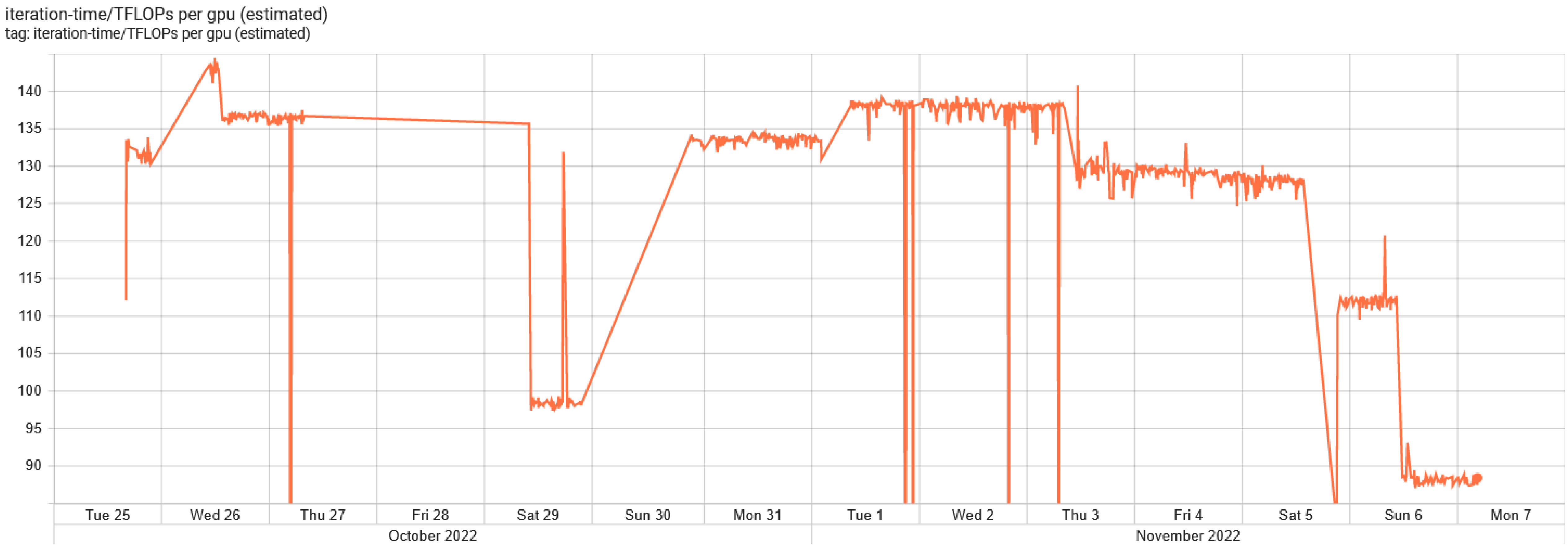
	Memory Consumption Formulation	Memory Consumption Specific Example K=12 Ψ=7.5B N_d=64	Comm Volume
Baseline	$(2+2+K)*\Psi$	120GB	1x
P_{os}	$2\Psi + 2\Psi + \frac{K*\Psi}{N_d}$	31.4GB	1x
P_{os+g}	$2\Psi + \frac{(2+K)*\Psi}{N_d}$	16.6GB	1x
P_{os+g+p}	$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB	1.5x



GPU Throughput from Small Scale Models



6.7B Decoder-only 3D Parallel Training on German Corpus



Loss Spikes and Tokenizer’s learned vocabulary


```
"\",":2688
"\\\\\\\\\\\\\\\\":4180
"-/":5242
"77":5380
"âĢ¦]":5613
"'t":5683
"ĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀ":5682
"Ġ[âĢ¦]":6225
"{{":6555
"----------------":6937
"\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\":6982
"________________":7807
"........":7862
"!!!":7962
"ĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀĀ":9391
"...]":9615
"ĠMÃ¼nster":10865
"\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\":11440
"Ġ#####":11990
"âĢ¦âĢ¦":12038
"ĠKlÃ¤ge":12063
"********":13435
"Ġ----":13777
"WhatsApp":13936
"ĠĠĠĠĠĠĠĠ":14265
"====":14546
",...":14860
```

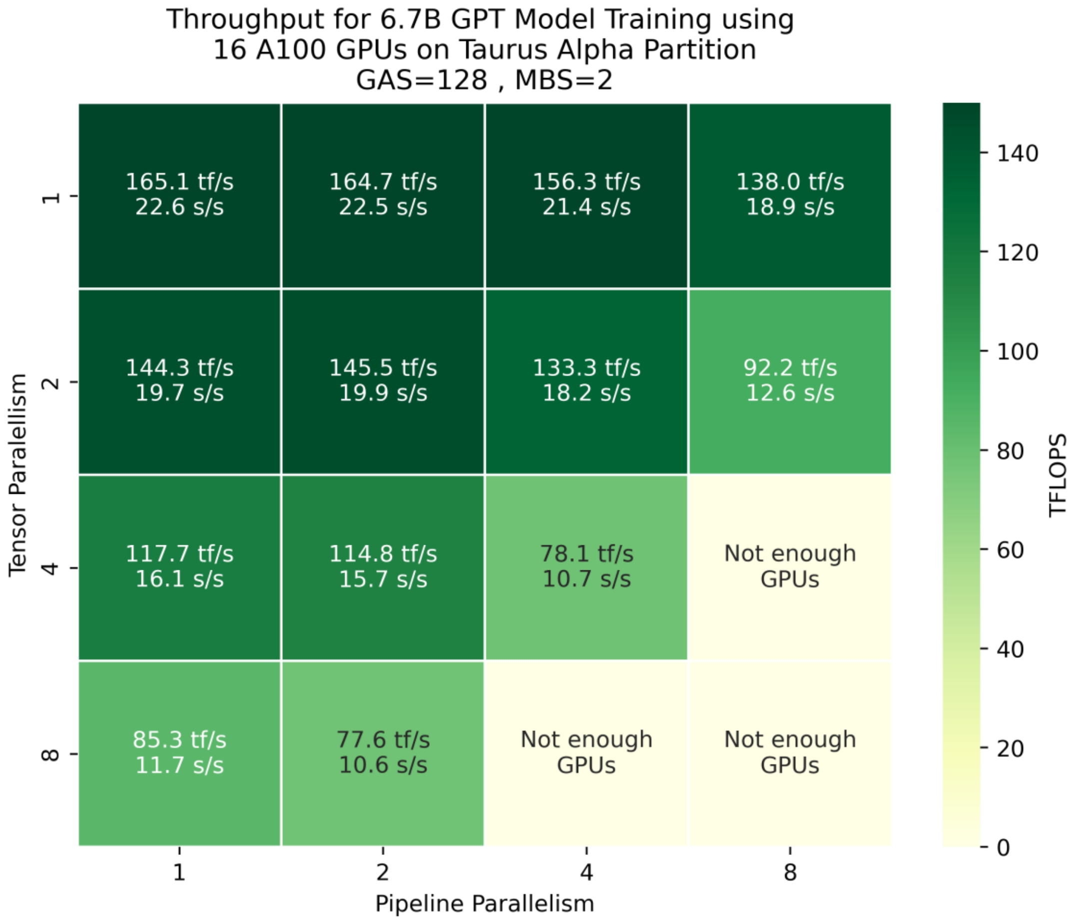

```
    12.6                          \>168
5      11.45          7.9                                  5.0                 3.1                              6.6                          \>168
6      22.90          8.35                                 2.3                 0.3                              0.4                          \>168

###### Properties of PSA and Diffusion Coefficients of Drugs by Different Test Methods

  sample   cross-linker (%)   cross-link density (10^--4^ mol/cm^3^)   temperature (K)   *D*-FT-IR (cm^2^/s)   *D*-MD (cm^2^/s)   *D*-polymer (cm^2^/s)   FFV (%)
  -------- ------------------ ---------------------------------------- ----------------- --------------------- ------------------ ----------------------- ---------
  1        0                  7.39                                     298               8.46 x 10^--8^        1.17 x 10^--7^     4.33 x 10^--8^          15.28
  2        1.43               7.42                                     298               6.83 x 10^--8^        9.22 x 10^--8^     3.34 x 10^--8^          15.92
  3        2.86               7.51                                     298               3.17 x 10^--8^        6.25 x 10^--8^     2.98 x 10^--8^          16.82
  4        5.72               7.63                                     298               2.64 x 10^--8^        4.85 x 10^--8^     2.61 x 10^--8^          16.62
  5        11.45              7.9                                      298               7.66 x 10^--9^        2.76 x 10^--8^     2.25 x 10^--8^          16.49
  6        22.90              8.35                                     298               1.25 x 10^--9^        4.31 x 10^--9^     1.82 x 10^--8^          16.10
  303      0                  7.39                                     303               9.35 x 10^--8^        1.28 x 10^--7^     6.72 x 10^--8^          16.64
  313      0                  7.39                                     313               1.04 x 10^--7^        1.54 x 10^--7^     7.25 x 10^--8^          17.41
  323      0                  7.39                                     323               1.67 x 10^--7^        2.41 x 10^--7^     1.312 x 10^--7^         18.53
  333      0                  7.39                                     333               4.16 x 10^--7^        6.03 x 10^--7^     3.23 x 10^--7^          19.83
  343      0                  7.39                                     343               7.59 x 10^--7^        9.42 x 10^--7^     4.51 x 10^--7^          20.52

The cross-link density of the samples was characterized with ^1^H NMR transverse relaxation parameters. The higher the cross-link density of the polymer, the faster the proton
transverse relaxation curve decayed. As shown in [Figure [3](#fig3){ref-type="fig"}](#fig3){ref-type="fig"}, samples 1--
```

Tokenizer Choice For LLM Training: Negligible or Crucial?

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, Nicolas Flores-Herr

The recent success of Large Language Models (LLMs) has been predominantly driven by curating the training dataset composition, scaling of model architectures and dataset sizes and advancements in pretraining objectives, leaving tokenizer influence as a blind spot. Shedding light on this underexplored area, we conduct a comprehensive study on the influence of tokenizer choice on LLM downstream performance by training 24 mono- and multilingual LLMs at a 2.6B parameter scale, ablating different tokenizer algorithms and parameterizations. Our studies highlight that the tokenizer choice can significantly impact the model's downstream performance and training costs. In particular, we find that the common tokenizer evaluation metrics fertility and parity are not always predictive of model downstream performance, rendering these metrics a questionable proxy for the model's downstream performance. Furthermore, we show that multilingual tokenizers trained on the five most frequent European languages require vocabulary size increases of factor three in comparison to English. While English-centric tokenizers have been applied to the training of multi-lingual LLMs in the past, we find that this approach results in a severe downstream performance degradation and additional training costs of up to 68%, due to an inefficient tokenization vocabulary.

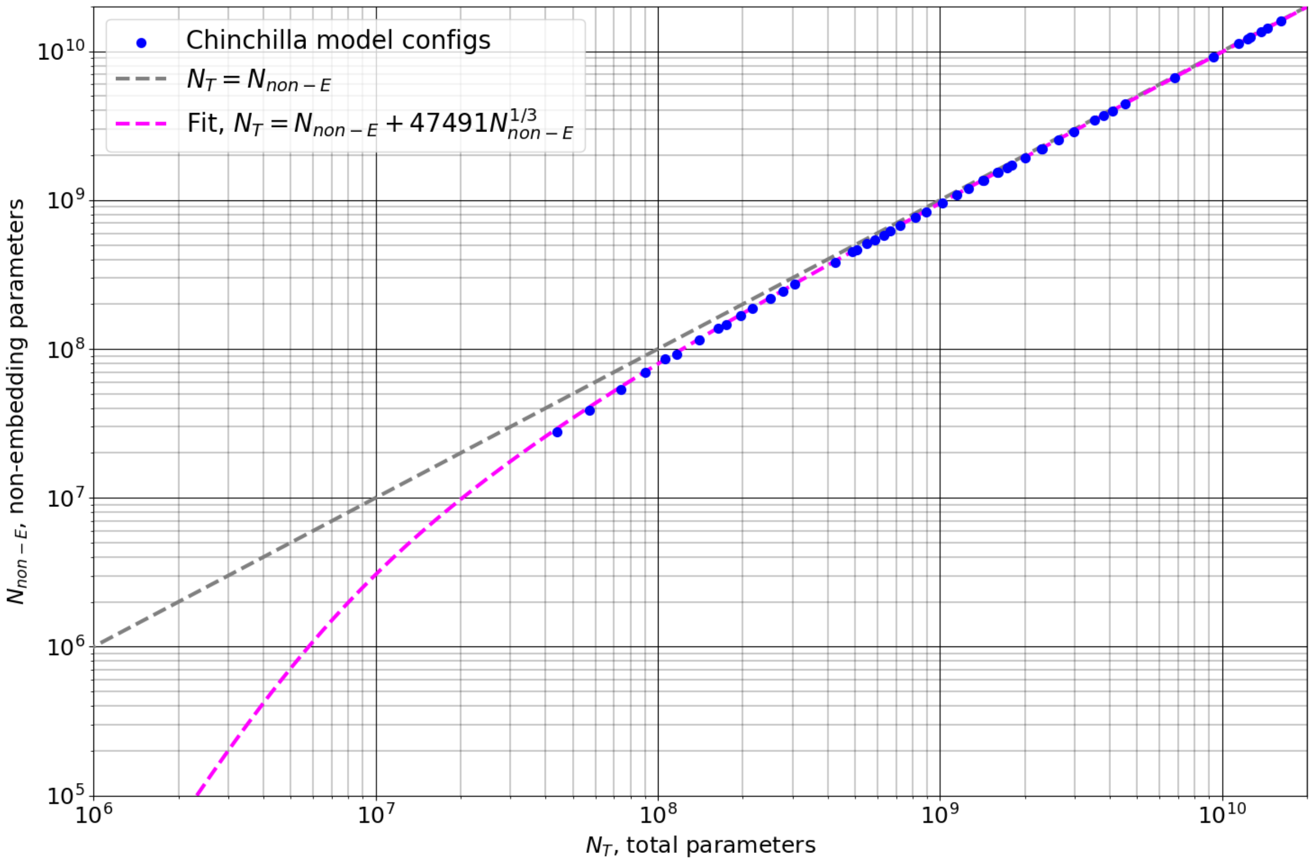
Parallelism Analysis



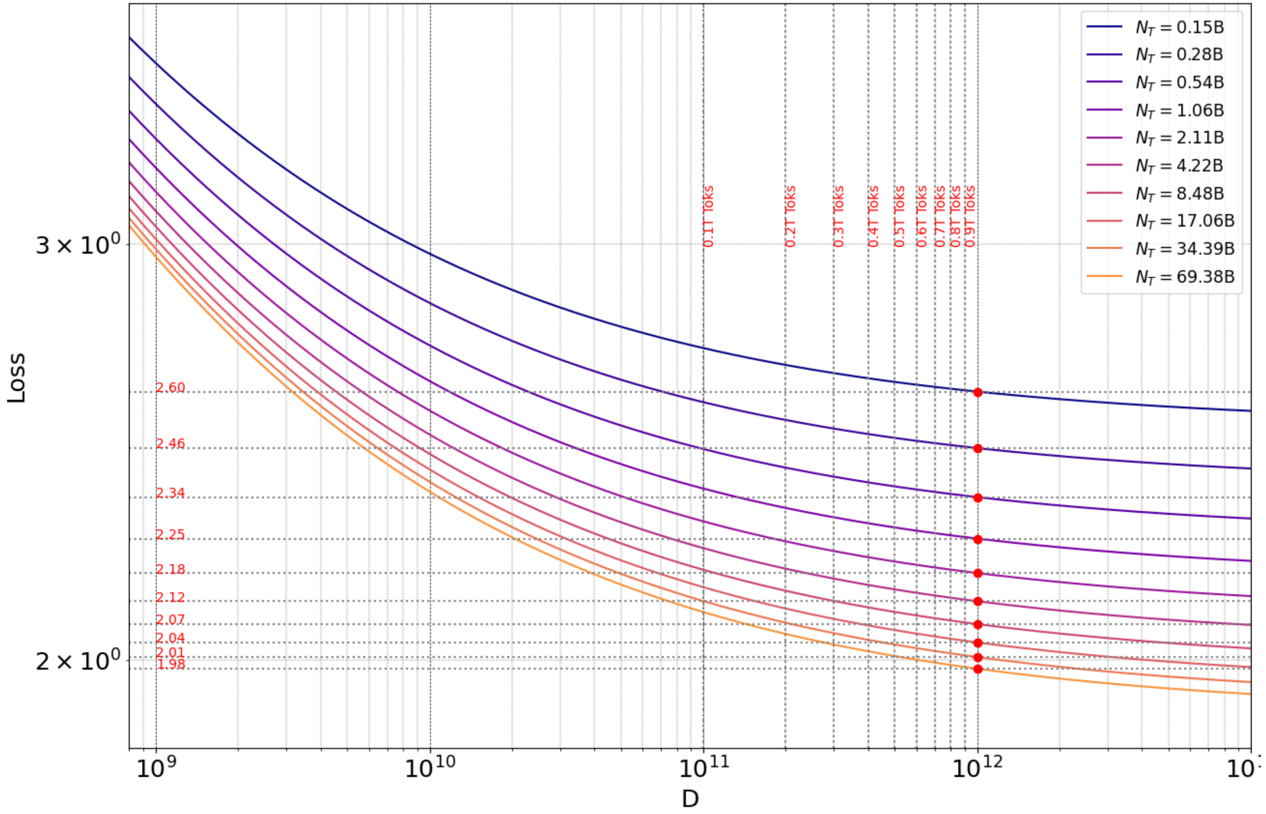
- Credits: Lena Jurkschat

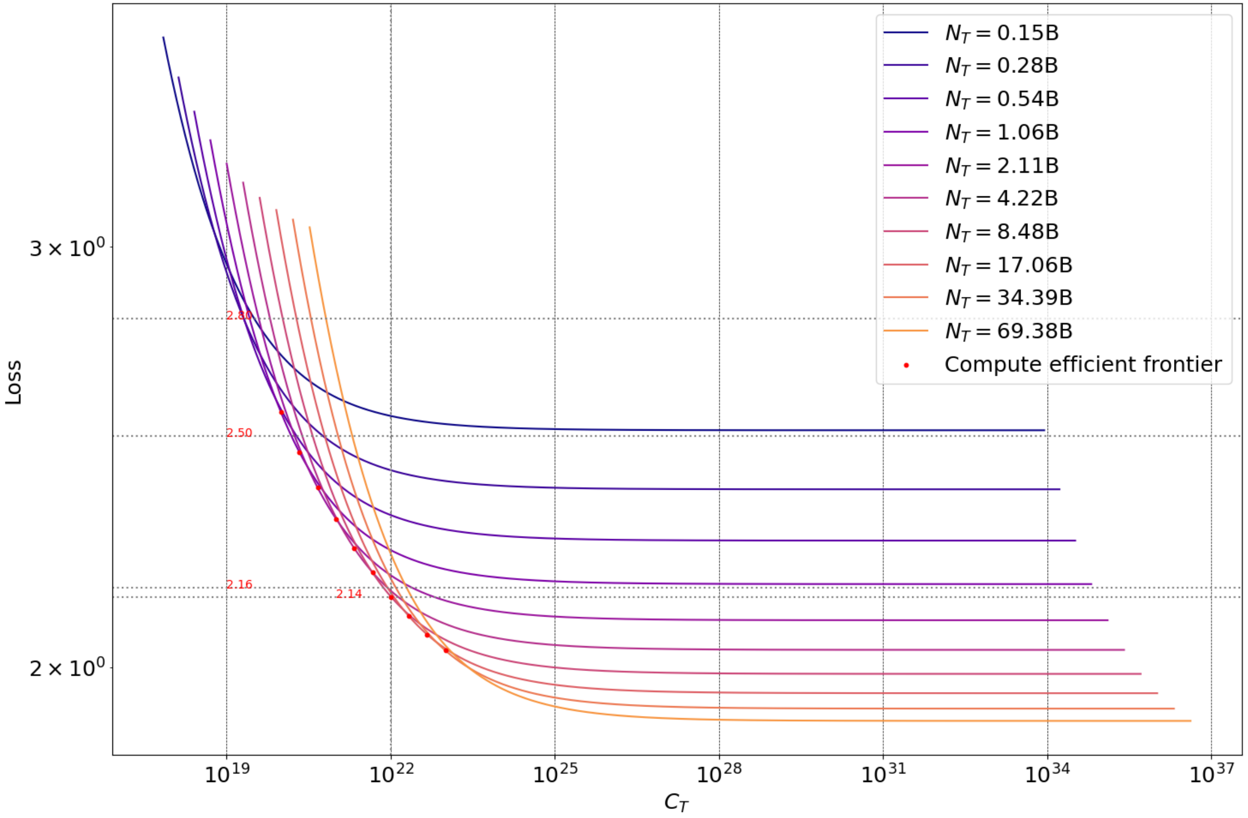
Scaling Laws Investigation

- Scaling laws help predict the trained model's behavior.
- Kaplan laws > Chinchilla laws > Reconciled laws (Takes into account the embedding parameters)
- Case: Compute Budget 46K A100 GPU Hours and 250-300B tokens of data.

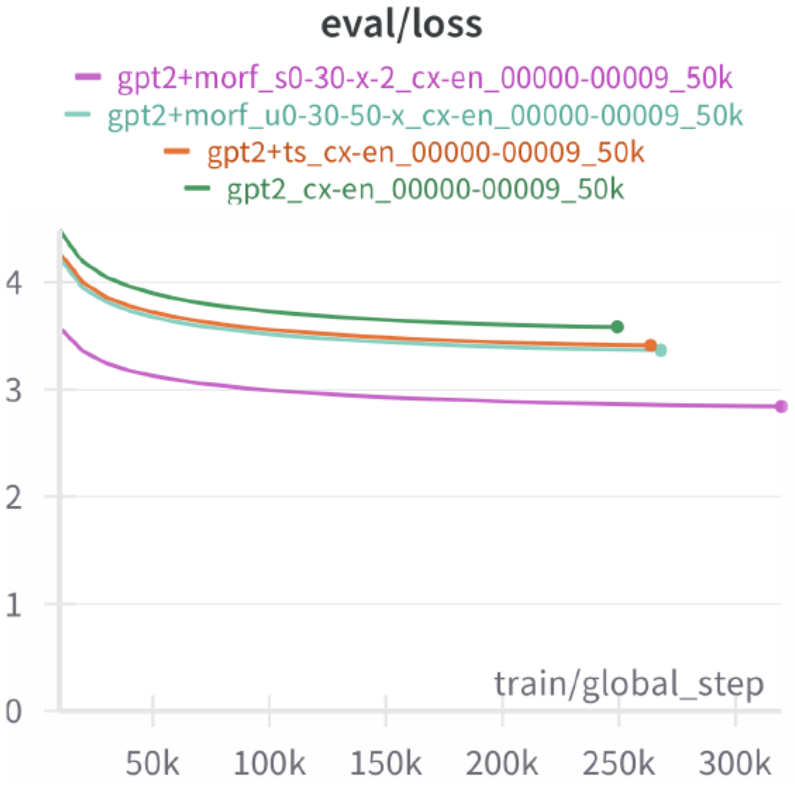


Scaling Laws Investigation

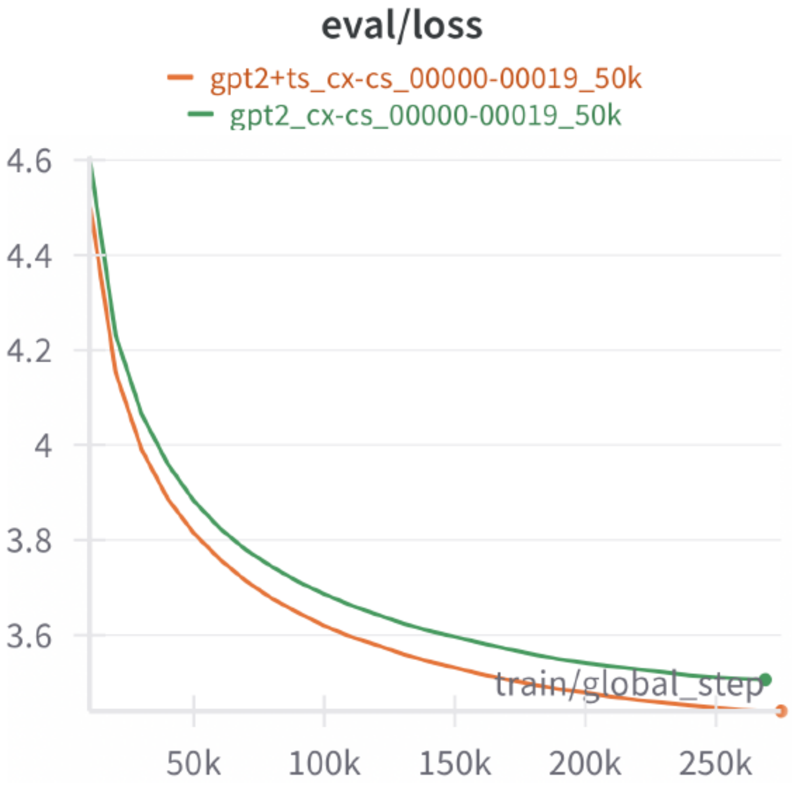




Morphologically Biased BPE Vocabulary



English Decoder-only Model

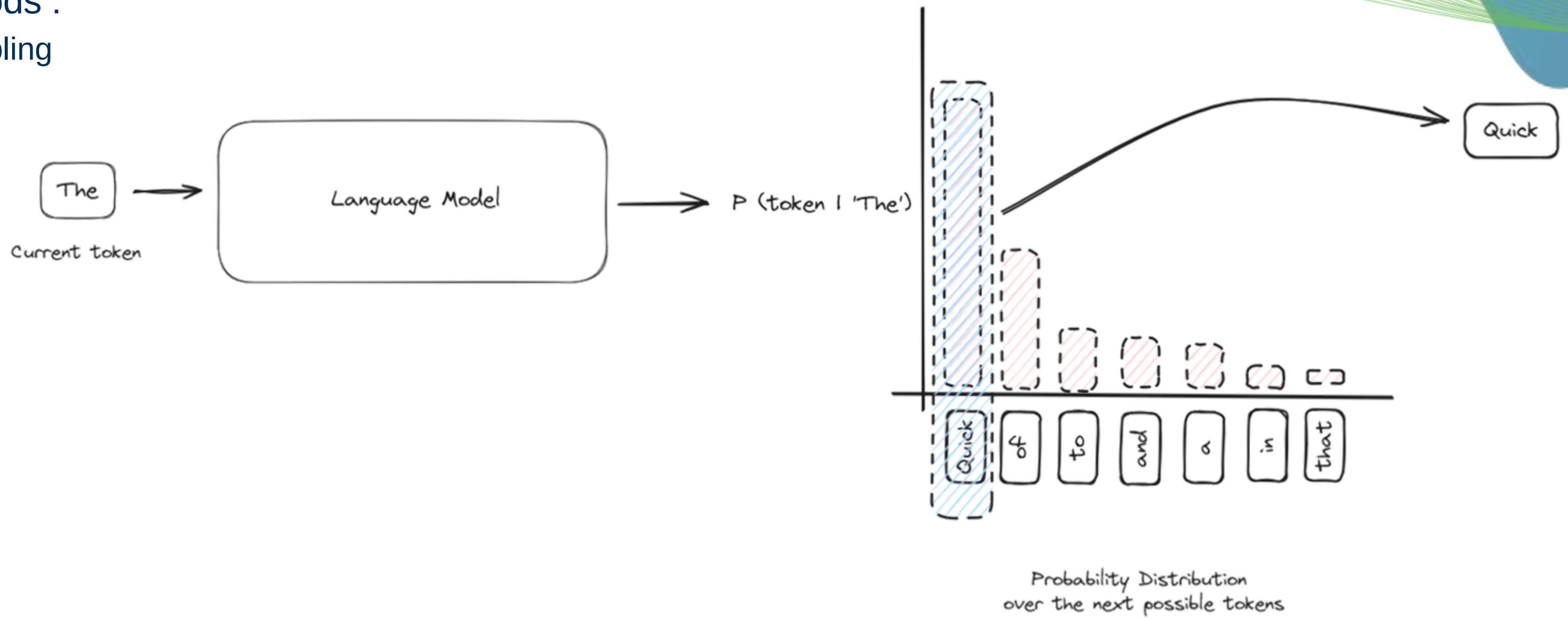


Czech Decoder-only Model

- Credits: Jonas Knobloch

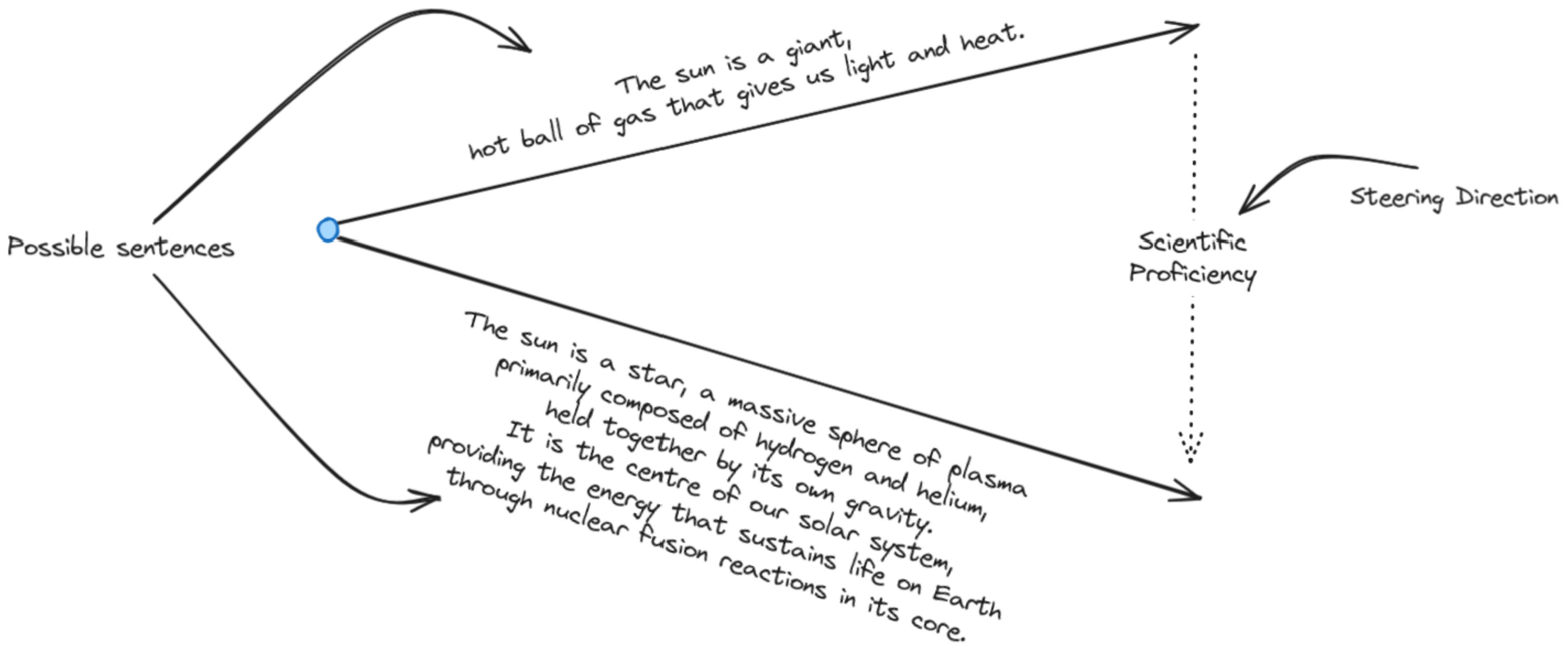
How LLM inference works

- LLMs do not output a token directly.
- They output a probability distribution over all the tokens and we use a sampling method to decide which token to use.
- Most common methods :
  - Temperature Sampling
  - Nucleus Sampling



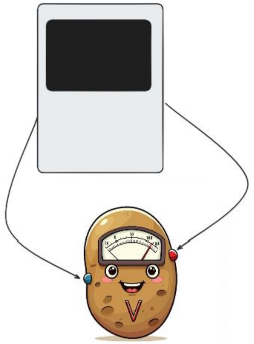
What does Steering mean?

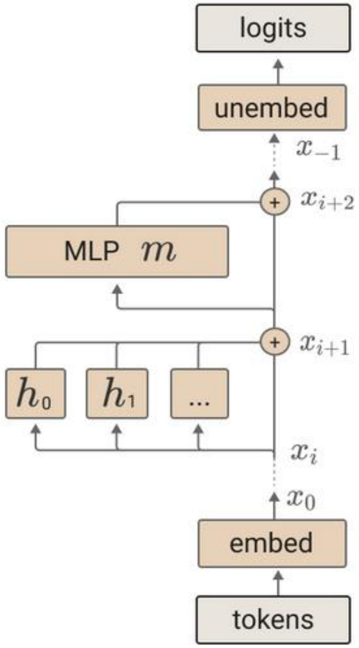
Using a method to modify the probability distribution of the token being predicted to avoid/favor particular tokens.

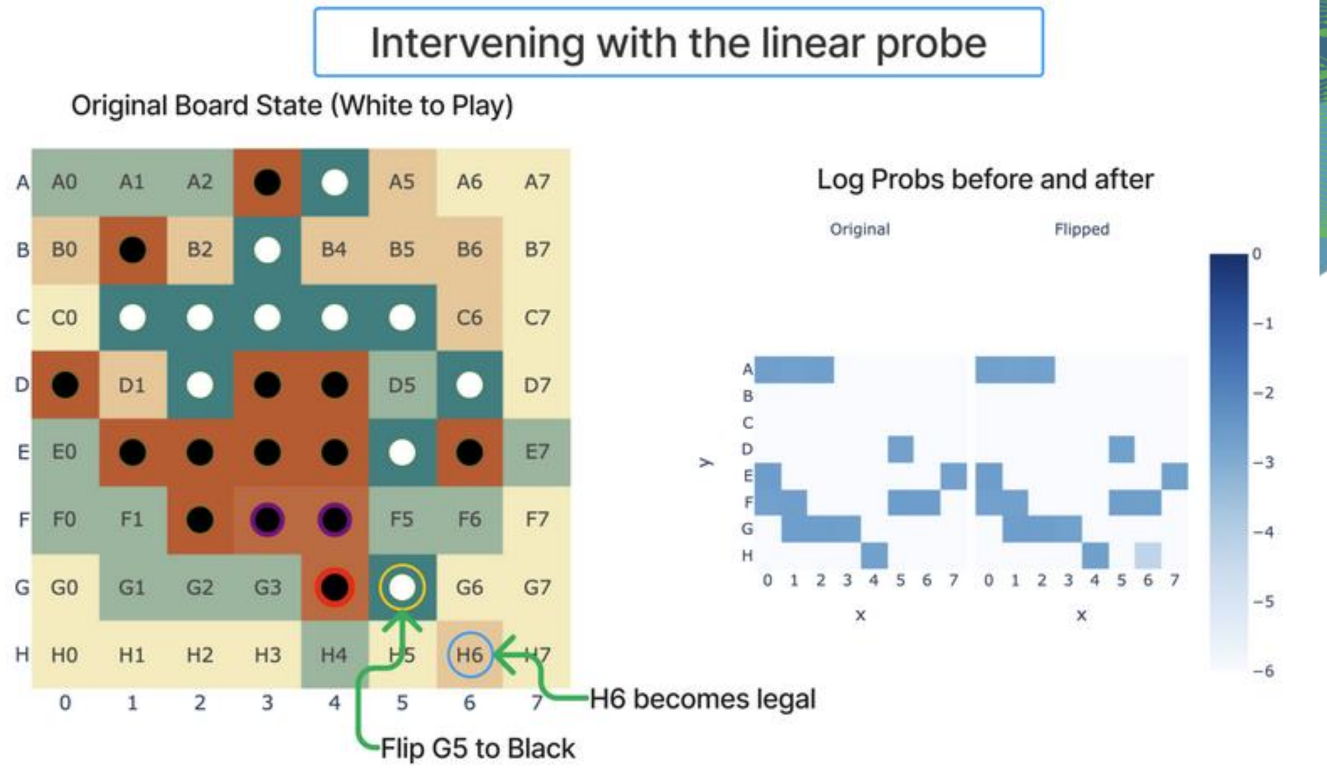


Probing

- Talk and probe in model’s own language, i.e, tensors.



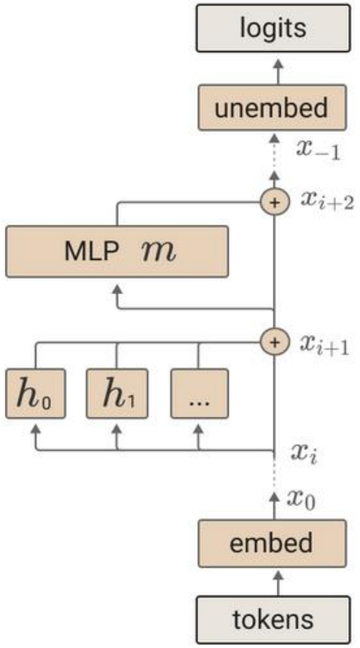


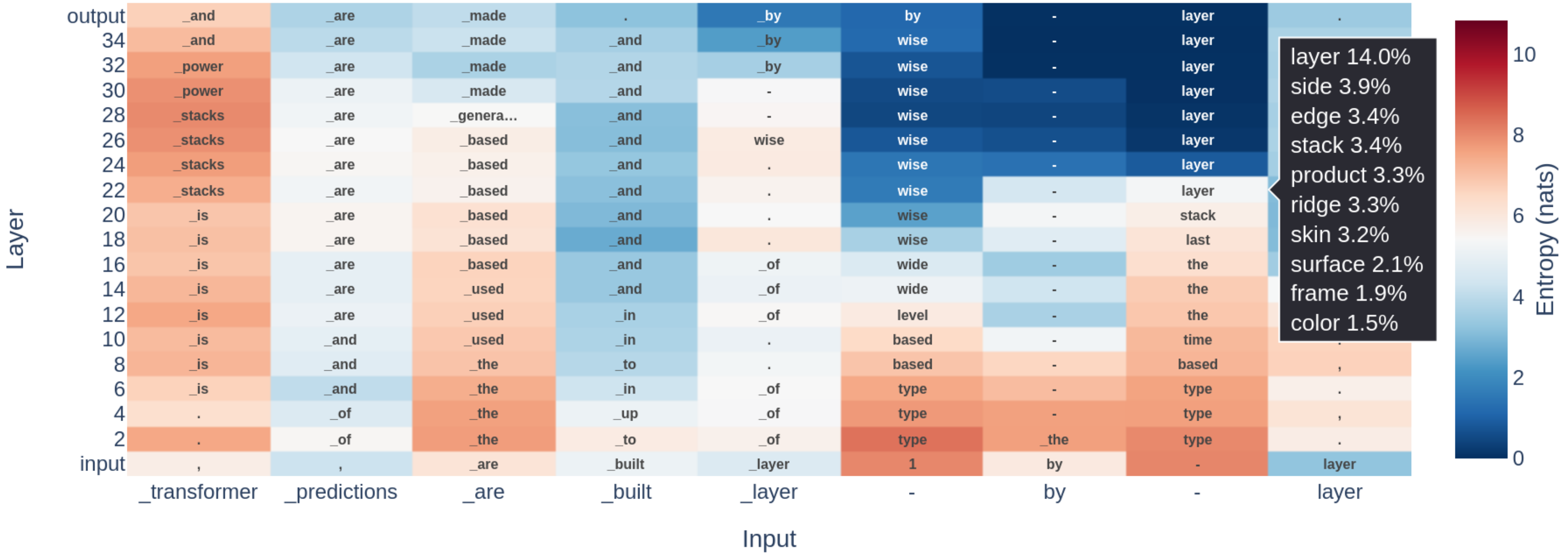


https://www.lesswrong.com/posts/nmxzr2zsjNtjaHh7x/actually-othello-gpt-has-a-linear-emergent-world

Probing

- Talk and probe in model’s own language, i.e, tensors.

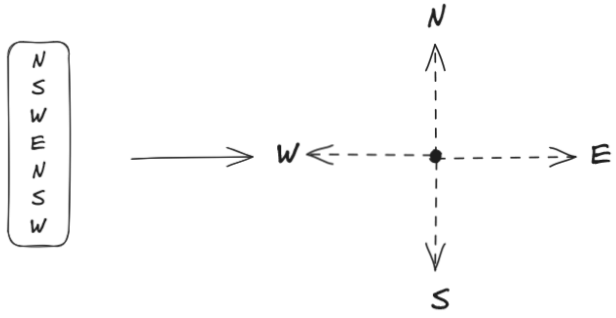


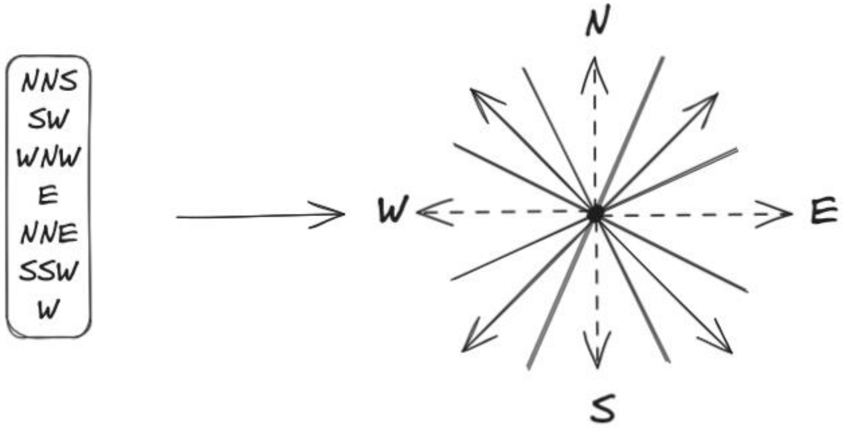


https://www.lesswrong.com/posts/nmxzr2zsjNtjaHh7x/actually-othello-gpt-has-a-linear-emergent-world

What is Superposition?

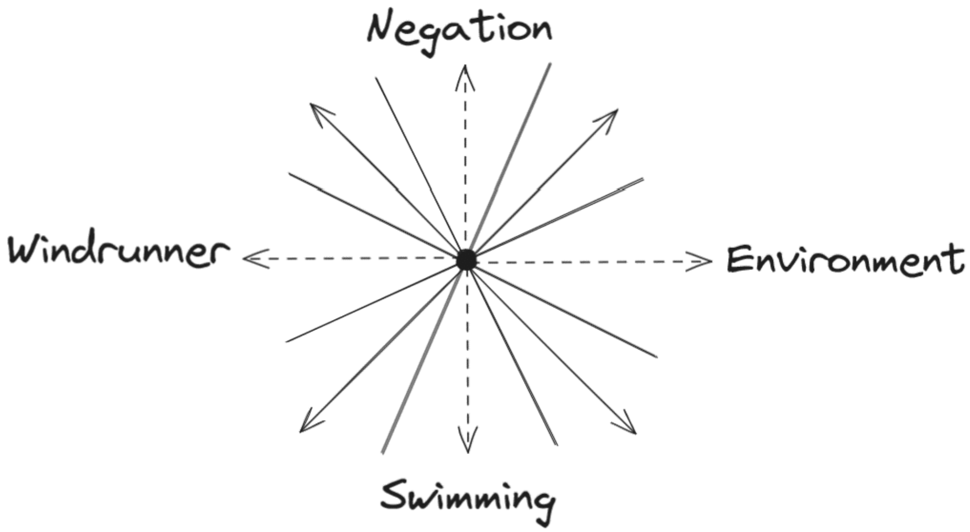
- Compressing more information than you have dimensions / directions.
- Form of lossy compression. Concepts could exist over multiple directions.





Why naïve Steering does not generalize?

- LLMs unfortunately use superposition.
- Linear Probes cannot separate out the concepts in the directions reliably.

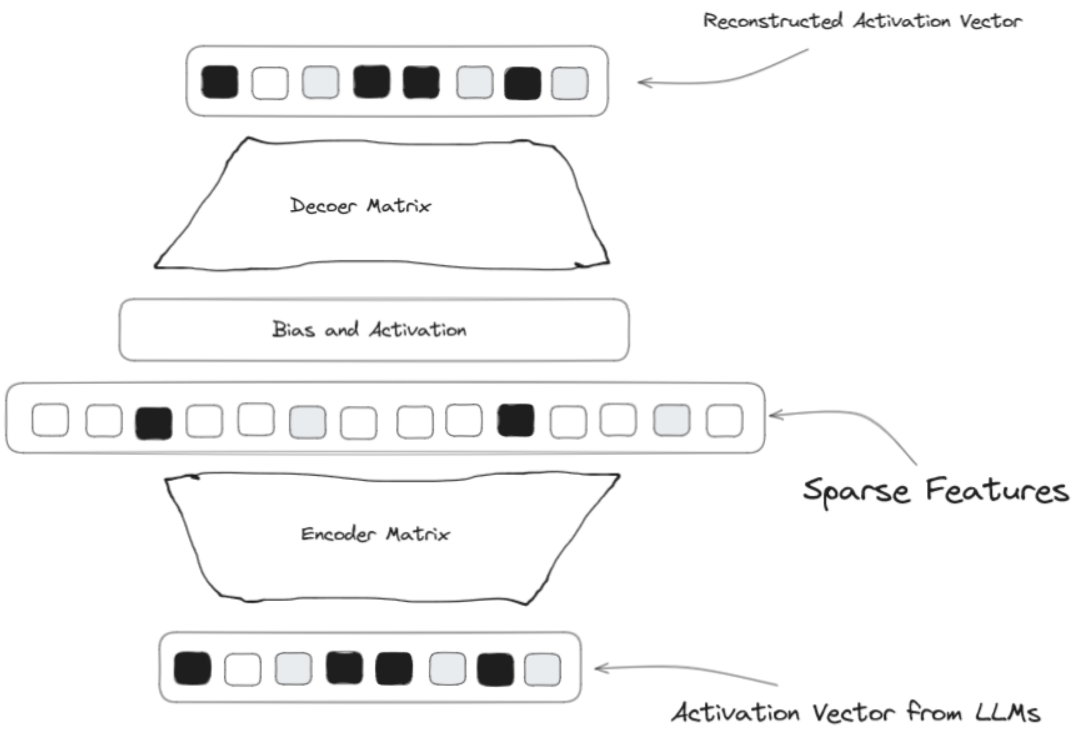


Sparse Autoencoders (SAEs)

- Autoencoders:

$$f(x) := \sigma\left(W_{enc}x + b_{enc}\right),$$
$$\hat{x}(f) := W_{dec}f + b_{dec}.$$

- Since we train the weights to encode and decode the input from the latent state, it becomes an autoencoder.
- The dimension of latent is much larger than the dimension of the inputs.
- SAEs typically shallow and wide.
- Add Sparsity Loss to encourage sparsity and we have SAEs



Reference: *Cunningham, Hoagy et al. "Sparse Autoencoders Find Highly Interpretable Features in Language Models." ArXiv abs/2309.08600 (2023).*

Implications of SAEs

- SAEs give a promising direction to focus on to make the models more safer.
- Imagine finding the directions responsible for the model being deceptive, lying, etc.. And then suppressing those particular directions.

Challenges of SAEs

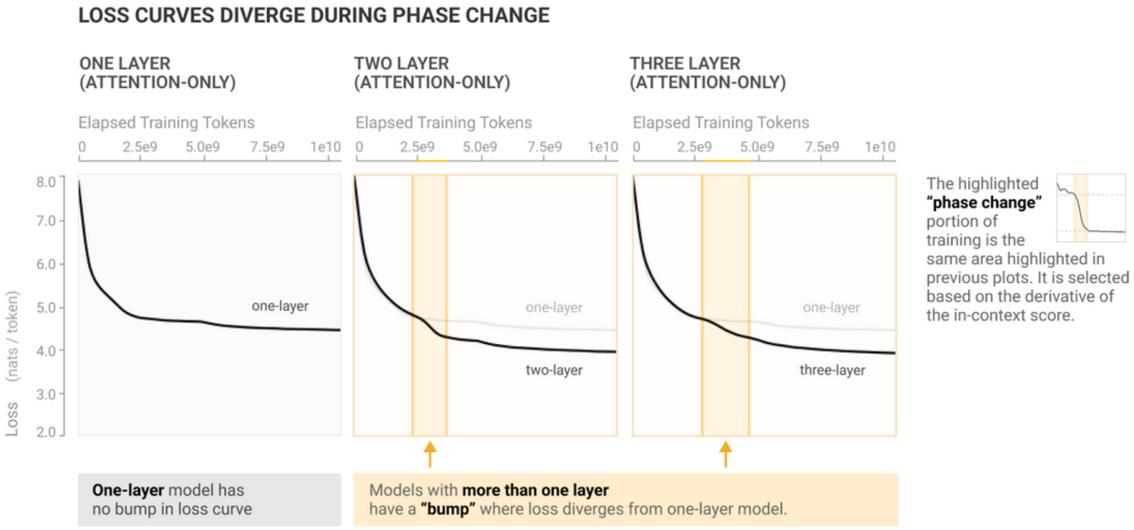
- SAEs are shallow but wide need tensor parallelism over pipeline parallelism.
- Not training on text/image data directly but on the internal activations of a LLM.
- Roughly, 100TB of disk space needed to store activations of a 9B scale model at single site and single layer.

Impact of SAEs

- Prior work on understanding internals and model (< 1 B params) steering relied on the assumption that the directions are decomposable.
- Golden Gate Claude (Claude Sonnet with SAE) demonstrates feasibility for LLMs.
- GemmaScope

Open for Questions

- How does the many post training methods such as finetuning, context length extension change the model weights?
- What is the algorithm that is learnt to solve n-digit addition ?
- Detecting / Fixing Jailbreaks to models.
- What happens when the model is induced to perform chain of thought?



Reference: *Lieberum, Tom et al. "Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2." ArXiv abs/2408.05147 (2024).*
https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html