ReFoRM Reading Group

Rethinking Foundations Real-World ML

Anay Mehrotra, Amin Saberi, Grigoris Velegkas

Welcome to ReFoRM!

What is this reading group about? Foundations of *"real-world"* ML

Welcome to ReFoRM!

What is this reading group about? Foundations of *"real-world"* ML

How is "real-world" ML different from "idealized" ML?

Idealized picture: $$\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}[\ell(z_i; \theta)]$$

Welcome to ReFoRM!

What is this reading group about? Foundations of *"real-world"* ML

How is "real-world" ML different from "idealized" ML?

Idealized picture: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}[\ell(z_i; \theta)]$

Decisions:

- How to choose the parameter space to avoid *overfitting*?
- What (convex) *loss function* ℓ to choose?
- Which *optimization algorithm* to use?

Welcome to ReFoRM!

What is this reading group about? Foundations of *"real-world"* ML

How is "real-world" ML different from "idealized" ML?

Idealized picture: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}[\ell(z_i; \theta)]$

Decisions:

- How to choose the parameter space to avoid *overfitting*?
- What (convex) *loss function* ℓ to choose?
- Which *optimization algorithm* to use?

Guarantees: Convergence rates, generalization bounds, uncertainty quantification (via confidence intervals), performance on different distributions,...

Welcome to ReFoRM!

What is this reading group about? Foundations of "real-world" ML

How is "real-world" ML different from "idealized" ML?

Real-world ML: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}\left[\ell(z_i; \theta)\right]$

Messy Dataset D

+

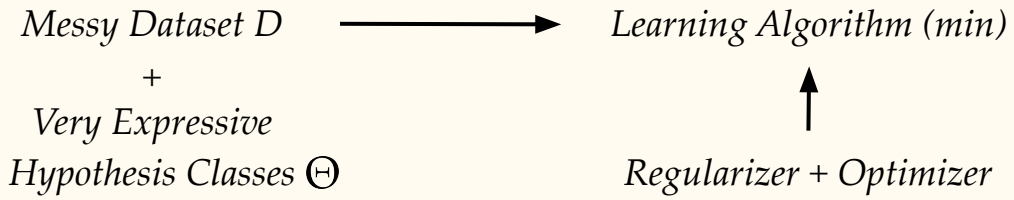
Very Expressive Hypothesis Classes Θ

Welcome to ReFoRM!

What is this reading group about? Foundations of "real-world" ML

How is "real-world" ML different from "idealized" ML?

Real-world ML: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}\left[\ell(z_i; \theta)\right]$

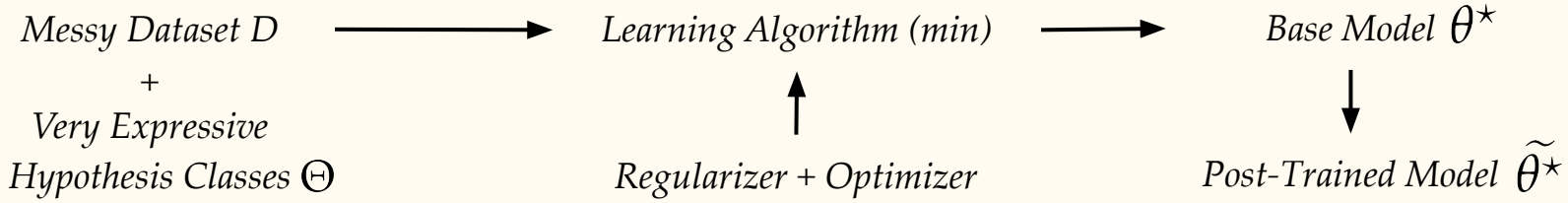


Welcome to ReFoRM!

What is this reading group about? Foundations of "real-world" ML

How is "real-world" ML different from "idealized" ML?

Real-world ML: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}\left[\ell(z_i; \theta)\right]$

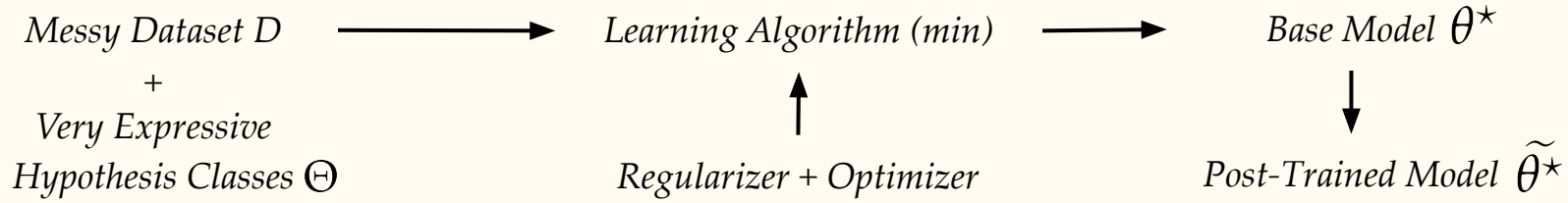


Welcome to ReFoRM!

What is this reading group about? Foundations of "real-world" ML

How is "real-world" ML different from "idealized" ML?

Real-world ML: $\theta^\star = \arg\min_{\theta \in \Theta} \mathbb{E}_{z \sim D}\left[\ell(z_i; \theta)\right]$



Implications: Unpredictability, theoretical wisdom might not apply, new considerations, need to understand new phenomena, …

Goal of this group

What do rigorous foundations for this new age of ML look like?

How can tools from statistics, CS theory, and operations inform a *better understanding* of machine learning algorithms and systems?

What are the right questions to ask, and phenomena to explain—at what *level of abstraction* should we be aiming to explain them?

What theoretical models not only *explain* unexpected phenomena, but also *predict* new phenomena that we can verify experimentally?

Intended format (thanks for signing up!)

Goal: Build intuition, leverage group's diversity, start collaborations (bringing new perspectives from everyone's field)

Sign up: https://tinyurl.com/reform-ml-s26

Goal(*s*) of the discussant (1-2 every week):

1. A single "deep dive" per week about one subject (can be multiple papers)
2. We have suggested several papers for each week, *more* than one can cover thoroughly in a week. Pick a small + focused paper set and read thoroughly
3. Prepare a 20-30 minute presentation, accessible to a second year PhD student, focusing on (a) *seeding discussion* and (b) *identifying gaps and connections,* and (c) *formulating open problems*

Everyone else: Read paper/watch talk/something! *Try to come with some familiarity*

Introductions!

What is your *name*?

What *program and year* are you in?

What *focus area* are you most interested in?

What are you *working* on? What do you *want to work* on?

What brought you to this reading group?

Outline for the Quarter

Introduce the theme for this quarter:

Understanding and Improving LLMs via a Theoretical Lens

The quarter is divided into three *sessions* (each two-week long)

Each Session's Goal: *Explore a sub-area in depth*

Understand the known results

Identify gaps

Formulate open problems

Sessions This Quarter

A) Models of Internal Structure of LLMs (≈ April 16th, 23rd, 30th)

B) Using Algorithms to Compress LLMs (≈ May 7th, 14th)

C) Systems to Improve LLMs (May 21st)

D) Tentative: RL-view on LLMs (May 28th)

Meetings This Quarter

Aprl 17nd (today!)	Low logit rank hypothesis
January 29th	Skipping due to ICML deadline
February 5th	Explanations of EoS / Sharpness & Generalization
February 12th	Double Descent
February 19th	Benign Overfitting
February 26th	Grokking
March 5th	Other emergent abilities
March 12th	Reserved for extra meeting on above / different topic

Session 1
Low Logit-Rank Hypothesis

Sequences of Logits Reveal the Low Rank Structure of Language Models

Noah Golowich*
Microsoft Research
noah.golowich@austin.utexas.edu

Allen Liu*
UC Berkeley
aliu42@berkeley.edu

Abhishek Shetty*
MIT
shetty@mit.edu*

LLM Abstractions

- LLMs are extremely complicated and we lack the mathematical tools to analyze them

- Conventional mathematical wisdom based on worst-case assumptions implies LLMs should *not* perform well

Q: *Are there any meaningful abstractions we can use that are analytically tractable and are aligned with empirical LLM behavior?*

Models to Understand Large Language Models

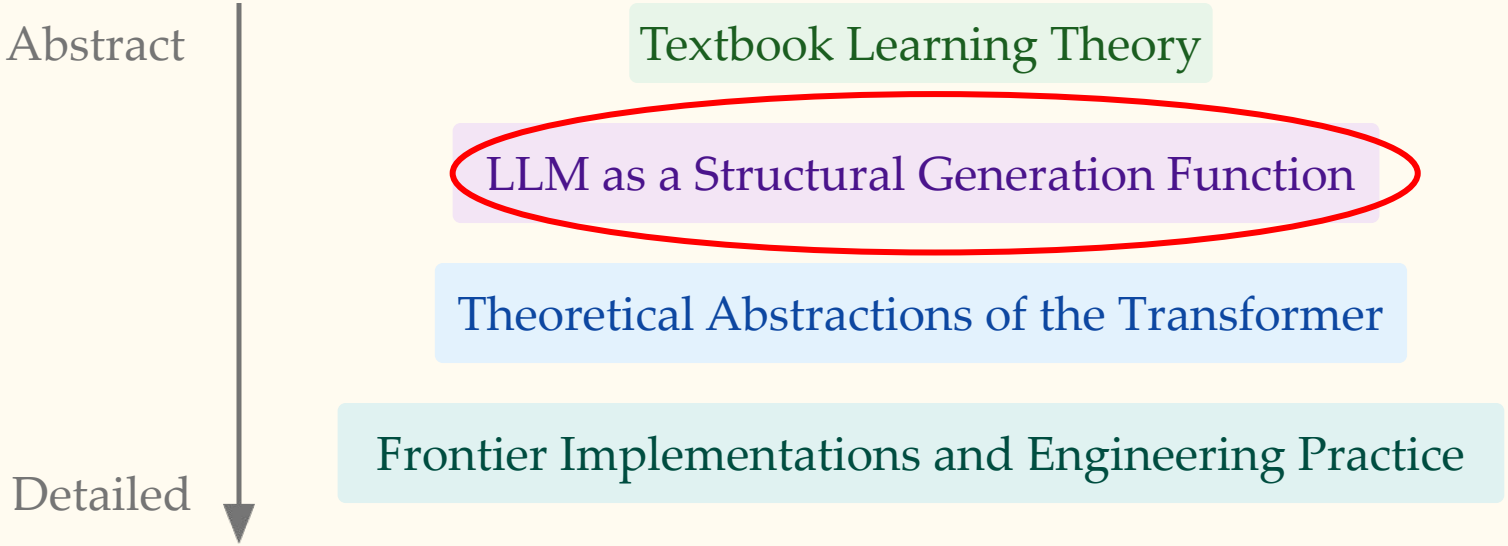
Large Language Models (LLMs) are extremely complicated

What is a good *model* to reason about LLMs?

- Needs to make *testable* predictions on *real* models
- Needs to be *tractable* for reasoning/analysis

No good answer yet!

Hierarchy of LLM Abstractions



Structure of LLMs

- Long line of work has studied the *structure of the representation* of the LLM (e.g., LoRA)

- In this talk we focus on the *structure of the generation function* that the LLMs induce

> **Q:** *How should one view the generation function so that its structure can allow for mathematical analysis and empirical verification?*

Logits in Transformer-Based Language Models

Given *history* h (sequence of tokens), a transformer-based LM is a *distribution* over the *next-tokens* v in the vocabulary V

Logits in Transformer-Based Language Models

Given *history* h (sequence of tokens), a transformer-based LM is a *distribution* over the *next-tokens* v in the vocabulary V

In particular,

$$\log \Pr[z|h] \propto w(z) \cdot g(h) \quad \text{for} \quad w(z), g(h) \in \mathbb{R}^h$$

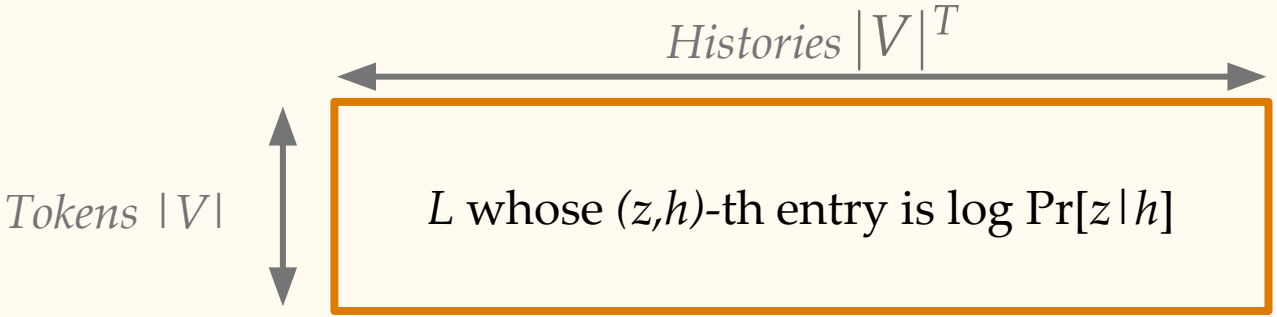
Logits in Transformer-Based Language Models

Given *history h* (sequence of tokens), a transformer-based LM is a *distribution* over the *next-tokens v* in the vocabulary V

In particular,

$$\log \Pr[z|h] \propto w(z) \cdot g(h) \quad \text{for} \quad w(z), g(h) \in \mathbb{R}^h$$

This implies that the *logit matrix L has rank at most h*



Low-Rank Structure Enables *Stealing* Models

Stealing Part of a Production Language Model

Nicholas Carlini [1] Daniel Paleka [2] Krishnamurthy (Dj) Dvijotham [1] Thomas Steinke [1] Jonathan Hayase [3]
A. Feder Cooper [1] Katherine Lee [1] Matthew Jagielski [1] Milad Nasr [1] Arthur Conmy [1] Itay Yona [1]
Eric Wallace [4] David Rolnick [5] Florian Tramèr [2]

"*[Utilizing the low-rank structure of language models…]* for *under $20 USD,* our attack *extracts the entire projection matrix* of OpenAI's `ada` and `babbage` language models… we also recover the *exact hidden dimension* size of the [last-layer of] `gpt-3.5-turbo` model"

Extended Logit-Matrix

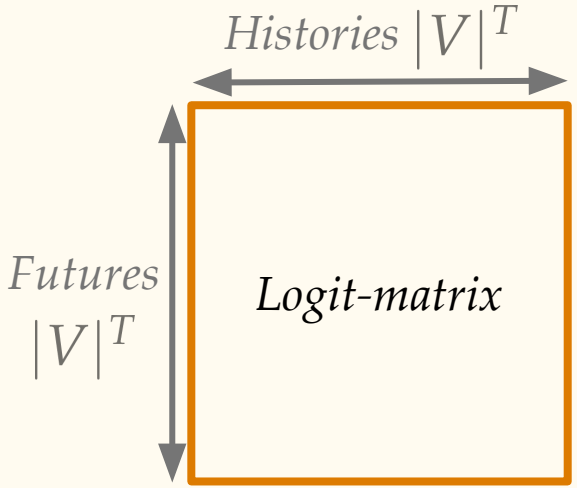
Extended Logit-Matrix L's entries are indexed by histories h and futures f, as

$$L_{f,h} := \log \Pr[f|h]$$

Question: *Is the extended logit-matrix also low rank?*

Not in the worst-case: Transformers' extended-logit matrices can have rank $\Omega(T)$

Challenge: The extended logit-matrix has exponential size and so hard to compute



Modified Extended Logit-Matrix

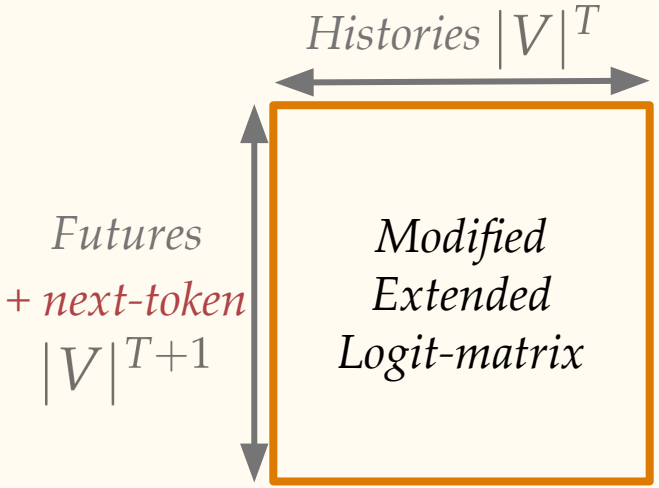
Modified Extended Logit-Matrix L: Entries indexed by histories h, futures f, and next-tokens z

$$L_{(z,f),h} := \log \Pr[z|h \circ f]$$

Question: *Is the extended logit-matrix also low rank?*

Not in the worst-case: Transformers' extended-logit matrices can have rank $\Omega(T)$

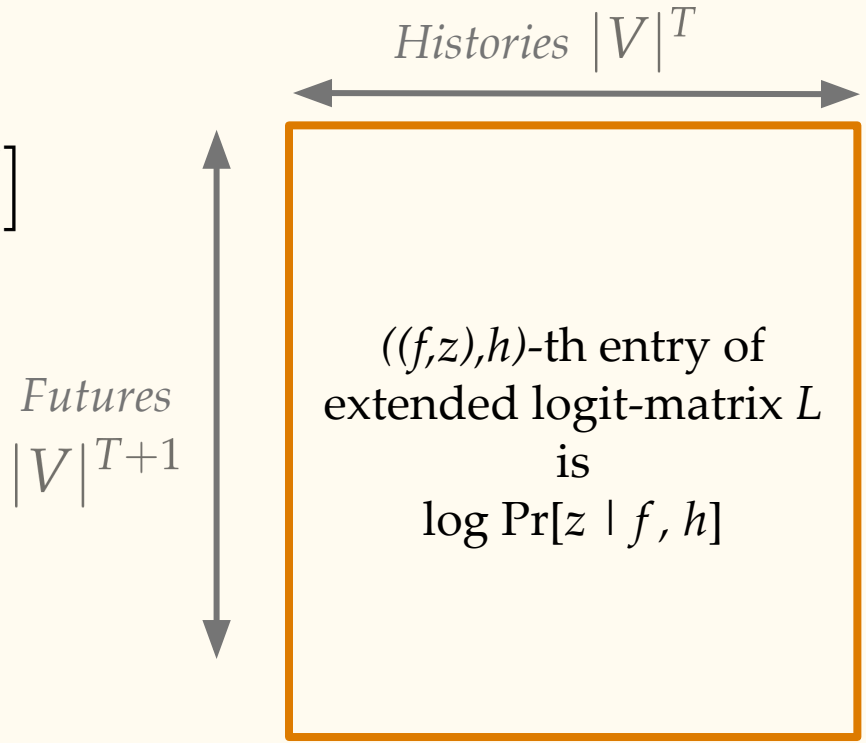
Challenge: The extended logit-matrix has exponential size and so hard to compute



(Modified) Extended Logit-Matrix

The extended logit-matrix's entries are

$$L_{(z,f),h} := \log \Pr[z|h \circ f]$$



Methodology of Empirical Results

- Sample *n strings S* from a dataset *D* (`wiki` split of `olmo-mix-1124`)
- Select length-*T prefixes* from *S* and let this be the set of *histories H*

- Sample *n strings S* from D
- Select length-*T suffixes* from *S* and let this be the set of *futures F*

- For each future *f*, vary *z* over 50 most-likely next tokens after *f*

Compute *singular value-decomposition* of the resulting extended logit-matrix

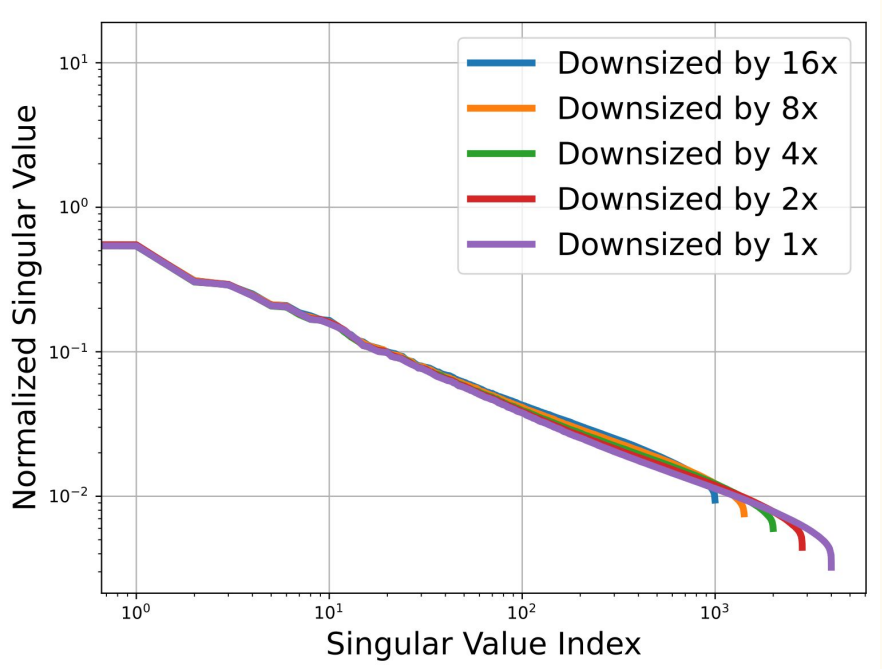
Empirical Results with OLMo-1b Model

Observation: *Power-law decay*

$$\sigma_i^2 \approx i^{-\alpha} \quad \text{for} \quad \alpha = 1.12$$

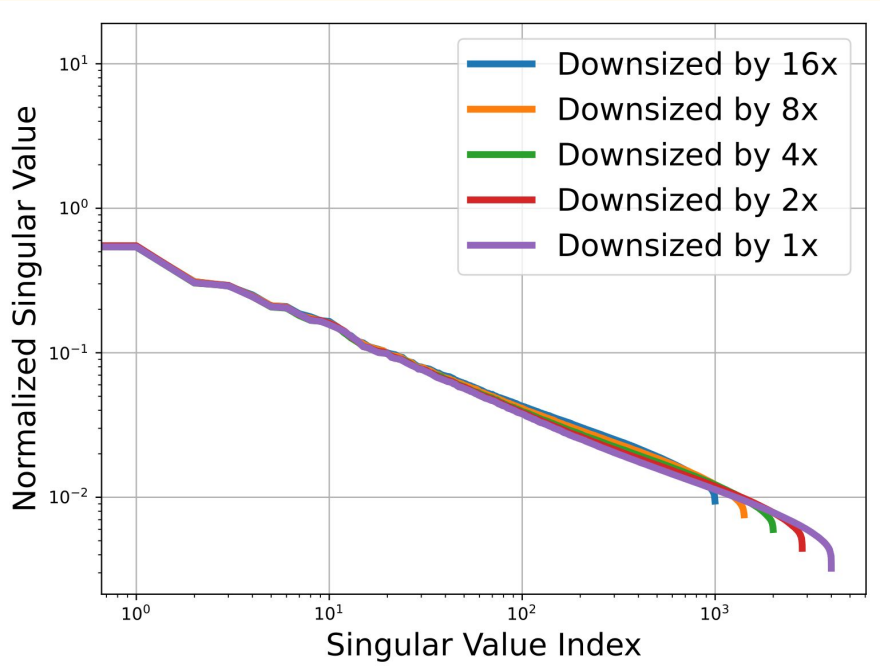
Implication: Since $\alpha > 1$, each L can approximated to 99% in Frobenius norm with O(1) singular values. That is, *L is well-approximated by a constant rank matrix.*

Robustness? Holds for different choices of n and k

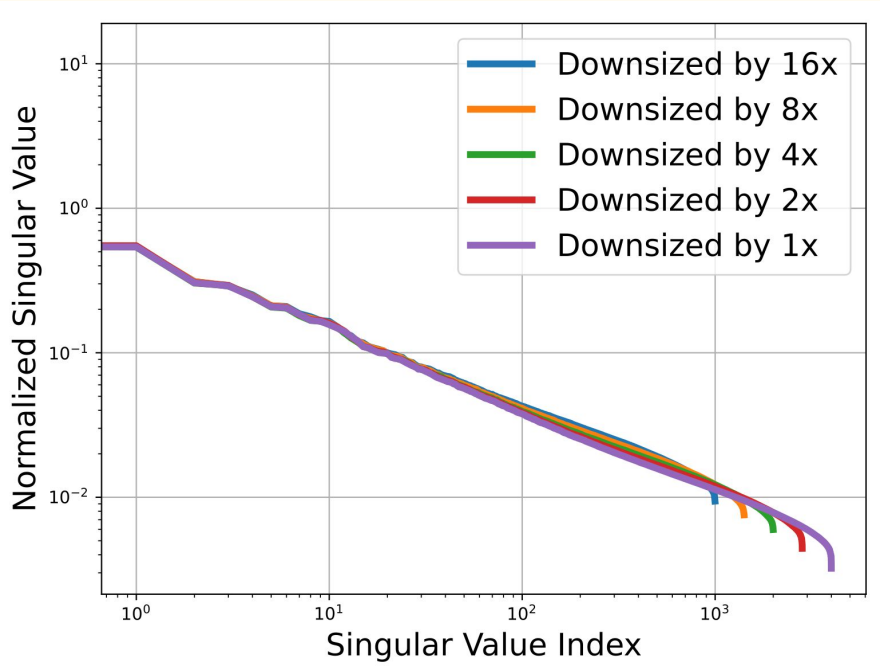


OLMo-1b

Is Low-Rankness an Artifact of the Architecture?



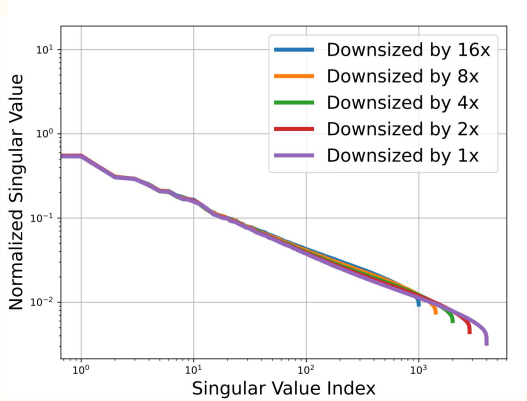
OLMo-1b (checkpoint 0) ($\alpha \approx 0.748$)



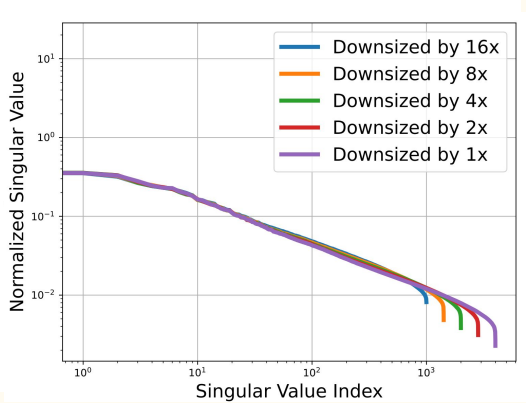
OLMo-1b ($\alpha \approx 1.122$)

Is Low-Rankness an Artifact of the Model?

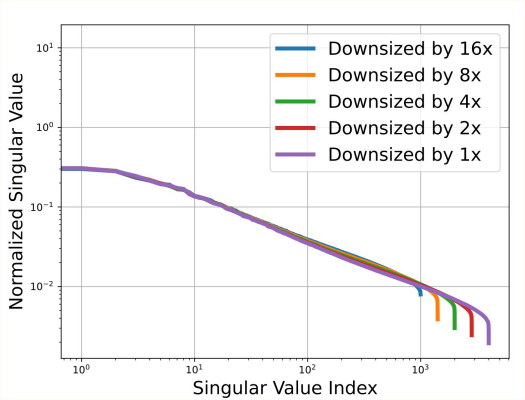
OLMo-1b
($\alpha \approx 1.122$)



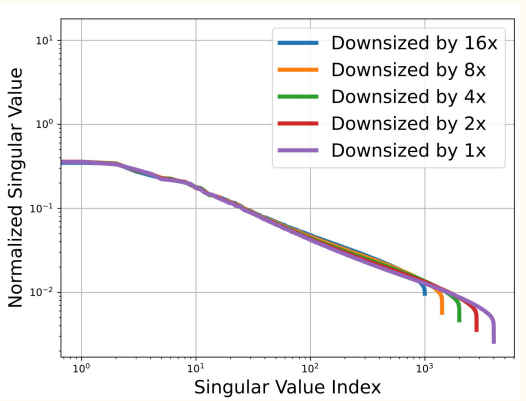
Gemma-1b
($\alpha \approx 1.130$)



Llama-1b
($\alpha \approx 1.146$)

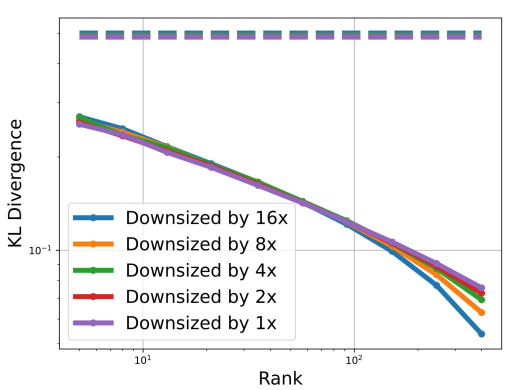


Mamba-1.4b
($\alpha \approx 1.126$)

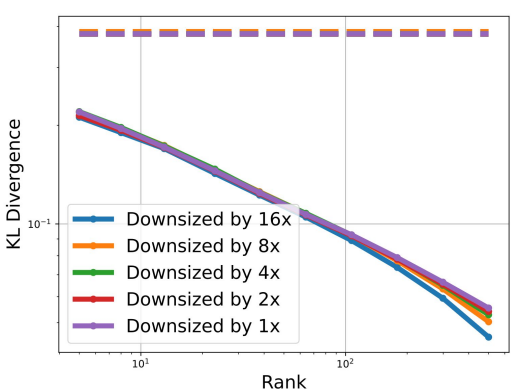


Is Low-Rankness an Artifact of the Metric?

OLMo-1b



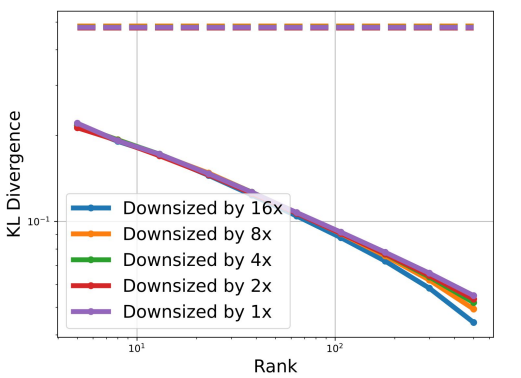
Gemma-1b



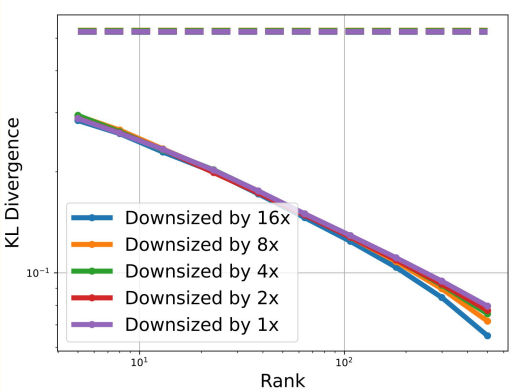
Metric: Average over *histories h* and *futures f* of the *KL-divergence* between the model's distribution and the distribution of its *best rank-r SVD approximation*

Is Low-Rankness an Artifact of the Metric?

Llama-1b



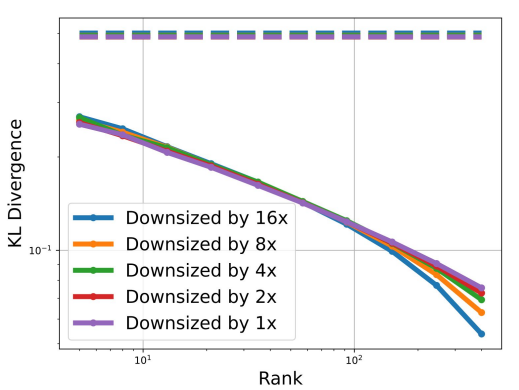
Mamba-1.4b



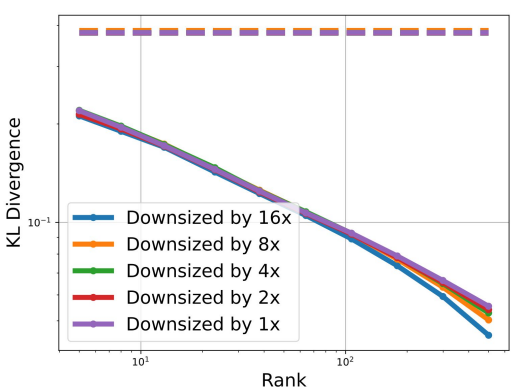
Metric: Average over *histories h* and *futures f* of the *KL-divergence* between the model's distribution and the distribution of its *best rank-r SVD approximation*

Is Low-Rankness an Artifact of the Metric?

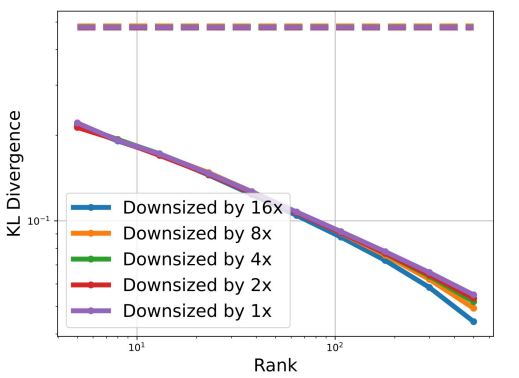
OLMo-1b



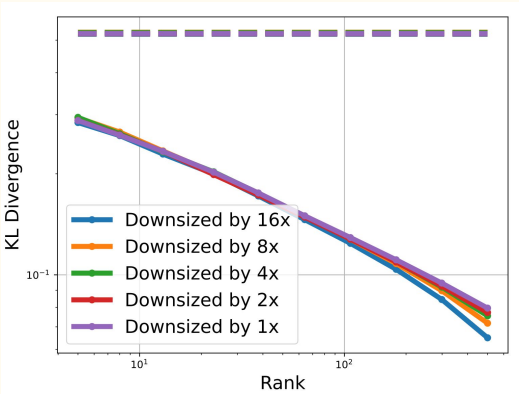
Gemma-1b



Llama-1b



Mamba-1.4b



LinGen: Decoding via a Linear Combination of Helper Prompts

Goal: Generate from a target prompt h^* *without* querying the model on h^*

Key Idea: Approximate **extended logit row** using some "helper" prompts:

$$L_{h^*} \approx \sum v_i L_{h_i}$$

LinGen Algorithm

1. **Choose helper prompts** $h_1, \ldots, h_m$
2. **Fit coefficients** v so the helpers reconstruct the target row
3. At step t, append the current generated prefix $y_{<t}$ to each helper and use the linear combination for the next token

Why LinGen works

Core intuition: low-rank predictive structure

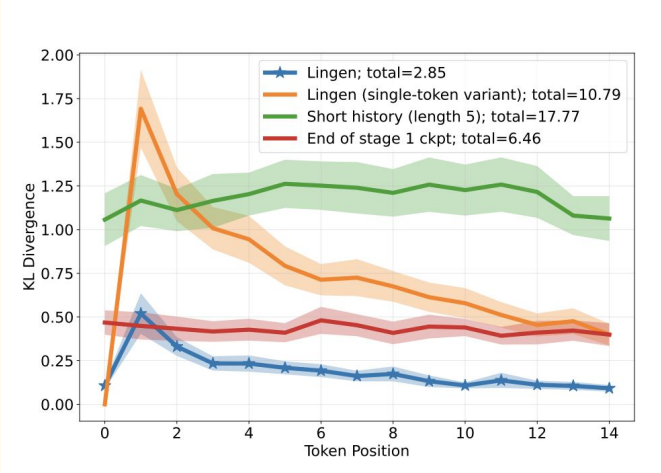
Since the model's **extended logit matrix** is approximately low rank:

- Many different histories are not independent.
- They lie near a shared **low-dimensional predictive space**.
- A target history can be approximated by a linear combination of other histories: $L_{h^*} \approx \sum v_i L_{h_i}$

Why the same coefficients keep working during generation

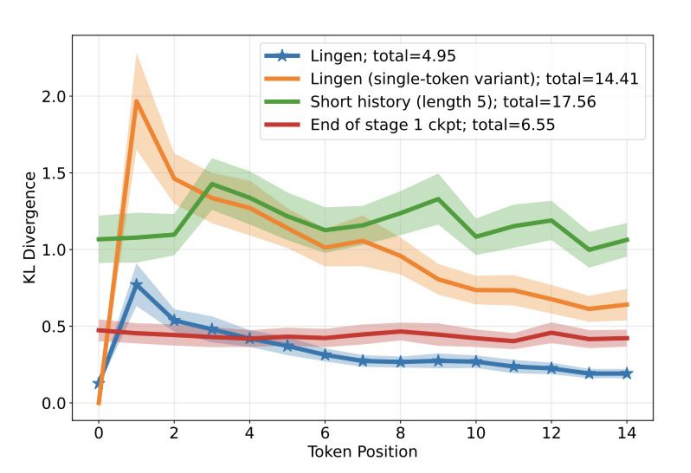
If this relationship holds over a broad set of futures F, then after appending the same generated prefix $y_{<t}$ to both the target and helper histories, the linear combination continues to provide a strong approximation for the next token.

Empirical Performance of LinGen



OLMo-1b

“in-distribution”



OLMo-1b

“out-of-distribution”

Learning a Low-Logit-Rank LM from Logit Queries

Oracle Access: For query any prefix, receive next-token logit vector
Assumption: sequence-level logit behavior is exactly low rank d

1. Adaptive Future Sets

Track small set of futures F□ to test if approximation is sufficient.

2. Basis Histories

DistSpanner selects histories whose logit rows span the useful space.

3. Verification Loop

Sample futures: if error is found, add to F□ and **restart**.

4. Global LP

Solve **one LP across timesteps** for stable, bounded coefficients.

5. Sampling

Approximate logits → softmax → sample next token autoregressively .

Theorem Takeaway

- Output is a **compact sampler** P close to target P
- **TV(P, P) ≤ ε** in poly-time/query under low-rank assumption

Caveat: *Theorem requires stronger approximation conditions than current empirical fits currently demonstrate. Parallel work relaxes them.*

Many Open Questions…

1. **Further implications of this model?**
2. **Why does the low-logit rankness arise during pre-training?**
3. **Transfer of "LinGen coefficients" from open source to frontier models?**