Alon Rosen

Low Logit Rank

My hovercraft is full of eels.[*]

The *low logit rank* phenomenon is an empirical observation about the input–output behavior of trained language models. It says that the model's behavior across different prompts lies close to a low-dimensional subspace, giving rise to linear relations between them.

For a single next-token prediction, low-dimensional logit structure follows from the final linear map on the hidden state. Here, the same structure extends to logits over longer token sequences.

This stronger structure allows the output on some target prompt to be reconstructed from queries on other prompts, so the target itself need not be queried. With chosen logit queries, it can also support efficient model learning and extraction.

The main points of this chapter are:

- Extended logit matrices contain next-token logits along multi-token continuations.
- In trained language models, these matrices are often approximately low rank.
- Low rank can let us probe behavior on a target prompt without querying the prompt itself.
- Logit queries can exploit this structure for efficient model learning and extraction.

[*]Monty Python, *Monty Python's Flying Circus*.

1 The low logit rank phenomenon

A Hungarian tourist enters the tobacconist at Thorpebank Road in London and, consulting his phrasebook, tells the storekeeper, "My hovercraft is full of eels." Somehow, he leaves with a box of matches. How could such prompting evoke the intended but unrelated response?

Empirically, trained language models produce logit patterns near a low-dimensional subspace. With approximate linear relations one can, for instance, recover a target prompt ("...matches?") from unrelated queries ("...eels."), or learn the model's output distribution from chosen logit queries.

1.1 The extended logit matrix

Let V be the token set. Histories and futures are strings $h, f \in V^*$, and $\ell(v \mid h) \in \mathbb{R}$ denotes the logit of token v after history h. Adding the same constant to every logit does not change the softmax probabilities, so we center:

$$\bar{\ell}(v \mid h) := \ell(v \mid h) - \frac{1}{|V|} \sum_{a \in V} \ell(a \mid h) \in \mathbb{R}.$$

Write hf for concatenation. Let $\boldsymbol{\ell}(\cdot \mid h) \in \mathbb{R}^{|V|}$ be a next-token logit vector. If $\boldsymbol{\ell}(\cdot \mid h) = W\boldsymbol{\phi}(h)$ with $W \in \mathbb{R}^{|V| \times d}$ and $\boldsymbol{\phi}(h) \in \mathbb{R}^d$, then centered logit vectors lie in a subspace of dimension at most d. Hence any $d+1$ such vectors must satisfy a nontrivial linear relation.

To see what changes beyond the next token, take $H = \{h_1, \ldots, h_k\}$ and $F = \{\epsilon, f_1, \ldots, f_k\}$. The relevant logits have the following $k \times (k+1)$ block form:

	ϵ	f_1	$\cdots$	f_k
h_1	$\bar{\boldsymbol{\ell}}(\cdot \mid h_1)$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_1 f_1)$	$\cdots$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_1 f_k)$
h_2	$\bar{\boldsymbol{\ell}}(\cdot \mid h_2)$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_2 f_1)$	$\cdots$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_2 f_k)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
h_k	$\bar{\boldsymbol{\ell}}(\cdot \mid h_k)$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_k f_1)$	$\cdots$	$\bar{\boldsymbol{\ell}}(\cdot \mid h_k f_k)$

Each block is a centered logit vector in $\mathbb{R}^{|V|}$. The first column is the ordinary next-token case; every other column appends the same future to all histories.

Definition 1 (Extended logit matrix)**.** *The matrix $L = L(H, F) \in \mathbb{R}^{|H| \times (|F||V|)}$ has rows indexed by $h \in H$ and columns indexed by $(f, v) \in F \times V$, with*

$$L_{h,(f,v)} = \bar{\ell}(v \mid hf).$$

If F contains only the empty future, this is the ordinary next-token matrix. Otherwise the output layer no longer bounds its rank, since the hidden state after hf depends jointly on h and f.

1.2 Approximate low rank

The full matrix is far too large to construct. If histories have length m and futures length n, it has $|V|^m$ rows. Each future f contributes a block of $|V|$ columns (f, v), one for every possible next token v, giving $|V|^n \cdot |V| = |V|^{n+1}$ columns in total.

Instead, about 10^4 histories and 10^4 futures are sampled from natural-language data. Each sampled history contributes one row. For each sampled future, the full block would contain $|V|$

token columns; only the 50 most likely tokens after that future are kept. The sampled matrix therefore has about 10^4 rows and $50 \cdot 10^4$ columns. The question is whether the same low-rank behavior persists as the sampled sets are enlarged.

Empirical observation. *For trained language models, sampled extended logit matrices appear to be* close to low rank*: they are well approximated by matrices of much smaller rank.*

To quantify "close to low rank," consider a sampled matrix A. Its singular value decomposition can be written as

$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top + \sum_{i=r+1}^{q} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_q > 0.$$

The vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ give successive independent directions, while σ_i measures the strength of the ith direction. Writing $A_r = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, this is the closest rank-r approximation to A, and the discarded squared singular values $\sum_{i>r} \sigma_i^2$ give its total squared entrywise error.

The singular values therefore determine how much rank is needed for a given accuracy. For trained models they decay roughly as a power law, with nearly the same decay as the sampled matrix is enlarged. Thus, for a fixed approximation accuracy, the required rank remains roughly unchanged as the matrix dimensions grow. There is no universal value of r: it depends on the model and the desired accuracy, and the exact rank of the matrix may still be large.

Because the sampled matrix is itself large, one need not compute its full SVD. Truncated methods estimate only the leading singular directions, while randomized variants first compress the matrix to a smaller subspace. This makes the leading spectrum practical to estimate without a full decomposition.

1.3 Prediction from linear relations

Write $L = L_r + E$, where L_r has rank r and E is the residual. A row relation in L_r holds across every future block, while E measures the error when the same relation is applied to L.

Proposition 1. *If L is close to rank r, then every continuation can be reconstructed approximately from r reference histories using coefficients independent of the future.*

Proof. Choose $h_1, \ldots, h_r$ whose rows form a basis for the row space of L_r. For every $h \in H$ there are coefficients $c_1, \ldots, c_r$ such that

$$L_r(h, \cdot) = \sum_{i=1}^{r} c_i L_r(h_i, \cdot).$$

For the full matrix,

$$L(h, \cdot) - \sum_{i=1}^{r} c_i L(h_i, \cdot) = E(h, \cdot) - \sum_{i=1}^{r} c_i E(h_i, \cdot).$$

Restricting this identity to any future block gives the corresponding continuation error. Thus a small residual E gives accurate reconstruction only when the coefficients are also controlled; large coefficients can amplify the residual. □

The key advantage of the extended matrix is that the same coefficients persist after every future continuation. Empirically, the resulting relations remain stable enough across generated prefixes to produce coherent multi-token continuations despite the approximation noise.

2 Applications

2.1 Indirect generation

Low logit rank can be used to reproduce the behavior of a target history h from a small set of reference histories. Suppose coefficients $c_1, \ldots, c_r$ have been found so that the target row is approximately the same linear combination of the reference rows over a sufficiently rich set of futures. The generation phase then need not query the model on h itself.

For a generated prefix g, query $h_1 g, \ldots, h_r g$ and form

$$\sum_{i=1}^{r} c_i \, \bar{\boldsymbol{\ell}}(\cdot \mid h_i g) \approx \bar{\boldsymbol{\ell}}(\cdot \mid hg).$$

Applying softmax gives an approximation to the next-token distribution after hg; sample the next token and repeat. The same coefficients are reused at every step. This persistence relies on the extended logit matrix. A relation among ordinary next-token logits may deteriorate after one generated token, whereas a relation among rows of the extended matrix is designed to persist along common continuations.

The reference histories can be unrelated to the target and can even be nonsensical token sequences. The resulting next-token distributions remain close enough to the target distribution over several steps to produce coherent continuations; using only the one-step logit relation degrades much more quickly after the first token.

This is a form of model extraction: the goal is to reproduce the model's output behavior, not to recover its parameters.

2.2 Learning from logit queries

A stronger question is whether low logit rank makes the entire generative distribution learnable. Here the goal is not to recover the weights of the original network. Given a model $\mathbb{M}$ over length-T token sequences, the learner should output an efficiently sampleable distribution $\widehat{\mathbb{M}}$ whose distribution over complete sequences is close to $\mathbb{M}$.

Ordinary samples do not suffice in the worst case. Even exact logit rank 2 can encode *learning parity with noise* (LPN), conjectured hard from independent noisy examples. Yet with chosen queries LPN is easy: repeatedly query each basis vector and recover the corresponding secret bit by majority vote.

The same contrast motivates logit-query access. For any chosen history h, a *logit query* returns the full next-token logit vector $\boldsymbol{\ell}(\cdot \mid h)$, equivalently the conditional next-token distribution up to an additive normalization. The learner can therefore choose which histories to probe and directly expose linear relations between them.

The formal assumption is stronger than observing a small SVD tail on one sampled matrix. At every position, the full logit matrix must be well approximated on average by a rank-d matrix under histories and futures drawn from the model, with a bounded low-dimensional factorization.

Proposition 2 (Learning from logit queries)**.** *With bounded approximate low logit rank and small error, chosen logit queries efficiently learn the full sequence distribution in total variation.*

Thus a polynomial-time algorithm using polynomially many chosen logit queries outputs an efficiently sampleable model with nearly the same sequence distribution. It recovers behavior, not the original network parameters.

3 Historical notes

The one-step rank restriction induced by the output projection is closely related to the softmax bottleneck. Carlini et al. and Finlayson, Ren and Swayamdipta used this type of structure to infer information about API-protected language models [1, 2].

Golowich, Liu and Shetty introduced the extended logit matrix and observed low approximate rank in modern language models [3]. They showed that the resulting linear relations support generation from unrelated or nonsensical histories. A follow-up proves efficient learning of the model's output distribution from chosen logit queries [4].

A separate phenomenon, called *subliminal learning*, shows that behavioral traits of a teacher model can pass to a student through semantically unrelated generated data during distillation [5]. This is related in spirit to hidden information in unrelated outputs, but is not known to follow from low logit rank.

References

[1] N. Carlini et al. *Stealing Part of a Production Language Model*. ICML, 2024.

[2] M. Finlayson, X. Ren, and S. Swayamdipta. *Logits of API-Protected LLMs Leak Proprietary Information*. COLM, 2024.

[3] N. Golowich, A. Liu, and A. Shetty. *Sequences of Logits Reveal the Low Rank Structure of Language Models*. arXiv:2510.24966, 2025.

[4] N. Golowich, A. Liu, and A. Shetty. *Provably Learning from Modern Language Models via Low Logit Rank*. arXiv:2512.09892, 2025.

[5] A. Cloud et al. *Language Models Transmit Behavioural Traits through Hidden Signals in Data*. Nature 652, 615–621, 2026.