Large Language Models are Locally Linear Mappings

Anonymous authors
Paper under double-blind review

Abstract

Despite significant progress in transformer interpretability, an understanding of the computational mechanisms of large language models (LLMs) remains a fundamental challenge. We demonstrate that the inference operation of LLMs can be mapped to an equivalent linear system that nearly exactly reconstructs the predicted output embedding for a given input sequence. Extending techniques from image diffusion models that exhibit local or piecewise linearity, we strategically detach components of the gradient computation with respect to an input sequence for a next-token prediction such that the Jacobian of the model reproduces the output with one linear operation per input token. We demonstrate this approach across models, including Qwen 3, Gemma 3, Llama 3, Phi 4, Mistral Ministral and OLMo 2, up to Llama 3.3 70B Q4. With the singular value decomposition of the detached Jacobian, we show that these LLMs operate in extremely low-dimensional subspaces where the largest singular vectors decode to distinct concepts related to possible output tokens. We examine the equivalent linear operation of each successive layer (and its attention and multilayer perceptron components) and observe the emergence of semantic concepts. We demonstrate that the detached Jacobian of middle layer representations can be used as a steering operator to insert semantic concepts into unrelated text, which could be useful for improving safety and decreasing bias. Despite their expressive power and global nonlinearity, modern LLMs can be interpreted through locally linear decompositions that provide insights into their internal representations and reveal interpretable semantic structures in the next-token prediction process.

1 Introduction

The transformer decoder has become the architecture of choice for large language models (Vaswani et al., 2017) and efforts toward a conceptual understanding of its mechanisms are ongoing. Significant insights include sparse autoencoders for conceptual activations in LLMs (Bricken et al., 2023), "white-box" alternative architectures (Yu et al., 2023), minimally sufficient architectures (He & Hofmann, 2023) and analytic results on generalization (Cowsik et al., 2024). While transformers are complex globally nonlinear functions of their input, we demonstrate how to compute an equivalent linear system that nearly exactly reconstructs the predicted output embedding for a given input sequence.

Our approach builds on two previous results: Elhage et al. (2021) found that attention-only networks with basic language generation abilities transfer semantic information across the network with interpretable circuits, including the "induction head", and Kadkhodaie et al. (2023) showed that powerful image denoising diffusion models can be made exactly locally or piecewise linear through several architectural constraints and interpreted as low-dimensional adaptive linear filters.

We extend these ideas by demonstrating local linearity for modern LLMs with gated linear activations. For many open-weight LLMs, the gradient operation with respect to the input can be manipulated at inference such that the output prediction is unchanged and also almost exactly locally linear. This numerical Jacobian computation captures the complete forward operation of the model, including activation functions and attention, although it is only valid for that particular input sequence (more "pointwise" linear than piecewise linear).

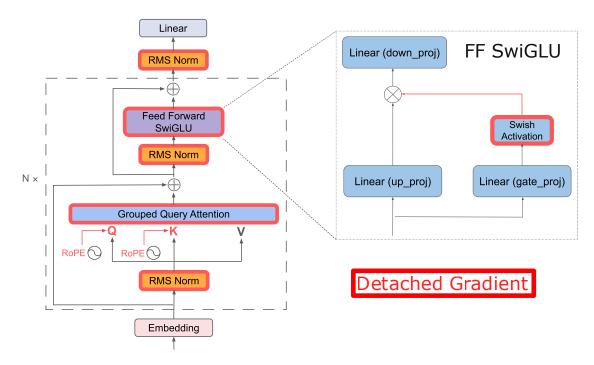


Figure 1: A schematic of the Llama 3 transformer decoder (Grattafiori et al., 2024; Nvidia, 2024). The gradient detach operations for components outlined in red effectively freeze the nonlinear activations for a given input sequence, creating a linear path for the gradient with respect to the input embedding vectors, but do not change the output. The output embedding prediction can be nearly exactly mapped to a linear system by the Jacobian autograd operation. The feedforward module with a SwiGLU activation function is shown in expanded form to demonstrate how the nonlinear Swish term can be detached from the gradient to form a linear path, achieving local linearity for a given input. The RMSNorm layers and and softmax attention blocks also must be detached from the gradient.

This approach allows us to analyze entire models, from input embeddings to predicted output embedding, as equivalent linear systems for a particular input sequence. By examining the singular value decomposition (SVD) of the equivalent linear system, we can measure the local dimensionality of the learned manifolds involved in next-token prediction, and we can decode the singular vectors into output tokens. This analysis can also be done layer by layer, or for individual attention and multilayer perceptron (MLP) modules, in order to observe how these models compose next-token predictions.

We demonstrate local linearity in model families including Qwen 3, Gemma 3, Llama 3, Phi 4, Mistral Ministral, Deepseek R1 0528 Qwen 3 8B and OLMo 2, at a range of sizes up to Llama 3.3 70B Q4. This approach does not require further model training, offering a new path to interpreting a wide range of openweight LLMs at a local level that could serve as a complement to other powerful interpretability methods.

2 Method

2.1 The Jacobian of a deep ReLU Network

Mohan et al. (2019) observed that deep ReLU networks for image denoising which utilize zero-bias linear layers are "adaptive linear" functions due to their homogeneity of order 1 at a given fixed input, which enables interpretation as a nearly-exact linear system. Given the homogeneity at a fixed input, the network's output can be nearly exactly reproduced by numerically computing the Jacobian matrix of the network at a particular input image \mathbf{x}_{im}^* and multiplying it by \mathbf{x}_{im}^* .

$$\mathbf{y}_{\mathbf{im}}^* = \mathbf{J}(\mathbf{x}_{\mathbf{im}}^*) \cdot \mathbf{x}_{\mathbf{im}}^* \tag{1}$$

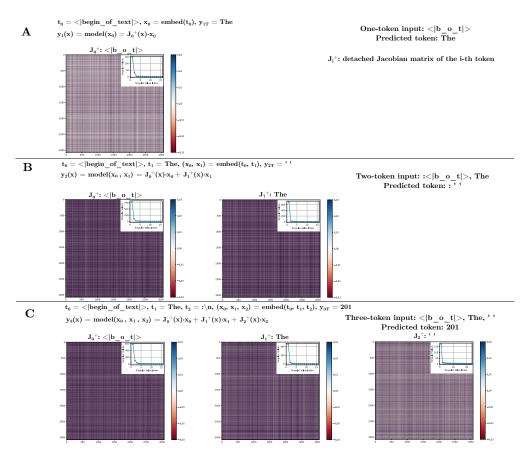


Figure 2: An overview of next-token prediction in the Llama 3.2 3B transformer decoder and decomposition of the predicted embedding vector computation using the detached Jacobian. Generating three tokens with only $\langle |BoT| \rangle$ as input produces "The 201". For each prediction, each input token $\mathbf{t_i}$ is mapped to an embedding vector $\mathbf{x_i}$, and the network generates the embedding of a next token. The phrase turns out to be "The 2019-2020 season". The detached Jacobian $\mathbf{J^+(x)}$ of the predicted output embedding with respect to the input embeddings is composed of a matrix corresponding to each input vector. Each detached Jacobian matrix $\mathbf{J_i^+(x)}$ is a function of the entire input sequence but operates only on its corresponding input embedding vector. The matrices tend to be extremely low rank, shown in the inset figures, and the matrix $\mathbf{J_0^+}$ varies across A), B) and C) above because the input sequences differ. Since the detached Jacobian captures the entirety of the model operation in a linear system (numerically, for a given input sequence), tools like the SVD can be used to interpret the model and its sub-components.

Due to the global nonlinearity of the network, the Jacobian must usually be computed again at every input of interest. The Jacobian may be the same for similar inputs in the same piecewise region of the response (Balestriero & Baraniuk, 2021).

2.2 The Jacobian of a transformer decoder

Many open weight LLMs also use linear layers with zero bias, like the ReLU network of Mohan et al. (2019). A transformer decoder predicts an output token embedding \mathbf{y} given a sequence of k input tokens $\mathbf{t} = (\mathbf{t_0}, \mathbf{t_1}..., \mathbf{t_k})$ mapped to input embedding vectors $\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}..., \mathbf{x_k})$, where \mathbf{t}^* and \mathbf{x}^* represent a particular sequence. The output embedding prediction is a nonlinear function of the input embedding vectors $\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_k}$, as LLMs utilize nonlinear gated activation functions for layer outputs (SwiGLU for Llama 3, GELU for Gemma 3 and Swish for Qwen 3) as well as normalization and softmax attention blocks.

Gated activations like $Swish(\mathbf{x}) = \mathbf{x} \cdot sigmoid(\mathbf{x})$, with a linear term and a nonlinear term, are also an "adaptive" linear function or, more generally, an adaptive homogeneous function of order 1 (Mohan et al., 2019). If the $sigmoid(\mathbf{x})$ term that gives rise to the nonlinearity is frozen for a specific numerical input, e.g. an embedding vector \mathbf{x}_0^* (Elhage et al., 2021) (or equivalently detached from the computational graph with respect to the input), then we have a linear function valid only at \mathbf{x}_0^* where (1) holds and we can numerically compute a Jacobian that carries out $Swish(\mathbf{x}_0^*)$ as a linear operation.

Below we show that computing the Jacobian after effectively substituting specific values for the nonlinear terms also works for other gated activation functions, zero-bias RMSNorm layers and softmax attention blocks. We further demonstrate that for a given input sequence the entire transformer decoder is an adaptive homogeneous function of order 1 where we can apply necessary gradient detachments and numerically compute a linear system that nearly exactly reproduces the transformer output embedding \mathbf{y}^* .

The Jacobian $\mathbf{J}(\mathbf{x})$ of a transformer is the set of matrices generated by taking the partial derivative of the decoder inference function $\mathbf{y}(\mathbf{x}) = f(\mathbf{x_0}, \mathbf{x_1}..., \mathbf{x_k})$, with respect to each element of each $\mathbf{x_i}$ (where $\mathbf{x_i}$ for Llama 3.2 3B has length 3072, for example, and therefore the Jacobian matrix for each embedding vector is a square matrix of this size).

We introduce a "detached" Jacobian J^+ , which is a set of matrices that captures the full nonlinear forward computation for a particular input sequence \mathbf{x}^* as a linear system. The detached Jacobian is the numerical Jacobian of the LLM forward operation when its gradient includes a specific set of detach() operations for the nonlinear terms in the normalization, activation and attention operations. Fig. 2 shows how each matrix of the detached Jacobian operates on its corresponding input embedding vector to provide a nearly-exact reconstruction of the LLM forward operation (with a relative error of 10^{-6} (the standard deviation of the reconstruction error divided by the standard deviation of the output embedding), see Fig. 3).

$$\mathbf{y}^* = \sum_{i=0}^k \mathbf{J}_i^+(\mathbf{x}^*) \cdot \mathbf{x}_i^* \tag{2}$$

While the traditional Jacobian \mathbf{J} for a particular input sequence \mathbf{x}^* is a locally linear approximation of the nonlinear LLM forward operation, it does not generate an exact reconstruction at \mathbf{x}^* since the transformer function is not linear. The detached Jacobian \mathbf{J}^+ evaluated at \mathbf{x}^* is the result of an alternative gradient path through the same network which is linear for the input \mathbf{x}^* . The detached Jacobian \mathbf{J}^+ only generates a near-exact reconstruction at \mathbf{x}^* and not in the local neighborhood due to the strong nonlinearity of the decoder inference function. The detached Jacobian matrices differ for every input sequence and must be computed numerically for every sequence.

2.3 Nonlinear layers as locally linear operations

In order to achieve local linearity, modifications must be made to the gradient computations of the RMSNorm operation, the activation function (SwiGLU in Llama 3.2) and the softmax term in the attention block output.

2.3.1 Normalization

Normalization layers like LayerNorm (Xu et al., 2019) or RMSNorm (Zhang & Sennrich, 2019) are non-linear with respect to their input because they include division by the square root of the variance of the input.

$$norm(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{var(\mathbf{x})}}\tag{3}$$

Mohan et al. (2019) devised a novel bias-free batch-norm layer which disconnects the variance term from the network's computational graph. Their batch-norm layer returns the same values as the standard batch-norm layer, but it is locally linear at inference as the nonlinear operation is removed from the gradient computation.

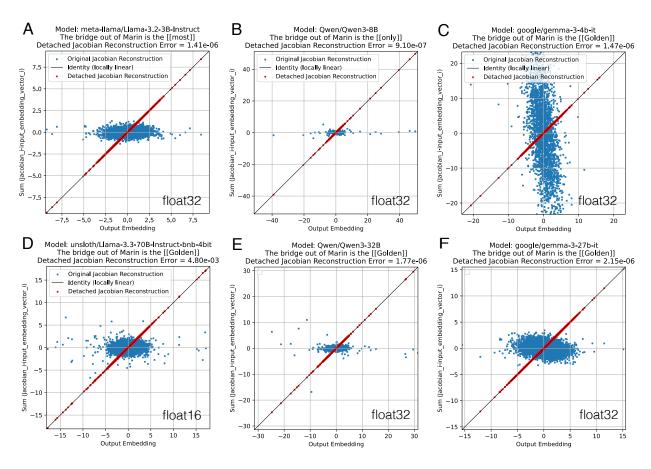


Figure 3: For the input sequence "The bridge out of Marin is the", the elements of the predicted output embedding vector of the model compared to the elements from the Jacobian reconstruction for both the original Jacobian (blue points) and detached Jacobian operations (red points), shown for a range of model families and sizes. Note that the detached Jacobian reconstructions nearly exactly match the predicted embedding in each case, with relative error (the standard deviation of the reconstruction error divided by the standard deviation of the output embedding) on the order of 10^{-3} for float16 precision and 10^{-6} for float32 precision. Llama 3 models are shown in A) and D), Qwen 3 models in B) and E), and Gemma 3 models in C) and F). These plots demonstrate the near-exact reconstruction of the detached Jacobian and therefore the local linearity of the mapping.

This is also similar to the "freezing" of nonlinear terms in attention-only transformers from Elhage et al. (2021).

We make a similar change for Llama 3.2 3B by altering how the gradient with respect to the input is computed at inference for RMSNorm. This is accomplished by substituting the value for the input vector \mathbf{x}^* for only the variance term as in (4). In PyTorch, this is accomplished by cloning and detaching the \mathbf{x} tensor within the variance operation, so its value will be treated as a constant. The gradient operation is still tracked for \mathbf{x} in the numerator, so that term will be treated as a variable by autograd.functional.jacobian. The gradient of the function is then computed at \mathbf{x}^* (we assume for simplicity a sequence of length 1).

$$norm_{LL}(\mathbf{x}) = \frac{\mathbf{x}}{\sqrt{var(\mathbf{x}^*)}} \tag{4}$$

We define the detached Jacobian as follows:

$$\mathbf{J}_{\mathbf{n_{LL}}}^{+} = \left[\frac{\partial}{\partial \mathbf{x}} norm_{LL}(\mathbf{x})\right]|_{\mathbf{x} = \mathbf{x}^{*}}$$
 (5)

We can rewrite the locally linear RMSNorm as follows:

$$norm_{LL}(\mathbf{x}^*) = \mathbf{J}_{\mathbf{n}_{\mathsf{T},\mathsf{T}}}^+(\mathbf{x}^*) \cdot \mathbf{x}^* \tag{6}$$

At inference, we now have a locally linear RMSNorm whose output is numerically identical to the one used in training. However, when we take the gradient with respect to the input vector \mathbf{x} in eval mode, the numerical output is the detached Jacobian matrix $\mathbf{J}_{\mathbf{n_{LL}}}^+$, which we can use to reconstruct the normalization output exactly as a linear system; no higher-order terms are needed.

The goal is to apply this same approach for other nonlinear functions in the decoder such that the entire computation from the input embedding vectors to the predicted output is locally linear, and we can compute and interpret the set of detached Jacobian matrices.

2.3.2 Activation functions

While Mohan et al. (2019) relied on ReLU activation functions, which do not require any changes to achieve local linearity, Llama 3.2 uses SwiGLU (Shazeer, 2020), Gemma 3 uses approximate GELU (Hendrycks & Gimpel, 2016) and Qwen 3 uses Swish for activation functions. Fortunately, there is a linear \mathbf{x} term in each of these, and the gradients can be cloned and detached from the nonlinear terms. This manipulation produces a locally linear SwiGLU layer with respect to the input \mathbf{x} . Below, $Swish(\mathbf{x}) = \mathbf{x} \cdot sigmoid(\mathbf{x})$ and \otimes is element-wise multiplication.

$$SwiGLU(\mathbf{x}) = Swish(\mathbf{W}\mathbf{x}) \otimes (\mathbf{Z}\mathbf{x}) \tag{7}$$

$$SwiGLU_{LL}(\mathbf{x}) = [Swish(\mathbf{W}\mathbf{x})]|_{\mathbf{x}=\mathbf{x}^*} \otimes (\mathbf{Z}\mathbf{x})$$
(8)

$$SwiGLU_{LL}(\mathbf{x}^*) = (\left[\frac{\partial}{\partial \mathbf{x}}SwiGLU_{LL}(x)\right]|_{\mathbf{x}=\mathbf{x}^*}) \cdot \mathbf{x}^*$$
(9)

$$SwiGLU_{LL}(\mathbf{x}^*) = \mathbf{J}_{\mathbf{SwiGLU}_{LL}}^+(\mathbf{x}^*) \cdot \mathbf{x}^*$$
(10)

Detaching the gradient from the Swish output thus allows for a locally linear form of SwiGLU at inference. A similar procedure may be carried out for GELU with Gemma 3 (see supplement).

2.3.3 Attention

The softmax operation at the output of the attention block can also be detached, with the linear relationship preserved through the subsequent multiplication with \mathbf{V} , which is a linear function of \mathbf{x} . Below, $\mathbf{Q} = \mathbf{W}_{\mathbf{Q}}\mathbf{x}$, $\mathbf{K} = \mathbf{W}_{\mathbf{K}}\mathbf{x}$ and $\mathbf{V} = \mathbf{W}_{\mathbf{V}}\mathbf{x}$.

$$Attn(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^{T}}{\sqrt{d_k}}) \cdot \mathbf{V}$$
(11)

$$Attn_{LL}(\mathbf{x}) = [softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})|_{\mathbf{Q}=\mathbf{Q}^*, \mathbf{K}=\mathbf{K}^*}] \cdot \mathbf{W}_{\mathbf{V}}\mathbf{x}$$
(12)

$$Attn_{LL}(\mathbf{x}^*) = (\left[\frac{\partial}{\partial \mathbf{x}} Attn_{LL}(\mathbf{x})\right]|_{\mathbf{x} = \mathbf{x}^*}) \cdot \mathbf{x}^*$$
(13)

$$Attn_{LL}(\mathbf{x}^*) = \mathbf{J}_{\mathbf{Attn}_{LL}}^+(\mathbf{x}^*) \cdot \mathbf{x}^*$$
 (14)

The linear \mathbf{x} term within \mathbf{V} makes it possible for the attention block to be locally linear at inference, as the gradient for the softmax output is detached.

2.3.4 The Transformer Decoder

With the above gradient detachments for the normalization layers, activation functions and attention blocks, the transformer decoder network is locally linear with respect to \mathbf{x}^* when evaluated at \mathbf{x}^* (shown here for length k).

$$\mathbf{y}^* = \sum_{i=0}^k \mathbf{J}_i^+(\mathbf{x}^*) \cdot \mathbf{x}_i^* \tag{15}$$

The output of the network incorporating the above gradient detachments is unchanged from the original architecture.

3 Results

3.1 Local linearity of the predicted output

In order to validate whether the detached Jacobian achieves local linearity, we can compare the predicted output embedding vector for a given input token sequence to the Jacobian reconstruction of the output.

Fig. 3 examines the original output and the output when incorporating the appropriate gradient detachments for local linearity for two sizes of Llama 3, Qwen 3 and Gemma 3. The reconstruction of the output embedding with the detached Jacobian matrices falls close to the identity line when compared with the output embedding, which is evidence for local linearity. The same comparison with the original Jacobian does not fall on the identity line. The standard deviation of the difference for the detached Jacobian divided by the standard deviation of the output embedding vector is on the order of only 10^{-6} , which quantifies how exactly the detached Jacobian reproduces the output). These plots therefore validate the local linearity of Llama 3, Qwen 3 and Gemma 3 with the appropriate gradient detachments for a particular input.

The numerical computation of the top k singular vectors of the Jacobian takes on the order of 10 seconds for an input sequence of 8 tokens for Llama 3.2 3B on a T4 GPU. At the other end of the spectrum, the singular vectors for the same sequence with Llama 3.3 70B Q4 on an H100 GPU takes more than a minute. (A method for computing the top singular vectors without forming the full matrix with Lanczos iteration is under development.)

3.2 Single-unit feature selectivity and invariance

Since the detached Jacobian applied to the input embedding reproduces the predicted output embedding vector, and the elements of the predicted output embedding vector are the units of the last transformer layer, the rows of the detached Jacobian matrices represent the input features to which the last layer units are selective and invariant for that particular input sequence (Kadkhodaie et al., 2023; Mohan et al., 2019).

The activation of a particular unit in the last layer is determined by the inner product of a row of the detached Jacobian and the input embedding vector. We can sort by the mangitude of row norms, then map the largest-magnitude rows of the detached Jacobian back to the input embedding space (by finding the nearest-neighbor embedding vectors from the tokenizer) to determine the tokens that cause each unit to be strongly positive or negative. We can see in Fig. 4 (in appendix A below) that the units respond strongly to the words of the prompt, including "bridge", "Marin" and "is". Decoding of the rows of the detached Jacobian for each token as well as the distribution of activations for this sequence is shown in Fig. 4A.

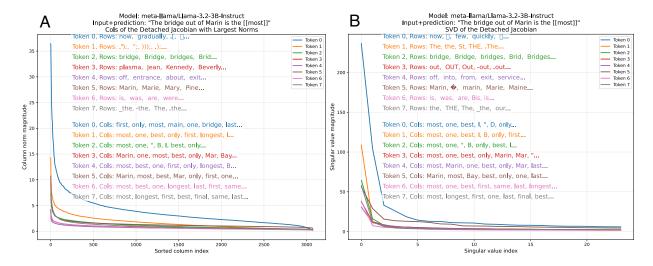


Figure 4: Given the sequence "The bridge out of Marin is the", the most likely prediction is "most" for Llama 3.2 3B. The detached Jacobian matrices for each token represent a near-exact local linearization of the predicted output embedding. A) We show the features which drive large responses in single units in the last decoder layer, which are the rows of the detached Jacobian with the largest norm values, and decode each of those into the most likely input embedding token. The blue list of words at the top are the ordered decoded "feature" input tokens from the largest rows of the detached Jacobian matrix for the beginning of sequence token, and the different colors show the decoded feature tokens for the other input tokens. A similar operation is carried out for columns of the largest norm values, which are decoded to the output token space. Note that the activation distribution of column magnitudes is fairly sparse, with only a few features driving the response. B) We take the singular value decomposition of the detached Jacobian matrix corresponding to each input token, which summarizes the modes driving the response, and decode the left and right singular vectors U and U to output and input embeddings, shown in colors. The singular value spectrum is extremely low rank, and decoding the U singular vectors returns candidate output tokens: "most" and "one" appear frequently. Decoding the U singular vectors returns variants of the input tokens like "bridge", "Marin" and "is", as well as others that are not clearly related to the input sequence.

The columns of the Jacobian can also be decoded to the output token space, and these turn out to be tokens that could be predicted, which include words like "most" or "first", which could be acceptable outputs.

3.3 Singular vectors of the detached Jacobian

An alternative approach is to look at the singular value decomposition of the detached Jacobian $J_i^+ = U\Sigma V^T$, following Mohan et al. (2019). Since the detached Jacobian represents the forward computation, and there are no higher order terms, the fact that the SVD is very low rank shows the entire forward computation can be approximated with only a few singular vectors operating on the input embeddings.

In Fig. 4B, the singular vectors are decoded for three different models of two sizes each, from 3B to 70B parameters. The right singular vectors V are decoded to input tokens in the same way the rows of the detached Jacobian were above, and we see almost identical decoding of the top tokens to the features driving the most active single units. The left singular vectors U can be decoded to output embedding tokens, and "most" is the strongest, as it was in the columns of the detached Jacobian matrices.

3.4 Singular vectors across model families

Fig 5 shows this same analysis for Llama 3, Qwen 3 and Gemma 3 across two different sizes of each. Note the low-rank structure of each of the detached Jacobians, as well as the differing decoding of the top singular vectors from each input embedding vector. The first or "beginning of sequence" token has the highest

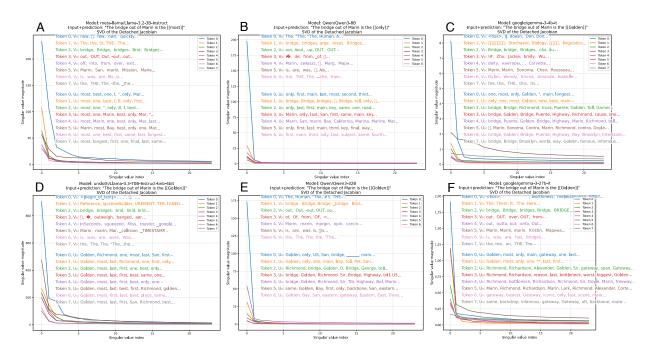


Figure 5: Singular value decomposition of the detached Jacobian for different families and sizes of language models (from 3B to 70B parameters) evaluating the input sequence "The bridge out of Marin is the", followed by a predicted token. The left singular vectors decode to tokens related to bridges and local geography, particularly the Golden Gate Bridge, while singular value spectra all have extremely low rank (see below for quantification). Each row shows top tokens associated with different singular vectors, demonstrating how models encode semantic knowledge about the input sequence and the prediction. See Fig. A1 for Deepseek R1 0528 Qwen 3 8B Distill, Phi 4, Mistral Ministral and OLMo 2.

magnitude in each spectrum reflecting how the positional encoding is entangled with semantic information in the detach Jacobian representation.

3.5 Layer output singular vectors

Tables 1, 3 and 2. show the decodings of the top three singular vectors of the detached Jacobians of the layer outputs for Llama 3.1 8B, Gemma 3 4B and Qwen 3 8B. Similar patterns emerge across models: the words "bridge" (and its variants), "Golden", "highway", "exit", "most" and "only" are highlighted to show their appearances in singular vector decodings. Early layers are excluded as the decodings are unintelligible. The emergence of intelligible tokens in later layers is shown in the tables as something like a phase change in the representation. Llama 3.1 shows a progression from generic bridge terms to specific San Francisco geographic terms and generating "Golden". Gemma 3 includes more diverse geographic terms including international locations but still results in "Golden". Qwen 3 produces more infrastructure and engineering concepts before producing "only".

Fig. 6A shows the normalized singular value spectra of the detached Jacobian at the output of every layer. Llama 3.2 3B has 28 transformer layers, and decoding the largest singular vectors shows that the word representation of these intermediate operations is not interpretable until later layers. From the decoding of the top singular vector by layer, "only" emerges in layer 19. From the map of the progression of the projection of the top two singular vectors onto the top two singular vectors of the last layer in Fig. 6B, we first see a shift at layer 11 toward the prediction.

Since the layer-by-layer operations are only linear and have no higher-order terms, the stable rank $R = (\sum_{i=1}^{L} S_i^2)/S_{max}^2$ serves as a measure of the effectively dimensionality of the subspace of the representation at a particular layer.

When looking at $W_{0_to_k}$, the cumulative layer transform up through layer k, the dimensionality of the detached Jacobian steadily decreases. When considering each layer i as its own individual transform W_i (where $W_{0_to_k} = \prod_{i=0}^k W_i$ for the simplified scenario of a single input token; there are other cross-token terms not shown here for mid-layer detached Jacobians for longer input sequences), we also see a large peak in dimensionality near the end.

		Deput token 1	Deput trians 2	Deput trikes 3	Deput teken 4	Topost tribers 5	Deput token 6	Sepat token 7
	m will in Not for	/Dk HeuderCode [OF ToPoint (typed upd Menay NCY	weit ve eller inn settiral alsolies	A sity peay its instead of a pend heat	ina insta olio. Cuity thi indo pray pend	spat 296 singleton releas via ob venue fed	ceratic Strip base ceracy Ridge murphic sediest RE	ceratic ceracy have folloystem ATEE reflect cerain fixip
Laper 14_1	weit : dealler skeden LineStyle Bridge yeine	ta dels own samplads arism :	id mar one fland 2	se weit gew Uplacide sysl rempose	heatmap yel plode arken pared Sqifterore	lenin lie 202 Depressind peds emps professed/bylar	ktop apan mktop respon situal	pre opend sintpir keedsy selets
Laper 14_2	olster Burgess (Ayperd Holge Subsystem Perudicens)	aga realist another aread	Kelly Jar Cunningham Jarvis Jeonat Hadism ortal	AnnilePlatEuten yackennously Fueldlatton Messitivity SetArtire Tick QOSpot	inn ills indrard Segment 440 vin Earls ager	KoratHandler Dickinson vida bio Sr MNN	ate atea stad at /is spired iyel	men y at aerid /is elk stope ATO
	wil i Mage gas iyah eval dedicated Mage	skis about c/typerof Trifficunds not /Dfa	bridge bridges Bridges Mond din	rgs Xnity atted IntPer rimon relong gred	rgs. Unity raining rises ross HD atted	wiki abolish i Kan Kutton	constit ATER Strip ceracy. Expert cenata ner ToF	service ge servey service ATER base ser
Laper 15_1	web Bridge squeeze dealler bridges water bridge	illin k Amedation sphe ATEG beidges	ge RAD JoseifuPlatEntics Dickinson Publishers NullEcoption Court	ner spil inesp yel sår ines gr	yel ge heatmap illin Jennsheien arken indiseries for	med CAPE Crastal Cape Surf (BitpContest Bristal Industrial	rdes apan ATEG ade upo respen	Pententur sur opend://navtypes.respen.cellForEce.ht
Laper 13 2		armed more pair agained and sail rading	Brare Consingham Inox Bow Issues	roling mond rol indeed olid links viewing	nes 255 spen morel Bil pet vital links		nt ng rwine nte kiek own enton	indi creat at our Extrus servid rade
Laper 16_5	i m mire bi yer wil reis iyah	sole /ress 850000 pdev sof spd +(types)	bridge bridges Bridge is also Bridges toll	Raid call Raid Raid SCCT Chicky SCCT S.	East rate East East to EXIT hote Cuty	northern Northern Northern inco ainers Tottenham Richmond Fold	constit Stelp not sensits ATEX marteles Station go	ceratic Kalasysicas consts go sole ser Statiter censory
Laper 16_1	Bridge bridge squeeze aksa KqEterrer weit falsa	h 700 ini III akra - Knit	ge RAD Dirkinson Uploside PushBarton elevada	Keyld NallKareption ole pared ner in improheseful	illin pared heatmap Upbaide go yel uppersaar SharedPir	Cuperine Handed Klicon Santa Honda uni avakeProsNth	bridge respons artigate to SWXVZ egod	coupes per Pentertor hereby was add Uplimide
Laper 16 2	Sevalidate BAD (Sygeof Subsystem /ems XUTKA saids	armi age ag reel auction /out dipl yil	vere leasure Braco Ecute employment closure mit vere	PushEntion Intalidate yackennously EAD <7 anilablytyped extl AlmostEqual	pat links link prog organis ress Titl	Cape CAPE Corneal Southampton Portland Marin WXYZ Portland	at regamen quiv /is own itself lef	900 251 shi dit errot on imir ren
Laper 17_6		noir /ress pr Bridge bridge PushButton ToP	bridge Bridge bridges Bridges h clocke Brig	reit Exit own Exit as favorites seal ties	rait tin Enit favories sale own Enit S.	northern Northern Northern Galdard Maria San	constitute are ATER ToF ORIZED reveales go sole	ceratic gr Eastine-Object say ONEDED cerain sole :
Laper 17_1	Bridge Marin bridge arLayout Brains bridges choke /connection	k yer EDENTITY 700 Affected Bridges Local	ge PushEntion ANNEL Impiration bust cerusion	NullSusption pel såe inesp illin vital undu	illin eh Equivalent Uplanide Equivalent ge Seguenta arken	Cupretise Hundred Oakland Hilron Beloley Santa Liver UC	egnal avigate tell sala sekles bridge Fac	hereby stope r was new deat rditary
Laper 17 2		austher inter hell to any aread new grader	hosting elector Brace resal head Hosting location Earlsboar	PushEstion pselecurously go availablytyped swick EnquateWriter Invalidate men-		Cape Consull Maria Preis CAPE Southern Jet Pre-	latter undry iyan ispesa miable ulur	dit 255 930 san matemiar errori on imili
Laper 18_5	bridge on 1 years old Nil ye San	bridge Bridge bridges sale remains brid bridge runner	bridge Bridge bridges Bridges brid Bridge bridge	bridge Bridge suit East 17 caser bridges rance	NY bridge Bridge swit East Nov of Marin	San Marin Galdand San SF Samameric van Northern	ORIZED creatic are sale same ATER ToF Easting-Object	ORIZED cerutic uer sole Easting-Object pr pits
Laper 18_1	Bridge bridge bridges Marin Bridge Bridges brid	Bridge Intiges OUNTRY EDINTITY Queries Choices	PushEuten ge Tos ANNEL Impiration Heate sole	ray vital pel ineep oft ater abat unba	alast for decrementation bridge arken Uphnide udi Equivalent	Stanford Capertine Oakland Berkeley Santa Kiliron California Pale	bridge bridges Bridge Bridges brid bridge egoal uncerator	hereby stope now to Uphoids :// Datum pun
Laper 18_2	Sevalidate analahlytyped Burke Naunders stid Lair agure	to another elaredore map her inter hell	Xan resal hosting element RF Boses	Puddletten perkennendy aralablytypel bradiskriv swirk initialiar/op new	Kan Karramento Guldand Maria Kan Northern san HF	San San Marin MI Cornwall Partile Brittany Robing	also ours itself lef nin at g leter	dit seun resut 251 au mon Adjacent 979
Laper 10_0	bridge RF seat San maly connecting Oakland usp	bridge Bridge bridges recening to I bridge choler FF	bridge Bridge bridges Bridges Bridge bridge brid	enit bridge Enit Maria NF Bridge cause bridges	NF Marin walk of Nan bridge East Eridge	Oakland Maria SF Berkeley Eay Northern of SF	cerutic ORIZED sole ner RustimeObject name Route Segments	covarie sole Rantina-Object say ORIZED (DGC)
Laper 19_1	Marin Bridge bridge 17 Golden Bridge bridge	Bridges Orbits shims they and HD Oubland system	Highway Route mate ANNEL Hwy bighway	undu et gew ner vital inesp Keyfel hip	Oakland Northern of see singe Northern ottenham Marin	Stanford Oakland Coperties Silicon Helseley Fals Liver Liver	bridge bridges Bridge bridge Bridges of SF recessing	hereby now of stope TokenType \// M NT
Laper 10 2		clarebore to new Oakland awriter heli progress	HF Kan where Oukland Golden Handord closure of	yackennosky Paskillation arrior analolskytypel hard japony	Shorich was et RetVal stretch IBHp connecting	Monterry Santa Marin San Facilie Santa San Bristol	men also lef itself latar g own	sews dit die inne istest own OLL infras
Laper 20_0	bridge toll mill Bridge bridge San SF Toll	bridge Bridge toll bridges Toll toll coming bridge	tell bridge Bridge tel latest bridges Tell Bridges	tell said East bridge Bridge tel Tell NF	tell said 10° Maria bridge East Bridge San	Morin Oakland RF Hay Puls Berkeley Northern of	tell bridge tel Four Tell Bridge ure connecting	bridge toll Bridge bridges covate Tell ser RuntimoChjert
	Marin of NF toll ar Layout Oakland Ereletry	Odds Oakland routes a thing whiles and intentor	Soute reste Biglioug Soute odus mate Buy	undu latest clove near kip et gow detester	Oakland stireham Marin Northern of Pale sumir Erg	Stanford Capertine Oakland Silven Pale Arbilles Senta Liver	tell bridge bridges Bridge of Tell recoving 82'	heavily now stops bridge of Brid Bridge M
Laper 20_2	toPleat PLEED Southern LA Los Firtcher wink bridge	to eluvature new ella ele proissala syn	NF Oubland Stanford assess New NF Ave of	analablytyped SMSOI yarkennouly Publishes immel	Shortch EstVal et WalgetStem LautEreur utan stretch	Parific Maria Prais Parific Brittany Monterry and Truman	g lotar com also count lef artic	dit was info Bul ink they CLL

Table 1: The top three singular vectors of the detached Jacobian for the layer outputs from Llama 3.1 8B for the sequence "The bridge out of Marin is the" with the prediction [[Golden]]. Legend: "Bridge", "only", "highway", "exit", "Golden", "most". This is a selection of the middle layers, see A.4 for full tables that extend to the last layer.

		Isput token I	Input token 2	Input token 3	Report token 4		Input tolon 6
	twoking CONSTANTS	vacc getch Period	first hardest factors exiting ramp	hardest fastest first exiting easiest Interstate	function whatsoever whilst Misc Hierarchy	fissk thm	this repeat
Layer 22_1	metaphor unc DERP OBJC stret .wp ISP	substr MROL bridge	hurry HIP opi Rockets TORT	screes	bere nation gim pick	eec eeub =BitConverter FTIVE .Guna equivalent	TRGL spNet
Layer 22_2	alk ole ool sos angan icon va	autics iker rocking	backstory weblog SVG JSON INCIDENT	ESA keyet .INSTANCE ECB	Theme Proxy height Entities Encode ultipart /System D		
Luyw 23_0	salopes CONSTANTS getch Uncomment massage TRY	metaphor bridge largest easiest cally first centerpiece	first hardest une fastest highway bottleneck	hardest first fastest casiest toughest weakest une quickest	Whilst whilst role CONTRACT Whilst	flock than	thm redit dates still
Layer 23_1	metaphor unc Derne makeshift OBJC	bridge Bridge bridge bridge	scenes WithError opi Timing presets Entering	scenes WithError	California California bere nation lower	ly at ett ath of tra ty	TRGL
Layer 23_2	ros lovers flutter	autics jams	weblog COMPONENT amot metaphor	fallest PGA ESA keyet JVM classic focus	specialised bespoke connexion macOS UNESCO SSH EU initialise		ity ation age my cretical ie ance acock
Layer 24_0	first third last most largest fourth culmination	metaphor largest conterpiece easiest first hardest bridge gaterany	hardest first easiest factors most once same	hardest first fastest casiest most longest same loast	Whilst whilst Firstly Whilst crate	culmination first fourth largest highest third same seventh	first fourth third last largest highest fifth most
Layer 24_1	metaphor makeshift .wp REAK	brid bridge bridge bridge Bridge	scenes LocalStorage WithError	scraes	entirety Malta main third favoured Yemen	ut ly ath ett of ality hereby	it ities
Layer 24_2	flatter ros Lingu		phenomena pauz annot metaphor	ESA fullest PGA machine focus ECB	bespoke macOS connexion specialised mount initialise UNESCO Patreon	GE CO	_
Layer 25_0	largest most first longest latest fastest last third	bridge bridges Bridge gateway	hardest once suit casiest first must fastest highway	hardest fastest easiest most once weakest first suit	Whilst Firstly dimensions Mahmond	note abortions edd	first fourth third last largest highest most weakest
Layer 25_1	bridge bridges Bridge Bridges brid	bridges bridge Bridge Bridge	suit suit suits eternity suit	esit	Malta main entirety Jordan third Klopp Nicar	eec fax streets hotlino Ustudent tats	ities it
Layer 25_2	bridges bridge Bridge bridge parliament	Exit out jame	INCIDENT symbolism	JVM migrations CAB largest upcoming machine	specialised initialise UNESCO bespoke connexion mount arist macOS		e cion ted
Layer 26_0	first most largest last longest latest gateway only	bridge bridger metaphor gateway connecting	highway first soil ones last hardest roads	highway hardost first weakest last road roads easiest	highway California highways road freeway most Highway Interetate	only first last fourth culmination longest highest same	first last only fourth third fifth minth same
Layer 26_1	bridge bridges metapher Bridges Bridge	bridges bridge structures brid bridge	charest jame Margine	jame wipes	Malta main Jordan Egyptian Romanian central	eec Untudent Begin streets hotline fax	ities TRGL
Layer 26_2	parliament structures bridges Parliament bridge	East out choke Exit panicked	symbolism metaphor	California California interetate Californ Maren Butter	UNESCO Brexit initialise bespoke ariet	Cumhurbakan	ly ted e
Layer 27_0	first last largest bridge longest most oldest latest	bridge bridges Bridge Bridges	last first suit highway bottleneck next road choke	last first weakest highway hardest bottleneck next road	bridge highway first California next largest last latest	first last only fourth culmination weakest highest same	first last fourth third bridge fifth second seventh
Layer 27_1	bridge bridges Bridge Bridges	bridges bridge Bridge bridge bridge	EXIT eait cuits (eait	exit EXIT exit exits	central Malta main Egyptian key Yemen Jordan	streets hotline ieres bridge cherry reason rise	TRGL ities
Layer 27_2	bridge bridge bridges Bridge structures	Exit out Exit out out	incident EXTRA incidents	California California Soattle Butter Seattle Californ	Tibetan Swiss Tibet arist Himal	Cumhurbakan	ly ted e cion lah
Layer 28_0	bridge longest largest first businst last object most	bridge bridges Bridge Bridge	highway sait bottleneck highways Highway last road exits	highway bettleneck Highway last highways soil read businet	California bridge Marin forry highway Oakland freeway	only first bridge last fourth culmination weakest same	first bridge last fourth third gateway highway fifth
Layer 28_1	bridge bridge Bridge Bridge	highway highways coast freeway roads road route	oult outs EXIT suit Exit	suit suits EXIV suit Euit suit	Jordan Memphis Nile Mississippi main Nicar	streets hotline ieres sang	TRGL h ities
Layer 28_2	bridge bridge bridges Bridge brid	Exit out Exit out out out	Saddan Moral Kuwait incident metaphor	Nile Kuwait Jordan Memphis	Slovenia Storen Croatia Manit Estonia Gael	Sy the	ly ted ments

Table 2: The top three singular vectors of the detached Jacobian for the layer outputs from Qwen 3 8B for the sequence "The bridge out of Marin is the" with the prediction [[only]]. Legend: "Bridge", "only", "highway", "exit", "Golden", "most".

		Sepat takes i	Sepat tokes 2	Japan token 3	Super token i	Sepan token S	Suport traken 6	Deput token 7
Laper 13_9	at only the alon that and	the that only at and most ralled either	probablement Guad alman Lass Yellow yaits	the at railed that only one	Nat Bar log loadwinds relievesaster brutal namage augry tell	Navillar Manathon ter Flower Brace Brazil Central Pam	- we set + seed of if	yerran
Laper 13_1	sh handful Khaq fame marlem pity mga	Ear GONE). Brt Green Brt	the and that the and if at	Empty Called Taj Input Fare Aqui Cr Fiel	the and that or or as	- maly one or at (/ ((?	(los) melatra socii imata lina	
Laper 11_2	irely issal (2):	0.3.1.2.4 mete 3		Hallowen gonna roller kiddon revegy neronitan Gonna getta	Assume = railed Called =	virtum delights (himselsle quaint triumple naturalist medieval	no facili el	solide el as
Laper 14_9	only that the at one either	only the that our at either only those	versio Bridge Good probablement transcrummahab Night	our only not at narbs tornorist sairide	Aleparta Cable HDMI rable Rades HDMI NevBar	mountain Mountain Hiking alpine Watershed Highland Hiking Mountain	us if at set railed	yeraras
Laper 14_1	the I relief were	Ear GONE ManLough Tota brus Seg Nova	the natist extremist that those treverist wonic anarch	Hank Manual Fair Quick Early Initial Alternatively Box		NASCAR Mega TikTok Walmari klussowyu Belani Urlan City	metatra (bus): Ima ionato ilenira ti	melikir an el
Laper 14_2	Tyl	0 1 5 2 4 3 revendes 6	and already after one X AL salp development	gona luko markandian faudir trek nevnitas skopper	tilatak Tiktak veldig isempable skissuare TikTak yon Beyoner	Eabimore Philadelphia Philadelphia skysemper Eabimore skysempon London ICAGO	ro I ainder la	
Laper 15_9	undy radied one probably either the very	only the one rither that called probably transportation	ionie Bridge bridge RQE Nope parate bridges Bridges	one unity probably railed airport transportation. highway contrader	Aiqueta aiquet Aiquet wastely Aiquet Incomp nathroad	Watershed Maple Reconstional Mountain Give Route Pine Glee.	ralled usly 1 mi	
Laper 15_1	frankism I relati imperialism ran politica	taw Neg taw North New Nevendo amon advance	one reference metal the complex terrorist supposed military	Blank Flight Can Quick Clock Onygon Cyclone Called	the that one railed those only not	Walmart AREAN airport Eshad NASCAE Nigorian Airport airports	(Soc) moletra lesa issaelo til	
Laper 15_2	11-	0 1 1 2 commutes 3 4 6	ralled Sept (Grand	Anime Economerer Isalis Scotie Apareola Gener Etry Gelfriand	Hydro Introduction Eddi Water Handbook Hotory Kirk Colde	maly timeless leabton (algor bike minimalist where	Empty partidas tion varshar grat	i m d
Laper 16_0	only one very first the route	only the one limit three two very	Ionair Bridge bridge Contrantal Greater Gilden bridge Ironic	sale route Ecute transportation one most seems very	Easts airport waterly roots detain Airport freeway sustleand	Reute Watershol watercolor Valley Pine hiking Reventional Hiking	make radied on make at	
Laper 16_1	tow North horsama New tangent Bitter Reg get Aug	North North tew ter largest laguou sugressates shortest	resetal sersic set military one likely very location	Flight Quick Are Fank Hank Onygen Air Ginger	the not located that one three only in	Treas sirport Mismi Walmart ritin NASCAE Nigorian rity	Clear) uphylline	melitie an el
Laper 15_2	Ast Appear	0.3.1.2.3 Provides 6.4	turbs Hydro Turbs Gro Mapa nordewed Mastel north	Anime Economerer Headphones Millenniah emericaal Gener Bratles	Hydra Northern Revised Introduction Route Nelson Percy Berkeley	only arathric timeless true rettage hippir wooden antique	BPC entropic uncharted disruption absighty	
Laper 17_9	undy first most our very tou the	only our first the two most very	ionie legesdary famed famous Genére namesake infamous Centranial	unity must one very ironic route probably not	airport blue Boute sable mate waterly tunnel sensie	Rente innie mat north gairway sonie reventional Valley	male on Brat North same at	
Laper 17_1	haroris tangent ten hornama north loophole	North north North ter terr supmanate supmanates beliess	unly very not seemic one likely military extremely	Flight adomain Funk Osygon Aer Osygon Turku Magosta	in the not unit there that located buildings	Treas cities UAE Keesit airport Nigorian military Xeedi	Ono	melikir an el
Laper 17_2). Paths Tab.).) Kndpoints Hosp	0 1 2 2 3 Provides 4 survendes	Turks TOC Pipeline Ariation City Route Water	Econometer Arme Anime maleus Attractive Descriptive eCommunic	Hydro Traffic Crossing John Water Highway Network Northern	alpine antique antique naturalist unity mountain never timeless	uncharted buov absighty HPC swirepic vortable	
Laper 14_0	undy first most our very tou the	only first our most fees that that	bridge Bridge bridges bunde counting Bridges	unity must one insale first very three	unity insule blue must break airport rocke galeway	ironic most only Route galernay arenic rountal watercolor	maly first own way inhastracture	
Laper 14_1	Implade New Lecutor FERTIXAccom/Token pound sobalk	below giant loophole layous swamp ManLougth Nee Locator	military unity reactal area likely heration brackes areas	Flight Fash Gina Magenta Acceler Going TeGo	in the located other od of sale state	Komit ritin UAE Nigotas Philippines Saudi Tenas Singapore	Lurgo gleiche sphyline name	solide sectionly aphyline t markets
Laper 14_2	(i.e.). Paths	0 Design 1 Style 3 Styling Provides	Water Pipeline Watercolor Pipelines Fountain Bracon Marathon Balloon	Acur Econometer mairon Apurceda hallowere Grackwindigheit Anime	Nelson Hydro Traffic Kirk John Water Brunck Margaret	alpine mountain Alpine skiers naturalist alpine antique renormation	uncharted extraple almighty brave disruption bratally	
Laper 19_0	undy Erst our most two highway the	only our link two that most arroad	Bridge bridge bridges Eridges Country Eridge Golden recoving	unity must one highway counted North Seat	highway andy North forway Highway resetal blue Route	subway Route mastal North East insale must Highway	mally first own bridge our way bighoup	name splitLength
Laper 19_1	Ring Cor Hoad reverse Rd Beacon	Bridge Bing Cor Road notorious iconic below Recense	reactal military area location areas most brack brackers	Park Flight ToGs Housely Thre Gina Learnhing	other in the empower government located buildings of	Kemit Souli Nigeian UAE Sugapor Oklahosa Florida Nigeia	gleiche Queste	solide senticulty t
Laper 19_2	versus rupani nggah inclusin advugado rompresa muphas	Croming Compre Leule Hard Indeas Cld Bridge	Ariarion Outreach Pipelines Turks Wastewater Figeline Brewing Bracon	Epilepay Watercolor Bathroom Ayurerda Gruze Bibliotera Economeror Mindhilurea	Bridge Traffic Berkeley Branch Grac Brid Md Nelson	alpine mentain wilderness Alpine mineral naturalist menutaine archaeological	non the which	
		one only most first second best	Bridge Bridge Bridges Crossing Bridges crossing bridging insuin	our andy much once highway first	our only must highway ours Highway rountal best	ment Highway Highland solvedon waterfront Cateragy Rente gateway	one only ones must largest first	name deletiment
Laper 20_1	Mon Mon Hor Tale Cor Trinity Bracon	Har Har Est Talk Mega Hainhow Cor	area one areas most foruntion only rounted military	ToGs Honortly Heuns thru Launching Alsh Flight	other anciest germanus constituent bedan dynasty highest oldest	Relaturian Circeland Minsk Euro Belatus Noida	valutastiste Pllanen	solidar t e y se
Laper 20_2	Zap Typical Ric Tab + 8 Ridge	Bridge Crossing Bridges Crossing Bridge Coast	Enegot minus Vertors Reinal spiral turbo spirals	Bhlistera Bridge Campground Bhlistheper Carm Brighton scuk BufOffart	Traffir Kirk Transit Bridge Linking Brid Md Lincoln	Jamairan Krepen Kasarbi Jamaira Southampton Nassan	non rosing via	-
Laper 25_0	bridge andy one must bighway first hout	heidge unity one must first highway heat	Bridge bridge bridge Bridge Bridge Croming crossing bridging	bridge Eridge highway bridges undy one must Highway	bridge highway Bridge Highway undy bridges must read	Bridge Bridge bridges freey Highway bighway Freey reusaling	bridge unity Eridge highway must Highway bridges rund	namer most simplest largest region best greatest strictest
	Bridge Bridge bridge bridges sungen Tol Tale							

Table 3: The top three singular vectors of the detached Jacobian for the layer outputs from Gemma 3 4B for the sequence "The bridge out of Marin is the" with the prediction [[Golden]]. Legend: "Bridge", "only", "highway", "exit", "Golden", "most".

3.6 The detached Jacobian as a conceptual steering operator

Steering vectors are a well-known technique for altering LLM outputs (Liu et al., 2023). Here we utilize the detached Jacobian from an intermediate layer for a phrase like "The Golden Gate", after the "Golden Gate Claude" demo (Templeton et al., 2024). The model predicts "Bridge", and this detached Jacobian is used as an operator to steer the continuation of a new phrase toward this concept. For a new input phrase, like "Here is a painting of the", the new embedding vectors \mathbf{x}_{new}^* are scaled by λ and multiplied by the detached Jacobian for the steering concept $\mathbf{J}_{\mathbf{L}}^+(\mathbf{x}_{steer}^*)$, and added to the layer activation $\mathbf{f}_{\mathbf{L}i}$ from the new input.

$$\mathbf{f_{Li}}(\mathbf{x}) = \lambda \cdot \mathbf{f_{Li}}(\mathbf{x_{new}^*}) + (1 - \lambda) \cdot \mathbf{J_{Li}^+}(\mathbf{x_{steer}^*}) \cdot \mathbf{x_{new}^*}$$
(16)

This intermediate representation is then fed back through the remaining layers of the network and the next token is decoded. The detached Jacobian must only be computed once for the steering concept, and therefore

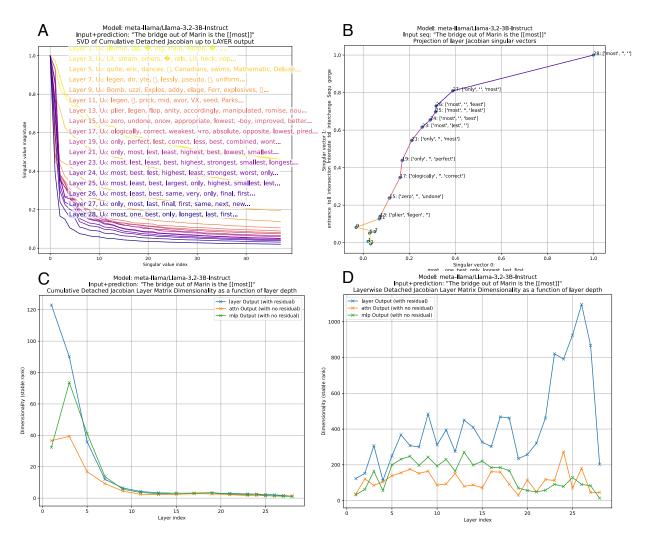


Figure 6: Since the transform representing the model forward operation is locally linear, we can also decompose each transformer layer as a linear operation as well. A) The singular value spectrum for the cumulative transform up to layer i. Note that later layers are lower rank than earlier layers. The top singular vectors of the later layers show a clear relation to the prediction of "most". B) The projection of the top two singular vectors onto the top two singular vectors of the final layer. The singular vectors of the first 10 layers are very different than those of the last layer, so the projections remain close to the origin. At layer 11, they begin to approach those of the output layer. C) A measurement of the dimensionality of the cumulative transform up to the output of each layer as the stable rank. Within each layer, the outputs of the attention and MLP modules (prior to adding the residual terms) can also be decomposed as linear mappings. The dimensionality decreases deeper into the network at each of these points, except for a slight increase for the attention and MLP module outputs in layer 3. D) The dimensionality of the detached Jacobian for the layer-wise transform at layer i for the layer output, as well as the attention module output and MLP module output.

this method is rather efficient. Table 4 shows how the detached Jacobian from an intermediate layer imposes the Golden Gate Bridge as the semantic output coherent with the rets of the input sentence, even when it is difficult to make a logical connection. Beyond demonstrating practical utility, the success of the steering operator provides validation that the detached Jacobian captures actual semantic representations.

4 Discussion

The detached Jacobian approach allows for locally linear representations of the transformer decode to be found for each input sequence, without changing the output. The intermediate outputs of each layer and sub-component are also exactly reproduced by the detached Jacobian function.

The detached Jacobian operation is only nearly exact at the specific operating point at which the matrices were computed by autograd. A short distance away in the input embedding neighborhood, the detached Jacbian will be extremely different because the manifold is highly curved. (Although local neighborhood validity is less applicable to LLMs which map tokens to embedding vectors, as inputs will only ever discretely sample the embedding space, and there are not obvious needs for exploring the local neighborhood beyond token embedding vectors). The manifold is not piecewise linear, but only has a linear equivalent exactly at the operating point, which can be found numerically for every input sequence.

Model	Layer intervention	Input sequence	Normal response	Steered response
Llama 3.1 8B IT	24 / 36	'I'm going to ari-	'I'm going to ari-	'I'm going to ari-
		zona to see the'	zona to see the	zona to see the
			Grand Canyon. I've	Grand Canyon,
			heard it's a must	and I'm planning
			see. I've also heard	to hike the Bright
			it's a bit of a trek to'	Golden Gate Bridge
				(I think that's the
				name of the trail) in
				the Grand Canyon.'
Qwen 3 8B IT	24 / 36	'Here is a painting	'Here is a paint-	'Here is a paint-
		of the'	ing of the same	ing of the Golden
			scene as in the pre-	Gate Bridge in San
			vious question, but	Francisco. The
			now the two peo-	Golden Gate Bridge
			ple are standing on	is one of the most
			the same side of the	famous bridges in
			building. '	the world. '
Gemma 3 12B IT	33 / 48	'I went to new york	'I went to new york	'I went to new york
		to see the'	to see the memo-	to see the 10th
			rial and museum. It	anniversary of the
			was a very moving	Broadway show,
			and emotional ex-	"The Golden Gate
			perience.	Bridge Bridge." It
				was a great show.'

Table 4: Detached Jacobian matrices as steering operators, pilot results with Llama 3.1 8B, Qwen 3 8B and Gemma 3 12B.

5 Conclusion

We have demonstrated that a number of open-weight LLMs can be made locally linear at inference, where a set of detached Jacobian matrices completely describes the forward computation, and that the predicted output embedding is unchanged from that of the original model. One potential safety application of this method is to examine large singular vectors for bias, misinformation or toxic content. The initial demonstrations of model steering may eventually be useful as a way to intervene and elicit safer outputs. Although this method requires intensive computation for decomposing single-token predictions, it should be possible to scale this investigation to well-known LLM datasets.

References

- Randall Balestriero and Richard Baraniuk. Fast jacobian-vector product for deep networks. arXiv preprint arXiv:2104.00219, 2021.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.
- Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. arXiv preprint arXiv:2403.02579, 2024.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.
- Bobby He and Thomas Hofmann. Simplifying transformer blocks. arXiv preprint arXiv:2311.01906, 2023.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.
- Zahra Kadkhodaie, Florentin Guth, Eero P Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representation. arXiv preprint arXiv:2310.02557, 2023.
- Sheng Liu, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. arXiv preprint arXiv:2311.06668, 2023.
- Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. arXiv preprint arXiv:1906.05478, 2019.
- Nvidia. Accelerating hugging face llama 2 and llama 3 models with transformer engine. https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/examples/te_llama/tutorial_accelerate_hf_llama_with_te.html, 2024.
- Noam Shazeer. Glu variants improve transformer. arXiv preprint arXiv:2002.05202, 2020.
- Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. (neurips), 2017. arXiv preprint arXiv:1706.03762, 10: S0140525X16001837, 2017.
- Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. Advances in neural information processing systems, 32, 2019.

Yaodong Yu, Sam Buchanan, Druv Pai, Tianzhe Chu, Ziyang Wu, Shengbang Tong, Benjamin Haeffele, and Yi Ma. White-box transformers via sparse rate reduction. *Advances in Neural Information Processing Systems*, 36:9422–9457, 2023.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. Advances in Neural Information Processing Systems, 32, 2019.

A Appendix

A.1 Locally linear approximate GELU

Gemma 3 uses the approximate GELU activation function. Below $\gamma = 0.44715$. Here is the derivation of the locally linear version of GELU used for Gemma 3 in the preceding analysis.

GELU
$$(\mathbf{x}) = \frac{1}{2} \mathbf{x} \left(1 + \tanh \left[\sqrt{2/\pi} \left(x + \gamma \mathbf{x}^3 \right) \right] \right)$$
 (17)

$$GELU_{LL}(\mathbf{x}) = \frac{1}{2}\mathbf{x}\left(1 + \tanh\left[\sqrt{2/\pi}\left(x + \gamma\mathbf{x}^{3}\right)\right]\right)|_{\mathbf{x} = \mathbf{x}^{*}}$$
(18)

$$GELU_{LL}(\mathbf{x}^*) = ([\frac{\partial}{\partial \mathbf{x}} GELU_{LL}(x)]|_{\mathbf{x} = \mathbf{x}^*}) \cdot \mathbf{x}^*$$
(19)

A.2 Code availability

Code is provided as a zip file (and will be made available on github).

A.3 Additional models

Local linearity for Deepseek R1 0528 Qwen 3 8B Distill, Phi 4, Mistral Ministral and OLMo 2 are shown on the following page. See Fig. A1.

A.4 Semantic Emergence in Transformer Layers

The first singular vectors of the detached Jacobians for the layer outputs from Llama 3.1 8B, Gemma 3 4B and Qwen 3 8B for the sequence "The bridge out of Marin is the" are shown below on the following page. These are expanded versions of Tables 1, 2 and 3 from the main text.

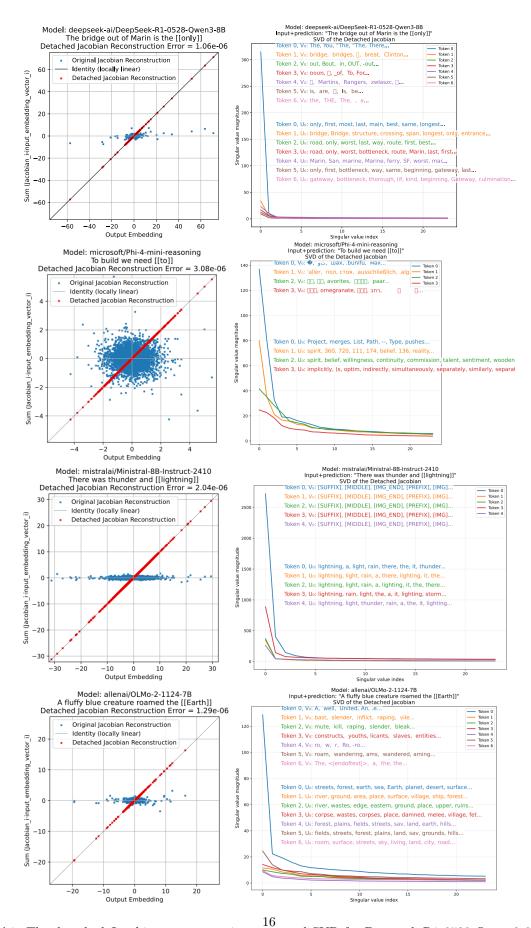


Figure A1: The detached Jacobian reconstruction error and SVD for Deepseek R1 0528 Qwen 3 8B, Phi 4 Mini 4B, Mistral Ministral 8B and OLMo2 7B.



Table 5: The top three singular vectors of the detached Jacobian for the layer outputs from Llama 3.1 8B for the sequence "The bridge out of Marin is the" with the prediction [[Golden]].

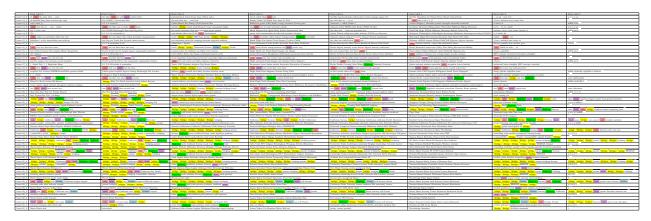


Table 6: The top three singular vectors of the detached Jacobian for the layer outputs from Gemma 3 4B for the sequence "The bridge out of Marin is the" with the prediction [[Golden]].

	Facest troken fit	Speed token 1	houst taken 3	Sanut token 3	Jacob token 4	Jacob token 5	local tribe 6
Laure 20 0	TRY NORMAL	Expert Colors 1	aledown euro slow consest nodeld		its connect normal whateveror Reflect research	Report Colores S	layerst carried smallest factors first weakest closest most
Laper 20_5	AUSE new blee metaphor JietFiles steet tat	corrier estructs Liter	villa fashion setatir deurem bias	whatsorver PORMAT OUESTION soraes	moneter coil than been nee		h H suNet F
Laper 20_2	ade flatter Fil mon inno and ren	by housing while	Estities campaign EventRus FILL		NVIC Locations level titular literal overse	artinadh H	a a spon r
Layer 21 0	TRY REGARD 4T	REGARD massage	ency slow exiting cothound fastest tight	rear slow factors hardest tight entiting slow desirates	Hierarchy ith whatevery Suit Firstly Role	shafter	first weakent largest fastest most longest easiest smallest
Layer 21 1	Great IGHL ocos UAGE est	corrier salety tax adors bestowed	Managed meds Chaires TORT Madaess machine Spare		here NYSE nation dellar	Guas instrate corresponding Producted each =BtConverter secified Franceworks	
Layer 21 2		tol Tonorrow goodless stash calar by syrun	HTTPS reinterest UTF REFER JSON Notifia		Proxy Preconditions contractual bright Eucode /euro Alamofre	II	N II M CANADA
Layer 22 0		succeptals Project	first hardest fastest exiting range	hardest factors first existing pagest Interstate	function whatevery whilst Miss Herardy	flash then	then remark
Layer 22 1	metaolog use DERP OBJC steet .we ISP	substr MROL bridge	heavy HIP oni Rockets TORT	somes	beer nation sim with	ess eachBitConverter ITIVE Guna restroirest	TRGL soNet
Laure 22 - 2	alk sie ool oor angen joon yn	autics ther recking	backstory weblog SVG JSON INCIDENT	ESA boost INSTANCE ECD	Theme Proxy bright Entities Encode ultimat /System D		
Laver 23 0	salones CONSTANTS witch Uncomment massage TRY	sectables beider largest easiest cally first contemiors	first hardest use fastest hisboom bettleseck	hardest first fastest majest touckest weakest unc ouickest	Whilst whilst role CONTRACT Whilst	dash then	thm redit dates still
Laver 23 1	metaolor une Derne makedelt OBJC	bridge Bridge bridge bridge	soraes WithError uni Timing presets Entering	somes WithError	California California bres nation lover	ly set ett ask of tea ty	TROL
Laver 23 2		autics laws	webbor COMPONENT annot metaphor	fullest PGA ESA levet JVM classic focus	specialized besocks commiss marOS UNESCO SSH EU initialize	,	It alion are ury certical in ance acook
Laver 21 0	first third last most largest fourth calmination	sartanhor largest crateroises easiest first hardest bridge gateway	hardest first easiest factest, most ones same	hardest first fastest easiest mast loagest same least	Whilst whilst Firstly Whilst crate	colonization first fourth largest highest third same seventh	first fourth third last largest highest fifth most
Layer 21 1	metaohor makeshift, wo REAK	brid bridge bridge bridge Bridge	soraes Localitanae WithError	const	entirety Malta main third favoured Yessen	st ly ath ett of ality hereby	it tim
Laper 21, 2		terr receip modes reselv mode	obraceroa rosa anaci setashor	ESA follow POA machine Scree ECD	hemody man OS committee manifoliant mount initialize UNDSCO Datases	to 19 and the to daily seemly	in tion
			hardest ones suit easiest first most fastest history			w.	first fourth third last largest highest most weakent
Layer 25_0		bridge bridges Bridge galenny		hardest factors easiest most ones weakest first rait	Whilst Firstly dimensions Mahmond	note abortions odd	
Laper 25_1	bridge bridges Bridge Bridges brid	bridges bridge Bridge Bridge	(sait sait exits elecuity sait	eit	Malta main entirety Jordan third Klopp Nicar	eer fax streets botline Untudent tats	ities it
Layer 25_2	bridges bridge Bridge bridge positionest	East pain jame	INCIDENT symbolism	JVM migrations CAB largest upcoming machine	specialised initialise UNESCO bespoke connexion mount arist macOS		e closs ted
Layer 25_0		bridge bridges metaphor gateway connecting	highway first exit ones last hardest roads	highway handest first weakest last road roads easiest	highway California highways road feeway saint Highway Intentate	eady first last fourth culmination longest highest same	first last undy fourth third fifth minth same
Layer 26_1	bridge bridges metaphor Bridge Bridge	bridges bridge structures held bridge	.chareet jame Margine	jame wipes	Malta main Jordan Egyptian Romanian central	eec Ostudent Begin streets hotline fax	ities TRGL
Layer 26_2	parliament structures bridges Parliament bridge	Ruit sait shoke Exit passished	symbolism metaphor	California California interetate Californ Massen Butter	UNESCO Resait initialise bespoke arist	Cumburbakas	ly ted e
Layer 27_0	first last largest bridge laugest most oblest latest	bridge bridges Bridge Bridges	last first rait highway bottleneck next road choke	last first weakest highway hardest hottleweck next road	bridge highway first California next largest last latest	first last only fourth culmination weakest highest same	first last fourth third beidge fifth second arresth
Layer 27_1	bridge bridges Bridge Bridges	bridges bridge Bridge bridge brid Bridge	EXIT suit cuits (suit	enit EXIT enit cuits	central Malta main Egyptian key Yenen Jordan	streets botline irres bridge cherry reason rise	TRGL ities
Layer 27_2	bridge bridge bridges Bridge structures	Exit out Exit suit suit	incident EXTRA incidents	California California Seattle Butter Seattle Californ	Tibetan Swiss Tibet arist Himal		ly ted e cion lah
Layer 28_0	bridge longest largest first businet last oldest suint	bridge bridges Bridge Bridge	highway out hettleweck highways Highway hat mad exite	highway buttleweck Highway last highways ext road busint	California bridge Maria ferry highway Oakland feeway	undy first bridge last fourth culmination weakest name	first bridge last fourth third gateway highway fifth
Layer 28_1	bridge bridges Bridge Bridge	highway highways coast frowway roads road roate	out outs EXIT out Exit	cuit suits EXIT cuit Enit cuit	Jordan Memphis Nile Mississippi main Nicar	streets hotline irres sang	TRGL h ities
Layer 28_2	bridge bridge bridges Bridge brid	Exit cuit Exit cuit cuit cuit	Saddam Moral Kuwait incident metaphor	Nile Kuwait Jordan Memphis	Slovenia Sloven Croatia Manit Estonia Gael	Sy the	ly ted meuts
Layer 29_0	bridge unity fourth last third longest fifth most	bridge bridges Bridge Bridges	only last first highway third highways said fourth	only last highway first businet buttleweck third hardest	California cody bridge first highway latest Oakland last	undy fourth last fifth third first seventh bridge	fourth third only fifth bridge first last seventh
Layer 29_1	bridge bridges Bridge Bridges	coast highway road driveway coastline roads highways feerway	cuite vait EXIT	cuite rait EXIT cuit	Jurdan Memphis Nile Frankfart Mississippi desert	only Only with which at cost only	TROL L
Layer 29_2	beidge heidges heidge structures beid structure	East wait Highway East	Saddam Moral Elvis metaphor incident	Nile Frankfurt Kunnit dosert	Slovenia Gael villagers Tibet Tibetan	checkpoint tunnel chances michannel crowing diabled tunnels	ly fully ments ted ser lab e
Layer 30_0	bridge most longest fourth third last unity fifth	bridge bridges Bridge Bridge	highway only bridge last first. Highway road highways	bridge undy highway last Highway foreway first busines	California Marin Bridge Oukland San California Ray	undy bridge fourth last fifth third first seventh	bridge fourth third fifth only first seventh lost
Layer 30_1	bridge bridges Bridges Bridge	coast freeway highway coastline road roads highways	heidge Bridge bridges beidge beid	bridge Maria underwater Bridge bridge public	Jurdan Memphis Mississippi Nile Missouri Nachville	sang irres streets Ostudent bridge	TRGL h ities
Layer 30_2	bridge structure structures bridges bridge brid	sail seab sailing Dermuda ship	Memphis Kumait Jordan Saddam Iowa	Nile Memphis Jordan Edmonton Kuwait	Gael Tibetan Tibet Nepal Kashmir Tanzania	bridges bridge crossing	ly rer e sign fully ted liness
Layer 31_0	bridge most only last longest first third largest	bridge bridges Bridge Bridge	only last highway fast bridge our Highway most	undy Maria lost highway bridge first California Highway	Marin California Sun Bay marine even bay coastal	only last first bridge third fourth fifth key	bridge third only fourth first last main fifth
Laver 31 1	coast airlines Interior airline interior Lua Speedsov	coast coastline coastal Coast Coastal roste	bridge Bridge bridges bridge underwater brid	Marin marine underwater bridge assatic man	Jordan Mississinni Menushis Nile Missouri Tennessee Jordan	bridge Bridge many streets bridges irres	TRGL h ities
Layer 31_2	bridge bridges bridge brid Bridge structure	ship sail sailing dock seab	Jordan Memphis Kumat Mississippi	Joedan Edmonton Nilv Memphis Cairo Kowait	Mara Finnish Finn Kashmir Gael	loidge leidges crossing	e by rer fully ted liness and sign
Laver 32 0	bridge most only first last longest third largest	bridge Bridge bridges Bridge bridge	only last first blabear most main route call	Marin colly lost highway route first most San	Maria San marin Mar California marine Bay mar	under last first main key most third subject	only main last first third key bridge subject
Layer 32 1	interior airline steam airlines Trail beest vacuum	most costal mostline route Coast Route beach	bridge man underwater connecting desk unblic member	Marin underwater marine man acceptic member Meer	Jurdan Mississiumi Nile Missouri Memohin 1889 main Cairo	under Only Vader under comelors by	ties TRGL Others
Laver 32 2	bridge bridge bridges Bridge brid Bridge	ship sail dock saling seals	Kweait Jordan Memohis Edmonton Mississioni Nile	Jordan Nile Edmonton Cairo Missouri	Honston Houston Birminsham Pittsburch Astron Chrysolet Harvey	bridge bridges crossing	e our ele v eon c or ein
Layer 33 0		brider Brider briders Brider mile	only last first hey main name most wait	Marin only lost San first lay same most	Marin marin San Mar Marina mar California marine	unity last first lavy subject third main final	cally subject third main first last key fourth
Layer 22, 1	planet interior cabin floors roots	coast coastline coastal Coast noute brack Coastal	sana subbit member library stations floating intervening deck		Jurdan Nile Mississiumi Menushis Cairo Missouri	own by about Vader about comelow	TROI rise
Layer 33 2		ship orbit aircraft sail yeard	Kurait Nile Education Saskatcherum Talea	Jordan Nile Edmonton Cairo Kurnit		wheels beider unite snow sky basy iron	est a ver fally the . d

Table 7: The top three singular vectors of the detached Jacobian for the layer outputs from Qwen 3 8B for the sequence "The bridge out of Marin is the" with the prediction [[only]].