Spectral Filters, Dark Signals, and Attention Sinks

Nicola Cancedda
FAIR @ Meta
ncan@meta.com

Abstract

Projecting intermediate representations onto the vocabulary is an increasingly popular interpretation tool for transformer-based LLMs, also known as the *logit lens* (Nostalgebraist). We propose a quantitative extension to this approach and define *spectral filters* on intermediate representations based on partitioning the singular vectors of the vocabulary embedding and unembedding matrices into bands. We find that the signals exchanged in the tail end of the spectrum, i.e. corresponding to the singular vectors with smallest singular values, are responsible for attention sinking (Xiao et al., 2023), of which we provide an explanation. We find that the negative log-likelihood of pretrained models can be kept low despite suppressing sizeable parts of the embedding spectrum in a layer-dependent way, as long as attention sinking is preserved. Finally, we discover that the representation of tokens that draw attention from many tokens have large projections on the tail end of the spectrum, and likely act as additional attention sinks.

1 Introduction

Large foundation models dominate the state of the art in numerous AI tasks. While we understand how these models work in terms of elementary operations, and black-box evaluations help characterize observable behaviours, we lack a clear understanding of the connection between the two.

There is a growing body of work providing insights into properties of model components, e.g. (Voita et al., 2019; Pimentel et al., 2020; Voita and Titov, 2020; Geva et al., 2022; Meng et al., 2022; Voita et al., 2023), as well as identifying and explaining fundamental phenomena, often with the support of simple models (Elhage et al., 2021, 2022; Olsson et al., 2022; Shai; Todd et al., 2023).

Most recent works assign a central role to the model's *residual stream* (RS) as the shared communication channel between model components.

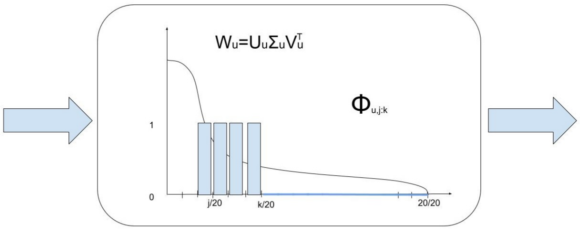

Figure 1: Spectral filters project signals exchanged between components onto selected subspaces as defined by the spectral decomposition of the vocabulary embedding and unembedding matrices of the model.

In this perspective, the probability distribution of a token is initialised from the projection of the embedding of the previous token through the unembedding matrix, and receives additive updates from attention heads and MLP components, each reading from the residual stream of the same or previous tokens. The role played by components is interpreted projecting their contribution on the probability distribution over vocabulary items, in what is referred to as the *logit lens* (Nostalgebraist; Geva et al., 2020). We extend this approach and introduce *logit spectroscopy*, the spectral analysis of the content of the residual stream and of the parameter matrices interacting with it. Equipped with this tool, we look at the part of the residual stream spectrum that is most likely to be neglected by the logit lens: the linear subspace spanned by the right singular vectors of the unembedding matrix with the *smallest* singular values. Drawing an analogy with "dark matter" in astrophysics, that interacts with light only indirectly, we dub projections onto this subspace *dark* parameters, features, activations etc.

We were motivated by the thought that LLMs could learn to use signals in the dark linear subspace to maintain global features responsible for long-range dependencies while minimizing their

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4792–4808
August 11-16, 2024 ©2024 Association for Computational Linguistics

interference with the next token prediction. We discovered instead that dark signals are instrumental to implementing the recently described phenomenon of *attention sinks* (Xiao et al., 2023), of which we provide a detailed account. We also show that the negative log likelihood of pretrained models can be kept low despite suppressing large swaths of the unembedding spectrum, as long as the dark signals required for attention sinking are untouched. Finally, we find a significant positive correlation between the average attention received by a token and the relative prevalence of dark signals in its residual stream, and find evidence that tokens uniformly receiving substantial attention are likely acting as additional attention sinks.

2 The LLaMa2 models

We chose LLaMa2 models (Touvron et al., 2023) as the object of our study as they were the most competitive models with open-access weights at the time we started this work. In particular we studied pretrained models (without instruction fine-tuning) with 7B, 13B, and 70B parameters.

Table 1 presents the key dimensional hyperparameters of these models. All LLaMa2 models share the same tokenizer, with 32,000 vocabulary items. Table 3, in Appendix A, summarizes the standard notation we use to refer to model components, and a high-level diagram of the architecture is in Figure 2.

We point out a few differences between LLaMa2 models and the original transformer architecture (Vaswani et al., 2017). RMSnorm (Zhang and Sennrich, 2019) is applied *before* every attention component, MLP component, and the final unembedding projection. RMSnorm includes a learned rescaling vector that is applied after the normalization proper. In all cases, we absorb this rescaling vector into the matrices downstream from the rescaling: this is mathematically equivalent and simplifies the analysis.

LLaMa2 models use SwiGLU activation functions (Shazeer, 2020). This means that MLP components have three parameter matrices instead of the more familiar two.

Finally, LLaMa2 models use rotary positional embeddings (Su et al., 2023). While important for the functioning of the model, their effect is independent of the phenomena we are focusing on.

3 Model parameter properties

Let $W_u = U_u \Sigma_u V_u^\top$ be the singular value decomposition of the vocabulary unembedding matrix (and similarly for W_e). Figure 3 shows the distribution of the singular values (SVs) of W_u for LLaMa2 13B. There is a single large SV[1] followed by a tail that declines only at the very end (the distribution for LLaMa2 7B and LLaMa2 70B is similar). The SV distribution of the W_e embedding matrix is also similar, but the top SV is only twice as large as the second one, with a longer "head" of relatively large SVs.

We use the adjective *U-dark* (*E-dark*) to characterize anything that happens in the linear subspace spanned by the 5% right singular vectors (RSVs) of W_u (W_e) with the smallest SVs, the dark basis.

We gain an initial insight into whether dark signals are exchanged by projecting rows and columns of parameter matrices on the dark bases. W_k, W_q, and W_v project the residual stream onto either the latent space used to compute attention scores or the latent space used to compute the attention output, so they "read" from the residual stream; similarly W_1 and W_3 map from the residual stream onto the activation layer of the MLP components. We project the columns of these matrices on the RSVs of W_u to estimate and visualize their aptitude to read from the dark subspace[2]. Conversely, W_o and W_2 map from latent spaces back into the residual stream, therefore we project their rows to check their aptitude to write into the dark subspace. We computed and plotted the norms of the d vectors of dimension d_h or d_m obtained with these projections, e.g.:

$$j_i = \begin{cases} ||(V_u^\top)_i W_y||_2, & y \in \{k, q, v, 1, 3\} \\ ||(V_u^\top)_i W_y^\top||_2, & y \in \{o, 2\} \end{cases} \tag{1}$$

We discovered a great variety in where, in the bases formed by the RSVs of W_u and W_e, model components are equipped to read from and write into. Figure 4 gives a sense of such variety.

The projections of MLP matrices also clearly indicate that some MLP components can write mostly into the dark subspace (see Fig.5 for the projection of 13B/L0/W_2 on W_u).

[1] Projecting unembeddings on the first singular vector shows that it is highly representative of token frequency.

[2] We adopt the convention that input vectors are row vectors and are multiplied by parameter matrices on the right.

Model	Layers	RS dim	MLP dim	Attn heads	KV heads
LLaMa2 7B	32	4096	11008	32	32
LLaMa2 13B	40	5120	13824	40	40
LLaMa2 70B	80	8192	14336	64	8

Table 1: Dimensional hyper-parameters of the LLaMa2 models. The 70B model uses grouped-query attention, with 64 attention heads sharing 8 key and value matrices.

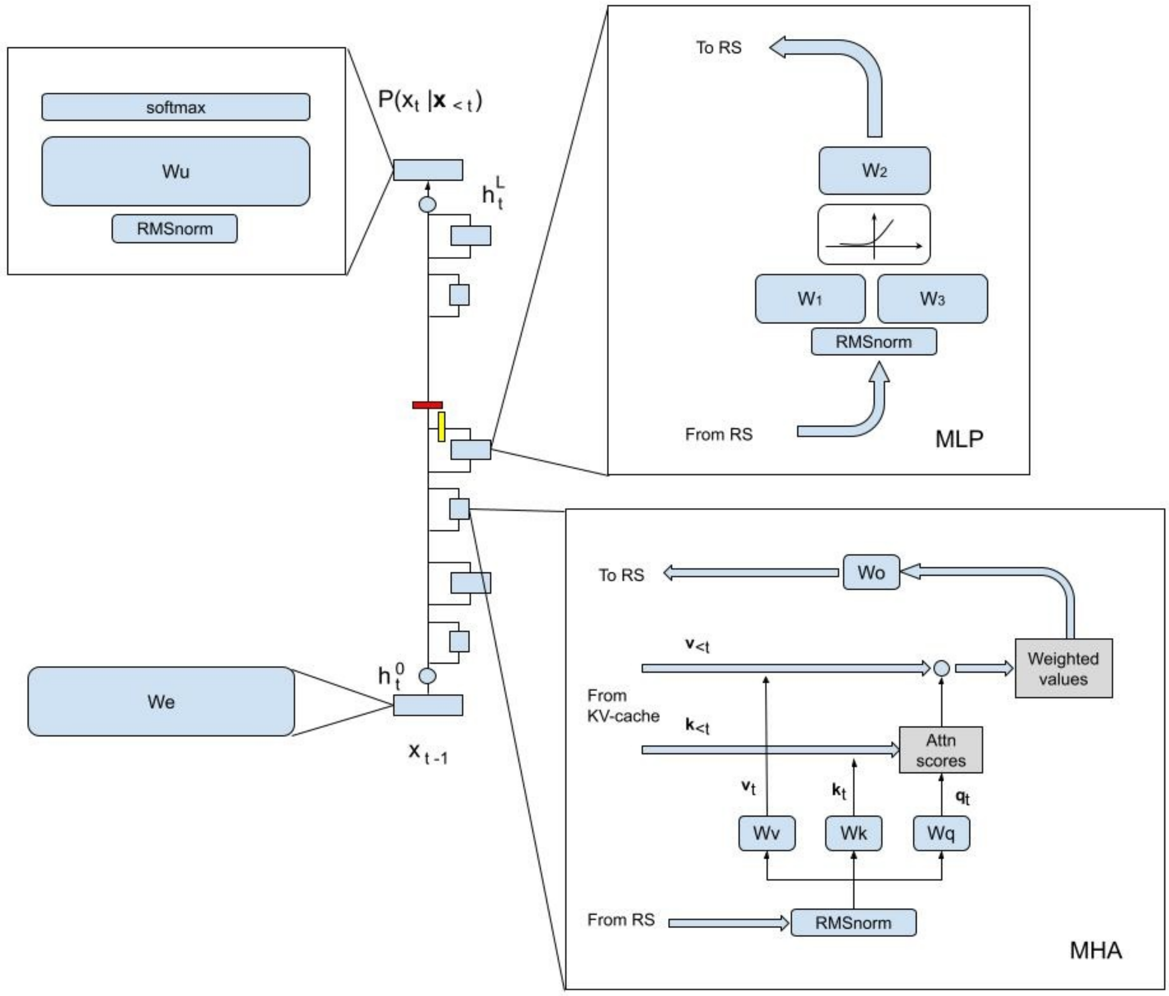

Figure 2: Schema of transformer-based autoregressive LMs. Initialized (bottom) with the embedding of x_{t-1}, the residual stream receives additive updates by attention and MLP components layer after layer. At the end a projection through the unembedding matrix W_u and a softmax yield the probability distribution from which the next token is sampled. The red and yellow blocks show the two positions where we applied spectral filters (Sec. 4).

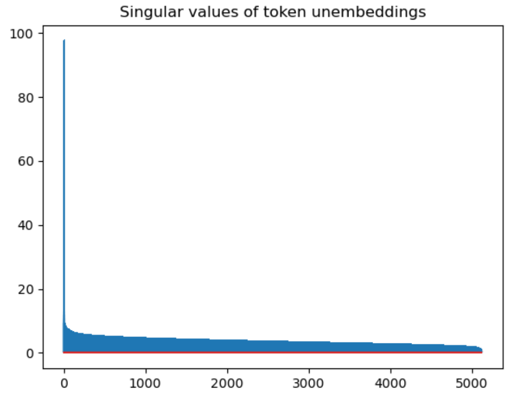

Figure 3: Distribution of the singular values of the unembedding matrix W_u of LLaMa2 13B. The *U-Dark* subspace is the one spanned by the last 5% right singular vectors.

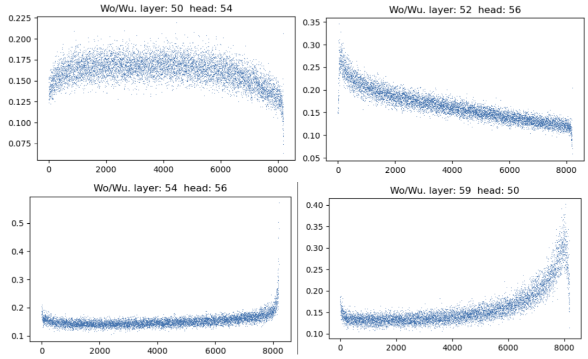

Figure 4: The projections of four W_o matrices of LLaMa2 70B on the RSVs of W_u. Different heads are equipped to write into different subspaces, with some targeting the dark subspace.

There are therefore components that are equipped for communicating through the dark subspace. In the following section we explore if such communication actually takes place and how it manifests itself.

4 Spectral filtering

One possible explanation for the existence of components that communicate through the dark subspace is that they are useless, and the model learned to divert their output to subspaces with little bearing on the vocabulary logits. To see if this is the case, we perform a series of experiments, where we patch some of the intermediate representations by projecting them onto the RSVs of W_u and W_e with largest singular vectors, therefore removing dark signals. We measure the average negative log-likelihood (NLL) of tokens in a sample of prompts: if dark signals are irrelevant noise then we expect NLL to be largely unaffected when they are filtered out.

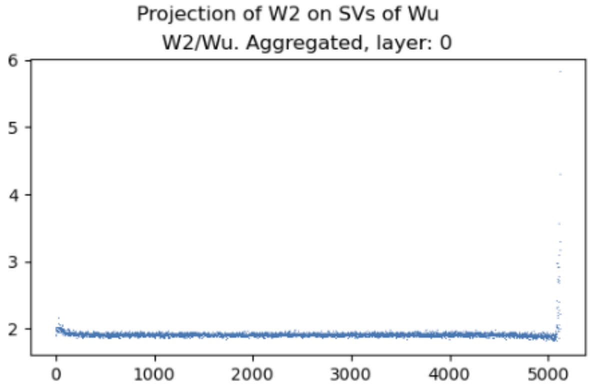

Figure 5: The projection of the rows of W_2 at L0 of LLaMa2 13B on the RSVs of W_u. Note the large values at the very right end of the spectrum, indicating the ability to write in the U-Dark space.

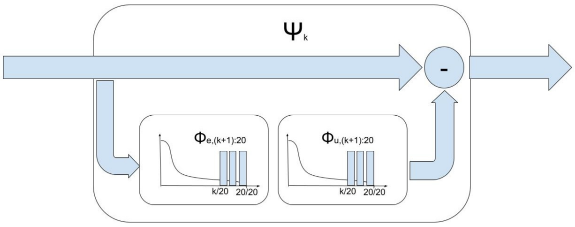

Figure 6: Ψ filters project vectors onto subspaces that are dark according to both the embedding and the unembedding matrix decomposition.

We split the singular vectors of W_u into 20 *bands*: $\{V_{u,1}, \ldots, V_{u,20}\}$, of cardinality $d/20$, and similarly for W_e. $V_{u,1}$ contains the 5% of singular vectors with largest singular values, $V_{u,2}$ the next 5%, and so on. Let $V_{u,j:k}$ be the matrix obtained concatenating $[V_{u,j} \ldots V_{u,k}]$, and (Fig. 1) let $\Phi_{u,j:k} = V_{u,j:k} V_{u,j:k}^\top$ (similarly for W_e). We can then project any d-dimensional vector onto the U-dark (E-dark) space by multiplying them by $\Phi_{u,20:20}$ (resp. $\Phi_{e,20:20}$).

We form a hierarchy of nested filters $\Phi_{u,1:k}$: multiplying a vector by $\Phi_{u,1:k}$ projects it onto the subspace of the $kd/20$ "least dark" dimensions. We also define filters that combine singular vectors of W_u and W_e (Fig. 6):

$$\Psi_k = (I - \Phi_{e,(k+1):20} \Phi_{u,(k+1):20}), k = 1, \ldots, 19 \tag{2}$$

and we set $\Psi_{20} = I$. Multiplying a vector by Ψ_k filters away projections onto subspaces of both W_e

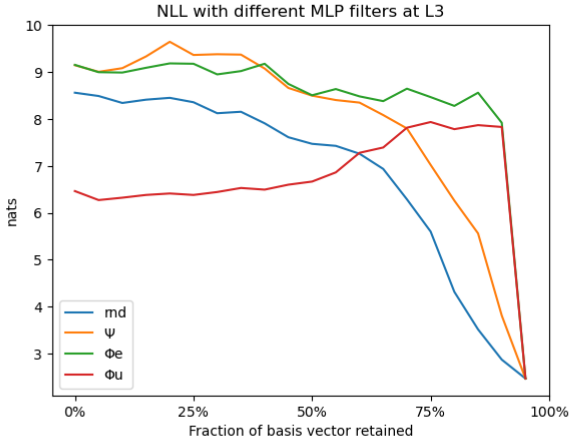

Figure 7: The effect of filtering 13B/L3/MLP with the filters defined in Section 4. 'Rnd' indicates filtering projection on subsets of a random orthonormal basis, for reference.

and W_u that get darker as k grows. Ψ_{19} filters out only vector components in the 5% dark subspace of *both* W_u and W_e. These definitions make $\Phi_{u,k}$, $\Phi_{e,k}$, and Ψ_k comparable when looking at the fraction of kept/discarded singular vectors, although, since it discards a double projection, Ψ_k retains more information. We create a dataset (ccnet-405) of 405 prompts (13,268 tokens) in English, from CCNET (Wenzek et al., 2019). We ran inference on ccnet-405 applying spectral filters to the output of MLP components, one by one (Fig. 2 (yellow box)) and measured the average NLL. We similarly measure NLL when filtering the RS after the contributions of a given target layer have been added to it (Fig. 2 (red box)). See App. B for implementation details.

4.1 Discussion

When filtering MLP layers of LLaMa2 13B, the effects of masking L0 and L3 dwarf all others.

When filtering the output of the MLP at L3 (Fig. 7) the loss remains poor until the last 5% of the spectrum is included for both the Φ spectral filters. The fact that the Ψ curve is always above the random filter indicates that this MLP exerts strong influence operating in the dark subspace. See Appendix C for a discussion of 13B/L0/MLP.

Figure 17 in App. C shows that MLP components with a similar propensity for writing dark signals exist also in LLaMa2 7B (L1) and 70B (L2 and L8).

We also look at what happens to the negative log-likelihood when filtering the residual stream after a given layer (Figures 2 (red box) and 8). Even a mild bottleneck after L3 results in a significant increase in NLL, in line with the observation that the MLP in L3 writes in the dark space. If we filter only signals that are dark to both W_e and W_u (Fig. 8(b)), we see bottlenecks continuing to be more harmful than random direction removal (Fig.8(a)) up until L20-25.

Dark signals exist and play an important role in achieving low perplexity, but what is this role?

5 Attention sinks

(Xiao et al., 2023) describe *attention sinks*: the special Beginning of Sentence (BoS) 'Token 0' receives a disproportionate amount of attention.[3] This happens because often an attention head should not activate in a given context, but the normalisation in the attention scores forces a constant amount of attention to be distributed to previous tokens. The model therefore learns to sink excess attention by allocating it to the BoS token and making allocating attention to the BoS token inconsequential, i.e. a 'no-op'.

Figure 9 shows how the norm of the Token 0 residual stream progresses over layers, and the contributions from MLPs and MHAs components. We plot the overall norm, and the norms of the projection on the U-Dark space and of the projection on its orthogonal complement (*U-Light*).[4]

After an initial phase of input enrichment that apparently does not need an attention sink, the MLP at L3 blasts off a vector of large norm and almost completely U-dark. This vector acts as an attention collector for heads in need of a sink according to the mechanism described in detail in Appendix D, and is kept around until the last few layers, where the combined action of MLPs and attention heads first erases it and then replaces it with the vector that encodes the probability distribution of the model over generation-initial tokens.

We confirmed that dark signals are primarily used to sink attention with an additional experiment (Fig. 10). We apply spectral filters at the exit of a layer but, rather than suppressing the filtered

[3] Xiao et al. (2023) include the first four tokens in their operational definition of attention sinks, for achieving effective streaming decoding. In our observations the BoS token is by far the one attracting the most attention.

[4] Note that the composition of the residual stream of Token 0 is always the same, irrespective of the prompt, because Llama 2 models initialize it with the embedding of the special <BoS> token, and since they are autoregressive there is no context that could influence it.

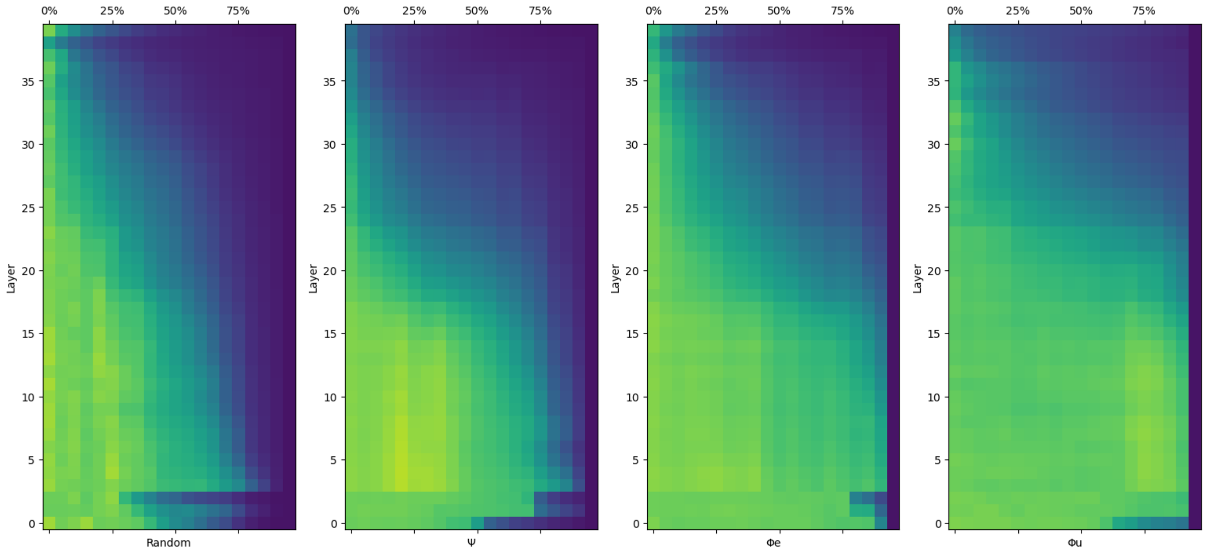

Figure 8: NLL of LLaMa2 13B on ccnet-405 when filtering the residual stream after a given layer (vertical axis), retaining an increasing number of SVs (horizontal axis) of (c) W_e or (d) W_u. (b) filters from the residual stream only its double projection onto both the W_e and W_u dark spaces. (a) shows, for comparison, the effect of adding more and more dimension in a random orthogonal base. See Fig.19 for similar heatmaps for LLaMa2 7B and 70B.

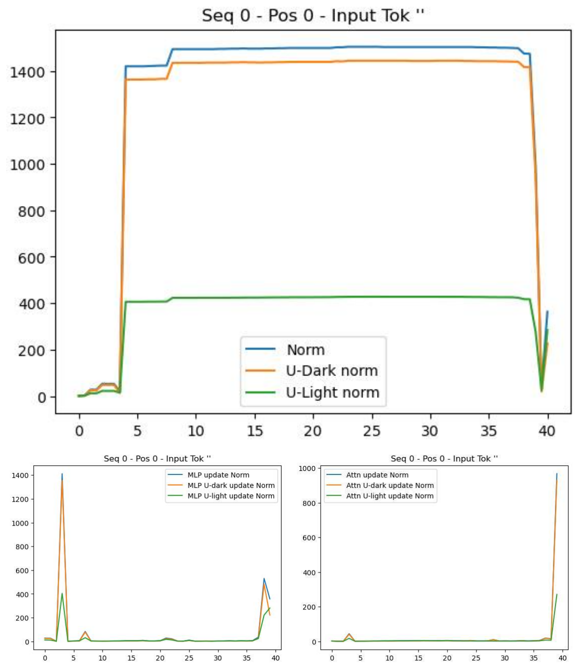

Figure 9: (Top) The composition of the RS of the BoS token for LLaMa2 13B as a function of the layer. (Bottom left) The norms of the contribution of MLP layers to the BoS RS. (Bottom right) The norms of the contribution of Multi-Head Attention components to the BoS RS.

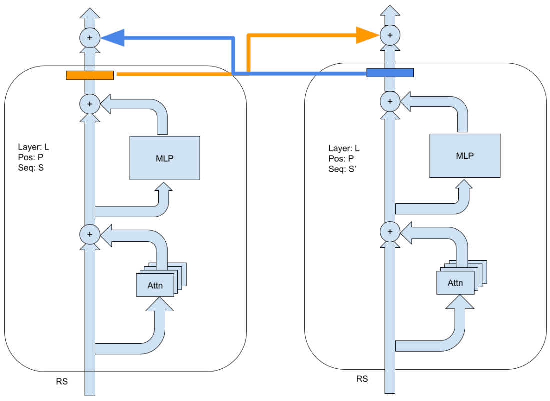

Figure 10: Shavings swap experiment: rather than being suppressed, the subspaces blocked by spectral filters are swapped with the corresponding ones from a token at the same position but in a different sample. This swap perturbs all residual streams except the one for Token 0, since this is identical for all samples due to the autoregressive nature of the model.

component, we add it into the corresponding residual stream of a different sample, at the same token position and layer. Since the content of the RS of the BoS is the same irrespective of the input, this procedure leaves it intact, while perturbing all other RSs.

The resulting NLL heatmap is in Fig.11 (left). Unlike in those in Fig. 8, there is no step-decrease in NLL when the last 5% of singular vectors is added. Conversely, if we apply spectral filters

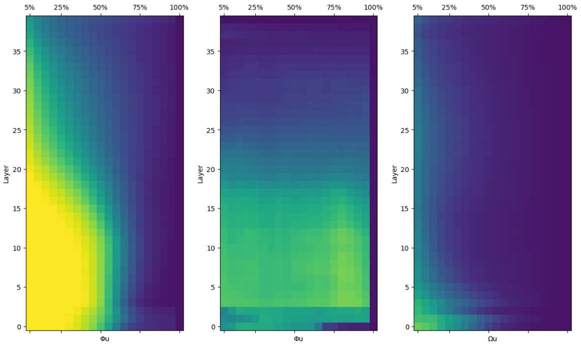

Figure 11: NLL of LLaMa2 13B on ccnet-405 when filtering the residual stream after a given layer (vertical axis). (Left:) Swapping the filtered vector components between RS at the same layer and token position, but in a different sample, perturbing all RSs except the BoS one. (Middle:) Filtering only the residual stream of the BoS token. (Right:) Applying the sink-preserving spectral filters $\Omega_{u,k}$ to a section of the residual stream of LLaMa2 13B right after a layer.

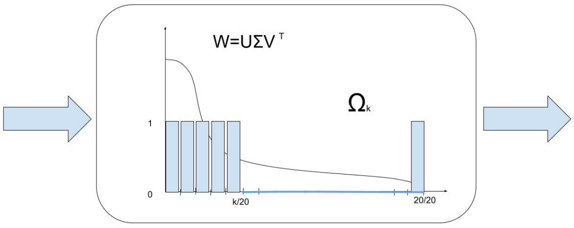

Figure 12: Ω filters project vectors on the head of the spectrum, but also on the tail end to preserve attention sinking.

only to Token 0, the step-decrease is clearly visible (Fig.11 (middle)). We conclude therefore that the primary function of the dark subspace is to enable the crucial attention sink mechanism.

6 Sink-preserving spectral filters

The finding that dark signals are essential to enable attention sinks leads to the question: what would be the impact of spectral filters that preserve dark signals but filter away one or more bands before them? Let $\Omega_{u,k}$ be a new family of spectral filters (Fig. 12):

$$\Omega_{u,k} = ([V_{u,1:k}; V_{u,20:20}])([V_{u,1:k}; V_{u,20:20}])^{\top} \tag{3}$$

Fig. 11 (right) shows that low values of NLL are achieved even when masking a significant number of components. As an example (Fig. 13) NLL grows only from 2.47 to 2.74 when suppressing 25% of the SVs by applying $\Omega_{u,14}$ after L12. We

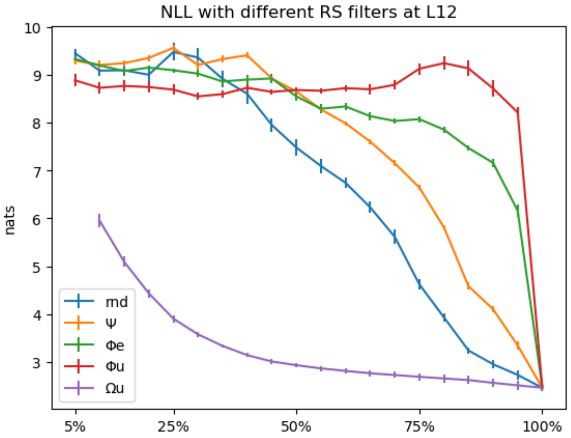

Figure 13: NLL by number of retained dimensions when applying different spectral filters at L12. Error bars indicate three standard deviations over five repeats with samples from the same source and of the same approximate size as ccnet-405.

validated these results by repeating the experiment with a sample of code from the DeepMind Code Contest dataset (Li et al., 2022) consisting of 117 prompts (14,641 tokens), Fig. 20, and with LLaMa2 7B and 70B (Fig. 19).

6.1 The effect of inhibiting attention sinks on generation quality

So far we have looked at the effect of spectral filters on the negative log-likelihood of prefixes that were fed as prompts to our models. It is interesting to also observe the impact on generated continuations when applying these same filters.

Layers 2-4 are the most fragile to random and Ψ filtering (Tab. 2), while generations remain coherent with Ω filtering irrespective of where the filter is applied, when 10-20% of the singular vectors are suppressed. We notice that the application of Ψ filters often results in the model entering repetitive patterns. This is consistent with the possibility that attention heads largely copy representations from the RSs of previous tokens, and inhibiting attention sinking results in over-copying. More examples can be found in App. E.

7 Dark Signals and Attention Bars

Inspecting attention matrices in our sample we see that there are frequently a few tokens that also receive a large amount of attention, giving rise to characteristic *attention bars* (e.g. Fig. 14). We hypothesized that such bars could correspond to tokens whose residual streams are similar to that of

Prompt: If you're interested in Grizzly Bear viewing in and around the northern Gulf of Alaska, consider a chartered flight with Trygg Air Alaska based out of King Salmon with service to and from Anchorage.
$\mathbf{\Psi}$**(90%)**: The EA, a GS, the E, A, for a total, the GS, the E, the S, the E, the S, the 15, the 16, the 15, the 16, the 15, the 17, the
Rnd (90%): What is the most beautiful thing that has ever been? What is the most beautiful thing that has ever been? It is not about the most beautiful thing that has ever been. The most beautiful thing that has ever been. It is not about the most beautiful thing that has ever been. It is not
$\mathbf{\Omega_u}$ (90%): The Grizzly Bear is one of the largest land carnivores in the world. Males can reach a weight of 800 pounds and females can reach a weight of 400 pounds. The Grizzly Bear is one of the most powerful predators in North America

Table 2: Sample generations by LLaMa2 13B with spectral filters applied right after L3.

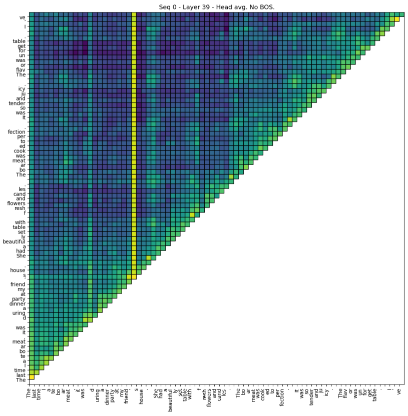

Figure 14: Attention matrix at L39 averaged across heads for a sample. Brighter columns correspond to tokens that are attended by many following tokens (LLaMa2 13B, Token 0 is removed for readability).

Token 0, and therefore become additional attention sinks.

We define High-Mean Low-Variance (HMLV) token-layer pairs those with mean received attention from subsequent tokens above $\tau_\mu = .018$ and variance below $\tau_\sigma = .01$. We selected thresholds manually so that most HMLV tokens appear as "attention bars" in attention matrices[5]. This heuristic yields 12,236 HMLV token-layer pairs for LLaMa2 13B on ccnet-405 out of a total of 404,748 token-layer pairs for the same tokens and layers considered. We ignore the BoS token, the last 4 tokens of each sequence, and layers 0-3, since the attention sink is not in place in the 13B model yet.

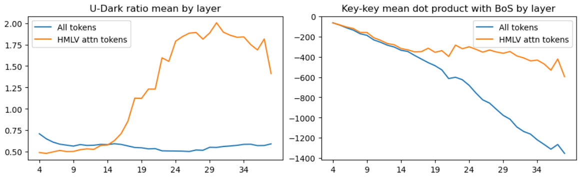

Figure 15: Mean U-Dark ratio (left) and key-key dot products with the BoS RS (right) for high-mean low-variance attention tokens and across all tokens, by layer.

We define the U-Dark ratio as a measure of the prevalence of U-Dark signals in a representation:

$$udr(h_t^l) = \frac{||\Phi_{u,20:20} h_t^l||}{||(I - \Phi_{u,20:20}) h_t^l||} \tag{4}$$

i.e. the norm ratio between the projection of the residual stream onto the U-dark space, and the projection on its orthogonal complement.

We also define the *key dot-product* as the average dot product between the keys projected through matrices W_k:

$$kdp(l) = \frac{1}{|H||T|} \sum_{h,t} \langle h_t^l W_{k,l}^h, h_0^l W_{k,l}^h \rangle \tag{5}$$

Intuitively, this value quantifies the extent to which a residual stream appears similar to the one of Token zero, as far as attention allocation is concerned.

We measured the average U-Dark ratio and key dot product for HMLVs, and contrasted them with the same statistics taken over all token-layer pairs. Results are in Fig. 15.

Starting around L13 the U-Dark ratio of HMLVs grows considerably, whereas it remains constantly below .75 for the overall population. At the same time, the gap in key dot product also grows, as the kdp for the overall population steadily decreases, while that for HMLVs stabilizes. Analogous plots for the LLaMa2 7B and 70B models show different profiles, but an even more marked difference between HMLVs and the general population (Appendix F). These measures support the hypotheses that HMLVs are token-layer pairs that appear very similar to Token 0, as far as attention allocation is concerned, and therefore become additional attention sinks.

8 Related work

In this work we adopt the framework introduced in (Elhage et al., 2021), which highlights the central role played by the residual stream as working

[5] Empirically, this heuristic appears more accurate for layers from L15 on. It might be better, but more laborious, to define layer-specific thresholds.

memory and communication channel across components, and suggests the presence of subspaces with differentiated functions. The *logit lens* was first introduced by (Nostalgebraist), and has been since used and extended in a number of interpretation studies, including (Dar et al., 2022; Geva et al., 2022; Belrose et al., 2023; Din et al., 2023; Katz and Belinkov, 2023; Voita et al., 2023). Of these, Dar et al. (2022) analyse models projecting parameter matrices onto the vocabulary space, an approach we are also following in Section 3.

Our hypothesis that the subspace most orthogonal to the vocabulary representation could encode non-sparse latent features that should not interfere with the distribution over the immediate next token appeared supported by (Elhage et al., 2022), where they show that, in basic models, important non-sparse features tend to be assigned principal components rather than being in superposition with sparse features. (Sharkey et al., 2022) showed that sparse autoencoders can recover a ground-truth assignment of features to linear subspaces by means of sparse autoencoders in an artificial and simplified setting and a 31M parameter LLM, with $d = 256$; (Cunningham et al., 2023) apply a similar method to larger Pythia 70M and Pythia 410M models. Our results with Ω filters with the much larger LLaMa2 models indicate that the residual stream could be used at less than full capacity, at least for some of the layers, raising the question of what model of superposition would best apply to them.

The concept of *Attention Sinks* was introduced by Xiao et al. (2023), who noticed that performance of streaming models dropped abruptly once the first tokens moved out of the attention sliding window context. We take here the next step in explaining how this process works, showing how it happens thanks to a very limited and specific portion of the spectrum.

Sharma et al. (2023) show that it is possible to preserve and even improve the performance of the GPT-J (Wang and Komatsuzaki, 2021) model while dropping large portions of component matrices based on the SVD decomposition of the matrices themselves. We also focus on spectral decomposition, but using the spectrum of the unembedding matrix, in accordance with the logit lens approach and the intuition that it represents a common ground for all the components in a model.

9 Conclusions and Future work

A better understanding of the inner workings of transformer models is important to find strategies to make them safer. In this work we explored a novel way to look at transformers, extending the *logit lens* approach into *logit spectroscopy* by introducing *spectral filters*. We explored the hypothesis that dark signals could be used to maintain global features while minimizing their interference with the next token, but we discovered that the main role of dark signals is in enabling attention sinking. We reconstructed how attention sinking works in LLaMa2 models, and we showed that, as long as attention sinking is preserved, they still achieve low negative log-likelihood even when significant portions of the unembedding spectrum are suppressed. Finally, we found a positive correlation between the attention received by a token and the relative prevalence of dark signals in its residual stream, especially in the upper layers.

The results on spectral filtering with dark signal preservation, combined with the observation that transformers are invariant to basis changes in the residual stream, suggest that it could be possible to first "canonicalize" a model so that its residual stream dimensions match the V_u columns, and then compress away from components dimensions that are low in the spectrum but are not used for attention sinking. We consider such a form of *spectral compression* an interesting direction for future work.

Finally, while our exploration of the connection between “attention bars”, prevalence of dark signals, and similarity with the Token 0 residual stream suggests that attention bars are additional attention sinks, increasing the granularity of the spectral bands in the dark subspace could lead to additional discoveries.

10 Limitations

We limited our attention to the LLaMa2 family of models, and only to the pretrained models prior to their instruction and safety fine-tuning. It is possible that other models do not realise attention sinks in the same way, or that they might not need them.[6]

The number of conditions tested in our detailed spectral filtering experiments meant that we could use only a text sample of limited size. While we did

[6] For example by using *off by one* softmax (Miller).

repeat some of the experiments using a code sample, we cannot exclude that there could be some differences when using samples from drastically different distributions, e.g. in languages using different scripts.

11 Ethical Considerations

While our aim in studying the behaviour of LLMs is to make them more transparent, controllable, and trustworthy, it is conceivable that increased understanding could also benefit malicious agents intending to undertake harmful actions. Besides these generic considerations, we cannot see problematic ethical issues arising from this work.

Acknowledgements

We thank the Transparent LLM team of Meta, and in particular Lena Voita, Yihong Chen, Javier Ferrando, for important suggestions and insightful discussions. We are also grateful to the anonymous reviewers for their constructive criticism and suggestions.

References

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2022. Analyzing transformers in embedding space. *arXiv preprint arXiv:2209.02535*.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. 2023. Jump to conclusions: Short-cutting transformers with linear transformations. *arXiv preprint arXiv:2303.09435*.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

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. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Shahar Katz and Yonatan Belinkov. 2023. Interpreting transformer's attention dynamic memory and visualizing the semantic information flow of gpt. *arXiv preprint arXiv:2305.13417*.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Evan Miller. Attention is off by one. Www.evanmiller.org/attention-is-off-by-one.html.

Nostalgebraist. interpreting gpt: the logit lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *Transformer Circuits Thread*.

Tiago Pimentel, Josef Valvoda, Rowan H Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622. Association for Computational Linguistics.

Adam Shai. Transformers represent belief state geometry in their residual stream.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. In *AI Alignment Forum*.

Pratyusha Sharma, Jordan T Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *arXiv preprint arXiv:2312.13558*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, page 127063.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

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

A Notation

Table 3 summarizes the notation used throughout the paper.

B Implementation details

Spectral filtering experiments were run on up to 8 NVIDIA A100 GPU servers with 80GB of memory each, the slowest (for the 70B models) taking up to 4 hours to complete. They were implemented modifying the public Llama inference code[7]. Inference used nucleus sampling with the default parameters (top-p = .9, T = .6).

C Additional spectral filtering experiments

When filtering the output of the MLP of LLaMa2 13B at L0 with filters $\Phi_{e,1:k}$ log-likelihood remains poor (around 9) (Fig.16) until about 90% of W_e's RSVs are retained. This indicates that the MLP in L0 exerts its effect by writing mostly into an E-dark subspace. The fact that the NLL recovers steadily starting from when about 40% of the RSVs of W_u are retained, and that the curve for the Ψ_k filters is close to the random masking, show that this MLP writes mostly in the target unembedding space.

Figure 17 plots the NLL loss when applying spectral filters to the MLP at L1 in LLaMa2 7B, and at L2 and L8 in LLaMa2 70B. The loss profiles of component 7B/L1/MLP (left) are similar to those of 13B/L3/MLP: it exerts its influence in a subspace that is dark to both embeddings and unembeddings, and indeed it is the MLP that creates the attention sink vector in the Token 0 residual stream. 70B/L2/MLP and 70B/L8/MLP are the components responsible for the attention sink in 70b (Fig. 18). Unlike with 7B and 13B, the attention sink vector in Token 0 is U-Dark but not E-Dark, and is formed in two steps at L2 and L8.

Figure 19 shows NLL loss heat-maps for LLaMa2 7B (top) and 70B (bottom) when spectral filters are applied to the RS at the layer on the vertical axis.

Figure 20 displays similar heat-maps for LLaMa2 13B, using a sample of code fragments from the solutions of the DeepMind Code Contest dataset rather than ccnet-405.

[7] https://github.com/facebookresearch/llama

V	Vocabulary
L<X>	Layer X
H<Y>	Head Y
d	Model dimension
d_h	Head dimension
d_m	MLP hidden layer dimension
$h_t^l \in \mathcal{R}^d$	Intermediate repr. for Token t at layer l
W_z, $z \in \{1, 2, 3\}$	MLP matrices ($W_1, W_3 \in \mathcal{R}^{d \times d_m}$), $W_2 \in \mathcal{R}^{d_m \times d}$
W_z, $z \in \{k, q, v, o\}$	Attention matrices ($W_q, W_k, W_v \in \mathcal{R}^{d \times d_h}$, $W_o \in \mathcal{R}^{d_h \times d}$)
W_z, $z \in \{e, u\}$	Embeddings and unembeddings ($\in \mathcal{R}^{\|V\| \times d}$)
$V_y, y \in \{e, u\}$	right singular vectors of emb. and unemb. matrices

Table 3: Notation for referring to model components.

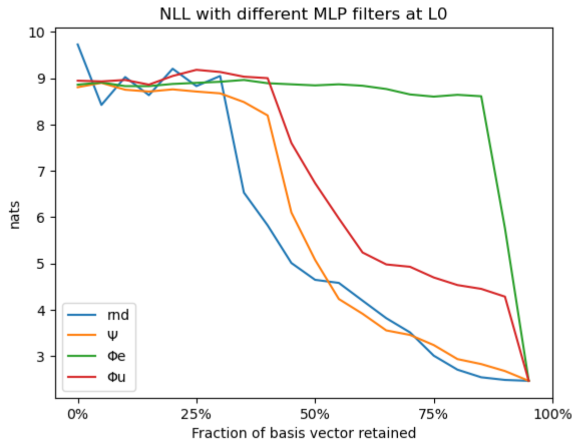

Figure 16: The effect of filtering 13B/L0/MLP with the filters defined in Section 4.

D Attention sink mechanics

In this appendix we offer a more detailed account of the attention sinking mechanism described in Section 5.

Consider the case when a given head H at layer $l \leq L$ is irrelevant when predicting a token t, perhaps because H specializes in a different kind of contexts, or in text from a different distribution (e.g. code, a different language, etc.).

Recall (Fig. 2) that the final representation h_t^L at a token t is the sum of the embedding of the previous token (BoS for Token 0) and of all contributions from MLPs and attention heads at all layers, including H. Since, in MHA components, nonlinearities are only involved in the determination of attention scores, the overall contribution of head H to the RS of t can be decomposed into the sum of micro-contributions, each corresponding to a specific token $s \leq t$. Each such micro-contribution is the projection of the hidden representation at a previous token s right before layer l through W_v^H and W_o^H scaled by the attention score $a_{s,t}$:

$$\delta(H, t) = \sum_{s \leq t} a_{s,t} h_s W_v^H W_o^H \tag{6}$$

where we omit the layer index l and transposition operators to avoid clutter. Since attention scores are the outputs of a softmax, they are non-negative and such that $\sum_{s \leq t} a_{s,t} = 1$, therefore the contribution of H to the RS of t is a convex combination of the projection of the RSs h_s of all tokens $s \leq t$ through the matrix product $W_v^H W_o^H$. Interestingly, LLaMa 2 parameters are such that the projection of the RS of Token 0 through $W_v^H W_o^H$ has much smaller norm than the projection of all other input tokens, and this for all H.

Since Token 0 has no token to its left, all its attention is focused on itself. This means that Fig. 9 (bottom right) can also be read as a plot of the norms for the combined micro-contributions from Token 0 of all attention heads. Fig. 21 (left) shows the same plot limited to the layers where the attention sink is in place, whereas Fig. 21 (right) shows a similar plot for a random token (most tokens look like this). Note the different scale on the vertical axis: projections from Token 0 are much smaller compared to other tokens. They are also

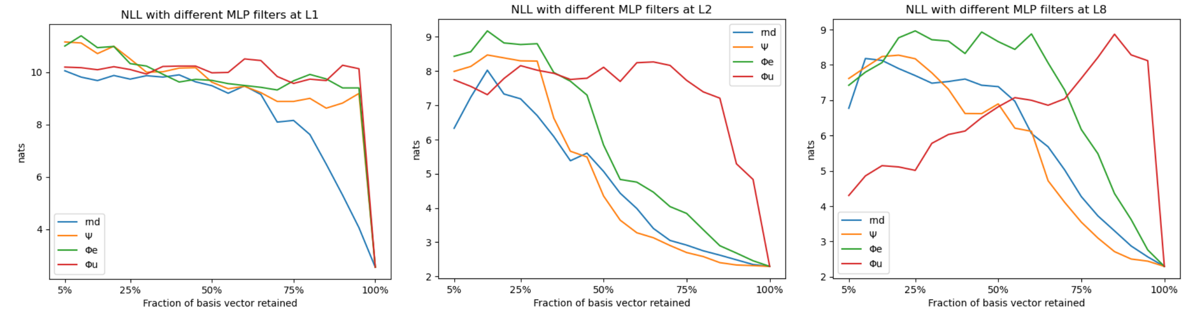

Figure 17: NLL filtering of MLPs in LLama2 7B (left) and 70B (middle) and (right) highlighting layers operating in the dark space.

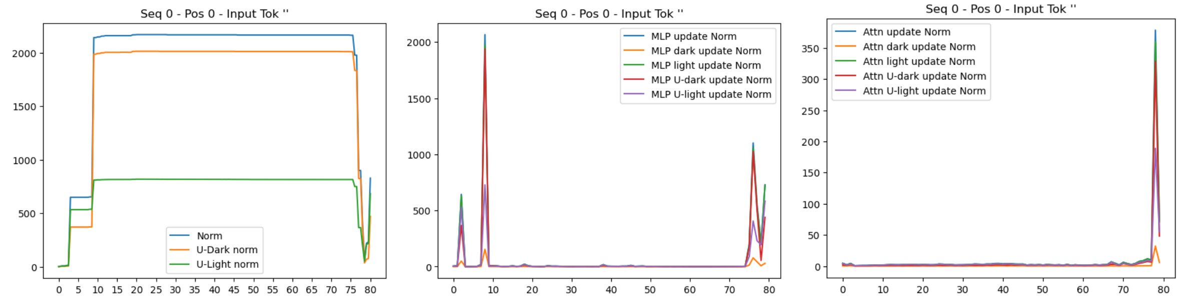

Figure 18: Token 0 residual stream in LLaMa2 70B. (left): RS norms by layer; (middle) norm of MLP contributions; (right) norm of MHA contributions.

darker. When they are added to the RS of a token t, they have negligible influence on the corresponding logits[8]. This in turn means that the (undesirable) perturbation from the irrelevant head H is minimized if the head assigns as much attention weight as possible to $s = 0$, i.e. $a_{0,t} = 1 - \epsilon$ for a small $\epsilon > 0$.

E Generation examples

Tables 4 and 5 show continuations from a same prompt generated by LLaMa2 13B models with three different filters. The left column is from a model filtered at a single layer with the filter Ψ that ablates the double projection on the 20% (resp. 10%) dark subspaces; the middle column is from a model where the filtering removes 20% (resp. 10%) random projections; the right column is generated with an equivalent Ω filtering. The row header indicates the layer immediately after which the filter was applied. Since the Ψ filter ablates the result of a double projection, we would expect its impact to be less noticeable. This is not the case: both the Ψ-filtered and the randomly-filtered models at L3[9] are disrupted already at 10%, but the randomly-filtered one starts recovering already when the filter is moved to L4, while the Ψ-filtered one keeps generating nonsense as the filter is moved up to L10 and beyond. Generations with Ω filtered-models remain largely coherent irrespective of where the filter is placed, for levels of suppression up to 20%.

F High-Mean Low-Variance received attention tokens

High-Mean Low-Variance token-layer pairs (HMLVs) are defined in Sec. 7 as token-layer pairs whose received attention over subsequent tokens has mean above $\tau_\mu = .018$ and variance below $\tau_\sigma = .01$. HMLVs tend to be associated to characteristic attention bars in attention matrices. Figure 22 shows the U-Dark ratio and Key dot-product metrics, by layer, for HMLVs and for the general population. HMLVs have much larger U-Dark Ratio than the general population, and also much larger key-dot product with the BoS token, supporting the hypotheses that they appear similar to the BoS token to subsequent tokens, and therefore behave as additional attention sinks.

[8] We verified that the reduction in entropy due to projections from Token 0, when sampling with $T = 1$, is always one or two orders of magnitude lower than the reduction gained by the same projections from other tokens.

[9] The impact on the random-filter at L3 suggests that L3 is where the residual stream is used at its full capacity.

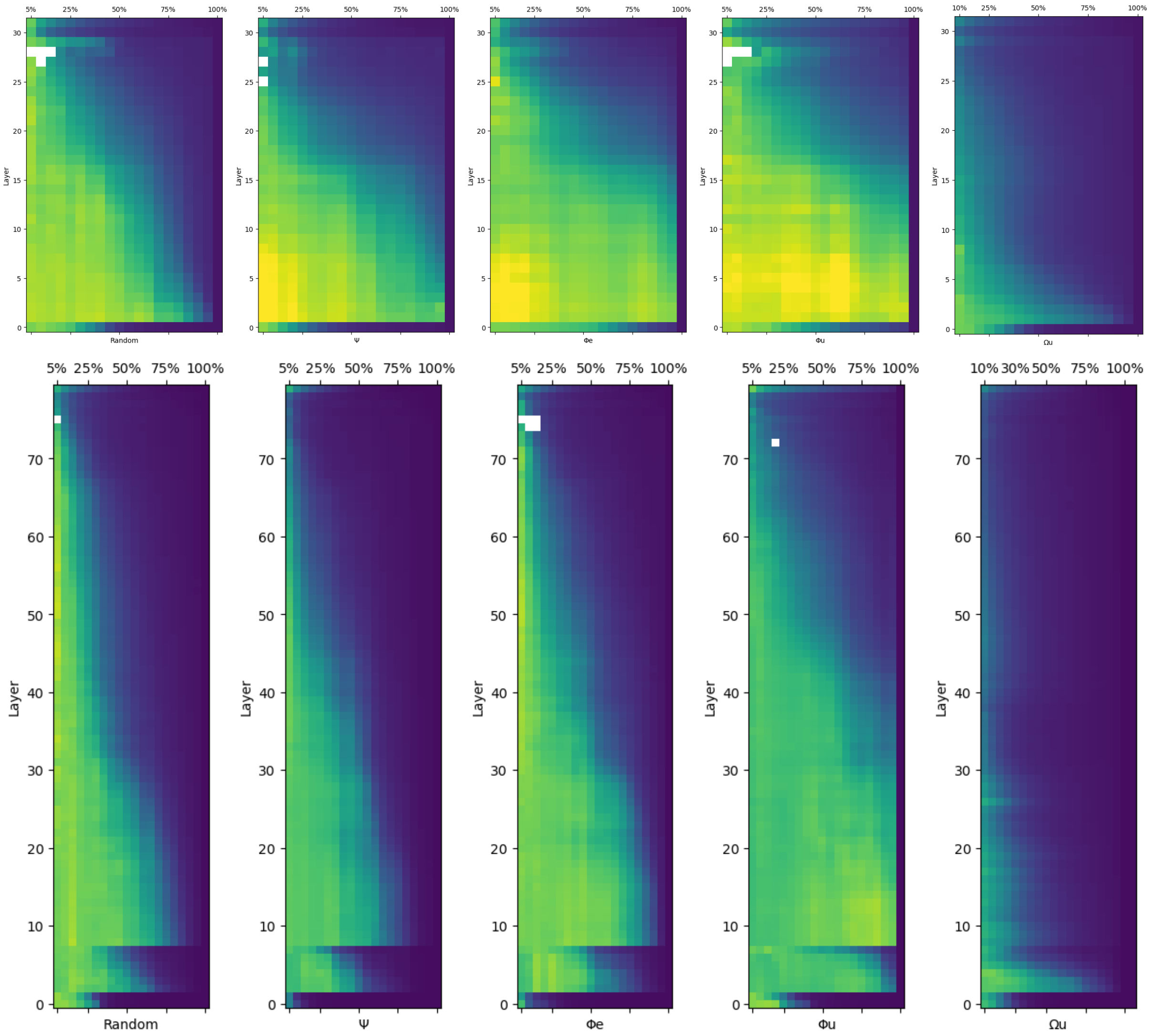

Figure 19: Negative Log-Likelihood when filtering the Residual Stream one layer at a time in LLaMa2 7B (top) and LLaMa2 70B (bottom). White cells correspond to configurations where NLL diverged for at least one of the samples.

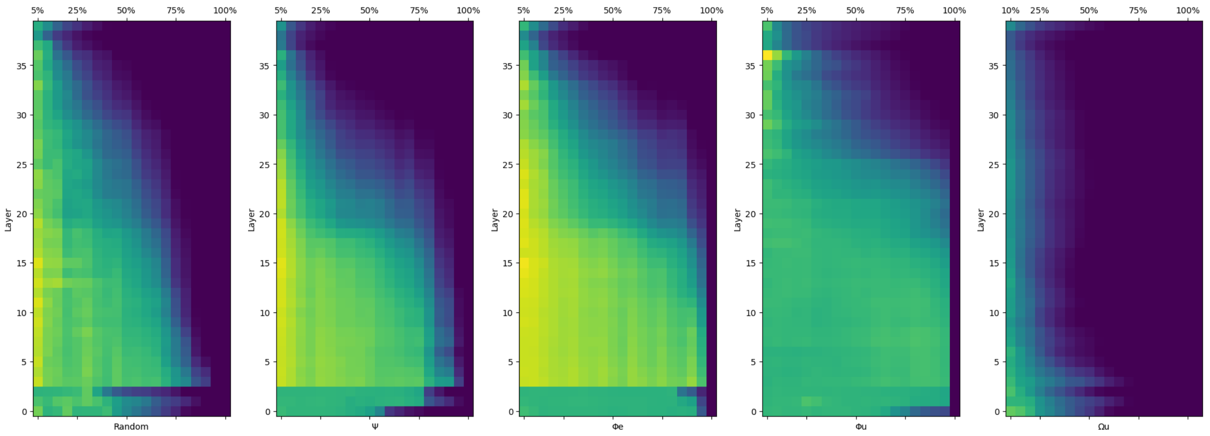

Figure 20: NLL filtering of MLPs in LLaMa2 13B on a sample of code from the DeepMind Code Contest dataset.

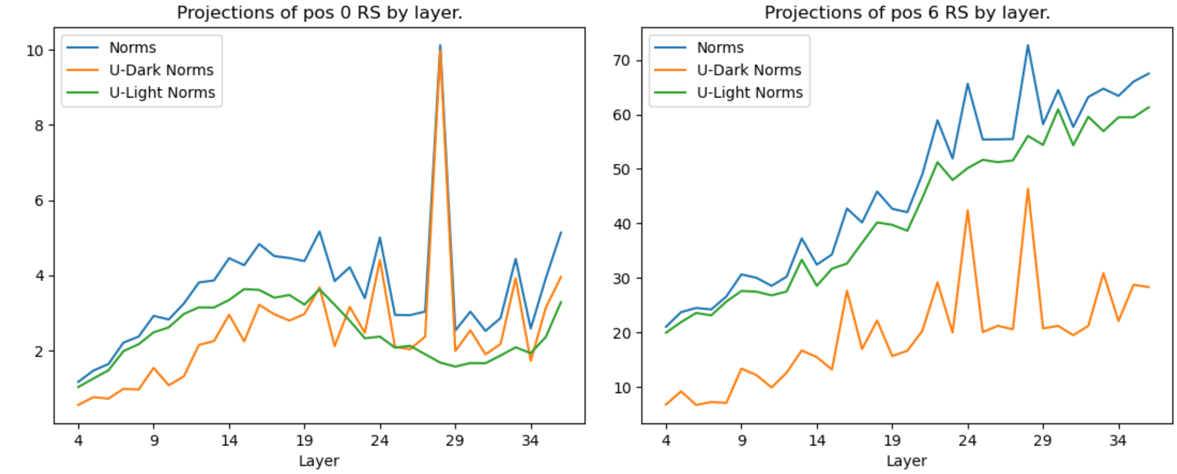

Figure 21: Norms of the combined projections of all attention heads from Token 0 (left) and from a generic random token (right). Note the different scale on the vertical axis: projections from Token 0 are much smaller, as well as darker.

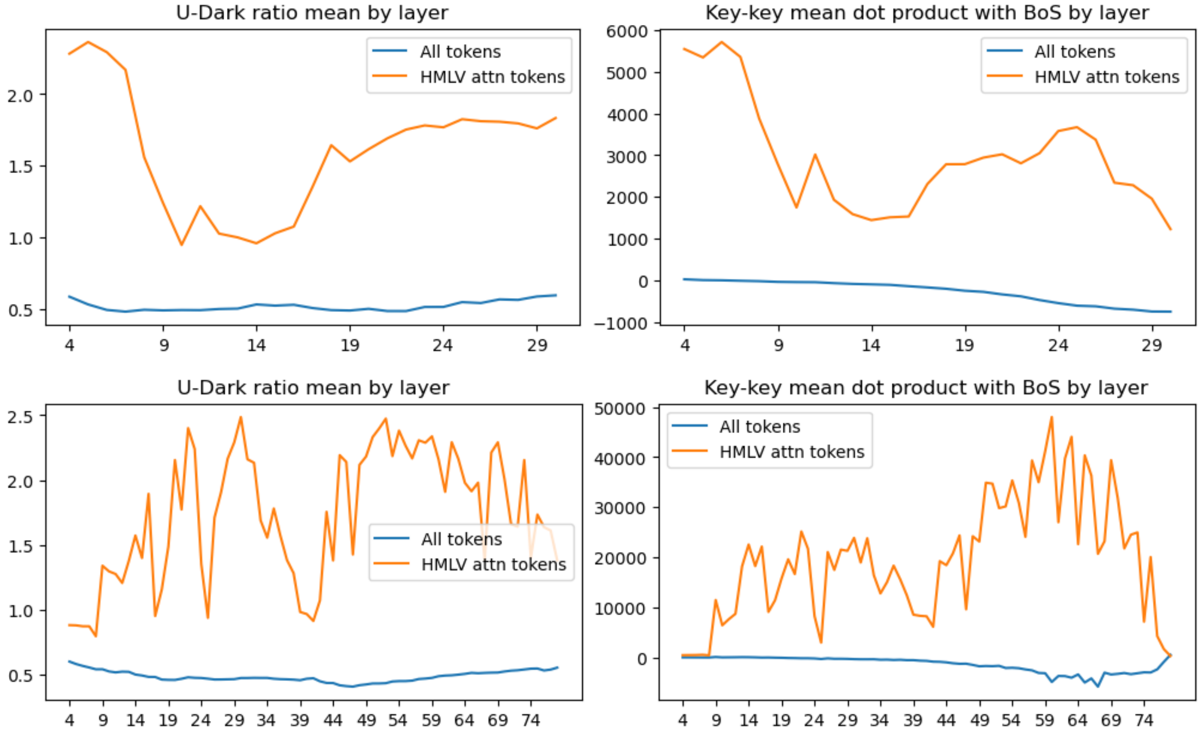

Figure 22: U-Dark Ratio and key dot-product metrics for the HMLVs and for the general population, by layers, for LLaMa2 7B (top) and 70B (bottom). See Section 7 for definitions.

Prompt: Tuatara are lizard-like reptiles that can be traced back to the time of the dinosaurs. For the first time in 200 years, researchers have discovered a baby tuatara on mainland New Zealand.			
	Ψ (80%)	Rnd (80%)	Ω_u (80%)
L3	When, “b/3,on,h,a The “t The A,e,c,t,a,w,d, The,a,d,th, The,th, The, The,t,d, The,d, on,	At the time of the early 2000-600, 2000 and 2000: 2000, 2000, 2000 and 2000: 2000, 2000,	Tuatara are one of the largest reptiles in the world and are native to New Zealand. However, due to predation, habitat loss and the introduction of mammalian predators, they are now limited to a few small islands. The tuatara is a small lizard-like re
L5	The first “tone” is the one, the first, the one, the first, the S, the first, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the	In the 1990s, the public is on the 100th and 100th. The 100th is the 100th. The 100th is the 100th. The 100th is the 1	A baby tuatara has been found for the first time in 200 years, according to a new study. Tuatara are lizard-like reptiles that can be traced back to the time of the dinosaurs, according to the study. The last time a tuat
L7	, in the, The , in the, the , in the, The , the , the, and , the , the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the,	Ali T and/ 2d. Ali T and 2d. Ali T and 2d. Ali T and 2d. A. Ali T and 2d. A. A. Ali T and 2d. A. A.	Tuatara are lizard-like reptiles that can be traced back to the time of the dinosaurs. For the first time in 200 years, researchers have discovered a baby tuatara on mainland New Zealand. Researchers from Massey University in New
L10	a y/t, a right The half, but, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the	A single baby tuatara is a rare sight. It’s been 18 years since a baby tuatou	The discovery was made by the Department of Conservation (DOC) and the University of Otago, which have been working on a project to monitor tuatara on the mainland. The last time a tuatara was found on the mainland was in the 1800s.
L14	Tuatara are lizard-like reptiles that can be traced back to the time of the dinosaurs. For the first time in 200 years, researchers have discovered a baby tuatara on mainland New Zealand. The E on a the, on the, the, at, 200, with, from, on, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the, the	The small, taura taura, is 3000 years old, and it is the most important reptile in the world. Taura Taura is a 3000-year-old reptile that is the most important re 3000 years old re	Tuatara are lizard-like reptiles that can be traced back to the time of the dinosaurs. For the first time in 200 years, researchers have discovered a baby tuatara on mainland New Zealand. The discovery of the tuatara, which is
L27	The newborn, which is only the 13th of its kind, was discovered on the “danger” Steph’, an, an, an, an, and the, and the, and the, and the, and the, and the, and the, and the, and the, and the	New Zealand’-Tiawan The island nation of New Zealand is a long way from the rest of the world, but that doesn’t mean it’s alone. It’s the world’s most isolated nation, but it’s also home to some of the world’s most unique and interesting	This is the first time in 200 years that a baby tuatara has been found on the mainland of New Zealand. The discovery was made by a group of researchers led by the University of Otago. The tuatara was found on the east coast of the North Island,
L36	The Tuatara is an iconic New Zealand native. The reptiles can live for more than 100 years and are found on the main islands of the country and a few smaller islands. The last time a tuatara was born on the mainland was in 1895,	The small lizard, weighing just 40 grams, was found in the South Island’s Kahurangi National Park. The species is usually found on the Chatham Islands, which are a group of small islands east of the South Island. The last time a tuatara	The discovery was made in a cave system in the remote Waitomo district, 300 kilometres south of Auckland. The young tuatara is about 10cm long and is the first recorded sighting of a juvenile tuatara on the mainland. The cave system is

Table 4: Generations from LLaMa2 13B with Residual Stream filtered at one layer, comparing removing 20% random projections, Ψ filtering, and Ω_u filtering. Note that generations can be a mixture of true and hallucinated content.

<table>
<tr><td colspan="4"><b>Prompt:</b> If you’re interested in Grizzly Bear viewing in and around the northern Gulf of Alaska, consider a chartered flight with Trygg Air Alaska based out of King Salmon with service to and from Anchorage.</td></tr>
<tr><td></td><td>Ψ (90%)</td><td>Rnd (90%)</td><td>Ω_u (90%)</td></tr>
<tr><td>L0</td><td>Trygg Air Alaska is a small air taxi company based in King Salmon, Alaska. The company provides chartered flights from King Salmon to Anchorage, Alaska, which is the nearest city to King Salmon with an airport. T§he company is owned by a man named Try</td><td>The Alaska Department of Fish and Game (ADFG) has been conducting an aerial survey of the Gulf of Alaska for Grizzly Bears since 2006. The survey is conducted each summer to estimate the number of bears in the Gulf. The survey area covers</td><td>The flight is a one-hour scenic trip that flies over the Alaska Peninsula and Bristol Bay. You’ll be able to view the Alaskan wilderness and Grizzly Bears in their natural habitat. The flight includes a stop at the Brooks Camp for a</td></tr>
<tr><td>L3</td><td>The EA, a GS, the E, A, for a total, the GS, the E, the S, the E, the S, the 15, the 16, the 15, the 16, the 15, the 17, the</td><td>What is the most beautiful thing that has ever been? What is the most beautiful thing that has ever been? It is not about the most beautiful thing that has ever been. The most beautiful thing that has ever been. It is not about the most beautiful thing that has ever been. It is not</td><td>The Grizzly Bear is one of the largest land carnivores in the world. Males can reach a weight of 800 pounds and females can reach a weight of 400 pounds. The Grizzly Bear is one of the most powerful predators in North America</td></tr>
<tr><td>L4</td><td>The ACS, or, the A.S.E. The A.S.S.A.S. on the E.S.A. The A.S.E.E.A.S. on the E.S.A. The A.S.S.A</td><td>The Grizzly Bear is a wild creature and not a domesticated pet, so when you’re dealing with a wild creature, you have to be careful with your approach and how you do it. The Grizzly Bear is a wild creature and not a domesticated pet, so when you’</td><td>Trygg Air Alaska is a family owned and operated air service based out of King Salmon, Alaska. We offer flightseeing tours and charter services to the Alaska Peninsula, Bristol Bay, and the Aleutian Islands. We have been flying in the region since 19</td></tr>
<tr><td>L6</td><td>The cost is about $1,500 per person, but the trip is about a month long. The Gulf of Alaska is the only place in the world the bears are a problem. The trip is the same, but the bears are not. The bears are the same</td><td>Trygg Air Alaska is a charter service offering air transportation between Anchorage, Alaska and King Salmon, Alaska. With a fleet of two Piper Cherokee 6 aircraft, Trygg Air Alaska offers an economical alternative to commercial airlines and private aircraft. The</td><td>Trygg Air Alaska is a small air charter service based out of King Salmon, Alaska. They provide flight services to and from Anchorage and the remote communities of King Salmon, Iliamna, and Bristol Bay. Trygg Air Alaska offers a range of services, including scen</td></tr>
<tr><td>L10</td><td>The Tail, the Beans, and the Salmon The Tail, the Beings, and the Salmon The Tail, the Beings, and the Salmon The Tail, the Beings, and the Salmon The Tail, the Beings, and the Salmon</td><td>Trygg Air Alaska is a family owned business, with the same values as our Alaskan ancestors: honesty, integrity and hard work. We’ve been in the air since 2010 and we’ve flown more than 500,000 miles over the beautiful</td><td>Trygg Air Alaska is a family owned and operated airline based in King Salmon, Alaska. Trygg Air Alaska has been providing charter services to the Grizzly Bear Viewing lodges in and around the Gulf of Alaska since 1989. We have the</td></tr>
<tr><td>L22</td><td>A Grizzly Bear viewing charter with Try The best time to see bears is from the end of the first week of the 2018/19 salmon run, through to the end of the 2019/2020 salmon run. The sal</td><td>Grizzly Bear viewing is an amazing experience. The bears are beautiful, majestic and awe-inspiring. The bears are also dangerous, and there are a lot of things to consider before you set out to view them. Trygg Air Alaska is dedicated to providing</td><td>Grizzly Bear viewing can be arranged in conjunction with fishing trips or as a separate trip. Fly-out trips can be arranged to view the Grizzly Bears at Brooks Falls, McNeil River, Katmai National Park, Lake Clark National Park,</td></tr>
<tr><td>L32</td><td>Trygg Air Alaska offers chartered flights for the purpose of viewing grizzly bears in and around the northern Gulf of Alaska. The northern gulf of Alaska is the only place in the world where grizzly bears are known to catch and eat the endanger</td><td>Grizzly Bear Viewing from the air The Gulf of Alaska is home to a large population of Grizzly Bears. There are approximately 1000 bears in the area, which is considered one of the largest Grizzly Bear populations in the world. In addition</td><td>Grizzly Bear Viewing, Fishing and More. Trygg Air Alaska is a charter air service based out of King Salmon, Alaska. Trygg Air Alaska has been operating for over 20 years. Trygg Air Alaska offers charters to the Katmai National</td></tr>
</table>

Table 5: Generations from LLaMa2 13B with Residual Stream filtered at one layer, comparing removing 10% random projections, Ψ filtering, and Ω_u filtering.