Transformer Feed-Forward Layers Are Key-Value Memories

Kaiwei Lei

06 June 2025

Outline

Motivations

- Wide application of DNNs, and transformer-based models in particular
- We know much about *how* it generates text, but little about *why* it generates a *specific* text.
- Interpretability research aims to bridge the gap.

Motivations

- Specifically, a *mechanistic* approach to interpretability attempts to "reverse engineer" the computations performed within the model, and extract human-explainable elements. (Elhage et al., 2021)
- In the context of transformers, the current research focuses on analysing the information flows within each component of the model architecture. (Sharkey et al., 2025)

Research Questions

- In a decoder-only, GPT-isque model, two-thirds of the parameters are found in MLP layers. What is the function of the MLP layer?
- Background: Three "'paradigm shifting" views

Outline

Residual stream and I/O behaviours

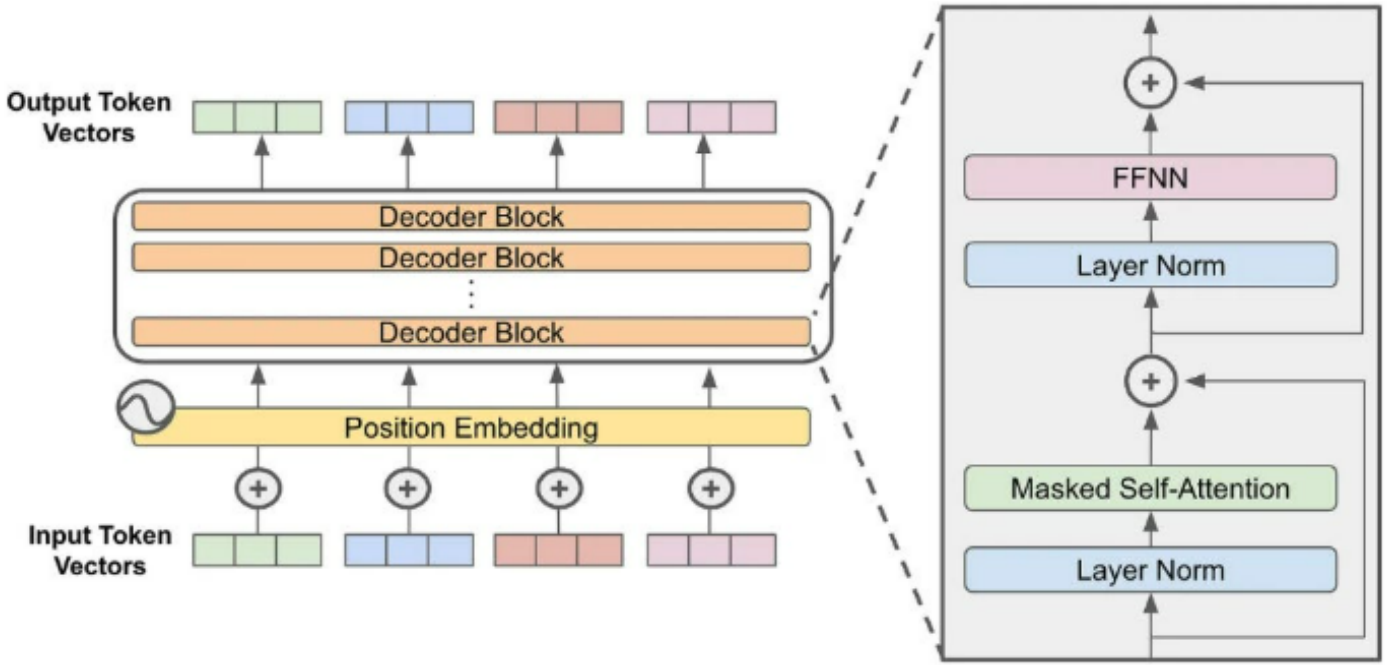

Figure: A typical illustration of a decoder architecture.

Residual stream and I/O behaviours

- We understand the model as taking a vectorised token as input, and "transforming" it through attention, MLP into a logit as output.

Residual stream and I/O behaviours

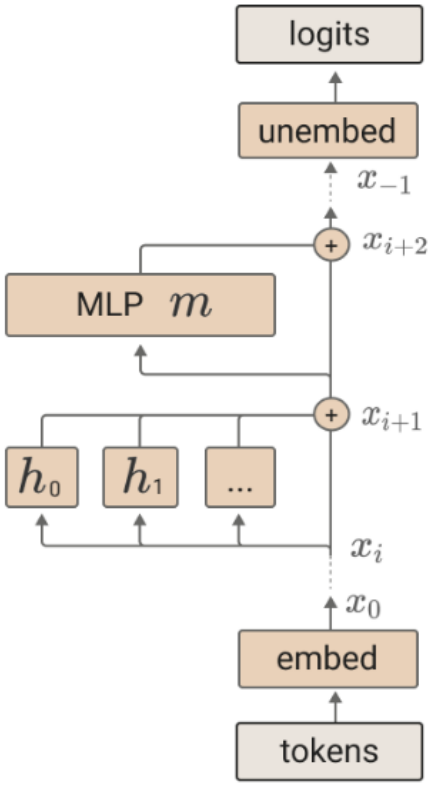

Figure: Residual stream view as per Elhage et al., 2021

Residual stream and I/O behaviours

- Because of the residual connection (always an addition), we can also view the generation process as one continuous flowing "residual stream" from embedding to unembedding, and each attention block/MLP layer reads signals from this stream, and in turn writes information into this stream.
- This viewpoint reorients our focus on each individual module, what information "activates" it, and what information it contains.

Neurons as units

1. Specifically, we can zoom into one MLP layer:
2. Each MLP layer in modern architecture has 3 important components: an upward projection M_1, an activation function f, and a downward projection M_2
3. Sizes: $M_1 : \mathit{hidden}, \mathit{inter}$, $M_2 : \mathit{inter}, \mathit{hidden}$, where the *hidden* dimension is the size of the residual stream
4. We define a **neuron** as a vector in the weight matrices along the dimension *inter*

Neurons as units

Given a residual stream x, the output of an MLP layer is decomposed as:

$$FF(x) = f(x \cdot M_1) \cdot M_2 \tag{1}$$

We can then further decompose the equation along the neuron dimensions, building on Cunningham et al. (2023):

$$FF(x) = f([xm_0, xm_1, ...xm_{inter}]) \cdot M_2 \tag{2}$$

$$= [f(xm_0), f(xm_1), ...f(xm_{inter})] \cdot M_2 \tag{3}$$

And further:

$$FF(x) = [mc_0, mc_1, ...mc_{inter}] \cdot M_2 \tag{4}$$

$$= \sum_{i}^{inter} mc_i \cdot v_i \tag{5}$$

where $mc_i = f(xm_i)$.

Neurons as units

- We may call mc_i the activation for neuron i
- The main claim of the Geva paper can be reformulated as: a given neuron activates on specific input text patterns, retrieves specific information stored in v and writes to residual stream.

Logit lens as interpretability tool

- First used in nostalgebraist (2020)
- If the next token is generated by multiplying the unembedding matrix with the final layer residual stream, what happens if we multiply the unembedding matrix with residual streams from earlier layers?
- More generally, we can project any model component to the vocabulary space using the unembedding matrix W_U

Checkpoint

To summarise, mechanistic interpretability makes the following claims:

- The behaviour of the whole model is the sum of each individual module.
- Each MLP module in turn is a linear combination of memory coefficients (neuron activations) multiplied with output vectors.
- The behaviour of each component can be explained by projecting it into vocabulary space.

Outline

Experimental Setup

- Model: Fairseq (applicable to any transformer based model)
- Dataset: training corpus Wikitext-103

Procedures

1. For each FFN block of a layer, 10 random neurons are sampled.
2. Every sentence in the training corpus of the model is turned into a list of "cloze-test" style strings (thus "I love you" would correspond to a list of ["I", "I love", "I love you"]).
3. Dot product is computed between each of the strings and the neurons. The 25 strings that yield the highest dot product with each neuron are saved and named "trigger examples".
4. For each list of trigger examples given a neuron, humans are then recruited to first identify a pattern/commonality among these examples, as well as classifying if the pattern is syntactical or semantical.

Procedures

- Note that the paper used dot product, but as we have seen already, large dot-product = high neuron activation
  ($mc = ReLU(x \cdot m) = x \cdot m$)
- But why only the highest activations?

Procedures

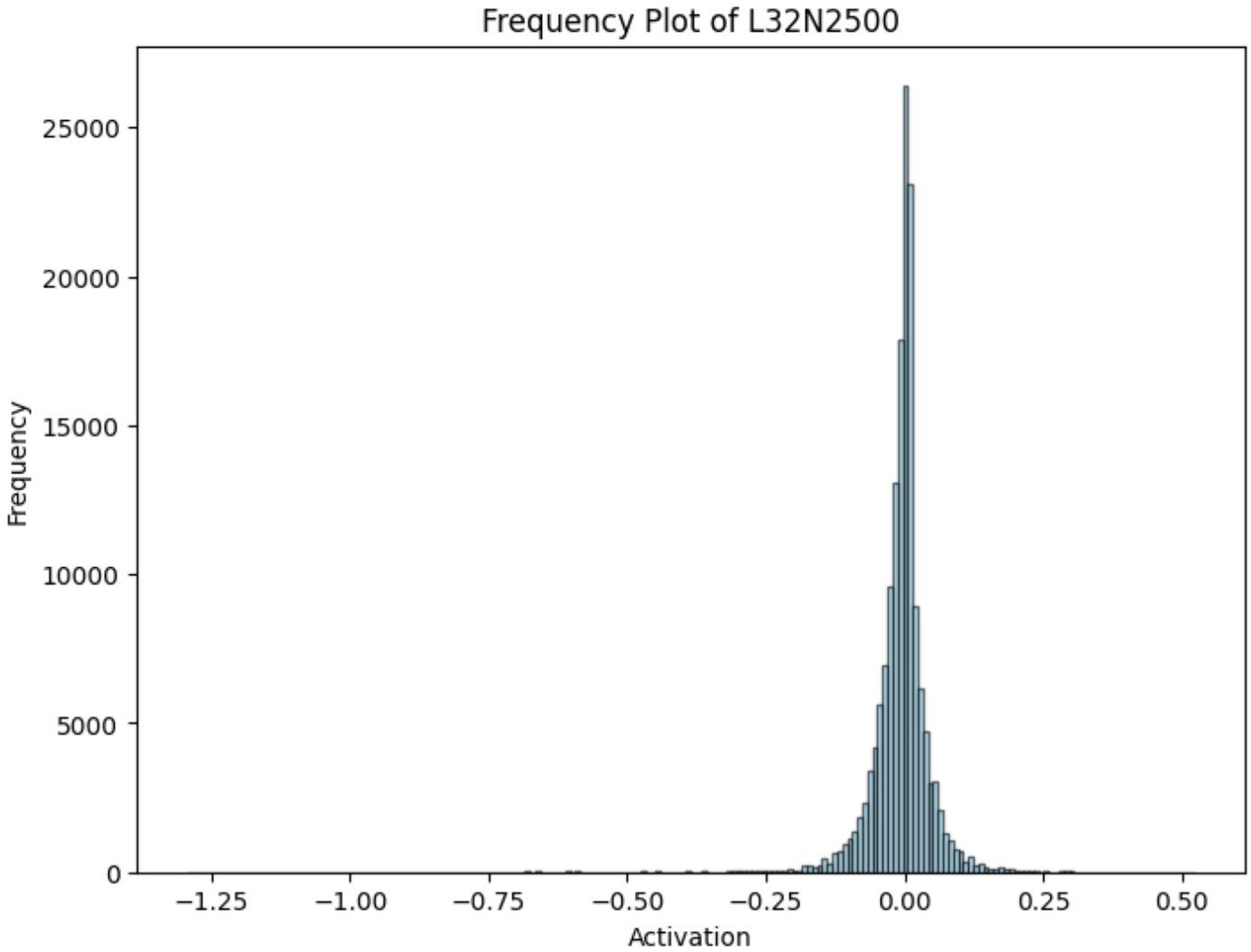

Figure: Activations distribution of Olmo-7B neuron L32N2500 on 10,000 random sentences from Dolma dataset.

Procedures

A typical neuron activation distribution is extremely leptokurtic: vast majority of activations bobbles around 0, indicating that the neuron is mostly dormant, only activated in very specific contexts.

Outline

Input behaviours

Key	Pattern	Example trigger prefixes
$\mathbf{k}^{1}_{449}$	Ends with *"substitutes"* (shallow)	*At the meeting, Elton said that "for artistic reasons there could be no substitutes* *In German service, they were used as substitutes* *Two weeks later, he came off the substitutes*
$\mathbf{k}^{6}_{2546}$	Military, ends with *"base"/"bases"* (shallow + semantic)	*On 1 April the SRSG authorised the SADF to leave their bases* *Aircraft from all four carriers attacked the Australian base* *Bombers flying missions to Rabaul and other Japanese bases*
$\mathbf{k}^{10}_{2997}$	a "part of" relation (semantic)	*In June 2012 she was named as one of the team that competed* *He was also a part of the Indian delegation* *Toy Story is also among the top ten in the BFI list of the 50 films you should*
$\mathbf{k}^{13}_{2989}$	Ends with a time range (semantic)	*Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7* *Weekend tolls are in effect from 7:00 pm Friday until* *The building is open to the public seven days a week, from 11:00 am to*
$\mathbf{k}^{16}_{1935}$	TV shows (semantic)	*Time shifting viewing added 57 percent to the episode's* *The first season set that the episode was included in was as part of the* *From the original NBC daytime version , archived*

Figure: Examples of neurons triggered by specific input patterns, some share the literal same words ("shallow" features), some share similar meanings ("semantic" features).

Input behaviours

- "key" in our language is the activation.
- Annotators were able to identify at least one pattern from the majority of the samples.
- The paper further posits that neurons in different layers capture different patterns: a "shallow" pattern is mainly found in lower layer neurons, whereas a semantic pattern, is attributed to upper layers.
- Note that due to attention, the pattern identifed is not limited to the current token, it could be tokens many positions before.

Output behaviours

Value	Prediction	Precision@50	Trigger example
$\mathbf{v}_{222}^{15}$	*each*	68%	*But when bees and wasps resemble each*
$\mathbf{v}_{752}^{16}$	*played*	16%	*Her first role was in Vijay Lalwani's psychological thriller Karthik Calling Karthik, where Padukone was cast as the supportive girlfriend of a depressed man (played*
$\mathbf{v}_{2601}^{13}$	*extratropical*	4%	*Most of the winter precipitation is the result of synoptic scale, low pressure weather systems (large scale storms such as extratropical*
$\mathbf{v}_{881}^{15}$	*part*	92%	*Comet served only briefly with the fleet, owing in large part*
$\mathbf{v}_{2070}^{16}$	*line*	84%	*Sailing from Lorient in October 1805 with one ship of the line*
$\mathbf{v}_{3186}^{12}$	*jail*	4%	*On May 11, 2011, four days after scoring 6 touchdowns for the Slaughter, Grady was sentenced to twenty days in jail*

Figure: Examples of retrieved value vectors, its top-1 predicted token, and the actual next token in the sentences.

Output behaviours

- The paper explores briefly the role of row vectors of M_2, i.e. the v_i
- For the top trigger examples of neuron m_i, the paper found that the top-1 token predicted by v_i, when unembedded, exhibit a certain degree of agreement with the **next** token in the trigger example
- In short: context prompt $\rightarrow$ highly activated neuron i $\rightarrow$ corresponding value i retrieved $\rightarrow$ predicted next token

Checkpoint

To recap:

- On the input side, neurons get activated on very specific tokens
- Neurons in different layers seem to demonstrate different types of functionality: some respond to syntactic info, some semantic
- On the output side, the corresponding output vectors seem to store information about the next tokens that the model wants to predict.

Outline

Are human annotators reliable?

- The experiments from the paper heavily relies on human annotators, where they are asked to identify a common pattern given a group of texts.
- Humans are known be psychologically biased to identify patterns even on noise data, a phenomenon known as "apophenia"

Layer division and token positions

- Geva paper posited some speculations on the divisions of MLP neurons between layers.
- The view sees the model "refines" its best guess at the next token gradually from the first layer to the last.
- This view has largely been abandoned in subsequent papers Geva et al. (2022), Geva et al. (2023)

Layer division and token positions

- Decoder-only models are trained with predicting the next-most-likely-token objective, and this leads to the duality of a neuron
- it encodes information about the current token, and at the same time predicts the next token position.
- These two functions might even conflict with each other: consider the prefix "London is the capital city of", after the attention layers, the neurons at the last position might contain associated information about London, which can be Buckingham Palace, Chelsea, or Dickens, but at the same time, the correct next token is " the" of " the United Kingdom", which is a stopword.

Layer division and token positions

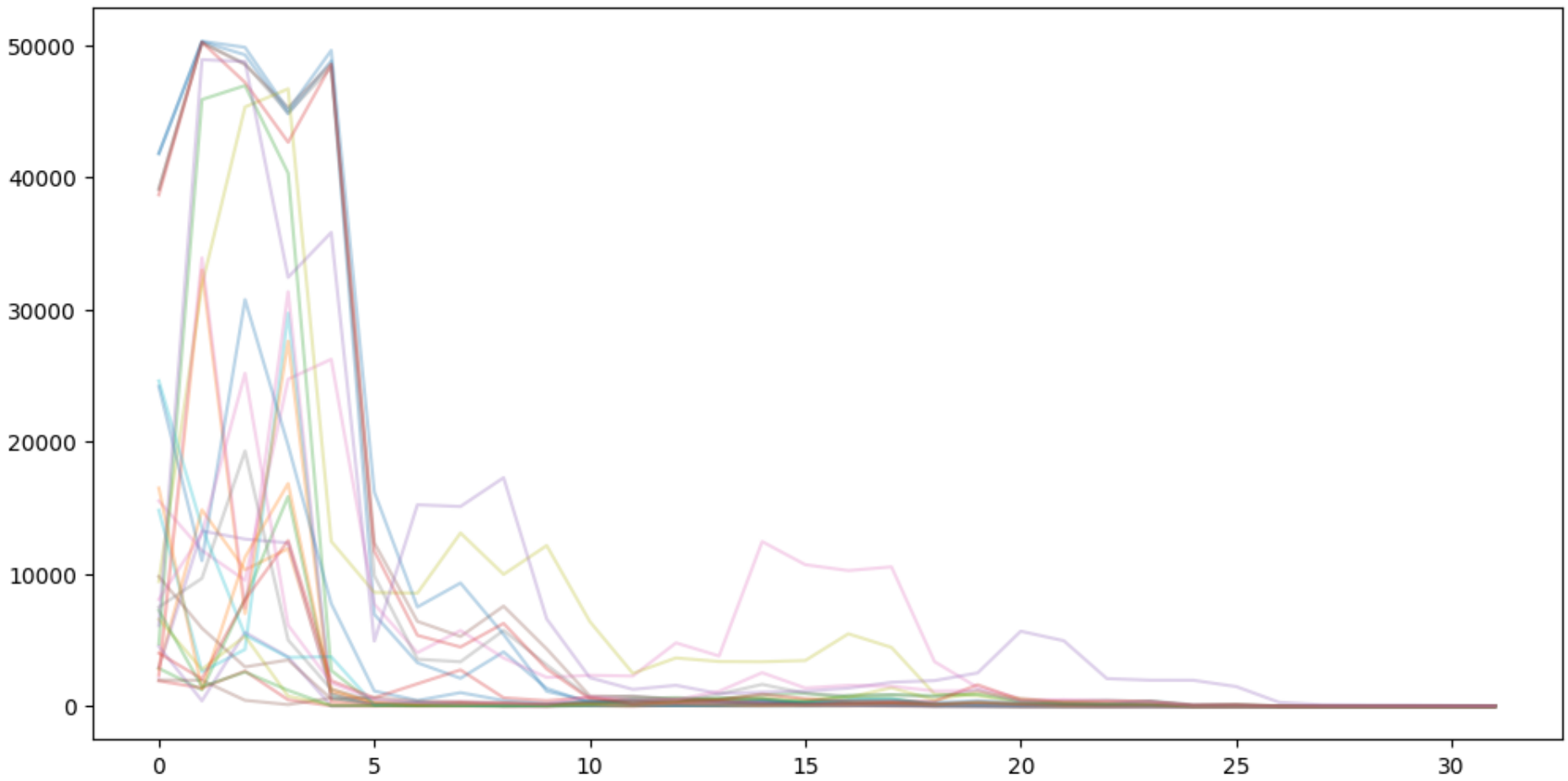

Figure: Ranks of the target answer token across layers of the sub question set: city_in_country. The refinement view does not explain the surety of model.

Outline

Open questions

Rather than offering the final answer, the paper raises more questions on neuron interpretability.

Neurons?

- Our unit of interpretability is neuron.
- Some researchers don't believe individual neurons are interpretable, but only groups of neurons. This view is based on the so-called "no priviledged basis" thesis.
- They use so-called Sparse Autoencoders (SAEs).

Finding neurons

Recall that Geva paper used activation-based thresholds.

- However, some neuron has activations up to 200, while some only ranges between -1 and 1.
- Does small activation equal non-activation or example of simply how this neuron naturally behaves?
- Other criteria available.

Philosophical question: Knowledge neurons?

What exactly have we found?

- A methodological point: we want to establish that the *function* of neuron x is to store information y: we need to confirm a) whenever x is activated, we observe y, and b) whenever x is not activated, we don't observe y.
- This strong claim is known as the "knowledge neuron" thesis, and it is controversial.
- Some paper claiming neuron to store information $\neq$ neuron to ablate, see Niu et al. (2024).

Thank you for listening!

Outline

References From Presentation

- Cunningham, H., Ewart, A., Riggs, L., Huben, R., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. https://arxiv.org/abs/2309.08600
- Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. (2021).A mathematical framework for transformer circuits. *Transformer Circuits Thread*, *1*(1), 12.
- Geva, M., Bastings, J., Filippova, K., & Globerson, A. (2023). Dissecting recall of factual associations in auto-regressive language models. https://arxiv.org/abs/2304.14767
- Geva, M., Caciularu, A., Wang, K. R., & Goldberg, Y. (2022). Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. https://arxiv.org/abs/2203.14680
- Niu, J., Liu, A., Zhu, Z., & Penn, G. (2024). What does the knowledge neuron thesis have to do with knowledge? https://arxiv.org/abs/2405.02421
- nostalgebraist. (2020) Interpreting gpt: The logit lens. *Lesswrong Forum*