Emergent Token–Neuron Circuits in GPT-2

Biclustering Transformer Activations to Measure Neural Modularity

Sam Turner

Theory of Predictive Modeling

CS 580/PHSCS 513R

Background + Problem Statement

Background

- Emergence
  - Smaller sub elements make a quantitatively different whole
  - More is different not just more
- LLMs
  - As LLMs scale, phase transitions occur in their performance (Wei, 2022)
  - Internal structure is a "black box"
  - Understanding these phase transitions is difficult as a result
- Spectral Biclustering
  - Regular clustering groups examples on single axis
  - Biclustering jointly groups attributes using linear algebra
  - These subgroups can show coherent groups with new meaning

Problem Statement

We investigate if Bi-clustering across neuron activations/input tokens can help reveal the internal structure of Large Language Models

Dataset + Models

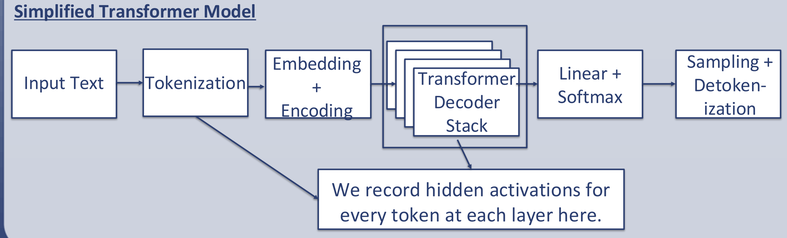
Dataset

- Source: Wikimedia Wikipedia dataset from HF
- Snapshot:11/01/2025 English
- Size: 11.6 GB
- Structure: Full text articles

Models

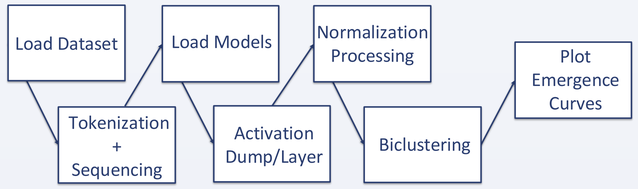
Model	Approx. Parameters	Transformer Layers Analyzed	Hidden Size	# Attention Heads	Activation Shape per Layer (tokens × dim)
GPT-2	~124M	12	768	12	20,000 × 768
GPT-2 Medium	~355M	24	1,024	16	20,000 × 1,024
GPT-2 Large	~774M	36	1,280	20	20,000 × 1,280
GPT-2 XL	~1.5B	48	1,600	25	20,000 × 1,600

Simplified Transformer Model



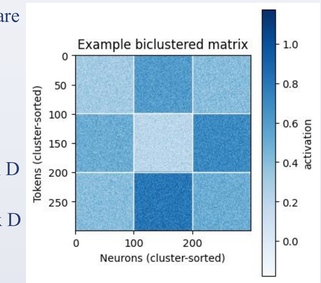
Methods

Pipeline



Overview

- Bi-clustering
  - We define a matrix M | $M \in R^{T \times N}$ where rows are **T**oken positions and columns are **N**euron activations
  - Spectral biclustering
    - Normalization (this helps heavily used neurons not dominate unnecessarily
      - Let D_r be the row of matrix D where $(D_r)_{ii} = \sum_j D_{ij}$
      - Let D_c be the row of matrix D where $(D_c)_{ii} = \sum_i D_{ij}$
      - Let $B = D_r^{-1/2} M D_c^{-1/2}$
  - SVD
    - $B \approx U_k \Sigma_k V_k^T$
    - $U_k \in R^{T \times k}$: top k left singular vectors (token-side)
    - $V_k \in R^{N \times k}$: top k right singular vectors (neuron-side)
    - Σ_k: diagonal of top k values
  - Clustering
    - Cluster and sort rows(tokens) by u_i and columns(neurons) by v_j

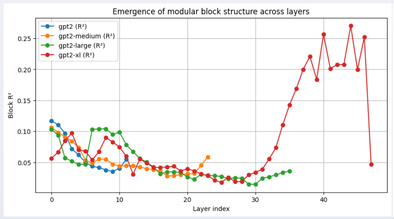


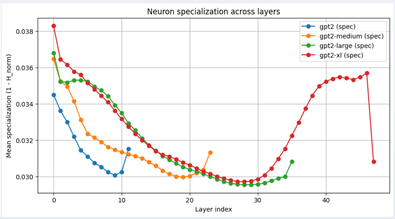
- Metrics Description
  - R^2
    - Measures the effectivity of the biclustering. It states that the biclusters account of X% of the activation matrix
    - $R^2 = 1 - \frac{\sum_{i,j}(M_{ij} - \widehat{M_{ij}})^2}{\sum_{i,j}(M_{ij} - \bar{M})^2}$
      - $\widehat{M} = reconstructed\ value\ from\ bicluster\ model$
      - $\bar{M}$ = overall mean across M matrix
  - Z-score
    - States how likely the resulting bicluster could come from a random partitioning of blocks
    - $Z = \frac{S_{obs} - \mu_{null}}{\sigma_{null}}$
- Search for Best Bicluster Params
  - We will run a loop over each model and each layer to find the number of row and column clusters which have the best R^2 and Z-score
  - This will provide us with the best cluster groups to do our analysis over
- Emergence Curves
  - Using the best K for rows and columns we will plot the R^2 values and the mean specialization to look at possible places of emergence in the models
- Neuron Circuit Analysis
  - Following the biclustering we will complete an analysis of the top activations in each cluster to see if we can attribute any semantic meaning to the cluster defined as possible neuron circuits in the model layers
  - We will then group layers by possible circuits to attempt better insight into interpretability

Results & Discussion

- **Biclustering Results**
  - Our biclustering showed a variety of different clustering sizes and activation levels
  - This bicluster map for layer 40 of gpt2-xl had some of the best activations out of our entire result set
  - For all the model sizes the best biclusterings were in the form of 12 rows and 12 columns

- **Emergence Curves**





  - From our graphs, we see that all models see a spike in the end with a decrease in the middle. This tells us that the models start with some distinct groupings followed by mixing in the middle of the model's layers and ends with adding more meaning into the layer's circuits.
  - GPTt-xl showed an extremely abnormal jump at the end compared to the other models
- **Neuron Circuits**
  - In analyzing the top clusters of the layers with the highest R2 clusters begin to show coherence especially in the later layers of the xl model
  - Cluster show focuses in the top tokens which repeatedly activate the group of neurons with distinct function groups emerging; however, these groups are not exclusive and do include other glue words
  - Layers do not specialize, but most show the similar groupings that develop over layers.
  - These results provide some insight into meaning in layers, they do not offer explanations for every layer. This could point to the idea that in the middle layers, the transformations applied to the token matrices are not connected with human semantic meaning

Cluster Family	Example Tokens	Meaning
Generic prose / glue	the; of; and; to; in; a; is; ,; .	Common function words and basic punctuation
Anarchism / politics	anarchism; anarchist; state; authority; capitalism; socialism; revolution; movement	Political ideology and state/authority concepts
Climate / surface / atmosphere	snow; surface; climate; solar; feedback; temperature; ice; sea; atmosphere; radiation	Earth and climate science topics
US geography / Geography	Alabama; Montgomery; state; county; city; population; area; census; United States	US state / city encyclopedia-style facts
Orthography / phonetics	vowel; consonant; letter; alphabet; script; IPA; Latin; Greek; phonetic; Unicode	Writing systems, letters, and sounds
Punctuation / layout	newline; paragraph break; .; ,; -; *; :; (;)	Sentence and paragraph boundaries; formatting
Mixed / residual	misc words; rare terms; blended topics	Leftover or blended feature space

- **Future Work**
  - Research was limited by GPU size and time, future work could use more neuron samples, other transformer models, or different cluster sizes
  - More work could be done in analyzing the token cluster for better understanding of meaning

Repository

Github Repository: https://github.com/sturner11/TokenNeuronBiclustering

References

- Katunchi. (2016, March 17). *Starling murmuration Primorsko* [Photograph]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Starling_murmuration_Primorsko2_%E2%80%94_%D0%BA%D0%BE%D0%BF%D0%B8%D1%8F.jpg
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD
- Dhillon, I. S., Mallela, S., & Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 89-98). ACM. https://www.cs.utexas.edu/~inderjit/public_papers/kdd_cocluster.pdf
- scikit-learn developers. (n.d.). *Biclustering documents with the Spectral Co-clustering algorithm*. In *scikit-learn 1.7.2 documentation*. Retrieved December 8, 2025, from https://scikit-learn.org/stable/auto_examples/bicluster/plot_bicluster_newsgroups.html