Investigating Lexical and Syntactic Complexity in Language Models via Activation Patching

#Mechanistic Interpretability #LLMs #Nueroscience

Smayan Khanna
Department of Computer Science
University of Chicago
smayan@uchicago.edu

Ian Zhang
Department of Computer Science
University of Chicago
ianhuzhang@uchicago.edu

This template is built on NeurIPS 2019 template.[1]
The content is based on Stanford CS224N's Custom Final Project content.[2]

1 Key Information to include

- Mentor: Dang Nguyen
- Sharing project: None
- TL;DR: Inspired by recent findings in mechanistic interpretability, this project explores whether activation patching can alter an LM's lexical richness and syntactic complexity by injecting activations from models trained on simpler data into more complex ones.

2 Research paper summary

Repeat the following template for each paper. Please do not exceed 1 page per paper.

Title	From Infant to Toddler to Preschooler- Analyzing Language Acquisition in Language Models
Venue	Stanford CS224N
Year	2024
URL	https://web.stanford.edu/class/cs224n//final-reports/256656329.pdf

Table 1: Paper 1 [1].

Background. The paper we chose, which was a final project for Stanford's CS224N course, focuses mainly on the subject of language acquisition in infants as they begin to master language through interaction and learning from parents and other individuals. Previous papers and research in the field have utilized computational language models to gain insight into human language understanding, including some that focus on child-directed speech and their differences from adult speech. This intersection between developmental neuroscience and computer science is a rich target of research, especially in unraveling the black boxes that our current understanding of language models, along with the human brain, tend to be.

Summary of contributions. Many papers exist which try to utilize language models to interpret human language understanding, including in children. However, the author notes that many of these previous authors studying the language development of children train their models on the entire dataset of child-directed text, effectively comparing child language master to adult language mastery. The current author showed that by training multiple models on the language of different age groups

[1] https://www.overleaf.com/latex/templates/neurips-2019/tprktwxmqmgk
[2] https://web.stanford.edu/class/cs224n/

of children, they could gain insights into the developmental pattern of babies as they grow and are exposed to language of increasing complexity. They show plausible paths for language development through childhood, and further show that more complex semantic understandings may not develop fully until after 6 years of age.

Limitations and discussion. One obvious limitation of this paper is that it only utilizes training language datasets for ages up to 6. This means that we miss out on interesting patterns of language development that happen mainly after 6 years of age.

Also, another limitation is that assumption that language models (BabyBERTa) accurately represent actual language development in infants. It also assumes that parents and other individuals adapt their language to different-aged children, and this is why children get better with speech. I am not fully convinced of this method, but I still think this is a very interesting way to approach the problem, and even to observe how the *model* adapts to different complexities of language.

Why this paper? We wanted to do a project that relates to Mechanistic Interpretability, but wanted to make sure that our project stayed within scope and reason for this course. This "paper" was presented as a project in Stanford's CS224N course, and seemed like a good introduction for us to examine the way models will adapt when trained on different difficulties of texts.

Wider research context. This paper focuses on modelling child language development, but we think that this can also fit into the wider scope of Mechanistic Interpretability. The author of this paper utilized a BabyBERTa model trained on increasingly complex texts to model human development. We think this is also a very good opportunity to see how the *model* changes as well, discovering or visualizing how different parts of the model adapt when trained on language with increasingly complex semantics or relationships.

Title	Activation Patching Reveals Language-Agnostic Concept Representations in Transformers
Venue	ICML Workshop for Mechanistic Interpretability
Year	2024
URL	`https://arxiv.org/pdf/2411.08745v3`

Table 2: Paper 2 [1].

Background. The paper we chose investigates how LLM's internally represent different concepts in different languages. The key question is whether transformers separate meaning from language, meaning that a concept (e.g., "dog") is encoded independently of whether it appears in English, French, or Chinese. The authors analyze this using activation patching, where they swap activations between different languages to see if the model still correctly processes meaning.

Summary of contributions.

- The authors show that LLMs store conceptual meaning separately from linguistic form in mid-to-late layers of the model. Lower layers encode language-specific information, while deeper layers represent abstract, language-independent meanings.
- By patching activations from one language into another, they demonstrate that mid-layer activations encode shared, language-agnostic representations of concepts. Translations are still coherent even if activations from different language translations are injected into mid-layer activations (provided the content of the translations are the same)
- They even show that injecting a "concept representation" from one language into another can improve translation accuracy which implies that there may be a shared space for semantic meaning.

Limitations and discussion. Firstly, the paper considered only very simple text for translations. Perhaps if the prompts are much more detailed, the internal representations for concepts in different languages vary greatly. Additionally some words in English don't have corresponding meanings in

German and vice versa - they didn't explore this asymmetry in their paper. Finally, while activation patching reveals correlations, it does not necessarily prove causation in how meaning is formed in LLMs.

Why this paper? This paper is relevant to our project because it uses activation patching to isolate internal representations, just like we aim to do with linguistic complexity (child vs. adult language). We certainly are inspired by these findings and hope we can use a similar methodology to discover whether linguistic complexity is encoded in specific layers of a model.

Wider research context. While this paper is "translation" heavy, it definitely has broader implications. If transformers encode conceptual meaning separately from language, this suggests that other abstract representations (like complexity, formality, or even age-related speech differences) may also be encoded distinctly.

3 Project description (1-2 pages)

Goal. Our goal for this project is to try to understand the developmental dynamics of how LLM's learn complex patterns in language. LLM's have recently emerged and been popularized due to their success on natural language tasks. However, these models are effectively black boxes; very little is actually known about how these models are able to succeed in tasks the way they do. We intend to fine-tune two LLM's: one on child-directed speech (children's storybooks etc.) and the other on formal adult text. Using mechanistic interpretability (MI) techniques, our aim is to extract activations layer by layer and identify where lexical richness and syntactic complexity diverge.

Specifically, we ask the following questions: How do different layers of a transformer language model encode linguistic complexity? Can activation patching causally manipulate these representations to shift a model's behavior towards simpler or more complex language processing? Does an LLM implicitly encode age-related linguistic differences?

We hope this study will provide insight into the following:

- **LLM Developmental Learning:** How training on child-directed vs. adult language affects learning representations.
- **Mechanistic Interpretability:** Uncovering neural LLM activations responsible for lexical richness, semantics and maturity/age of language.

A more lofty and secondary goal is to use this project to compare LLM learning to human language acquisition. Perhaps we will be able to sequentially isolate specific layers in an LLM that evolve as the complexity of the data increases. By tracking how different layers change in response to varying linguistic complexity, we may be able to draw parallels to cognitive language development in humans. If certain layers demonstrate gradual shifts in representation—such as moving from simpler phrase structures to more abstract syntactic patterns—this could suggest that LLMs undergo a form of hierarchical learning similar to human developmental stages.

Task. Using activation patching, we will:

1. Extract activations from a given layer in the child-trained model.
2. Inject those activations into the corresponding layer of the adult-trained model.
3. Measure whether this intervention shifts lexical and syntactic complexity.

Data. We plan to use the BabyLLM (10M words) dataset for child-directed text, along with formal datasets such as Wikipedia and news corpora for adult-directed text. Preprocessing will include: tokenization and cleaning of raw text and length normalization to ensure comparable inputs.

Baselines. For activation patching, our baselines will consist of the fine-tuned models that have not been edited any further, and we will compare results to those which have modified layers through activation patching. We will make comparisons using metrics defined below in "Evaluation". We will also compare both fine-tuned (but unpatched) models with each other, to discover where the values and weights of the models significantly diverge.

Evaluation. We hope to provide a variety of visualizations and comparisons between models trained on simple language, and those trained on complex language. We might also provide quantitative measures of difference in the transformers' weights, layers, etc. to emphasize locations of major differences between the trained models. We plan to use external libraries such as spaCy to quantitatively measure a generated text's lexical richness, as well as other tools such as K-Means clustering on word embeddings to display word complexity and model evolution. We may also use other metrics of lexical richness (such as Type-Token Ratio), sentence length and proportion of dependent clauses, and potentially qualittative measures of performance as well.

Analysis. We plan to visualize model neurons, weights, attention layers, etc. using the TransformerLens library, providing an effective way to visually compare the differences in the models fine-tuned on simple versus complex language. After activation patching, we will also utilize metrics above in "Evaluation" to analyze which layers within our models bring the most weight in terms of the resulting output's lexical richness, diversity, semantic complexity, etc.

Ethical considerations. The main ethical considerations for this project are relating to data privacy and the existence of biases. Because we are using datasets of language texts that are sourced from real conversations, we run the risk of revealing information about the original human subjects. We hope to mitigate this risk by using public datasets such as HuggingFace's BabyLLM datasets (which are likely to be anonymized/processed), but also to exercise caution when sourcing other data, anonymizing and removing identifiable parts as necessary.

We are also aware that humans may exhibit bias inherently in their language. Thus, our model trained on human speech may reflect these biases. We hope that our task of interpretability is less impacted by bias than say, text generation or generative language modelling, but will take care that any potential risks of bias in our results will be clearly explained.

References

[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Association for Computational Linguistics (ACL)*, 2018.