prompt_question: you are an expert in the research of llm mechanistic interpretability, how do you utilize residual stream for research in mechanistic interpretability

ANS from llama3.1-405b provided at the Emergency Llama 3.1 Hackathon at AGI House:

As an expert in LLM (Large Language Model) mechanistic interpretability, I'd be delighted to share how I utilize residual streams for research in this area.

Background

Mechanistic interpretability aims to understand how complex models like LLMs work by identifying the underlying mechanisms and components that contribute to their behavior. Residual streams, a concept introduced in the Transformer architecture, have been instrumental in advancing our understanding of these models.

What are residual streams?

In the Transformer architecture, residual streams refer to the concatenation of the input embeddings with the output of each encoder layer. This allows the model to preserve the original input information while incorporating the learned representations from each layer. The residual stream is computed as:

`residual_stream = input_embeddings + layer_output`

Utilizing residual streams for mechanistic interpretability

To utilize residual streams for mechanistic interpretability research, I employ the following methods:

1. **Layer-wise analysis**: By examining the residual stream at each layer, I can identify how the model's representations evolve throughout the network. This helps to pinpoint the specific layers and components responsible for particular linguistic phenomena or behaviors.
2. **Feature importance**: By analyzing the residual stream, I can determine the importance of individual input features (e.g., specific words or tokens) on the model's predictions. This is achieved by measuring the magnitude of the residual stream's values corresponding to each input feature.
3. **Path attribution**: I use the residual stream to attribute the model's predictions to specific paths or sequences of computations within the network. This involves analyzing how the residual stream changes as the input propagates through the model, allowing me to identify the most influential components.
4. **Hidden state analysis**: By examining the residual stream in conjunction with the hidden states of the model, I can gain insights into how the model's internal representations are transformed and updated during processing.
5. **Ablation studies**: I perform ablation studies by selectively removing or modifying components of the residual stream to assess their impact on the model's behavior. This helps to identify the essential components and mechanisms contributing to specific phenomena.
6. **Comparative analysis**: I compare the residual streams of different models or model variants to identify similarities and differences in their mechanisms and representations. This facilitates the transfer of insights and knowledge between models.

Tools and techniques

To facilitate the analysis of residual streams, I employ a range of tools and techniques, including:

1. **Tensor decomposition**: I use tensor decomposition methods (e.g., PCA, t-SNE) to reduce the dimensionality of the residual stream and identify patterns and structures.

2. **Visualization**: I utilize visualization tools (e.g., heatmaps, scatter plots) to represent the residual stream and its transformations throughout the model.
3. **Clustering**: I apply clustering algorithms to group similar components or patterns in the residual stream, facilitating the identification of common mechanisms.
4. **Correlation analysis**: I perform correlation analysis to examine the relationships between the residual stream and other model components or outputs.

By leveraging residual streams and these analysis techniques, I can gain a deeper understanding of the complex mechanisms underlying LLMs and shed light on the intricate processes that govern their behavior.