There is a lot of literature showing that LLMs/transformers/neural networks are more linear than we think:
- https://arxiv.org/pdf/2311.04930
- http://arxiv.org/html/2303.09435v2 (especially in the middle layers)
- https://arxiv.org/abs/2311.03658
- https://arxiv.org/pdf/2506.06609
These papers show:
- The mapping from one layer's activation to the next is approximately linear
- The mapping from the activation at one token to the activation at the next token is approximately linear (especially at middle layers)
- You can map from one model's activations to another model's activations with a linear mapping

The goal of this project would be to characterize the linearity of transformer/LLM representations and see how this can help us to build simple models for LLM/transformer behavior/output generation/training, to which we can then apply the rich mathematical literature that exists around linear mappings (dynamical systems, Markov chains, etc.), thus helping to provide a more comprehensive theory of how LLMs/transformers work.