# Thesis

Albert Ramus Sidenius Garde (s183969)

June 5, 2024

# Contents

# Chapter 1

# Introduction

Large Language Models (LLMs) based on the transformer architecture (**?**) have shown exceptional performance across a range of tasks, yet their complexity renders their inner workings opaque. This opacity challenges our ability to understand, trust, and safely deploy these models in real-world applications. The field of mechanistic interpretability aims to address this issue by providing insights into the behaviour of these models. Transformer models contain both attention and multi-layer perceptron (MLP) layers, and the latter is the focus of this thesis. Most attempts at interpreting MLP neurons have focused on understanding the behaviour of individual neurons, but as demonstrated in **?**, the behaviour of individual neurons often doesn't map onto human understandable concepts. **?** shows a possible way forward by suggesting the features of SAEs (Sparse Autoencoders) as alternative units of interpretability. This thesis aims to apply the N2G (**?**) method to the features of SAEs, with the goal of understanding the behaviour of these features and the potential for using this understanding to interpret the behaviour of LLMs. If SAEs truly do provide more interpretable features and N2G truly does provide a useful representation of feature behaviour, we would expect this to be reflected when comparing N2G graphs for individual neurons against those for features. This means that the results of this thesis will inform the usefulness of these two methods.

Info on what comparisons we expect to make.

- Recently LLMs have bla bla bla
- Many benefits, but also worries of harm
  - List a few types of harm
- One issue is models are opaque. We cannot understand their behaviour to foresee or guarantee against harmful behaviour, and we have few options to correct harmful behaviours we do know about.
- Interpretability aims to address this issue by providing insights into the internal workings of these models.
- Many approaches. See **?** for a review
- We focus on...
- Specifically these existing methods
- And we do this
  - Provide a review of work on sparse autoencoders
  - Perform ¡insert experiment here¿