# Thesis

Albert Ramus Sidenius Garde (s183969)

February 15, 2024

# Contents

**1 Introduction** **2**

# 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. To rectify this, the overlapping fields of Explainable AI (XAI) and Mechanistic Interpretability (MI) have emerged. The former focuses on developing methods to explain the predictions of LLMs, while the latter focuses on understanding the inner workings of LLMs. Transformer models contain both attention and MLP (multi-layer perceptron) 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.