**Universidad ORT Uruguay**
**Facultad de Ingeniería**

# Exploring Attention Patterns and Neural Activations in Transformer Architectures for Sequence Classification in Context Free Grammars

Entregado como requisito para la obtención del título de Ingeniería en Sistemas

Matías Molinolo De Ferrari - 231323

Tutores: Dr. Sergio Yovine, Dr. Franz Mayr

**2024**

# Declaración de Autoría

Yo, Matias Molinolo De Ferrari, declaro que el trabajo que se presenta en esta obra es de mi propia mano. Puedo asegurar que:

- La obra fue producida en su totalidad mientras realizaba el Proyecto Final de Ingeniería en Sistemas;
- Cuando he consultado el trabajo publicado por otros, lo he atribuido con claridad;
- Cuando he citado obras de otros, he indicado las fuentes. Con excepción de estas citas, la obra es enteramente mía;
- En la obra, he acusado recibo de las ayudas recibidas;
- Cuando la obra se basa en trabajo realizado conjuntamente con otros, he explicado claramente qué fue contribuido por otros, y qué fue contribuido por mi;
- Ninguna parte de este trabajo ha sido publicada previamente a su entrega, excepto donde se han realizado las aclaraciones correspondientes.

Matias Molinolo De Ferrari

**dd-10-2024**

# Agradecimientos

{DEDICATORIA}

{AGRADECIMIENTOS}

# Abstract

Cuerpo del Abstract.

# Abstract Español

Cuerpo del Abstract.

# Palabras clave

tag1; tag2; tag3

# Key words

tag1; tag2; tag3

# Contents

# 1 Introduction

Large Language Models (LLMs) have been a topic of interest in the field of Computer Science for the past few years, and more recently, with the release of ChatGPT [1] by OpenAI, they have become a topic of interest for the general public too.

These models are based on an architecture called Transformer [2], a type of Artificial Neural Network (ANN) well suited to process sequences, such as text. These models have grown exponentially, as seen in 1.1, both in size and complexity, in the last years, and have shown to achieve state-of-the-art results in a wide variety of Natural Language Processing (NLP) and Natural Language Understanding (NLU) tasks.

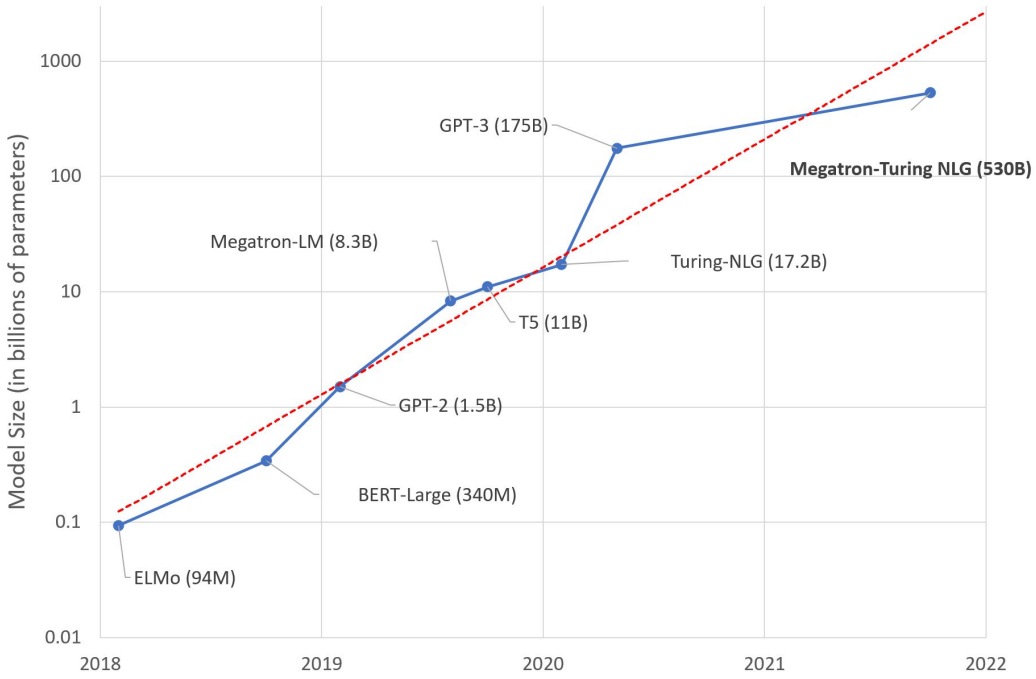


Figure 1.1: Model Sizes (2018–2021) [3]

However, despite their state-of-the-art performance, the inner workings of these models are not yet fully understood and these models are still considered opaque or *black-box* [4], which is a problem for their adoption in critical applications, such as healthcare or finance where decisions need to be explainable and interpretable.

Moreover, there is still a boundary to the capabilities of these models, regarding which problems can or cannot be solved by them and what can be learned and expressed by these models.

A discussion on the differences between learning and expressivity in these models is necessary to define our research question, as the gist of this problem is linguistics.

The objective of this thesis is to explore the attention patterns and neural activations in Transformer architectures for sequence classification, more specifically, in context-free grammars. This work is based on the hypothesis that the attention patterns and neural activations in these models can be used to explain the decisions made by the model, and that these explanations can be used to improve the model's performance and interpretability.

**Outline**

Chapter 1 will introduce the concepts behind formal languages and context-free grammars, focusing especially on the Chomsky Hierarchy and Dyck-$k$ languages. Chapter 2 will introduce the Transformer architecture, with a focus towards the attention mechanism. Chapter 3 will discuss related works and the state-of-the-art in the field of explainable AI and formal languages Chapter 4 focuses on the experimental setup, the dataset used, the model architecture, the training process and the obtained results. Chapter 5 will discuss the results obtained and the implications of these results. Chapter 6 will summarize the work done and propose future work.

# 2 Bibliography

[1] OpenAI, Nov 2022. [Online]. Available: https://openai.com/index/chatgpt

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[3] J. Simon, "Large language models: A new moore's law?" Oct 2021. [Online]. Available: https://huggingface.co/blog/large-language-models

[4] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, J. Su, K. Duh, and X. Carreras, Eds. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. [Online]. Available: https://aclanthology.org/D16-1011

# 3 Annexes

## 3.1 Annex 1

## 3.2 Annex 2