GPT (to be changed)

Michael Gathara
University of Alabama at Birmingham

Akshar Patel
University of Alabama at Birmingham

Vaishak Menon
University of Alabama at Birmingham

Jason Liu
University of Alabama at Birmingham

Elizabeth Molen
University of Alabama at Birmingham

Trenton Davis
University of Alabama at Birmingham

***Abstract*—Large language models (LLMs) are a subset of artificial intelligence (AI) models that have the ability to intake, understand, and generate human readable text. Since their inception, these models have found rapid success and widespread adoption in a vast range of industries and use cases including software engineering, healthcare, insurance, and supply-chain management. The most powerful LLMs are often built on an architectural framework called a transformer which calculates importance of different parts of the input and weighs the relationship between them in order to produce more coherent and relevant outputs. A specific type of LLM- generative pretrained transformers (GPTs)- stand out as one of the most used and adopted models as they are tuned to be highly effective at understanding conversational context and generating coherent and relevant text. Built on top of the transformer architecture, GPTs are often more finely tuned to perform specific tasks after first being trained on a large corupus of less specific textual data- hence the pre-trained portion of the name. While the transformer has remained the core of these models, many different optimizations have been proposed and implemented over the past several years including flash attention, blank, and blank. In this work, we implement a scratch-built GPT equipped with some of these optimizations, train it on several different datasets, and then test it on several tasks and analyze some of the features of the model.**

I. Introduction

In their seminal 2017 paper, Vaswani et al. introduced the transformer architecture and set the stage for the many iterations of LLMs that have come in the years since [**?**]. This original paper focused on the core attention component of the transformer which calculates the interactions between different input tokens and uses this contextual knowledge to more dynamically generate relevant output text.

II. Methods

A. Model Architecture

B. Training

C. Evaluation and Introspection

III. Results

IV. Discussion

V. Conclusion