Imperial College London

MEng Individual Project

Department of Computing

Imperial College of Science, Technology and Medicine

A New Scalable Runtime for LLM Inference

Author:
Hamish McCreanor

Supervisor:
Peter Pietzuch

January 6, 2025

Submitted in partial fulfillment of the requirements for the MEng Computing of Imperial College London

Contents

1 Introduction **2**

2 Background **3**
2.1 Preliminaries 3
2.1.1 Transformer Architecture 3
2.1.2 KV Cache 3
2.2 Related Work 3
2.2.1 PagedAttention 3
2.2.2 FlashAttention 3

3 Project Plan **4**
3.1 Identifying Existing Performance Bottlenecks 4
3.2 Runtime Implementation 4

4 Evaluation Plan **5**

5 Ethical Issues **6**

Chapter 1

Introduction

Chapter 2

Background

2.1 Preliminaries

2.1.1 Transformer Architecture

2.1.2 KV Cache

2.2 Related Work

2.2.1 PagedAttention

2.2.2 FlashAttention

Chapter 3

Project Plan

3.1 Identifying Existing Performance Bottlenecks

3.2 Runtime Implementation

Chapter 4

Evaluation Plan

Chapter 5

Ethical Issues