A Survey on Private Inference for Large Language Models

Michael Shell
School of Electrical and
Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: http://www.michaelshell.org/contact.html

Homer Simpson
Twentieth Century Fox
Springfield, USA
Email: homer@thesimpsons.com

James Kirk
and Montgomery Scott
Starfleet Academy
San Francisco, California 96678-2391
Telephone: (800) 555–1212
Fax: (888) 555–1212

***Abstract*—The abstract goes here.**

1. Introduction

This demo file is intended to serve as a "starter file" for IEEE Computer Society conference papers produced under LaTeX using IEEEtran.cls version 1.8b and later. I wish you the best of success.

mds
August 26, 2015

2. Background

2.1. Transformer-based Language Models

KV Cache

2.2. Inference Process of LLMs

Prefill & Decode
Difference with general server-client model

2.3. LLM Serving Systems

Academic and industrial systems
Trends in LLM serving systems

- Paging, Chucked Prefill
- PD disaggregation

2.4. Privacy Threats

Disclosed privacy threats
Private Cloud Compute

3. Taxonomy

5 points of PCC.

4. Concurrency Optimization

4.1. Parallel Processing

Speculative inference. Pipeline Parallelism. Sequence Parallelism. Tensor Parallelism.

4.2. Batch Processing

Iteration-level batch, chunked prefill, prepack prefill.

5. I/O and Memory Optimization

5.1. Memory Management

paging, disk offloading, prefix caching, MQA, GQA.

5.2. Transmission

Duplication. Pulling. request migration. disaggrated serving.

5.3. Scheduler

priority-based, stateful scheduler. local schduler, instance flip.

Global profiling. request-level prediction

6. Discussion

6.1. Stateful Scheduler

6.2. Verification

7. Conclusion

The conclusion goes here.

Acknowledgments

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.