CSCE 689-609
Special Topics in Programming
Large Language Models (LLMs)

Jeff Huang
jeff@cse.tamu.edu
o2lab.github.io

Course information

○ https://classroom.google.com (using your tamu account)

Class code: **fjzt7kf**

Grading policy

Lecture Participation	*10%*
Homework Assignments	*20%*
Paper Review and Presentation	*20%*
Course Project	*50%*

The grading scale will be:

- A = 90 - 100
- B = 80 - 90
- C = 70 - 80
- D = 60 - 70
- F = < 60

Homeworks

HW0 (1pt):

- Evaluate top-3 leading frontier LLMs with your own tests
  - Prepare 10 super difficult questions

HW1 (4pt):

- Write a Python program that uses LLMs to find and patch security vulnerabilities
  - In a Mock Challenge Project from AIxCC

HW2 (5pt):

- Reproduce GPT-2

HW3 (10pt):

- Write a personal AI assistant that can
  - Write and send emails on your behalf
  - Schedule meetings for you
  - Search the Internet
  - Read multiple PDF files and answer questions
  - Ask you questions, e.g., for your private information or when uncertain

Course project

Scope:

- build a new application based on LLMs
- improve an existing LLM-based technique
- apply an existing technique to a new domain
- other relevant ideas

- Project proposals:
  - propose your own project or choose one of the selected projects
  - submit a proposal and have it approved by me

- Required submissions:
  - All code, tests, documents in github
  - A final project report

Selected projects

- **Train and evaluate a variant of GPT-2** (based on a modified transformer)
- **Local LLM-based tools**
  - Develop a sensitive data cleaner
  - Develop a private data blocker in browser
  - Develop a real-time audio transcription in browser
  - Develop a form auto-fill in browser
  - Develop a video summarizer for Youtube
- **Cybersecurity tools**
  - Develop a LLM-based tool to detect and patch bugs in Linux kernel
- **LLM inference performance**
  - Optimize a local LLM using existing or new quantization techniques
  - Speculative decoding for whisper models
  - Inference Llama 3.1 in one file of pure C (based on karpathy/llama2.c)

Selected papers

- **Attention is All You Need**. 2017. https://arxiv.org/pdf/1706.03762
- **Scaling Laws for Neural Language Models**. 2020. https://arxiv.org/pdf/2001.08361
- **Chain of Thought**: 2022. https://arxiv.org/abs/2201.11903
- **ReAct**: https://arxiv.org/abs/2210.03629
- **Speculative Decoding**. 2022. https://arxiv.org/abs/2211.17192
- **AWQ: Activation-aware Weight Quantization**. 2023. https://arxiv.org/pdf/2306.00978
- **SmoothQuant** 2023. https://arxiv.org/abs/2211.10438
- **Flash Attention**. 2023. https://arxiv.org/abs/2307.08691
- **Paged Attention**. 2023. https://arxiv.org/pdf/2309.06180
- **The llama 3 herd of models**. 2024. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

- More: https://github.com/Hannibal046/Awesome-LLM

Important Notes

- Zero tolerance: cheating & plagiarism
- Late penalty: 2% per hour

- Materials:
    - Transformer Explainer https://poloclub.github.io/transformer-explainer/
    - Illustrated Transformer: https://jalammar.github.io/illustrated-transformer/
    - Chatbot Area: https://arena.lmsys.org/
    -
    - LiteLLM https://docs.litellm.ai/docs/
-
- Due
    - HW0