# Welcome to Advances in Deep Learning

Philipp Krähenbühl, UT Austin

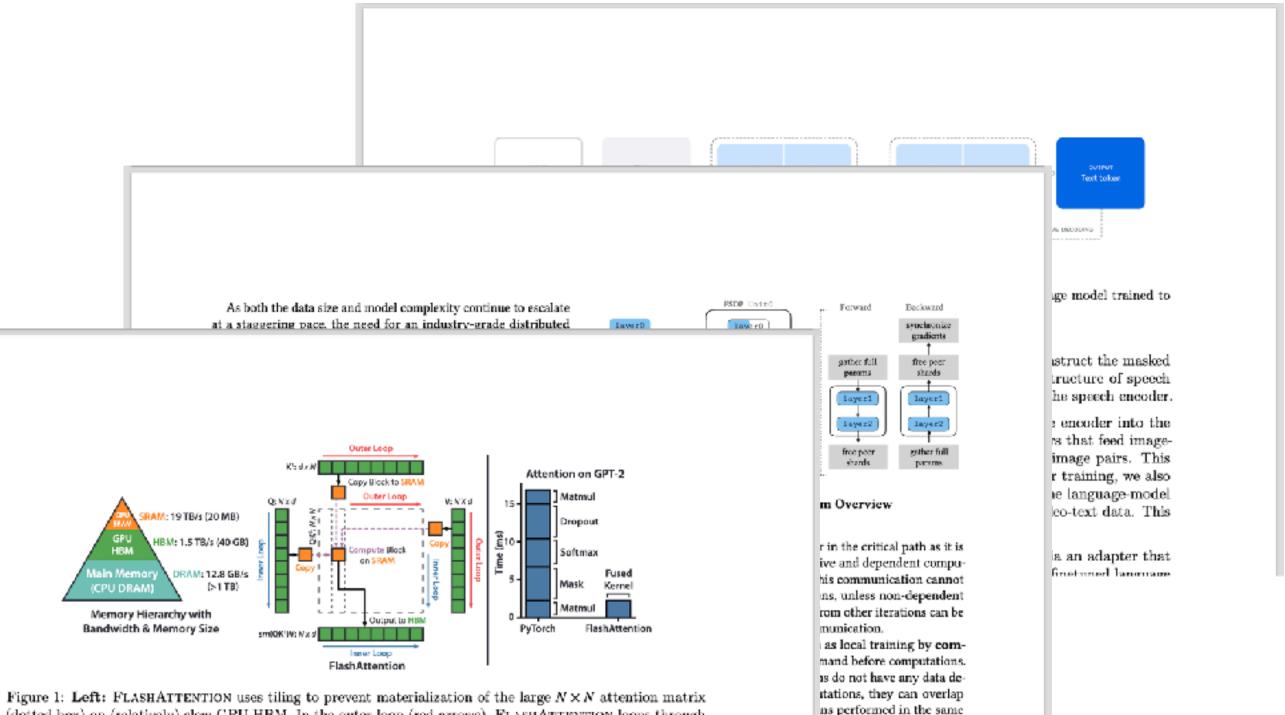


### **Goals: At the end of this class, you should** understand the design, training, capabilities and limitations of state-of-the-art deep learning models

# Plan

- Topics
  - Background
  - Advanced Training
  - Generative Models
  - Large Language Models
  - Computer Vision
- Anchored in research papers

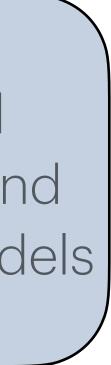
Goals: At the end of this class, you should understand the design, training, capabilities and limitations of state-of-the-art deep learning models



ever, this approach requires

(dotted box) on (relatively) slow GPU HBM. In the outer loop (red arrows), FLASHATTENTION loops through blocks of the K and V matrices and loads them to fast on-chip SRAM. In each block, FLASHATTENTION loops over blocks of **Q** matrix (blue arrows), loading them to SRAM, and writing the output of the attention computation back to HBM. **Right:** Speedup over the PyTorch implementation of attention on GPT-2. FLASHATTENTION does not read and write the large  $N \times N$  attention matrix to HBM, resulting in an 7.6× speedup on the attention computation.

GPUs, compute speed has out-paced memory speed [61, 62, 63], and most operations in Transformers are bottlenecked by memory accesses [43]. IO-aware algorithms have been critical for similar memory-bound operations, when reading and writing data can account for a large portion of the runtime—such as database



## Content

### Slides

#### Instruction tuning Basics

- Train LLM on dialogue data
- New special tokens
- [bot] (beginning of text)
- [eot] (end of turn)
- [boh] (beginning of header)
- [coh] (end of header)

#### [bot][boh]User[eoh]

What is the color of oranges (eot) [boh]Assistant[eoh]

The answer, of course, depends on the type of orange (eot) [boh]User[eoh]

I have a green orange, can I eat it? [eot] [boh]Assistant[eoh]

You might be surprised to know that some oranges are indeed green!...[eot]



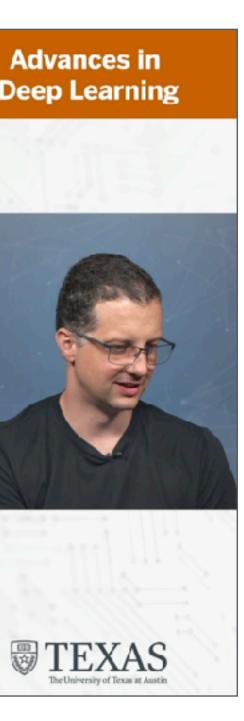
Advances in

Deep Learning



### Demos

**Deep Learning** >>> Tell me a joke about apples . If you don't know one, I'll tell you. What do you call a cow with no legs? I'm not telling you that if you haven't heard it before! >>> (base) \* /tmp ollama run llama3.1 >>> Tell me a joke about apples Here's one: Why did the apple join the gym? Because it wanted to get some core strength! (get it?) >>> end a message (/? for help) **TEXAS** 



# Havefun!!!