AQLM

- Installation required building from scratch, since cpython is utilized,

- pypi install unideal since there were dependency conflicts with our current work. Clone repo and modify install to not require triton and numba (linux libraries).

- issues with specifying compile version. We pass in c++17 explicitly to resolve this.

- Immediately run into DLL error. Spent a lot of time finding what dependency issues there were. Turns out we need to manually include cuda toolkit dlls in code before loading the aqlm dll. Fun

LoRA

Only base lora is codewise compatible with aqlm, so we cannot use derivatives such as adalora.

Priorities: flash attention, figure out how to implement it, look into onyx/deepspeed for inference

NOTE: significant changes made to transformers library: move the changes to external when possible. In modelling\_llama and is\_flash\_attn\_2\_available

Reinforcement learning from human feedback: apply it to conversational modelling

Wow, ok so transformers initializes an aqlm classifier wrongly

Activation beacon not supported since flash attention 2 doesn’t integrate with it

Information Retrieval is a no go, no library support L

Experiments:

- Train generative chat model with tucore gcn inspired architecture

- hyperparameter tune soda\_lora model

Torch compile

Look into calculating attention while ignoring padding: Flash attention takes batched inputs and concatenates their unpadded forms before calculating attention.

Thankfully dgl is compatible

However need to resolve llama issues:

<https://github.com/Dao-AILab/flash-attention/issues/595> cite this ltr

https://arxiv.org/abs/2312.11918