Team 2
Jack Hanke, Daniel Plotkin, Nicole Birova
1 May 2025

NLP Project Proposal

Team 2 will train a 1B parameter decoder-only LLM from scratch named *qt*, with a focus on basic conversation and instruction following. The team was inspired by the *Survey on Large Language Models* paper to create an LLM on a student budget. We plan on following much of the engineering advice proposed in *Cramming: Training a Language Model on a Single GPU in One Day* for student-scale specific language model creation. We additionally plan to quantize the final model so that inference can be run on a consumer-grade graphics card with 16GB VRAM.

Creating an LLM involves preprocessing the data, pretraining, finetuning, quantization, and benchmarking. For pretraining, our focus on conversational ability led us to the *bookscorpus* dataset (~1B Tokens, 6GB via Hugging Face). Pretraining (FP32 params) will be done on a Google Colab 80GB VRAM NVIDIA A100 (paying by hour out of pocket, pretraining cost ~$15 per group member). We confirmed that a FP32 1B parameter model, training with the Adam optimizer, will fit on such a card along with batched data. Usual checkpointing, clipping, early stopping, and progress logging will be implemented. We plan to initialize output logits by adding an initial bias term (log probability of token) based on token frequency from bookscorpus to speed up pretraining and reduce initial entropy. We will report the train, validation, and test cross entropy loss and perplexity on the bookscorpus as our pretraining metric.

For finetuning, we will instruction tune our model on the DialogSum dataset (12k dialogue/summary pairs via Hugging Face)**.** We will conduct full parameter fine tuning, and if time allows reduce the computation by implementing LoRA. For quantization, we are planning on casting our model to FP16, and Int8 if necessary. As we are interested in basic conversation and instruction following, we choose performance on the 11 taks in the GLUE benchmark as our downstream task performance metric. Finally, we will baseline our smaller language model from the class homeworks on cross entropy loss, perplexity, and GLUE performance to compare our efforts.

Overall, the team hopes to explore the world of large generative language models, especially when this world is so often treated as a task strictly for large organizations to take on.