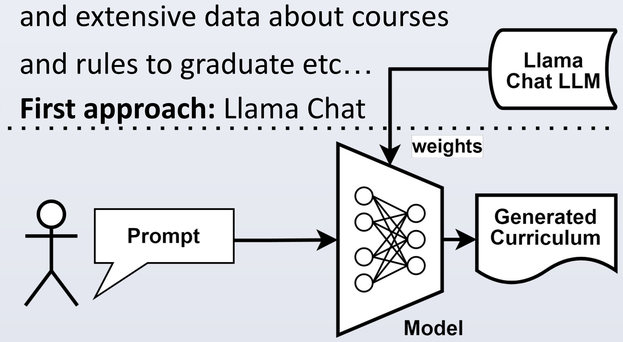
Implementation of a Georgia Tech curriculum advisor chatbot based on qLoRA fine-tuning of Llama 7B

Brieuc POPPER <bpopper3@gatech.edu>, Francisco Nicolás PÉREZ FERNÁNDEZ <ffernandez35@gatech.edu>

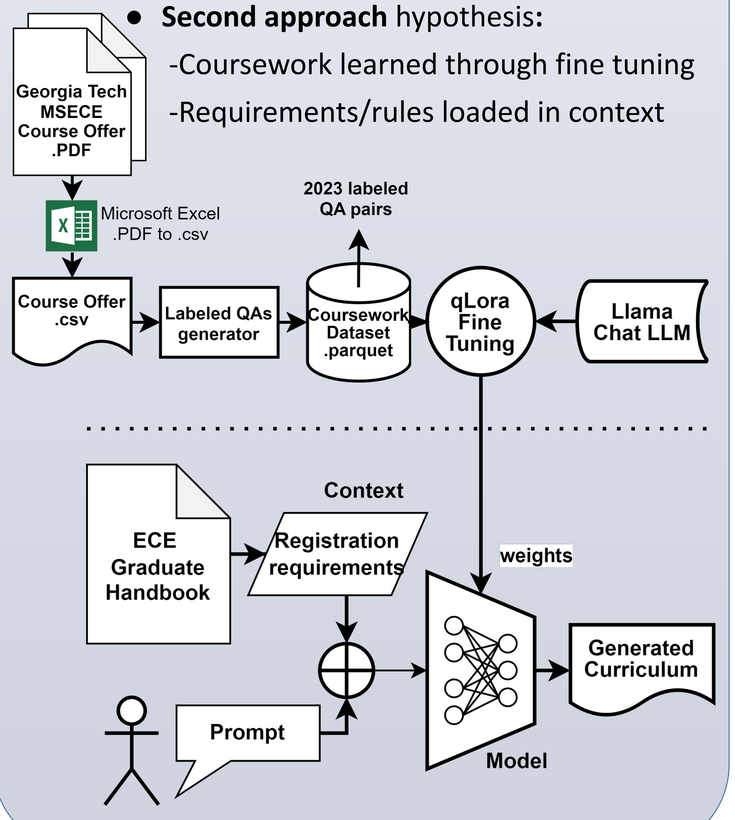
Introduction

- **Goal** : Create a chatbot that acts as a **curriculum advisor** for Georgia Tech's school of ECE
- **How?** Using open-source Llama-7B large language model and using fine-tuning and context.
- **Challenges :** Need to teach the model very precise and extensive data about courses and rules to graduate etc…
- **First approach:** Llama Chat



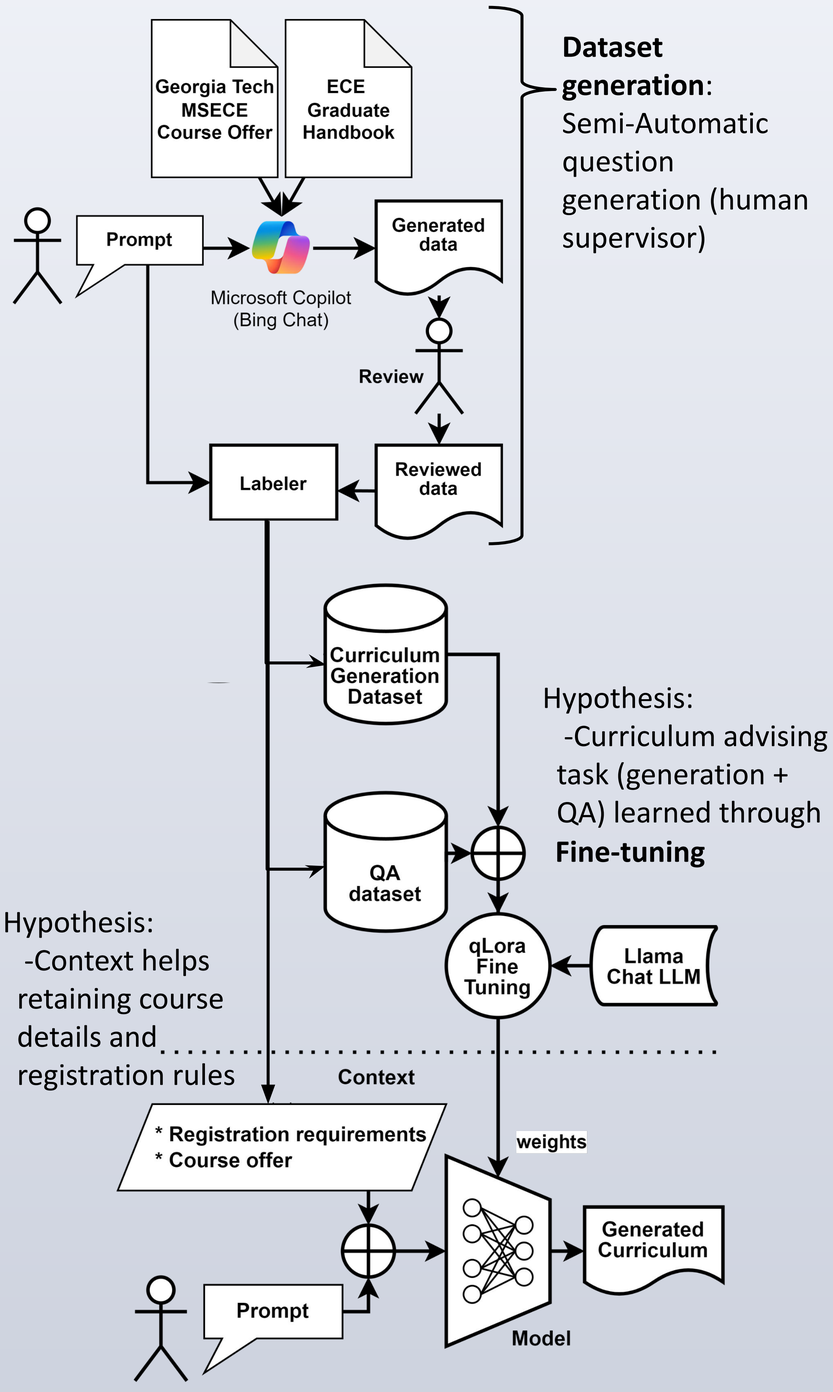
Approach

- **Second approach** hypothesis**:**
  -Coursework learned through fine tuning
  -Requirements/rules loaded in context



- **Third approach: Zero-Shot.** Hypothesis:
  -Course offer and requirements learned from context

- **Fourth approach: Fine-tuning + context**



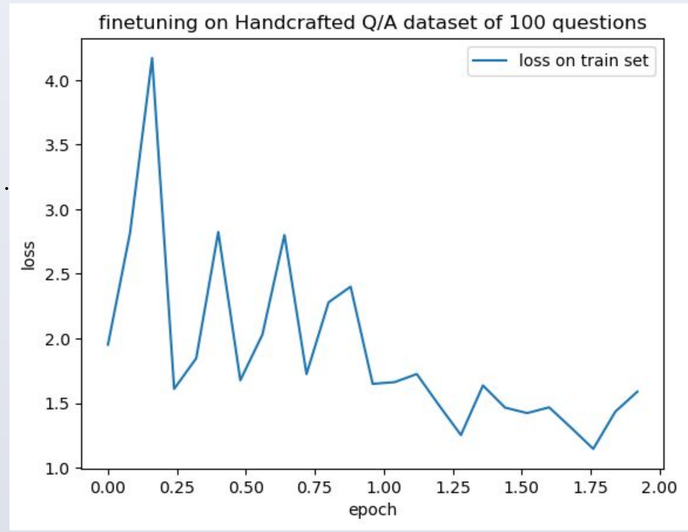
Dataset generation: Semi-Automatic question generation (human supervisor)

Hypothesis:
-Curriculum advising task (generation + QA) learned through **Fine-tuning**

Hypothesis:
-Context helps retaining course details and registration rules

Experiments

- Creating a set of **questions** to ask every model
- Recording every model's answer and perform **qualitative analysis**
- Creating a **scoring metric** to get quantitative results
- Evaluating them all vs Microsoft Copilot (previously Bing AI chat, uses **Retrieval Augmented Generation**)



Conclusions

Case/Model	Correctness	Completeness	Score
Llama Chat 7B (vanilla)	3.00	7.00	**2.45**
Learning from fine-tune	2.91	5.09	**2.27**
Just Context	3.54	5.50	**2.91**
Fine-tuning+Context	4.81	6.82	**4.36**
Microsoft Copilot	5.95	7.27	**5.54**

- **Context is limited:** large context ⟶ need more **GPU memory** & **inference time,** not reliable **long-range**
- Best solution: **Retrieval-augmented generation** or quality fine-tuning on a **diverse labelled data**

References

Understanding Finetuning for Factual Knowledge Extraction from Language Models, 2023
Training language models to follow instructions with human feedback, 2022
Language models are few-shot learners, 2020
LLaMA: Open and Efficient Foundation Language Models, 2023