Capstone Bi-Monthly Update

Date: 09/13/2023

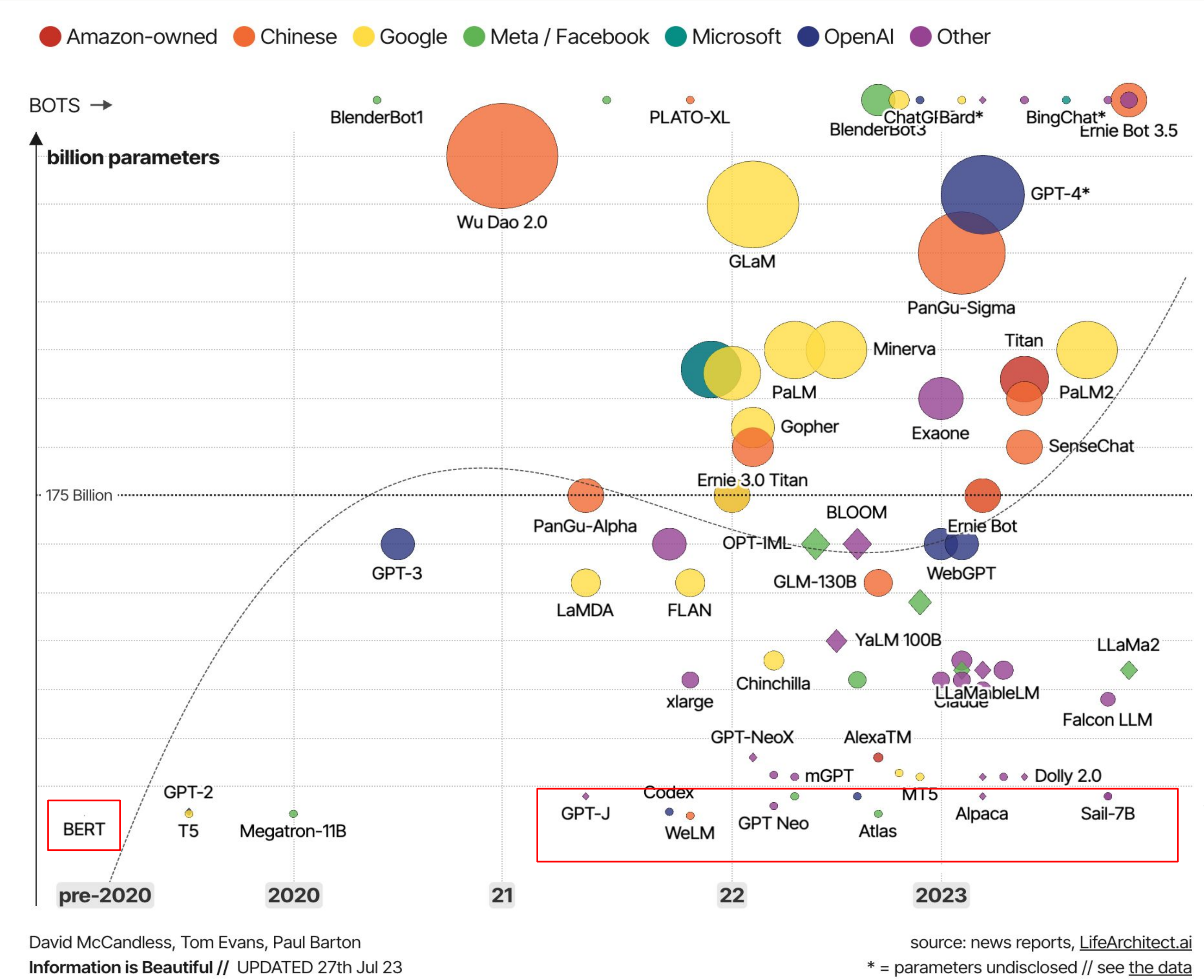
Team: Aditya, Elisa, Jenny, Zhanyi
Mentor: Professor Ye

Duke

Agenda

- Share findings of literature survey
- Get feedback on potential datasets
- Clarification on the library for fine-tuning and how HuggingFace can be incorporated

The Development of Small Models Hasn’t Been The Focus Of The LLM Community



size = no. of parameters open-access

← Our Area of Focus

Duke

LLM Models Overview

Model	Year	# of Parameters	Suitable Use Cases	GLUE Score (if available)	Training Dataset
BERT (Base)	2019	110M	+ Text Classification, NER, POS, Q&A - Text Generation, Machine Translation	79.5	English Wikipedia; Toronto Book Corpus (Total: 3,300M words)
DistilBERT	2020	66M	+ Text Classification, NER, POS, Q&A - Text Generation, Machine Translation	77.0	English Wikipedia; Toronto Book Corpus (Total: 3,300M words)
LLaMa2	2023 (July)	7B, 13B, 70B	+ Text Classification, NER, POS, Q&A - Text Generation, Machine Translation, Code Generation		English CommonCrawl; C4; English WikiPidia; Github;Arxiv

Duke

LLM Models Overview

Model	Year	# of Parameters	Suitable Use Cases	GLUE Score (if available)	Training Dataset
Cerabras-GPT	2023 (March)	13B	WIP	WIP	WIP
Alpaca	2023 (March)	7B	+ Text Classification, NER, POS, Q&A - Text Generation, Machine Translation, Code Generation	WIP	Stanford Alpaca Dataset (52002 pairs of prompt and output generated by OpenAI's text-davinci-003 model)
Koala-13B	2023 (April)	13B	WIP	WIP	WIP

LLM Models Overview

Model	Year	# of Parameters	Suitable Use Cases	GLUE Score (if available)	Training Dataset
Dolly 2.0	2023 (April)	12B (smaller models available at size 2.8B)	+ Q&A	N/A	instruction / response (finetuning: 15k samples; human-generated)
Sail-7B	2023 (June)	7B	WIP	WIP	Alpaca-GPT4 Dataset (English & Chinese); ranked responses from GPT-4, GPT-3.5 and OPT-IML; 9K unnatural Instruction data generated by GPT-4
Open LLM	2023 (June)	13B	WIP	WIP	

Duke

BERT (2019)

- Designed to pre-train deep bidirectional representations from unlabeled text
- Easy and Fast to fine-tune: model has an unified architecture across downstream tasks which was rare at the time.

Training Procedure:

- Step 1 - Masked Language Model: randomly masks some of the tokens from the input
- Step 2 - Next Sentence Prediction (NSP): the model concatenates two masked sentences as inputs during pretraining. The model then has to predict if the two sentences were following each other or not.

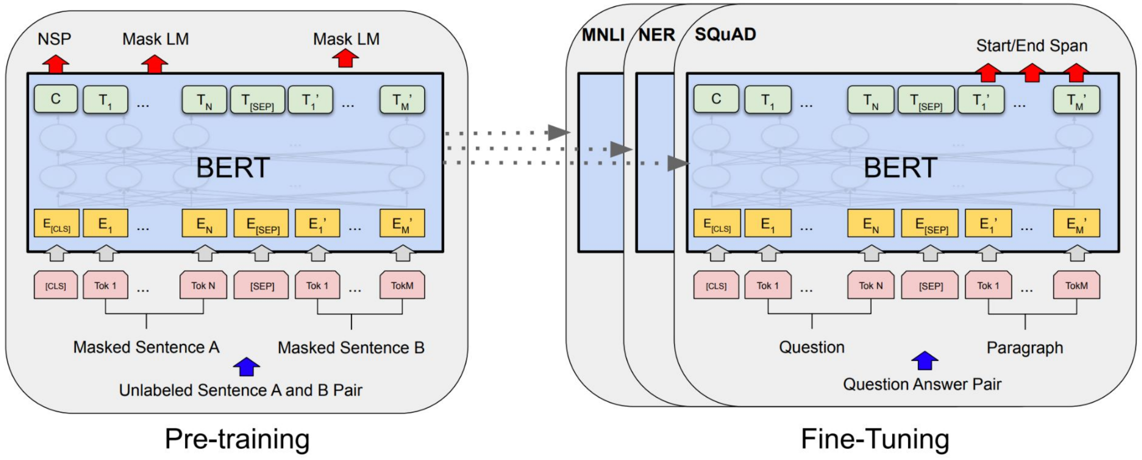


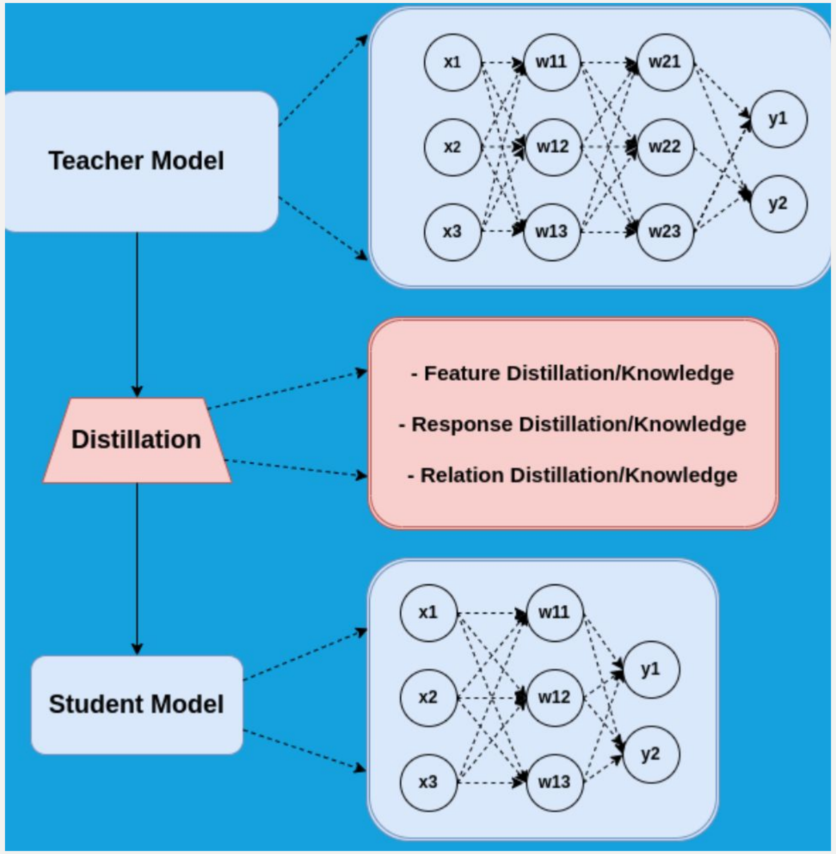
Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Duke

DistilBERT

- 40% less parameters, runs 60% faster while preserving over 95% of BERT’s performance
- A compression technique called knowledge distillation used to create the model (teacher-student training)
- Changes to the architecture: reduced number of layers by a factor of 2, token-type embeddings and poolers removed

Teacher-Student Model



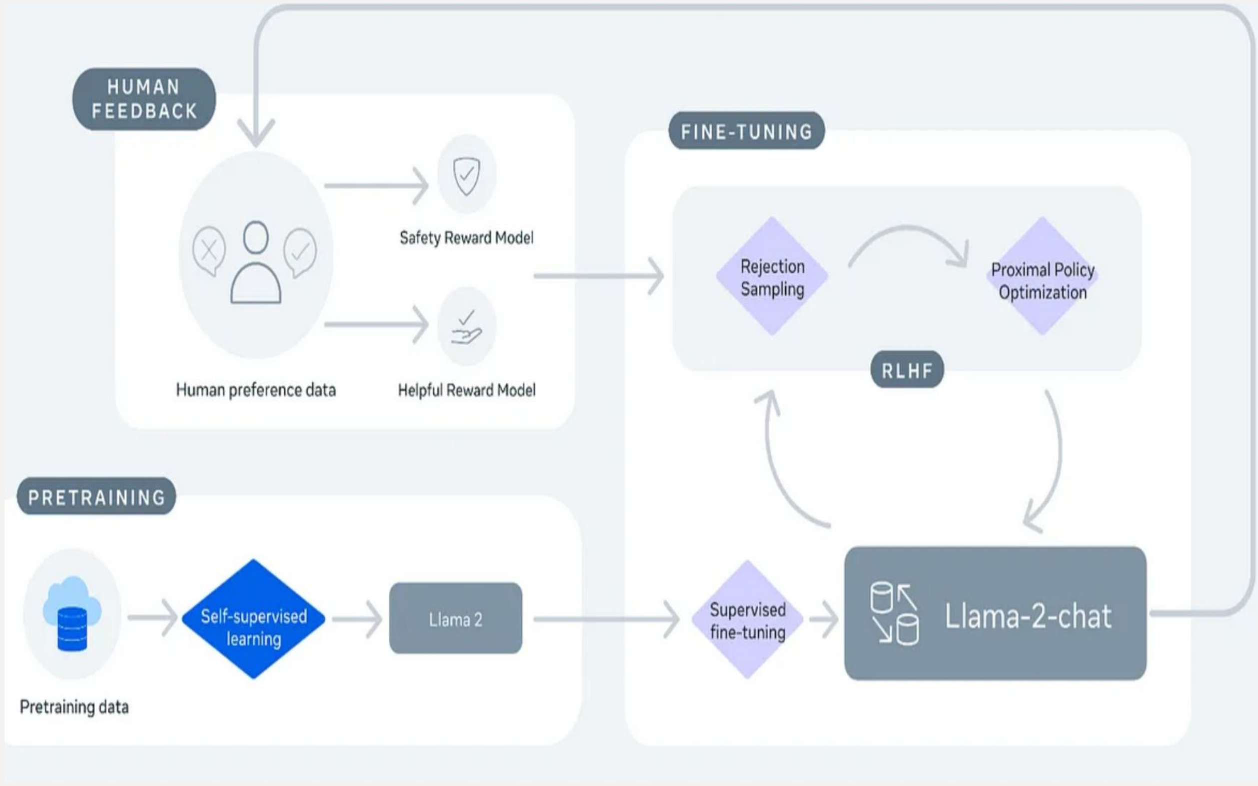
Duke

Llama 2

- Llama 2 is a open source model developed by Meta with three parameter sizes: 7B, 13B, 70B

- Llama 2 adopted grouped-query attention to enhance inference scalability and pre-training data size increase by 40% compared to Llama 1.

- RLHF (Reinforcement Learning with Human Feedback) eliminates diversity in responses to factual prompts but retains more diversity when generating responses to creative prompts.

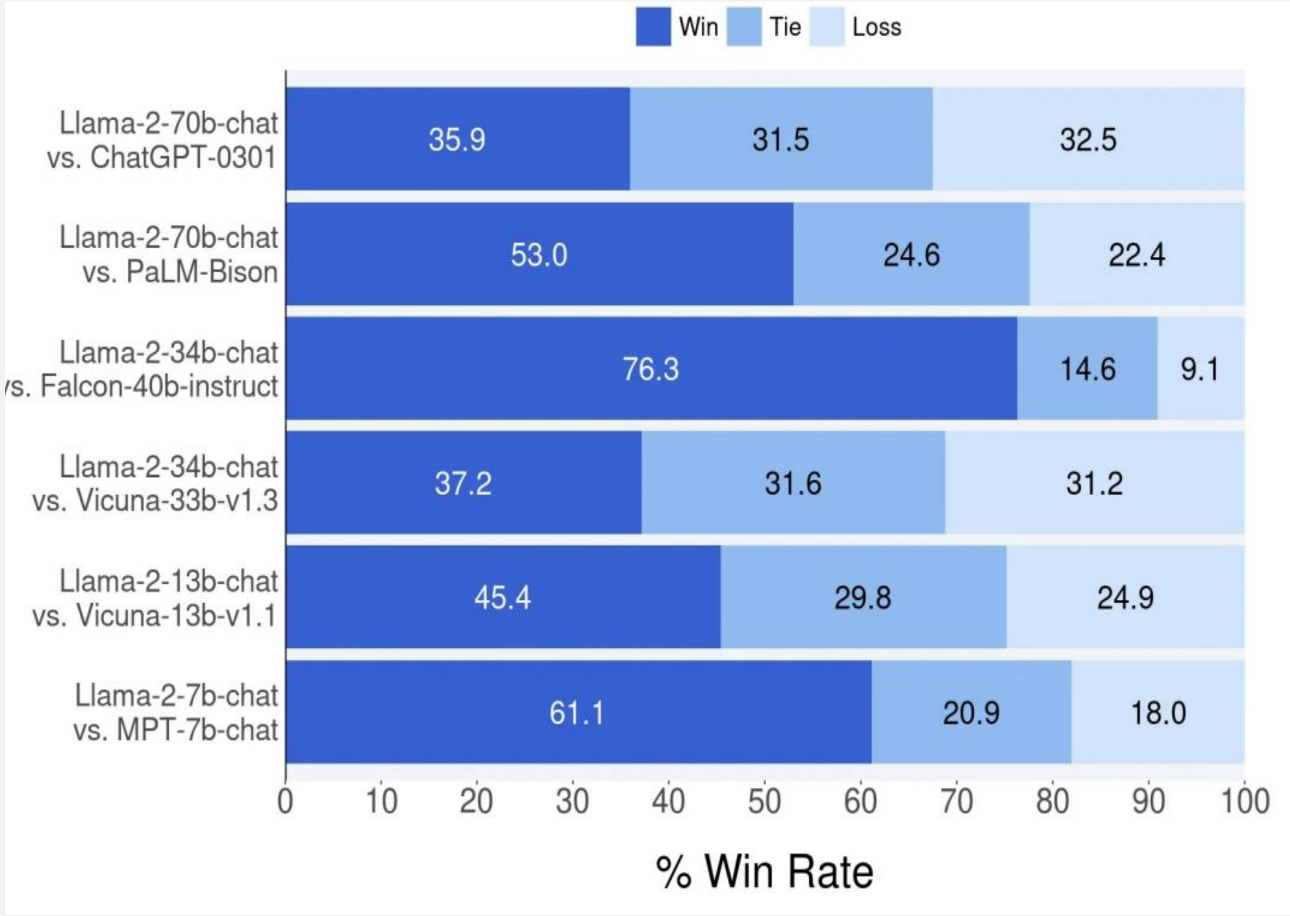


Llama 2

- LLAMA-2 Chat the outperform open-source models by a significant margin(60–75%) on both single-turn and multi-turn prompts and comparable to ChatGPT.

- The initial version of Llama 2-Chat predominantly focused on English-language data.

- Measures have been taken to increase the safety of Llama 2 models, including safety-specific data annotation, tuning, red-teaming, and iterative evaluations.



Duke

MPT 7B - Commercial Open-Source Models

- One of the many commercially open-source models released by companies
- Developed by Mosaic ML (commercial use allowed)
- Foundational model trained on 1T tokens, text and code
- Advantages:
  - Handles long input
  - Optimized for fast training and inference
  - Open-sourced training code
- Fine-tuned variations for different tasks
  - Instruct
  - Chat
  - StoryWriter (Can handle 65k input tokens)

Duke

HuggingFace

- Provides a very easy to use interface to fine-tune Large Language models.
- In addition, it also hosts various datasets and models that can be easily used.

Open Questions:

1) What is the expectation of the fine-tuning library? Is it so that we can later on integrate Parameter Efficient Fine-Tuning and Model Compressions techniques?
2) At what capacity have you used HuggingFace in the past for your projects?

Potential Training Datasets

1. Label Box (https://app.labelbox.com/catalog)

2. Hugging Face Dataset (https://huggingface.co/datasets)

3. Kaggle Datasets (https://www.kaggle.com/datasets)

    - Malware Detection(https://www.kaggle.com/c/malware-detection/data)
    - Malicious and Benign Websites (https://www.kaggle.com/datasets/xwolf12/malicious-and-benign-websites)

4. ML for Cybersecurity Datasets (https://github.com/jivoi/awesome-ml-for-cybersecurity)

5. Samples of Security Related Data (http://www.secrepo.com/)

Next Steps

- Literature Review on Models (contd.)
  - Keep it short. Spend 2-3 days (~3-7B)
- Literature Review on PEFT (Jenny, Aditya)
  - SOTA: Lora and qlora techniques
  - Learn: quantization technique (optional: other model compression and latency reduction techniques)
- Setting up Infrastructure For Small LLM Fine-Tuning (Zhanyi, Elisa)
  - Just set up infrastructure for any small LLM model on a small dataset
  - Github repo, Colab workspace,
  - Use PyTorch
- Continue exploring datasets in cybersecurity space
  - Look into wikiSQL and spyderDB

Duke