FinTech Frankenstein 

By:
Sapan Patel, Aditya Natham, Akaash Dash, Max Zhao, Erik Larson

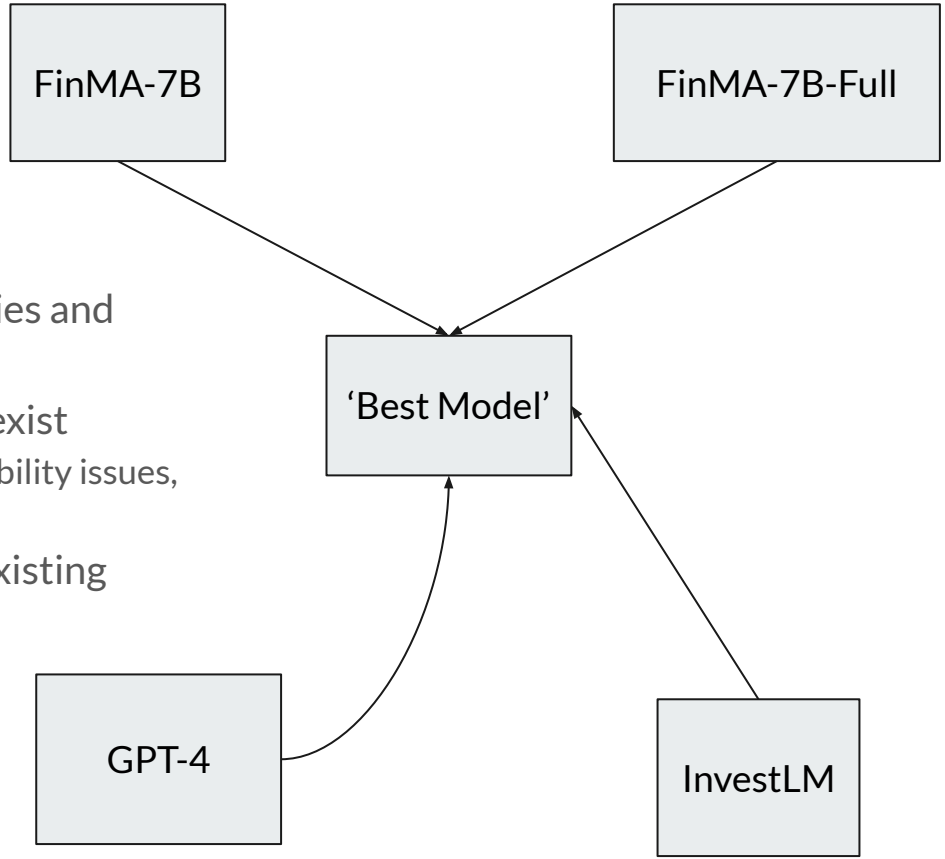
State of FinTech LLMs

- LLMs have exploded in popularity!
- Currently moving very rapidly through FinTech world which has been relatively underdeveloped until recently

Model	Backbone	Size	Open Source		Instruct	Language	Evaluation		Release Date
			Model	Data			NLP	Fin	
finBERT (Araci, 2019)	BERT	110M	✓	✓	✗	English	✓	✗	08/27/19
FinBERT (Yang et al., 2020)	BERT	110M	✓	✗	✗	English	✓	✗	06/15/20
Mengzi-fin (Zhang et al., 2021)	RoBERTa	103M	✓	✗	✗	Chinese	✓	✗	10/13/21
FLANG (Shah et al., 2022)	ELECTRA	110M	✓	✓	✗	English	✓	✗	10/31/22
BBT-FinT5 (Lu et al., 2023)	T5	220M	✓	✓	✗	Chinese	✓	✗	02/18/23
BloombergGPT (Wu et al., 2023)	BLOOM	50B	✗	✗	✗	English	✓	✗	03/30/23
FinMA	LLaMA	7/13B	✓	✓	✓	English	✓	✓	06/01/23

Best Model

- With every new model comes new abilities and new weaknesses
- One singular best model improbable to exist
  - Seesaw effect, different focuses, replicability issues, etc.
- What if we combined the best parts of existing models to make a ‘singular best model?’



Problem Statement

- How can we dynamically choose the best model for a prompt and receive the most effective response?

hotpot.ai/art-generator

Project General Overview

- Research Finance LLM benchmark suites
- Explore different LLM performances on the suites
- Build a classifier to decide a prompt category/task/goal
- Design a router to query the best suited LLM based on the prompt class
- Deploy website for user interactions

Determining Strengths and Weaknesses

- How can we determine the best and worst parts of each model?
  - Benchmark suites to analyze performance across a range of financial tasks and inputs
- Many suites exist, and they have evolved over time
  - FPB
  - FiQA
  - FLUE
  - FLARE

Data	Task	Raw	Instruction	Data Types	Modalities	License
FPB	sentiment analysis	4,845	48,450	news	text	CC BY-SA 3.0
FiQA-SA	sentiment analysis	1,173	11,730	news headlines,tweets	text	Public
Headline	news headline classification	11,412	11,412	news headlines	text	CC BY-SA 3.0
NER	named entity recognition	1,366	13,660	financial agreements	text	CC BY-SA 3.0
FinQA	question answering	8,281	8,281	earnings reports	text,table	MIT License
ConvFinQA	question answering	3,892	3,892	earnings reports	text,table	MIT License
BigData22	stock movement prediction	7,164	7,164	tweets,historical prices	text,time series	Public
ACL18	stock movement prediction	27,053	27,053	tweets,historical prices	text,time series	MIT License
CIKM18	stock movement prediction	4,967	4,967	tweets,historical prices	text,time series	Public

Model Selection/Strengths

- **Used:**

  GPT-4, FinMA-7B, FinMA-7B-Full

- **Didn't use:**

  FinMA-30B

- **Wanted to use:**

  InvestLM

Dataset	Metrics	GPT NeoX	OPT 66B	BLOOM	Chat GPT	GPT 4	Bloomberg GPT	FinMA 7B	FinMA 30B	FinMA 7B-full
FPB	Acc	-	-	-	0.78	0.76	-	0.86	**0.87**	0.87
	F1	0.45	0.49	0.50	0.78	0.78	0.51	0.86	**0.88**	0.87
FiQA-SA	F1	0.51	0.52	0.53	-	-	0.75	0.84	**0.87**	0.79
Headline	AvgF1	0.73	0.79	0.77	0.77	0.86	0.82	**0.98**	0.97	0.97
NER	EntityF1	0.61	0.57	0.56	0.77	**0.83**	0.61	0.75	0.62	0.69
FinQA	EmAcc	-	-	-	0.58	**0.63**	-	0.06	0.11	0.04
ConvFinQA	EmAcc	0.28	0.30	0.36	0.60	**0.76**	0.43	0.25	0.40	0.20
BigData22	Acc	-	-	-	0.53	**0.54**	-	0.48	0.47	0.49
	MCC	-	-	-	-0.025	0.03	-	0.04	**0.04**	0.01
ACL18	Acc	-	-	-	0.50	0.52	-	0.50	0.49	**0.56**
	MCC	-	-	-	0.005	0.02	-	0.00	0.00	**0.10**
CIKM18	Acc	-	-	-	0.55	**0.57**	-	0.56	0.43	0.53
	MCC	-	-	-	0.01	**0.02**	-	-0.02	-0.05	-0.03

BERT Text Classification Process

- BERT is a Bidirectional Encoder Representations from Transformers
  - large open source NLP model that is trained on vast amounts of general data
  - can be fine-tuned for a specific purpose
- We will use BERT to classify the prompt types by these steps:

1. Gather training data consisting of prompts to fine-tune BERT
2. Fine-tune BERT on said training data
3. Evaluate fine-tuned BERT
4. Integrate it into our project

Data Collection

- The chosen FLARE benchmark suite is available on HuggingFace in a dataset collection

```
[ ]  flare_datasets = [
        "ChanceFocus/flare-fpb",
        "ChanceFocus/flare-fiqasa",
        "ChanceFocus/flare-headlines",
        "ChanceFocus/flare-finqa",
        "ChanceFocus/flare-convfinqa",
        "ChanceFocus/flare-finer-ord",
        "ChanceFocus/flare-sm-bigdata",
        "ChanceFocus/flare-sm-acl",
        "ChanceFocus/flare-sm-cikm",
        "ChanceFocus/flare-german",
        "ChanceFocus/flare-australian",
        "ChanceFocus/flare-edtsum",
        "ChanceFocus/flare-fomc",
        "ChanceFocus/flare-ner"
    ]
```

Function to download dataset

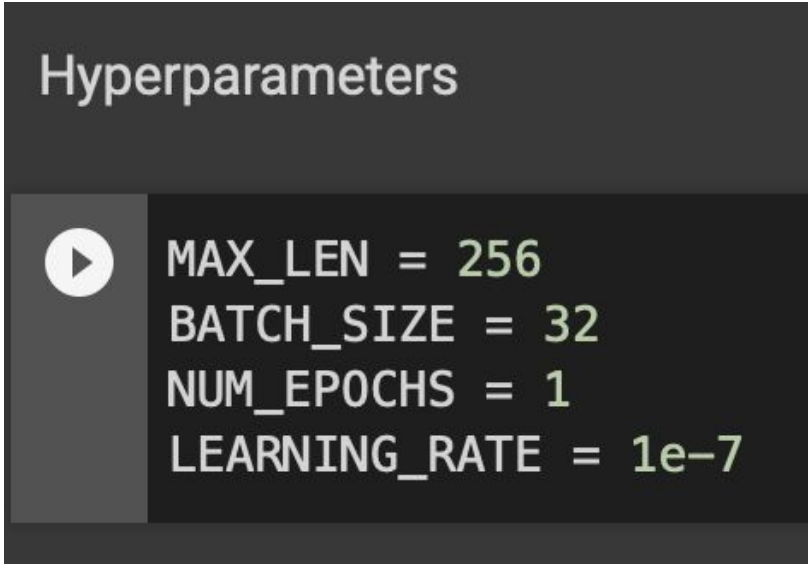
```
[ ]  def download_dataset(dataset_name, selected_columns):
        dataset = load_dataset(dataset_name)
        dataset.save_to_disk(dataset_name)
        print(f"Dataset {dataset_name} and saved to {dataset_name}")
```

Download all datasets

```
[ ]  for dataset in flare_datasets:
        download_dataset(dataset, columns)
```

Fine Tuning

- Used DataLoader and Dataset to create 80/20 train/validate split
- Trained bert-base-uncased tokenizer and classifier with AdamW optimizer



Fine Tuning Results

```
Accuracy: 0.8598601160013648, Precision: 0.8598601160013648, Recall: 0.8598601160013648
```

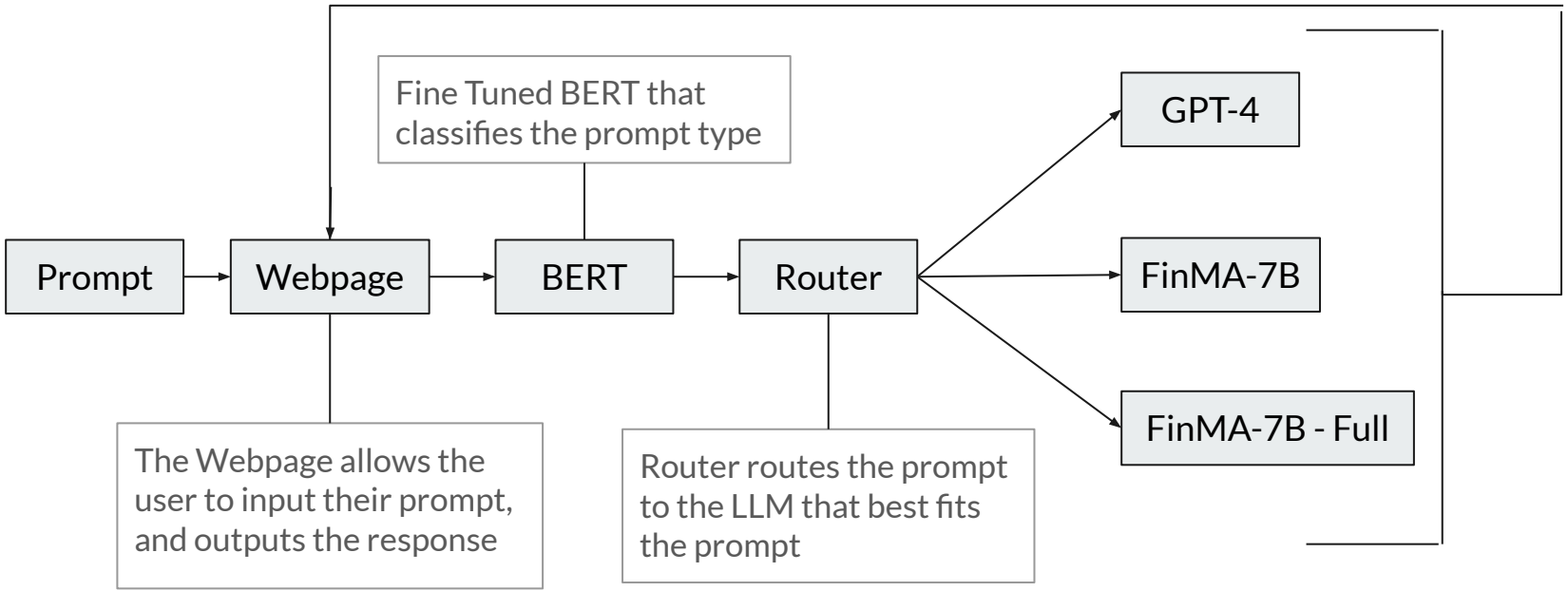
Router Model

- After we have fine tuned BERT, the last step is to create the router
- The router takes the classified prompt, identifies the LLM whose strengths align with the class, and inputs it
- After the LLM returns an output, the router will then propagate the response

Model	fpb	fiqasa	headlines	finqa	conv
GPT-4	0.78	0.8	0.86	0.63	
FinMA-7B-Full	0.88	0.79	0.97	0.04	
FinMA-7B	0.86	0.84	0.98	0.06	

```
def query(prompt):
    routed_model = routes[router.predict(prompt).replace('ChanceFocus/flare-', '')].idxmax()
    return routed_model, models[routed_model].query(prompt)
```

Project Architecture



Final Results

Benchmark	Accuracy	Benchmark	Accuracy
FPB	0.7	SM-ACL	0.4
FiQA-SA	0.8	SM-CIKM	0.1
Headlines	1.0	German	0.3
FinQA	0.4	Australian	0.3
ConvFinQA	0.7	EDTSum	0.4
Finer-ORD	0.7	FOMC	0.1
SM-BigData	0.2	NER	0.8

Discussion

- Model does a very good job of picking the correct model, ensures that prompts will receive the best response
- Testing not truly representative due to the extremely limited nature
- Our responses would be better than a user picked model due to the nature of always picking the strongest model for the task

Current Limitations

- BERT is limited to a 256 tokens max due to colab pro limits
- English-only despite capabilities for more
- Could have explored more with the hyperparameters
- Testing was not as controlled as ideal due to computing limitations

Looking Forward

- Add more LLMs to cover a larger spectrum of categories
- Continue Fine Tuning BERT to account for more categories
- Combining multiple LLMs when answering a question
- Add more recently created benchmarks, such as FinanceBench

References

Anil, R. (2023, May 17). *PALM 2 Technical Report*. arXiv.org. https://arxiv.org/abs/2305.10403

OpenAI. (2023, March 15). *GPT-4 Technical Report*. arXiv.org. https://arxiv.org/abs/2303.08774

Shah, A. (2023, May 26). *Zero is Not Hero Yet: Benchmarking Zero-Shot Performance of LLMs for Financial Tasks*. arXiv.org. https://arxiv.org/abs/2305.16633

Touvron, H. (2023, February 27). *LLAMA: Open and Efficient Foundation Language Models*. arXiv.org. https://arxiv.org/abs/2302.13971

Wu, S. (2023, March 30). *BloombergGPT: A large language model for finance*. arXiv.org. https://arxiv.org/abs/2303.17564

Zheng, S. (2023, September 28). *GPT-Fathom: Benchmarking Large Language Models to Decipher the Evolutionary Path towards GPT-4 and Beyond*. arXiv.org. https://arxiv.org/abs/2309.16583