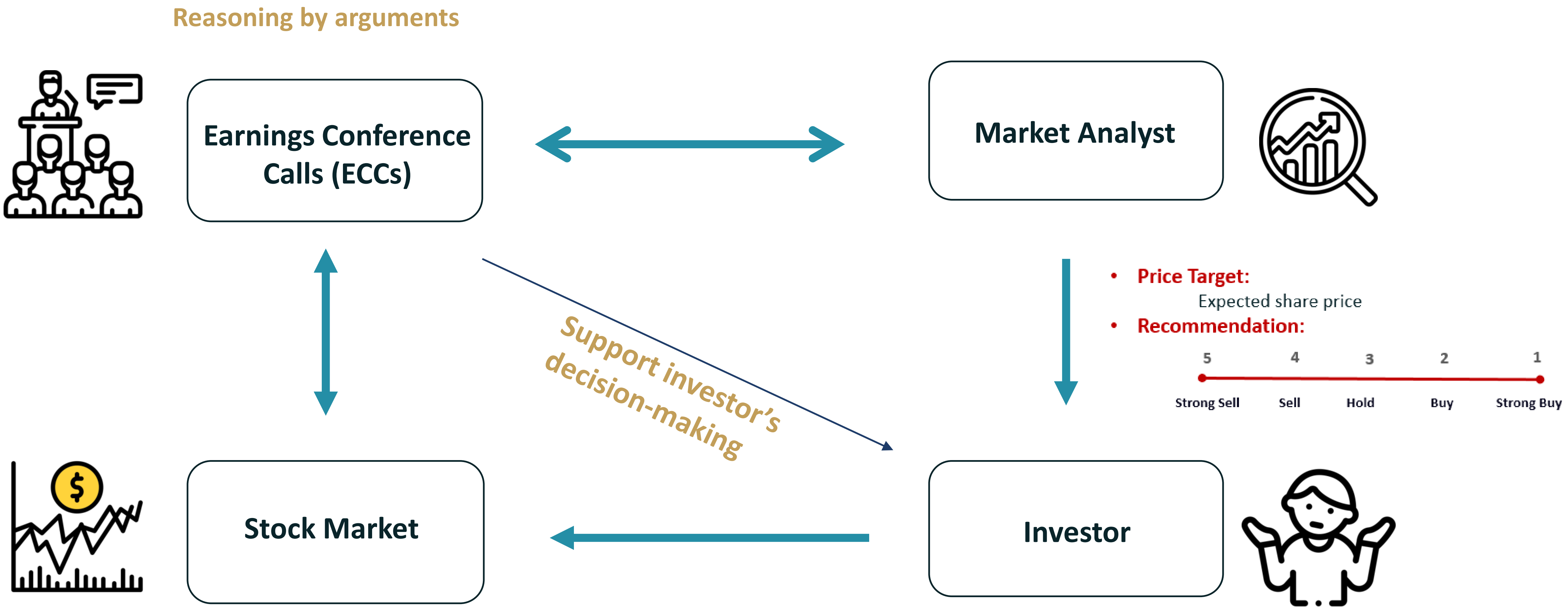
Exploring Large Language Models in Financial Argument Relation Identification

Yasser Otiefy, Alaa Alhamzeh

University Of Passau
yasser.otiefy@uni-passau.de, alaa.alhamzeh@uni-passau.de

Motivation


Reasoning by arguments
Earnings Conference Calls (ECCs)
Market Analyst
Price Target:
Expected share price
Recommendation:
5 4 3 2 1
Strong Sell Sell Hold Buy Strong Buy
Support investor's decision-making
Stock Market
Investor

Computational argumentation tasks

The global market **for power transmission and distribution infrastructure is expected to remain buoyant in 2023**

Demand is forecast **to be driven** in Europe and North America **by the need** for equipment replacement, improved grid reliability and efficiency and further grid interconnections.

Claim

Support

Premise

Is the argument persuasive?
Well-supported?
…

Argument Identification → Unit Classification → Structure Classification → Argument Assessment → Argument Generation

Argument Identification
- Argument
- Non-argument

Unit Classification
- Premise(s)
- Claim (conclusion)

Structure Classification
- Premise(s)
- Claim
- Relation (support/attack)

Argument Assessment
- Quality assessment
- Stance detection
- ….

Argument Generation
- Argument summarization
- Argument synthesis
- Claim synthesis
- ….

Argument Mining (AM)

Financial Argumentation Data

- Earnings conference calls for major tech companies
- Annotated on the sentence level to cover the argument structure, and argument quality:

Argument structure corpus - *FinArg*

Alaa Alhamzeh et al. **It’s Time to Reason: Annotating Argumentation Structures in Financial Earnings Calls: The FinArg Dataset**, Financial NLP workshop FinNLP@EMNLP 2022.
Download: Github - Alaa-Ah/The-FinArg-Dataset-Argument-Mining-in-Financial-Earnings-Calls.

Argument quality corpus - *FinArg Quality*

Alaa Alhamzeh **Argument Quality Assessment in Financial Earnings Conference Calls** – International Conference on Database and Expert Systems Applications DEXA 2023.
Download: GitHub - Alaa-Ah/The-FinArgQuality-dataset-Quality-of-managers-arguments-in-Eearnings-Conference-Calls.

Problem statement

Argument Relation Identification

- Claim [SEP] Premise
- Negative sampling
- 10K samples
- Binary classification task on balanced data
- Poorly studied task in the literature
- FinArg-1 shared task

The global market **for power transmission and distribution infrastructure is expected to remain buoyant in 2023**

Claim

Support

Premise

Demand is forecast **to be driven** in Europe and North America **by the need** for equipment replacement, improved grid reliability and efficiency and further grid interconnections.

FinArg-1 @ NTCIR-17

Experimental Setup

General-purpose models

- Vicuna
- Bloom
- Llama
- …

Financial-fine-tuned models

- FinBert
- Deberta-finetuned-finance-text-classification
- ….

Debate-fine-tuned models

- ArgumentMining- EN-ARI-Debate
- Roberta-argument
- ….

GPT -4 Zero shot learning

Experimental Setup

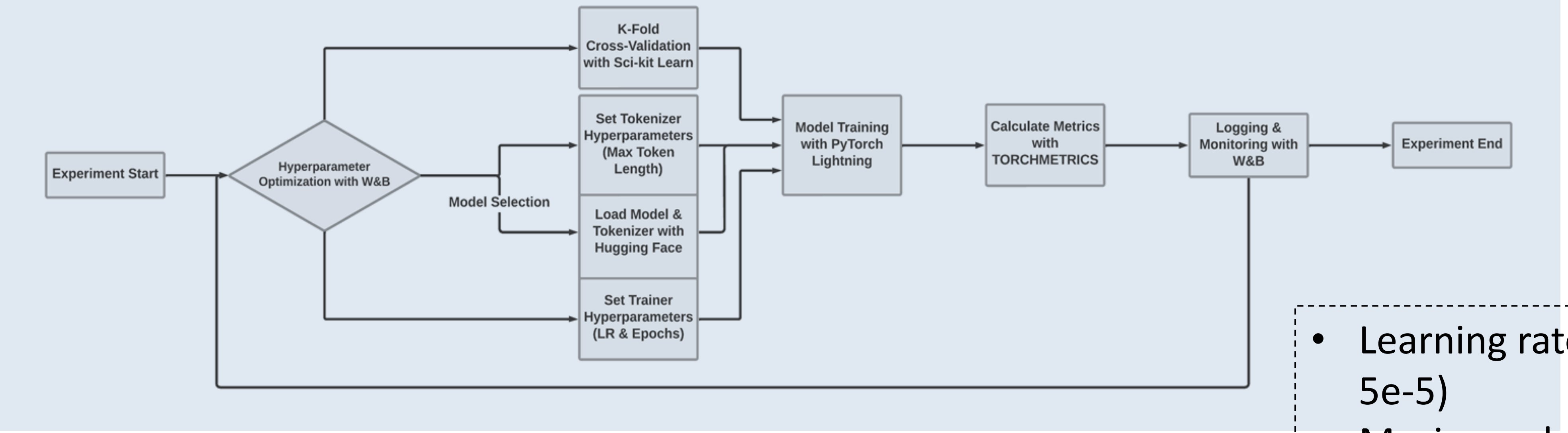


Figure: Workflow of open-source models' experiments

- Learning rate (2e-5, 3e-5, 5e-5)
- Maximum length of the tokenizer (64, 128, 256)
- Number of epochs (2 to 5)
- 5-fold cross validation
- Weight&Bias platform

Results - Open Source models

Model	Accuracy	F1-score	Precision	Recall	Model Type
Vicuna-13b_rm_oasst-hh	0.764 ± 0.05	**0.751 ± 0.05**	0.767 ± 0.05	0.764 ± 0.05	General-Purpose Models
Vicuna-13b-v1.5	0.762 ± 0.05	0.750 ± 0.05	0.762 ± 0.05	0.762 ± 0.05	
Bloom-7b1	0.675 ± 0.04	0.659 ± 0.06	0.677 ± 0.04	0.674 ± 0.04	
meta-llama/Meta-Llama-3-8B	0.642 ± 0.02	0.638 ± 0.02	0.643 ± 0.02	0.642 ± 0.02	
Bloom-1b1	0.567 ± 0.04	0.549 ± 0.05	0.572 ± 0.04	0.567 ± 0.04	
Bloomz-7b1	0.567 ± 0.02	0.534 ± 0.03	0.573 ± 0.02	0.567 ± 0.02	
Bloom-560m	0.531 ± 0.02	0.507 ± 0.03	0.530 ± 0.02	0.531 ± 0.02	
Bert-base-uncased	0.532 ± 0.01	0.503 ± 0.03	0.541 ± 0.02	0.532 ± 0.01	
GPT4-x-Alpaca	0.558 ± 0.04	0.536 ± 0.04	0.561 ± 0.04	0.558 ± 0.04	
LLaMa-2-7B-Guanaco-QLoRA-GPTQ	0.517 ± 0.01	0.468 ± 0.06	0.504 ± 0.09	0.517 ± 0.01	
Roberta-base	0.547 ± 0.03	0.479 ± 0.09	0.563 ± 0.13	0.547 ± 0.03	
ArgumentMining-EN-ARI-Debate	0.753 ± 0.02	**0.751 ± 0.02**	0.753 ± 0.01	0.753 ± 0.02	Debate-fine-tuned Models
ArgumentMining-EN-AC-Essay-Fin	0.622 ± 0.04	0.615 ± 0.04	0.627 ± 0.02	0.622 ± 0.02	
Roberta-base-150T-argumentative-sentence-detector	0.578 ± 0.01	0.569 ± 0.01	0.584 ± 0.02	0.578 ± 0.02	
ArgumentMining-EN-CN-ARI-Essay-Fin	0.532 ± 0.01	0.492 ± 0.07	0.540 ± 0.06	0.532 ± 0.01	
ArgumentMining-EN-AC-Financial	0.530 ± 0.02	0.480 ± 0.08	0.536 ± 0.09	0.530 ± 0.02	
FinancialBERT-Sentiment-Analysis	0.518 ± 0.02	**0.514 ± 0.02**	0.518 ± 0.02	0.518 ± 0.02	Financial-fine-tuned Models
Roberta-Earning-Call-Transcript-Classification	0.503 ± 0.01	0.371 ± 0.07	0.359 ± 0.14	0.503 ± 0.01	
Finbert	0.516 ± 0.02	0.507 ± 0.03	0.517 ± 0.02	0.516 ± 0.02	
Deberta-v3-base-finetuned-finance-text-classification	0.554 ± 0.01	0.505 ± 0.03	0.589 ± 0.02	0.554 ± 0.01	

Table: Argument relation identification using 5-fold cross-validation. All models are fine-tuned using Lr=5e-5, and 5 epochs, except Bloomz-7b1, for 2 epochs

Results - GPT-4

Our prompt:

“

You are a helpful assistant. Given the following claim and premise, please classify the relation between them as either Related or Unrelated. Please only generate one of the two labels:

Claim: ….

Premise: …..

”

Class	Precision	Recall	F1-score	Support
Related	0.85	0.75	0.79	4899
Unrelated	0.77	0.87	0.82	4899
Accuracy			0.81	9798
Macro Avg	0.81	0.81	0.81	9798
Weighted Avg	0.81	0.81	0.81	9798

Table: Performance of GPT4 zero shot learning

Discussions - Model category

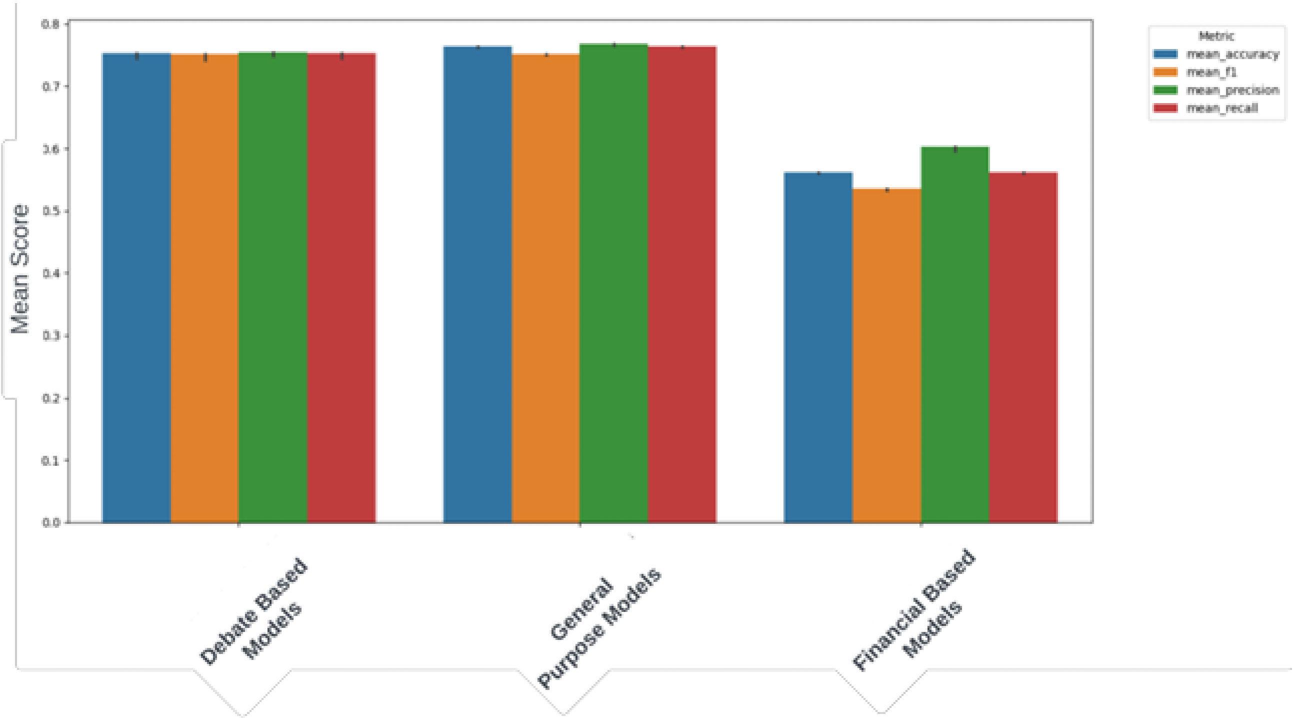


Figure: Mean performance by model category

Discussions – Impact of model size

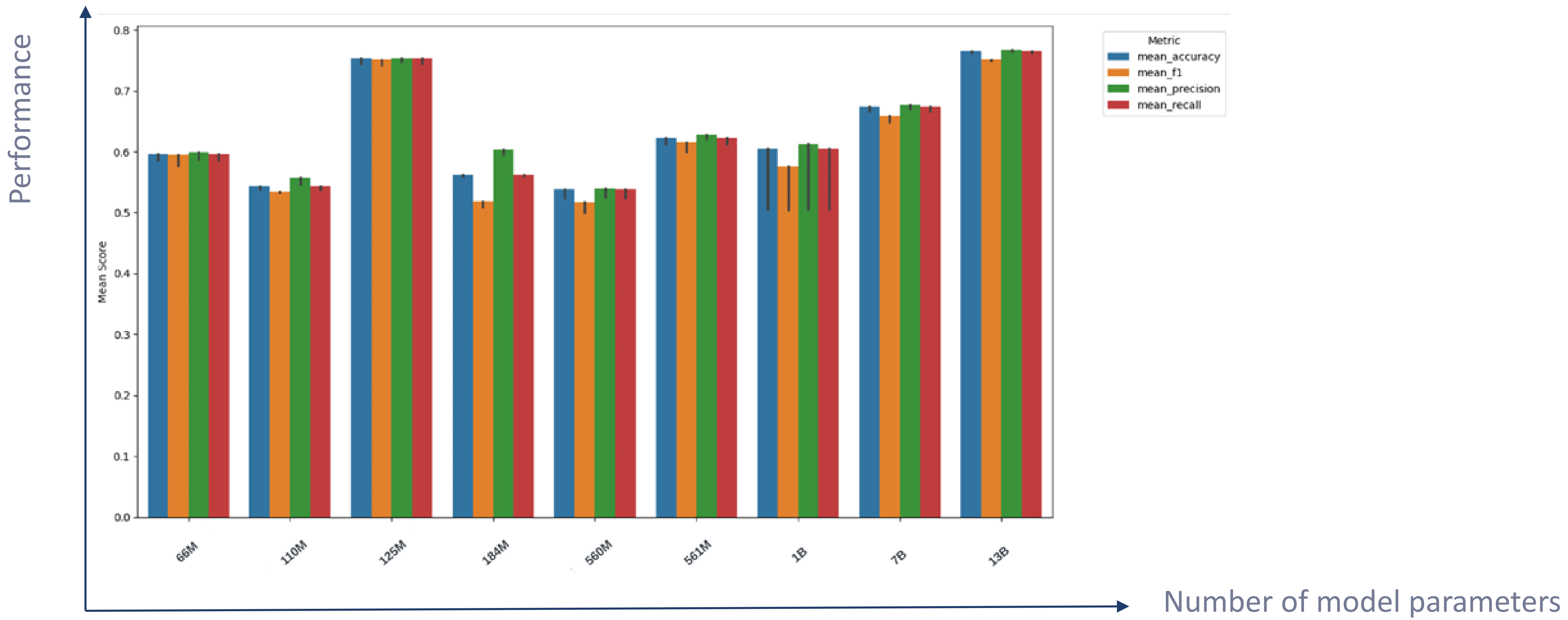


Figure: Mean performance by model size

Discussions - Hyperparameters

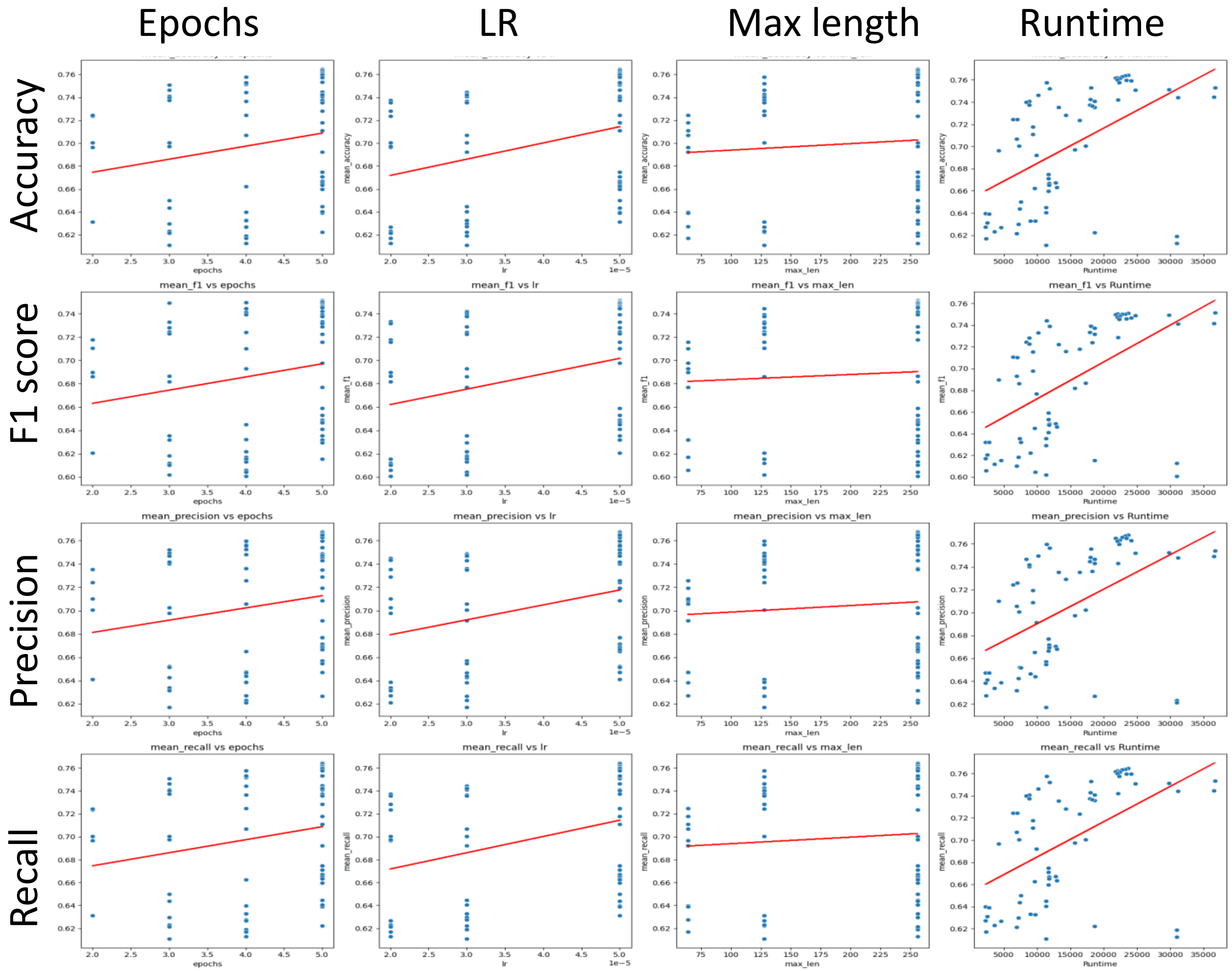


Figure: Model performance by hyperparameters settings and runtime

Conclusion & Future work

- GPT-4 achieved the highest F1-score (0.81) in zero-shot learning
- Significance of zero-shot learning for complex language tasks in finance
- Applications: include into a RAG framework, real-time analysis of financial text, and assist decision-making
- Interpretation tools like Google Patchscopes
- Model merging

Thank you for your attention