The Problem

We'd like you to please demonstrate a LLM driven prototype that can run through the ConvFinQA dataset and provide your answers to the queries for the subset.

Inside the train.json file are questions and answers, along supporting data and table of figures, here's a snippet of the json:

```
qa": {
      "question": "what was the percentage change in the net cash from operating activities from 2008 to 2009”,
       "answer": "14.1%",
}
```

Please produce a report on metrics with accuracies of your system, and short a write up of your findings and short comings.

The Approach:

Given the time constraints, my decision was to test two different OpenAI models (one from the completion api set, one from the chat api set), writing a software that can be easily modified for testing different models and different LLM (e.g. Anthropic, LLAMA, etc).

The question is embedded in a prompt
The response is then processed by another prompt to ensure that the answer will be just a number.

The analysis of the answers must consider differences in how the response are represented, from single differences in number formatting, to differences in rounding and truncations to evident errors and miscalculations

The accuracy of the system will be measured against different post-processing techniques to evaluate the correctness of the answers.

Challenges: How to evaluate if the answer is correct?

Comparing the answer from the system and the one from the dataset:

- The two answers are different
- The two answers are the same but with different representation
  - *42.46% vs 0.43 (Double_PNC/2014/page_99.pdf)*
- The two answers are different due rounding
  - *7.99 vs 8.1% (Single_DRE/2007/page_39.pdf-4)*
- The two answers are different due truncation
  - *58.4 vs 58% (Double_NKE/2015/page_37.pdf)*
- The two answers are different but they have string similarity
  - *7220 vs 7020 (Single_HII/2015/page_120.pdf-1)*
  - *595840 vs 594840 (Single_AMT/2002/page_104.pdf-2)*
- The two answers are the same but the sign
  - *-3919.0 vs 3919 (Double_AWK/2013/page_123.pdf)*
- The two answers are the same but the decimals are not correct
  - *3.17 vs 31.7% (Double_NCLH/2017/page_57.pdf)*
- *The answer is correct but the format is different*
  - *2017: $201 million, 2016: $189 million, 2015: $191 million vs 581 (Single_ECL/2017/page_96.pdf-4)*

Precision

Precision is a measure of the accuracy of the positive predictions made by a model. It tells us what proportion of positive identifications was actually correct. Precision is particularly useful when the cost of a false positive is high. For example, in email spam detection, a high precision rate means that few legitimate emails are incorrectly classified as spam.

Recall (Sensitivity)

Recall, also known as sensitivity, measures the ability of a model to find all the relevant cases (i.e. actual positives). It is the proportion of actual positives that were identified correctly. Recall is especially important when the cost of missing a positive instance (false negative) is high.

Choosing Between Precision and Recall

The importance of precision versus recall depends on the specific costs associated with false positives versus false negatives:

- High Precision: Important when false positives are more costly than false negatives.
- High Recall: Crucial when false negatives carry a heavier cost than false positives.

F1 Score

Since there is often a trade-off between precision and recall, the F1 Score is used as a harmonic mean of the two that balances both. The F1 Score is particularly useful when you need a single metric to gauge the performance of a model where both recall and precision are important:

Metrics: Accuracy/Precision, Recall, F1-Score

	Answer	Total	Count	%	TP	FP	FN	Precision	Recall	F1
5	A#range5	1311	825	62.929062	825	486	0	0.629291	1.000000	0.772472
4	A#round0	1311	807	61.556064	807	504	18	0.615561	0.978182	0.755618
6	A#range4	1311	794	60.564455	794	517	31	0.605645	0.962424	0.743446
7	A#range3	1311	743	56.674294	743	568	82	0.566743	0.900606	0.695693
8	A#range2	1311	703	53.623188	703	608	122	0.536232	0.852121	0.658240
9	A#range1	1311	650	49.580473	650	661	175	0.495805	0.787879	0.608614
3	A#round1	1311	578	44.088482	578	733	247	0.440885	0.700606	0.541199
2	A#round2	1311	333	25.400458	333	978	492	0.254005	0.403636	0.311798
1	A#Processed	1311	326	24.866514	326	985	499	0.248665	0.395152	0.305243
0	A#Exact	1311	69	5.263158	69	1242	756	0.052632	0.083636	0.064607

roundN - round N decimals
rangeN - abs(result) < (N/10)

Exact - answer after LLM Query
Processed - answer after processing LLM Query answer

Conclusion

Post processing the answer and applying a rounding improve drastically the Recall.

The right choice between rounding and ranging improves the F1 score
From 0.06 to 0.77

	Answer	Total	Count	%	TP	FP	FN	Precision	Recall	F1
5	A#range5	1311	825	62.929062	825	486	0	0.629291	1.000000	0.772472
4	A#round0	1311	807	61.556064	807	504	18	0.615561	0.978182	0.755618
6	A#range4	1311	794	60.564455	794	517	31	0.605645	0.962424	0.743446
7	A#range3	1311	743	56.674294	743	568	82	0.566743	0.900606	0.695693
8	A#range2	1311	703	53.623188	703	608	122	0.536232	0.852121	0.658240
9	A#range1	1311	650	49.580473	650	661	175	0.495805	0.787879	0.608614
3	A#round1	1311	578	44.088482	578	733	247	0.440885	0.700606	0.541199
2	A#round2	1311	333	25.400458	333	978	492	0.254005	0.403636	0.311798
1	A#Processed	1311	326	24.866514	326	985	499	0.248665	0.395152	0.305243
0	A#Exact	1311	69	5.263158	69	1242	756	0.052632	0.083636	0.064607

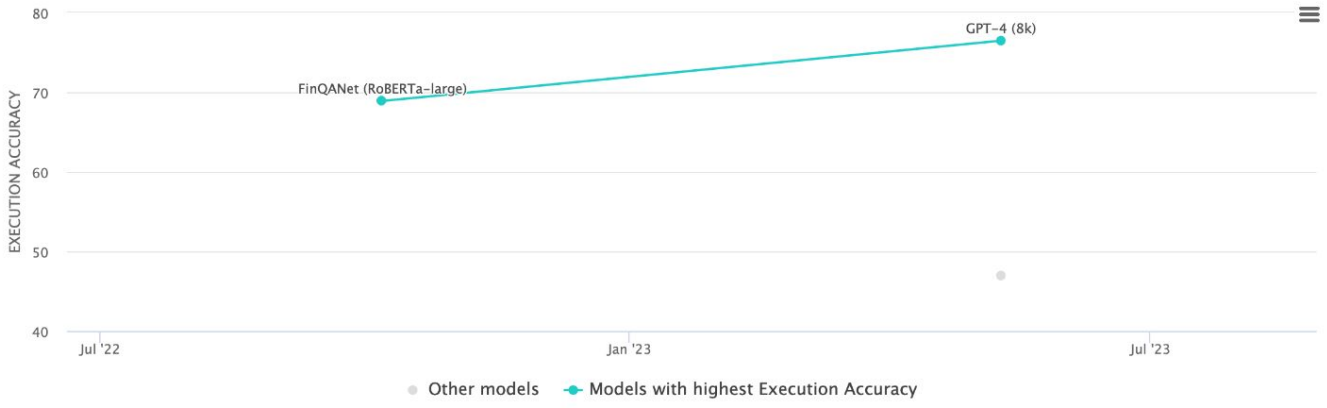
The main challenge is finding a right balance between prompt, rag techniques and post processing to evaluate the model precision and recall. What we want is the right balance between the ability of the model to give us the right answer even if the right answer is not exactly in the same format of the training set.

Appendix

Question Answering on ConvFinQA

Leaderboard Dataset

View Execution Accuracy by Date



Filter: untagged

Edit Leaderboard

Rank	Model	Execution Accuracy	Paper	Code	Result	Year	Tags
1	**GPT-4** (8k)	76.48	Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks			2023	
2	**FinQANet** (RoBERTa-large)	68.9	ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering			2022	
3	**General Crowd**	46.90	Are ChatGPT and GPT-4 General-Purpose Solvers for			2023	

Bibliograpy:

ConvFinQA: Exploring the Chain of Numerical Reasoning in Conversational Finance Question Answering

Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks