Ideas for how to benchmark:

1. We can follow the previous format used:
    a. We currently have a table that displays the topics that we will give the AI questions to answer. We would specify the type of questions and the number of questions per each question type.
        i. Here is how the table looks:

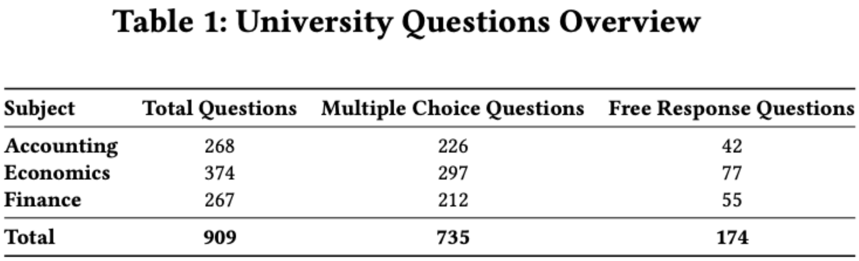

Table 1: University Questions Overview

Subject	Total Questions	Multiple Choice Questions	Free Response Questions
Accounting	268	226	42
Economics	374	297	77
Finance	267	212	55
Total	**909**	**735**	**174**

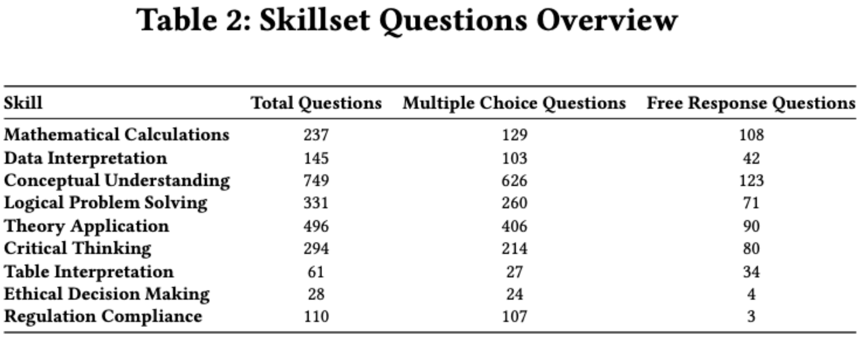

Table 2: Skillset Questions Overview

Skill	Total Questions	Multiple Choice Questions	Free Response Questions
Mathematical Calculations	237	129	108
Data Interpretation	145	103	42
Conceptual Understanding	749	626	123
Logical Problem Solving	331	260	71
Theory Application	496	406	90
Critical Thinking	294	214	80
Table Interpretation	61	27	34
Ethical Decision Making	28	24	4
Regulation Compliance	110	107	3

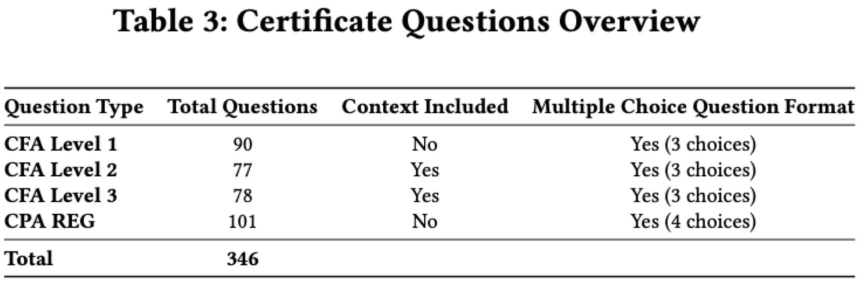

Table 3: Certificate Questions Overview

Question Type	Total Questions	Context Included	Multiple Choice Question Format
CFA Level 1	90	No	Yes (3 choices)
CFA Level 2	77	Yes	Yes (3 choices)
CFA Level 3	78	Yes	Yes (3 choices)
CPA REG	101	No	Yes (4 choices)
Total	**346**		

1. We can then have a program that will work with the FinGPT AI to automate the question-asking process. From here, with each response the AI gives, we will check with our resources (preferably we have a dataset of answers to our questions) and check for 3 things:
    a. The accuracy of the response.
        i. This will check how many right and wrong answers and get the average accuracy.
    b. The FActScore of the response.
        i. This will check how factual each response is and put it into a percentage.
    c. The weighted average of the above.
        i. We add the two scores and then divide it by 2 to get this.
2. Here is a benchmark we have for other AI's:

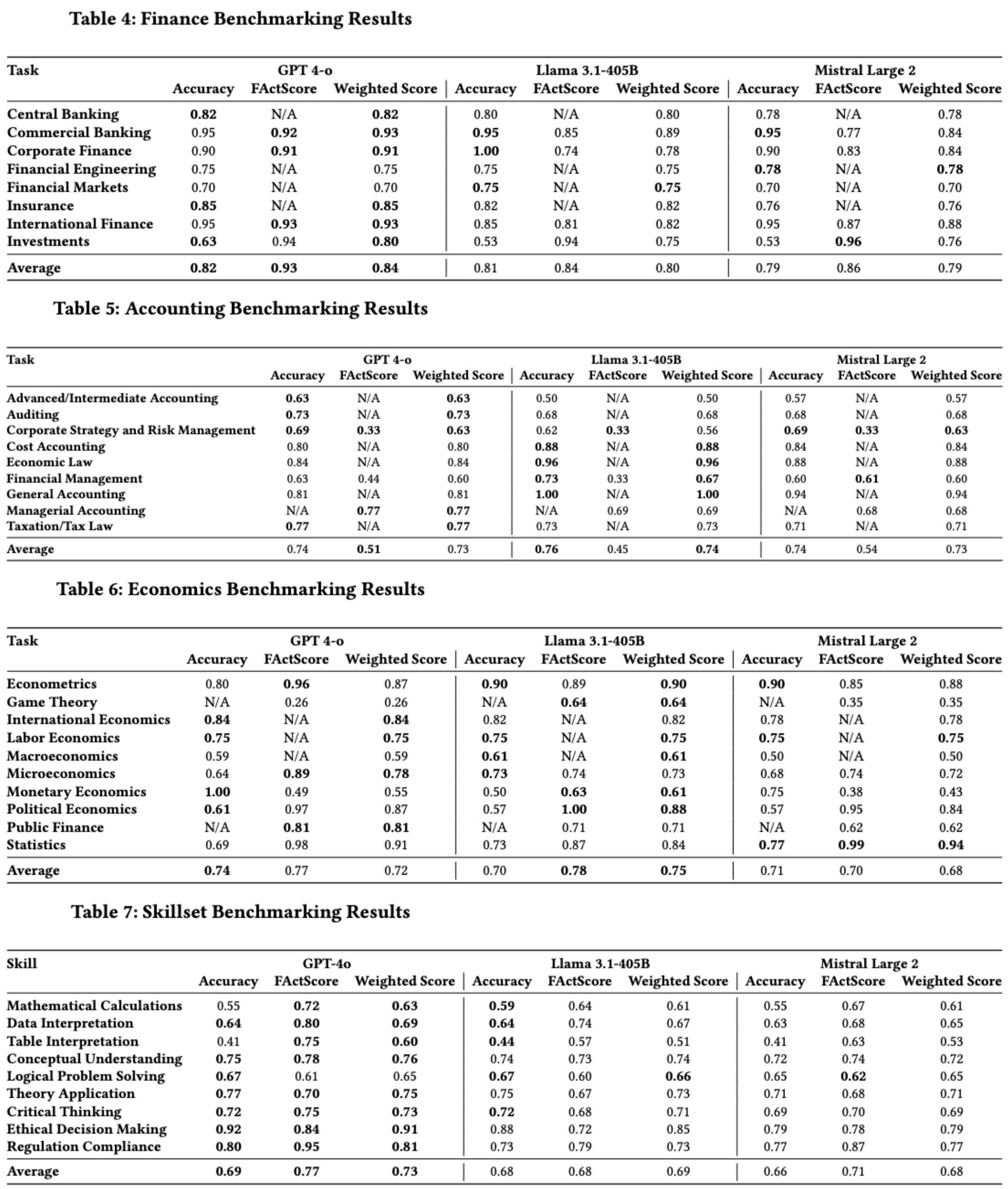

Table 4: Finance Benchmarking Results

Task	GPT 4-o			Llama 3.1-405B			Mistral Large 2		
	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score
Central Banking	**0.82**	N/A	**0.82**	0.80	N/A	0.80	0.78	N/A	0.78
Commercial Banking	0.95	**0.92**	**0.93**	**0.95**	0.85	0.89	**0.95**	0.77	0.84
Corporate Finance	0.90	**0.91**	**0.91**	**1.00**	0.74	0.78	0.90	0.83	0.84
Financial Engineering	0.75	N/A	0.75	0.75	N/A	0.75	**0.78**	N/A	**0.78**
Financial Markets	0.70	N/A	0.70	**0.75**	N/A	**0.75**	0.70	N/A	0.70
Insurance	**0.85**	N/A	**0.85**	0.82	N/A	0.82	0.76	N/A	0.76
International Finance	0.95	**0.93**	**0.93**	0.85	0.81	0.82	0.95	0.87	0.88
Investments	**0.63**	0.94	**0.80**	0.53	0.94	0.75	0.53	**0.96**	0.76
Average	**0.82**	**0.93**	**0.84**	0.81	0.84	0.80	0.79	0.86	0.79

Table 5: Accounting Benchmarking Results

Task	GPT 4-o			Llama 3.1-405B			Mistral Large 2		
	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score
Advanced/Intermediate Accounting	**0.63**	N/A	**0.63**	0.50	N/A	0.50	0.57	N/A	0.57
Auditing	**0.73**	N/A	**0.73**	0.68	N/A	0.68	0.68	N/A	0.68
Corporate Strategy and Risk Management	**0.69**	**0.33**	**0.63**	0.62	**0.33**	0.56	**0.69**	**0.33**	**0.63**
Cost Accounting	0.80	N/A	0.80	**0.88**	N/A	**0.88**	0.84	N/A	0.84
Economic Law	0.84	N/A	0.84	**0.96**	N/A	**0.96**	0.88	N/A	0.88
Financial Management	0.63	0.44	0.60	**0.73**	0.33	**0.67**	0.60	**0.61**	0.60
General Accounting	0.81	N/A	0.81	**1.00**	N/A	**1.00**	0.94	N/A	0.94
Managerial Accounting	N/A	**0.77**	**0.77**	N/A	0.69	0.69	N/A	0.68	0.68
Taxation/Tax Law	**0.77**	N/A	**0.77**	0.73	N/A	0.73	0.71	N/A	0.71
Average	0.74	**0.51**	0.73	**0.76**	0.45	**0.74**	0.74	0.54	0.73

Table 6: Economics Benchmarking Results

Task	GPT 4-o			Llama 3.1-405B			Mistral Large 2		
	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score
Econometrics	0.80	**0.96**	0.87	**0.90**	0.89	**0.90**	**0.90**	0.85	0.88
Game Theory	N/A	0.26	0.26	N/A	**0.64**	**0.64**	N/A	0.35	0.35
International Economics	**0.84**	N/A	**0.84**	0.82	N/A	0.82	0.78	N/A	0.78
Labor Economics	**0.75**	N/A	**0.75**	**0.75**	N/A	**0.75**	**0.75**	N/A	**0.75**
Macroeconomics	0.59	N/A	0.59	**0.61**	N/A	**0.61**	0.50	N/A	0.50
Microeconomics	0.64	**0.89**	**0.78**	**0.73**	0.74	0.73	0.68	0.74	0.72
Monetary Economics	**1.00**	0.49	0.55	0.50	**0.63**	**0.61**	0.75	0.38	0.43
Political Economics	**0.61**	0.97	0.87	0.57	**1.00**	**0.88**	0.57	0.95	0.84
Public Finance	N/A	**0.81**	**0.81**	N/A	0.71	0.71	N/A	0.62	0.62
Statistics	0.69	0.98	0.91	0.73	0.87	0.84	**0.77**	**0.99**	**0.94**
Average	**0.74**	0.77	0.72	0.70	**0.78**	**0.75**	0.71	0.70	0.68

Table 7: Skillset Benchmarking Results

Skill	GPT-4o			Llama 3.1-405B			Mistral Large 2		
	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score	Accuracy	FActScore	Weighted Score
Mathematical Calculations	0.55	**0.72**	**0.63**	**0.59**	0.64	0.61	0.55	0.67	0.61
Data Interpretation	**0.64**	**0.80**	**0.69**	**0.64**	0.74	0.67	0.63	0.68	0.65
Table Interpretation	0.41	**0.75**	**0.60**	**0.44**	0.57	0.51	0.41	0.63	0.53
Conceptual Understanding	**0.75**	**0.78**	**0.76**	0.74	0.73	0.74	0.72	0.74	0.72
Logical Problem Solving	**0.67**	0.61	0.65	**0.67**	0.60	**0.66**	0.65	**0.62**	0.65
Theory Application	**0.77**	**0.70**	**0.75**	0.75	0.67	0.73	0.71	0.68	0.71
Critical Thinking	**0.72**	**0.75**	**0.73**	**0.72**	0.68	0.71	0.69	0.70	0.69
Ethical Decision Making	**0.92**	**0.84**	**0.91**	0.88	0.72	0.85	0.79	0.78	0.79
Regulation Compliance	**0.80**	**0.95**	**0.81**	0.73	0.79	0.73	0.77	0.87	0.77
Average	**0.69**	**0.77**	**0.73**	0.68	0.68	0.69	0.66	0.71	0.68

a.

3. We can also have the program identify where the FinGPT AI produces hallucinations and misinformation. Once it identifies this, we will want to store this information along with the question that produced it. Doing this should allow us to find patterns on how to reduce the amount of hallucinations and misinformation that are being produced by the AI. Currently, we have a method of reducing this by giving the AI links to websites that have the correct information which it can pull into its responses. However, we can also try to improve this process so that it never happens which is what we are hoping to achieve at the end of this project.