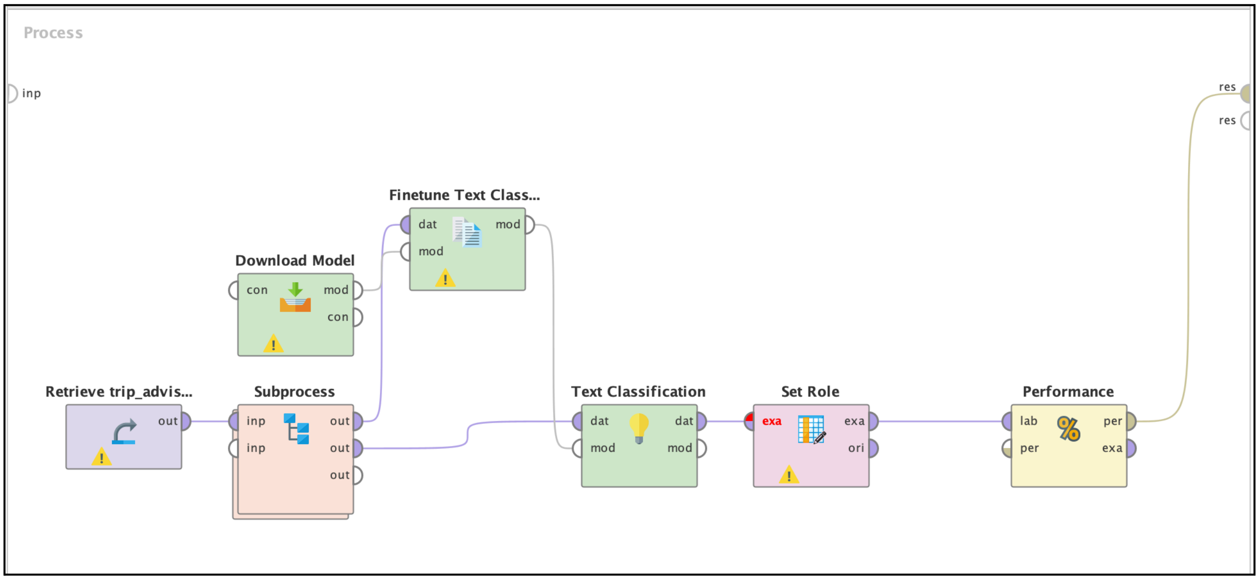
Assignment 3

Aneesh Soni

Part 1

Model Design



Performance Matrix: Accuracy of 69.25%

accuracy: 69.25%

	true 4	true 5	true 2	true 3	true 1	class precision
pred. 4	69	17	0	20	1	64.49%
pred. 5	42	165	1	1	2	78.20%
pred. 2	1	0	7	4	4	43.75%
pred. 3	9	0	10	25	3	53.19%
pred. 1	0	0	6	2	11	57.89%
class recall	57.02%	90.66%	29.17%	48.08%	52.38%	

Execution Time: 15 minutes 49 seconds

homework_3_part_1 (1 results. Process results)
Completed: Nov 6, 2024 9:39:24 PM (execution time: 15:49)

Performance Vector (Performance)
Result not stored in repository.

```
PerformanceVector:
accuracy: 69.25%
ConfusionMatrix:
True:   4       5       2       3       1
4:      69      17      0       20      1
5:      42      165     1       1       2
2:      1       0       7       4       4
3:      9       0       10      25      3
1:      0       0       6       2       11
```

Explanation:

Please comment on the differences between the results produced by the fine-tuned model you just built in this assignment and the pre-trained model produced in Part 4 (the "Generative Models" part) of the Lab. Also, how good are the accuracy performance results of these models, as compared with the results produced in Part 3 of Assignment 2 (i.e. are they better or worse than those other results)? Please, justify your answer.

The model in Part 4 of the lab solely relied on a pre-trained model (aka Generative Model) to predict the ratings whereas for Part 1 of this assignment we fine tuned the "nlptown/bert-base-multilingual-uncased-sentiment" model. Part 4 of the lab just output a column with the rating such as "1 star", so I went ahead and truncated the results to just their integer form and then calculated the accuracy (total correct predictions / # of total predictions) and this came out 60% for Part 4 of lab 1. So fine tuning did improve the model, increasing the accuracy to 69.25%.

When compared to the models in Part 3 of Assignment 2 the accuracy performance of this model produced far better results overall. In fact, this model is better at predicting ratings throughout the rating spectrum. We can observe that the class precision for each prediction rating is substantially higher (column highlighted in green) than every model in Part 3 of Assignment 2.

Class Precision for Assignment 3 vs Assignment 2

	Assignment 3 Part 1	Assignment 2 Part 3a	Assignment 2 Part 3b	Assignment 2 Part 3c
Pred 1	57.89%	10%	0%	39.39%
Pred 2	43.75%	0%	0%	0%
Pred 3	53.19%	20%	18.18%	27.27%
Pred 4	64.49%	31.31%	23.68%	18.75%
Pred 5	78.20%	51.75%	58.15%	53.80%

Interestingly enough if we take a look at the class recall (table on next page), we see that the fine tuning approach for Part 1 in this homework assignment (highlighted in green) is better for classes 2, 3 and 4 however for class 1 and class 5 assignment 2 part 3c has better recall meaning that it maximizes the number of positives correctly classified and minimizes false negatives. While all models do quite well handling class 5, the marked improvement across other classes while maintaining fairly high recall for class 5 in the model for assignment 3 showcases its relative strength compared to other modeling techniques. It might be worth exploring modeling techniques that leverage aspects of some of the strong performers in assignment 2 and merging them with assignment 3.

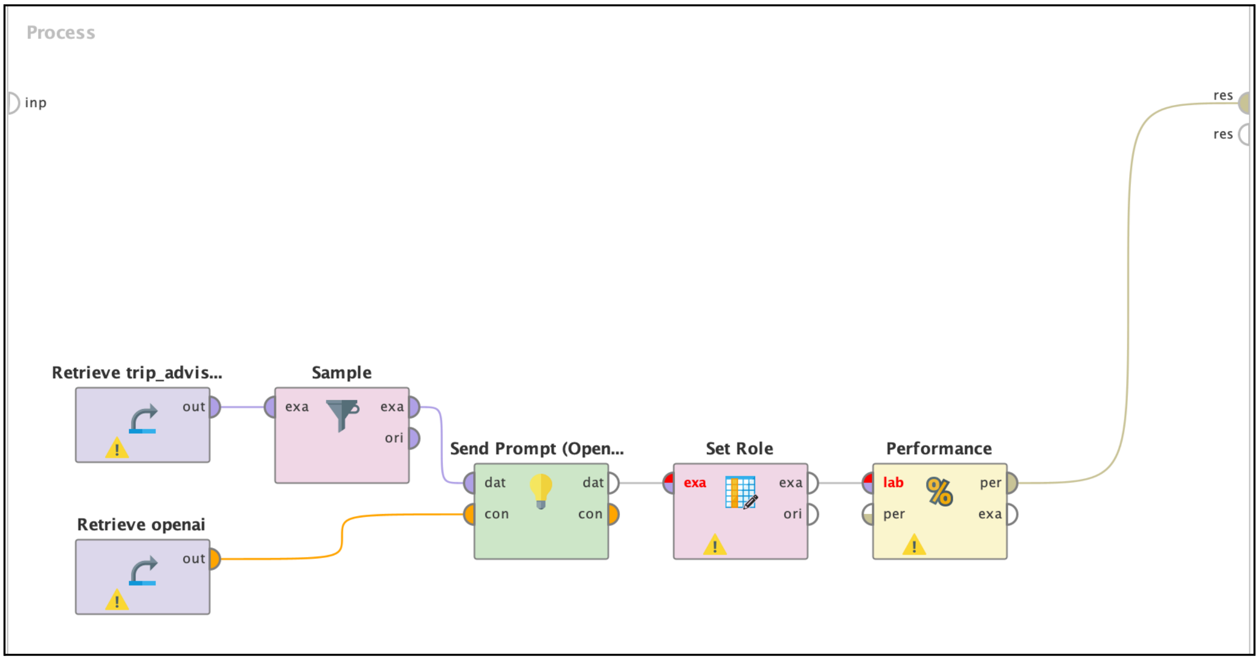
Class Recall for Assignment 3 vs Assignment 2

	Assignment 3 Part 1	Assignment 2 Part 3a	Assignment 2 Part 3b	Assignment 2 Part 3c
True 1	52.38%	9.52%	0%	61.90%
True 2	29.17%	0%	0%	0%
True 3	48.08%	7.69%	3.85%	11.54%
True 4	57.02%	25.62%	14.88%	2.48%
True 5	90.66%	73.08%	100%	97.25%

(continued on next page)

Part 2 Zero-Shot

Model Design



Performance Matrix: Accuracy of 67%

Table View   Plot View

Changes to a table showing the confusion matrix.

accuracy: 67.00%

	true 2	true 5	true 4	true 1	true 3	class precision
pred. 2	2	0	0	0	6	25.00%
pred. 5	0	47	17	0	0	73.44%
pred. 4	0	4	11	0	1	68.75%
pred. 1	1	0	0	3	0	75.00%
pred. 3	0	0	4	0	4	50.00%
class recall	66.67%	92.16%	34.38%	100.00%	36.36%	

Execution Time: 8 seconds

homework_3_part_2 (1 results. Process results)
Completed: Nov 3, 2024 4:32:41 PM (execution time: 8 s)

Performance Vector (Performance)
Result not stored in repository.

```
PerformanceVector:
accuracy: 67.00%
ConfusionMatrix:
True:   2       5       4       1       3
2:      2       0       0       0       6
5:      0       47      17      0       0
4:      0       4       11      0       1
1:      1       0       0       3       0
3:      0       0       4       0       4
```

Explanation:

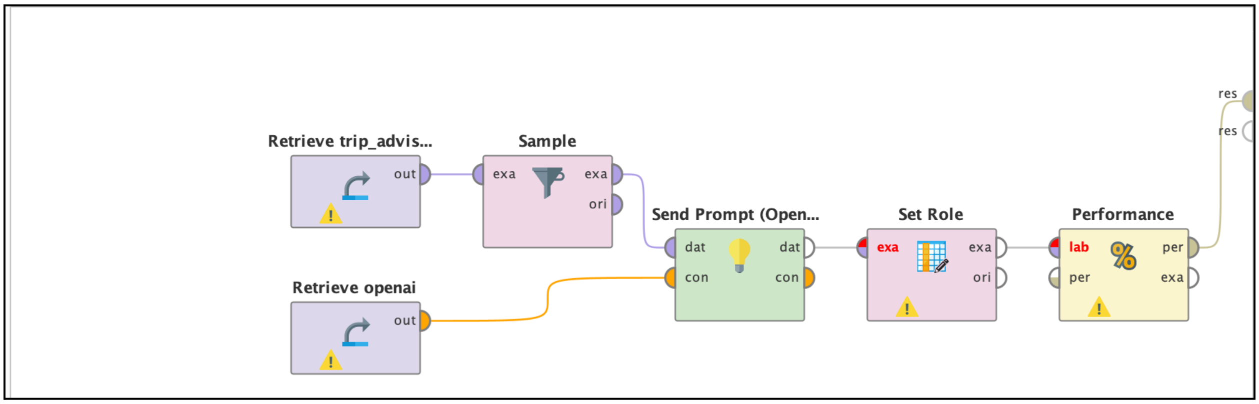
Also, briefly comment on the performance results, i.e. how is the performance compared to the finetuning models in Part 1. What are the potential reasons for such differences?

The zero-shot approach using OpenAI's GPT-3.5 Turbo (67% accuracy) is quite similar accuracy wise to the finetuned model (69.25% accuracy) in Part 1. It's highly likely that OpenAI's models have already ingested a large corpus of reviews from a number of internet platforms (i.e. yelp, google reviews, etc.) and as a result has a greater knowledge base. This indicates that these proprietary closed sourced solutions such as the one from OpenAI are quite performant out of the box and might not require a lot of fine-tuning to retrieve similar results as the nlp bert models. Nonetheless, this does show that open source models can be quite performant if you put in the effort to create finetuned solutions. Furthermore, to get a better understanding of how these models might go head to head it would be interesting to compare results without finetuning the bert model as well as a finetuned version of the zero-shot approach.

(continued on next page)

Part 2 Few-Shot

Model Design



Performance Matrix: Accuracy 66%

Table View Plot View

accuracy: 66.00%

	true 2	true 5	true 4	true 1	true 3	class precision
pred. 2	1	0	0	0	4	20.00%
pred. 5	0	43	15	0	0	74.14%
pred. 4	0	8	12	0	0	60.00%
pred. 1	2	0	0	3	0	60.00%
pred. 3	0	0	5	0	7	58.33%
class recall	33.33%	84.31%	37.50%	100.00%	63.64%	

Execution Time: 7 seconds

homework_3_part_2_few_shot (1 results. Process results)
Completed: Nov 3, 2024 5:22:17 PM (execution time: 7 s)

Performance Vector (Performance)
Result not stored in repository.

```
PerformanceVector:
accuracy: 66.00%
ConfusionMatrix:
True:   2       5       4       1       3
2:      1       0       0       0       4
5:      0       43      15      0       0
4:      0       8       12      0       0
1:      2       0       0       3       0
3:      0       0       5       0       7
```

Explanation:

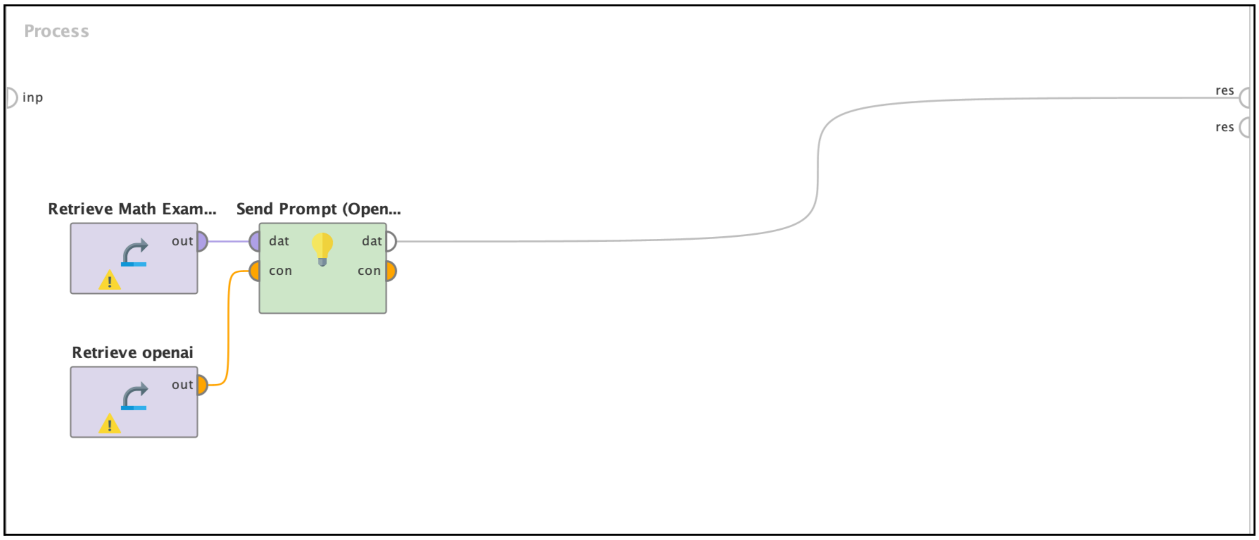
<u>Also, briefly comment on the performance results. How would these performance results compare with those achieved by the Zero-shot model from Step 2 of Part 2? How would you explain the performance differences?</u>

Surprisingly, the zero-shot approach slightly outperformed the few-shot model (67% vs. 66% accuracy), which was interesting given that few-shot prompts typically improve performance by adding relevant examples. The difference might be due to constraints imposed by the few-shot examples, which can lead to overfitting to those specific review styles rather than generalizing well to new inputs. In contrast, the zero-shot model interpreted each review in isolation, potentially leading to a more flexible and slightly better generalization. For example, the zero-shot model achieved higher precision for class 1 (75.00%) and better recall for class 2 (66.67%), indicating it was somewhat better at recognizing negative reviews than the few-shot model. The zero-shot model's isolated interpretation of each review may have contributed to better generalization, particularly for lower-rated reviews, where the few-shot model struggled with precision and recall in classes 2 and 3. Nonetheless, the data volumes are quite low for this test and it's worth running a larger set to see the results we can achieve.

(continued on next page)

Part 3 (a)

Model Design



Model Results:

Open in Turbo Prep | Auto Model | Interactive Analysis

Filter (5 / 5 examples): all

Row No.	Question	Answer	CoT	openai
1	Randy has 6...	85	Half of the number of Ran...	90 trees. Randy has 60 mango trees and half of that is 30. 5 less than 30 is 25, so Randy has 25 coconut trees. ...
2	Joy can read...	5	In one hour, there are 3 s...	6 hours
3	Mr. Sam sha...	5250	Tony got twice $1750 whi...	$5250
4	Tim rides hi...	16	He bikes 20 * 2 = 40 mil...	Tim spends a total of 32 hours biking a week.
5	Bella bought...	38	The number of truck stam...	33 stamps

Explanation:

Run this model and report how many questions ChatGPT answered correctly. Also, briefly comment on the performance results [note that the answer for each question is shown in the dataset in the "Answer" column].

When running the model the first time I got the results shown in the image **above**. We can see that ⅕ of the questions were answered correctly, specifically question 3 was right as OpenAI successfully arrived at 5250. Additionally, it was interesting to see that OpenAI provided rationale/logic for certain questions and for other questions it just provided the answer. As a result this led me to rerunning the model several times without adjusting the settings and ensuring the temperature was still left at 0. After doing so I got the results shown in the image **below**. Everytime, I re-ran the model I noticed that we got question 3 right (i.e. successfully arriving on 5250) and we always arrived at 33 stamps for question 5 (i.e. incorrect). Furthermore, I noticed that the model would sometimes get the answer correctly but the associated explanation arrived at a different answer. GPT3.5 Turbo doesn't have the ability to "think" at quite the same level as some of the newer models with reasoning capabilities such as o1-preview and o1-mini.

Rerunning the Model Multiple Times:

homework_3_part_3_part_a (1 results. Process results)
Completed: Nov 4, 2024 1:29:20 AM (execution time: 7 s)

ExampleSet (Send Prompt (OpenAI))
Result not stored in repository.

4 attributes:

Role	Name	Type	Range	Missings	Comment
–	Question	polynominal	= {Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck , Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?, Mr. Sam shared a certain amount of money between his two sons, Ken and Tony. If Ken got $1750, and T, Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango tr, Tim rides his bike back and forth to work for each of his 5 workdays. His work is 20 miles away. He }	no missing values	–
–	Answer	integer	= [?...?]; mean =?	no missing values	–
–	CoT	polynominal	= {Half of the number of Randy's mango trees is 60/2 = 30 trees. So Randy has 30 - 5 = 25 coconut trees, He bikes 20 * 2 = 40 miles each day for work So he bikes 40 * 5 = 200 miles for work That means he b, In one hour, there are 3 sets of 20 minutes. So, Joy can read 8 x 3 = 24 pages in an hour. It will t, The number of truck stamps is 11 + 9 = 20. The number of rose stamps is 20 – 13 = 7. Bella bought 11, Tony got twice $1750 which is 2*$1750 = $3500 The total amount shared was $1750+$3500 = $5250. The a}	no missing values	–
–	openai	polynominal	= {$5250, 33 stamps, 6 hours, 90 trees., Tim spends a total of 32 hours biking a week.}	no missing values	–

homework_3_part_3_part_a (1 results. Process results)
Completed: Nov 4, 2024 1:32:33 AM (execution time: 2 s)

ExampleSet (Send Prompt (OpenAI))
Result not stored in repository.

4 attributes:

Role	Name	Type	Range	Missings	Comment
–	Question	polynominal	= {Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck , Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?, Mr. Sam shared a certain amount of money between his two sons, Ken and Tony. If Ken got $1750, and T, Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango tr, Tim rides his bike back and forth to work for each of his 5 workdays. His work is 20 miles away. He }	no missing values	–
–	Answer	integer	= [?...?]; mean =?	no missing values	–
–	CoT	polynominal	= {Half of the number of Randy's mango trees is 60/2 = 30 trees. So Randy has 30 - 5 = 25 coconut trees, He bikes 20 * 2 = 40 miles each day for work So he bikes 40 * 5 = 200 miles for work That means he b, In one hour, there are 3 sets of 20 minutes. So, Joy can read 8 x 3 = 24 pages in an hour. It will t, The number of truck stamps is 11 + 9 = 20. The number of rose stamps is 20 – 13 = 7. Bella bought 11, Tony got twice $1750 which is 2*$1750 = $3500 The total amount shared was $1750+$3500 = $5250. The a}	no missing values	–
–	openai	polynominal	= {$5250, 12 hours, 33 stamps, 6 hours, 85 trees. Half of 60 mango trees is 30, and 5 less than that is 25 coconut trees. So, Randy has 60}	no missing values	–

homework_3_part_3_part_a (1 results. Process results)
Completed: Nov 4, 2024 1:32:44 AM (execution time: 1 s)

ExampleSet (Send Prompt (OpenAI))
Result not stored in repository.

4 attributes:

Role	Name	Type	Range	Missings	Comment
–	Question	polynominal	= {Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck , Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?, Mr. Sam shared a certain amount of money between his two sons, Ken and Tony. If Ken got $1750, and T, Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango tr, Tim rides his bike back and forth to work for each of his 5 workdays. His work is 20 miles away. He }	no missing values	–
–	Answer	integer	= [?...?]; mean =?	no missing values	–
–	CoT	polynominal	= {Half of the number of Randy's mango trees is 60/2 = 30 trees. So Randy has 30 - 5 = 25 coconut trees, He bikes 20 * 2 = 40 miles each day for work So he bikes 40 * 5 = 200 miles for work That means he b, In one hour, there are 3 sets of 20 minutes. So, Joy can read 8 x 3 = 24 pages in an hour. It will t, The number of truck stamps is 11 + 9 = 20. The number of rose stamps is 20 – 13 = 7. Bella bought 11, Tony got twice $1750 which is 2*$1750 = $3500 The total amount shared was $1750+$3500 = $5250. The a}	no missing values	–
–	openai	polynominal	= {$5250, 12 hours, 33 stamps, 6 hours, 90 trees.}	no missing values	–

Execution Time: 3 seconds

homework_3_part_3_part_a (1 results. Process results)
Completed: Nov 3, 2024 6:11:15 PM (execution time: 3 s)

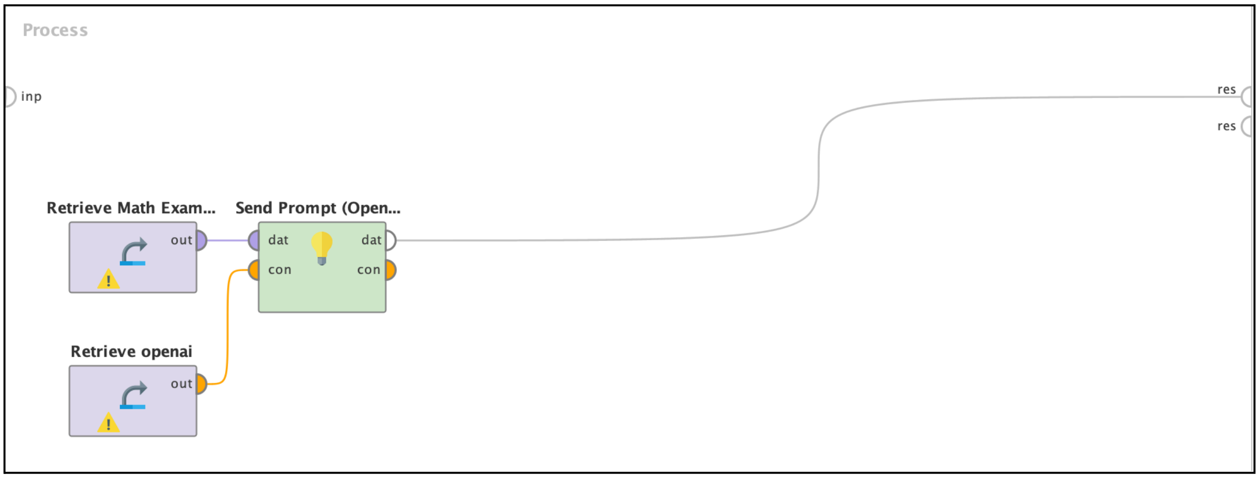
ExampleSet (Send Prompt (OpenAI))
Result not stored in repository.

Data table
Number of examples = 5
4 attributes:

Role	Name	Type	Range	Missings	Comment
–	Question	polynominal	= {Bella bought stamps at the post office. Some of the stamps had a snowflake design, some had a truck , Joy can read 8 pages of a book in 20 minutes. How many hours will it take her to read 120 pages?, Mr. Sam shared a certain amount of money between his two sons, Ken and Tony. If Ken got $1750, and T, Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango tr, Tim rides his bike back and forth to work for each of his 5 workdays. His work is 20 miles away. He }	no missing values	–
–	Answer	integer	= [?...?]; mean =?	no missing values	–
–	CoT	polynominal	= {Half of the number of Randy's mango trees is 60/2 = 30 trees. So Randy has 30 - 5 = 25 coconut trees, He bikes 20 * 2 = 40 miles each day for work So he bikes 40 * 5 = 200 miles for work That means he b, In one hour, there are 3 sets of 20 minutes. So, Joy can read 8 x 3 = 24 pages in an hour. It will t, The number of truck stamps is 11 + 9 = 20. The number of rose stamps is 20 – 13 = 7. Bella bought 11, Tony got twice $1750 which is 2*$1750 = $3500 The total amount shared was $1750+$3500 = $5250. The a}	no missing values	–

Part 3 (b)

Model Design



Model Results Using Chain-of-Thought Processing:

Row No.	OpenAI Response	Correct?
1	Randy has 60 mango trees. Half of 60 is 30, and 5 less than that is 25 coconut trees. So, Randy has 60 mango trees + 25 coconut trees = 85 trees in total on his farm.	YES
2	If Joy can read 8 pages in 20 minutes, she can read 1 page in 20/8 = 2.5 minutes. To read 120 pages, she will need 120 * 2.5 = 300 minutes. There are 60 minutes in an hour, so it will take her 300/60 = 5 hours to read 120 pages. The answer is 5 hours.	YES
3	If Ken got $1750, then Tony got $1750 * 2 = $3500. Therefore, the total amount of money shared was $1750 + $3500 = $5250.	YES
4	Tim rides 20 miles to work and 20 miles back, so he bikes 40 miles each workday. He works 5 days a week, so he bikes 40 miles x 5 days = 200 miles for work. Adding the 200 miles for the weekend bike ride, Tim bikes a total of 200 miles + 200 miles = 400 miles in a week. At a speed of 25 mph, Tim spends 400 miles / 25 mph = 16 hours biking each week. Therefore, Tim spends 16 hours biking a week.	YES
5	Bella bought 11 snowflake stamps. She bought 9 more truck stamps than snowflake stamps, so she bought 11 + 9 = 20 truck stamps. She bought 13 fewer rose stamps than truck stamps, so she bought 20 - 13 = 7 rose stamps. In total, Bella bought 11 + 20 + 7 = 38 stamps.	YES

Explanation:

Also, briefly comment on the performance results, i.e. how is the performance compared with the results from Part 3(a) (where the CoT prompt engineering technique was not used)? How would you explain the differences?

This time around, when leveraging chain-of-thought prompt engineering, we got all 5 questions correct and is a substantial improvement versus Part 3(a) where the model only answered 1 out of 5 questions correctly. This improvement can be attributed to the chain-of-thought prompt guiding the model to break down each question step-by-step, allowing it to approach the problems logically and systematically. Without CoT, GPT-3.5 Turbo sometimes skipped steps or arrived at the correct answer without a clear logical path, leading to errors in multi-step problems. The CoT prompt enabled the model to process the questions more thoughtfully, making it easier to handle complex reasoning and calculations accurately. This difference demonstrates that CoT prompt engineering is especially effective for tasks that require multiple logical steps, helping GPT-3.5 Turbo better apply consistent reasoning to arrive at correct answers. In a way, using prompt-engineering we created a “mini” version of OpenAI’s latest o1 models which has reasoning capabilities.