FIT5212 Data analysis for semi-structured data - S1 2023

Assignment 2

Name: Tsz Yan CHUNG

Student ID: 32973381

Task 1: Recommender System Challenge

1. **Introduction**

This report aims to present the finding of a recommender system built around books. The objective is to build an effective model to make accurate prediction on users’ rating on some unseen records of books, based on their previous rating habit. This report will mainly cover this recommender system and the rating prediction model under the collaborative filtering setting.

1. **Methodology**

In the first step of this report, comparisons between algorithms will be discussed, and the best performing models will be picked for further analysis. For clarification, files applied in this report are:  
- trainset: 56,199 x 4, columns: [ID, user\_id, item\_id, book\_name], no null value.

- testset: 286,136 x 4, columns: [user\_id, item\_id, rating, book\_name], no null value.

- metadata: 1,636,235 x 9, columns: [Name, pagesNumber, Publisher, CountsOfReview, PublishYear, PublishYear, Language, Authors, Rating, item\_id], with null values, where data is missing at random.

Different algorithms were applied in this report. The memory-based collaborative filtering algorithm adopted different detailed algorithms and tested with various epoch and learning rate combinations. As for model-based collaborative filtering algorithm, it adopted to a basic matrix factorization neural network model, and an advanced matrix factorization neural network model that takes on extra features, separately.

1. **Algorithms**
   1. **Memory-based collaborative filtering**

**Step 1: Deciding algorithm (Hug, n.d.; Li 2018).**

In the tutorial activity, the SVD algorithm was introduced. However, there are other algorithms not covered in the tutorial while offered by the surprise library. In this step, we will go through some of them and decide on the two best algorithms to further investigate their result.

Algorithms adopted include:

* Baseline model: a basic model that computes for the relationship between user and items
* Singular value decomposition (SVD) method covered in the tutorial
* KNN Basic: basic nearest-neighbors algorithm covered in the tutorial
* KNNWithMeans: In addition to KNNBasic, the mean rating of each user was also taken into consideration.
* KNNWithZScore: Apply z-score normalization to each user’s KNN prediction results
* KNNBaseline: KNN prediction, while taking the baseline linear regression’s prediction result into account
* Non-negative Matrix Factorization (NMF): A similar matrix factorization algorithm to the SVD, with the constraints that the latent features must not be negative
* NormalPredictor: Basic predictor that assumes that data is normally distributed with no learning process.

By iterating through the above algorithms, the following table of evaluation result was returned:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | test\_rmse | test\_mae | fit\_time | test\_time |
| BaselineOnly | 0.885706 | 0.706468 | 1.035147 | 0.373926 |
| SVD | 0.893029 | 0.708788 | 5.417379 | 0.606191 |
| KNNBaseline | 0.922535 | 0.725123 | 2.072465 | 3.3104 |
| KNNWithZScore | 0.950551 | 0.748652 | 1.328606 | 3.305496 |
| KNNWithMeans | 0.950937 | 0.751212 | 1.000784 | 3.531999 |
| NMF | 0.996664 | 0.792071 | 13.23311 | 0.328419 |
| KNNBasic | 0.999318 | 0.789699 | 0.868566 | 3.218344 |
| NormalPredictor | 1.344041 | 1.074687 | 0.37043 | 0.512161 |

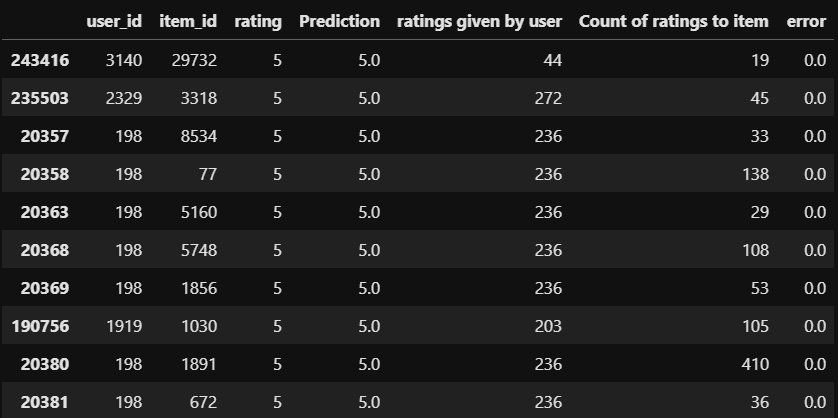
Which shows the BaselineOnly model and the SVD are the best performing algorithms for the train dataset, with the lowest value of RMSE and MAE, possibly due to its relatively simple patterns and relationship between users, items, and ratings.

As a result, in the following memory-based Collaborative Filtering analysis, the BaselineOnly model and the SVD algorithms will be applied.

**Step 2: Apply algorithms.**

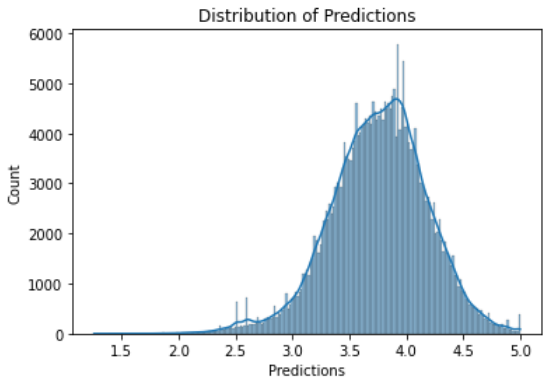
BaselineOnly:

Best predictions from BaselineOnly model:



Worst predictions from BaselineOnly model: A screenshot of a graph

Description automatically generated with low confidence



From the prediction analysis, results from the BaselineOnly model are left-skewed and unable to give accurate prediction on low rating results.

SVD

For this algorithm, the singular value decomposition (SVD) method was applied. The SVD does not require extra features to be input, and evaluate imply based on users’ and items’ feedback, so, only the test set and train set were utilized. In other words, this is relatively easy to implement and computationally effective, while maintaining decent prediction result.

Alterable parameters include:

* Learning rate of model
* Number of training iteration
* Magnitude of regularization term to relieve the issue of overfitting.

The train set was iterated with different combinations of model parameters (with regularization term remained constant), and resulted in the following two graphs regarding its relationships with Root Mean Square Error (RMSE) and Mean Absolute Error (MAE):

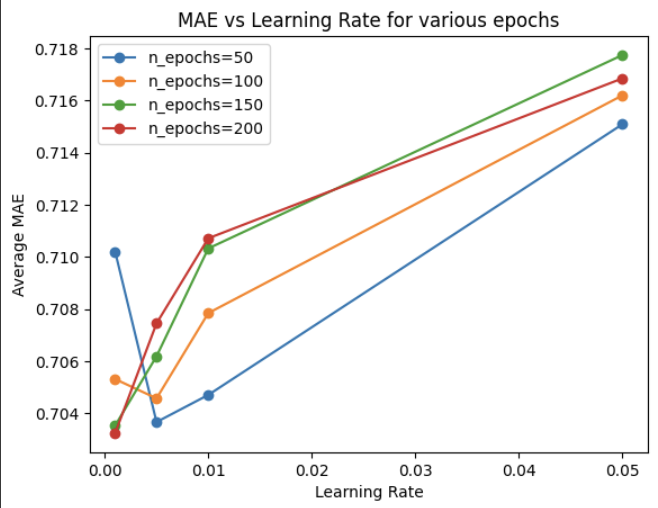
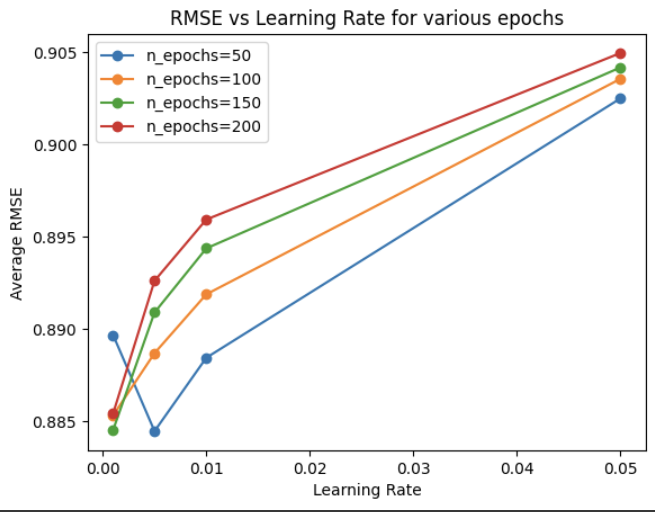
Fig 1. Average MAE vs learning rate for different number of iterations

Fig 2. Average RMSE vs learning rate for different number of iterations



From the above graphs, it can be concluded that the following combination of parameters outperformed the rest of the other combination, with significant lower RMSE and lower MAE values:

1. Learning rate = 0.001, number of iterations = 200 (best performance mode)
2. Learning rate = 0.001, number of iterations = 150
3. Learning rate = 0.005, number of iterations = 50

**Step 3: Summary of memory-based Collaborative Filtering results**

|  |  |
| --- | --- |
| Algorithms | Kaggle result (MAE) |
| BaselineOnly | 0.70462 |
| SVD (lr = 0.001, n\_epochs = 200) | 0.69876 |
| SVD (lr = 0.001, n\_epochs = 150) | 0.7033 |
| SVD (lr = 0.005, n\_epochs = 50) | 0.69905 |

Generally, the memory-based algorithms have very similar result on the test set. Further post-processing methods were applied to the prediction result, for instance, rounding predictions results to integer values or 2 decimal places. However, it was not able to see significant improvement in the MAE results, the results are even worse for most of the time. Therefore, no further post-processing will be applied.

Prediction from the best SVD model

Predictions made from the best performing SVD model experienced similar bias to the BaselineOnly model, prediction results are left skewed and low rating predictions were not made accurately. Indicating that memory-based models are potentially influenced easily to make accurate predictions on biased data.

A picture containing plot, diagram, screenshot, line

Description automatically generated

**3.2 Model-based collaborative filtering**

Algorithm 1: Basic Matrix Factorization neural network (Basic MF)

For the Basic MF model, only the users, items, and rating columns in train set were considered. The Basic MF with bias neural network class from tutorial activity were the model adopted in this task.

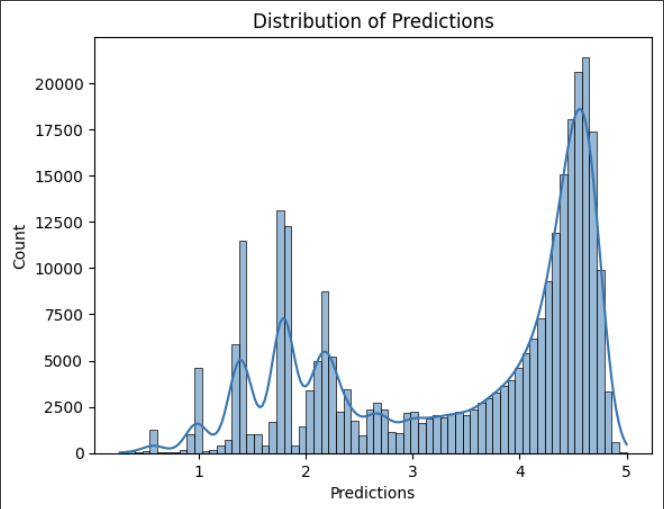
Learning rate was kept at 0.01, embedding size kept at 100, with regularization term as 5 x 10-5, in order to compensate the potential overfitting issue caused by the high learning rate.

* Normalization

As the prediction results given by the neural network model ranged outside of the 1 – 5 zone, a basic rescaling method was applied, where:

And no further processing was conducted on the prediction results, and the prediction results on the test set will be concluded together with the Advanced MF model.

Fig 3. Comparison of prediction results to their true label in the train set

A picture containing text, diagram, line, plot

Description automatically generated

Algorithm 2: Advance Matrix Factorization neural network with extra features (Advance MF)

In this model, on top of the users, items, and rating columns in the train set, information in the metadata file will also be considered.

Step 1: Metadata preprocessing

Data in the metadata file is not organized and contain various data type, the following preprocessing steps were applied:

Item\_id: dropping all records where item\_id is null.

Language:

1. Grouping all variation of English (en-US, en-GB, en-CA) to “eng”.
2. Impute for null value from author column, assuming an author only write in one language.
3. Remove rare language (count less than 50) or null value as “other”
4. Assign categorical value to the remaining language value

Authors:

1. Get count of books written by each author in the dataset.
2. Divide the range from 0 to max count in 100 bins, assign counts to bins accordingly.
3. Further convert them as categorical value, where null values were represented by “0”.

CountOfReview:

1. Divide the range from 0 to max count in 100 bins, assign counts to bins accordingly.
2. Convert them into categorical value, where null values were represented by “0”.

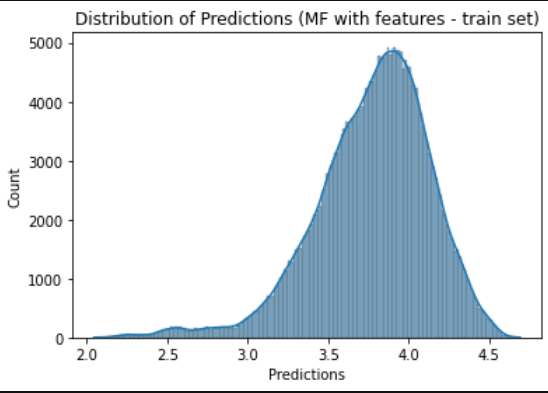
Dropped columns: pagesNumber, Publisher, PublishYear, Name

Step 2: Define MFWithFeatures class

The code of this class referenced the model built in the tutorial activity. However, part of it in the \_\_init\_\_ and forward method were altered so that it matches with the parameters in this book rating prediction work case.

The model was run with 50 training epochs with a learning rate of 0.01, and a regularization term of 5 x 10-5. The same normalization method applied in Basic MF model was also applied to the Advance MF results.

Distribution of prediction results on the trainset from Advance MF model:



The results from the Advance MF model present

Kaggle results:

|  |  |
| --- | --- |
| Algorithms | Kaggle results (MAE) |
| Basic MF | 1.0528 |
| Advance MF | 0.7789 |

Comparing Basic MF and Advance MF

For Basic MF, the distribution of prediction results demonstrated a similar pattern to the true labels of the trainset, potentially indicating an overfitting issue. Thus, leading to relatively poor performance on the test set.

As for Advance MF, it exhibits similar characteristics to the memory-based models, with making majority of the predictions on the far end on the scale. One possible reason is the categorical values in the merged dataset are in range far larger than the range of rating (1 to 5), hence resulting in bias results. If all categorical values are limited to 5, it is likely that the biasness in the prediction results could be relieved.

**4. Conclusion**

To conclude, this report has covered various models: BaselineOnly, SVD, BasicMF, and AdvancedMF with extra features.

Surprisingly, the simplest model, BaselineOnly, outperformed the more complex ones slightly in our experiments. While the performances of all models were generally similar, BasicMF demonstrated some overfitting issues, suggesting it may not generalize well to unseen data.

One potential explanation for the BaselineOnly model's superior performance is that the patterns in the dataset might be simpler than expected, reducing the advantage of more complex models. It's also noteworthy that the AdvancedMF did not perform as expectedly better over the other models.

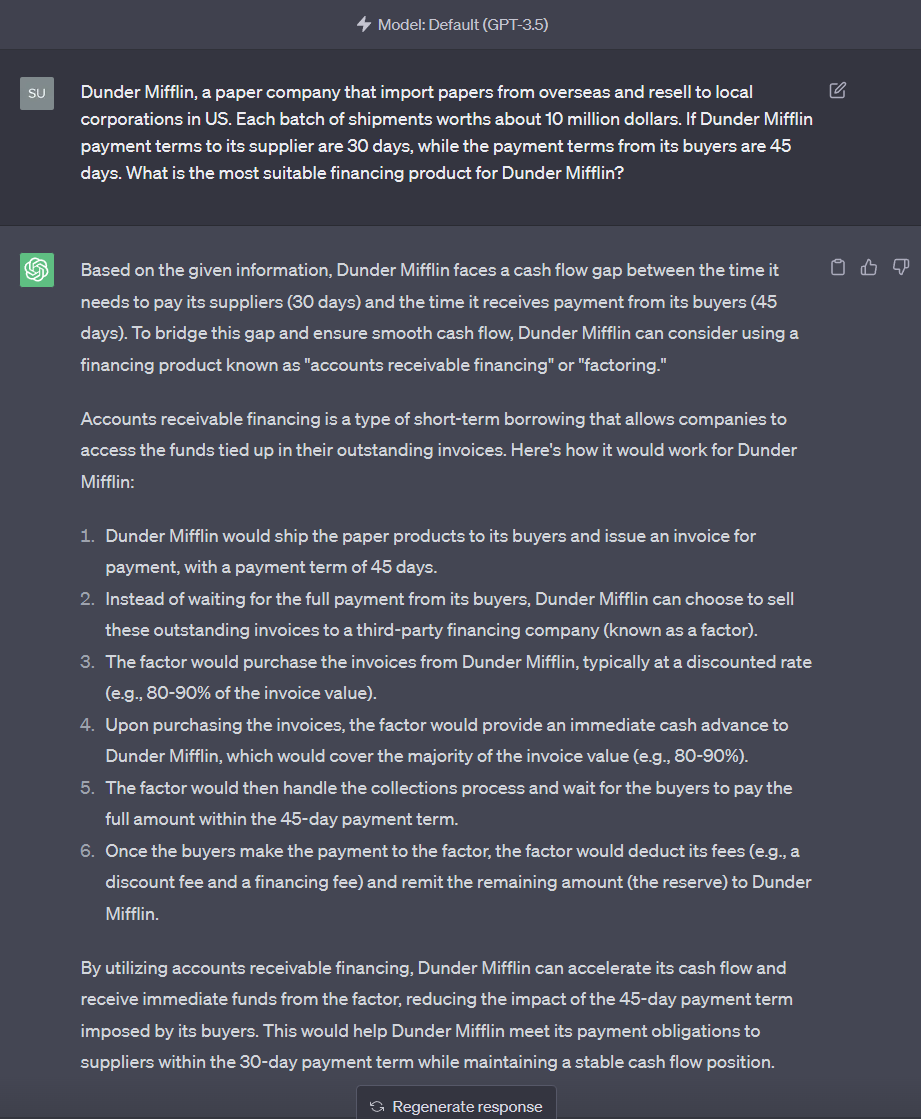
This could be indicative of improper implementation or insufficient tuning of the model's hyperparameters or may suggest that the extra features included did not significantly contribute to the prediction accuracy.

Future work should focus on reviewing the implementation of AdvancedMF and exploring other potential features implementation method or more complex modeling approaches that could better capture the patterns in the data.

**Task 2: ChatGPT**

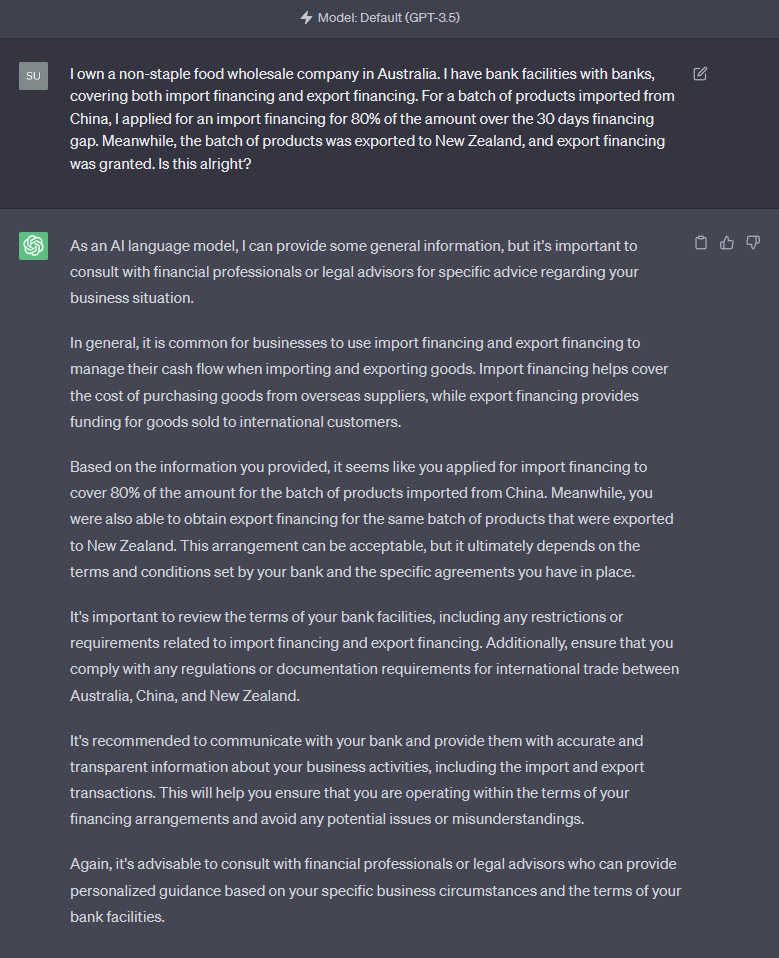
Domain 1: Banking and Finance industry

1. Wednesday 24 May 2023 18:25



In the banking industry, factoring is considered as a financing method that requires excessive commitments from the banks and the client during the audit and execution process. While it is definitely a possible option, it is not the common practice to adopt factoring. Due to the nature of factoring requires the company to pledge their account-receivable during the financing period, which may significantly stagnant cash flow and potentially cause issues.

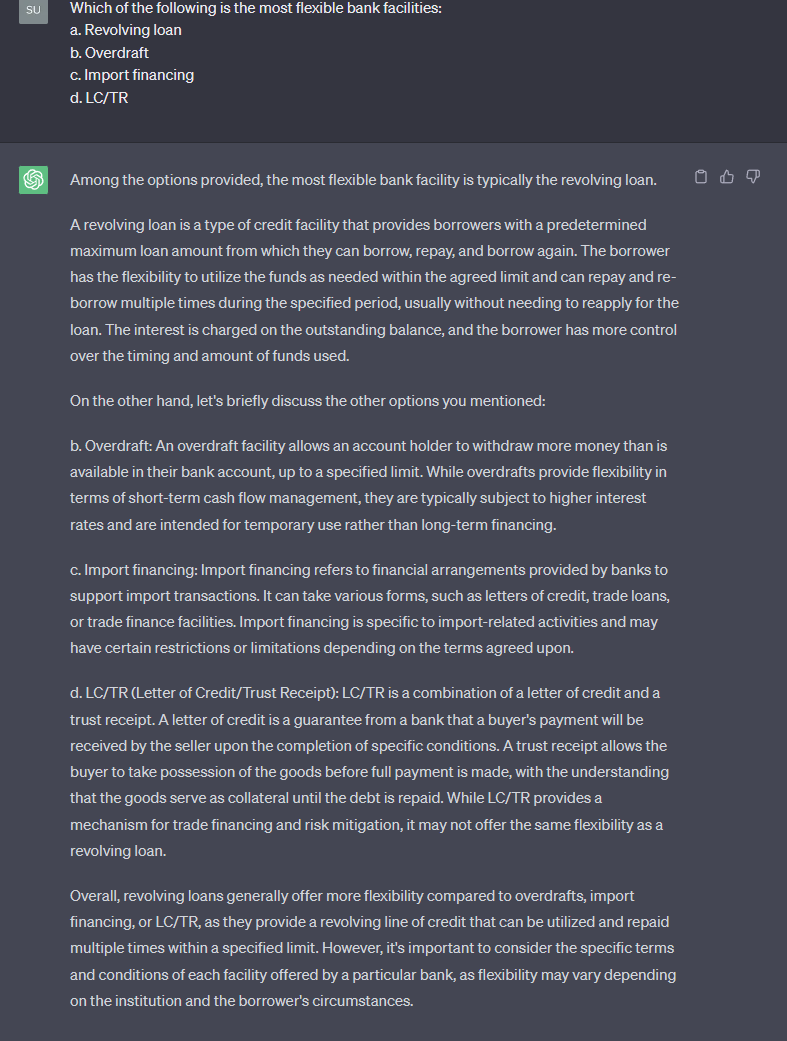
1. Wednesday 24 May 2023 18:38



ChatGPT has failed to notice the potential double financing issue in the example, where a company is receiving financing from a same batch of goods, that may violate facility contract with the financial institute and may cause severe issue.

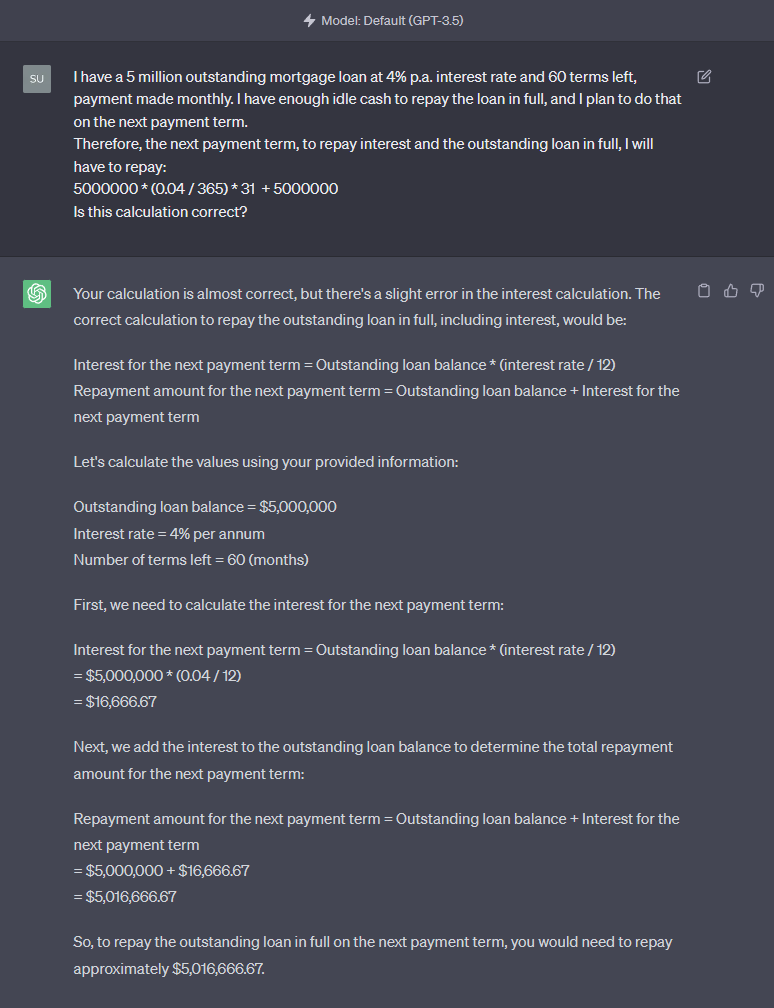
This confusion could also be caused by the attention mechanism embedded in the model, leading the model to focus on other sections of the prompt and failed to notice other elements in the text.

1. Wednesday 24 May 2023 18:51



While revolving loan is also flexible, in the list of four, overdraft has to be the most flexible facility of them all. The model gives hallucinating responses which also take into consideration of interest rate, repayment terms, neglecting the main element in the question is about the flexibility.

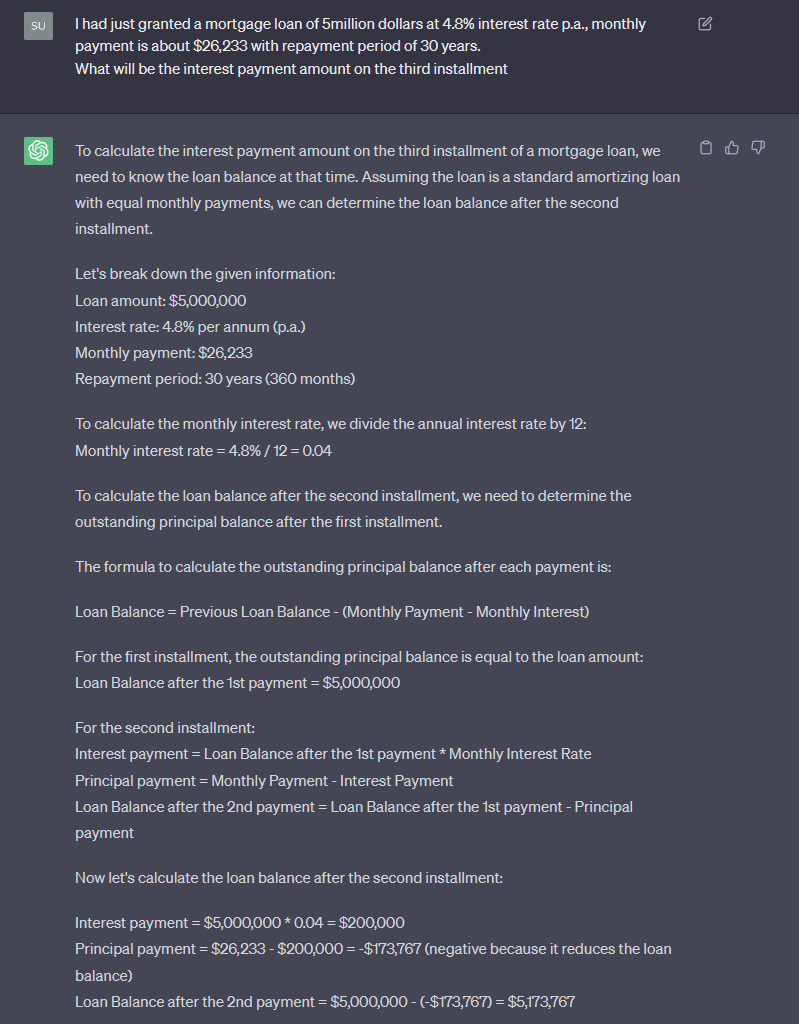
4. Wednesday 24 May 2023 19:12



ChatGPT did its calculation correctly, while this is not primarily a mathematical calculation question.

In the banking industry, an early prepayment penalty is very common. My input prompt may have misled it to focus only on the calculation part, while the main issue lies within the concepts behind the action of repaying a loan before maturity.

5. Wednesday 24 May 2023 19:23



Finally, mathematical calculations question to challenge ChatGPT while it is known that it is relatively weak at. The error was made at very early stage where monthly interest 4.8%/12 should be 0.004 instead of 0.04.

As ChatGPT model was primarily built as a large language model with the objective to understand natural language and to generate human-like responses. To achieve that, Bubeck et al. suggested the issue of reverse reasoning and verification, implying the model’s tendency to generate a full output even if the initialization phase is already incorrect.

Domain 2: General (Commonsense, Logical, Temporal

1. Wednesday, 24 May 2023 17:22

A screenshot of a computer

Description automatically generatedA picture containing symmetry, line, pattern, colorfulness

Description automatically generated

Bubeck et al. (2023) suggested that Text-based games pose a significant challenge for language models as it is challenging for them to comprehend natural language, reason about current game state, while generating valid comments all at the same time.

They may also encounter similar traits to a RNN model that may lose track of elements during the computation process (Kiddon, 2016), and hence, lead to the reasoning failure occurred in this example.

2. Wednesday, 24 May 2023 17:26

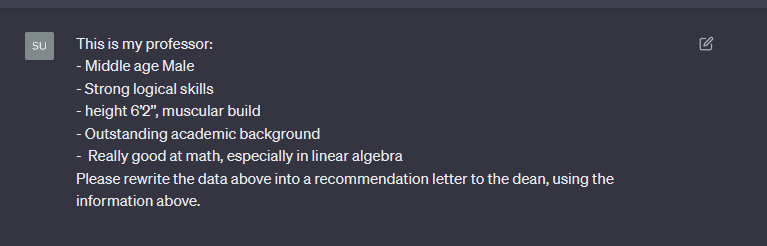
A screenshot of a computer

Description automatically generated with medium confidence

In this hypothetical drawing scenario, GPT, as a large language model, is not able to recognize the influence of physics in painting, indicate its limitation on processing spatial ideas, in a structure that could be rare in its training data.

According to Davis (2023), commonsense knowledge and reasoning have unclear scope and involved highly intelligent reasoning process which is easily affect by noise in data. Due to the massive dataset that ChatGPT was built upon, it may have suffered from some subpar quality of data in this response that led to this reasoning error.

1. Wednesday, 24 May 2023 17:37 (Bubeck et.al., 2023)

(As GPT’s response is long, I will analysis before giving the screenshot)

This question has referenced Bubeck et.al. (2023) example in their paper about employing GPT 4 to produce medical record of a female patient.

Their paper refered such situation as hallucination, imply the language model creating new information on his own, in this case, creating information of the professor’s leadership ability and other personal traits, that wasn’t include in the prompt.

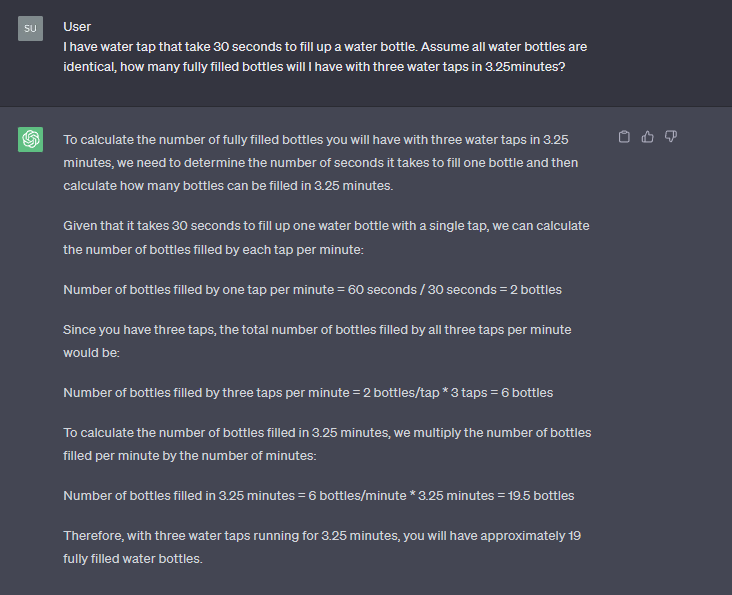
While hallucination could be useful in some situations, especially when it is asked to create some creative contents, it could cause potential issue when the industry requires only precise and accurate information, like the healthcare sector.

A screenshot of a computer

Description automatically generated with low confidence A picture containing text, screenshot, font

Description automatically generated

1. Wednesday, 24 May 2023 17:50

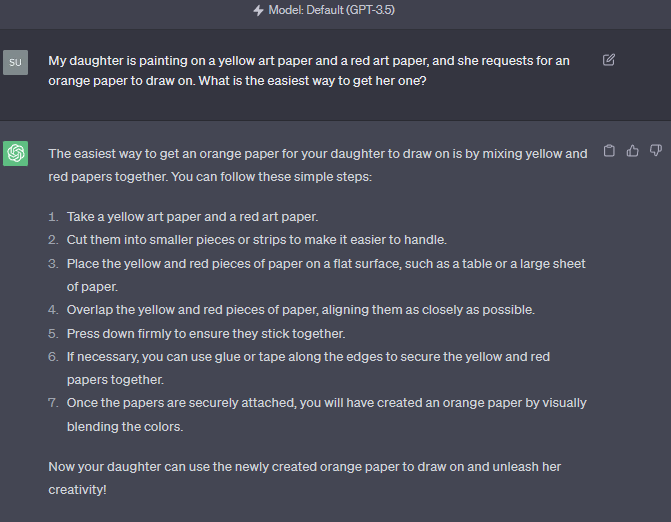


Liu et.al. (2023) suggest that ChatGPT can solve well-known logical comprehension reason task accurately, but may struggle when the question is “out-of-distribution”.

In this example, I specifically asked for the amount of fully filled bottles, while ChatGPT is not able to distinguish between commonsense reasoning problem and numerical calculation question, and hence lead to such inaccurate output.

It directly jumped into calculation to handle my question, which failed to recongize the nature of the question.

1. Wednesday, 24 May 2023 18:02



This example is another demonstration of ChatGPT’s commonsense reasoning failure.

From its training data, it is able to recognize the combination of colours, as formula. Therefore, when it comes to this question, it directly apply its knowledge that the mixture of yellow and red would become orange, possibly tricked by the attention mechanism that correlated yellow red and orange too quickly while disregarding the physical constraints on the objects itselves.

References:

Bubeck, Sebastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, et al. 2023. “Sparks of Artificial General Intelligence: Early Experiments with GPT-4.” *ArXiv Preprint ArXiv:2303.12712*. https://arxiv.org/abs/2303.12712.

Davis, Ernest. 2023. “Benchmarks for Automated Commonsense Reasoning: A Survey.” *ArXiv Preprint ArXiv:2302.04752*, February. <https://arxiv.org/pdf/2302.04752.pdf>.

Hug, Nicolas. n.d. “Home.” Surprise Home Page. https://surpriselib.com/.

Li, Susan. 2018. “Building and Testing Recommender Systems with Surprise, Step-By-Step.” Medium. Towards Data Science. December 26, 2018. https://towardsdatascience.com/building-and-testing-recommender-systems-with-surprise-step-by-step-d4ba702ef80b.

Liu, Hanmeng, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. “Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4.” *ArXiv Preprint ArXiv:2304.03439*, April. https://arxiv.org/pdf/2304.03439.pdf.