May:

- Task 1: Based on the current database JSON sample file of the product scheme, develop a Python script that auto extracts the data into a csv file which contains only the data we need used for ML purpose. The script assumes using a JSON file as an input, and it can auto download the image data to the target directory from the given link inside the JSON data. (Xuping, Shadan)
  - Task length estimation: 1 week
- Task 2: Define new labels for recommendation systems based on the current data scheme, while fetching some input with the company. The label needs to fulfill the following: (Xuping, Shiyao, Shadan)
  - Appears to be useful in the recommendation systems.
  - It's possible to be predicted by ML models based on the raw data that comes with the product.
  - It's not inside the raw data that comes with the product as those are already there and have no need to be predicted again. However this excludes the case where the label is not standlized, I already sent ShopHopper an email to confirm this.
  - Task length estimation: 2 - 3 weeks
- Task 2.5: Label the new defined labels in the csv file: (Archi)
  - Task length estimation: on going with the task 2
- Task 3: Search for possible ML models/strategies that can be used for label prediction in this task, and based on the most of the searched applicable methods, define the data pre-processing methodology. Need to define which raw data that comes with the product will be used as possible model input first. (All members of the team - however this is also an ongoing task so if a member have other tasks in hand he/she may consider focus less on this task during the completion of the other task)
  - For deep learning approaches the image should be applied with padding as images don't have a consistent width-height ratio.

    - For the text processing part, there's probably a need for keyword filtering for small phrases before encode the text and feed it to the model. For the body_html data, it needs to be processed to become pure text as well.

    - Task length estimation: 3 to 4 weeks as a starting period to have some possible approaches to ongoing through the later month (note this task can also be considered as an ongoing task through the project if the progress of other tasks are ideal, as there could always be approaches to explore).

- Task 4. After the preprocessing method is defined, implement the base version of the script to apply the pre-processing of the data extracted in Task 1. (Shadan, Mingwei, Luis)
    - Task length estimation: 2 weeks (note the finish date of this task can be extended into June).

- Additional notes for tasks in May:

    - Task 1, 2, and 3 can be started simultaneously on day 1, for task 3 we already have some labels designed to be predicted so it doesn't have to wait until task 2 is finished.

    - Task 4 should be able to start when task 3 is half completed, it doesn't need to wait for task 3 to be fully completed.

June to July (or to early/mid August):

- Task 5. Implement the ML model and its training script and make sure it's working with the prepared data. And train the model based on the training script implemented in the previous step while tuning the parameters until it gives acceptable results. And after that (This task will include all members)

    - Note that despite the possible large structural difference between different ML approaches as the model probably takes multiple inputs, some sub sections of the model could be using the same sub model or module, such as word embedding. So members are encouraged to help each other despite them working on different models. Members are also encouraged to help another member's method that he/she is working on with getting input from the other member.

    - It is strongly recommended to make sure your training script is working and is actually training using a very small amount of data, before loading the whole dataset into it as it requires more computation to load the actual dataset.

    - Note that the GPU machine in the lab can be booked for this task.

    - There could be some other model to explore as well, different members can search/design a unique ML model/approach (can use a different set of input from the company data) and each of them may work on training that model by that person and the result can be compared. Multiple team members are okay to work on one approach if they are able to communicate closely with each other, or can work on their own approach too.

    - Members shall, compare the performance of different approaches in the following aspects in order to select the best approach from the following aspects:
        - Accuracy of the label predicting.
        - Computational cost
        - How easily can it be integrated in the company's tech stack

    - Task length estimation: around or over 2 month as this is the main implementation/testing part of the project, and note this task also refers to task 3's ongoing - which means members can always explore new possible ML approaches to predict labels and try to implement/test them.

August:

- Task 6. After an acceptable approach is proved, if time allows, the team can work closely with the company's team and define or actually implement a pipeline to integrate the AL model to the company's technology stack. (This task will include all members)

    - Task length estimation: around 2 weeks, and note this task should not be started unless task 5 is fully finished and an acceptable proof of concept is available.