



NUS Logo **KENNETH**

TITLE  
subtitle

Our names **KENNETH**

Overview/ Problem we are trying to solve/ Why is it commercially important/ Show some examples where different users ask the same question in different ways

**KENNETH**

Proposed method -  
Generating questions +  
Finetuning **KENNETH**

Dataset - (i) What data we have (ii) How we augmented it (iii) show some well generated questions (cherry pick) **RAIMI**

Some questions (from the test set) that the model (trained with generated questions) did well **AZMI**

Azmi's Performance (bar chart)  
**AZMI**

Pipeline (some diagram)  
If there is space, then we will include the T5 architecture  
**RAIMI**

Evaluation/ MRR metric explanation  
**KENNETH**

Future improvements (other ways of augmentation)  
**RAIMI**

Conclusion  
**KENNETH**

Our student email  
Github link + QR code  
**KENNETH**

## Overview

Information retrieval models generally require large amount of training data to perform well. Unfortunately, most companies do not have sufficient data for training as they might not have data collection pipelines in place to collect training data. To solve the lack of training data, we generate questions for the training set by using the Text-to-text Transfer Transformer (T5) model. We believe that the performance of the information retrieval (DistillBert) model will improve after being trained on the generated questions.

## Proposed Method

Definitions:

$Q_O$ : Original questions given to us.

$A_O$ : The universe of possible paragraphs/ documents that can be returned to the user of the information retrieval system

$Q_G$ : The questions generated by T5

$Q_{O+G}$ : The final set of questions used to fine tune the information retrieval model

To finetune a T5 model that is capable of generating  $Q_G$ , the T5 model is first fine tuned against the SquAD dataset. During the fine tuning process, the aim of the T5 model is to generate questions based on a given span. The fine tuned T5 model is then used to generate  $Q_G$  from  $A_O$ . Finally,  $Q_{O+G}$  and  $A_O$  are used to finetune the DistillBert (information retrieval) model.

## Dataset

We used a dataset from the Ministry of Health's frequently-asked questions on **Covid-19 Vaccination and Booster**<sup>1</sup> which informs Singapore citizens how to manage Covid-19. This dataset consists of 188 pairs of questions and answers.

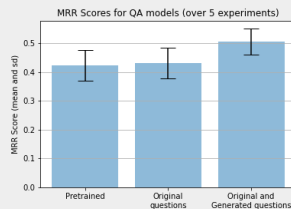
This dataset will be used to (1) fine-tune the information retrieval model to fit our domain and (2) generate new questions based on the given answers.

<sup>1</sup> <https://ask.gov.sg/agency/moh/topics/COVID-19/2020/vaccination%20and%20booster>

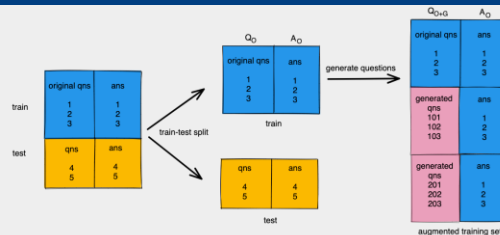
## Question Generation in Action

Question		
I am a dialysis patient. I tested positive on an Antigen Rapid Test (ART) self-test. What should I do?		
Model	Answer	Score
Pre-trained	You may refer to the instructions on the Antigen Rapid Test (ART) self-test kits. Click here for more information about ART self-testing.	0.666508
Original questions	You may refer to the instructions on the Antigen Rapid Test (ART) self-test kits. Click here for more information about ART self-testing.	0.657461
Original and Generated questions	Upon testing positive on ART self-test, please do the following: a.)Go for a confirmatory PCR test at any SASH clinic or Community Test Centre (CTC). It is important to also inform them that you are a dialysis patient so that you fall under Protocol 1; b.)Please report your ART positive result to your usual renal provider (i.e. the dialysis centre); c.)Your dialysis centre should then make arrangements to ensure that you attend your next dialysis session without delay. Please contact your dialysis centre to ensure that you do not miss your dialysis session whilst you wait for your PCR test results.	0.602046

## Performance



## Pipeline



The training set is augmented with more question-answer pairs (red), where the questions are generated by from the existing answer set in the training set. In our implementation, we generated 10 sets of questions (shown above as 2). This augmented training set, together with the test set, are used to fine tune the model.

## Evaluation Metric

Mean Reciprocal Rank (MRR) is generally used as the evaluation metric for information retrieval models. Given a scenario with 2 queries, the MRR score is  $[(1/3) + (1/1)]/2 = 2/3$

Question	Proposed answer	Correct answer	Reciprocal rank
Q1	A2, A5, A3	A3	1/3
Q2	A1, A4, A5	A1	1/1

## Future Improvements

Future improvements that can be done include:

- Generating more than 10 sets of questions for every answer
- Use other data augmentation techniques like synonym replacement and random deletions to the questions
- Using multiple question generators (eg. GPT2) to generate the questions

## Conclusion

We have shown that the performance of the information retrieval model is best when trained with questions generated from the T5 model. We believe that the process of generating new questions is similar to the act of augmenting data. Ultimately, the augmented data serves to represent different ways that the same question can be asked. By exposing the information retrieval model to such questions, we believe that the information retrieval model can perform much better in practical scenarios. If you have any questions, do contact us.

Raimi bin Karim: [raimi.karim@u.nus.edu](mailto:raimi.karim@u.nus.edu)  
Azmi bin Mohamed Ridwan: [E0576209@u.nus.edu](mailto:E0576209@u.nus.edu)  
Wang Tian Ming Kenneth: [E0573193@u.nus.edu](mailto:E0573193@u.nus.edu)  
Github link: <https://github.com/kenneth-wang/cs5260>

## Title

- Improve question answering through question generation
- Adaptive Information Retrieval Systems (AIRs)

## Subtitle

- Making information retrieval smart again
- Data augmentation meets information retrieval
- Improve the performance of information retrieval systems through question generation

# Potential questions

- Why is it considered “adaptive”?

# Azmi Temp

<b>Model</b>	<b>MRR Score</b>	<b>Standard Deviation</b>
Pre-trained	0.4238	0.0532
Using Original Questions only	0.4322	0.0535
Using Original and Generated Questions	0.5063	0.0456

Question		
I am a dialysis patient. I tested positive on an Antigen Rapid Test (ART) self-test. What should I do?		
Model	Answer	Score
Pre-trained	You may refer to the instructions on the Antigen Rapid Test (ART) self-test kits. Click here for more information about ART self-testing.	0.666508
Original questions	You may refer to the instructions on the Antigen Rapid Test (ART) self-test kits. Click here for more information about ART self-testing.	0.657461
Original and Generated questions	Upon testing positive on ART self-test, please do the following: a.\tGo for a confirmatory PCR test at any SASH clinic or Community Test Centre (CTC). It is important to also inform them that you are a dialysis patient so that you fall under Protocol 1; b.\tPlease report your ART positive result to your usual renal provider (i.e. the dialysis centre); c.\tYour dialysis centre should then make arrangements to ensure that you attend your next dialysis session without delay. Please contact your dialysis centre to ensure that you do not miss your dialysis session whilst you wait for your PCR test results.	0.602046