An Intelligent Question Suggestion System for Adaptive Learning (IQSAL)

Ashima Fatima Seik Mugibur Raghman,
21BAI1830

Background and Significance of Study

- Traditional methods of assessing exam question similarity and cognitive difficulty are often **time-consuming, subjective, and lack the precision** required for personalized learning.

- The advent of **digital and adaptive learning platforms** and the growing emphasis on individualized education, has created a huge *market gap* for question recommendation systems. This is more so with isolated *self-learning* becoming the norm as a consequence of *remote learning* the during COVID-19 pandemic.

- Another key factor behind the creation of this system is to reduce the **lack of support in exam preparation** for students by schools due *systematic disadvantages* such as financial, geographic location, disabilities, etc.

- Students require a **intensive practice on personal weaker aspects** in order for more effective studying rather than the current practice of solving entire subject practice papers.

Problem Statement and Research Background

The Intelligent Question Suggestion System for Adaptive Learning (IQSAL) is a software solution to enhance the effectiveness of exam preparation, here specific to the British GCSE and A-level examinations, by providing an adaptive and personalized learning experience through intelligently recommending practice questions tailored to individual learning needs.

By leveraging NLP algorithms based on Bloom's taxonomy and word similarity models, IQSAL identifies the cognitive skills required for each question and establishes the semantic relationships between topics, ensuring targeted and relevant question recommendations.

Background:

- NLP research has extensively explored solutions to text similarity in domains such as FAQ query-question similarity, exam question classification, essay plagiarizing, exam cheating, question generation (multiple-choice), short answer generation, etc.

RESEARCH GAP

- In this project, I am to focus on 2 aspects of questions to assess their similarity
  1. The cognitive difficulty level of the question
  2. The semantic similarity of the question to assess similar subject and topic matter.

Key Objectives

The key objectives of this system are:

- To tailor question recommendations based on user performance for a real-time adaptive learning experience.
- To develop a system that accurately categorizes questions using Bloom's taxonomy so that recommended questions match required level of cognitive skill.
- To implement a word similarity model for identifying relevant topics and suggest related questions if required by user to focus on areas of weakness in exam content.
- Create an intuitive user interface for seamless interaction and feedback.

Educational Taxonomies

Educational taxonomies are frameworks that **categorize and classify learning objectives**, providing a structured approach to define and assess cognitive skills in educational settings.

1. BLOOM'S TAXONOMY

Categorizes cognitive skills into 6 hierarchical levels from simple to complex.

2. ANDERSON'S TAXONOMY

A development on Bloom's taxonomy. Emphasizes **action verbs** and **precise language** in learning objectives.

From this taxonomy, we have **a matrix of verbs** developed in Chang[1] to perform text classification on.

RESEARCH GAP: current research on taxonomies in questions has only focused on <u>the identification to provide a balanced question set in examinations </u>for proper assessment purposes and no research exists on using taxonomies for <u>question recommendation according to cognitive difficulty level</u>

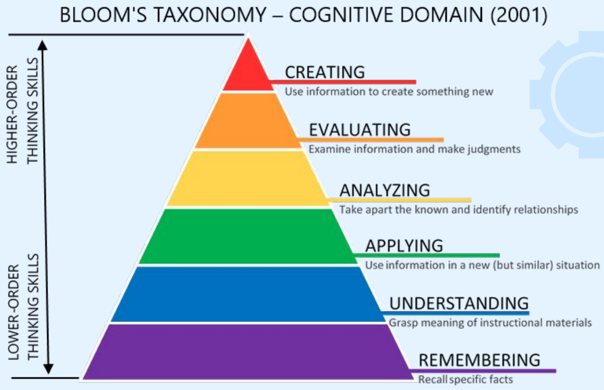


TABLE I.
ANDERSON'S REVISIONS ON BLOOM'S TAXONOMY

Category	Cognitive Verb list of Anderson Taxonomy	
	Description	*Verb list*
Categories	Recall or retrieve previous learned information.	defines, describes, identifies, knows, labels, lists, matches, names, outlines, recalls, recognizes, reproduces, selects, states
Understand	Comprehending the meaning, and interpretation of instructions and problems.	comprehends, converts, defends, distinguishes, estimates, explains, extends, generalizes, gives an example, infers, interprets, paraphrases
Apply	Use a concept in a new situation or unprompted use of an abstraction	applies, changes, computes, constructs, demonstrates, discovers, manipulates, modifies, operates, predicts
Analyze	Separates material or concepts into component parts	analyzes, breaks down, compares, contrasts, diagrams, deconstructs, differentiates, discriminates, distinguishes, identifies, illustrates
Evaluate	Make judgments about the value of ideas or materials	appraises, compares, concludes, contrasts, criticizes, critiques, defends, describes, discriminates, evaluates, explains, interprets
Create	Builds a structure or pattern from diverse elements	categorizes, combines, compiles, composes, creates, devises, designs, explains, generates, modifies, organizes, plans, rearranges, reconstructs

Literature Review on Bloom's Taxonomy Exam Question Classification

Studies	Year	Processing and Classification Method	Dataset	Result
Jayakodi et al.	2016	Rule-based (POS tag + WordNet + Cosine Similarity)	53:35, University-level Applied Sciences	Accuracy: 71%
Setyaningsih and Listiowarni	2021	Naïve Bayes and Laplace Smoothing	600:250 High school Biology	Accuracy: 75.94%
Osadi et al.	2017	Ensemble Classifier (Rule-based + SVM + NB + KNN)	60:40 University-level Programming Questions	Accuracy: 82.5%
Aninditya et al.	2019	Text Mining (N-gram + TF-IDF + NB)	8:2 (ratio) University-level Information Sciences	Accuracy: 85%
Thing Goh et. al	2023	Rule-based (SPOS + WordNet)	200 University-level Engineering	Accuracy: 83%

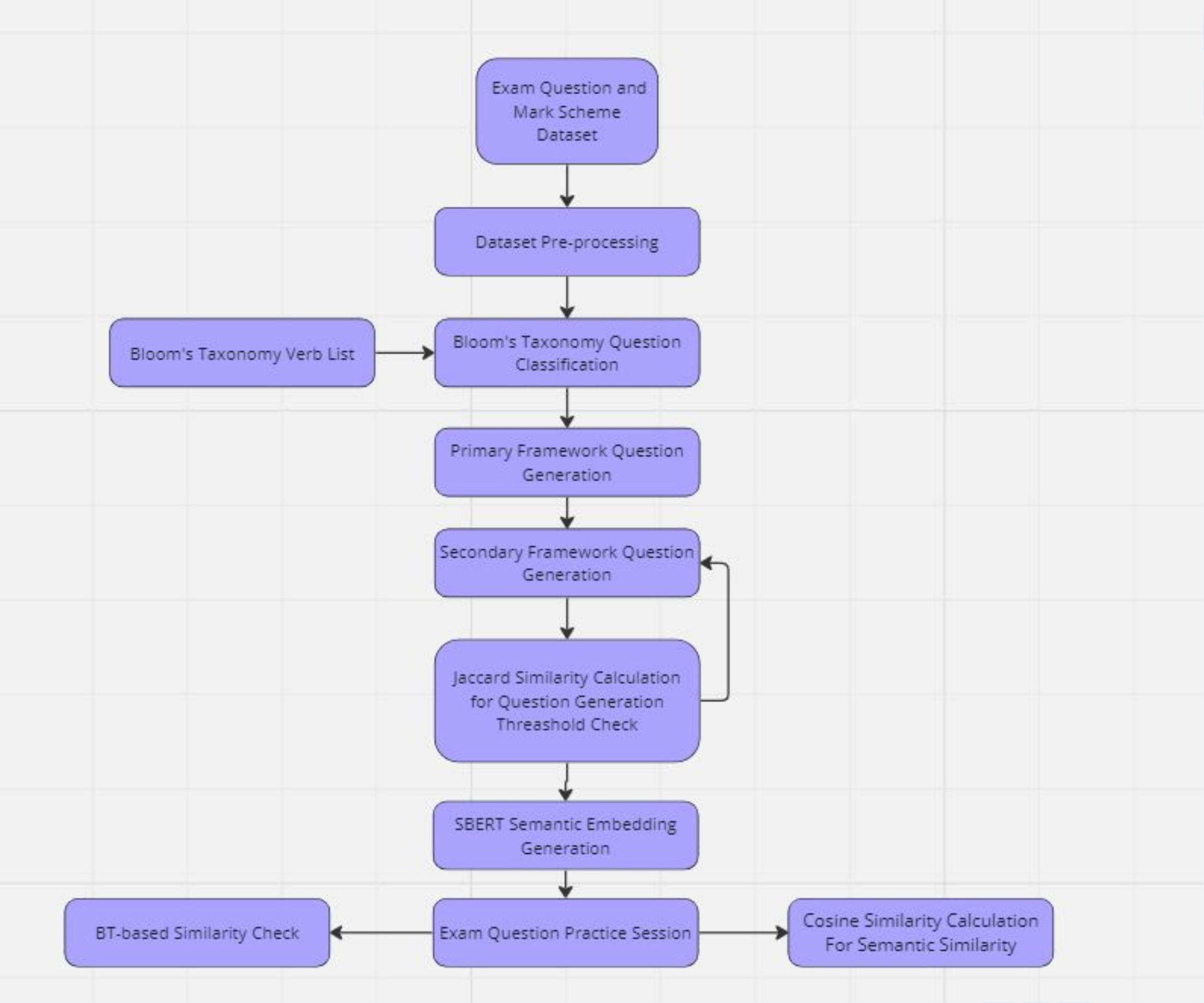
CURRENT CHALLENGE/RESEARCH GAP: There is **no one single public dataset** for exam question that is consistently used leading to uncertainties in comparative studies of performance. Dataset are often span different sizes, levels of study, languages, and subject matters. In order to overcome this problem, I will be using the **internationally well-known British I/GCSE exam papers** which is an **established exam series** due to which subsequent research on the same dataset can allow for improved comparative studies.

Literature Review on Semantic Analysis

Studies	Year	Methodology	AIM
Lopez-Gazpio et al.	2019	N-gram Attention Model	Sentence Similarity
Setyaningsih and Listiowarni	2014	Ontology-based VSM	Question Similarity
Fauzi et al.	2017	N-gram + Cosine Similarity	Essay Scoring
Pratama, et al.	2019	N-gram + Term Frequency + Cosine Similarity	Essay Answer Similarity
Sakata et. al	2019	BERT-based	FAQ Query-Question Similarity

In addition, various other methods such as bag-of-words model, soft-cosine similarity, Jaccard similarity, Levenshtein distance (LD), Singular Value Decomposition (SVD), Long-Short term memory (LSTM) and Bi-direction Long-Short term memory (biLSTM) were used in the context of text/sentence similarity.

Architecture Diagram



Modules

The modules involved in this project are:

1. Data Extraction and Dataset Creation
2. Dataset Loading and Pre-processing
3. Bloom’s Taxonomy Classification
4. Generalization of Scenario-based Exam Questions
5. Semantic Similarity Calculation
6. Adaptive Learning Function

Data Extraction and Dataset Creation

Exam question and Mark Scheme data were sourced from high school Biology Exam Papers (AQA GCSE Biology Paper 1H). The PDFs were downloaded and text data was extracted using PyPDF and then regular expression were used to remove unnecessary information. In order to effectively obtained the question and mark scheme text alone, LangChain prompt-based pipeline was built utilizing Claude API. Extracted text information was exported into an excel file along with manually added metadata about the exam paper and question ID.

For ease of use, only text based question with little intra-question dependencies were added to the dataset and questions utilizing diagrams and math equations were excluded.

Dataset Loading and Pre-processing

Dataset was then loaded as a Pandas dataframe and further data cleaning was done. Data was separately tokenized and added as an additional column to the dataset for function that required such pre-processing

Bloom’s Taxonomy Classification

A dictionary of Bloom’s Taxonomy categories and they verbs/question words was created.

The tokenized preprocessed question text was was then parsed to find matches of verbs/question words and frequency of match with each category were stored. At the end of parsing, category with maximum value was chosen as the Bloom’s Taxonomy category for the question.

Initially, word synets were used to additional use synonyms of the verb/question words list but this proved ineffective and so was removed.

The categorical labels were then mapped to numeric labels for easier future processing.

```
[12] bloom_taxonomy_mappings = {
         "Recall or retrieve": ["define", "describe", "identify", "know", "label", "list", "match", "name", "outline", "recall", "r
         "Understand": ["comprehend", "convert", "defend", "distinguish", "estimate", "explain", "extend", "generalize", "give an e
         "Apply": ["apply", "change", "compute", "construct", "demonstrate", "discover", "manipulate", "modify", "operate", "predic
         "Analyze": ["analyze", "break down", "compare", "contrast", "diagram", "deconstruct", "differentiate", "discriminate", "di
         "Evaluate": ["appraise", "compare", "conclude", "contrast", "criticize", "critique", "defend", "describe", "discriminate",
         "Create": ["categorize", "combine", "compile", "compose", "create", "devise", "design", "explain", "generate", "modify", "
     }
```

Generalization of Scenario based Question

Exam Questions were mostly scenario based and hence contained a large amount of text that was irrelevant to the syllabus or the true intention of the question. This meant that classic semantic analysis would provide futile as the true intended question words were in the minority or sometimes, even non-existing. This would mean that same question with different scenarios applied to them would be processed as vastly different question while actually different questions with similar scenarios would be preprocessed as similar question.

This meant that classic semantic embedding based similarity calculation would be the wrong approach.

To overcome this issue, I aimed to extract a truer "Framework Question" which would be the basic question void of any specific scenario example.

To do this, I created a LangChain prompt-based pipeline to extract the fundamental framework question.

Theory - Causality

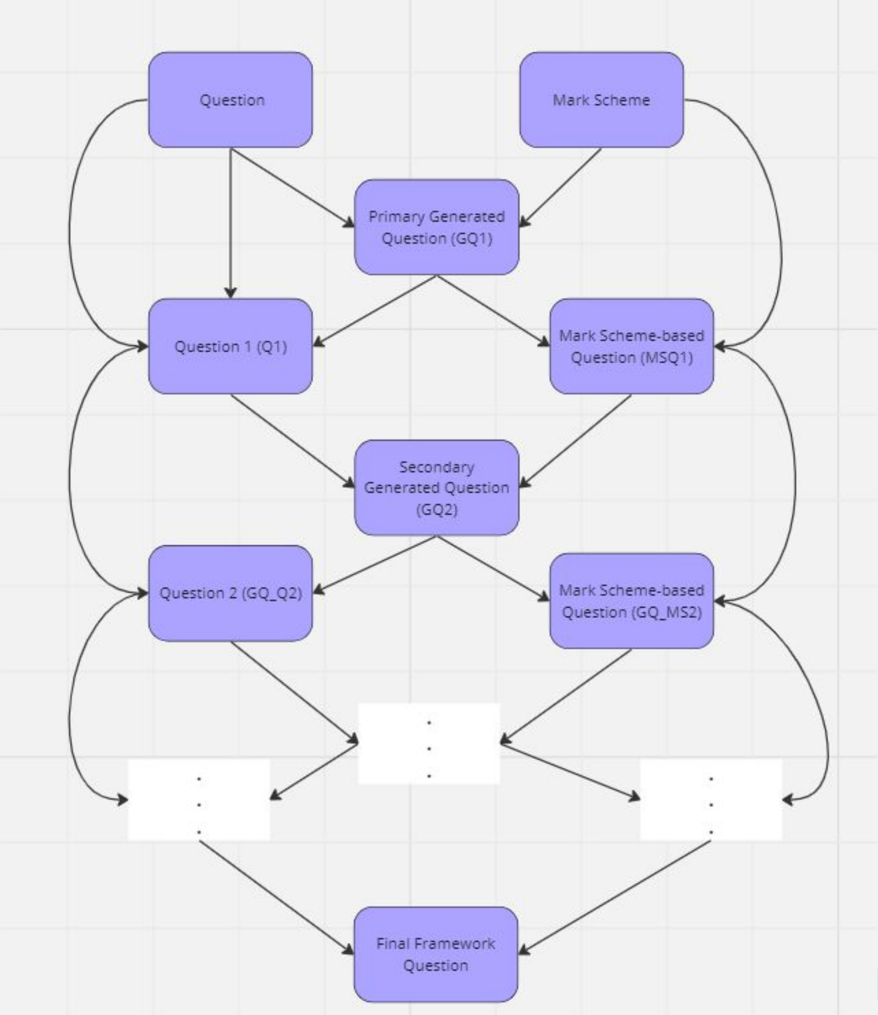
By only utilizing the exam question to extract the fundamental question, generated questions tended to focus too much on the scenario. To overcome this problem, the concept of causality was utilized.

By understanding that **questions were the cause**, that <u>induced</u> the **effect of an answer** (mark scheme). I attempted to approach the problem with the idea that student users, in using an exam similarity program would in fact be attempting to find questions which produce similar answers and that the similarity focus is not on the question itself alone, but on the mark scheme.

Hence, in this scenario, shifting the focus on mark scheme similarity would be more effective than focusing on question similarity. In other words, deciding suitable question (cause) via the intended answer (effect), would be more effective than a vice versa approach.

Thus, the generated framework question must be extracted from both mark scheme the mark scheme and the question so that the focus would be on topic information in the mark scheme and while inclusion of the question would prevent over-fitting on a specific mark scheme.

Framework Question Generation Diagram



Question	Mark Scheme	Q1	MSQ1	GQ1	GQ_Q2	GQ_MS2	GQ2
Eating food containing Salmonella bacteria can cause illness. Two symptoms of infection by Salmonella are vomiting and diarrhoea. What causes these symptoms?	toxins / poisons (secreted by / from / in bacteria)	What type of substances secreted by bacteria can cause vomiting and diarrhoea?	"Explain the role of toxins in bacteria and how they can negatively impact human health."	What is the impact of bacterial toxins on human health?	What are the general effects of bacterial infections on living organisms?	What are the overall effects of bacterial toxins on living organisms?	What are the overall effects of bacterial interactions on living organisms?

An framework of iterative question generation was created, with question generation halting when the Jaccard Similarity between the generated question (GQ) with the previous Question (Q) and GQ with the previous Mark Scheme based Question (MSQ) reached a certain threshold.

Semantic Similarity

Semantic similarity was applied on the final generated question i.e. the framework question using SBERT's transformer-based semantic embedding function and then cosine similarity was calculated to create a similarity matrix.

Adaptive Learning

To create the adaptive learning module, I first initialized an empty set to add questions as they are output to prevent repetition in question recommendation. I then initialized functions to obtain questions within the same Bloom's taxonomy category and questions with semantic similarity.

Initially, the practice session starts with a random question. Users can then choose to move onto another random question, get a question of similar difficulty (based on Bloom's Taxonomy) or on a similar topic (based on semantic similarity). They can then choose to end the practice session after the end of any question.

Result

```
Current Question:
Urea is transported in the blood plasma.
Name two other substances transported in the blood plasma.
Enter 'next' for another random question, 'difficulty' for questions of similar difficulty, 'topic' for more questions on a similar topic, or 'end' to end practice session: next
Current Question:
Which two blood vessels carry deoxygenated blood?
Aorta
Coronary artery
Pulmonary artery
Pulmonary vein
Vena cava
Enter 'next' for another random question, 'difficulty' for questions of similar difficulty, 'topic' for more questions on a similar topic, or 'end' to end practice session: difficulty
Current Question:
Eating food containing Salmonella bacteria can cause illness.
Two symptoms of infection by Salmonella are vomiting and diarrhoea.
What causes these symptoms?
Enter 'next' for another random question, 'difficulty' for questions of similar difficulty, 'topic' for more questions on a similar topic, or 'end' to end practice session: topic
Current Question:
Eating food containing Salmonella bacteria can cause illness.
Two symptoms of infection by Salmonella are vomiting and diarrhoea.
What causes these symptoms?
Enter 'next' for another random question, 'difficulty' for questions of similar difficulty, 'topic' for more questions on a similar topic, or 'end' to end practice session: end
```

Initially, a random question is requested. Next, a question of a similar difficulty level is requested and then a question on a similar topic is requested. Lastly, the practice session is ended.

Conclusion and Evaluation

Thus, the program attains the intended objective of providing practice exam questions based on both question difficulty and topic similarity.

For further improvements,

- Connect the program to Streamlite to create a cloud based hosted site for user usage.
- Improvements can then be made through implementing reinforcement learning based recommendation algorithms to streamline the recommendation process via content-filtering based approaches.
- Lastly, instead of the user manually inputting their requirement and apply more sophisticated methods for assessing semantic similarity can be implemented

Bibliography

K. Jayakodi, M. Bandara, I. Perera, and D. Meedeniya, “WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom’s Taxonomy,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 11, no. 04, p. 142, Apr. 2016, doi: https://doi.org/10.3991/ijet.v11i04.5654.

E. R. Setyaningsih and I. Listiowarni, “Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing,” *IEEE Xplore*, Apr. 01, 2021. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9431862 (accessed Dec. 08, 2021).

K. A., M. G., and W. V., “Ensemble Classifier based Approach for Classification of Examination Questions into Bloom’s Taxonomy Cognitive Levels,” *International Journal of Computer Applications*, vol. 162, no. 4, pp. 1–6, Mar. 2017, doi: https://doi.org/10.5120/ijca2017913328.

A. Anindityа, M. A. Hasibuan, and E. Sutoyo, “Text Mining Approach Using TF-IDF and Naive Bayes for Classification of Exam Questions Based on Cognitive Level of Bloom’s Taxonomy,” *IEEE Xplore*, Nov. 01, 2019. https://ieeexplore.ieee.org/abstract/document/8980428 (accessed Dec. 07, 2021).

Thing Thing Goh, Nor, H. Mohamed, Mohd Nazri Ismail, and H. Chua, “Semantic Similarity Analysis for Examination Questions Classification Using WordNet,” *Applied sciences*, vol. 13, no. 14, pp. 8323–8323, Jul. 2023, doi: https://doi.org/10.3390/app13148323.

I. Lopez-Gazpio, M. Maritxalar, M. Lapata, and E. Agirre, “Word n-gram attention models for sentence similarity and inference,” *Expert Systems with Applications*, vol. 132, pp. 1–11, Oct. 2019, doi: https://doi.org/10.1016/j.eswa.2019.04.054.

J. Yu, D. Li, J. Hou, Y. Liu, and Z. Yang, “Similarity Measure of Test Questions Based on Ontology and VSM,” *The Open automation and control systems journal*, vol. 6, no. 1, pp. 262–267, Dec. 2014, doi: https://doi.org/10.2174/1874444301406010262.

M. A. Fauzi, D. C. Utomo, B. D. Setiawan, and E. S. Pramukantoro, “Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning,” *Proceedings of the International Conference on Advances in Image Processing*, Aug. 2017, doi: https://doi.org/10.1145/3133264.3133303

W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi, “FAQ Retrieval using Query-Question Similarity and BERT-Based Query-Answer Relevance,” *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2019, doi: https://doi.org/10.1145/3331184.3331326.