

H278050: Understanding the Effectiveness of Large Language Models in Question Generation

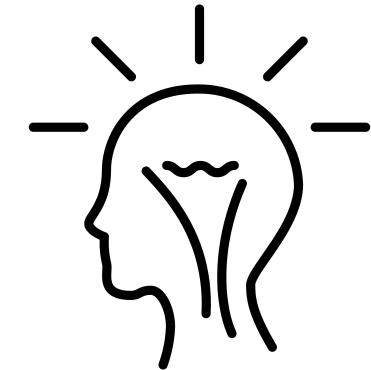
Sumanth Yalamarty

NUS School of Computing

Motivation

Motivation

- Quality assessments, **critical** for effective learning
- But, require a lot of human **time** and **effort** to craft
- Question generation system is **beneficial**
- These generated questions can serve:
 - as **extra practice** to students to clear up doubts, understand the content at a deeper level
 - as **new ideas** for educators when constructing quizzes, examinations



Motivation

- Recently, **large language models** (LLMs) have shown **impressive performance** improvements on a host of natural language tasks
- This **motivates** the attempt to use these models for
 - **generating questions in educational settings**
 - understanding **if they can actually help**



Objectives

Objectives

- **How to generate questions?**
 - Exploration of different ways to interact with LLMs
 - Techniques in **prompt engineering, retrieval augmented generation**
 - Understanding avenues to **improve model performance**
- **How to assess the generated questions?**
 - Identifying a **rubric** for judging question quality
 - Human evaluation studies
 - Different **stakeholders** in the setting

Related Work

The Past

Advancements in NLP and LLMs

- LLMs have been showing immense progress in natural language tasks
- Success can be attributed to their **massive scale**, both in **training data**, and **usage of resources**
- Challenges arise from their **size and proprietary nature** (Schick and Schütze, 2021)
- **New ways** to effectively utilize these models without direct modification: **Prompt-based Learning** (Liu et al., 2023; Gao et al., 2024)
- This approach has proven effective (Petroni et al., 2019; Brown et al., 2020; Radford et al., 2019)

The Past

LLMs in Generating Educational MCQs

- GPT-4 has been found **capable** of generating **high-quality multiple-choice questions** for various educational domains (Elkins et al., 2023; Bhat et al., 2022; Doughty et al., 2023)
- Questions have been **positively** evaluated, highlighting their **utility in creating assessments**
- Limitations still exist in generating questions for **advanced subjects** and **implementing context control mechanisms**

The Past

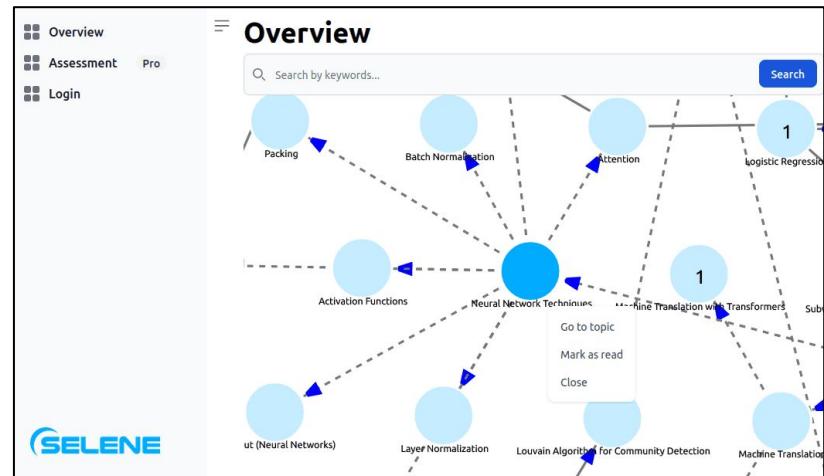
Crafting High-Quality MCQs

- Development of effective MCQs requires **deep subject knowledge, time, and expertise**
- Aspects can include **well-structured content, efficient distractors, and clear, relevant options** (Kar et al., 2015; Salam et al., 2020).
- Adhering to these practices in question writing is crucial for assessing **higher-order thinking skills** and ensuring the educational value of MCQs (Catanzano et al., 2022; Coughlin and Featherstone, 2017)
- **Sets a standard for LLM-generated questions** in educational assessments

The Future

SELENE, an innovative online platform

- Designed for **self-paced learning in AI**
 - Repository of **Jupyter notebooks** for ML/AI education
 - Web frontend that **visualizes learning paths**
- Partly inspired from MR Kabir, 2023



The Future

- Aims to function as a **large-scale interactive textbook** accessible globally
- Future plans to incorporate **self-assessments**, including **MCQs generated by LLMs**, to enhance learning efficiency and quality

Highlights an area for further development and research

Subword-based Tokenization

Overview

In the realm of natural language processing (NLP), the task of breaking down textual data into smaller units, known as tokens, forms the foundation of many sophisticated algorithms and models. Traditional tokenization methods often treat words as indivisible units, but this approach encounters challenges with morphologically rich languages, out-of-vocabulary words, and handling rare or unseen terms. To address these issues, subword-based tokenization has emerged as a powerful technique, revolutionizing the way neural networks process and understand language.

Subword-based tokenization involves dividing words into smaller, meaningful subunits or morphemes, thereby capturing the internal structure of words. This methodology not only enhances the model's ability to handle previously unseen or rare words but also improves generalization by recognizing similarities between related terms. By breaking words down into subword units, models become more robust and adaptable, effectively mitigating the issues of data sparsity and vocabulary mismatch that often plague conventional tokenization approaches.

The importance of subword-based tokenization in the context of deep learning cannot be overstated. Deep learning models, particularly those based on recurrent and transformer architectures, thrive on vast amounts of data for training. Subword tokenization enables these models to effectively leverage diverse linguistic patterns and structures, leading to superior performance across various NLP tasks. Moreover, it facilitates transfer learning and cross-lingual applications, as subword representations capture language-agnostic features and can be applied across different languages with minimal modifications.

In this era of ever-expanding linguistic diversity and complexity, subword-based tokenization stands as a cornerstone technology, enabling deep learning models to grasp the nuances of human language with unprecedented accuracy and efficiency. This introduction sets the stage for exploring the intricacies and significance of subword-based tokenization in the landscape of modern NLP and deep learning.

Navigation

Parent Topic

[Tokenization](#)

Notebooks

[Introduction to Subword-based Tokenization](#)
[Subword-based Tokenization: Approaches](#)

Subtopics

[Byte-Pair Encoding](#)
[WordPiece](#)
[Unigram \(Tokenization\)](#)

Recommended Background

[Natural Language Processing](#)
[Text Preprocessing](#)
[Regular Expressions](#)

Related Topics

[Character-based Tokenization](#)
[Word-based Tokenization](#)

Follow-up Topics

[BERT](#)
[Transformers](#)

Test your Knowledge

(probably some more information here, e.g., how many quizzes already done with some statistics regarding success rate)

[Start Quiz!](#)

Methods

Setup

Variables Controlled

Question Type

(Multiple Choice Questions)

Rationale

- **Constrained format** which **simplifies** automated **evaluation**
- **Assess knowledge** through plausible **distractors**
- LLMs have been **previously subjected** to this task

Subject Matter

(Natural Language Processing and Text Mining)

Rationale

- **Complexity** and **relevance** to **advanced undergraduate coursework**
- A lot of **bridging-the-gap knowledge** since these are broad areas
- Availability of **domain expertise** from NUS courses, **CS4248** and **CS5246**, led by **Dr. Christian Von Der Weth**

Naive Approach

Employed a baseline approach by directly prompting **ChatGPT** via its web interface to generate questions, termed as **Zero Shot Prompting**

Example prompt used:

Generate 10 MCQ questions about {topic}

where {topic} is a specific concept like
vector space model or **naive bayes classifier**

Naive Approach



You

Generate 10 MCQ questions about naive bayes classifier.



ChatGPT

1. What is the Naive Bayes classifier primarily used for?
 - A) Regression analysis
 - B) Clustering data points
 - C) Text classification
 - D) Image recognition
2. What assumption does the Naive Bayes classifier make about the features in a dataset?
 - A) They are highly correlated.
 - B) They are independent of each other.
 - C) They follow a Gaussian distribution.
 - D) They are all equally important.
3. In the context of the Naive Bayes classifier, what does the term "naive" refer to?
 - A) It refers to the simplicity of the algorithm.
 - B) It refers to the assumption of independence among features.
 - C) It indicates the use of Bayes' theorem.
 - D) It denotes a lack of prior knowledge about the dataset.
4. Which probability distribution is commonly used for continuous features in the Naive Bayes classifier?

Snippet of ChatGPT's Response for **Naive Bayes Classifier**

Naive Approach



You

Generate 10 MCQ questions about vector space model.



ChatGPT

1. Which of the following best describes the Vector Space Model (VSM)?
 - A) A technique used in linear algebra for solving systems of equations.
 - B) A mathematical model used in information retrieval to represent text documents.
 - C) A computational method for clustering data points in multidimensional space.
 - D) A statistical approach for analyzing categorical variables in datasets.
2. In the Vector Space Model, how are documents and terms represented?
 - A) Documents as vectors and terms as scalars.
 - B) Documents as scalars and terms as vectors.
 - C) Both documents and terms as vectors.
 - D) Both documents and terms as scalars.
3. Which of the following is a key concept in the Vector Space Model for representing text documents?
 - A) Singular Value Decomposition (SVD)
 - B) Term Frequency-Inverse Document Frequency (TF-IDF)
 - C) Principal Component Analysis (PCA)
 - D) Euclidean Distance
4. What does TF-IDF stand for in the context of the Vector Space Model?

Snippet of ChatGPT's Response for **Vector Space Model**

Observations

The generated questions were **generic**

- Focused on **factual recall** about broad topics such as **definitions** and **high-level applications**
- Lacked **specificity** and **alignment with learning outcomes**

Limitations

Limited Context Control

Inability to tailor questions to specific contexts or learning objectives

Absence of Learning Objectives

Questions did not align with or assess mastery of targeted concepts or skills

Prompt Engineered Approach

Improvements

System Prompt

- OpenAI cookbook
- **Lecturer** persona
- Reference textbook as a source of authority
- Expected **response format**

Topic Controlled User Prompt

- Taken from the textbook itself

Choice of Model

- **GPT-4** via the OpenAI API
- **1st on LMSys Leaderboard** (Chiang et al., 2024)

Explicit Instructions

Comprehensive Response

- Not only the question, but also the answer and explanation
- **Chain of Thought** prompting (Wei et al., 2023)

Prompt Engineered Approach

The “new” system prompt

You are a lecturer for an advanced undergraduate natural language processing course. Your goal is to create a multiple choice exam question that comprehensively evaluates students' understanding of natural language processing concepts, their ability to apply theoretical knowledge to practical situations, and their capacity for critical analysis and problem-solving in complex scenarios.

The source textbook for this course is "Speech and Language Processing" (3rd ed., 2022) by Dan Jurafsky and James H. Martin.

For each question, you should:

- Provide a detailed solution that explains the thought process, reasoning, and step-by-step approach required to arrive at the correct answer.
- The solution should demonstrate a deep understanding of the underlying concepts and their practical applications.

The question itself should meet the following criteria:

- Be a multiple choice question (MCQ) with 5 choices in markdown format:
 1. Choice 1
 2. Choice 2
 3. Choice 3
 4. Choice 4
 5. Choice 5
- Incorporate both theoretical concepts and practical applications of natural language processing topics covered in the course.
- Require a unique synthesis of ideas from multiple topics, concepts, and sources, going beyond questions commonly found in standard textbooks.
- Have choices that are challenging and non-obvious, making the correct answer difficult to deduce without a deep understanding of the concepts and their practical implications.
- Your output should only be in markdown format, with the following headers:

Question

Solution

Correct Answer

Reasoning

- Inline equations should use the markdown format: $a = b + c$
- Block equations should use the markdown format:
$$a = b + c$$

Prompt Engineered Approach

The “new” user prompt

Create a multiple choice question (MCQ) and solution
that covers one or more of the following topics:
{topics}

where **{topics}** is taken directly from the contents of the textbook; for example

- Logistic Regression
- Classification with Logistic Regression
- Multinomial logistic regression
- The cross-entropy loss function ...

Observations

Empirically, it was observed that

- **Quality improved** dramatically
- Responses adhered to **specified format**
- Only **1 in 1001** samples generated had a **gibberish response**
- No syntax errors* (Prone to Sampling Bias)

Limitations

Reliance on Internal Knowledge

Potentially outdated, false, cannot be relied upon

Topics cannot be too specific

Arbitrary scales to control question difficulty not effective

Observations

Example Responses

Question

Given a trigram language model (words are denoted as w and their position in a sentence as subscripts), we aim to estimate the probability of the word "algorithm" given its context. The context in this scenario is the bigram "a novel". This model employs Kneser-Ney smoothing, which is known for its effectiveness in handling unseen words or N-grams, making it particularly adept at dealing with sparse data situations. Assuming the following data is available:

- The count of the bigram "a novel" is 20.
- The count of the trigram "a novel algorithm" is 3.
- The total number of unique bigrams in the dataset is 5000.
- The continuation count, count_c , of "algorithm" (i.e., the number of unique bigrams that "algorithm" completes) is 150.
- The discount value, d , applied in Kneser-Ney smoothing for adjusting counts is 0.75.

Calculate the probability, $P(w_3 = \text{"algorithm"} | w_1 = \text{"a"}, w_2 = \text{"novel"})$, for the given trigram using the Kneser-Ney smoothing formula. Select the closest option:

1. 0.015
2. 0.025
3. 0.030
4. 0.050
5. 0.075

Question

Given a dataset for a binary classification problem, you've decided to use logistic regression with regularization to prevent overfitting. Your logistic regression model uses a sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ for predictions, where $z = w_0 + w_1x_1 + w_2x_2 + \dots + w_nx_n$ and L2 regularization for its cost function. The regularized cost function $J(w)$ for logistic regression is given as:

$$J(w) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(\sigma(z^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2$$

where m is the number of training examples, n is the number of features, $y^{(i)}$ is the actual label for the i -th training example, w are the parameters (weights), λ is the regularization parameter, and $\sigma(z^{(i)})$ is the prediction for the i -th training example.

To train this logistic regression model using gradient descent, you will need to compute the gradient of $J(w)$ with respect to each weight w_j . Which of the following correctly represents the derivative of $J(w)$ with respect to w_j (for $j \geq 1$)?

1. $\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^m (\sigma(z^{(i)}) - y^{(i)}) x_j^{(i)}$
2. $\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^m (\sigma(z^{(i)}) - y^{(i)}) x_j^{(i)} + \frac{\lambda}{m} w_j$
3. $\frac{\partial J}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m [y^{(i)}(1 - \sigma(z^{(i)})) - (1 - y^{(i)})\sigma(z^{(i)})] x_j^{(i)}$
4. $\frac{\partial J}{\partial w_j} = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(\sigma(z^{(i)})) - (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))] x_j^{(i)} + \frac{\lambda}{2m} w_j^2$
5. $\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^m (\sigma(z^{(i)}) - y^{(i)}) + \frac{\lambda}{2m} w_j^2$

Observations

Gibberish Response

I the: [...]

∴ f. 1. the Abortion Introduction

Our restaurant's and introduce new menu Changes.

Question

A 2. Solution

B Engineers applying Natural Language Processing Programming implementations the with

When long 3. Monkeys' document the at developers Large

The Heart 4. Connection

5.

C Meta-Reflections

D

E ANSWERS:

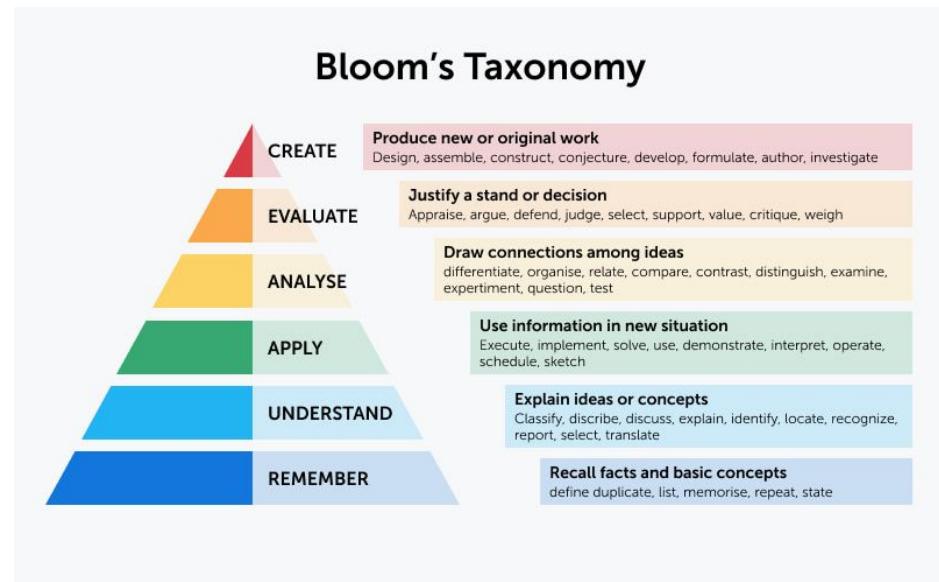
The defouad the on Secular ArtS

In the bacterial Correct Reaction

F Rationale

Using Learning Objectives

- Focus on aligning question generation with learning outcomes using **Revised Bloom's Taxonomy** (LW et al., 2001; Bloom et al., 1956)
- Taxonomy levels: **Remember, Understand, Apply, Analyze, Evaluate, Create**
- Tailored questions to specific **cognitive processing levels** to match learning objectives



Using Learning Objectives

Create a multiple choice question (MCQ) and solution that covers one or more of the following topics:
{topics}

The question should have the learning objective
{taxonomy} from the Revised Bloom's Taxonomy.

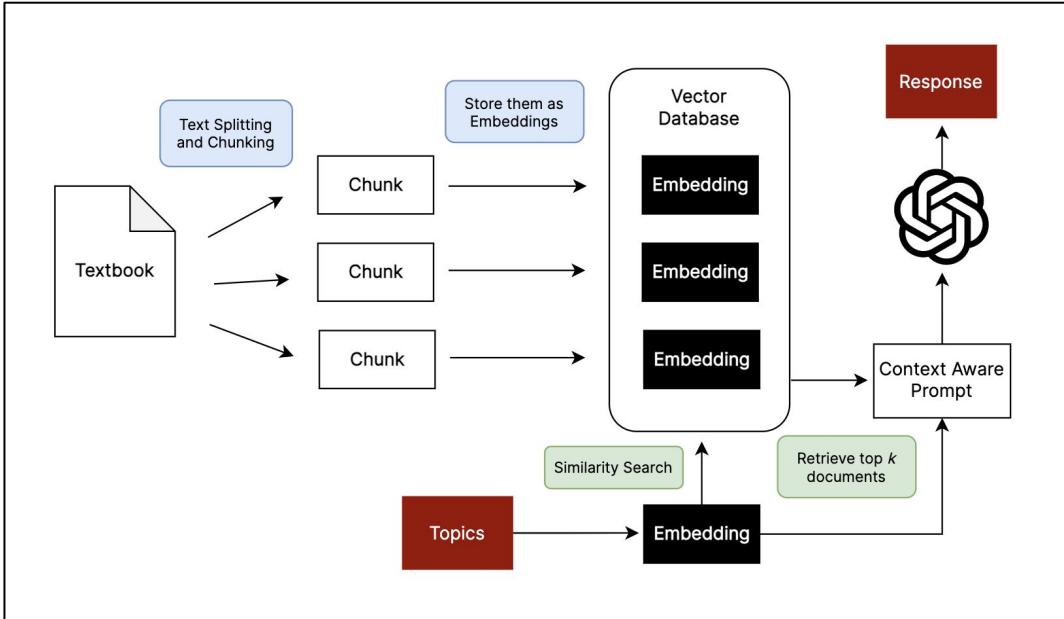
where {taxonomy} is one of Remember, Understand, Apply, Analyze, Evaluate, or Create

On Evaluation

- Difficult to perform automated qualitative analysis
- Human evaluation necessary

RAG Vanilla Approach

As a context control mechanism



- **Directly parse Speech and Language Processing textbook**
- **Store the embeddings in a vector database**
- Semantically query topics to retrieve **top k similar documents**
- **Context for the prompt**

Vanilla Approach Flow

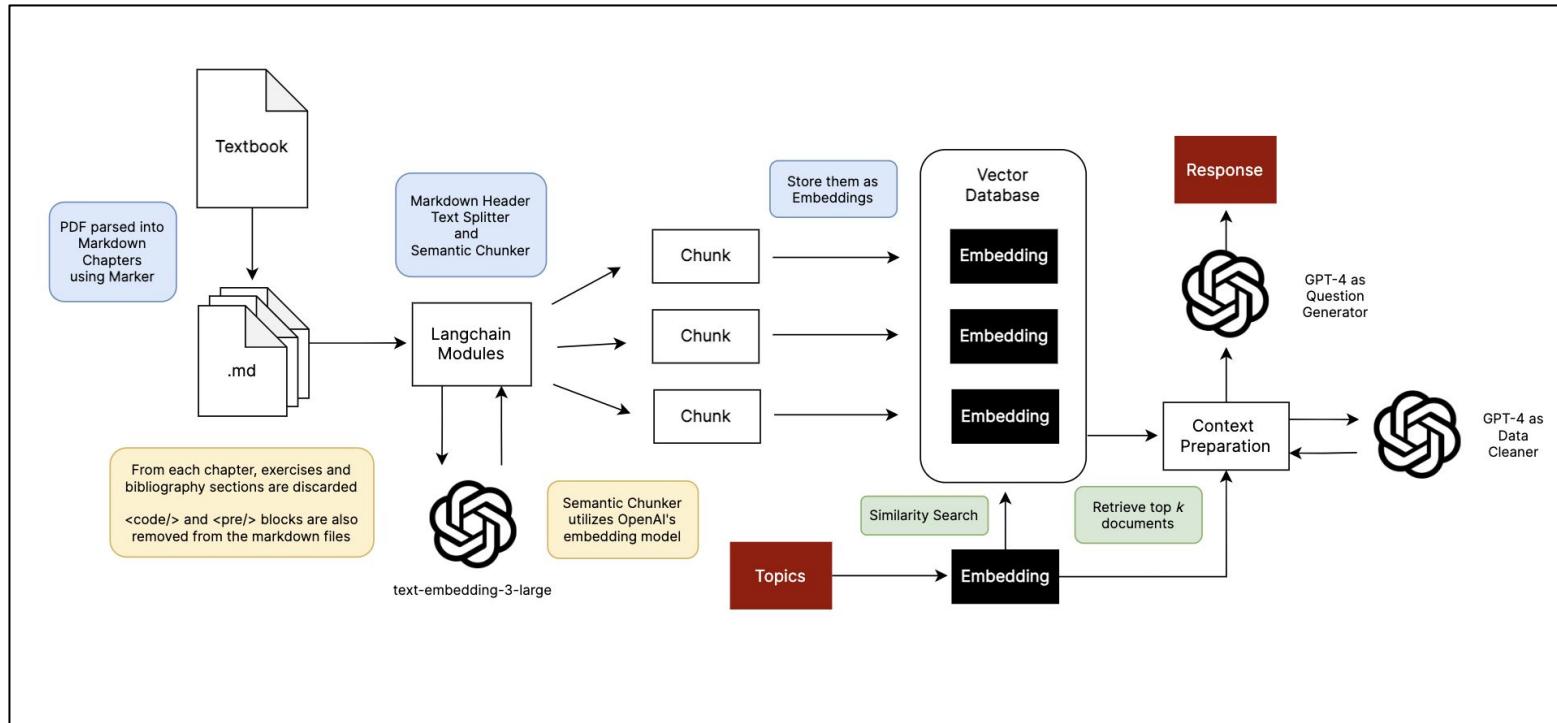
RAG Vanilla Approach

Empirically, it was observed that

- **Erroneous PDF Parsing**
 - PDFs are notoriously **hard to parse**
 - Elements like **code, tables, mathematical equations** are **confusing**
- **Poor Chunking Strategy**
 - Direct dumping leads to **material being cut-off** at arbitrary positions
- **Lack of Grammatical Coherence**
 - Contexts **may not make grammatical sense**
 - Risk of passing **incorrect knowledge** to LLM

A better architecture was needed to address these issues

RAG Advanced Approach



Advanced Approach Flow

RAG Advanced Approach

Improvements

Parsing Strategy

- Usage of an **OCR supported parser, Marker**
- Convert **individual chapters** to **markdown files**
- Remove **irrelevant information** at the end of each chapter, like exercises

Chunking Strategy

- Separate markdown content into **paragraphs**
- Combination of **MarkdownHeaderSplitter** and **SemanticChunker** based on **OpenAI embedding model**
- **Reference** headers, sections from **textbook**

Context Preparation

- Usage of GPT-4 as a **context cleaner**
- **GPT-4 effective at understanding unstructured data** (Bolding et al., 2023)

RAG Approach User Prompt (General)

Create a multiple choice question (MCQ) and solution that covers one or more of the following topics:
{topics}

Utilize the relevant textbook content:
{context}

where **{topics}** is taken from the textbook and **{context}** is the content retrieved after querying the vector database

RAG Advanced Approach

Empirically, it was observed that

- Type of content retrieved from textbook matters a lot;
still prone to topic specificity
- Query to vector database incorporated content from
different chapters
- Responses **seemed to be more grounded***, with references to the textbook content used in explanations

On Evaluation

- **Difficult to perform automated qualitative analysis**
- **Human evaluation necessary**

RAG Advanced Approach

Example Response

Question

In the context of developing a neural language model for a text completion application, you are deciding between using a feedforward neural language model and a transformer-based language model. Given the requirements for high accuracy in word prediction, the ability to handle long contexts, and a preference for models that can generalize well over various contexts, which model would be more suitable for your application?

1. Feedforward neural language model, because it is less complex and faster to train compared to transformer-based models.
2. Transformer-based language model, due to its ability to handle much longer histories and generalize better over contexts of similar words.
3. Feedforward neural language model, as it is more interpretable than transformer-based models, making it easier to debug and refine.
4. Transformer-based language model, because it is simpler and requires less energy to train than feedforward neural language models.
5. Feedforward neural language model, because it uses embeddings that allow it to generalize better to unseen data compared to transformer-based models.

Solution

The correct answer is 2. Transformer-based language model, due to its ability to handle much longer histories and generalize better over contexts of similar words.

Correct Answer

2. Transformer-based language model, due to its ability to handle much longer histories and generalize better over contexts of similar words.

Reasoning

The decision between using a feedforward neural language model and a transformer-based language model for a text completion application hinges on several factors: accuracy in word prediction, the ability to handle long contexts, and generalization capabilities over various contexts.

From Context 1, we learn that neural language models, including feedforward neural language models, have advantages over n-gram models, such as handling longer histories and generalizing better over contexts. However, compared to transformer-based models, feedforward neural language models have limitations in their complexity, speed, and energy requirements for training. They approximate the probability of a word given the entire prior context based on a limited number of previous words (for example, using a 4-gram model), and while they use embeddings to generalize to unseen data, they don't inherently handle very long contexts as well as transformer-based models.

From Context 2, transformer-based language models are described as having a wide context window, allowing them to draw on much longer histories (up to 4096 tokens) to predict upcoming words. This characteristic makes them particularly suitable for applications requiring high accuracy in word prediction over long contexts. Furthermore, transformer-based models are noted for their ability to generalize well over various contexts, which is crucial for applications like text completion where diverse input text can be encountered.

Therefore, given the requirements for handling long contexts and generalizing over various contexts, along with the goal of achieving high accuracy in word prediction, the transformer-based language model is the more suitable choice. This is directly supported by the information provided in Context 2 about the capabilities of transformer-based models, making option 2 the correct answer.

RAG Advanced Approach

Textbook v/s Provided Context

10.11 Summary

This chapter has introduced the transformer, and how it can be applied to build large language models. Here's a summary of the main points that we covered:

BIBLIOGRAPHICAL AND HISTORICAL NOTES 241

- Transformers are non-recurrent networks based on **self-attention**. A self-attention layer maps input sequences to output sequences of the same length, using attention heads that model how the surrounding words are relevant for the processing of the current word.
- A transformer block consists of a single attention layer followed by a feed-forward layer with residual connections and layer normalizations following each. Transformer blocks can be stacked to make deeper and more powerful networks.
- Language models can be built out of stacks of transformer blocks, with a linear and softmax max layer at the top.
- Transformer-based language models have a wide context window (as wide as 4096 tokens for current models) allowing them to draw on enormous amounts of context to predict upcoming words.
- Many NLP tasks—such as question answering, summarization, sentiment, and machine translation—can be cast as tasks of word prediction and hence addressed with Large language models.
- The choice of which word to generate in large language models is generally done by using a **sampling** algorithm.
- Because of their ability to be used in so many ways, language models also have the potential to cause harms. Some harms include hallucinations, bias, stereotypes, misinformation and propaganda, and violations of privacy and copyright.

Provided Context

Context 1

This chapter has introduced the transformer and its application in constructing large language models. Here's a summary of the key points discussed:

- Transformers are non-recurrent networks that utilize self-attention. A self-attention layer transforms input sequences into output sequences of identical length, employing attention heads to model the relevance of surrounding words for processing the current word.
- A transformer block is composed of a single attention layer followed by a feedforward layer, with residual connections and layer normalizations after each.

References: Header 2 10.11 Summary

Context 2

Transformer blocks can be stacked to create deeper and more robust networks. Language models can be developed from stacks of transformer blocks, topped with a linear and softmax layer. Transformer-based language models benefit from a wide context window, capable of handling up to 4096 tokens, which allows them to leverage vast amounts of context for predicting forthcoming words. Many NLP tasks, such as question answering, summarization, sentiment analysis, and machine translation, can be formulated as word prediction challenges, making them suitable for large language models. The selection of words in large language models is typically managed through a sampling algorithm.

References: Header 2 10.11 Summary

Evaluations

Evaluations

- Multiple stakeholders in a **classroom setting**
- Understanding their **perspectives** is important to **establish foundation**
- Identified **target groups**
 - **Students**, and **end-users** who are the main learners
 - **Teaching assistants**, who are the **grassroot communicators** from the teaching team; often the closest in learning journey
 - **Lecturers/Educators**
- Hence, **three different evaluations** done to get a broader view

User Study

Objective: How receptive are the questions generated by LLMs in learning? How are questions rated on several metrics by the end-users?

Setup

Two canvas quizzes were created, one with LLM-generated questions, one with feedback survey

Questions were from different topics from the first half of the semester

Each question given an extra option to flag in case of it being “invalid, unclear, flawed, or not answerable”

Students were given the disclaimer of questions being LLM-generated in the announcement and brief

Participants

Students from the Text Mining course at NUS, CS5246 this semester

In total, 10 students participated in the study

User Study

Feedback survey consisted of **14 Likert Scale questions**

Students were asked to judge both **questions** and **answers + explanations** that were generated by the LLM

Strongly disagree	Disagree	Neutral	Agree	Strongly Agree
Example Likert Scale Statement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Statement	Tag
The LLM-generated questions were clear and easy to understand.	clarity
The difficulty level of the LLM-generated questions was appropriate for the course level.	difficulty
The LLM-generated questions effectively assessed my understanding of the course material.	coverage
The LLM-generated questions challenged me to think critically and apply concepts.	challenge
The assessment format was conducive to demonstrating my knowledge.	usefulness
I found the LLM-generated questions to be engaging and interesting.	engagement
I felt motivated to do well on the assessment with LLM-generated questions.	motivation
Overall, I had a positive experience with the LLM-generated questions.	sentiment

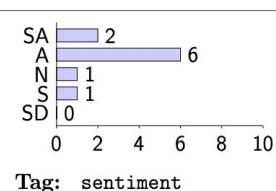
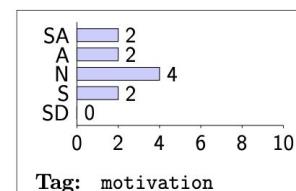
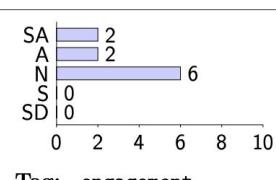
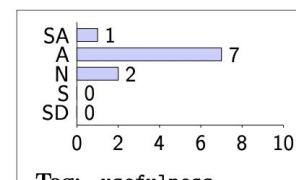
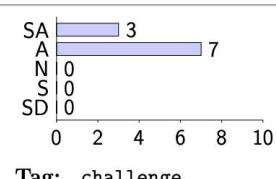
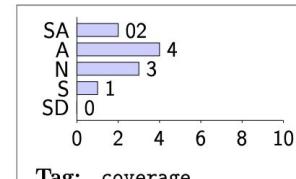
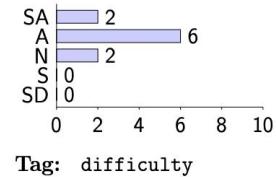
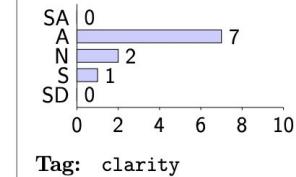
Statement	Tag
The LLM-generated answers and explanations were clear and easy to understand.	clarity_a
The difficulty level of the LLM-generated answers and explanations was appropriate for the course level.	difficulty_a
The LLM-generated answers and explanations effectively assessed my understanding of the course material.	coverage_a
The LLM-generated answers and explanations challenged me to think critically and apply concepts.	challenge_a
I found the LLM-generated answers and explanations to be engaging and interesting.	engagement_a
Overall, I had a positive experience with the LLM-generated answers and explanations.	sentiment_a

Results

Overall, the perception was positive

Most scales had **majority of the students agreeing** with the statements

For the Questions



LEGEND

SA	Strongly Agree
A	Agree
N	Neutral
D	Disagree
SD	Strongly Disagree

Results

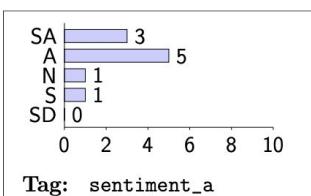
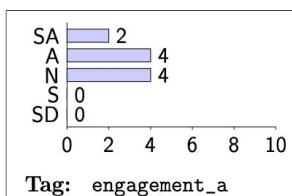
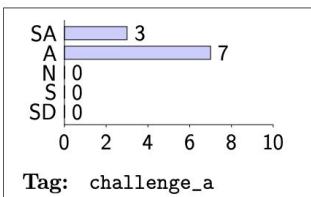
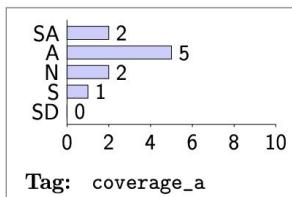
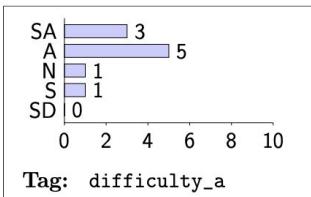
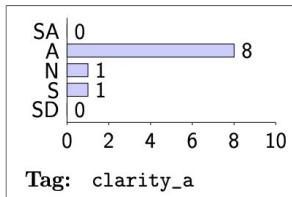
Qualitative feedback from the students included

“Inclusion of more examples”

“Clarification of definitions used by the question”

“All questions sounded pretty similar”

For the Answers + Explanations



Taxonomy Study

Objective: Understanding the effect of mentioning learning objectives in prompts to help educators craft questions using LLMs

Setup

18 questions were generated across **2 subjects**, NLP and Text Mining, and **6 learning objectives** from **Revised Bloom's Taxonomy**

For each generated question, the participant had to respond to multiple feedback questions

Participants

The subject of this study was **Dr. Christian** himself, the supervisor of this project.

He **teaches both NLP and Text Mining courses** at NUS.

Taxonomy Study

Actual Responses can be found here:

- [CS4248](#)
- [CS5246](#)

Generated Questions

Description	Guide
Is the question invalid, unclear, flawed, or not answerable?	Yes/No
What is the most likely learning objective that is met by this question?	Create Evaluate Analyze Apply Understand Remember
What is another likely learning objective that is met by this question?	Create Evaluate Analyze Apply Understand Remember
What is the most likely difficulty level of the question?	High Medium Low
Where do you think this question belongs to?	Nowhere, the question still needs human effort to correct Non-graded assessments, tutorial discussions Graded assessments, exam papers
Additional Feedback	Textual Response

Results

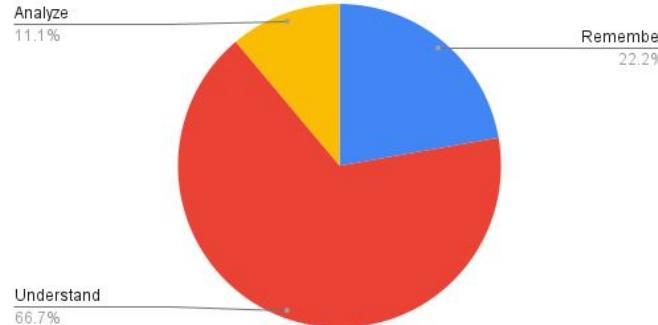
With respect to the study conducted, it was found that **indicating the learning objective is not effective**

Identified learning objective **not the same as prompted** learning objective in most cases

Limitation of the taxonomy:
Interpretations of the definitions can be subjective

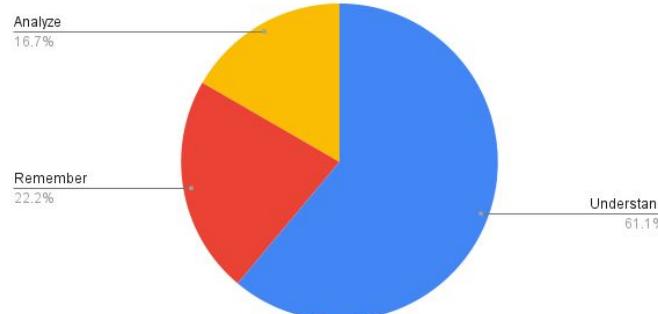
NLP

Count of What is the most likely learning objective that is met by this question?



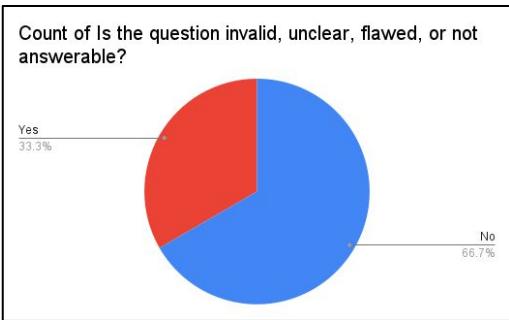
Text Mining

Count of What is the most likely learning objective that is met by this question?

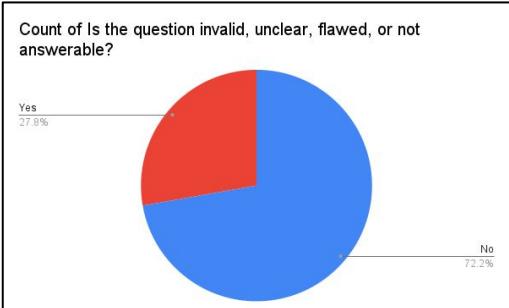
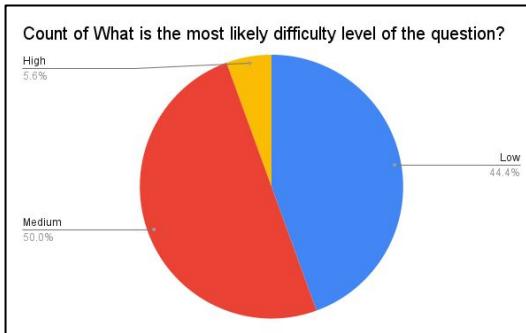


Results

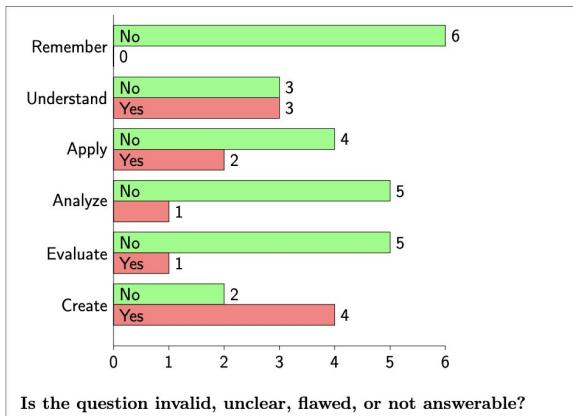
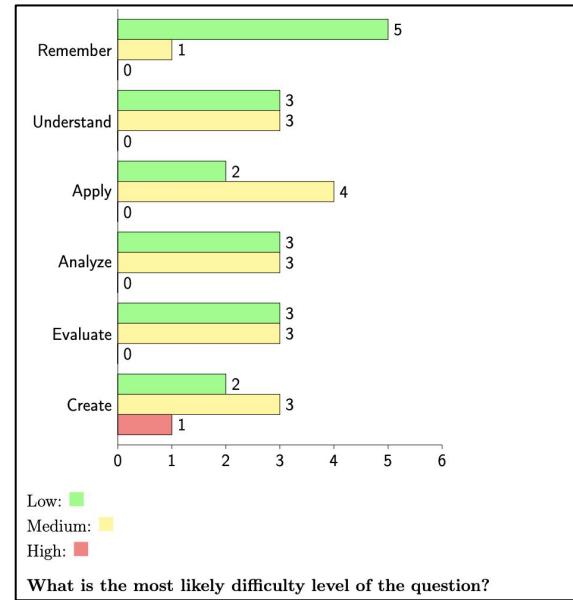
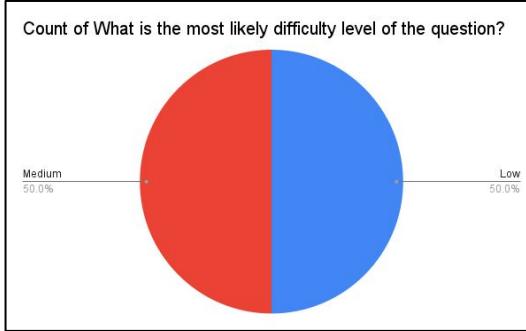
- Difficulty Level
- Clarity of the questions



NLP



Text Mining

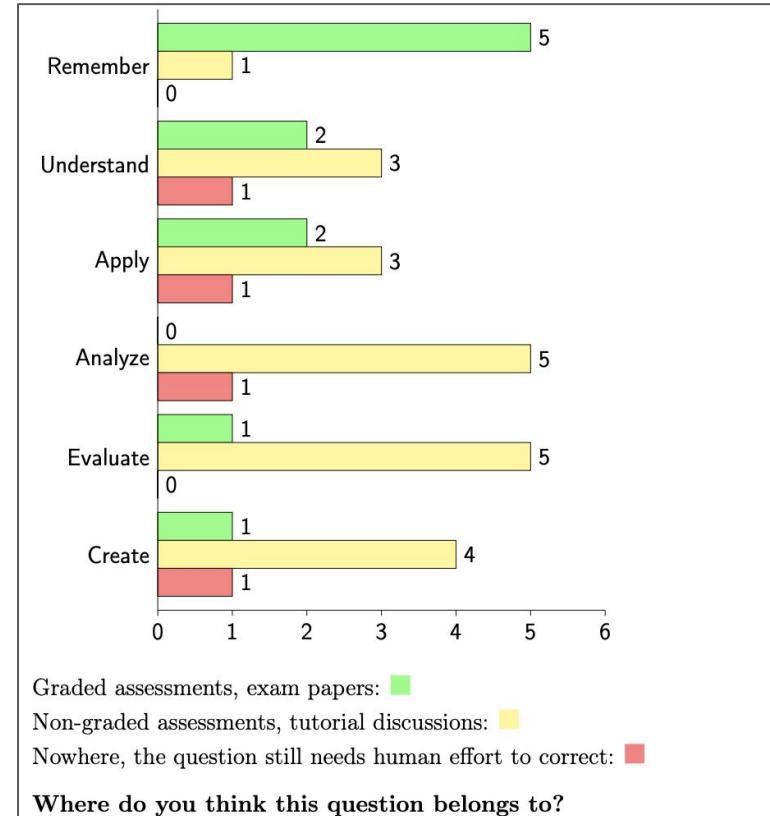


Results

Dr. Christian mentioned that he **wouldn't use** any question "**as is**" for a **graded examination** apart from very few that are sure shot correct and clear

Many to most questions would be **good for tutorials or practice quizzes** to engage **discussion and critical thinking**

Most questions were certainly **good for ideation** which can already **help a lot when making quizzes**



Results

Count of Where do you think this question belongs to?

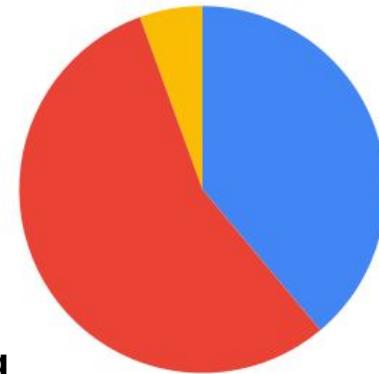
- Non-graded assessments, tutorial discussions
- Nowhere, the question still needs human effort to correct
- Graded assessments, exam papers



NLP

Count of Where do you think this question belongs to?

- Graded assessments, exam papers
- Non-graded assessments, tutorial discussions
- Nowhere, the question still needs human effort to correct



Text Mining

Expert Study

Objective: Understanding the quality of generated questions across different approaches

Setup

21 questions were generated using 3 approaches:
Prompt Engineered **Gemma 7B**, **GPT-4**, and
Advanced RAG

The questions varied widely across **NLP topics** and
were **chosen randomly from a pool** of existing
generated questions

3 different scorecards were given to participants to
provide feedback

Participants

The participants for this study were **Ph.D. students** from NUS who had a **strong background in Natural Language Processing**

2 Ph.D. students took part in this study

Expert Study

- The Question Scorecard

Description	Guide
How answerable is the question? (Answerability) Inspired from Nema and Khapra, 2018	1 - All important information is missing and it is impossible to answer the question 2 - Most of the important information is missing and I can't infer the answer to the question 3 - Some important information is missing leading to multiple answers 4 - Most of the important information is present and I can infer the answer 5 - All important information is present and I can answer the question
What's your choice?	One of the options from 1 - 5
How would you rate the difficulty of the question?	1 - Very easy 2 - Easy 3 - Medium 4 - Difficult 5 - Very difficult

- Answer + Explanation Scorecard

Description	Guide
Do you think the given correct answer is correct?	Yes/No
Do you think the given explanation is correct?	Yes/No

- Overall Response Scorecard

Description	Guide
How would you rate the overall quality of the question, given answer, and options combined?	1 - Poor quality overall 2 - Below average quality 3 - Average quality 4 - High quality 5 - Excellent quality
Where do you think this question belongs to?	Nowhere, the question still needs human effort to correct Non-graded assessments, tutorial discussions Graded assessments, exam papers
Additional Feedback	Textual Response

Expert Study

Actual Responses can be found here:

- [Participant 1](#)
- [Participant 2](#)

[Template Evaluation Sheet](#)

[Participant Guide](#)

[Generated Questions + Answers](#)

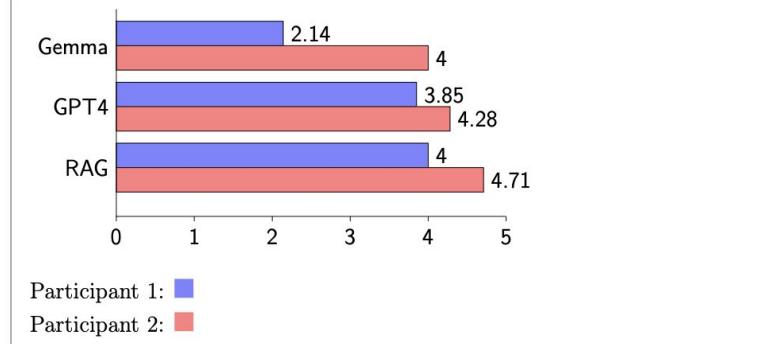
Rubrics

- **Answerability** (Nema and Khapra, 2018)
- **Overall Quality**
- Given **answer** being correct
- Given **explanation** being correct

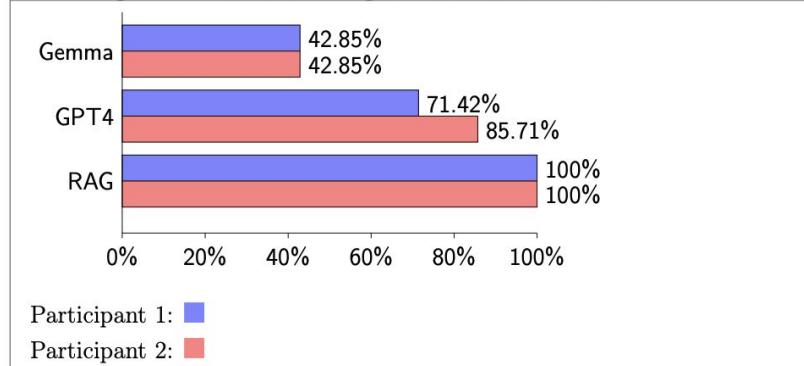
Last three metrics inspired from Olney, 2023

Results

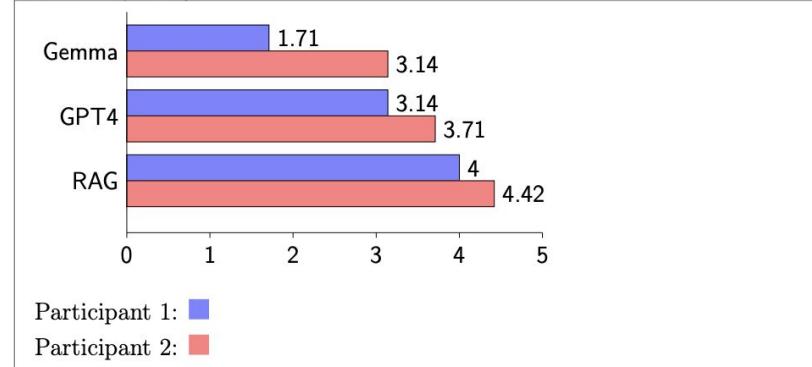
Answerability



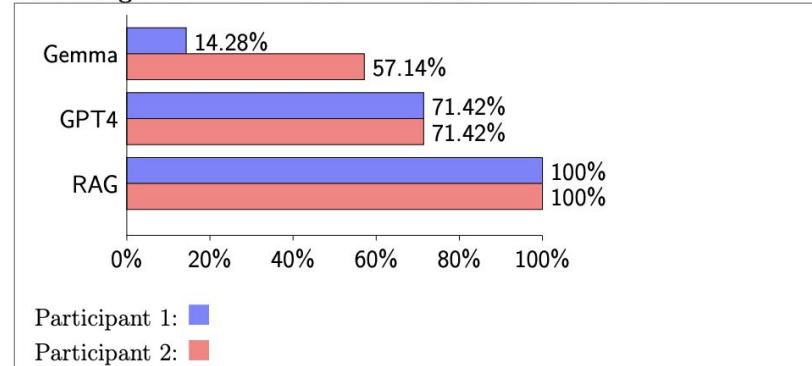
Percentage of times Given Explanation is Correct?



Overall Quality



Percentage of times Given Correct Answer is Correct?



Results

Overall, **Gemma** showed **poor performance**. This motivated the need to consider a methodology to try **fine tuning**. A **proof of concept** was done.

GPT4 generated **good content** but only at the **surface level**. The participants pointed out that when they read the content, it **seemed repetitive**.

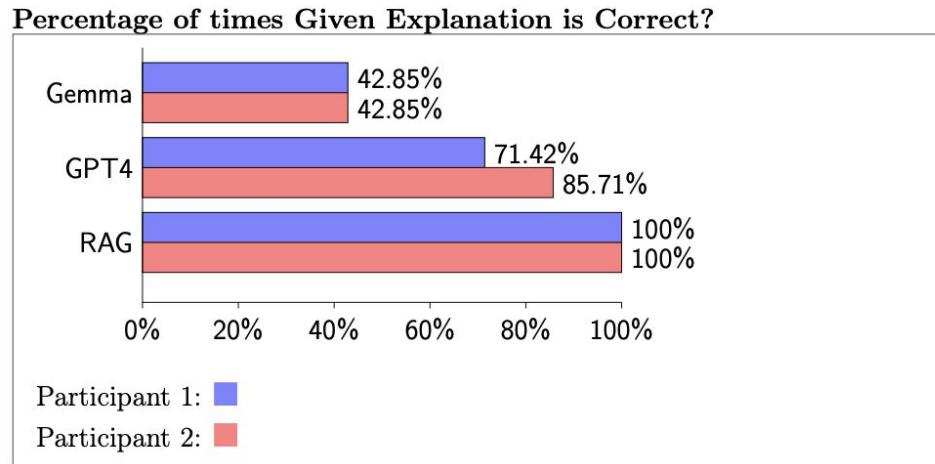
RAG certainly helped in **grounding the truth**, being the best performing method overall. The **context retrieved from the textbook** allowed for **better explanations** by the LLM. Some of the **most favored questions** in the entire study were generated using the RAG pipeline.

RAG method were the **highest rated**, (100% by Participant 1, and 71.4% by Participant 2) to be **used in graded assessment** settings.

Results

RAG certainly helped in grounding the truth, being the best performing method overall. The context retrieved from the textbook allowed for better explanations by the LLM. Some of the most favored questions in the entire study were generated using the RAG pipeline.

RAG method were the highest rated, (100% by Participant 1, and 71.4% by Participant 2) to be used in graded assessment settings.



Conclusions

Conclusions

Large language models (LLMs) like GPT-4 **show promise** in generating educational questions **but have limitations**

Generated questions are **not directly suitable for exams** but are **valuable for ideation, practice quizzes, and tutorials**

Human oversight is crucial to ensure the **accuracy and appropriateness**; hence it **should be made clear** that the questions are AI-generated

With further development and quality assurance, LLMs could significantly enhance learning experiences

Ethical Considerations

- Transparency and Explainability
- Bias and Fairness
- Privacy and Data Protection
- Accountability and Oversight

Future Work

Fine Tuning Experiment

GPT-4 is the most likely **go-to model** based on **empirical performance**

Ranked **1st** on **LMSys Leaderboard** (Chiang et al., 2024)

But, it is **closed-source**

**Not suitable for custom modifications, training
on custom datasets**

Fine Tuning Experiment

On Open Source Models

- Open source models are available, **but performance is questionable**
- Top LMSys rankings are taken by proprietary closed-source models
- Gemma-7B (Team et al., 2024) model chosen on the basis of manual observations

Fine Tuning Experiment

Can Fine Tuning help?

- Models usually trained as **generalists**
- **Pre-Training then Fine Tuning** is a **successful paradigm** in natural language tasks
- Inspiration from **Alpaca from Stanford** (Taori et al., 2023) and **Minigpt-4** (Zhu et al., 2023)
- **Synthetic data generated from a superior model can improve performance**

Fine Tuning Experiment

Setup

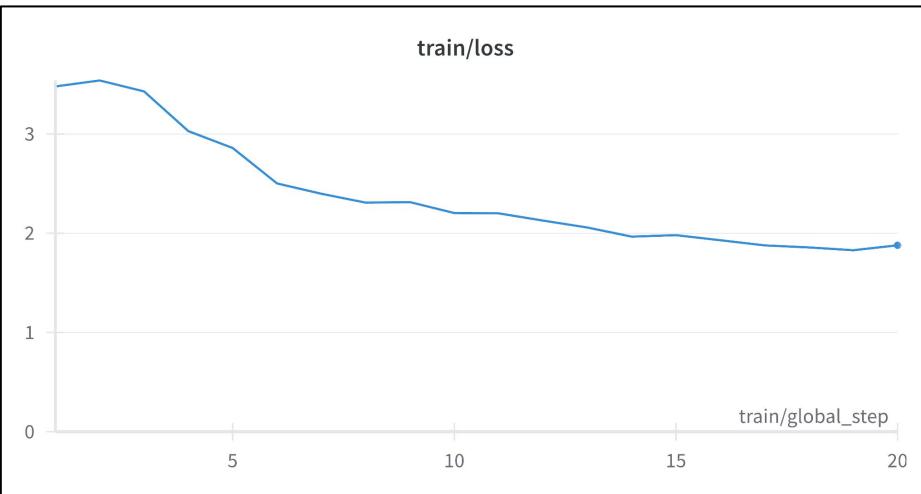
Using GPT-4, a dataset of **1000 LLM generated questions** was created. These **samples were cleaned** to ensure there are no syntax errors. The dataset was split **80/20 for training and validation**.

The fine tuning methodology used was **Quantized-LoRa**, which belongs to the family of fine tuning techniques called **PEFT (Parameter Efficient Fine Tuning)**.

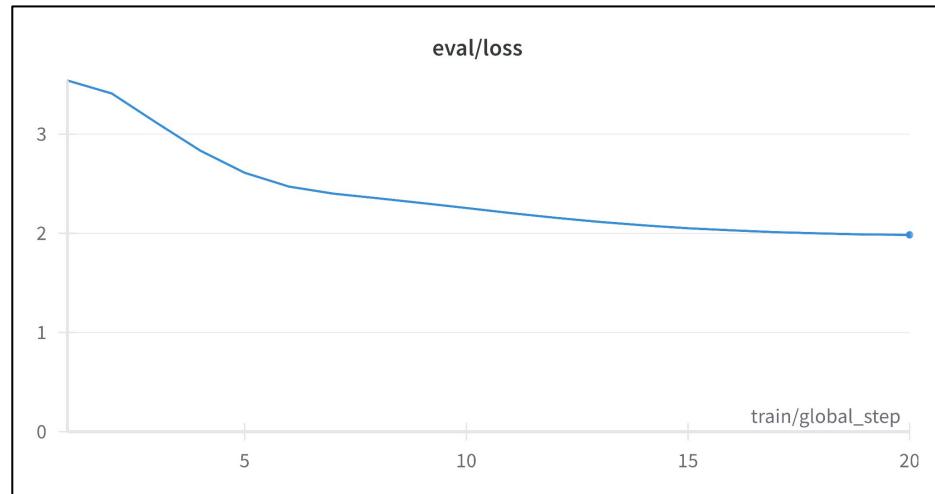
The model chosen for fine tuning was the **Gemma 7B Instruct model**, with **4-bit quantization**, the same one which was used for **zero-shot generation**.

Observations

Loss Graphs



Training Loss



Validation Loss

The losses seemed to **fall** fairly across **1 epoch** distributed over **20 steps**

But it **seemed to flat-line near 2**, which was **still very high**

Observations

Concluding statements

Empirically observed, **post fine tuning generation showed improvements** only in **correct formatting** of the generated questions

However, **more work needs to be done** in this area to make concrete claims

This experiment is a simple **proof-of-concept** that fine tuning might work in order to create a specialized model for question generation

Practical limitations, like **cost, resources required to generate large scale datasets**, that limited further exploration

Thank you for listening!

Q&A

References

- Bhat, S., Nguyen, H. A., Moore, S., Stamper, J. C., Sakr, M., & Nyberg, E. (2022). Towards automated generation and evaluation of questions in educational domains. *Educational Data Mining*. <https://api.semanticscholar.org/CorpusID:251180837>
- Bloom, B. S., Engelhart, M. B., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: the classification of educational goals. handbook 1: Cognitive domain*. Longmans Green.
- Bolding, Q., Liao, B., Denis, B. J., Luo, J., & Monz, C. (2023). Ask language model to clean your noisy translation data.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners.
- Catanzano, T. M., Jordan, S. G., & Lewis, P. (2022). Great question! the art and science of crafting high-quality multiple-choice questions. *Journal of the American College of Radiology : JACR*. <https://api.semanticscholar.org/CorpusID:247429556>
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot arena: An open platform for evaluating llms by human preference.
- Coughlin, P. A., & Featherstone, C. R. (2017). How to write a high quality multiple choice question (mcq): A guide for clinicians. *European journal of vascular and endovascular surgery : the official journal of the European Society for Vascular Surgery*, 54 (5), 654–658. <https://api.semanticscholar.org/CorpusID:29784206>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Doughty, J. A., Wan, Z., Bompelli, A., Qayum, J., Wang, T., Zhang, J., Zheng, Y., Doyle, A., Sridhar, P., Agarwal, A., Bogart, C., Keylor, E., Kultur, C., Savelka, J., & Sakr, M. F. (2023). A comparative study of ai-generated (gpt-4) and human-crafted mcqs in programming education. *Proceedings of the 26th Australasian Computing Education Conference*. <https://api.semanticscholar.org/CorpusID:265709732>
- Elkins, S., Kochmar, E., Cheung, J. C. K., & Serban, I. (2023). How useful are educational questions generated by large language models? *International Conference on Artificial Intelligence in Education*. <https://api.semanticscholar.org/CorpusID:258108262>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.
- Kar, S. S., Lakshminarayanan, S., & Mahalakshmy, T. (2015). Basic principles of constructing multiple choice questions. *Indian Journal of Community and Family Medicine*, 1, 65–69. <https://api.semanticscholar.org/CorpusID:86760184>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.
- Liu, Z., Yu, X., Fang, Y., & Zhang, X. (2023). Graphprompt: Unifying pre-training and downstream tasks for graph neural networks.
- LW, A., DR, K., PW, A., KA, C., Mayer, R., PR, P., Raths, J., & MC, W. (2001, January). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*.
- Nema, P., & Khapra, M. M. (2018). Towards a better metric for evaluating question generation systems.
- Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Müller, A. H., & Riedel, S. (2019). Language models as knowledge bases?
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. <https://api.semanticscholar.org/CorpusID:160025533>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.
- Salam, A., Yousuf, R., & Bakar, S. M. A. (2020). Multiple choice questions in medical education: How to construct high quality questions. <https://api.semanticscholar.org/CorpusID:213149071>
- Schick, T., & Schütze, H. (2021). It's not just size that matters: Small language models are also few-shot learners.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Stanford alpaca: An instruction-following llama model.
- Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Huszenot, L., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Hélio, A., ... Kenealy, K. (2024). Gemma: Open models based on gemini research and technology.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models.
- Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). Minigpt-4: Enhancing vision-language understanding with advanced large language models.