Journal of High School Science

Advanced Question Answer Generation (QAG) for mastering vocabulary proficiency using Generative AI.

Ramamoorthy P

Submitted: February 18, 2025, Revised: version 1, April 10, 2025, version 2, April 20, 2025
Accepted: April 29, 2025

Abstract

Question Answer Generation (QAG) is an essential task in Natural Language Processing (NLP) and Artificial Intelligence (AI). QAG has widespread applications, particularly in the education domain such as, creating content for educational tools and tutoring systems, as well as generating questions for the Scholastic Assessment Test (SAT). Currently, creating such content is largely manual, time-intensive, and is reliant on human expertise. This research focuses on evaluating Meta's Llama3 for generating Words in Context questions which is one of the categories in the SAT's reading and writing section under the Craft and Structure section. The research demonstrates how the Llama 3 8B model can be leveraged to generate context-relevant SAT questions based on a given genre and word combination. Specifically, In-Context Learning (Prompt Engineering) techniques along with Semantic Similarity (SS) embedding models are used as part of the experiments. Baseline test against the model resulted in a low accuracy of 47%. After experimenting with different prompt engineering techniques and dataset fine tuning using SS, Retrieval Augmented Generation (RAG) yielded better results with 74% accuracy. This research shows that tutors can leverage RAG + SS technique to generate Word in Context questions in an automated fashion and thereby reducing the manual effort in the generation part. Evaluation requires human assessment but that is aided by automated reasoning. Codebase that resulted from this research can be used to generate such questions for any genre and word combination, and delivered through a mobile/web-based application to help students master SAT vocabulary. The approach used allows for genre input from various categories, such as science and history; among others; to generate creative content. This research paves the way for leveraging Gen AI for other Reading/Writing SAT sections.

Keywords

Generative AI, Machine learning, Natural Language Processing, Deep learning, Automated Question Generation, Llama 3, Scholastic Assessment Test, Semantic similarity, Retrieval Augmented Generation, Prompt engineering

Pragyan Ramamoorthy. Dougherty Valley High School, 709 Jadecrest Court, San Ramon, California, 94582, USA.
pragyan0314@gmail.com

https://doi.org/10.64336/001c.137480

1. Introduction

Vocabulary skills are essential in order to obtain a high score in the Scholastic Assessment Test (SAT) Reading and Writing section of the test. Leveraging Gen AI for generating SAT style questions that enhance vocabulary proficiency can construct scalable preparation tools at a lower cost to help test takers. The key question that this research addresses is, can Llama 3 (1), an open-source Large Language Model (LLM), generate SAT style Words in Context (2) type questions; as shown in Table 1; using in-context Learning (Prompt Engineering) with the necessary accuracy and sophistication?

Historically, Question Answer Generation (QAG) for educational purposes has long been a focal point in AI research. The evolution from prior Natural Language Processing (NLP) techniques, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM), to the superior Transformers architecture has allowed text generation to become far more versatile due to the Transformers ability to focus on specific words, regardless of their positions in a sentence, using the attention mechanism (3). Transformers architecture based LLMs such as, Llama and ChatGPT, have demonstrated proficiency in generating meaningful and context-relevant content (4). Despite these advancements, many existing systems lack precision in incorporating given words in passages, often relying on basic approaches that overlook contextual relevance. Moreover, the complexity of SAT questions, which frequently demand analytical reasoning and comprehension, remain under explored in the scope of current content-generation frameworks. This research aims to address these gaps by employing the Llama 3 model for content generation, forming a pipeline that incorporates contextual relevance into the content. Table 1 shows a sample Words in Context question from College Board, the organization that administers the SAT (5). This research attempts to generate questions in the format and sophistication exhibited by that in Table 1.

Table 1. Sample words in Context question from College Board

Question	Explanation
In recommending Bao Phi's collection Sông I Sing, a librarian noted that pieces by the spoken-word poet don't lose their _____ nature when printed: the language has the same pleasant musical quality on the page as it does when performed by Phi. Which choice completes the text with the most logical and precise word or phrase? A) jarring B) scholarly C) melodic D) personal	Correct Answer: Choice C is the best answer. "Melodic," referring to a pleasant arrangement of sounds, effectively signals the later use in the passage of "pleasant musical quality" to refer to Phi's spoken-word poetry when read rather than heard. Distractor Explanations: Choice A is incorrect because "jarring," meaning disagreeable or upsetting, suggests the opposite of what the passage says about the "pleasant musical quality" of Phi's spoken-word poetry, whether read or heard. Choice B is incorrect because "scholarly" does not effectively signal the later use in the passage of "pleasant musical quality" to refer to Phi's spoken-word poetry.

	Choice D is incorrect because “personal” does not effectively signal the later use in the passage of “pleasant musical quality” to refer to Phi’s spoken-word poetry.

2. Related work

QAG has been an area of active research in Natural Language Processing (NLP), with various methods explored to improve the quality and relevance of generated questions. Traditional approaches used rule-based and template-driven methods (6), which lacked adaptability and scalability, especially for complex educational assessments such as the SAT.

Transformer-Based Approaches

The introduction of deep learning and Transformer-based architectures (3) significantly improved QAG by allowing models to generate contextually aware and dynamic questions. Models such as BERT (7) and T5 (8) have been fine-tuned on educational datasets to generate high-quality questions. However, these models depend on large amounts of labeled training data and often struggle to integrate external knowledge dynamically, which limits their effectiveness in crafting SAT-style "Words in Context" questions.

Large Language Models (LLMs) for question generation

With the emergence of large-scale language models like GPT-3 (9) and ChatGPT, question generation has seen notable advancements. These models produce more coherent and nuanced questions, yet they present challenges such as hallucinations, limited domain adaptation, and dependence on proprietary Application Programming Interface (API)s, which restrict accessibility and customization. Open-source models like Llama 3 have been explored as viable alternatives due to their flexibility and cost-effectiveness, but they still require strategic fine-tuning for optimal educational application.

Retrieval-Augmented Generation (RAG) and Semantic Similarity (SS)

Retrieval-Augmented Generation (10) has emerged as a promising method to fix factual errors by using external knowledge for the generative process. Research has demonstrated its ability to improve answer quality in QAG tasks, but its application to educational domain remains relatively unexplored. Additionally, leveraging semantic similarity techniques (11) has shown potential in refining question generation by ensuring that generated content aligns closely with the intended context.

Limitations of existing approaches

Despite the advancements in model architecture, existing methods struggle to generate questions that meet the specific structural and cognitive demands of standardized tests like the SAT. Transformer-based models often produce factually inconsistent or semantically misaligned content when asked to generate content on topics they were not exposed to during pre-training. This is seen in the baseline tests with a 45% hallucination rate. RAG reduces hallucination to 20% by using retrieved context. This use

case in QAG category is unique, in that the goal is to optimize Words in Context type question generation for given a genre and word combination that has limited prior research, especially for the SAT. However, this research builds on prior work that showed the effectiveness of RAG in such a scenario (10).

3. Methods

3.1 Model selection

The Llama 3 8B Instruct model was chosen for this research since it is one of the leading open-source LLMs (1). In addition, Llama can run on inexpensive machines for inference. This model is a decoder only model. It outperformed against competing models such as Gemma 2 9B and Mistral 7B in the General evaluation category, as shown in Figure 1. Commercial models, such as ChatGPT, etc., were not considered due to cost implications. Furthermore, for building an inexpensive website/mobile-app for students, an open-source model is not limited by permissive licensing.

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B
General	MMLU (5-shot)	69.4	**72.3**	61.1
	MMLU (0-shot, CoT)	**73.0**	72.3	60.5
	MMLU-Pro (5-shot, CoT)	**48.3**	-	36.9
	IFEval	**80.4**	73.6	57.6

Figure 1. Performance of Lama 3 on key benchmark evaluations under the General Category (1).

Llama 3 is available in three sizes, 8B (Billion), 70B and 40B. The 8B size was chosen because it encodes the content much more efficiently, primarily because of its tokenizer that has a vocabulary of 128K tokens (1). That efficiency leads to better model performance making it possible to run inference on machines that are relatively inexpensive, such as the Nvidia A100 (12), a 1 GPU, machine. Llama 3 is pretrained on 15T (Trillion) tokens that were all collected from publicly available sources and is further trained on sequences of 8,192 tokens using a mask to ensure that self-attention does not cross document boundaries (1).

Figure 2 is a high-level overview of how Llama 3's Transformer architecture works during inference in the context of generating SAT word in Context questions (1). The process begins with tokenization, where the input prompt, for example, "*Generate a paragraph on the Black Death with the word surreptitious in it*" is broken down into smaller units called tokens. These tokens are then converted into numerical embeddings, which serve as the model's way of understanding the relationships between words. At this stage, the model identifies key terms like "Black Death" and "surreptitious" recognizing their significance in shaping the response. These numerical representations are then passed through multiple layers of the model for deeper processing.

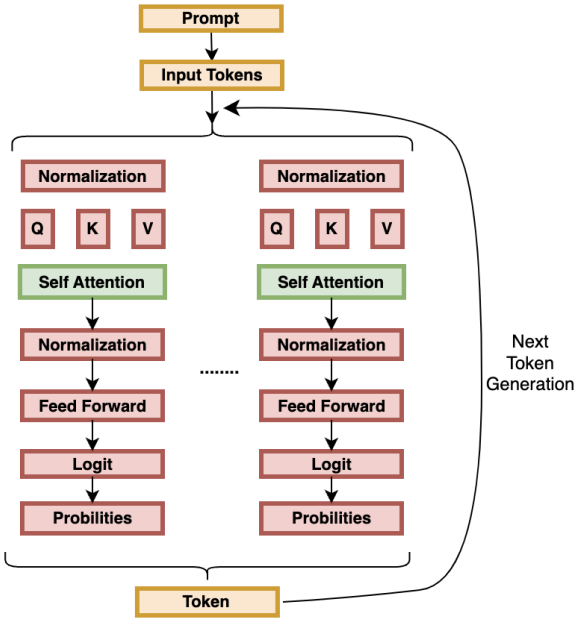

Figure 2. Llama 3 architecture

The attention heads operate in a similar way to other transformer-based architectures. Llama uses multi-head self-attention mechanisms to process tokenized inputs, assigning varying levels of focus to different tokens based on their relevance to the current context. KV caching (Key-Value caching) is a technique used by Llama, to speed up inference by storing previously computed key (K) and value (V) tensors. Instead of recomputing attention for all tokens in a sequence, the model caches these tensors and reuses them for new incoming tokens.

Next, the model applies self-attention and feedforward networks to analyze the input context. Using multi-headed self-attention, it determines how strongly each word relates to others. For example, it understands that "*The Black Death*" should be the main focus and that "*surreptitious*" must be incorporated naturally. The model retrieves relevant knowledge from its pre-trained data on *"The Black Death."* Rotary positional embeddings (RoPE) help the model maintain coherence in sentence structure, ensuring that the generated text flows logically (1). The model refines its response through multiple layers, reinforcing context and meaning.

Finally, during the text generation phase, the model employs autoregressive decoding, predicting and generating one word at a time. It begins constructing the response by appending to the input prompt: "*The Black Death, a pandemic that ravaged Europe…*" while ensuring grammatical accuracy and logical progression. As it continues, it places "*surreptitious*" where it fits naturally: "*The Black Death, a pandemic that ravaged Europe in the 14th century, was a surreptitious force…*" The model generates each subsequent

word by selecting the most probable next token, maintaining fluency to the extent it has learned through pre-training. The model concludes generations after it reaches the maximum number of new tokens count that is passed as part of the model configuration.

3.2 Dataset

The Words in Context evaluation in the SAT focuses on words that fall under "general academic words" or "rich vocabulary". These words are often ones that test takers are unaccustomed to. For example, instead of the word 'walk', its nuanced form 'saunter' could be used. The datasets used in the research consisted of 1) SAT words – This dataset (csv format) has 1000 SAT words (13). Examples are Abate, Aberration, Abhor, Abject, Abjure, etc. 2) SAT Genre – content is derived from a variety of topics that aligns with the type of genres that College Board leverages. This dataset has a total of 300 genres. Examples are, Emergence of Homo sapiens, Use of fire by early humans, Development of stone tools, Agricultural Revolution, Establishment of Jericho, etc. 3) Baseline Test Cases – Test cases were generated by selecting random words and genres from the words and genre datasets. This is a naïve approach to evaluate how the model behaves and introduce other approaches such as semantic similarity as required. Figure 3 shows a sample of these test cases.

```
import pandas
from pandas import DataFrame
import os

# let's load the test cases file
test_cases_file='eval_word_genre.csv'
test_cases = pandas.read_csv(test_cases_file)

test_cases.head(5)
```

	word	genre	definition
0	Resilient	Global population surpasses 8 billion	Able to recover quickly from difficult condit...
1	Efficacy	Rise of social media	The ability to produce a desired or intended ...
2	Obstinate	Egyptian pyramids	Stubbornly refusing to change one's opinion o...
3	Crass	Rise of Islam	Lacking sensitivity, refinement, or intellige...
4	Whimsical	Trojan War	Playfully quaint or unusual in an appealing w...

Figure 3. Sample Test cases

Experiment test cases were created with genre and word pairings that have a higher semantic score. The process at a high-level was; depending on the number of test cases that needed to be created, a genre and 20 words were selected at random, a semantic score for each word against the genre was created, and the word that had the highest score and additional was > 0.5 was chosen. If none of the words resulted in a score > 0.5, the process repeated.

3.3 Setup

A Nvidia A100 single-GPU machine from Lambda Labs (12) was used for the experiment. However, a machine with 16GB of memory is good enough to run inference. For faster inference, A100s are a good choice. The specifications were; GPU: 1x NVIDIA A100 SXM, Memory: 40GB, Ram: 220 GiB.

3.4 Experiment pipeline

The pipeline in Figure 4 below shows the steps that were used for the experiments.

Figure 4. Experiment pipeline

The pipeline consisted of three core steps, 1) Input step, this section included the component that read from input datasets (word and genre), selected the genre and word combination using semantic similarity, and assembled answer choices. 2) Inference step, this section formatted the input into a prompt, leveraged the input configuration to run inference against the model using hugging face's https://huggingface.co/ libraries, as shown in figure 5 below. 3) Post processing step, this section validated the output, ran post processing logic, routed for manual review, and finally the output was used for delivering the questions through a web app or mobile app. The Code for the pipeline and the datasets can

be found in the code repository (14). All the experiments outlined in the paper can be easily reproduced using the code.

```
import csv
import numpy as np

paragraph_length=150
replacement_mask='_________________'

# function for running inference against the model
def run_inference(prompt, paragraph_length):
    inputs = tokenizer([prompt], return_tensors="pt").to(device)
    outputs=model.generate(**inputs, generation_config=generation_config)
    transition_scores = model.compute_transition_scores(outputs.sequences, outputs.scores, normalize_logits=True)
    input_length = 1 if model.config.is_encoder_decoder else inputs.input_ids.shape[1]
    complete_text=''
    for t in outputs.sequences:
        complete_text += tokenizer.decode(t)

    generated_tokens = outputs.sequences[:,input_length:]
    generated_text = ''
    for t in generated_tokens:
        generated_text += tokenizer.decode(t)

    return [generated_text, generated_tokens, transition_scores, complete_text]
```

Figure 5. Code sample shows how Hugging Face Transformer libraries were used to run inference against the Llama mode.

3.5 Evaluation method

For evaluating the quality of text generated, Perplexity, ROUGE score and BLEU score (15) are some of the popular metrics that can be leveraged. These metrics help in assessing the model's performance by measuring various aspects of the generated text, such as fluency, coherence, accuracy, and relevance. For these metrics to work, labeled examples, called the ground truth are required. For the use case described in this paper, creating ground truth will not render valid evaluation since a gold standard paragraph for given a word and genre combination does not and cannot exist. Several different paragraphs using the word and genre combination could still be correct. Hence, this paper recommends manual evaluation along with leveraging automated reasoning that is generated using the model, to validate if the word fits in a logical and precise manner in the paragraph. The metrics used for evaluating the experiments are described in this manuscript. The experiment that results in a high accuracy and precision, is the best approach for the use case. Note that the metrics of recall and F1 scores are not relevant for the evaluation as this experiment is not a classification task.

3.5.1 *Word inclusion*

This simply measures the number of paragraphs that include the requested word. The model does not include the word when it is unable to form a sentence with that particular genre and word combination. This is because of the pre-training data not containing such context or information.

3.5.2 *Manual assessment*

This assessment examines fluency and coherence using two qualitative metrics, word subject agreement and sentiment matching. Word subject agreement simply states that the word must be applicable to the correct subject; as an example, a flower cannot be talkative. When the model uses the word 'aggrandize' in the context of 'landmass', it lacks coherence and is flagged as not valid. Sentiment matching checks whether the word fits the mood. For example, a cemetery should not normally be described as 'jolly'. Paragraphs are marked as hallucination when subject agreement and sentiment matching is not valid.

3.5.3 *Precision*

This metric is also called as positive prediction, a measure of actual correct prediction. Precision = # of valid outputs - # of outputs containing hallucinations / (# of valid outputs + # of outputs containing hallucinations). Precision helps with the evaluation of the model performance after hallucinations are purged.

3.5.4 *Accuracy*

This metric is the proportion of correctly generated paragraphs out of the total number of test cases. This measures the overall performance of the model. Accuracy = # of valid outputs - # of outputs containing hallucinations / (# of valid outputs + # of outputs containing hallucinations + # of Paragraphs without the word).

3.6 Experiments

This section describes the experiments conducted using various In-Context Learning techniques to incrementally make the responses better suit the goal of generating content that meets the SAT "Word-In-Context" rigor.

3.6.1 *Baseline*

The purpose of the baseline was to gain insights into the quality of the content the model generated. The test included 100 test cases, each test case was a random genre and word combination for which the model needed to generate a 150-word paragraph on the given genre with the word included. The 150-word length was chosen based on sample questions from the College Board. Question lengths varied from 50 – 150 words (5) and the higher-end was chosen for the baseline. Based on the analysis of the results, the word count can be adjusted in further experimentation.

Table 2. Sample baseline output

Question	Automated Explanation	Manual Assessment Explanation
Prompt: **Generate a paragraph on The Black Death with the word surreptitious in it.** Response: The Black Death, a pandemic that ravaged Europe in the 14th century, was a **surreptitious** force that spread rapidly and silently, catching many off	Prompt: Is surreptitious the most logical and precise word in the following paragraph? Response should be, answer: yes or answer: no. Here is the paragraph: The Black Death, a … Answer: yes. The word surreptitious is used to describe the way the disease spread,	The word "surreptitious" means something done secretly or without open acknowledgment. The response correctly uses "surreptitious" to describe the Black Death as an unseen and unacknowledged force spreading in a way that was not immediately noticed. This is supported by historical evidence, as the spread of the Black Death was often ignored. Additionally, the passage uses the word's meaning

guard. The disease, caused by the bubonic plague, was transmitted through the bites of infected fleas that lived on rats and other rodents. As the disease spread, it was often kept secret by those who were infected, as they feared being ostracized or even killed by those who did not understand the cause of the illness.	which was silent and unnoticed, and the way it was kept secret by those who were infected. The word accurately conveys the idea of something being done secretly or stealthily, which is the intended meaning in this context. It is a precise and logical choice.	by explaining how infected individuals hid their condition out of fear, further showing the secretive nature of the disease's transmission. This paragraph effectively uses "surreptitious" in a logical and contextually appropriate manner. However, the classical application of the word is for events, not for abstractions, such as 'force'.
Prompt: Generate a paragraph on Chinese Pagodas with the word assiduous in it. Response: Chinese pagodas are iconic structures that have been a part of the country's rich cultural heritage for centuries. These **assiduous** structures, with their intricate carvings, ornate decorations, and towering heights, have been a symbol of Chinese architecture and spirituality. The pagodas' unique design, which combines elements of Buddhism, Taoism, and Confucianism, reflects the country's complex cultural landscape. The **assiduous** attention to detail in the construction of these structures is evident in the intricate carvings, which often depict scenes from Chinese mythology and folklore. The pagodas' towering heights, which can reach up to"	Prompt: Is assiduous the most logical and precise word in the following paragraph? Response should be, answer: yes or answer: no. Here is the paragraph: Chinese pagodas are… Answer: yes. The word "assiduous" is used two times in the paragraph, and each time it is used to describe the structures, the attention to detail, and the craftsmanship of the artisans. The word accurately conveys the idea of careful and diligent effort, which is a fitting description of the intricate carvings and the construction of the pagodas.	The model hallucinates in trying to fit the word ***assiduous*** in the paragraph. Definition of assiduous is "showing great care and attention to something, often in a persistent and diligent manner". In the paragraph, the model describes the Chinese Pagoda structures as assiduous, which is incorrect because the model uses assiduous to describe the structures. In the second instance, the word is used to describe the creators for their assiduous attention, which is correct. Overall, the paragraph is not valid due to hallucination in the first instance. In addition, the model includes the word more than once, which is not ideal. Reducing the word-count may help.

Table 3 shows the findings upon running 100 test cases and analyzing the results. The Out-of-the-box baseline model performance is poor with an accuracy of 47% and precision of 0.69. The complete test output can be found in the code repository (14).

Table 3. Summary of Baseline Tests

Word Included	Word Not Included	Word in a Different Form	Word Inclusion (%)	Valid Paragraph	Hallucination Count # (%)	Precision	Accuracy
68	32	2	68%	47	21 (45%)	0.69	47%

The baseline test findings demonstrated that the Text generation was not deterministic. Each time the same prompt was run, the response was different but similar in meaning. In 32 out of 100 cases, the response did not include the word that the model was asked to include. The model hallucinated 45% of the time. This happened when the model was asked to generate a paragraph with an odd combination of word and genre. For example, "The Chinese

Pagodas" genre and the word "assiduous" represent an odd combination. In many cases (75 out of 100) tokens representing the word were experiencing split. Further exploration is needed to check if the sub-word tokenization was affecting the coherence of the generated paragraph. For example, the word assiduous was split into three parts when generating and each sub-word had a 100% next word probability. Previous research (19) did not identify this anomaly. When prompted for a longer output, such as 150 words, the model generated paragraphs with the word included more than once. There were 20 such occurrences. In addition, the paragraphs were becoming too 'wordy'. Given the SAT Words in Context questions have 50 to 150 words, selecting 100 words would be a good balance. In some cases, the model used a different form of the word. For example, when prompted for the word 'laud', the model came up with the word 'laudably' in order to fit the sentence. Similarly, when prompted for the word 'engross', the model came up with the word 'engrossing'. In such cases, the word was changed accordingly in post processing. Some of the responses included random text. Random quotes, citations, etc., were also found.

The following sections describe the changes that were made for addressing the above deficiencies, starting with identifying an inference configuration to facilitate response evaluation in a predictable manner, leveraging semantic similarity for selecting genre and word combination, and logic for selecting answer choices.

3.6.2 *Inference configuration*

Figure 6 shows the configuration that was used for all the experiments. It is set to produce deterministic output and return output logits, and the probability of word generation, among other metrics.

```
from transformers import GenerationConfig
generation_config = GenerationConfig(
        # number of tokens to generate
        max_new_tokens=150,
        # only choose from the top k most likely words
        top_k=50,
        # Whether or not to use sampling ; use greedy decoding otherwise.
        do_sample=True,
        # parameter that controls the randomness or creativity of the generated text
        temperature=0.001,
        # sets the pad tokens to whatever it is in the tokenizer
        pad_token_id=tokenizer.eos_token_id,
        # output unnormalized outputs
        output_logits=True,
        # output the probabilities
        output_scores=True,
        # passes hidden state along with output
        output_hidden_states=True,
        #returns output as a dict
        return_dict_in_generate=True
    )
```

Figure 6. Configuration used for the experiments

```
print(generation_config)
```

```
GenerationConfig {
  "do_sample": true,
  "max_new_tokens": 150,
  "output_hidden_states": true,
  "output_logits": true,
  "output_scores": true,
  "pad_token_id": 128009,
  "return_dict_in_generate": true,
  "temperature": 0.001
}
```

Figure 6 (continued). Configuration used for the experiments

3.6.3 *Hallucination reduction*

The model hallucinated because it was not exposed to certain word genre pairings during pretraining. To mitigate this discrepancy, semantically similar word genre pairs were selected. A semantic similarity score for a given genre and 20 random words was generated. The combination that achieved the highest semantic score was selected. Semantic score generation was performed by using an embedding model, BAAI/bge-large-en-v1.5 (18) and Lambda Index. The embedding model has a 1024-dimensional embedding, meaning it generates embedding vectors with 1024 values for each piece of text it processes. Similarity scores ranged from 0 to 1, with 0 implying no semantic similarity and 1 being indicative of a very high semantic similarity.

Test cases creation using Semantic Similarity was outlined in section 3.2. As an example, Table 4 shows words that are scored against the genre 'Persian Wars'. The word "volatile" has a greater semantic score compared to other words. Using this word when paired with this genre hence gives the model a better chance of resulting in a coherent paragraph, thus minimizing hallucination. A screenshot of the source code used for the semantic similarity analysis is shown in Figure 7.

Table 4. Semantic score for the genre 'Persian Wars' against 10 random words.

Word	Similarity Score
volatile	0.51964003
defiant	0.502297435
flout	0.498239049
implicit	0.493628775
illustrate	0.4931558
reprehensible	0.492156056
overt	0.490368928
languid	0.474414434
testimony	0.464038252
utilitarian	0.451558958

```
from llama_index.embeddings.huggingface import HuggingFaceEmbedding
embed_model = HuggingFaceEmbedding( model_name="BAAI/bge-large-en-v1.5", trust_remote_code=True)
```

```
output_df = DataFrame(columns=['genre', 'word', 'similarity_score', 'definition', 'invalid_answer_choices'])
test_cases_count=1000
word_count=10
invalid_choice_count=3
counter = 0
while True:
    words=[]
    genre = (genre_df['genre'][randrange(genre_df.shape[0])]).lower()
    for i in range (word_count):
        key = randrange(vocab_df.shape[0])
        words.append((vocab_df['word'][randrange(vocab_df.shape[0])]).lower())

    highest_score=0
    similarity_scores={}
    selected_word=''
    passing_count=0
    for i in range (len(words)):
        result = await evaluator.aevaluate(
            response=genre,
            reference=words[i],
        )
        similarity_scores.update({result.score:words[i]})
        if (result.passing):
            passing_count += 1
        # print("{},{}".format(words[i],result.score))

    # we need atleast one match with greater than 50%
    if passing_count == 0:
        continue

    # sort the ditionary so that we can pick the word that has highest similarity and pick
    # the bottom 3 for invalid choices
    scores = list(similarity_scores.keys())
    scores.sort()
    sorted_scores = {i: similarity_scores[i] for i in scores}

    invalid_choices=[]
    for k in range(invalid_choice_count):
        invalid_choices.append(sorted_scores[scores[k]])

    # word with highest similarity
    selected_word = sorted_scores[scores[len(scores) - 1]]

    # capture the word definition
    word_definition = vocab_df[vocab_df['word'].str.lower() ==
                       selected_word.lower()].squeeze()['definition']

    # store in the dataframe
    output_df = output_df.append({'genre':genre,'word': selected_word,
        'similarity_score':scores[len(scores) - 1], 'definition': word_definition, 'invalid_answer_choices':inval
        ignore_index=True)

    print("{} - Genre: {}; Word: {} ".format(counter, genre, selected_word))
    counter += 1

    # break if test cases count has been reached
    if counter >= test_cases_count:
        break
```

Figure 7. Screenshot of the source code that checks semantic similarity and creates a dataset.

The examples described below show how semantic scores play a role in reducing hallucination. Three examples below are texts generated for the words ‘volatile’, ‘testimony’ and ‘utilitarian’, selected from the top and bottom of the list in Table 4 with the genre ‘Persian wars’. The first sample has a higher semantic score and hence generates a coherent paragraph compared to samples 2 and 3.

Sample 1. Genre: The Persian Wars; Word: volatile

“The Persian Wars, fought between the Persian Empire and the city-states of Greece, were a series of conflicts that lasted from 499 to 449 BCE. The wars were sparked by the Persian Empire's desire to expand its territories and impose its rule on the Greeks. The **volatile** situation was further exacerbated by the fact that the Persians were a powerful and well-organized empire, while the Greeks were a loose confederation of city-states.”.

The paragraph generated was coherent and the word ‘volatile’ was used correctly.

Sample 2. Genre: The Persian Wars; Word: testimony

“The Persian Wars, fought between the Persian Empire and the city-states of Greece, were a series of conflicts that lasted from 499 to 449 BCE. The wars were sparked by the Ionian Revolt, a rebellion by the Greek city-states of Ionia against Persian rule. The Persian Empire, under the rule of Darius I, responded with a massive invasion of Greece, which was met with fierce resistance from the Greeks.”.

The model was not able to include the word in the paragraph.

Sample 3. Genre: The Persian Wars; Word: utilitarian

“The Persian Wars, fought between the Persian Empire and the city-states of Greece, were a series of conflicts that lasted from 499 to 449 BCE. The wars were sparked by the Persian Empire's desire to expand its territory and impose its utilitarian brand of governance on the Greek city-states. The Persians, under the rule of Darius I, saw the Greeks as inferior and believed that they could be easily conquered and assimilated into their empire.”.

Based on the text above, the model hallucinates. Utilitarianism is a moral philosophy that argues that the right action is the one that maximizes overall happiness or well-being. The Persians are described as seeking to impose their own system of government and culture on the Greeks, which implies that they were trying to impose their own values and beliefs on the Greeks, rather than simply seeking to expand their territory.

The above observations suggest that genres and words that score greater with particular genres have a greater probability to include the word in the paragraph and to generate coherent paragraphs.

	genre	word	similarity_score	definition	answer_choices
0	buddhist stupas	sublime	0.570303	awe	sublime,manifest,vague,skeptic
1	evolution of first tetrapods (four-limbed vert...	eclectic	0.539813	deriving ideas or style from a diverse range o...	oblivious,tirade,vigilant,eclectic
2	appearance of the first trees	inventive	0.513281	imaginative, creative, able to think of new th...	castigate,taint,utilitarian,inventive
3	emergence of earth's first atmosphere	viable	0.525630	capable of working or being successfully imple...	stolid,abstain,belittle,viable
4	alexander the great's conquests	fervor	0.524804	great warmth and intensity of feeling; She spo...	belittle,vapid,fervor,vague

Figure 8. Revised word and genre combination test cases based on semantic similarity

3.6.4 *Answer choice selection*

Words in Context offers four answer choices. One of them is the correct choice and the other three are distractors. Of the distractors, one of them is an antonym and the other two are words that are not relevant for the context. This can be seen from the choices presented in Table 1. The correct choice is the word ‘melodic’. The word ‘Jarring’ is an antonym. The other two word choices, ‘personal’ and ‘scholarly’, are not relevant.

In this research, the Llama model was used for generating the antonym. The other two distractor words were the ones that had lower semantic similarity scores against the genre. For example, in Sample #1 (Genre: Korean War) in Table 9, the correct choice is the word ‘impasse’. Its antonym is the word ‘breakthrough’. Two other incorrect choices – the words ‘sycophant’ and ‘blanch’ are the ones that have lower similarity scores of 0.45 and 0.46 respectively.

3.6.5 *In-Context Learning (Prompt Engineering) experimentation*

Given that the baseline model underperformed with an accuracy of 47%, this research further explored if In-Context Learning (16,17) enhanced the quality of the responses. The three main objectives of In-Context Learning were 1) Providing additional context – feeding the LLM with more information when querying the model. For example, sharing the word meaning to prevent hallucination. 2) Reducing ambiguity - given that a word can have different meanings depending on the context, it can cause confusion at text generation time. 3) Real-world grounding - Although LLMs can process language very well, they may lack a true "understanding" of the real world, which can limit their ability to interpret certain concepts. Some of the Prompt Engineering techniques that were explored included the Zero-shot and the Few-shot.

Experiment 1: Zero-shot

Zero-shot is a prompt engineering technique used to provide instructions to the model to perform a task it has not been specifically trained to do (16, 17). To perform a specific task, the model must be trained using labeled examples. However, with Zero-shot, the idea is to make use of the model's general understanding and leverage the additional context provided through the prompt for generating output to fit SAT requirements. Below is a sample zero-shot prompt and output. As part of the prompt, a template output is provided for the model to follow: *"Question: …… The Answer is: Insert Paragraph here"*.

Table 5. Sample Zero-Shot output

Question	Automated Explanation	Manual Assessment Explanation
Sample #1: Genre: Invention of the printing press; word: catalyst Generate a paragraph on invention of the printing press with the word catalyst in it. Response: The invention of the printing press by Johannes Gutenberg in the 15th century was a significant _____ for the spread of knowledge and literacy during the Renaissance. Prior to the invention of the printing press, books were laboriously copied by hand, a process that was time-consuming and expensive. The printing press revolutionized the dissemination of information, making it possible for books to be mass-produced and distributed widely.	Is catalyst the most logical and precise word in following paragraph? Response should be, answer: yes or answer: no. Response: This led to a significant increase in the availability of written materials, which in turn contributed to a rise in literacy rates and a broader dissemination of knowledge. Answer: yes. The word "catalyst" is the most logical and precise word in this paragraph because it accurately describes the role of the printing press in accelerating the spread of knowledge and literacy during the Renaissance.	The word "catalyst" refers to something that accelerates or facilitates a process without being consumed in the process itself. The response correctly applies "catalyst" to describe the printing press as a transformative force that accelerated the dissemination of knowledge and literacy. The explanation of how books were previously copied by hand and how the printing press changed that process provides strong contextual support for why "catalyst" is an appropriate choice in this paragraph. The paragraph successfully uses both the literal and metaphorical meaning of "catalyst," making it a valid usage.
Sample #2: Genre: African tribal huts; word: dilapidated Generate a paragraph on African tribal huts with the word dilapidated in it. Response: The African tribal huts, once a symbol of vibrant culture and community, now stood as _____ structures, worn down by time and neglect. The once-sturdy thatched roofs now hung precariously, threatening to collapse at any moment, while the wooden frames creaked and groaned in the gentle breeze.	Prompt: Is dilapidated the most logical and precise word in following paragraph? Response should be, answer: yes or answer: no. Here is the paragraph: The African tribal huts… Response: The once-vibrant colors had faded, leaving the huts a dull, weathered gray. The once-thriving community had long since abandoned the area, leaving the huts to stand as a testament to the transience of human existence. Answer: yes. The word "dilapidated" accurately describes the state of the huts, which are worn down, worn out,	The word "dilapidated" means worn out, or in a state of disrepair. The paragraph correctly applies "dilapidated" to describe the structural decline of the huts due to time and neglect. The supporting details, such as the precariously hanging roofs and creaking wooden frames, reinforce the visual and contextual meaning of "dilapidated." This ensures the word is used in a way that is both accurate and meaningful, making the paragraph a valid response.

	and in disrepair. The word effectively conveys the idea that the huts are no longer in good…	

Table 6 shows the findings upon running 100 test cases and analyzing the results. Accuracy improved slightly compared to the baseline prompt, from 47% precision to 62%.

Table 6. Summary of Zero-shot Tests

Word Included	Word Not Included	Word in a Different Form	Word Inclusion (%)	Valid Paragraph	Hallucination Count # (%)	Precision	Accuracy
73	27	1	73%	62	11 (18%)	0.85	62%

Experiment 2: Few-shot

Few-shot is a prompt engineering technique where the model is provided with a few examples to use as the context (8, 9). The difference between zero-shot learning and Few-shot learning is that a handful of labeled examples are provided to help the model understand the task's output structure. Table 7 lists some Few-Shot prompts which include a labeled example.

Table 7. Few-shot Output

Question	Automated Explanation	Manual Assessment Explanation
Sample #1: Genre: Invention of the printing press; word: catalyst Prompt: Generate a paragraph on The Iguazu Falls with the word revered in it. The Answer is: The Iguazu Falls, which lie on the border between Argentina and Brazil, are a popular and revered tourist destination. The waterfalls have been visited by people from all over the world for over a century, and they are often cited as one of the most impressive natural wonders on Earth. Generate a paragraph on invention of the printing press with the word catalyst in it. Response: The invention of the printing press by Johannes Gutenberg in the 15th century was a major _____ for the spread of knowledge and literacy during the Renaissance. Prior to the invention of the printing press, books were copied by hand, a time-consuming and laborious process that limited the	Prompt: Is catalyst the most logical and precise word in following paragraph? Response should be, answer: yes or answer: no. Here is the paragraph: The invention of the printing press... Response: The invention of the printing press was a turning point in the history of communication, and it played a crucial role in shaping the course of human history. Answer: yes. The word "catalyst" is used to describe the printing press as something that accelerates or triggers a significant change or process, which is a precise and logical usage of the term. The word	"Catalyst" is something that accelerates or triggers a major change. The paragraph correctly applies "catalyst" by showing how the printing press accelerated the spread of literacy and knowledge. The contrast between the slow, labor-intensive process of hand-copying books and the efficiency of mass printing strengthens the argument that the printing press was indeed a catalyst for societal transformation. Response is very similar to Zero-shot output.

availability of written materials. The printing press allowed for mass production of books, making written knowledge more widely available and contributing to a significant increase in literacy rates.	"catalyst" is often used in scientific and technical contexts to describe a substance or agent that speeds up...	
Sample #2: Genre: African tribal huts; word: dilapidated Prompt: Generate a paragraph on The Iguazu Falls with the word revered in it. The Answer is: The Iguazu Falls, which lie on the border between Argentina and Brazil, are a popular and revered tourist destination. The waterfalls have been visited by people from all over the world for over a century, and they are often cited as one of the most impressive natural wonders on Earth. Generate a paragraph on African tribal huts with the word dilapidated in it. Response: The African tribal huts, with their thatched roofs and mud walls, are a common sight in many rural areas of the continent. However, many of these huts are now _____ and in disrepair, a testament to the challenges faced by many African communities in terms of poverty and lack of resources. Despite their humble appearance, these huts are often the heart of the community, serving as a place of shelter, social gathering, and cultural celebration.	Prompt: Is dilapidated the most logical and precise word in following paragraph? Response should be, answer: yes, or answer: no. Here is the paragraph: The African tribal huts... Response: Answer: yes. The word "dilapidated" is the most logical and precise word to describe the state of the huts in the paragraph. It accurately conveys the idea that the huts are in a state of disrepair and decay, which is supported by the context of the paragraph. Other options, such as "ruined" or "destroyed", might be too strong, while "old" or "worn out" might not convey the same level of decay and …	"Dilapidated" means something that is in a state of disrepair due to neglect or aging. The paragraph correctly uses "dilapidated" to describe the deteriorating state of the huts, reinforced by vivid imagery such as collapsing roofs and creaking wooden frames. These details effectively convey a sense of decay, making the usage of "dilapidated" appropriate and well-integrated into the context. Response is very similar to Zero-short prompt response.

Table 8 shows findings upon running the same 100 test cases used in Zero-shot and analyzing the results. The result was very similar to the zero-shot prompt. Zero-shot and few-shot are similar approaches, and the results are expected to be in the same range because the prompt is provided with formatting instructions only with no new information on the genre or word.

Table 8. Summary of Few-shot Tests

Word Included	Word Not Included	Word in a Different Form	Word Inclusion (%)	Valid Paragraph	Hallucination Count # (%)	Precision	Accuracy
93	7	1	93%	60	33 (55%)	0.65	60%

3.6.6 *Retrieval Augmented Generation (RAG)*

Upon analyzing the output from the above approaches, it was evident that the model needed more context; beyond the instructions provided in the input; in order to generate high quality responses. Many of the genre and word combinations were odd and the model may not have been trained with such data. For example, asking to include the word assiduous in a 150-word paragraph on the Chinese Pagodas is unlikely to result in a cogent paragraph, as shown in Table 2. That is where RAG (16,17) was anticipated to fill the gap. RAG can provide context dynamically, so as to nudge the token generation in a certain direction. This is particularly helpful when the pre-trained data is out of date, or the pre-trained data lacks the required knowledge or information. The context for RAG involves passing information as part of the prompt or storing relevant documents that provide more context in a vector database. At query time, semantically similar documents are retrieved as context for the LLM to generate a response. For the experiments in this paper, Word meaning was passed dynamically as part of the prompt without having to leverage a vector database. Table 9 shows a few samples where word meaning was passed as part of the prompts. Note that answer choices are included here and not in other experiments above, in order to avoid repetition.

Table 9. RAG Output

Question	Automated Explanation	Manual Assessment Explanation
Sample #1: Genre: Korean war; word: impasse Prompt: Given the meaning of the word impasse: a situation in which a conflict or problem cannot be resolved because of a lack of agreement or a stalemate; Generate a paragraph on Korean war with the word impasse in it. Make sure your paragraph is one single paragraph that is formally worded. If you are done generating the paragraph, stop. Make sure to use the given word impasse as is and only use it once. Response: The Korean War, which lasted from 1950 to 1953, was a stalemate that reached an _____ in the early 1950s. The war began when North Korean forces, backed by China and the Soviet Union, crossed the 38th parallel and invaded South Korea, prompting a response from the United Nations and the United States. Answer Choices: A) sycophant B) blanch C) impasse	Prompt: Is impasse the most logical and precise word in following paragraph? Response should be, answer: yes, or answer: no. Here is the paragraph: The Korean War… Answer: yes. The word "impasse" is the most logical and precise word in this paragraph because it accurately describes the situation where the war reached a point where neither side could gain a decisive advantage, and the conflict became stuck or stalled.	The word "impasse" refers to a deadlock or situation where no progress can be made due to disagreement. In this context, the Korean War resulted in a military and political impasse because neither side was able to secure a decisive victory, leading to a prolonged ceasefire without a formal peace treaty. The paragraph effectively conveys this meaning by linking the term to the unresolved conflict and ongoing division between North and South Korea. Answer choices include a correct choice, which is 'impasse'. An antonym, which is 'breakthrough'. Two other incorrect choices that have lower similarity score: 1/ 'sycophant' with a similarity score of 0.45; 2/ 'blanch' with a similarity score of 0.46.

D) breakthrough		
Sample #2: Genre: Invention of the printing press; word: catalyst Prompt: Given the meaning of the word catalyst: a substance that speeds up a chemical reaction without being consumed or altered in the process; Generate a paragraph on invention of the printing press with the word catalyst in it. Make sure your paragraph is one single paragraph that is formally worded. If you are done generating the paragraph, stop. Make sure to use the given word catalyst as is and only use it once. Response: The invention of the printing press by Johannes Gutenberg in the 15th century was a groundbreaking event in the history of communication, marking a significant turning point in the dissemination of knowledge and ideas. The printing press, a mechanical device that allowed for the mass production of printed materials, was a _____ for the spread of literacy and education, enabling the widespread dissemination of written works and facilitating the exchange of ideas across geographical boundaries. Answer Choices: A) stagnant B) deflate C) inhibitor D) catalyst	Prompt: Is catalyst the most logical and precise word in following paragraph? Response should be, answer: yes, or answer: no. Here is the paragraph: The invention of the … Answer: yes. The word "catalyst" is the most logical and precise word in this paragraph because it accurately describes the role of the printing press in accelerating the spread of literacy and education. The word "catalyst" implies a triggering or accelerating effect, which is exactly what the printing press did in facilitating the dissemination of knowledge and ideas. Other words, such as "driver", "engine", or "factor", might also convey the idea, but "catalyst" is…..	The word "catalyst" describes something that accelerates or initiates significant change without itself being altered. The paragraph appropriately applies "catalyst" to the printing press, as its invention drastically accelerated the spread of literacy and knowledge during the Renaissance. The supporting details, how the printing press allowed for mass production and dissemination of books, reinforce why the term "catalyst" is well-suited in this context. Answer choices include a correct choice, which is ‘catalyst’. An antonym, which is ‘inhibitor’. Two other unrelated words that have lower similarity score against the genre: 1/ ‘stagnant’ with a similarity score of 0.43 2/ ‘deflate’ with a similarity score of 0.44.
Sample #3: Genre: African tribal huts; word: dilapidated Prompt: Given the meaning of the word dilapidated: a word that is used to describe something that is in a state of disrepair or decay; Generate a paragraph on african tribal huts with the word dilapidated in it. Make sure your paragraph is one single paragraph that is formally worded. If you are done generating the paragraph, stop. Make sure to use the given word dilapidated as is and only use it once. Response: The African tribal huts, once a symbol of vibrant	Prompt: Is dilapidated the most logical and precise word in following paragraph? Response should be, answer: yes, or answer: no. Here is the paragraph: The African tribal huts… Answer: yes. The word "dilapidated" accurately describes the state of the huts, which are worn out, weathered, and in disrepair. It is a precise and logical choice for this context. The word effectively conveys the idea that the huts are no longer in good condition and	The word "dilapidated" refers to something in a state of disrepair due to aging. The paragraph uses "dilapidated" properly by describing the deteriorating huts with vivid details, frayed roofs, cracked wooden frames, and faded colors, showing a strong representation of decay. The context aligns perfectly with the meaning of "dilapidated," making this a well-constructed and accurate usage of the word. Answer choices include a correct choice, which is ‘dilapidated’. An antonym, which is ‘refurbished’. Two other unrelated words that have lower

culture and community, now stood as _____ structures, their thatched roofs worn and frayed, their wooden frames weathered and cracked. The once-vibrant colors that adorned the huts had faded, leaving behind a dull, muted palette that seemed to reflect the decline of the tribe's fortunes. Answer Choices: A) expunge B) dilapidated C) vindicate D) refurbished		similarity score against the genre: 1/ 'expunge' with a similarity score of 0.39 2/ 'vindicated' with a similarity score of 0.41.

Table 10 shows findings from the RAG testing upon running 100 test cases and analyzing the results. The accuracy improved from 62% obtained with Zero-shot, to 74%. Hence, passing the meaning of the word in the paragraph significantly improved the accuracy of the QAG.

Table 10. Summary of RAG Tests

Word Included	Word Not Included	Word in a Different Form	Word Inclusion (%)	Valid Paragraph	Hallucination Count # (%)	Precision	Accuracy
89	11	7	89%	74	15 (20%)	0.83	74%

4. Experiment results

Upon analyzing the results, in the Baseline test, the model hallucinated when it was forced to generate content that it had not previously encountered during pre-training. This meant that it generated a plausible-looking answer but an incoherent paragraph because it lacked grounding in real-world data for that specific prediction. This hallucination originated from the model "guessing" patterns based on its pre-trained data.

From a semantic perspective, the model was unable to incorporate words and genres that were not common. As an example, the model was asked to generate a paragraph on the Trojan war with the word 'whimsical' in it. Its response was *"The Trojan War, a legendary conflict from ancient Greek mythology, was a whimsical tale of love, betrayal, and war…"*. A war cannot be called whimsical, but most humans can be creative enough to use whimsical to describe another noun. The model however, failed to do that, calling either the war, or the gods of Olympus, whimsical. On the other hand, the words and genres that were semantically easy to mesh, had a high success rate across all the test cases. The relation between the words and genres used directly correlated with hallucinations; however, this still varied significantly between trials. Zero-shot and Few-shot prompts only provided the model with instructions on the format of the output. They did not provide context that aided

the quality of the model output. When the model was provided with more context through RAG, the accuracy improved to 74%.

At times, the word was split into many different tokens (sub words). For example, the word "assiduous" was prone to split, whereas "adept" was tokenized as a single token. Modern LLMs, including Llama, uses byte-pair encoding (BPE) to understand words that they are not originally trained on. This exhibits a remarkable ability to recover word meaning (19). Sub word tokenization is therefore, not a concern.

Of the In-Context learning techniques, RAG outperformed other approaches with an accuracy of 74%. This was primarily because of feeding the model with the meaning of the word. Accuracy is anticipated to vary depending on the how closely the word and genre are related. If they are related, the accuracy is expected to greater, and vice versa. In other words, the higher the accuracy, more are the valid paragraphs that are generated.

Table 11 shows how various in-context learning approaches compare. Based on the results obtained, RAG performs better with Accuracy of 74% and precision of 0.83. In other words, RAG is effective in producing more valid paragraphs compared to other approaches. Results are shown visually in Figure 9.

Table 11. Summary of all the tests. The RAG Test used the open-source Llama 3 8B model coupled with the Retrieval Augmented Generation (RAG) prompt and the Semantic Similarity (SS) algorithm.

	Word Included	Word Not Included	Word in a Different Form	Word Inclusion (%)	Valid Paragraph	Hallucination Count # (%)	Precision	Accuracy
Baseline	68	32	2	68%	47	21 (45%)	0.69	47%
Zero-shot	73	27	1	73%	62	11 (18%)	0.85	62%
Few-shot	93	7	1	93%	60	33 (55%)	0.65	60%
RAG	89	11	7	89%	74	15 (20%)	0.83	74%

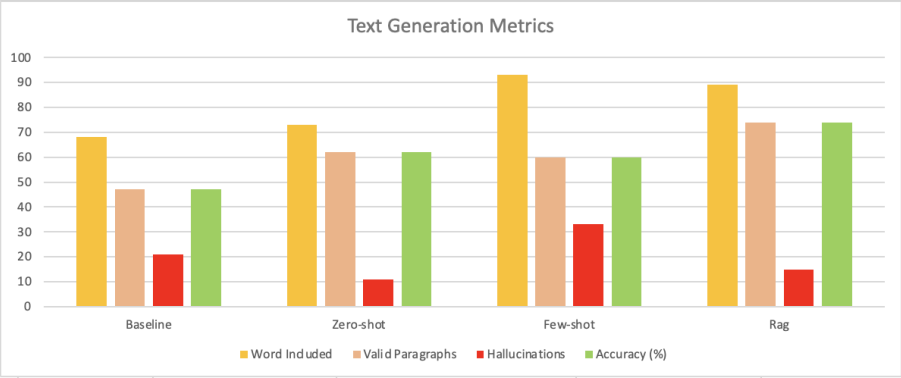

Figure 9. Output Metrics shows RAG + SS outperforming other techniques.

5. Discussion

Non-AI approaches are available for generating such QAG content. For example, books can be downloaded from http://gutenberg.org and programmatically searched for paragraphs that include typical SAT words. This approach will be able to provide 100% generation accuracy (provided the sourced content is accurate) but there are a few limitations to this approach. Books tend to spread relevant context across many paragraphs or pages, and it may be a complex task to build a cohesive paragraph with the necessary context. College Board uses various genres for the questions and searching for such content in books or articles can be convoluted. Associating the word to a genre helps with memorization. Leveraging Gen AI for Words in Context may hence be more advantageous over non-AI approaches.

Commercial Gen AI models such as Chat GPT and Perplexity provide good web-based experience with abstracted pre-processing and post-processing logic. These models make web-based access free. However, API access is charged. Since API access is not free, it becomes cost prohibitive for generating content at scale for such use cases. The key differentiating factors of open-source AI based over commercial models are, 1) cost - for this research, upwards of 6000 paragraphs were generated in total for different experiments and trials, and each paragraph required 7 calls to the model. Leveraging ChatGPT API would have costed $2300 vs $500 that is cost using Llama on A100 GPU. 2) flexibility - The code base uses Hugging Face libraries and users can easily try out other popular open-source models to generate content with minimal changes such as changing the model's name, authentication token, etc.

Results from the open-source Llama 3 8B model coupled with the Retrieval Augmented Generation (RAG) prompt and the Semantic Similarity (SS) algorithm show promise for Words in Context type question generation. The approach used in this paper has broader implications within the world of education. It can be used to create questions in any humanitarian field, lessening the burden on educators by reducing their manual work. Moreover, the use of semantic similarity to manipulate prompts based on the users' requirements can greatly reduce hallucinations in the model. Specifically in smaller models, innovative prompt fine tuning such as using semantic similarity, can make them more viable for such use cases.

Fine-Tuning of the model is unlikely to produce better results for this use case because an exhaustive – bordering on endless – permutations and combinations of genre and words will be required in the fine-tuning data. Instead of forcing the model to learn odd combinations, using the RAG with Semantic Similarity approach is recommended for choosing the word and genre combination.

Given that the SAT has different question types in the Reading/Writing section, such as "Information and Ideas", "Expression of Ideas", etc., there is more that can be explored using open-source Gen AI models. Generation of these questions will bring more innovation and make SAT studying far more accessible and inexpensive. This can also extend to many other fields like employee training, ethics training and other general tests.

6. Perspectives

Codebase and the artifacts that were created as part of this research are published to the following Git repository (14) and is flexible for solution providers to leverage using their own words and genres. Upon running the code, a JSON file is created with paragraphs, answer choices, etc., that can be rendered through a website or mobile app. A screenshot of the website that was created as part of this research is shown in Figure 10.

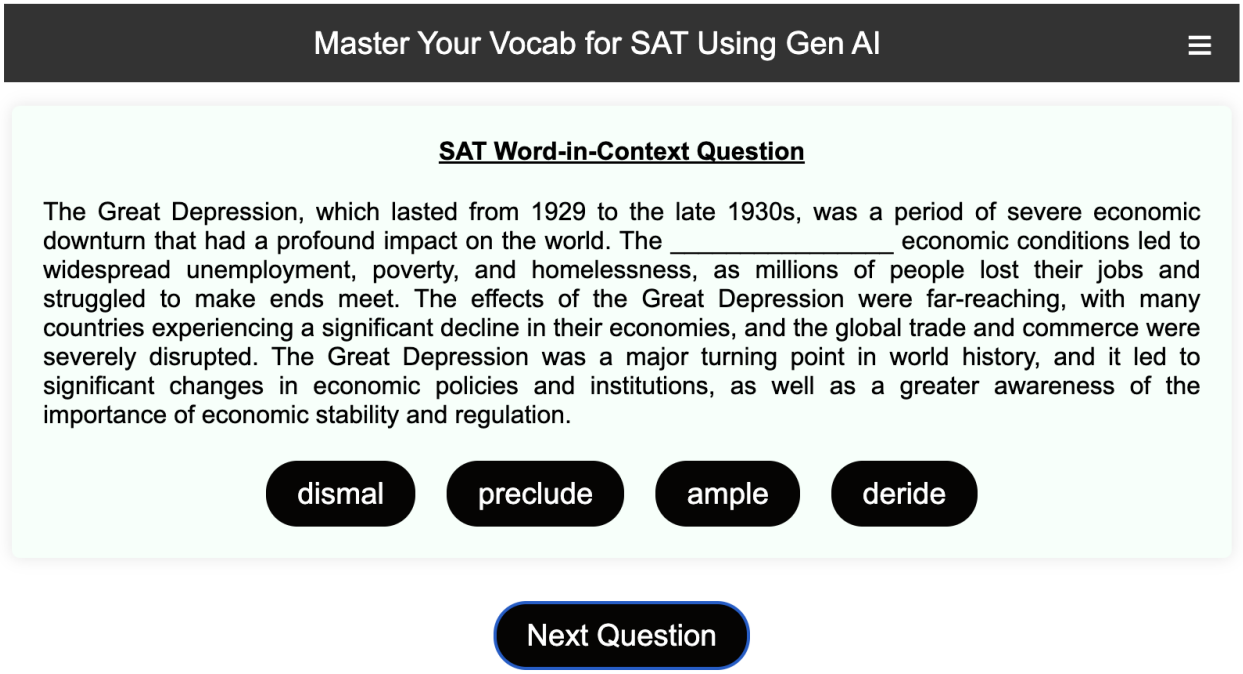

Figure 10. Example of an application (https://www.acesat.ai) that utilizes the approach outlined in this research.

7. Limitations

Despite the improvement in accuracy and precision using the open-source Llama 3 8B model coupled with the Retrieval Augmented Generation (RAG) prompt and the Semantic Similarity (SS) algorithm, evaluation still required human involvement. Automated reasoning helps human assessment but for the approach to scale to larger QAG volumes, automated evaluation will be ideal, which was not achieved using this model. The Llama 3 Instruction model is an instruction fine-tuned version of the base model, and at times, responses included unwarranted sentences such as, “here is your output:”. This required post processing. In some cases, the model responded with the given word in a different form. For example, when the model was asked to use the word ‘reliant’, the model instead responded with the word ‘reliance’. In such cases, post processing is required. In a few cases, the paragraph did not end satisfactorily and in the post processing, the last sentence that was not fully formed needed to be removed. In this paper, distractor answer choices were generated by random word

selection having a lower similarity score. Instead, incorporating words that go with the theme (genre) of the paragraph but not a perfect fit would be ideal, as they would not represent 'obviously incorrect' choices to the test-taker. For example, in Table 1, the correct choice was the word 'melodic'. 'Jarring' is the antonym and two other words, 'personal' and 'scholarly', blend well, yet they are not the correct choice and can be eliminated by intelligent guessing. In contrast, the answer choices generated by the model for Sample 2 in Table 11, consist of the correct answer choice, i.e. 'catalyst'. The antonym is 'inhibit'. The two other distractor or unrelated choices are 'stagnant' and 'deflate'. Since, the distractor words in this case are consistent with the paragraph theme, they cannot be eliminated outright by guessing. In those instances where the word precedes an article such as a, an, etc., the article will have to be included in the answer choices so that users cannot easily guess. This is a post processing step that needs to be added.

8. Conclusion

Experiments were performed to answer the research question: can Llama 3, an open-source LLM, generate SAT style 'Words in Context' type questions with necessary accuracy and sophistication? Based on the findings, the open-source Llama 3 8B model coupled with the Retrieval Augmented Generation (RAG) prompt and the Semantic Similarity (SS) algorithm was found to be effective for this use case with a precision of 0.83 and an accuracy of 74%. The model reduced the time required for generating content, which, in turn, reduced the manual effort required in creating the content. From a content evaluation perspective, there is still an element of manual evaluation that is required. However, this can also be aided by automated reasoning. The model captures the necessary sophistication needed for SAT-style questions, evidenced through manual assessment. In post processing, paragraphs that do not include words are removed, and further processing can be performed to automate evaluation. While the results are promising, future work should explore Fine-Tuning approaches. The model needs more targeted fine-tuning, especially in training the model to integrate words naturally into passages for a given word and genre combination. Structuring the training data more effectively, showing an example of usage of the word based on the genre, could further enhance coherence and paragraph sophistication. The success of this research in generating 'Words in Context' questions show promise for leveraging open-source Gen AI for other types of questions in the SAT Reading and Writing section.

9. Acknowledgements

I would like to thank my mentor, Ellie Fassman, a PhD student at Cornell University for her valuable guidance on this research.

10. References

1. Grattafiori A., Dubey A., Jauhri A., Pandey A., Kadian A., Al-Dahle A., et al. (2024). "The llama 3 herd of Models". arXiv: 2407.21783v3 [cs.AI], 1-3. https://doi.org/10.48550/arXiv.2407.21783

2. The Reading and Writing Section – SAT Suite | College Board. https://satsuite.collegeboard.org/sat/whats-on-the-test/reading-writing

3. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., et al. (2023). "Attention is all you need". arXiv:1706.03762v7 [cs.CL], 1-10. https://doi.org/10.48550/arXiv.1706.03762

4. Bhowmick A. K., Jagmohan A., Vempaty A., Dey P., Hall L., Hartman J., et al. (2023). "Automating question generation from educational text". arXiv:2309.15004 [cs.CL], 1-4. https://doi.org/10.48550/arXiv.2309.15004

5. College Board: *Digital Sat sample questions and explanations (Page 5, RW question 6)* https://satsuite.collegeboard.org/media/pdf/digital-sat-sample-questions.pdf

6. Heilman M., Smith A. N. (2010). "Good Question! Statistical Ranking for Question Generation". "Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL", 609–616. Aclanthology.org. https://aclanthology.org/N10-1086.pdf

7. Devlin J., Chang M.W., Lee K., Toutanova K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2 [cs.CL], 1-9, https://arxiv.org/pdf/1810.04805

8. Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". Journal of Machine Learning Research 21. arXiv:1910.10683v4 [cs.LG], 1-6. https://arxiv.org/pdf/1910.10683

9. Brown T. B., Mann B., Ryder N., Subbiah M. (2020). "Language Models are Few-Shot Learners". arXiv:2005.14165v4 [cs.CL], 1-8. https://arxiv.org/pdf/2005.14165

10. Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y., et al. (2024). "Retrieval-Augmented Generation for Large Language Models: A Survey". arXiv:2312.10997v5 [cs.CL], 1-5,14-17. https://arxiv.org/pdf/2312.10997

11. Reimers N., Gurevych I. (2019). "Sentence-BERT: Sentence embeddings using Siamese Bert-Networks".arXiv:1908.10084 [cs.CL], 1-2. https://doi.org/10.48550/arXiv.1908.10084

12. Nvidia A100 tensor core GPU architecture. NVIDIA A100 Tensor Core GPU Architecture. (2023). https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf

13. The 1000 most common sat words. The 1000 Most Common SAT Words. https://img.sparknotes.com/content/testprep/pdf/sat.vocab.pdf

14. Ramamoorthy, P. https://github.com/PragyanR/GenAI_for_SAT_Prep/tree/main

15. Papineni K., Roukos S., Ward T., Zhu W. J. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation". "Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318". Aclanthology.org. https://aclanthology.org/P02-1040.pdf

16. Sahoo P., Singh A. K., Saha S., Mondal S. (2025). "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications". arXiv:2402.07927v2 [cs.AI], 1-3,7. https://arxiv.org/pdf/2402.07927

17. Introducing Meta Llama 3: The most capable openly available LLM to date. AI at Meta. https://ai.meta.com/blog/meta-llama-3/

18. Xiao S., Liu Z., Zhang P. (2024). "C-Pack: Packed Resources For General Chinese Embeddings". arXiv:2309.07597v5 [cs.CL], 1-3. https://arxiv.org/pdf/2309.07597

19. Kaplan G., Oren M., Reif Y., Schwartz R. (2024). "From Tokens to Words: On the Inner Lexicon of LLMs. From tokens to words: On the inner lexicon of llms". arXiv:2410.05864v2 [cs.CL], 1-5. https://arxiv.org/html/2410.05864v2

20. Words in context and command of evidence (Page 13). https://satsuite.collegeboard.org/media/pdf/redesigned-sat-k12-teacher-training-facilitators-guide-2-words-context-command-evidence.pdf