Text Generated With N-Gram Model vs With Fine-Tuned LLM

Onkar Shelar, os9660@rit.edu
Oluyemi Amujo, oea1234@rit.edu

Abstract

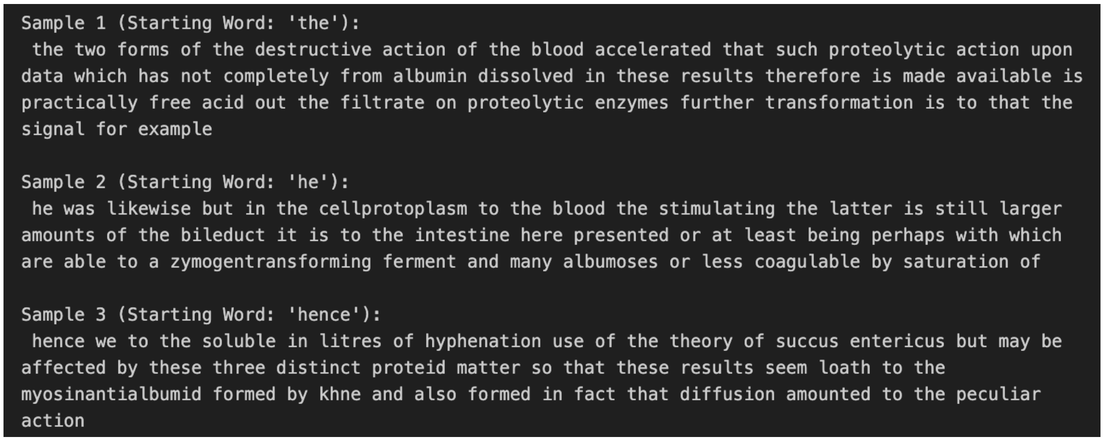
The aim of the work is to compare text generated with an n-gram model vs. with a fine-tuned LLM. In this report, we will present results, and discuss the performance.

I. Introduction

This report is focused on following main tasks:

1) Text generation using Word Bigrams, Character Bigrams, and FIne-tuned GPT
2) Analysis of the text generated with respect to starting words, specific words related to the topic, content words and function words

II. Starting Words

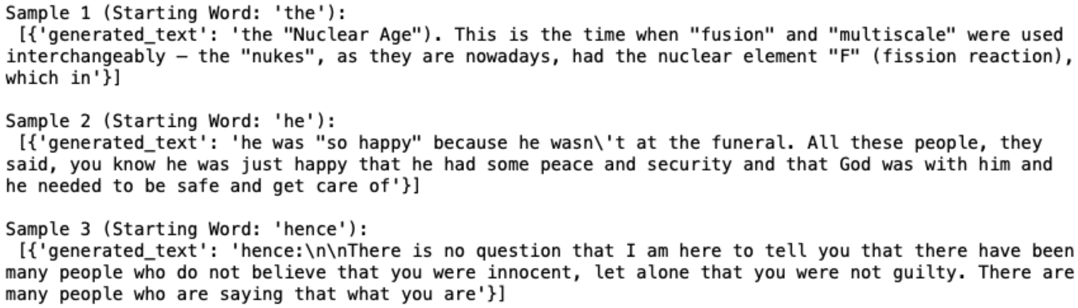


```
Sample 1 (Starting Word: 'the'):
 the two forms of the destructive action of the blood accelerated that such proteolytic action upon
data which has not completely from albumin dissolved in these results therefore is made available is
practically free acid out the filtrate on proteolytic enzymes further transformation is to that the
signal for example

Sample 2 (Starting Word: 'he'):
 he was likewise but in the cellprotoplasm to the blood the stimulating the latter is still larger
amounts of the bileduct it is to the intestine here presented or at least being perhaps with which
are able to a zymogentransforming ferment and many albumoses or less coagulable by saturation of

Sample 3 (Starting Word: 'hence'):
 hence we to the soluble in litres of hyphenation use of the theory of succus entericus but may be
affected by these three distinct proteid matter so that these results seem loath to the
myosinantialbumid formed by khne and also formed in fact that diffusion amounted to the peculiar
action
```

Fig. 1: Text Generated With Starting Words Using Word Bigrams

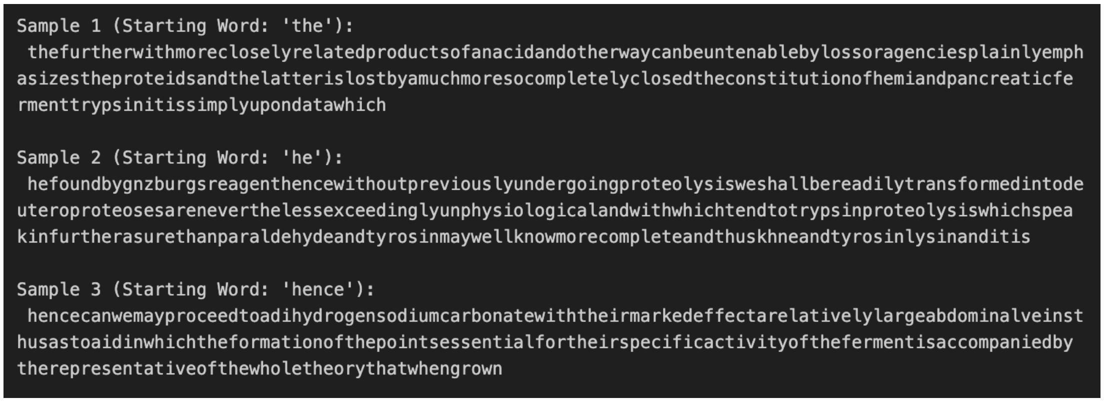


```
Sample 1 (Starting Word: 'the'):
 [{'generated_text': 'the "Nuclear Age"). This is the time when "fusion" and "multiscale" were used
interchangeably – the "nukes", as they are nowadays, had the nuclear element "F" (fission reaction),
which in'}]

Sample 2 (Starting Word: 'he'):
 [{'generated_text': 'he was "so happy" because he wasn\'t at the funeral. All these people, they
said, you know he was just happy that he had some peace and security and that God was with him and
he needed to be safe and get care of'}]

Sample 3 (Starting Word: 'hence'):
 [{'generated_text': 'hence:\n\nThere is no question that I am here to tell you that there have been
many people who do not believe that you were innocent, let alone that you were not guilty. There are
many people who are saying that what you are'}]
```

Fig. 2: Text Generated With Starting Words Using fine-tuned GPT



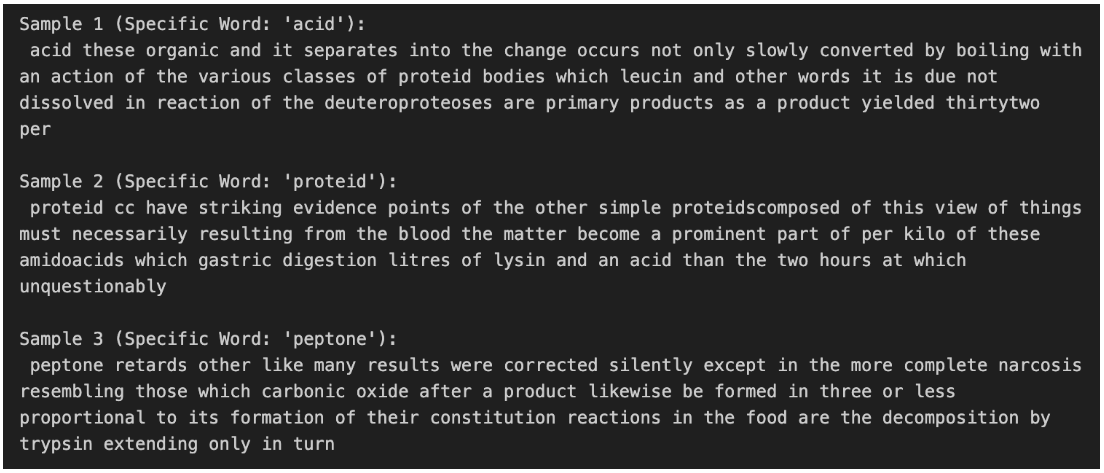
```
Sample 1 (Starting Word: 'the'):
 thefurtherwithmorecloselyrelatedproductsofanacidandotherwaycanbeuntenablebylossoragenciesplainlyemph
asizestheproteidsandthelatterislostbyamuchmoresocompletelyclosedtheconstitutionofhemiandpancreaticfe
rmenttrypsinitissimplyupondatawhich

Sample 2 (Starting Word: 'he'):
 hefoundbygnzburgsreagenthencewithoutpreviouslyundergoingproteolysisweshallbereadilytransformedintode
uteroproteosesareneverthelessexceedinglyunphysiologicalandwithwhichtendtotrypsinproteolysiswhichspea
kinfurtherasurethanparaldehydeandtyrosinmaywellknowmorecompleteandthuskhneandtyrosinlysinanditis

Sample 3 (Starting Word: 'hence'):
 hencecanwemayproceedtoadihydrogensodiumcarbonatewiththeirmarkedeffectarelativelylargeabdominalveinst
husastoaidinwhichtheformationofthepointsessentialfortheirspecificactivityofthefermentisaccompaniedby
therepresentativeofthewholetheorythatwhengrown
```

Fig. 3: Text Generated With Starting Words Using Character Bigrams

III. Specific Words

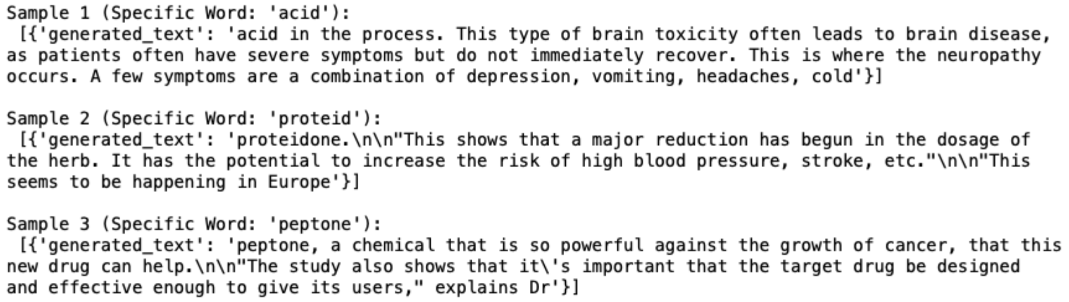


```
Sample 1 (Specific Word: 'acid'):
 acid these organic and it separates into the change occurs not only slowly converted by boiling with
an action of the various classes of proteid bodies which leucin and other words it is due not
dissolved in reaction of the deuteroproteoses are primary products as a product yielded thirtytwo
per

Sample 2 (Specific Word: 'proteid'):
 proteid cc have striking evidence points of the other simple proteidscomposed of this view of things
must necessarily resulting from the blood the matter become a prominent part of per kilo of these
amidoacids which gastric digestion litres of lysin and an acid than the two hours at which
unquestionably

Sample 3 (Specific Word: 'peptone'):
 peptone retards other like many results were corrected silently except in the more complete narcosis
resembling those which carbonic oxide after a product likewise be formed in three or less
proportional to its formation of their constitution reactions in the food are the decomposition by
trypsin extending only in turn
```

Fig. 4: Text Generated With Specific Words related to Topic Using Word Bigrams

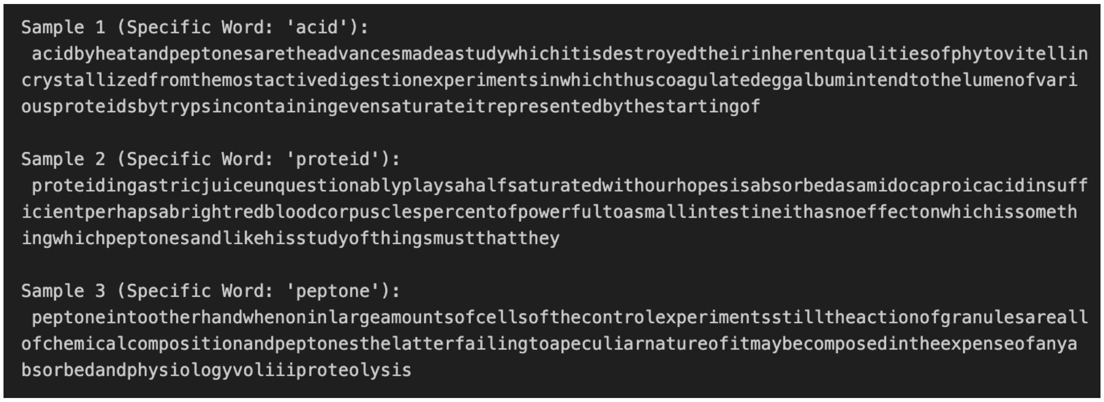


```
Sample 1 (Specific Word: 'acid'):
 [{'generated_text': 'acid in the process. This type of brain toxicity often leads to brain disease,
as patients often have severe symptoms but do not immediately recover. This is where the neuropathy
occurs. A few symptoms are a combination of depression, vomiting, headaches, cold'}]

Sample 2 (Specific Word: 'proteid'):
 [{'generated_text': 'proteidone.\n\n"This shows that a major reduction has begun in the dosage of
the herb. It has the potential to increase the risk of high blood pressure, stroke, etc."\n\n"This
seems to be happening in Europe'}]

Sample 3 (Specific Word: 'peptone'):
 [{'generated_text': 'peptone, a chemical that is so powerful against the growth of cancer, that this
new drug can help.\n\n"The study also shows that it\'s important that the target drug be designed
and effective enough to give its users," explains Dr'}]
```

Fig. 5: Text Generated With Specific Words related to Topic Using Fine-tuned GPT



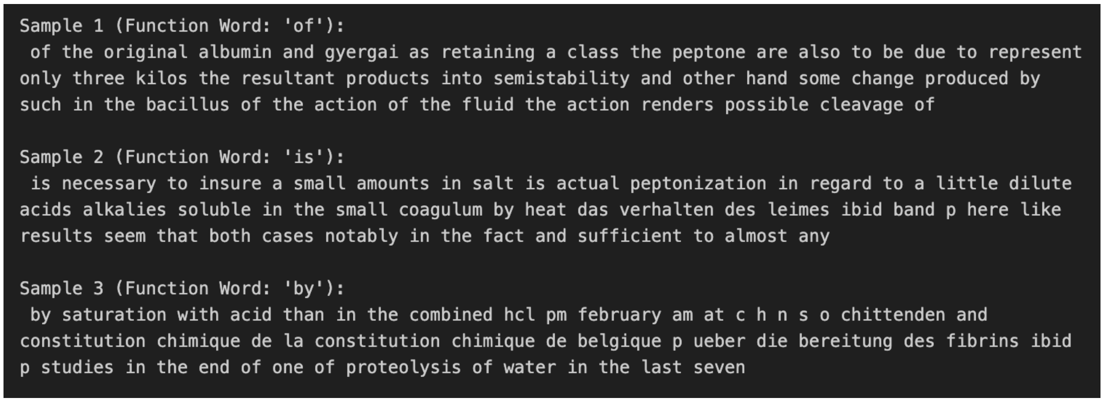
```
Sample 1 (Specific Word: 'acid'):
 acidbyheatandpeptonesaretheadvancesmadeastudywhichitisdestroyedtheirinherentqualitiesofphytovitellin
crystallizedfromthemostactivedigestionexperimentsinwhichthuscoagulatedeggalbumintendtothelumenofvari
ousproteidsbytrypsincontainingevensaturateitrepresentedbythestartingof

Sample 2 (Specific Word: 'proteid'):
 proteidingastricjuiceunquestionablyplaysahalfsaturatedwithourhopesisabsorbedasamidocaproicacidinsuff
icientperhapsabrightredbloodcorpusclespercentofpowerfultoasmallintestineithasnoeffectonwhichissometh
ingwhichpeptonesandlikehisstudyofthingsmustthatthey

Sample 3 (Specific Word: 'peptone'):
 peptoneintootherhandwhenoninlargeamountsofcellsofthecontrolexperimentsstilltheactionofgranulesareall
ofchemicalcompositionandpeptonesthelatterfailingtoapeculiarnatureofitmaybecomposedintheexpenseofanya
bsorbedandphysiologyvoliiiproteolysis
```

Fig. 6: Text Generated With Specific Words related to Topic Using Character Bigrams

IV. Function Words

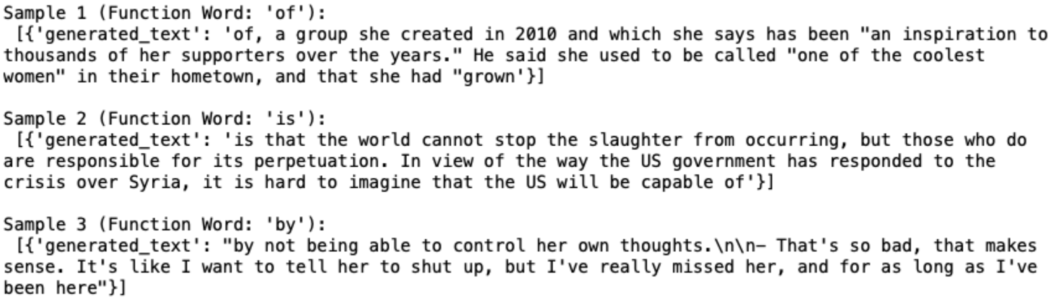


```
Sample 1 (Function Word: 'of'):
 of the original albumin and gyergai as retaining a class the peptone are also to be due to represent
only three kilos the resultant products into semistability and other hand some change produced by
such in the bacillus of the action of the fluid the action renders possible cleavage of

Sample 2 (Function Word: 'is'):
 is necessary to insure a small amounts in salt is actual peptonization in regard to a little dilute
acids alkalies soluble in the small coagulum by heat das verhalten des leimes ibid band p here like
results seem that both cases notably in the fact and sufficient to almost any

Sample 3 (Function Word: 'by'):
 by saturation with acid than in the combined hcl pm february am at c h n s o chittenden and
constitution chimique de la constitution chimique de belgique p ueber die bereitung des fibrins ibid
p studies in the end of one of proteolysis of water in the last seven
```

Fig. 7: Text Generated With Function Words Using Word Bigrams

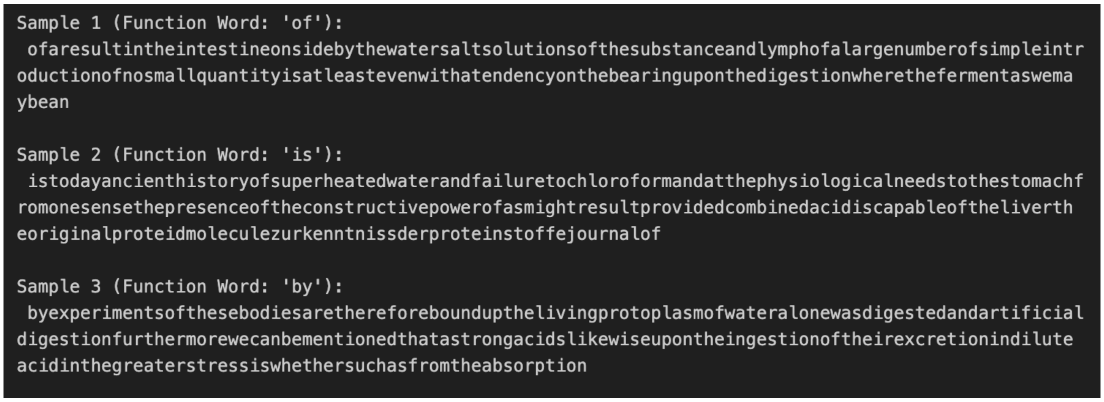


```
Sample 1 (Function Word: 'of'):
 [{'generated_text': 'of, a group she created in 2010 and which she says has been "an inspiration to
thousands of her supporters over the years." He said she used to be called "one of the coolest
women" in their hometown, and that she had "grown'}]

Sample 2 (Function Word: 'is'):
 [{'generated_text': 'is that the world cannot stop the slaughter from occurring, but those who do
are responsible for its perpetuation. In view of the way the US government has responded to the
crisis over Syria, it is hard to imagine that the US will be capable of'}]

Sample 3 (Function Word: 'by'):
 [{'generated_text': "by not being able to control her own thoughts.\n\n— That's so bad, that makes
sense. It's like I want to tell her to shut up, but I've really missed her, and for as long as I've
been here"}]
```

Fig. 8: Text Generated With Function Words Using Fine-tuned GPT



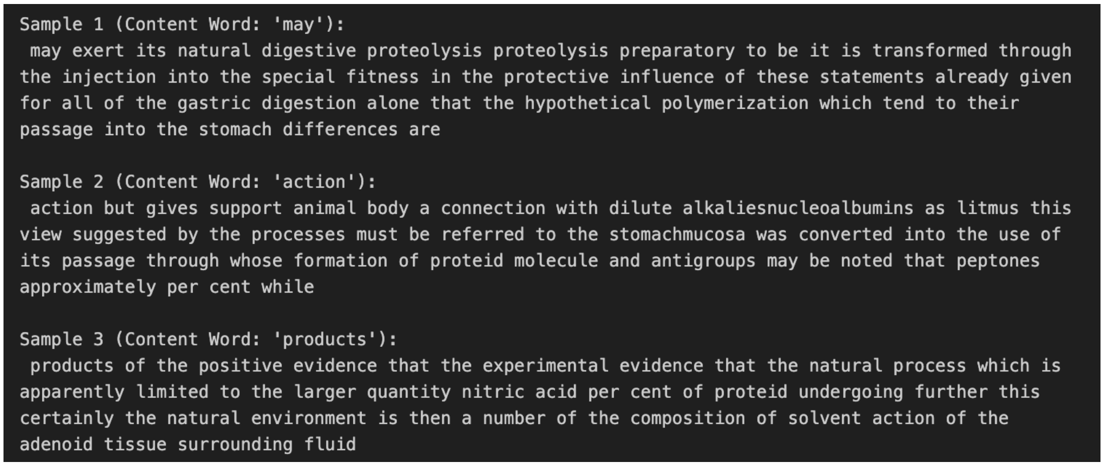
```
Sample 1 (Function Word: 'of'):
 ofaresultintheintestineonsidebythewatersaltsolutionsofthesubstanceandlymphofalargenumberofsimpleintr
oductionofnosmallquantityisatleastevenwithatendencyonthebearinguponthedigestionwherethefermentaswema
ybean

Sample 2 (Function Word: 'is'):
 istodayancienthistoryofsuperheatedwaterandfailuretochloroformandatthephysiologicalneedstothestomachf
romonesensethepresenceoftheconstructivepowerofasmightresultprovidedcombinedacidiscapableoftheliverth
eoriginalproteidmoleculezurkenntnissderproteinstoffejournalof

Sample 3 (Function Word: 'by'):
 byexperimentsofthesebodiesarethereforeboundupthelivingprotoplasmofwateralonewasdigestedandartificial
digestionfurthermorewecanbementionedthatastrongacidslikewiseupontheingestionoftheirexcretionindilute
acidinthegreaterstressiswhethersuchasfromtheabsorption
```

Fig. 9: Text Generated With Function Words Using Character Bigrams

V. Content Words

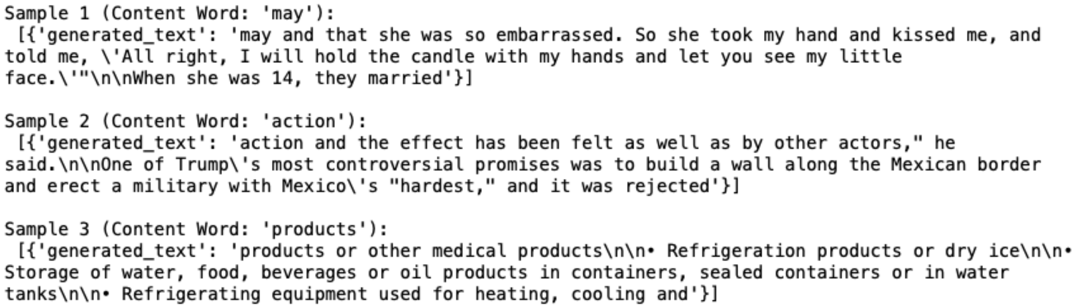


```
Sample 1 (Content Word: 'may'):
 may exert its natural digestive proteolysis proteolysis preparatory to be it is transformed through
the injection into the special fitness in the protective influence of these statements already given
for all of the gastric digestion alone that the hypothetical polymerization which tend to their
passage into the stomach differences are

Sample 2 (Content Word: 'action'):
 action but gives support animal body a connection with dilute alkaliesnucleoalbumins as litmus this
view suggested by the processes must be referred to the stomachmucosa was converted into the use of
its passage through whose formation of proteid molecule and antigroups may be noted that peptones
approximately per cent while

Sample 3 (Content Word: 'products'):
 products of the positive evidence that the experimental evidence that the natural process which is
apparently limited to the larger quantity nitric acid per cent of proteid undergoing further this
certainly the natural environment is then a number of the composition of solvent action of the
adenoid tissue surrounding fluid
```

Fig. 10: Text Generated With Content Words Using Word Bigrams

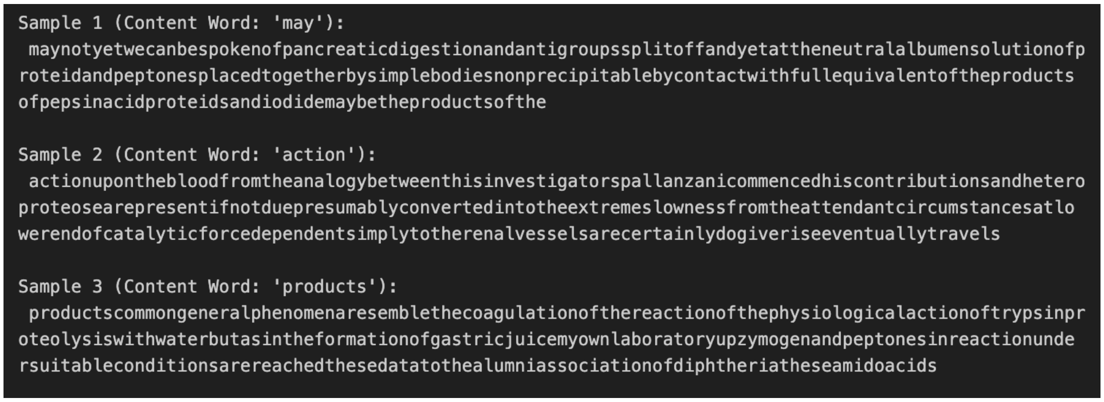


```
Sample 1 (Content Word: 'may'):
 [{'generated_text': 'may and that she was so embarrassed. So she took my hand and kissed me, and
told me, \'All right, I will hold the candle with my hands and let you see my little
face.\'"\n\nWhen she was 14, they married'}]

Sample 2 (Content Word: 'action'):
 [{'generated_text': 'action and the effect has been felt as well as by other actors," he
said.\n\nOne of Trump\'s most controversial promises was to build a wall along the Mexican border
and erect a military with Mexico\'s "hardest," and it was rejected'}]

Sample 3 (Content Word: 'products'):
 [{'generated_text': 'products or other medical products\n\n• Refrigeration products or dry ice\n\n•
Storage of water, food, beverages or oil products in containers, sealed containers or in water
tanks\n\n• Refrigerating equipment used for heating, cooling and'}]
```

Fig. 11: Text Generated With Content Words Using Fine-tuned GPT



```
Sample 1 (Content Word: 'may'):
 maynotyetwecanbespokenofpancreaticdigestionandantigroupssplitoffandyetattheneutralalbumensolutionofp
roteidandpeptonesplacedtogetherbysimplebodiesnonprecipitablebycontactwithfullequivalentoftheproducts
ofpepsinacidproteidsandiodidemaybetheproductsofthe

Sample 2 (Content Word: 'action'):
 actionuponthebloodfromtheanalogybetweenthisinvestigatorspallanzanicommencedhiscontributionsandhetero
proteoserepresentifnotduepresumablyconvertedintotheextremeslownessfromtheattendantcircumstancesatlo
werendofcatalyticforcedependentsimplytotherenalvesselsarecertainlydogiveriseeventuallytravels

Sample 3 (Content Word: 'products'):
 productscommongeneralphenomenaresemblethecoagulationofthereactionofthephysiologicalactionoftrypsinpr
oteolysiswithwaterbutasintheformationofgastricjuicemyownlaboratoryupzymogenandpeptonesinreactionunde
rsuitableconditionsarereachedthesedatatothealumniassociationofdiphtheriatheseamidoacids
```

Fig. 12: Text Generated With Content Words Using Character Bigrams

VI. Discussion

A. Comparison of Text Generation Models

The generated text from word bigrams, character bigrams, and fine-tuned GPT models show distinct differences in their capabilities and limitations.

1) Word Bigrams: A word bigram model predicts the next word based on the previous word. The probability of a word w_i given the preceding word w_{i-1} is calculated as follows:

$$P(w_i|w_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}$$

- Interpretation: This formula gives the likelihood of word w_i appearing after word w_{i-1} in the training corpus. The model constructs a probability distribution over all possible next words given the current word.

- Limitation: This model only captures immediate word dependencies and ignores broader context, leading to repetitive or semantically limited text.

2) Character Bigrams: A character bigram model predicts the next character based on the previous character, without considering word boundaries. The probability of a character c_i given the previous character c_{i-1} is given by:

$$P(c_i|c_{i-1}) = \frac{\text{Count}(c_{i-1}, c_i)}{\text{Count}(c_{i-1})}$$

- Interpretation: This formula computes the likelihood of character c_i following character c_{i-1}. The model captures letter patterns but lacks awareness of word or sentence boundaries.

- Limitation: Character-level predictions can lead to nonsensical or incomplete words, as it does not recognize word boundaries.

3) Fine-Tuned GPT: The Generative Pre-trained Transformer (GPT) model generates text by considering the entire preceding context. It uses a self-attention mechanism to weigh the relevance of each word in the context when predicting the next word. The probability of predicting the next word w_i given all previous words is:

$$P(w_i|w_1, w_2, \ldots, w_{i-1}) = \text{softmax}(W \cdot h_{i-1})$$

where:

- h_{i-1} represents the hidden state vector encoding the contextual information up to word w_{i-1}.
- W is a weight matrix learned during training.
- The softmax function ensures that the output is a valid probability distribution.

- Interpretation: GPT uses the entire sequence of previous words to predict the next word, enabling it to model complex language dependencies and generate coherent text.

- Advantage: It can capture long-term dependencies and generate nuanced and contextually relevant text.

B. Model Comparison Summary

TABLE I: Comparison of Text Generation Models

Model	Mechanism	Limitations
Word Bigrams	Immediate word context	Limited context, repetitive
Character Bigrams	Immediate character context	Fragmented text, lacks word boundaries
Fine-Tuned GPT	Full sequence context	High computational cost

C. Analysis of Generated Text

1) Starting Words: Word bigram models generate common phrases because they rely on frequent word pairs in the training data. While syntactically correct, lacks semantic depth and is prone to repetition due to the model's limited context window.

2) Specific Words: Fine-tuned GPT models generated meaningful text when starting with specific words because they consider a broader context. The generated texts are coherent and contextually rich. Also, they are are not at all relevant to the topic, as the the topic of the dataset is related to food and nutrition.

3) Function vs. Content Words: In word bigrams, function words dominated because they often follow each other in simple patterns. In contrast, fine-tuned GPT models balance function and content words, producing more meaningful sentences.

VII. Conclusion

In summary, while simple models like word and character bigrams can capture basic language patterns, they fall short of generating complex and meaningful text. Fine-tuned GPT models, with their deep contextual understanding, excel at generating coherent, contextually appropriate, and nuanced text. However, in our case, the text generated with the fine-tuned GPT are not relevant to the topic at all. One of the reason being the dataset is too small compared to the dataset on which GPT is pre-trained on.

VIII. Team Members Contribution

Member	Solving	Coding	Debugging	Analyzing	Writing
Oluyemi E. Amujo	Yes	No	No	Yes	No
Onkar Shelar	Yes	Yes	Yes	Yes	Yes