Generating an N-Gram Language Model

By:

Joshua Souane

Elena Ajayi

Omar Shajib

Jomo Smith

Jordan Hayes

Rosa DiLiberto

**Introduction:**

In this project, our focus was on developing an N-gram Language Model using a web crawler code from the previous project and leveraging the corpus extracted from the St. John’s website. This corpus provided the basis for creating a variety of N-gram models that produced diverse outcomes. Engaging ourselves in N-gram language models with real-world web data deepened our understanding of the language model. The methodology and results derived from each were then outlined and assessed.

Part 1: Language and Model Training

**Model 1: Corpus Only**

*Architecture*

This method employs an N-gram model to generate text continuations from a provided prompt, utilizing a derived corpus and tokenizing it with NLTK library's word\_tokenize function. NLTK's word\_tokenize technique splits the text into individual words while preserving contractions and punctuation marks as separate tokens. We opted for this library over developing our own solution due to time constraints, recognizing NLTK's superiority in efficiency and accuracy.

An N-gram model of a specified order (N) is then built from the tokenized corpus, with N denoting the number of words in each n-gram. The build\_ngram\_model function creates a dictionary where keys represent tuples of N consecutive words (n-grams), and values consist of lists of words following those n-grams in the corpus. Subsequently, the generate\_continuation function takes a prompt as input and generates a continuation by iteratively selecting the next word based on the previous N words using the N-gram model.

In case the N-gram model encounters an n-gram not present in the vocabulary, it ceases text generation. If it fails to generate any continuations, a simple message is returned, read as follows:

*“I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!”*

**Corpus Only Results:**

**Prompt 1:** In the fall semester, the business department offers courses such as

**Continuation:** In the fall semester, the business department offers courses such as sales , purchasing , HR , accounting , assurance , and neither is the Center and increase your proficiency in

**Prompt 2:** To apply for financial aid and scholarships, students must

**Continuation:** To apply for financial aid and scholarships, students must have , as well as through an electronic platform that enables us to offer SJU 's discounts to all St.

**Prompt 3:** to apply for a graduate degree students must

**Continuation:** to apply for a graduate degree students must meet the New York Locations Global Locations Career Development Health & Wellness Student Success Parent & Family Faculty , Administrators

**Prompt 4:** Graduate students have research opportunities in areas like

**Continuation:** I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

**Prompt 5:** On-campus housing options provide amenities such as

**Continuation:** On-campus housing options provide amenities such as dances , sports activities and round table conversations Cultural orientations for the content of these are ; NYC Carbon Challenge

**Prompt 6:** The university's policies on academic integrity include guidelines like

**Continuation:** I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

**Prompt 7:** Students can engage in extracurricular activities such as

**Continuation:** Students can engage in extracurricular activities such as estate , gift , and that the candidate may wish to share with the information they need to navigate and

**Prompt 8:** Support for international students includes services like

**Continuation**: I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

**Prompt 9:** Submitting a thesis or dissertation requires steps such as

**Continuation:** Submitting a thesis or dissertation requires steps such as literature , philosophy , and ethical responsibilities as a student is exposed to new location '' / > < stEvt

**Prompt 10:** The university's career center offers services like

**Continuation:** I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

Immediately evident is the model's inadequacy in context relevant predictions. In the 6/10 prompts that had predicted words, another shortcoming in our model. The only prompt with relevant output was prompt #1. In an effort to enhance the corpus's quality, additional iterations of the n-gram were pursued. The optimizations done for the next n-gram models generated interesting outputs using two different GPT approaches.

**Model 2 : N-gram model with GPT-2 Base**

As a preliminary experiment, a model was constructed using GPT-2 as the foundation, without incorporating the corpus. This was solely aimed at evaluating the nature of output generated in response to the prompts.

*Architecture*

This attempt employs the gpt\_2\_simple library to generate text continuations using the GPT-2 model. The generate\_gpt2\_continuation function, central to the script's functionality, initiates a TensorFlow session and loads the pre-trained GPT-2 model, specifically the 124M version. With this model in place, the function generates a continuation based on a provided prompt, with the option to specify the length of the generated text. For each prompt, the function calls generate\_gpt2\_continuation to generate a continuation, printing both the prompt and its corresponding continuation in the console. Through the execution of the main function, text continuations are seamlessly generated and displayed for each prompt, showcasing the capability of the GPT-2 model.

*GPT-2 Base Results:*

**Prompt 1:** In the fall semester, the business department offers courses such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** In the fall semester, the business department offers courses such as how to drive a car, how to find a right-hand-drive mechanic, how to handle a car with "very low mileage," and how to show drivers what a car does to get around.

"For most people, the car

**Prompt 2:** To apply for financial aid and scholarships, students must

Loading pretrained model models/124M/model.ckpt

**Continuation:** To apply for financial aid and scholarships, students must take a course of study intended to integrate the academic experience of a graduate student in the teaching profession.

The program offers a wide range of opportunities to students who are interested in learning to work on technical and political issues.

Students who want to

**Prompt 3:** to apply for a graduate degree students must

**Continuation:** to apply for a graduate degree students must submit their work and/or papers to the department's Graduate Program in Computer Science.Students with post-secondary degrees must complete the Graduate Program in Information Technology before entering the department.Post-secondary graduates with a bachelor's degree must

**Prompt 4:** Graduate students have research opportunities in areas like

**Continuation:** Graduate students have research opportunities in areas like Environmental Health, Environmental Leadership, Environmental Health, Environmental Conservation, and Environmental Security. Related bachelor's degree in environmental health, environmental governance, environmental governance, environmental governance, environmental governance, environmental governance, environmental governance, environmental governance,

**Prompt 5:** On-campus housing options provide amenities such as

**Continuation:** On-campus housing options provide amenities such as clean air, a breeze, and natural light. But many of the services are also on-campus. "For many young people, seeing their parents, relatives and school or college being impacted by an on-campus housing decision is simply not

**Prompt 6:** The university's policies on academic integrity include guidelines like

**Continuation:** The university's policies on academic integrity include guidelines like "No racial geometry, racial slurs, etc.""We look at what we're doing and we don't think it's fair for students to be discriminated against based on race or sex," said Rana Kondo, executive director of the

**Prompt 7:** Students can engage in extracurricular activities such as

**Continuation:** Students can engage in extracurricular activities such as taking a walk or a run, all while receiving the same level of education and support. Students can also use a computer or mobile device to access information and communicate with others. The 2018-2019 National Education Standards for the All-Aboard

**Prompt 8:** Support for international students includes services like

**Continuation:** Support for international students includes services like the Refugee Admissions Office, where they can help to select students for resettlement in the United States. USC's Refugee Admissions Office is a nonprofit 501(c)(3), non-profit organization that provides services for international students. This

**Prompt 9:** Submitting a thesis or dissertation requires steps such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** Submitting a thesis or dissertation requires steps such as identifying a researcher who is qualified, sought out and well-connected and establishing a budget. In addition, there are some requirements for publication. First, the study should be in English language. Second, the study should not be written in German

**Prompt 10:** The university's career center offers services like

**Continuation:** The university's career center offers services like graduate training, internship and internships that bring you a diverse range of career opportunities. The School of the Arts offers a range of undergraduate and graduate degrees from the arts, humanities and social sciences. Students who seek to be a creative and creative revolutionary must

**Module 3: GPT Base Fine Tuned With Corpus**

*Architecture*

Ultimately, the idea emerged that fine-tuning our GPT-2 model with the corpus obtained from the web crawler could yield more contextually relevant outputs. By integrating this corpus, the model could generate responses closely resembling those that might originate from a language model trained solely on St. John's websites. In essence, this approach enables the artificial generation of responses that closely approximate the style and content typical of St. John's website content.

This code accomplishes two main tasks: fine-tuning the GPT-2 model on a given corpus and generating text continuations based on prompts using the fine-tuned model.

Firstly, the fine\_tune\_gpt2 function takes a path to a corpus file as input and fine-tunes the pre-trained GPT-2 model on this corpus. The number of training steps can be adjusted as needed.

Secondly, the generate\_gpt2\_continuation function generates text continuations based on a provided prompt using the fine-tuned GPT-2 model. The length of the generated text can be specified, with a default length of 50 tokens.

*GPT 2 Base Fined Tuned with Cropus Results:*

Loading checkpoint models/124M/model.ckpt

Loading dataset...

100%|██████████| 1/1 [01:28<00:00, 88.04s/it]

dataset has 17668215 tokens

Training...

[1 | 132.04] loss=7.42 avg=7.42

[2 | 260.36] loss=9.05 avg=8.24

[3 | 395.25] loss=5.36 avg=7.27

[4 | 522.30] loss=6.94 avg=7.19

[5 | 649.74] loss=6.55 avg=7.06

Saving checkpoint/run1/model-5

**Prompt 1:** In the fall semester, the business department offers courses such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** In the fall semester, the business department offers courses such as "Business Basics for Business Education," "Business Learning for Entrepreneurs," "Business Learning for Students," and the "Business Learning for Entrepreneurship and the Business Education of Entrepreneurship."The courses are free and accessible, but students are

**Prompt 2:** To apply for financial aid and scholarships, students must

Loading pretrained model models/124M/model.ckpt

**Continuation:** To apply for financial aid and scholarships, students must seize a state-level job with a proven track record of recruiting and recruiting. They must earn at least 2,500 hours of service in the military, and have been awarded at least 3,000 academic credits. The student needs to have

**Prompt 3:** to apply for a graduate degree students must

Loading pretrained model models/124M/model.ckpt

**Continuation:** to apply for a graduate degree students must have read, written, and discussed at least four courses and completed at least seven days of study. If you have not read, written, or discussed at least three courses and completed at least seven days of study, you may be ineligible to apply.

**Prompt 4:** Graduate students have research opportunities in areas like

Loading pretrained model models/124M/model.ckpt

**Continuation:** Graduate students have research opportunities in areas like bioengineering and engineering in the fields of solar and space sciences. This page is part of the MIT Media Lab's "MIT Newsroom" project.<|endoftext|>We've seen a reduction in the number of children in Canada this year, with marked

**Prompt 5:** On-campus housing options provide amenities such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** On-campus housing options provide amenities such as a warm and inviting setting and a quality dining experience. Student housing has become an essential part of our neighborhoods. Resources and resources The University of Texas is committed to providing a safe, affordable, and comfortable environment for all.

**Prompt 6:** The university's policies on academic integrity include guidelines like

Loading pretrained model models/124M/model.ckpt

**Continuation:** The university's policies on academic integrity include guidelines like the one the university's vice president for university affairs and faculty said were "probably the most stringent" in the nation. An extra semester is allowed, but students are not allowed to take on more than one course.

"It's a very

**Prompt 7:** Students can engage in extracurricular activities such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** Students can engage in extracurricular activities such as sports, music, sporting events and the like.

It's a hypothetical scenario but it's one that may make sense. The NCAA has already denied that it has money to spend on social media.

The NCAA said that free publicity on

**Prompt 8:** Support for international students includes services like

Loading pretrained model models/124M/model.ckpt

**Continuation:** Support for international students includes services like a summer camp, staying with family, and exploring work opportunities.

"We feel we have a good chance at supporting our students and will be working with the university to provide them with the resources they need to succeed in their chosen field," said

**Prompt 9:** Submitting a thesis or dissertation requires steps such as

Loading pretrained model models/124M/model.ckpt

**Continuation:** Submitting a thesis or dissertation requires steps such as an application and a research project. The project is in line with the research objective and the needs of the article.

The Graduate Research Fellowship (GRF) or the Career Opportunities Fellowship (CPRF) are also available in order to

**Prompt 10:** The university's career center offers services like

Loading pretrained model models/124M/model.ckpt

**Continuation:** The university's career center offers services like a practical internship program, technical training, and professional development programs. It also has a student-centered faculty with an emphasis on student learning.The main purpose of the institute is to find the best music and dance courses for students. The heavily

While not flawless, the generated continuations bear closer resemblance to content typically found on St. John's website. It's worth noting that the model was trained for only five epochs, a relatively low number. With further training, say 50 epochs, the results would likely be more finely tuned to the provided corpus, resulting in even more domain-specific responses.

**Part 2: Tokenization Techniques:**

There were three tokenization techniques used to test and find the best n-gram model, word tokenization, character tokenization and word embeddings. Word tokenization generated the best model with the most context relevant output. A shortcoming that was present across all model integrations was that the model could not generate predictions for prompts #4, #6, #8, and #10. This issue is due to one of the *n* words in the n-gram model not being in the vocabulary, therefore the model has no word to generate a proportion on. To solve this issue, we tried the two different techniques, character tokenization and word embeddings.

*Word Tokenization:*

Word tokenization generated our best models. The words generated were words from the corpus and were context relevant, two important factors of a successful n-gram model and were both short coming in our other two tokenization approaches.

*Character Tokenization:*

Our group created a model using character tokenization, therefore having the model predict proportions on characters as tokens rather than words. This solution did not improve our model, in fact it had worse output. Instead of generating next words, it generated single letters, which is not the output we were intending.

*Word Embeddings:*

We attempted to create a model that would use word embeddings to generate the next word of our prompts but decided it would be too complex to suit what our n-gram model needed. We were unable to construct a successful model with word embeddings.

**Part 3: Choosing N**

With our corpus only model, we tested different numbers for *n* to see which n-gram model would generate the best continuation for each prompt. We tested our model at *n=1, n=2, n=3,* a hybrid approach with either *n=2* or *n=3,* and finally a higher order approach with the max of *n=5.* We wanted to test all the different model approaches with lower value n’s as well as higher value’s to see which would lead to better model performance. Our results are as follows:

*n=1 results:*

  Prompt: In the fall semester, the business department offers courses such as

Continuation: In the fall semester, the business department offers courses such as part of professional or potential parental , ( t� �ݵ�1um�� C�� PLQ '' '' �� : �|R $ bk�xv�7���/� ��dRj�N

Prompt: To apply for financial aid and scholarships, students must

Continuation: To apply for financial aid and scholarships, students must have partnered with it. ” 0B ­áÄ¹³ò1Y ­Æ´EYé €d£v X ) K ? MW^ ` ����� ) C��� $ ��Q���\_t��:4\��

Prompt: to apply for a graduate degree students must

Continuation: to apply for a graduate degree students must submit a ) ��Y��Z % +AF '' 2 \* ���� # ���q� ! �l��Ө�� '' ��Ӹ�X� $ � \* ��

Prompt: Graduate students have research opportunities in areas like

Continuation: Graduate students have research opportunities in areas like public policy or Schools & �����an , ��� ? x�F��\�3�����R��.6��T� % �Hs�� @ nz��G��m��0��u��mO���A�����F '' ��m ���R���~���ǷN���W�i7ߥF��R e�D�I > ���

Prompt: On-campus housing options provide amenities such as

Continuation: On-campus housing options provide amenities such as @ �� \* �1Σ �5���~ '' �\��I�m } ���� ��5T�G�Tet̕�� < H � ( 3���� ��թZ '' ���o AR�� `

Prompt: The university's policies on academic integrity include guidelines like

Continuation: The university's policies on academic integrity include guidelines like more in Marketing Amber Hatcher St. John's Information Public School of Business Administration The Road St. John 's Success Parent

Prompt: Students can engage in extracurricular activities such as

Continuation: Students can engage in extracurricular activities such as early lessons of Conduct The Global Programs of their emerging careers . { �-�k��� { 1 �u��� % � ��i޹ڗY��-l0X��0��V

Prompt: Support for international students includes services like

Continuation: Support for international students includes services like public service to use data-driven approaches to explore the list of Creativity and Sciences Psychology : SԙP��߸�� ( # 6�

Prompt: Submitting a thesis or dissertation requires steps such as

Continuation: Submitting a thesis or dissertation requires steps such as a St. John 's Success Story Value in Africa . Professor of Equal Opportunity “ ¯gÓsĞÆcøôéóù �ü ) Libraries Research

Prompt: The university's career center offers services like

Continuation: The university's career center offers services like Hayes and Society Fr . ! V�RgUb4�赨 N�L�t�E�� �Bͣ�����J�HVR�2 { �O~ i�W2�B� ��u+��Q4+� g�� ����u���\_�y���t ���xO����M �O� \* \* B

In the output above, you can the output of the model with *n=1* predicted unknown characters and words that don’t make sense. We can decide from this that our model does not perform well with *n=1.*

*n=2 results:*

Prompt: In the fall semester, the business department offers courses such as

Continuation: In the fall semester, the business department offers courses such as dances , sports activities and round table conversations Cultural orientations for the field . Administrative Studies BS MA Minor BA/MA

Prompt: To apply for financial aid and scholarships, students must

Continuation: To apply for financial aid and scholarships, students must meet both of the marketing analytics provides a foundation in the world and experience the homes of so many of

Prompt: to apply for a graduate degree students must

Continuation: to apply for a graduate degree students must be exam P ) �� '' � [ i�� �S�w�����u�̉� ' B�2yAd�ӳ '' �j�9Jg-� $ TU�����b ` �S�� $

Prompt: Graduate students have research opportunities in areas like

Continuation: I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

Prompt: On-campus housing options provide amenities such as

Continuation: On-campus housing options provide amenities such as flooding . The University conducts compost tea brewing workshops in the issue of fair trade all my life . I

Prompt: The university's policies on academic integrity include guidelines like

Continuation: I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

Prompt: Students can engage in extracurricular activities such as

Continuation: Students can engage in extracurricular activities such as : internal audit , fraud examination , information technology audit , information technology audit , and competencies needed for success

Prompt: Support for international students includes services like

Continuation: I am sorry I don't have sufficient information to answer this one. Please Feed Me More Corpus !!!!

Prompt: Submitting a thesis or dissertation requires steps such as

Continuation: Submitting a thesis or dissertation requires steps such as poverty , illness , or in professional positions in industry , government officials , and recommend modern global supply chain

Prompt: The university's career center offers services like

Above is the output of the model with *n=2*. Here, the output of the predicted words is much better and makes contextual sense with the prompt. Out of the 10 prompts, only 2 really make contextual sense, but 6 out of the 10 prompts were able to generate words from the corpus which is better than our model with *n=1.*

*n=3 results:*

Next, we tried our n-gram model with *n=3*. There were many prompts that did not have the vocabulary to generate a prediction, but the two prompts that did generate text predicted words that had good contextual accuracy.

*Hybrid Approach, n=2/3 results:*

Prompt: In the fall semester, the business department offers courses such as

Continuation: In the fall semester, the business department offers courses such as Regulations S-K , S-X , S-T , and S-B along with exemptions provided under Regulations A and D. Financial reporting

Prompt: To apply for financial aid and scholarships, students must

Continuation: To apply for financial aid and scholarships, students must be in NYC . Career Outcomes The Master of Business Administration , please contact : Office of Graduate Admission The

Prompt: to apply for a graduate degree students must

Continuation: to apply for a graduate degree students must have , as a method of forecasting future financial and operating results , and as a means of diagnosing managerial

Prompt: Graduate students have research opportunities in areas like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: On-campus housing options provide amenities such as

Continuation: On-campus housing options provide amenities such as dances , sports activities and movies . It ’ s up to you what you do with it. ” Wendell

Prompt: The university's policies on academic integrity include guidelines like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: Students can engage in extracurricular activities such as

Continuation: Students can engage in extracurricular activities such as role-plays to maximize English speaking skills . They will also attend classes taught by well-regarded professors in International Business |

Prompt: Support for international students includes services like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: Submitting a thesis or dissertation requires steps such as

Continuation: Submitting a thesis or dissertation requires steps such as How to Write the College Application Essay , Tips for Summary Writing and The Art of Small Talk Apply Online

Prompt: The university's career center offers services like

Continuation: No continuation found. More diverse corpus might be needed.

  For our model with a hybrid approach, it slightly improved the generated responses. There were 3 prompts that the corpus did not have the dictionary for, and the other prompts did not generate great contextual output. Our model even generated. the same output for two different prompts, which is not good for our model.

*Higher-Order Approach results:*

Prompt: In the fall semester, the business department offers courses such as

Continuation: In the fall semester, the business department offers courses such as accounting , actuary science , business analytics , economics , finance , risk management , and more—gives you the skills , knowledge , and abilities that employers want . Those qualities translate across innumerable roles and fields . Why St. John's At St. John ’ s , the opportunity to

Prompt: To apply for financial aid and scholarships, students must

Continuation: To apply for financial aid and scholarships, students must be between 16-17 years old , and proficient in English . ( Students should have the equivalent of a B2 level on the CEFR . ) If English is not your native language , please review the chart found here to determine whether your English proficiency meets this requirement .

Prompt: to apply for a graduate degree students must

Continuation: to apply for a graduate degree students must have , as a minimum academic credential , the equivalent of a United States ' bachelor 's degree and demonstrated competency in pre-calculus mathematics and basic computer use . Students are evaluated based on their undergraduate academic performance , score on the GMAT , professional accomplishments , letters of recommendation

Prompt: Graduate students have research opportunities in areas like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: On-campus housing options provide amenities such as

Continuation: On-campus housing options provide amenities such as Wall Street , the Metropolitan Museum of Art , and Broadway . Learn More Semester Visiting Students Are you interested in spending a semester at St. John ’ s University , and enables students to select from a wide range of enterprise risk management concentrations , such as , operational

Prompt: The university's policies on academic integrity include guidelines like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: Students can engage in extracurricular activities such as

Continuation: Students can engage in extracurricular activities such as role-plays to maximize English speaking skills . They will also attend classes taught by well-regarded professors in International Business concepts such as finance , marketing and design , technology and intellectual property rights . To apply the language and concepts learned in these workshops , and learn team-building skills ,

Prompt: Support for international students includes services like

Continuation: No continuation found. More diverse corpus might be needed.

Prompt: Submitting a thesis or dissertation requires steps such as

Continuation: Submitting a thesis or dissertation requires steps such as sales , use , excise , and a variety of other transaction taxes ; transfer taxes such as estate , gift , and inheritance taxes ; value-added taxes ; and property taxes . TAX 654 Transfer Pricing ( 3 credit hours ) Prerequisite : TAX 603/TAX 635/TAX 655/equivalent . This

Prompt: The university's career center offers services like

Continuation: No continuation found. More diverse corpus might be needed.

Above is the output from our model with a higher order approach. The higher order approach allows the n to go from 2, which we saw generate pretty good responses, but not the most contextually accurate because of the number of words being used to predict as well as going up to *n=6,* a higher number of words being used to predict and generate output with more context. This model ended up generating the best output of the models in a words and context sense.

**Part 4 Model Application:**

In the end, the higher-order technique was the most effective. It produced the most contextually relevant results and highlighted the value of a diverse n-gram range for capturing language nuances. However, the limitations of our corpus size were evident, as it needed to generate a sufficient vocabulary for questions about specific issues, such as academic integrity policies or support for overseas students. This emphasized the importance of a broader, more diverse corpus—a conclusion consistent with creating robust language models.

**Part 5 Integration with AI (ChatGPT or Google colab):**

In our efforts to create a successful N-gram model, we were meticulous in our wording and closely followed the established guidelines. To build the basis, our first prompt sought to improve the web crawler.  
  
This project necessitated testing a model with ten carefully written prompts intended to assess the quality and relevancy of the generated responses. We assumed that the N-gram model was a version of ChatGPT that had been combined with a web crawler to improve parsing, text extraction, and decision-making capabilities. This integration was designed to leverage AI's capabilities to optimize procedures and enhance efficiency in data-driven jobs.  
  
The prompts used in this application were explicitly designed to test and develop the web crawler's AI functionality. Each challenge targeted a particular aspect of the crawler's operations or attempted to improve its algorithmic architecture. The following are the precise prompts used and the results they produced:

1. **"I need help optimizing my web crawler."** – The response provided fell short of explicitly meeting the optimization objective but prompted further clarification in subsequent interactions.
2. **"To generate a model that allows the 'n' to switch between 2 or 3 depending on which is better for the model."** – The initial response did not allow for dynamic alternation between bigrams and trigrams; however, the model was refined post-feedback to incorporate this functionality.
3. **"Can you adjust this prompt so that it is a mixture of an n-gram model with 2 or 3 as the 'n'?"** – The code generated from this prompt accurately established an n-gram model that dynamically adjusted 'n' based on the input, albeit requiring some initial clarification.
4. **"What about a higher-order approach?"** – This prompt inspired the creation of a model capable of initiating with higher n-grams and scaling down if the data was inadequate, showcasing an effective high-order strategy.
5. **"I heard that you are experienced with N-gram models, I wish to put that statement to the test."** – This prompt did not yield any code but served more as an exploration of capabilities, highlighting the need for more explicit prompts.
6. **"Do you think any optimization is necessary here?"** – This led to the optimization of the current N-gram model code, resulting in a functioning version, albeit without significant improvement in runtime.
7. **"Parameterization can help with finer control in my GPT-2 code; can you please explain that in more detail?"** – This elicited an in-depth explanation and a new code snippet. Initially, the code was incomplete but was subsequently revised.
8. **"Hey ChatGPT, that code you just provided did not work; please amend it."** – This direct request for correction resulted in a satisfactory response where the AI provided a functional code snippet.

The iterative process of developing the N-gram Language Model highlighted several distinct problems, each presenting a learning curve and a potential for advancement. As we worked through these issues, the efficacy of our model grew increasingly dependent on the accuracy of our prompts and the breadth of the corpus we used.

One crucial problem was the model's failure to create text continuations for some prompts due to a lack of comparable n-grams in the corpus. This constraint demonstrated the corpus' breadth and depth, forcing additional iterations of web crawling to improve the dataset. Despite an increased corpus size, the results could have been better, prompting us to investigate alternative methodologies.

Another challenge was implementing word embeddings, a technique we thought would increase model performance. Unfortunately, this did not produce the expected results, which led us to a critical realization: the complexity of language and context understanding within AI models necessitated careful study and a nuanced approach.

The most significant problem stemmed from the disparity between the AI-generated code and our project's operational requirements. This disparity necessitated a thorough calibration of prompts to ensure the AI's responses were theoretically valid and practical. It required us to personalize our questions with granular depth, contextual clarity, and technological specificity.

Addressing these issues frequently requires rapid iterative modifications and a thorough understanding of the AI's working habits. As we improved our inquiry formulation, we established a feedback loop that constantly improved the AI's output. A clear pattern emerged: the specificity of our prompts was directly proportionate to the utility of the resulting code. The AI-generated solutions became more precise with each iteration and closely aligned with our operational goals.

*Conclusion:*

Our project demonstrated the importance of thorough and precise rapid engineering in AI-driven language model creation by highlighting the issues faced and solutions implemented. The project's success depended on the synergistic interaction between a well-structured corpus, meticulously constructed prompts, and AI's adaptability to complicated linguistic demands. Our combined efforts have helped to advance the state-of-the-art in N-gram Language Models, paving the way for future advances in this exciting field of research.