



Module 3: How does AI generate text?

Purpose

A deeper dive into machine learning, specifically focusing on small and large language models, predictive text, and generative AI. Students will develop a foundational understanding of how models learn from examples to make predictions, emphasizing the idea that larger datasets contribute to more robust models. The lesson aims to demystify the process of AI-generated text and highlight the challenges of obtaining sufficient input/output pairs for effective model training.

Objectives

- Students will be able to explain how small language models work.
- Students will be able to define large language models and predictive text.
- Students will be able to practice using Generative AI with various text input, and edit their inputs to achieve their desired output.

Key Vocabulary

- **Language model:** An algorithm that uses probability to predict a sequence of words in a sentence. Additionally, large language models are language models that have been trained using massive datasets
- **Machine Learning:** Machine learning is when a computer program can learn from information and use that knowledge to make decisions or predictions.
- **Generative AI:** a type of artificial intelligence that can create new and original content, such as images, music, or text, based on patterns it has learned from existing data.
- **Token:** the number of individual words and punctuation marks in a text selection.
- **Knowledge Management:** Knowledge management is the process of collecting, organizing, and sharing information so that it can be used effectively. It involves creating a system to store information, making sure the information is accurate and up-to-date, and making it easy for people to find and use the information they need.
- **Natural Language Processing:** Natural language processing is a type of computer technology that helps computers understand human language. It allows computers to read, interpret, and respond to human language in a way that is similar to how people communicate with each other.
- **Information Management:** Information management is the process of collecting, storing, and organizing information so that it can be used effectively. It involves creating a system to store information, making sure the information is accurate and up-to-date, and making it easy for people to find and use the information they need.
- **Database:** A database is a collection of information that is organized in a way that makes it easy to search and retrieve. It is like a digital filing cabinet that stores information in tables, which can

be linked together to create a complete picture of the data. Databases are used to store all kinds of information, from customer data to scientific research.

Key Ideas

- Machine learning is essentially the same as showing a model of your desired output, and examples of that model, and letting the machine figure out patterns on its own. The more examples you give, the better the computer gets at making decisions on its own. Generative AI (like ChatGPT) and autocomplete generate text by using probability to predict the next word in a sentence
- In order to generate text that seems like it was written by a human, we must first train the artificial intelligence on large amounts of existing data, observation of humanity, and text. Tools like ChatGPT have collected massive amounts of writing from the internet up until the year 2021 in order to simulate human writing.
- Once the AI has been trained with data, it does not need to be trained again to perform the same tasks. The training data remains in the AI's learned parameters.

Module Breakdown

Mini Lesson (10 minutes)	How do chatbots, like ChatGPT work?
Activity (25 minutes)	Green Eggs and Ham: AI's Version
Debrief & Assess (15 minutes)	Discussion and Assessment
Extensions (additional time)	CS Academy app on a webpage Text editor app

Mini-Lesson (10 minutes)

1. Introduction & Discussion

How do AI Chatbots, like ChatGPT work?

- Bots like ChatGPT (GPT-3.5) work by using a type of artificial intelligence called a language model. These language models are trained on vast amounts of text data to understand and generate human-like language.
- They work by leveraging advanced language models and machine learning techniques to understand and generate human-like text based on the input they receive.

What can we do with AI: Generative Text?

There are some well known generative text AI's already at work. You probably see them often. Here are some examples.

- Predictive text with Gboard: Google's Gboard integrates Google's AI technology to offer predictive text suggestions, autocorrections, and emoji predictions. It even allows for accommodations, such as Glide Typing and Voice Typing.
- OpenAI's GPT-3 and its chatbots-ChatGPT. People can use this to look up information, ask the "database of the internet" a more specific question, or even get helpful tips in common areas like math, science, social studies, and even writing, literacy, and language.
- BERT: Google's BERT is designed to understand the context of words in search queries and provide more relevant search results. This can be understood as a much more accurate search for exactly what a user was looking for, without having to look through many links, in search of one that is not only credible, but accurate. People do also use it for text generative tasks, much like GPT-3.

- Grammarly Keyboard: Grammarly's mobile keyboard extension uses AI to offer grammar and spelling corrections, as well as vocabulary enhancements. It can also provide style suggestions and tone adjustments.

2. Video about these processes

[How AI Works: Chatbots and Large Language Models \(Part 1\)](#)

Discussion after video:

What machine learning and language models have you encountered?

- Expected Responses: predictive text in Google/iPhone, AI extensions on web browsers, AI extensions in text message app keyboards, filters on social media apps.

What is machine learning and how does it generate text?

- Expected Response: Machine learning is the way artificial intelligence can be trained by a user by using input and data, and giving the AI rules and parameters. It can generate text based on all of the text it is allowed to access, or has access to in general.

Activity (25 minutes) Green Eggs and Ham - AI's Version

3. Explanation

In this two part activity, students will look at data and analyze the text around it. In part 1, students will work with the story Green Eggs & Ham by Dr. Suess. Students will be looking at individual words, and figuring out which word patterns emerge in the text. Students will begin to see that throughout these stories, you can usually predict which words should follow almost any word in the text (aside from articles such as: the, and, a, etc).

For example: Dr. Suess writes: "I do not like them, Sam I am." many times. If a student sees the words "I do not like them" they would likely predict the words that follow would be "Sam I am." If a student is searching for the word "Sam," then they would likely begin to notice the pattern "Sam I am".

When students do this, they are sorting and organizing the data given to them, just like AI does. While this doesn't take forever, it does take some time and mental effort to search for the same word in a simple text for older students. This is one of the benefits of AI, AI can do this menial task, very quickly.

Below are two sets of instructions, the first to do this unplugged, using physical print copies of the book and paper, the second to this plugged, using a computer program.

In the second part of the activity, students will sort and organize the same text digitally, through a text generation app called CS Academy Text Editor. This part allows students to practice using an AI (the text editor) that is trained using the data set (words) in the story Green Eggs and Ham. This is done with something called N-grams.

4. Directions for Unplugged Activity

Prep for Unplugged Activity

- Prep printed copies of the story Green Eggs and Ham by Dr. Seuss
 - Resource Link:
 - [Green Eggs and Ham Digital Text](#)
 - Green Eggs and Ham Keyword Handouts: [PDF](#) | [Spreadsheet](#)
 - Answer Set: [Green Eggs and Ham 2-gram frequencies](#)
 - Digital Asset: [1 token Text Generation for Green Eggs and Ham](#)
 - Students can also view the text digitally, they can type CTRL + F to search the text for a specific word of their choosing.
 - This allows the students to find the exact word they chose, and the phrases that follow more quickly and accurately.
 - Students can also view a few of the pages or the whole book to look for data.

- Accommodations: bolding or underlining the word chosen, or allowing the student to highlight the word and its subsequent word or phrase.
- Here is a list of all words contained in this story, for reference when choosing words:

a	am	and	anywhere	are	be	boat	box	car	could
dark	do	eat	eggs	fox	goat	good	green	ham	here
house	I	if	in	let	like	may	me	mouse	not
on	or	rain	Sam	say	see	so	thank	that	the
them	there	they	train	tree	try	will	with	would	you
. (period)	, (comma)	! (excl)	? (question)	Sam-i-am					

- Each student should be given **one word** to search for. We encourage you to select the most appropriate words for each of your students based on their ability level. The PDF handout above has the words sorted from most frequent to least frequent.
 - The most frequent word, “Not” appears 83 times in the story, while the least frequent word, “If”, only appears once.
 - Due to time constraints, you may decide that certain words show up too frequently for your students to complete within the allotted time. If you choose to eliminate some words, you can find them already completed in the answer key above.
 - If you choose to give students words with extremely low frequencies, you can ask them to complete additional words.
- Students should look through the text and search for their word, highlight or underline it.
- Then they should write a list of the words that come next after the keyword they highlighted above.
 - See sample of a student’s notes who chose to search for the word “Here” in the story Green Eggs and Ham:

- Then they should count the number of times that each next word appears.

Keyword	Here
1	or
2	or
3	or
4	or
5	they
6	or
7	or
8	in
9	or
10	or
11	and

or → 8
 they → 1
 in → 1
 and → 1

- 'or' appears 8 times after 'here' in the story.
 - 'they' appears 1 time after 'here' in the story.
 - 'in' appears 1 time after 'here' in the story.
 - 'and' appears 1 time after 'here' in the story.
- Together as a class, generate a nonsensical Dr. Seuss story by making a sentence, This is you acting like a generative AI. Start with the word 'Green', write it on the board. Ask for the student who had the word 'Green' and have them tell you what is the next most frequent word, 'Eggs'. Write that word on the board. Then ask for the student who has that word 'Eggs' and have them tell you what is the next most frequent word, 'And'. Write that word on the board. Then ask for the student who has that word 'And' and have them tell you what is the next most frequent word, 'Ham'. You can continue this pattern for a few more words or start with a new word.
 - It may seem like the sentence comes out nonsensical and this is the point. When we only can look back in the text one word, it is not going to make a complete sentence.

5. Discussion

- **Is this activity enjoyable to you? Would you want to do this all the time? Why or why not?**
 - Expected Response: Many students would likely say that they don't enjoy the activity much, as it is just searching for a single word. Some may enjoy it somewhat, as these books are preferred by some as childhood favorites. Note that this is a human characteristic, however. AI does not 'enjoy or not enjoy' doing this, unlike us humans who have very particular preferences.
- **How would you feel if I asked you to do this with something like a very long book, such as The Hobbit? A very long song, like "We Didn't Start the Fire?"**
 - Expected Response: Many students will say they would definitely not enjoy this, as they may not even prefer that style of book or music. That would be much longer and there is really no reward.

6. Explanation for Plugged Demonstration

Links for Demonstration: [CS Text Generation \(select multiple tokens\): Selecting Multiple Tokens](#)

In this activity, an **N-gram** is defined as the number of words being found from a data set. In the unplugged activity example above, the N-gram was 2. You asked students what is the next most frequent word after 'green', therefore they were looking for 2 words or N-gram 2. If you said - what is the next most frequent word after 'green eggs and' that is N-gram 4.

When you increase the N-gram you are more likely to get a more sensible phrase. However, it is incredibly difficult for humans to figure out from all of the possible combinations of words what comes next with larger N-grams. That is why we use computers to do this work.

In text generative AI apps and large language models, the machine is analyzing all of the data of all of the text it has access to and doing mathematical calculations of what word is most likely to come next.

In this demonstration activity the computer model you are going to be using is only for demonstration purposes, it is simply demonstrating how the AI model can predict what word comes next much faster than a human can. It also demonstrates how using more words to predict (a higher N-gram), provides a better result.

This Small Language Model Demo has the full text of all of Dr. Seuss books. Use Green Eggs and Ham first, so that students can make the connection between the plugged and the unplugged activities in this module. Then allow them to try others.

- The Grinch
- Fox and Socks
- The Cat and the Hat
- Green Eggs and Ham

Want to learn more about N-Grams and how this model works? Go here:

<https://docs.google.com/document/d/1emkWEECTdACS7Kylx9YNd78u5Zbq7i9OvmYKwikwOFQ/edit?usp=sharing>

Directions for Plugged Demonstration

- Show the model on the board.
- Start with N: 2 on the bottom and click generate, you will notice in the middle it generates the next word after and on the right it shows all of the possible options for the next word.
- Then increase the N: 3, ask students to predict what is going to happen.
- Now the model is using two words to predict the next word, giving a slightly better result.
- Then increase the N: 4 to ask students to predict what is going to happen.
- Now the model is using three words to predict the next word, giving a slightly better result.
- Continue this for as many as you wish.
- Students should compare the results they got when they manually counted to the results the model got. Were they right, did they miss any or make a mistake? Which was faster?

Green Eggs and Ham

Choose the dataset

Explain some stuff

Generate next word/token

Explain some stuff

1: i am sam . i am sam . sam i am . that sam-i-am . that sam-i-am : i do not like that sam-i-am do you like green eggs and ham i do not like them , sam-i-am . i do not like green eggs and ham . would you like them here or there ? i would not like them here or there . i would not like them anywhere . i do not like green eggs and ham . i do not like them , sam-i-am . would you like them in a house ? would you like them with a mouse ? i do not like them in a house . i do not like them with a mouse . i do not like them here or there . i do not like

1: i am
2: i am
3: i am
4: i do
5: i do
6: i do
7: i would
8: i would
9: i do
10: i do
11: i do
12: i do
13: i do
14: i do
15: i do
16: i do
17: i would
18: i would

2

-

+

Generate

Change the N in N-gram

- Texting, chatbots, autofill, autocorrect, google docs, email, predictive search, and so many more.
- AI has collected years of written human text and now has a database of information. When I start typing to my friend “Good” based on tons of data, it auto predicts that I may type “morning” “afternoon” “evening” because that is what most people type after the word “Good”. It also may be taking into account data about the time of day it is to help make the best guess at what you want to say.

Assessment

8. Assessment

- Which of these is an example of text-generative AI?
 - a. Typing in a question on www.google.com to search for an answer to a question
 - b. Chatbots (like ChatGPT)**
 - c. Language Translators (like Google Translate)
 - d. Text to Speech
- In order to create an artificial intelligence that writes like you, what data could you use to train it?
 - a. Expected Responses: Try training it using words that sound much like you. Ask it to give you more eloquent words, or ask it to make something easier to understand. Ask it to translate it to another language. You can even ask it to check for something easy, like spelling errors and grammatical errors.
- Predict what word will follow this phrase based on “One Fish, Two Fish, Red Fish, Blue Fish”
 - a. Predict the next word: _____.

Extensions

- Demonstration CS Academy app on a webpage: <https://cs.cmu.edu/~pvirtue/hungry.html>
- (Demonstration)Text editor app: https://www.cs.cmu.edu/~pvirtue/text_generation.html
- Language & Writing Application: Students collectively produce a story, letter, or any kind of writing together. This can be done in a passed notebook, or even a digital platform, like Google Slides. Students “predict text” and tell a story by first reading what their classmates wrote first. Then, have the next student add a preset amount of words or sentences to the story. Do students notice the story tends to follow a pattern, much like the N-Gram game above? The more input an “AI” receives, the better it can follow the pattern.
- Alternative: Play this like the well loved “Toss the Ball” game for a quick Speaking and Listening check. Have students verbally say the preset amount of words. Students must listen closely to the entire story, especially if the data set (preset amount of words) is higher.
- Math Application: Don’t Hit Oh! Students can play the same “Toss the Ball” style of game, with a twist. The goal of this game is to NOT hit zero (Oh!) while playing with math calculation. Students will work together to take turns adding to an equation in order to make the total of the equation never reach 0. Here is a sample of how this game is played:
 - To begin, select a student to start the game, and hand them a ball.

Student A: Says aloud “200”

 - Begins equation for class.
 - Choose a whole number.
 - Tosses the ball to a new student.

Student B: Says aloud “Divided by...”

 - Chooses the function that follows Student A’s number.
 - Student B can say plus, minus, times, divided by, or even add an exponent.
 - Tosses the ball to the next student.

Student C: Says aloud "100"

- Chooses the next number in the equation, following Student B's function.
- Tosses the ball to the next student.

Since the total has not yet reached zero (as it is now at 2), the game continues in this way until the total is forced to reach zero or into the negatives.

Student D: Says aloud "Plus"

- Chooses the next function in the equation.
- Tosses the ball to the next student.

Knowing the total is at 2 right now, as $200 \div 100 = 2$, this student could accidentally take the total past zero if they choose to subtract too much, or choose another function that could cause the total to drop below zero. They will need to keep track of the total in order to not end the game by forcing zero to be reached.

Student E: Says aloud "19"

- Chooses the next number in the equation.
 - Tosses the ball to the next student.
 - Continue the game as time permits, until a student is forced to give a response that is equal to or less than zero. Once that has happened, the game ends. Each time, the student who next catches the ball's goal is to NOT allow the total remaining to ever hit or dip below zero.
 - This is a collaborative equation that is being generated by students by taking turns.
Students will need to:
 - Listen closely and keep track of the total at all times, in case they are chosen to be the next student to catch the ball.
 - Choose a number or function that works with the previous student's response, that does not force the total to dip to at or below zero.
 - Calculate the equation each piece at a time, while holding the previous total in their brains in order to keep the total above zero.
 - Code.org [Lesson Plan](#) for Chatbots and Large Language Models (LLMs)
 - Code.org [Written Voice Simulator Widget Widget](#)
 - Code.org [Written Voice Simulator - Student Activity Handout](#)
-