

# Large Language Models <sup>1</sup>

Aleks Lillis

June 2025

<sup>1</sup>Hey Mrs. Murphy... I used AI for this project!

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	What are Large Language Models . . . . .	2
1.2	Examples . . . . .	3
1.3	Brief Statistics . . . . .	4
<b>2</b>	<b>LLMs Text Generation</b>	<b>6</b>
2.1	Language . . . . .	6
2.2	Meaning . . . . .	6
2.3	Text Generation Process . . . . .	7
2.3.1	The Power of the Prompt . . . . .	7
2.3.2	Probabilistic Models . . . . .	9
<b>3</b>	<b>Model Training</b>	<b>10</b>
3.1	What is Training . . . . .	10
3.2	How are Models Trained . . . . .	10
3.2.1	Pretraining . . . . .	10
3.2.2	Fine-tuning . . . . .	10
3.2.3	Reinforcement Learning . . . . .	11
3.3	Training Data and Bias . . . . .	11
<b>4</b>	<b>Training Example</b>	<b>13</b>
4.1	Process Overview . . . . .	13
4.1.1	What the process Was . . . . .	13
4.1.2	Generated Content . . . . .	13
4.2	Analysis . . . . .	18
4.2.1	Randomness . . . . .	18
4.2.2	Trained Data . . . . .	18

# Introduction

## 1.1 What are Large Language Models

In recent times artificial intelligence (AI) has been rapidly developing. With this large AI booms large language models (LLMs) have become a well known thing with many people mistaking them to be synonymous with AI as a whole. This is inaccurate because LLMs are a subset of AI which focus specifically on human language.

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.

Above is the definition of LLMs by International Business Machines Corporation, more widely known as IBM. To put it simple, LLMs are models embedded into programs which are designed to not only understand human text but also respond by generating content. Generated content is based on the data that trains the model (which will be gone over more in depth in a later section) and the prompt that is provided.

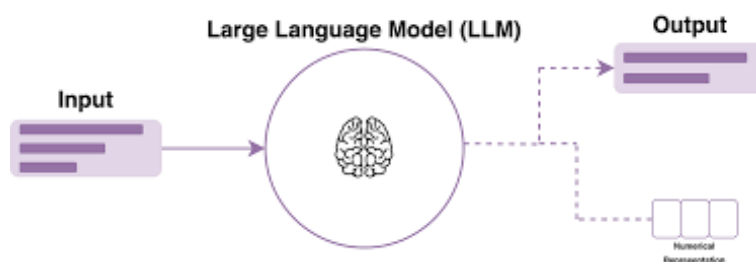


Figure 1.1: Simple LLM diagram [3]

To better understand what LLMs are it is very important to take into consideration what they aren't. LLMs primarily generate context based on patterns uncovered during training and fine-tuning phases of their development. These models can respond to questions by predicting what follows but they do not truly think, that is the job of reasoning models. Reasoning models simulate ac-

tual logic processes in order to solve problems. LLMs are more likely to come up with inaccurate answers because they are models predicting what the response should be.

## 1.2 Examples

LLMs are used in many different applications. What's nice about modern models is that they can be easily accessible through various mediums by: being built into applications, easily trained remotely from open-source access points, and just being incorporated through a remote server that can send over responses from specific prompts. Some specific usages of LLMs are as follow:

- Text Generation
- Content Summarization
- AI Assistance
- Code Generation
- Sentiment Analysis
- Language Translation
- Grading
- Teaching

The largest use of LLMs in modern day is the use of chat agents that are able to respond and generate context from given prompts but are also able to add prior information from the chat log and use it as context for future generation (while not guaranteed in every model this is typically an implemented feature). Very well-known chat agents are:

- ChatGPT

ChatGPT is an easily accessible model which has rapidly developed recently and is the face of AI chat bots.

- DeepSeek

DeepSeek was created by a Chinese AI startup and stirred up the world of AI by having similar capabilities at a much lower cost.

- Claude

Claude models are forms of conversational AI designed with the goal of being ethical and safe.

- Meta AI

Meta AI is a conversational generative AI that is popular mainly due to its integration to Meta platforms such as Facebook, Instagram, WhatsApp and more.

- Google Gemini

Google Gemini is Google's AI chatbot which is integrated into many of Google's different apps and it is widely popular due to its summarization of search queries on Google's search engine.

- Microsoft Copilot

Copilot is Microsoft's conversational AI model which is integrated into windows computers and other products that Microsoft has.

- Khan Academy Khanmigo

Khanmigo is an AI powered chatbot which focuses on helping students learn from a wide range of courses.



Figure 1.2: An image of the "friendly" icon of the khanmigo AI [4]

## 1.3 Brief Statistics

The chatbot market is a relatively new but very quickly growing market. According to the Chatbot Global Market Report 2025 the market was valued at \$8.27 billion in 2024 and is projected to have a growth to \$10.32 billion by the end of 2025. Seeing this compound annual growth rate of roughly 25%, the chatbot market is expected to experience an exponential growth to \$29.5 billion by 2029 [8].

The market is not the only place where AI numbers have seen a huge increase in recent times. AI usage has had a wild increase in the last few years globally. AI is very appealing in the work place due to it being able to very easily complete repetitive tasks that would otherwise take up a lot of time for someone to do it themselves. In a study done by the Melbourne Business School as of 2025 %58 of employees have reported intentionally (using specific tools to aid in task completion) using AI at work on a regular basis where applicable [12]. In a survey done in August 2024, about one out of three respondents reported using AI tools daily or at least a few times per week. One interesting data point was that AI usage was reported to be higher at home than at work but more frequent occurrence was more common at work [7]. This large-scale usage of AI tools has mixed impacts but there is no use that it is a very rapidly developing field and becoming a much larger part of our lives.

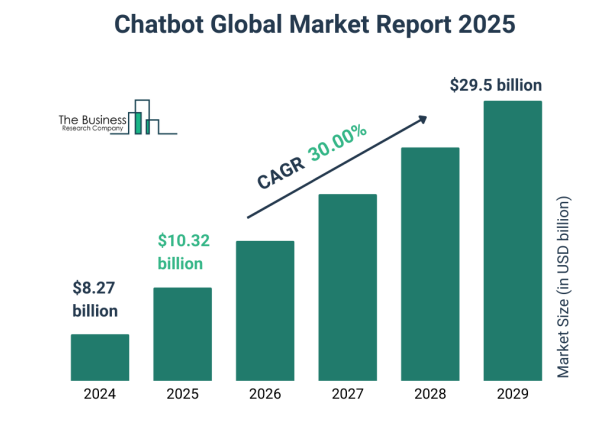


Figure 1.3: A graph depicting the chatbot market growth [3]

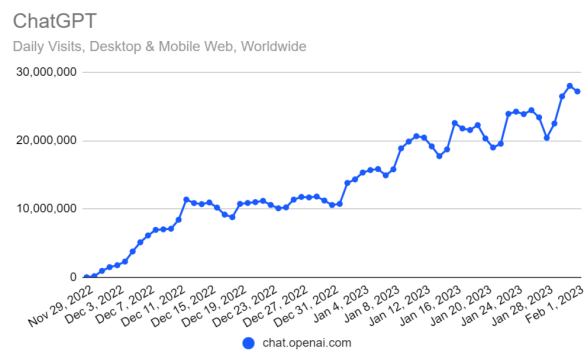


Figure 1.4: A graph showing the fast increase of AI usage [16]

# LLMs Text Generation

## 2.1 Language

As mentioned earlier, the main purpose of LLMs is to generate text—but what, fundamentally, is language? Bernard Bloch and George L. Trager define it as “a system of arbitrary vocal symbols by means of which a social group cooperates.” This classic definition emphasizes spoken language and its social function. However, language has since broadened to include written, gestural, and even computational forms. At its core, language is a structured system of symbols—whether phonetic or textual—used to convey meaning, share knowledge, and build connections. Understanding this helps frame what LLMs are truly modeling: not intelligence in the human sense, but patterns in the way people communicate meaning through text.

## 2.2 Meaning



Figure 2.1: The famous sculpture of “The Thinker” which embodies a strong philosophical connection to meaning [9]

Does the text that LLMs produce actually convey any meaning and can invoke emotion? Lets start first by comparing this generated text to human language. In human language meaning is deeply tied to context and sentence intention. A sentence or statement isn't powerful because of the words that it uses but mostly because of what we mean by them. Large Language Models, however, do not have the ability to convey this level of meaning and have as high of a level of understanding of what is being conveyed. Instead, LLMs are able to generate text by using highly advanced probabilistic models based on statistical patterns from trained data to complete sentences.

## 2.3 Text Generation Process

How do Large Language Models actually generate text?

### 2.3.1 The Power of the Prompt

There are many different types of prompts that LLMs are capable of responding to.

- Summarization

LLMs are able to summarize text. This can be achieved by using advanced algorithms for generating a shorter version of a text that still keeps the main points. Summarization primarily reliant on **word frequency** (words that appear more often are likely to be more important), **sentence position** (some sentences are more important in conveying the point of a text such as thesis statements or conclusions), **specific names** (sentences which specify information like names, locations, etc. are more likely to be important), and much more [14]. As well as this, larger texts are typically broken down into smaller chunks of text in order to summarize each one into broader summaries of the entire text by keeping the most important information [1].

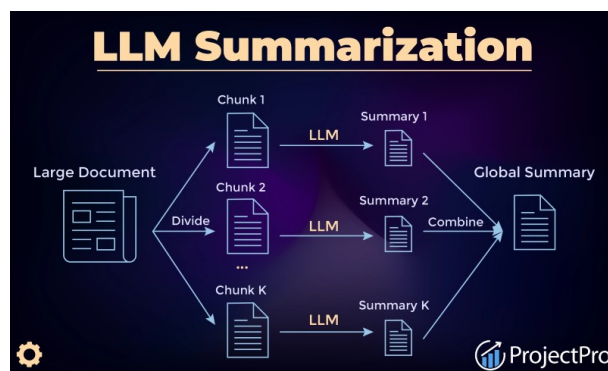


Figure 2.2: Simple diagram breaking down LLM summarization [1]



- Question Answering

AI models are able to answer questions which is one of the primary ways that LLMs are used. Well, LLMs don't actually recognize and know how to respond to a question by thinking. The way LLMs actually answer is by continuing text. The main way that this is able to be done is when the non-AI portion of the program transforms the original part into text that can be continued to provide an answer. This is because the model can finish and return a completed version of rephrased sentence which is the generated answer.

- Text Completion

As briefly touched upon in the question answering section LLMs excel at completing text. This is because at their core LLMs operate by guessing what should be added in order to complete or continue text. Many different types of prompts ultimately end up relying on text completion by having whatever program that the model is hosted on modify a user prompt to be an incomplete statement that must be completed.

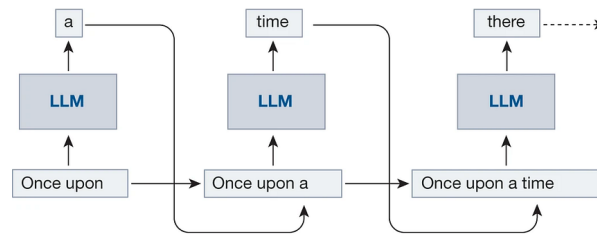


Figure 2.3: Simple diagram breaking down LLM text completion [13]

- Computer Code Generation

LLMs can also be useful in generating computer code. The process of doing that is very similar to the process that models generate text. Models can be fine tuned in training to focus on large datasets with lots of computer code that is assumed to be accurate. When a prompt is received by a model, it will generate and combine code fragments that are likely relevant to the prompt and then output the fully generated code [6]. Although there are many worries about AI coding and how it could destroy the computer science market by replacing many different employees there are many reasons why this type of computer code generation is not effective with how AI is developed currently. This is because of many concerns such as AI repeating poor security practices based on some of the code which it might be based on, legal concerns since AI may use code which is from copyrighted material, and lack of efficiency by having models repeatedly implement inefficient and ineffective algorithms [15].

These are just a few of the ways that artificial intelligence can be prompted to generate specific types of content.

### 2.3.2 Probabilistic Models

In this report it has been mentioned time and time again that generated content is simply what is most likely to follow based on the context of a prompt or question. These models are able to do this because of the combined power of the information it knows from its training period (a topic for the next chapter) and the context of a specific prompt. The model takes this information and, at each step of the generation process, it estimates what the most likely next word will be. Context is important because the estimation is not just based on all prior words but it is based on all prior words within the prompt and newly generated words as well. The prediction of words could pick only the most probably next word but this can cause problems by resulting in repetition and absolutely no variance under using the same prompt with the same training process. Because of this, LLMs tend to incorporate probability distributions by compiling the most likely next words and sampling by occasionally selecting words outside of the most likely. Most models have a temperature property which controls just how random a model is able to be with lower values making the model more deterministic and higher values making a model more creative and able to generate diverse content [10].

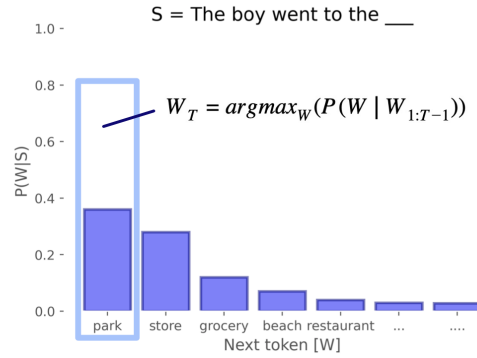


Figure 2.4: Simple diagram showing a probability distribution for most likely next word [10]

# Model Training

## 3.1 What is Training

Training is how large language models actually gather information. This is the process that actually allows models to learn patterns, structures, and find relationships between words and phrases. Just as children learn language by being immersed through conversation, social interaction, stories, and more, training is a process that is enabled by exposing models to large amounts of data. What makes training difficult for large language models is that models can see different words, for this example "dog", in millions of sentences. Through this exposure models can learn how the word is used but these models do not know information about what a dog is unlike children who most likely learn the word by associating it with the animal that it can see, hear, or feel.

## 3.2 How are Models Trained

### 3.2.1 Pretraining

Pretraining is the most fundamental part of training and creating large language models. This is where models are actually provided with thousands, millions, billions, or even trillions (depending, of course, on the scale of the model) words and data points. During this type of training large language models are not necessarily learning new facts or comprehending any information but they are enabled to create and better understand complex relationships with words and phrases.

### 3.2.2 Fine-tuning

Fine-tuning is another process that models go through after pretraining. This is a process where models adapt the general knowledge that is gained and transform it into models that are tailored for specific applications and use cases. The main goal of this phase is to optimize the model to complete specific tasks by adjusting weights based on task-specific data, improve accuracy and relevance of the model in specific fields, and also reduce biases that could have been introduced during pretraining that could make the model much more reliable for

real-world use [2].

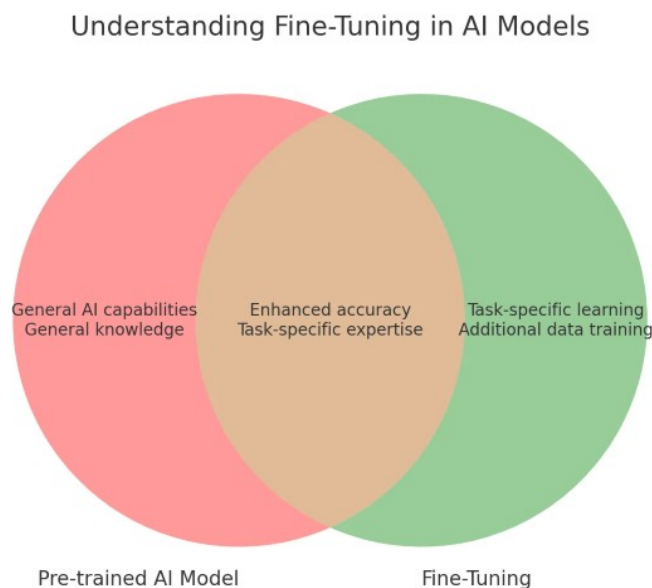


Figure 3.1: Diagram comparing pretraining and fine-tuning. [2]

### 3.2.3 Reinforcement Learning

Some models go through a reinforcement learning phase where outputs are rated for accuracy and after many iterations of rating responses a model will be weighted towards generating responses that are more likely to get a higher rating [5].

## 3.3 Training Data and Bias

A lot of different types of content is included in training data for large language models. The types vary largely but can mostly be summarized as either websites or books. Websites are useful data sets because websites can feature important information such as **translations** (which can be useful for translation tasks), **computer program content** (which is very useful for generating code), **FAQ sections** (which is very useful for question prompts), **general information** (which is very important data in order for the large language model to make accurate responses), and much more. One well-known source for training data is colossal clean crawled corpus which contains text from billions of extracted data. However, when choosing data to train a model it is very important to take into consideration the bias and accuracy that the selected data presents.

Bias can be presented if the training data doesn't accurately represent the real world or the population that model is intended to serve and can be introduced in many different ways. One common example of how bias can be introduced is through misrepresenting specific groups of people which can create inconsistent data. As well as this, it is extremely important to train models with accurate information. This is due to the fact that models are basing their responses to prompts on what is most likely to follow prior text. It is very possible, and realistic, that a model will respond with content that does statistically follow the text based on the input data but the response could be a complete lie that is presented within some portion of the trained data [11].

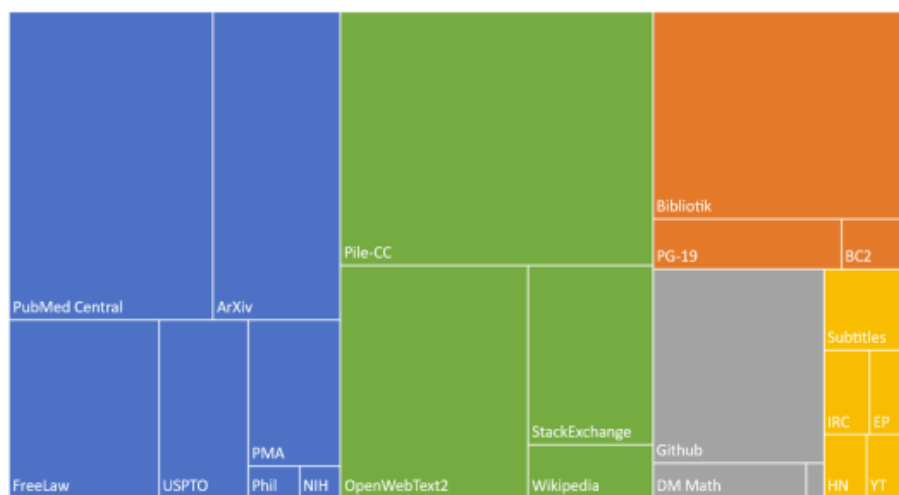


Figure 3.2: Sample corpus of data scraped for the internet used for training purposes [11]

# Training Example

## 4.1 Process Overview

### 4.1.1 What the process Was

For this report, I fine-tuned my own version of GPT-2 (an older AI model) to generate content based on my text. I provided essays from the last year and had the model generate its own essay. Overall the essays I included ultimately gave the model 6,826 tokens (words) to work with and create relationships. I will be comparing the content generated from my own model with 0.7 temperature (moderate level of randomness and creativity), 0.2 temperature (much less randomness and creativity), and ChatGPT. Both the custom fine-tuned models used the same prompt and were instructed to generate 500 tokens.

#### Custom Prompt

Topic: The Usage of AI in Education, Essay:

### 4.1.2 Generated Content

Here is the generated content from the 0.7 temperature model.

Essay:

While technology has advanced over time in terms of its wide access and wide access to all people it also has a part in cognitively stimulating activities that involve reasoning and practice. Though some people argue that this type of instruction is useless and question its application to the real world, its value is important for early aged students before they move on to working primarily with technology to write.

In recent history technological advancements have had a major impact on our society and ways of communication. One of the largest impacts that this has had is the ability to communicate

with one another in a near-instantaneous manner. This communication is able to be done through various means such as social media or texting apps. The wide access to such applications allows for a wide range of people, including younger people, but it also allows for people to anonymously say or do hurtful things to other individuals through cyberbullying. With this in mind the First Amendment right to free speech does not protect cyber-bullies because cyber bullying has lasting impacts on individuals who are impacted by it, in some cases cyberbullying is able to be held legally accountable, and the first amendment is in place to prevent the government from restricting the truth or construction criticism, not for allowing young people to be exposed to such hateful content. Cyberbullying can happen to anyone at any time, including younger kids. Source E shows that 42% of kids have been bullied while online (Source E). This can have a major impact on a younger individual who is still developing and learning about themselves. This can cause someone to start to feel insecure about themselves or grow one of their insecurities. Many people think that most people can move on with their lives after experiencing cyberbullying but younger victims are likely to react in ways which can have a negative impact on other people. This includes the victim cyber-bullying other people back to make themselves feel better, potentially seeking revenge on the bully, or by isolating themselves (Source A). This is supported by the fact that 53% of kids admit to having said something mean or hurtful to another person online which is likely caused because they felt bad about themselves and didn't know how to react (Source E). Saying things which can harm other people is not able to be protected by the first amendment because any website or platform is able to handle reports of such activity and take them down.

Here is the generated content from the 0.2 temperature model.

Essay:

" the need for curriculum that emphasizes

critical thinking, the study of language and world languages, and the development of humanities and STEM learning is necessary to advance the way humans think."

This is supported with another small anecdote about how on an election day for his second Congress he was meeting with voters and when he asked a voter who they planned on voting even though the woman couldn't pronounce or spell her name without aid she pulled out a paper so that she knew who to cast her vote for which had her name written on it. The next day at work she greeted customers and asked if she could cast her vote but when she asked another customer to wait until she was done talking to the customer she pulled out a paper so that she knew who to cast her vote for which had her name written on it. This evidence supports the idea that the purpose of this essay was to teach humanities and STEM topics to students in order to help them become better informed citizens. The handwriting-exercises implemented during class are necessary to help students develop their cognitive skills, be prepared for the future, and be better at literary analysis.

argument that the use of cognitive skills in education is key is valid and it is valid for students to be taught about cognitive skills while promoting humanities and STEM learning. The handwriting exercises used in this essay are to show how the brain is developmentally and culturally developing as a result of these early learning experiences.

Birds brain development is not quick and easy to adapt to new situations. Many people think that most people can move on with their lives after experiencing life as a stranger without incidenting on the path to victimization. However, some people think that most people can move on with their lives after experiencing life as a stranger without incidenting on the path to victimization. This is because most of the cognitive skills that a typical person would be able to move on with their lives after experiencing life as a stranger in America are completely gone.



The majority of people who hear about cognitive skills being important in education are unable to get right away and say, "Wow, that student could have gone on to become a successful engineer, physicist, or other popular professions without going through the extra steps of applying cognitive skills." Cognitive skills are absolutely essential for students to be taught in school. To support this idea of how important cognitive skills are in the modern

Here is the ChatGPT response from the same prompt:

Artificial Intelligence (AI) is transforming the landscape of education, offering powerful tools that enhance learning experiences, streamline administrative tasks, and personalize student support. From AI-powered tutoring systems to automated grading software, the integration of AI into classrooms and online platforms is reshaping how educators teach and how students learn. While AI presents several challenges, its potential to revolutionize education makes it a valuable asset for the 21st century.

One of the most impactful uses of AI in education is personalized learning. Traditional education often adopts a one-size-fits-all approach, which may not meet the diverse needs of individual students. AI can analyze a student's learning patterns, strengths, and weaknesses to tailor instruction accordingly. For example, adaptive learning platforms such as Khan Academy and DreamBox use algorithms to modify lesson difficulty in real time, ensuring that students receive content at an appropriate challenge level. This increases engagement and improves learning outcomes, especially for students who may struggle in conventional settings.

AI also assists teachers by automating repetitive tasks, allowing them to focus on more meaningful interactions with students. Tasks such as grading multiple-choice tests, monitoring student progress, or handling basic questions through chatbots can be efficiently

managed by AI systems. This not only saves time but also reduces teacher burnout. Moreover, AI can provide teachers with data-driven insights about student performance, helping them identify who may need additional help and which teaching strategies are most effective.

In addition to supporting teachers and students, AI plays a significant role in expanding access to education. Virtual teaching assistants and language processing technologies can help bridge communication gaps for students who speak different languages or have disabilities. Tools like speech-to-text and AI-generated subtitles make content more accessible, promoting inclusivity and equal opportunity in learning environments.

However, the use of AI in education is not without ethical and practical concerns. One major issue is data privacy. AI systems rely on large amounts of student data to function effectively, raising questions about how this data is collected, stored, and used. There is also the risk of algorithmic bias, where AI unintentionally favors certain groups of students over others due to flawed training data. Furthermore, overreliance on AI could depersonalize education and weaken critical human relationships between teachers and students.

Despite these concerns, the responsible and thoughtful integration of AI into education offers significant benefits. It is essential for schools and policymakers to establish clear guidelines for ethical AI use, ensuring transparency, accountability, and data protection. Educators should also receive training to understand how to effectively use AI tools while maintaining a human-centered approach to teaching.

In conclusion, AI has the potential to transform education by personalizing learning, reducing administrative burdens, and increasing accessibility. While challenges such as data privacy and ethical use remain, they can be addressed with proper oversight and policy.

When used responsibly, AI can be a powerful ally in creating a more efficient, inclusive, and student-centered educational system.

## 4.2 Analysis

Unfortunately, as seen in the bizarre amalgamations generated by my fine-tuned model, my model was not able to properly generate unique essays. This primarily due to the fact that although I had trained the model with over 6,500 words that is barely able to compare to the millions or billions of tokens that larger models are trained on. If the detail of the essays actually staying on track with the prompt is ignored this text generation was successful in showing how temperature can impact randomness but also in general how the trained data influences how the text is actually generated.

### 4.2.1 Randomness

Although there wasn't a crazy change in the generated text exhibited there was still a change in the randomness of the generated text between the different temperature models. Specifically, in the 0.7 temperature model (more random) the topic is changed much more than the 0.2 temperature model (more deterministic). This is due to the model being much more open to choosing different words following the same word compared to the other model. As well as this, the 0.2 model focused a lot more on the prompt by mentioning education a lot more unlike the other model because it was a lot less random in its generation.

### 4.2.2 Trained Data

In the generated text there is a lot of repetition because of the lack of training data for the models but at the very least this displays how the model was able to successfully train itself on my data and my custom models picked up much of my writing style over the last year which can be seen by common phrases that are included in my writing overall writing such as "with this in mind" being included and stating evidence by saying "this is supported by." In contrast, ChatGPT's content includes a lot more general text which is less likely to be written by a student within 40 minutes for a timed writing class-period, especially myself.

# Bibliography

- [1] en. Jan. 2025. URL: <https://www.projectpro.io/article/llm-summarization/1082>.
- [2] en. Feb. 2025. URL: <https://www.sapien.io/blog/fine-tuning-vs-pre-training-key-differences-for-language-models>.
- [3] URL: <https://images.app.goo.gl/KskyibfBQ4kLSzA67>.
- [4] URL: <https://support.khanacademy.org/hc/en-us/articles/13987846292621-Where-did-the-name-Khanmigo-come-from>.
- [5] en-US. URL: <https://aws.amazon.com/what-is/reinforcement-learning/>.
- [6] Admin. *How does a Large Language Model (LLM) writes Code - AI Verse Info*. en-US. Jan. 2024. URL: <https://aiverseinfo.com/how-llm-writes-code/>.
- [7] Kseniia Burmagina. *Unveiling the Future: AI Usage Stats You Need to Know*. en-US. Jan. 2025. URL: <https://elfsight.com/blog/ai-usage-statistics/>.
- [8] *Chatbot Market Report 2025: Statistics and Key Trends by 2034*. en. URL: <https://www.thebusinessresearchcompany.com/report/chatbot-global-market-report>.
- [9] Wikipedia contributors. *The Thinker*. en. June 2025. URL: [https://en.wikipedia.org/wiki/The\\_Thinker](https://en.wikipedia.org/wiki/The_Thinker).
- [10] Shelton Fitch. *The surprising power of next word Prediction: Large Language Models Explained, Part 1*. en-US. June 2024. URL: [https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/#:~:text="](https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/#:~:text=)
- [11] Daniel Jurafsky and James H. Martin. *Speech and language processing*. Jan. 2025. URL: <https://web.stanford.edu/~jurafsky/slp3/10.pdf>.
- [12] "Key findings on AI at work and in education<sub>2025</sub>". In: (2025). URL: <https://www.mbs.edu/faculty-and-research/trust-and-ai/Key-findings-on-AI-at-work-and-in-education>.

- [13] Javaid Nabi. “All you need to know about LLM Text Generation - Javaid Nabi - medium”. en. In: *Medium* (Nov. 2024). URL: <https://medium.com/@javaid.nabi/all-you-need-to-know-about-llm-text-generation-03b138e0ed19>.
- [14] Rajni Singh. “Large language models and text summarization: a powerful combination”. en. In: *Medium* (Nov. 2024). URL: <https://medium.com/@singhrajni/large-language-models-and-text-summarization-a-powerful-combination-6400e7643b70>.
- [15] Jenny Smith. “Disadvantages of AI generated code — Medium”. en. In: *Medium* (Nov. 2024). URL: <https://jenny-smith.medium.com/disadvantages-of-ai-generated-code-466485d016d8>.
- [16] Krunal Vaghasiya. *The Latest ChatGPT Statistics and User Trends (2022-2025)*. en-US. June 2024. URL: <https://wisernotify.com/blog/chatgpt-users/>.