Plot Bot

Luke Olson, Alex Busch, Nathan Liew, Sami Aloymari, Muhammad Sharif

Introduction - Background and Novelty

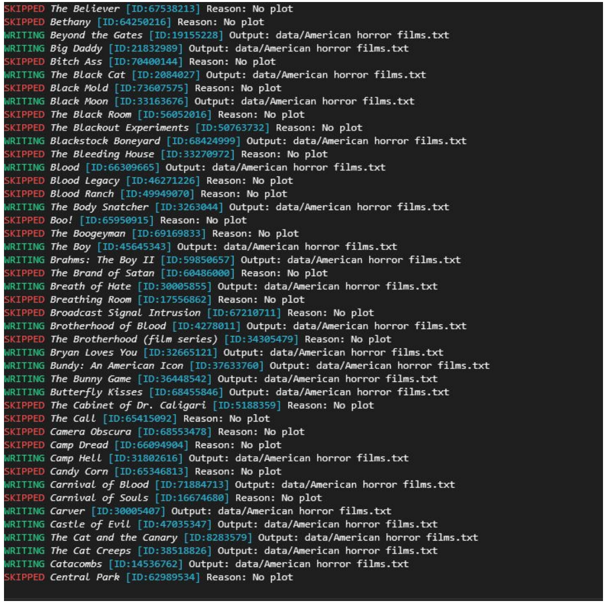
- Why did we do this?
  - Provide script writers with tools so that they can improve their work and creative process.
  - Offer a set of tools that make it easy for users to train a language model on a specific genre
  - Showcase nanoGPT's capabilities
  - Encourage experimentation with GPT models
  - Gain a deeper understanding of transformer based large language models
  - All in all, really fun idea to us
- How is this new?
  - Media generation through machine learning is a new and quickly developing field
  - Thanks to nanoGPT's low complexity and accessibility, the latest language processing advancements are implemented more quickly than other models

Methods – Overview

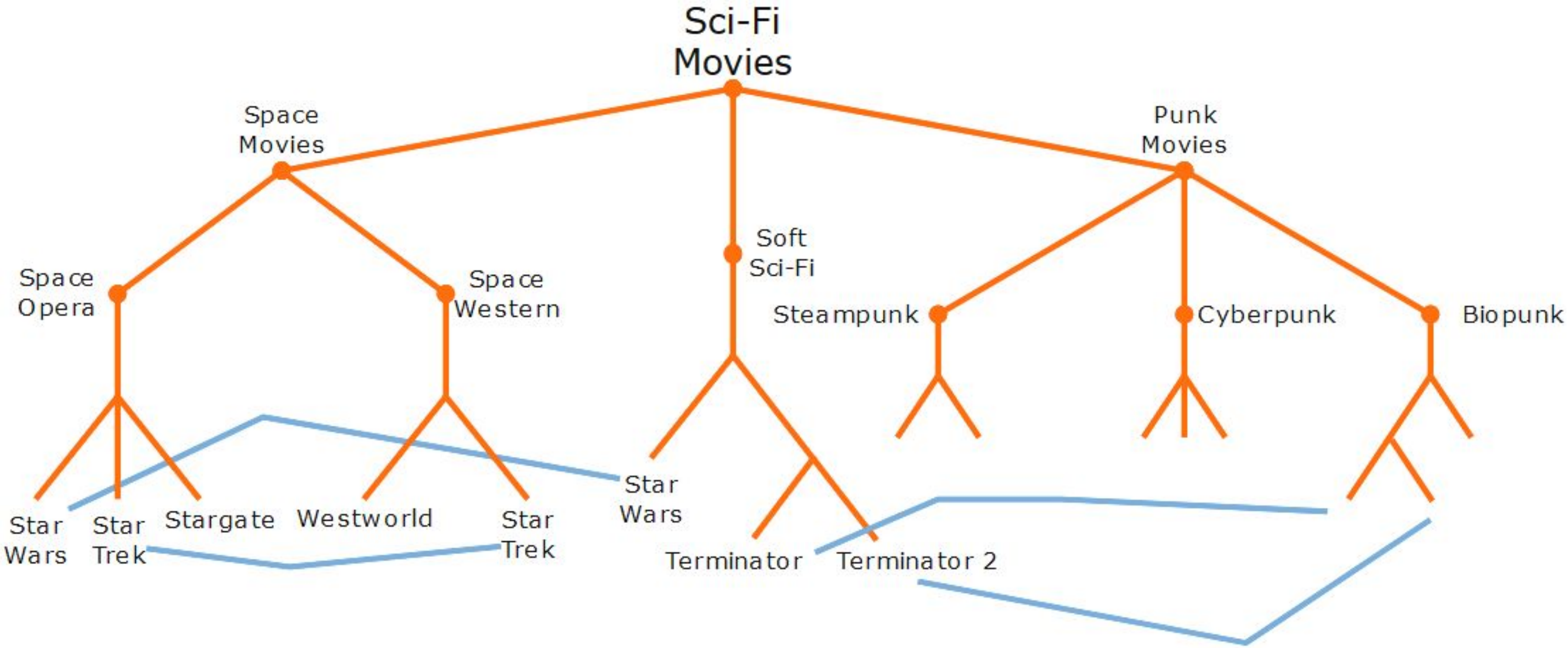
- Pull movie data from Wikipedia API
- Parse plot separated by film title, preprocess data, and compile into a single file
  - Example: All the fantasy film in one text
- Train it to output a new plots using nanoGPT (transformer)
- When the loss function is minimize, we assume that the plot is done
- Evaluate grammar and sentence similarity to check if the results are coherent
- Try combining different genres (Fantasy, Robot , Western etc.).
- Use ChatGPT to summarize output plot and DALL-E to produce a poster

Data - Retrieval

- Data was retrieved from the Wikipedia API utilizing a Python script.
- Script starts in a specified root category/film genre and then recursively iterates through each of its subcategories.
- Certain movies were included under multiple sub-categories, the script keeps track of which movies have already been retrieved.
    - Films that are already stored within the script's memory are skipped
    - Films that do not have a plot included on its Wikipedia page are skipped



```
SKIPPED The Believer [ID:67538213] Reason: No plot
SKIPPED Bethany [ID:64250216] Reason: No plot
WRITING Beyond the Gates [ID:19155228] Output: data/American horror films.txt
WRITING Big Daddy [ID:21832989] Output: data/American horror films.txt
SKIPPED Bitch Ass [ID:70400144] Reason: No plot
WRITING The Black Cat [ID:2084027] Output: data/American horror films.txt
SKIPPED Black Mold [ID:73607575] Reason: No plot
WRITING Black Moon [ID:33163676] Output: data/American horror films.txt
SKIPPED The Black Room [ID:56052016] Reason: No plot
SKIPPED The Blackout Experiments [ID:50763732] Reason: No plot
WRITING Blackstock Boneyard [ID:68424999] Output: data/American horror films.txt
SKIPPED The Bleeding House [ID:33270972] Reason: No plot
WRITING Blood [ID:66309665] Output: data/American horror films.txt
SKIPPED Blood Legacy [ID:46271226] Reason: No plot
SKIPPED Blood Ranch [ID:49949070] Reason: No plot
WRITING The Body Snatcher [ID:3263044] Output: data/American horror films.txt
SKIPPED Boo! [ID:65950915] Reason: No plot
SKIPPED The Boogeyman [ID:69169833] Reason: No plot
WRITING The Boy [ID:45645343] Output: data/American horror films.txt
WRITING Brahms: The Boy II [ID:59850657] Output: data/American horror films.txt
SKIPPED The Brand of Satan [ID:60486000] Reason: No plot
WRITING Breath of Hate [ID:30005855] Output: data/American horror films.txt
SKIPPED Breathing Room [ID:17556862] Reason: No plot
SKIPPED Broadcast Signal Intrusion [ID:67210711] Reason: No plot
WRITING Brotherhood of Blood [ID:4278011] Output: data/American horror films.txt
SKIPPED The Brotherhood (film series) [ID:34305479] Reason: No plot
WRITING Bryan Loves You [ID:32665121] Output: data/American horror films.txt
WRITING Bundy: An American Icon [ID:37633760] Output: data/American horror films.txt
WRITING The Bunny Game [ID:36448542] Output: data/American horror films.txt
WRITING Butterfly Kisses [ID:68455846] Output: data/American horror films.txt
SKIPPED The Cabinet of Dr. Caligari [ID:5188359] Reason: No plot
SKIPPED The Call [ID:65415092] Reason: No plot
SKIPPED Camera Obscura [ID:68553478] Reason: No plot
SKIPPED Camp Dread [ID:66094904] Reason: No plot
WRITING Camp Hell [ID:31802616] Output: data/American horror films.txt
SKIPPED Candy Corn [ID:65346813] Reason: No plot
WRITING Carnival of Blood [ID:71884713] Output: data/American horror films.txt
SKIPPED Carnival of Souls [ID:16674680] Reason: No plot
WRITING Carver [ID:30005407] Output: data/American horror films.txt
WRITING Castle of Evil [ID:47035347] Output: data/American horror films.txt
WRITING The Cat and the Canary [ID:8283579] Output: data/American horror films.txt
WRITING The Cat Creeps [ID:38518826] Output: data/American horror films.txt
WRITING Catacombs [ID:14536762] Output: data/American horror films.txt
SKIPPED Central Park [ID:62989534] Reason: No plot
```


Sci-Fi Movies
Space Movies
Punk Movies
Soft Sci-Fi
Space Opera
Space Western
Steampunk
Cyberpunk
Biopunk
Star Wars
Star Trek
Stargate
Westworld
Star Trek
Star Wars
Terminator
Terminator 2

Data - PreProcessing

- Data retrieved stored the raw contents of the entire wikipedia page. The plot needed to be parsed from the page.
- Text scraped from Wikipedia includes hyperlinks and reference info that need to be preprocessed/cleaned.
- Preprocessed data was appended to a single file that contains all films within that genre separated by film title.

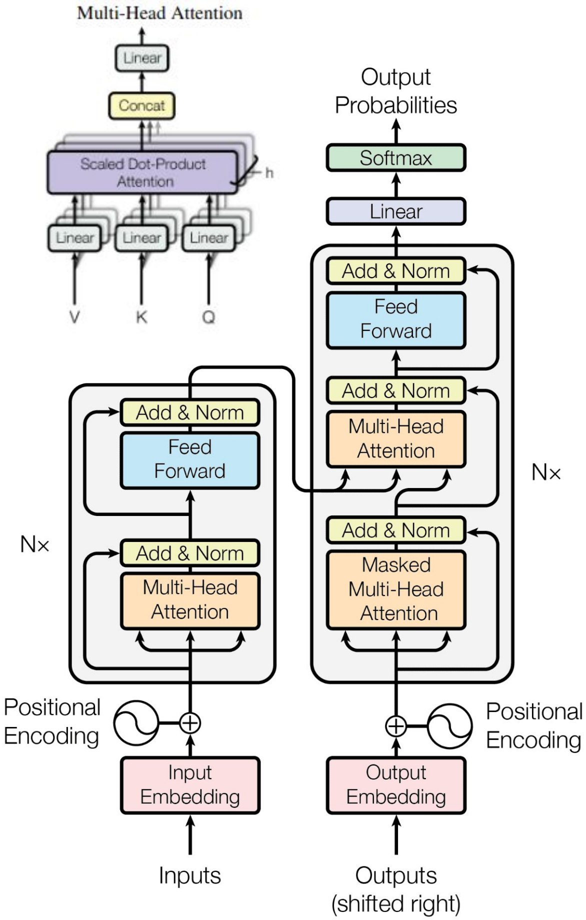
```
Set in the near future, ''2B'' portrays a familiar
decaying world on the cusp of "great transformation and
awesome wonders". The [[story|plot]] is based upon real
[[science]] and evolving technologies. When the world's
first transhuman is created by a renegade corporate CEO
and bioscientist, the foundations of society's beliefs
are threatened in a transhuman world where man merges
with
technology.<ref>[http://www.woodstockfilmfestival.com/f
estival2009/details.php?id=20876 World Premiere],
[[Woodstock Film Festival]], Retrieved 2012-06-05</ref>
```

→

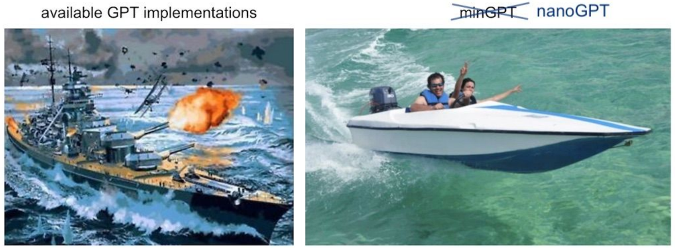
```
Set in the near future, "2B" portrays a
familiar decaying world on the cusp of "great
transformation and awesome wonders". The
story is based upon real science and evolving
technologies. When the world's first
transhuman is created by a renegade corporate
CEO and bioscientist, the foundations of
society's beliefs are threatened in a
transhuman world where man merges with
technology.
```

Baseline - Transformer Architecture

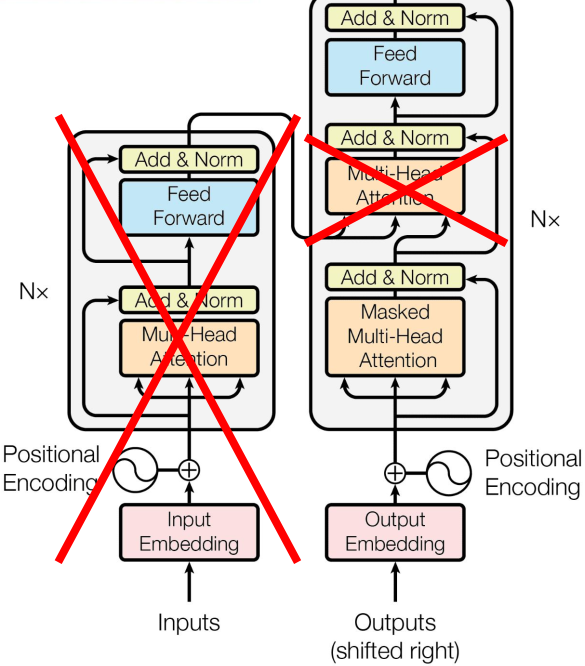
- **Transformer** based neural network architecture.
- This model has proven to be very successful in natural language processing tasks and powers GPT.
- Utilizes a series of stacked layers each with a **self-attention** mechanism.
- Works by assigning an attention score to each token that is then used to determine its relevance to the next output.
    - An attention score is calculated through the dot product of key & query vectors, which is then normalized and multiplied with the value vector
- This approach is much more efficient and allows the model to capture context at a much longer range as opposed to older n-gram language models.



nanoGPT



- Fastest open-source repository for training/fine tuning medium-sized GPTs.
- Provides us with a raw transformer that we can configure and train.
- Minimal and well documented code base, extensible.
- Only provides a decoder, does not include an encoder.
  - Generates an infinite stream of text based upon its trained input.
  - Cannot be prompted, unconditioned output.

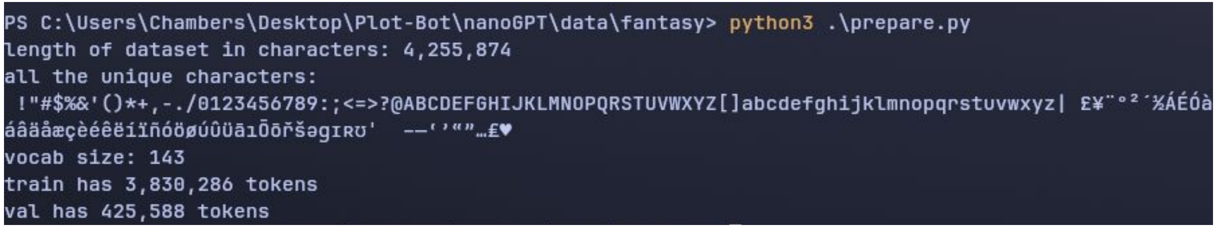


nanoGPT - Configurable Parameters

- **Block size** - Adjusts range of self-attention mechanism.
- **Batch size** - Controls number of training sequences that are processed together.
- **Learning rate** - Controls the step size of the optimization algorithm.
- **Linear decay** - Is a technique used to decrease the learning rate over time during training.
- **Max iters** - Specifies the number of training iterations to perform
- **Temperature** - Adjusts the randomness of the generated output.
- **Seed** - Fixed initial value used for random number generation, allows us to reproduce outputs.
  - Kinda like Minecraft world seeds.

nanoGPT - Data Preparation

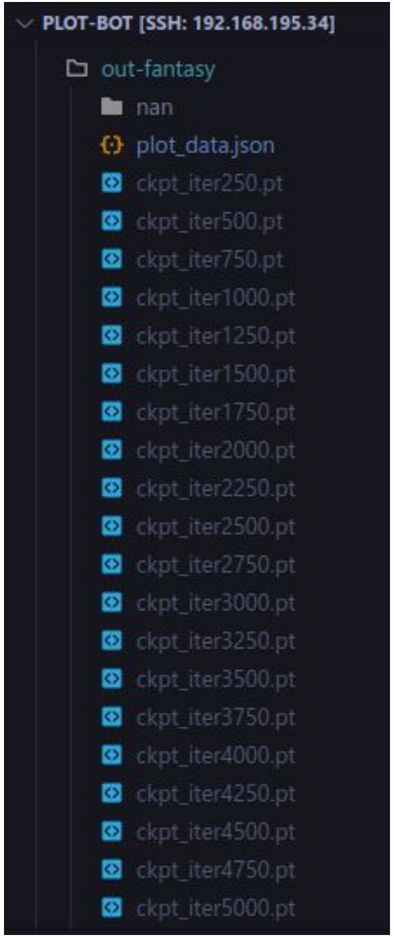
- Training datasets needed to be prepared for training the model.
- **Tokenization** - Text data is split into individual tokens.
  - In the case of our model, the token were individual characters.
- **Encoding** - Encode tokens within vocabulary into a numerical format.
- Determine the language’s vocabulary (list of all unique characters).



```
PS C:\Users\Chambers\Desktop\Plot-Bot\nanoGPT\data\fantasy> python3 .\prepare.py
length of dataset in characters: 4,255,874
all the unique characters:
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[]abcdefghijklmnopqrstuvwxyz| £¥¨°²´½ÁÉÓà
áâäåæçèéêëíïñóöøúûüāıŌōřšəgɪʀʊ'  –—‘’“”…€♥
vocab size: 143
train has 3,830,286 tokens
val has 425,588 tokens
```

nanoGPT - Checkpoint System

- A checkpoint of the language model’s training progress is generated after an elapsed specified number of iterations.
- Checkpoints are used to resume the language models training or generate/sample it’s output.
- Used as data points when performing our evaluations.
- We modified the system to allow us to generate samples every 250 iterations that could then be used to evaluate grammar and sentence similarity.



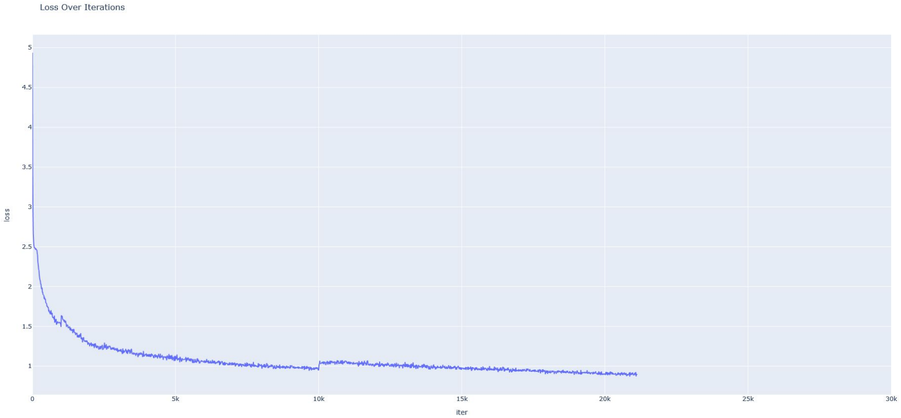
```
PLOT-BOT [SSH: 192.168.195.34]
  out-fantasy
    nan
    plot_data.json
    ckpt_iter250.pt
    ckpt_iter500.pt
    ckpt_iter750.pt
    ckpt_iter1000.pt
    ckpt_iter1250.pt
    ckpt_iter1500.pt
    ckpt_iter1750.pt
    ckpt_iter2000.pt
    ckpt_iter2250.pt
    ckpt_iter2500.pt
    ckpt_iter2750.pt
    ckpt_iter3000.pt
    ckpt_iter3250.pt
    ckpt_iter3500.pt
    ckpt_iter3750.pt
    ckpt_iter4000.pt
    ckpt_iter4250.pt
    ckpt_iter4500.pt
    ckpt_iter4750.pt
    ckpt_iter5000.pt
```

Training - Science Fiction Films

- Initial dataset training, more of a practice run.
- Graphing of loss function had not been implemented at the time of training.
- Dataset contained 5,413,828 characters, largest dataset.
- Vocab size of 149 tokens.
- Trained and sampled at 40,000 iterations.
- Took around two days to complete.

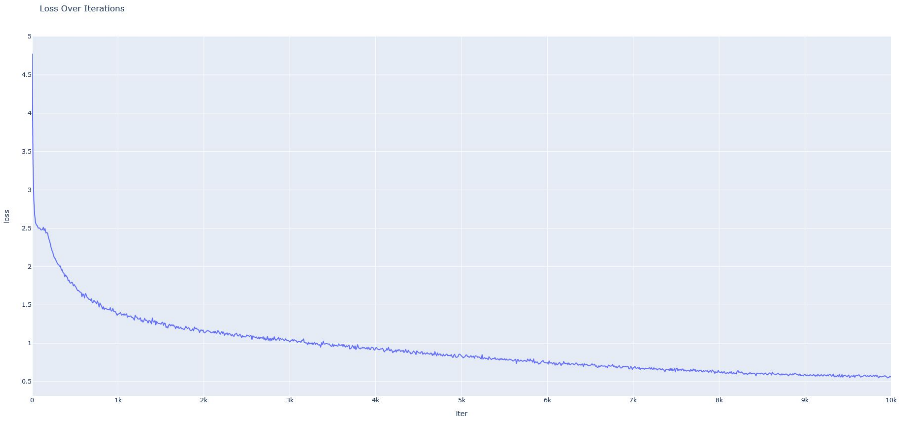
Training - Fantasy Films

- This was our final language model that we performed our evaluations on
- Dataset contained 4,255,874 characters
- Vocab size of 143 tokens
- Trained for 30,000 iterations
- Sampled model at around 20,000 iterations
- We believe that the cut off was caused by too high of a learning rate and overfitting
- Final loss: 0.889
- Took around a day to complete



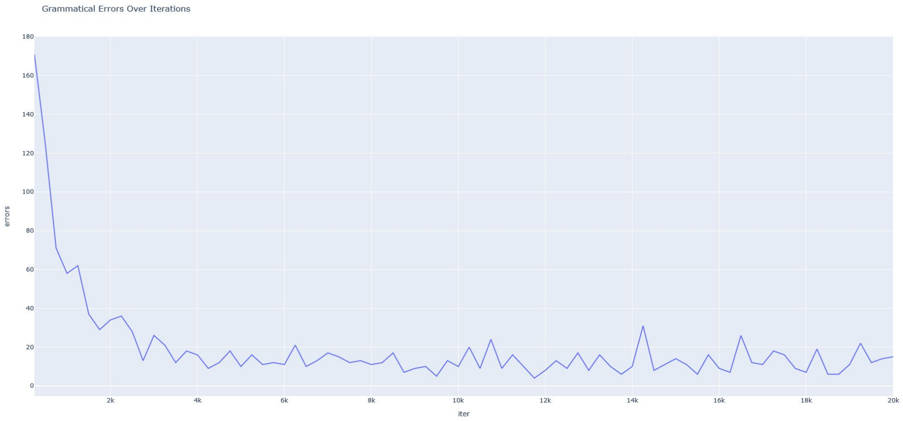
Training - Western Robot Disaster Films

- Combination of multiple datasets, compiled into one file
- Dataset contained 811,970 characters, much smaller
- Vocab size of 104 tokens
- Trained and sampled at 10,000 iterations
- Final loss: 0.569
- Only took a few hours to complete



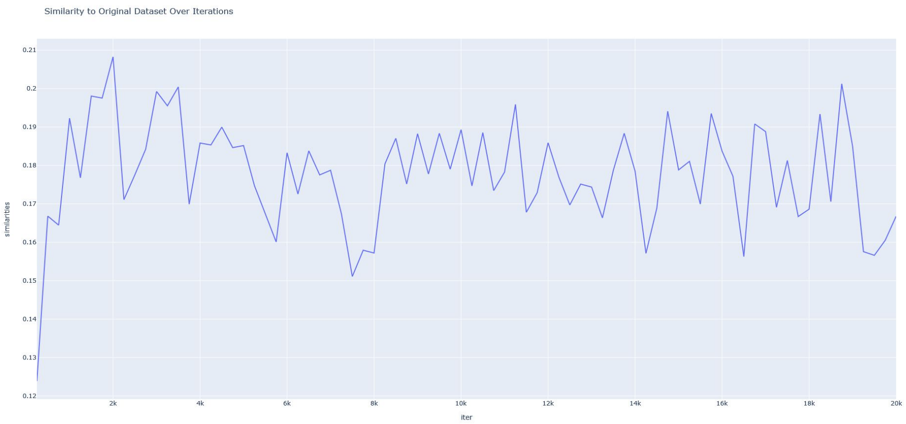
Evaluation - Grammar

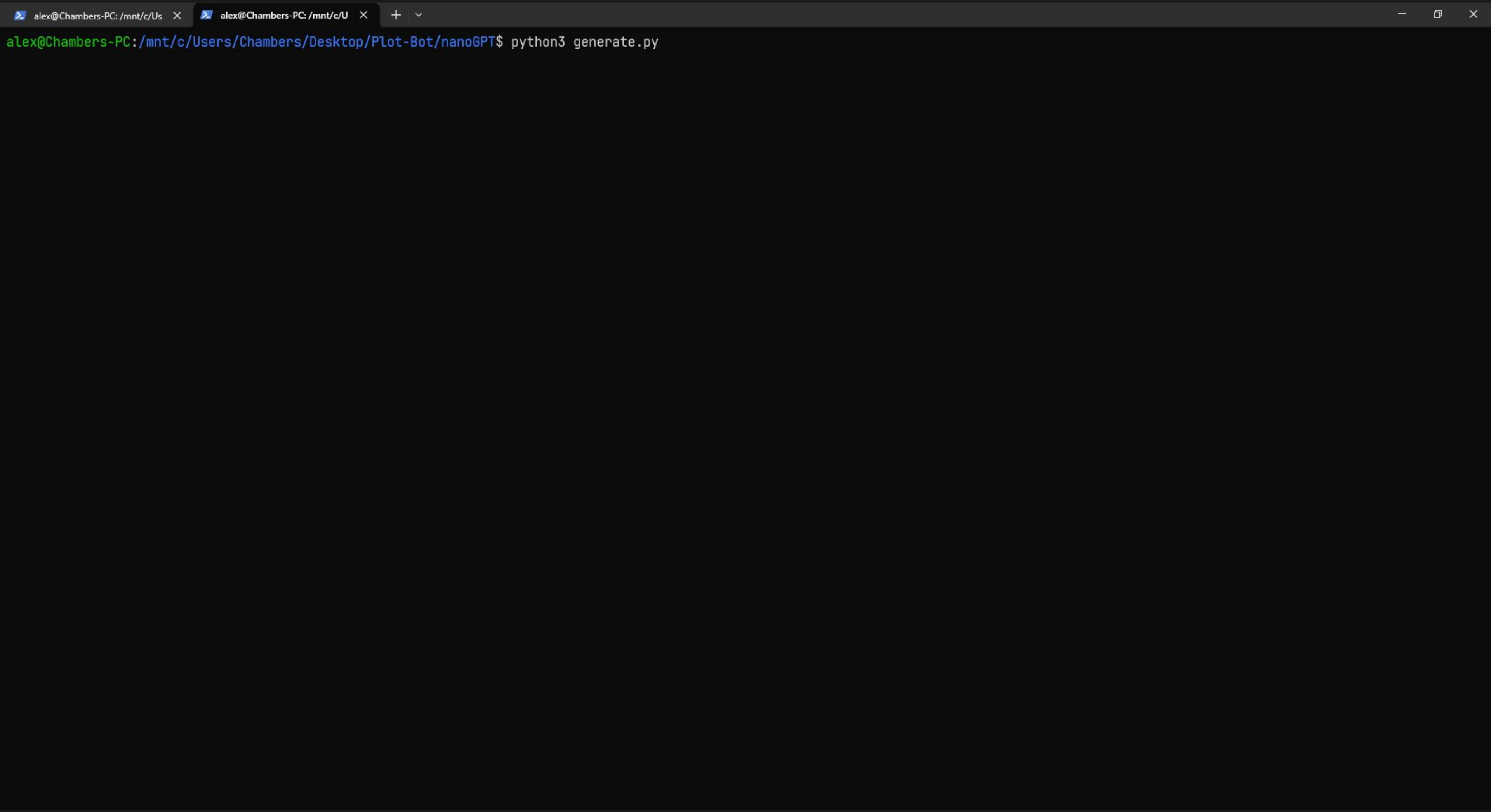
- Utilizes LanguageTool API
- Semantically checks the output
- We are evaluating the errors throughout the entirety of the output
- The fewer errors, the better
- Plateau: ~15 - 30 errors at 20k iterations
- For comparison: I wrote 1500 words over a fantasy topic as fast as I could without backspacing, and ended with ~33 errors



Evaluation - Sentence Similarity

- Tokenizes 2 documents by sentence:
  - Script trained off of
  - Sampling output
- Encodes the tokens with a BERT model modified for sentence derivation
- Utilizes encoded vectors to calculate cosine similarity between documents
- We need a happy medium:
  - Higher similarity shows that sentences are structured similarly, and therefore should make sense
  - Lower similarity shows that our output is unique
- Average: 0.1658
- Story I wrote: 0.1159


Similarity to Original Dataset Over Iterations



```
alex@Chambers-PC:/mnt/c/Users/Chambers/Desktop/Plot-Bot/nanoGPT$ python3 generate.py
```

Results

Early Attempt

Moon

LM: Science Fiction Films

After a disguised Starfleet of the Moon facility, the station's reporter and the Moon becomes a highly extraterrestrial attack by the moon and sends a missile at the same time to save the Moon and the Moon arrive at the same time to find the Moon is a clone of the Moon. The Moon arrives and gives the Moon to the Moon to return to the Moon. The Moon arrives and attacks the Moon and stops the Moon from destroying the Moon. The Moon arrives and begins to explode.

Revenge of The Hotel

LM: Science Fiction Films

"Revenge of the Hotel" follows the story of Tony Kerrier and a group of survivors facing an onslaught of challenges, including terrorist aliens, enemy forces, and various life-threatening incidents. As they strive to save others and themselves, they encounter mysterious entities and make shocking discoveries. The film takes the characters through a series of intense confrontations, as they grapple with their past, confront their fears, and forge unexpected alliances. The movie is packed with action, suspense, and emotional moments, as the characters face one dangerous situation after another in their fight for survival.

Judy Awakens

LM: Western Robot Disaster Films

"Judy Awakens" is an action-packed story set in the aftermath of a disastrous earthquake that causes a series of accidents and emergencies. The characters, including John, Claudia, Fin, and Jesse, must navigate through various dangerous situations, such as malfunctioning equipment, fires, and collapsing buildings. The film features numerous dramatic rescue attempts, including a daring helicopter operation, as well as personal struggles and relationships among the characters. As the story unfolds, the characters form alliances and encounter unexpected twists, ultimately learning the importance of teamwork and resilience in the face of adversity.

Spongebob’s Ghost

LM: Fantasy Films

"SpongeBob's Ghost" follows the eccentric adventure of SpongeBob as he tries to complete a mission assigned by his stepmother. Along the way, he encounters a mysterious girl named Havey Valley, who helps him in his quest. As they face various challenges, including gangsters led by Captain Kyra and supernatural threats, they must locate a powerful artifact called the Kragle. Amidst their journey, they uncover hidden truths and confront a menacing witch. The story intertwines themes of friendship, courage, and self-discovery.

Kung Fu

LM: Fantasy Films

In Kung Fu, Puppy Character Kung Cass and her best friend Randy embark on a thrilling adventure to stop three hybrid demons. They travel to the future and team up with vibrant character Boo, John Shoos, and Annie Munny. After foiling the villain Viktoria's plan, the group faces various challenges with their new friends Rocky, Chase, and Bullwinkle. Together, they fight against the Rourans and make their way to London World. Throughout their journey, they uncover the truth about Bullwinkle's past and encounter magical creatures, ultimately reuniting with their loved ones.

The Queen of Water Town

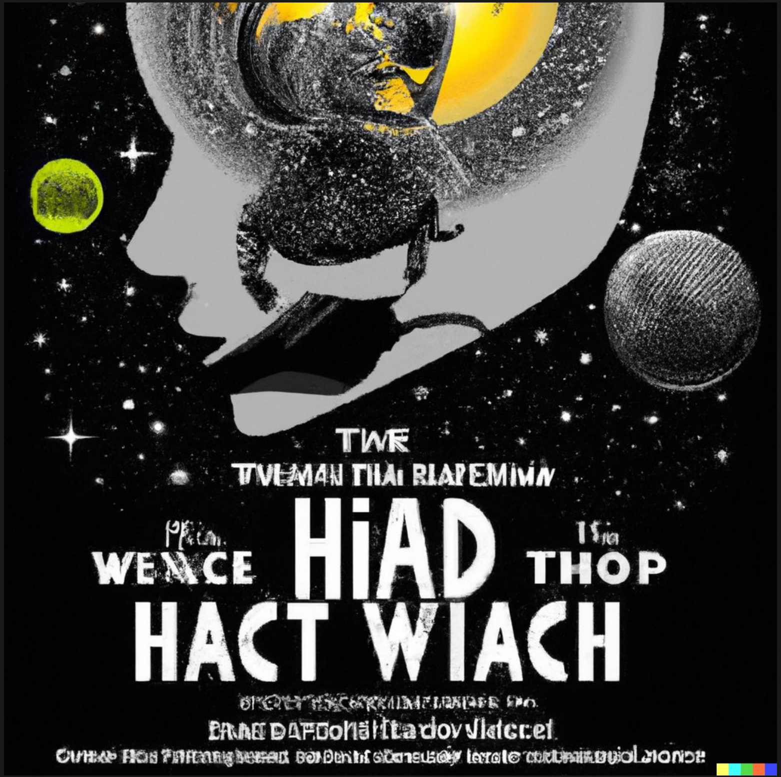
LM: Fantasy Films

In "The Queen of Water Town," the story follows a group of characters through a magical world filled with danger and enchantment. The protagonist, Longbottom, gets involved with the wizard King Van Clare, who seeks to create chaos. Characters like Lumpy, Heffalump, and Belladonna each face their own challenges and adversaries. Throughout the story, they navigate various locations, including enchanted forests, deserts, and a mysterious city. The characters discover alliances, uncover hidden powers, and confront their own fears, ultimately banding together to defeat the evil forces that threaten their world.

Space Witch's Head

LM: Fantasy Films

In "The Space Witch's Head," Scott Harold and his partner Ran discover a picture of Elizabeth Switch and her husband. Scott receives a new chance to fly and encounters various characters like Mr. Chairman, Larry, and Cheetah. Scott is warned about a transformation and seeks to find a solution. Meanwhile, Tarzan and Jane face challenges of their own as they try to protect their land. The two groups eventually meet and must work together to save their worlds.



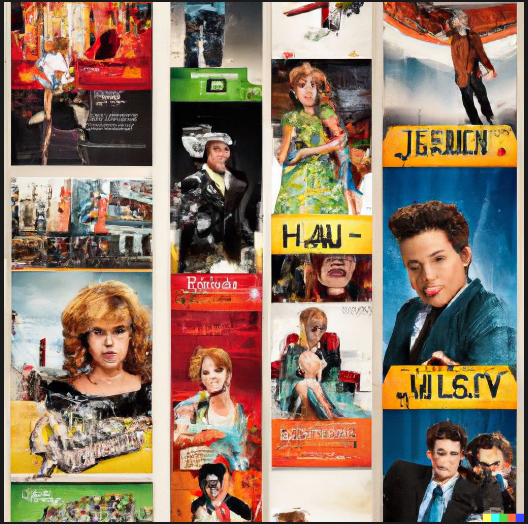
HALL
THAW VPACE
WICHD SHAW

HAPT
TIME AFAID
WITHACICH
WEEAD HACH

TTOLE VIGAL HOP

ACURITY

LION ON CRIAS


JEEDICN
HAU-
WILS.IV

Discussion - How can outputs be improved?

- **Dataset quality -** An ideal dataset is a large, diverse, and relevant to what you are planning to achieve.
- **Training duration -** Adequate training time is needed for the model to learn and generalize well. Insufficient training will result in underfitting, while excessive training may lead to overfitting.
- **Language model quality -** With a more advanced model such as GPT, prompts can be used to let the model know what parts do and don't work. This also allows for more experimentation and fine tuning
- **Computing power -** A large GPU is required to generate the output in a reasonable amount of time, each checkpoint takes up a relatively large amount of disk space as well

Conclusion - Where do we go from here?

- Can be enjoyed by the general public through generating interesting scripts using their favorite genres, characters, and/or movies
- We can send this to script writers to help lessen the workload of their jobs
- With more training time and data…
  - Can be paired with voice-generating AI to visualize how a script for a movie can look/sound
  - Can be used to write entire, objectively good scripts
  - May even be used to phase out most script writers
  - Generated movie plots can be summarized and fed into DALL-E to generate posters for the movie

Questions?