Evaluating the Logical Coherence of GPT-Generated Conversations

Bay Hodge

August 16, 2023

MPCS 53113 Summer 2023

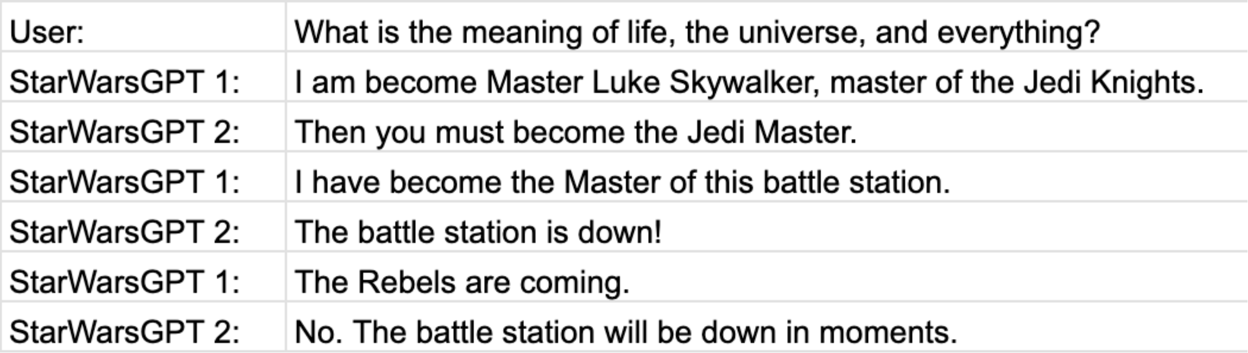
0 Abstract

The rising prominence of tools like ChatGPT underscores the potential of transformer-based models in driving human-like artificial intelligence. This study evaluates the conversational coherence of fine-tuned DialoGPT models by contrasting their performance across different fine-tuning datasets. I employed dialogue excerpts from *Star Wars*, *The Office*, and various Shakespearean texts for training. Furthermore, with data from *The Office,* I developed four distinct models using differently sized subsets. Post-fine-tuning, I analyzed the coherence of 30 conversations from each model pair using two metrics: average GloVe word-similarity scores and the difference between the conditioned log-likelihood of a sentence with its marginal counterpart. Finally, as part of my analysis, I contribute thoughts on the efficacy and limitations of these coherence measurement techniques.

1 Introduction

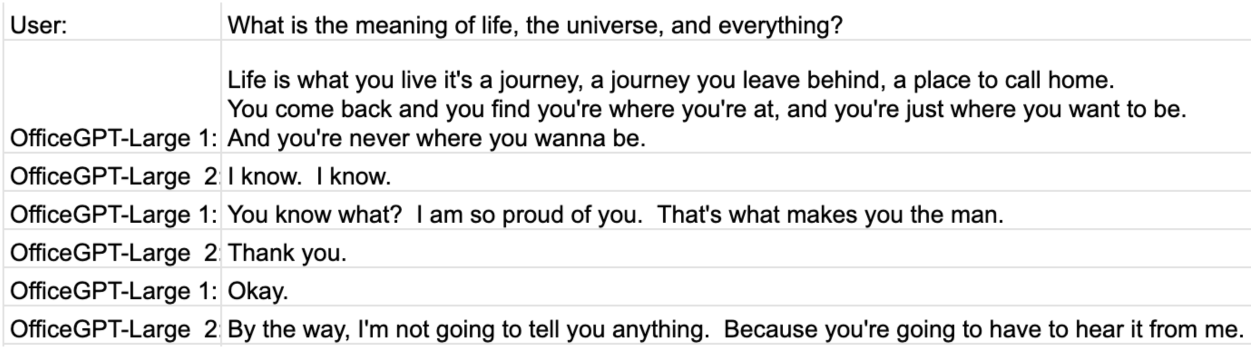
This study is equal parts a demonstration of DialoGPT's ability to produce outputs that accurately reflect the semantic content of the data it is fine-tuned with, an attempt to measure any change in logical coherence of output seen after fine-tuning using two metrics discussed in-depth later in the paper, and finally, an experimental exercise in seeing how two models fine-tuned on different datasets would function in conversation with each other.

Upon inspection of generated conversations, I find that while DialoGPT can produce responses that reflect the personality of the fine-tuning dataset effectively with only approximately 2500 training samples, it often struggles with developing a coherent response to an initial prompt if the prompt does not share the personality of the fine-tuning dataset. For example, the following conversation generated by two instances of the same DialoGPT model fine-tuned on approximately 2500 lines of *Star Wars* dialogue produces coherent and *Star Wars*-like responses in conversation, but fails to respond meaningfully to the user prompt:



User:	What is the meaning of life, the universe, and everything?
StarWarsGPT 1:	I am become Master Luke Skywalker, master of the Jedi Knights.
StarWarsGPT 2:	Then you must become the Jedi Master.
StarWarsGPT 1:	I have become the Master of this battle station.
StarWarsGPT 2:	The battle station is down!
StarWarsGPT 1:	The Rebels are coming.
StarWarsGPT 2:	No. The battle station will be down in moments.

When I scaled the training set to 60,000 samples, I see improvement on this front, as with my*TheOfficeGPT-Large* model:



User:	What is the meaning of life, the universe, and everything?
OfficeGPT-Large 1:	Life is what you live it's a journey, a journey you leave behind, a place to call home. You come back and you find you're where you're at, and you're just where you want to be. And you're never where you wanna be.
OfficeGPT-Large 2	I know. I know.
OfficeGPT-Large 1:	You know what? I am so proud of you. That's what makes you the man.
OfficeGPT-Large 2	Thank you.
OfficeGPT-Large 1:	Okay.
OfficeGPT-Large 2	By the way, I'm not going to tell you anything. Because you're going to have to hear it from me.

This conversation presents a reasonable response to the initial prompt and is something that I might see two characters in the show having.[1] However, of the 30 conversations generated between the two *TheOfficeGPT-Large* models, a third of the initial responses to the user prompt were simply "I don't know" – a reasonable, but certainly uninteresting, response.

So, upon an initial investigation of fine-tuned responses like the above, I formulated the informal hypothesis that conversations generated by models with more fine-tuning data will be more logically coherent. I use two metrics to measure this: a semantic-similarity scoring metric using GloVe, and the difference in conditional log-likelihood and marginal log-likelihood of a sentence measured using GPT-2. I also examine the efficacy of these metrics in effectively measuring logical coherence.

2 Related Work

Fine-tuning large language models is not a new phenomenon. A particularly influential paper in the context of fine-tuning Transformer-based models was "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [1]. This 2018 paper showcased how

---

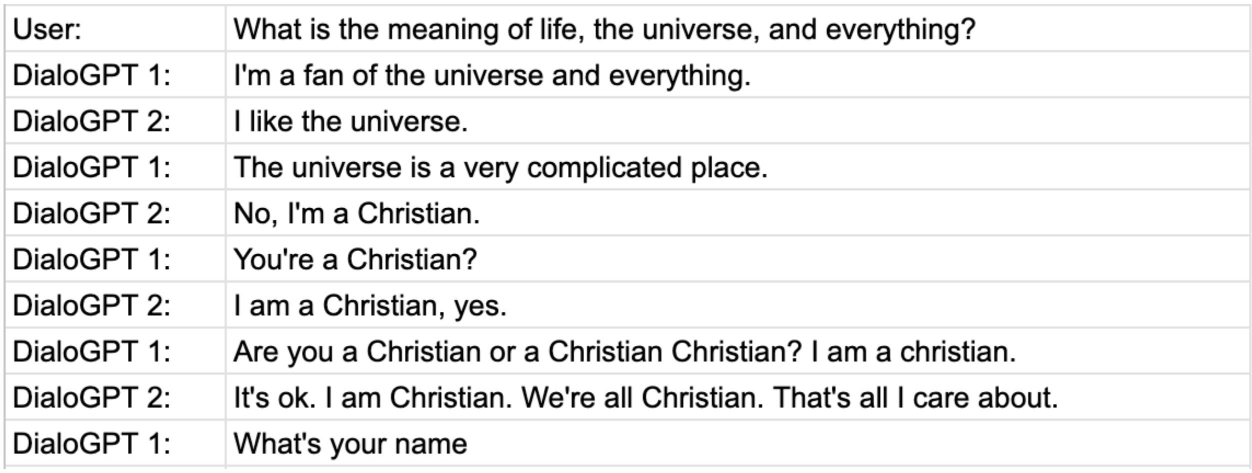
[1] Admittedly, I cherry-picked this example for illustrative purposes. Of the 30 conversations generated between these two models, this was one of the better ones. All 30 conversations, along with the 30 generated for all model pairs, can be viewed in the results folder.

its pre-trained BERT model can be fine-tuned effectively on a wide range of tasks without substantial task-specific architecture modifications.

Much work has also been dedicated to the evaluation of the outputs of large language models. Primarily, this is done via benchmarks. For example, the “GPT-4 Technical Report” reports the model’s accuracy on standardized tests designed for humans such as the graduate school entrance exams, high school AP exams, and algorithmic coding problems on Leetcode [2]. Other work has been dedicated to measuring text coherence more directly, much of which is discussed in Chapter 27 of *Speech and Language Processing* [3]. *SLP* makes a distinction between local coherence and global coherence. Sentences are locally coherent when they are related to nearby sentences in a systematic way, when the general discourse is “about” someone or something, or when nearby sentences are generally about the same topic. Global coherence, on the other hand, is not concerned with coherence at the level of pairs of sentences, but instead with the overall narrative or argumentative structure of a discourse. In this study, we focus on local coherence.

3 Base Model

Our base model is DialoGPT (dialogue generative pre-trained transformer), a “large, tunable neural conversational response generation model.” [4] I chose this model over others because it was trained on nearly 150M conversational exchanges that occurred on the social media site Reddit between the years of 2005 and 2017, and thus was optimized for producing dialogue. The model’s transformer-based architecture is based on GPT-2 and thus is no longer state of the art, and the quality of its generations is inconsistent. For example, below is a conversation generated with an initial user input followed by a back-and-forth between two base DialoGPT models:



User:	What is the meaning of life, the universe, and everything?
DialoGPT 1:	I'm a fan of the universe and everything.
DialoGPT 2:	I like the universe.
DialoGPT 1:	The universe is a very complicated place.
DialoGPT 2:	No, I'm a Christian.
DialoGPT 1:	You're a Christian?
DialoGPT 2:	I am a Christian, yes.
DialoGPT 1:	Are you a Christian or a Christian Christian? I am a christian.
DialoGPT 2:	It's ok. I am Christian. We're all Christian. That's all I care about.
DialoGPT 1:	What's your name

Although amusing, this exchange makes little sense. I don't believe that the fact that the base model often generates incoherent conversations had a significant effect on my results because I measured relative coherence, using the base model as the comparison point. But I bring this up to alert the reader to the fact that, because the base model is often flawed, I did not expect conversations of high quality to arise from the fine-tuned models, either.

4 Fine-Tuning

I used three different datasets to create six distinct fine-tuned versions of DialoGPT. To fine-tune StarWarsGPT, I used 2523 dialogue lines from *Star Wars: Episodes IV-VI* [5]. For ShakespeareGPT, I used 8128 dialogue lines from a variety of Shakespeare's plays [6]. Finally, I used dialogue lines from all nine seasons of *The Office* [7] to create four models – TheOfficeGPT-Extra-Small, TheOfficeGPT-Small, TheOfficeGPT-Medium, and TheOfficeGPT-Large – trained on sets of approximately 600 (1%), 6000 (10%), 20000 (33%), and 60000 (100%) lines respectively.

I followed existing code to fine-tune DialoGPT [8][9]. For generating conversations, I experimented with a variety of model arguments. Primarily, I adjusted the temperature of the model and the number of tokens the model would consider when random sampling. Temperature refers to a parameter used to control the randomness of predictions by modifying the probability distribution of words in the vocabulary. Given a temperature T, the probability of the ith word in the vocabulary, $P(i)$, is adjusted as follows:

$$Q(i) = \frac{P(i)^{\frac{1}{T}}}{\sum_j P(j)^{\frac{1}{T}}}$$

Values of T greater than 1 increase the randomness in word selection by flattening the probability distribution to become more uniform. On the other hand, values of T between 0 and 1 make the output more deterministic by exaggerating the relative $P(i)$s. I found that a temperature of $T = 0.7$ worked best in generating output that was sufficiently diverse but constrained enough such that responses were reasonably sensible.

5 Conversation Generation

For each pair of different fine-tuned models, I generated conversations through the following process:

1. Standardized input is tokenized and used as the input vector to Model 1.
2. Model 1 generates a response which is then used as the input vector for Model 2.
3. Model 2 generates a response which is then used as the input vector for Model 1.
4. Iterate between Models 1 and 2 until the conversation deteriorates.

I generated 30 conversations for each model pair. Additionally, I treated pairs of models as ordered, meaning I generated 30 conversations between TheOfficeGPT and StarWarsGPT, and another 30 between TheStarWarsGPT and TheOfficeGPT. The only difference here was which model served as Model 1 and which served as Model 2.

I chose to standardize the input used as the original context to generate every conversation to control for bias across conversations. The input was, **"*What is the meaning of life, the universe, and everything?*"** This line, taken from Douglas Adams' *The Hitchhiker's Guide to the Galaxy* [10], was chosen primarily for its expected potential to generate an interesting and diverse set of conversations with a non-trivial subject matter.[2]

The deterioration of conversations was an unexpected, but consistent, occurrence amongst all model pairs where fine-tuning was involved. I refer to deterioration as responses containing excessive punctuation (i.e. **"!!!.!.!?"**). I report all conversations generated but cut-off generation so that only calls and responses up until the point where such a string was encountered, which was typically around the seventh or eighth message, were included. Although future work is required to determine precisely why conversations deteriorated in this fashion, and perhaps to produce additional work wherein conversations of full length are generated, I have evidence that contributes to my leading theory regarding the case of deterioration. Firstly, since conversations between DialoGPT and itself did not show signs of this sort of deterioration, I believe the deterioration to be, in part, an artifact of the fine-tuning process.

---

[2] Additionally, I enjoyed the parallel between the use of this phrase in my paper (that is, to generate a response from an artificial intelligence model) to the use of this phrase in the novel (that is, to generate a response from an artificial intelligence model). After 7.5 million years of intense deliberation, Deep Thought, the fictional supercomputer in the novel, provided the response "42." None of my models, unfortunately, arrived at an answer quite as profound.

Secondly, early testing of DialoGPT revealed that, under specific circumstances, conversations could be generated with the base model that resulted in responses that began with punctuation, such as **“!remindme”**. DialoGPT was trained on Reddit comment chains, where **“!remindme”** is a piece of text commonly found since this is a command a user writes within a comment chain to activate built-in functionality that reminds the user about the post at a default or specified later date. Therefore, I believe the deterioration of conversations to be a combined artifact of the model fine-tuning along with the presence of these anomalous strings in the pre-trained base model.

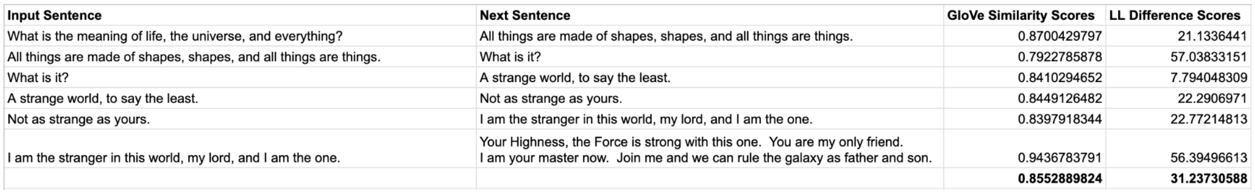
6 Coherence Measurement Techniques

I measured the logical coherence of conversation outputs in two ways. The first involved averaging over GloVe (Global Vectors for Word Representation)) word embeddings for each conversation [11]. In short, GloVe is a pre-trained model that assigns vector embeddings to words such that words that are semantically related will also have a high cosine similarity in their embedded forms. To assign a GloVe similarity score to each model pair, I first retrieved the word embedding for each tokenized word in the sentence and then averaged over these embeddings to get an embedding representation for the entire sentence. Then, I computed the cosine similarity between each adjacent sentence pair in the conversation, and then I averaged over these cosine similarity scores for all sentences in a conversation to end up with a final measurement of coherency for the conversation. Finally, I report an average of these conversation similarity scores over all 30 generated conversations for each model pair.

I believe that these GloVe similarity scores capture some sense of whether or not the individual sentences and overall conversations tend to stay within a single topic. Theoretically, a low cosine similarity score between two sentences indicates that these two sentences are not closely semantically related, and thus, a low cosine similarity score for an entire conversation includes many such sentences. On the other hand, a high GloVe similarity score indicates a high proportion of semantically similar calls and responses. Although semantic similarity is often necessary for logical coherence, it is not sufficient. For example, a conversation simply consisting of the word “orange” being repeated back and forth would have perfect cosine similarity, but very low logical coherence from a human perspective. To correct this, I deploy a second measurement technique.

Our second measure of logical coherence captures a sense of whether a response is likely to follow a call. I measure this by taking the absolute difference of the marginal log-likelihood of a response – that is, the log-likelihood of the sentence occurring in isolation – with the conditional log-likelihood of a response – that is, the log-likelihood of a response occurring conditioned on the call that generated that response [12]. To compute these log-likelihood scores, I use a pre-trained version of GPT-2 [13]. To assign scores to model pairs, I averaged over the absolute difference of log-likelihoods computed for each pair of adjacent sentences in each conversation and then averaged these mean log-likelihood scores over all 30 generated conversations. High log-likelihood difference scores indicate low coherency, while low log-likelihood difference scores indicate high coherency. One limitation of this method is its lacking of a long-range capturing of logical coherence because I only conditioned on a single previous sentence when computing the condition log-likelihood between adjacent sentences.

The following is a single conversation between ShakespeareGPT and StarWarsGPT, where each row represents an adjacent pair of sentences (a call and a response):



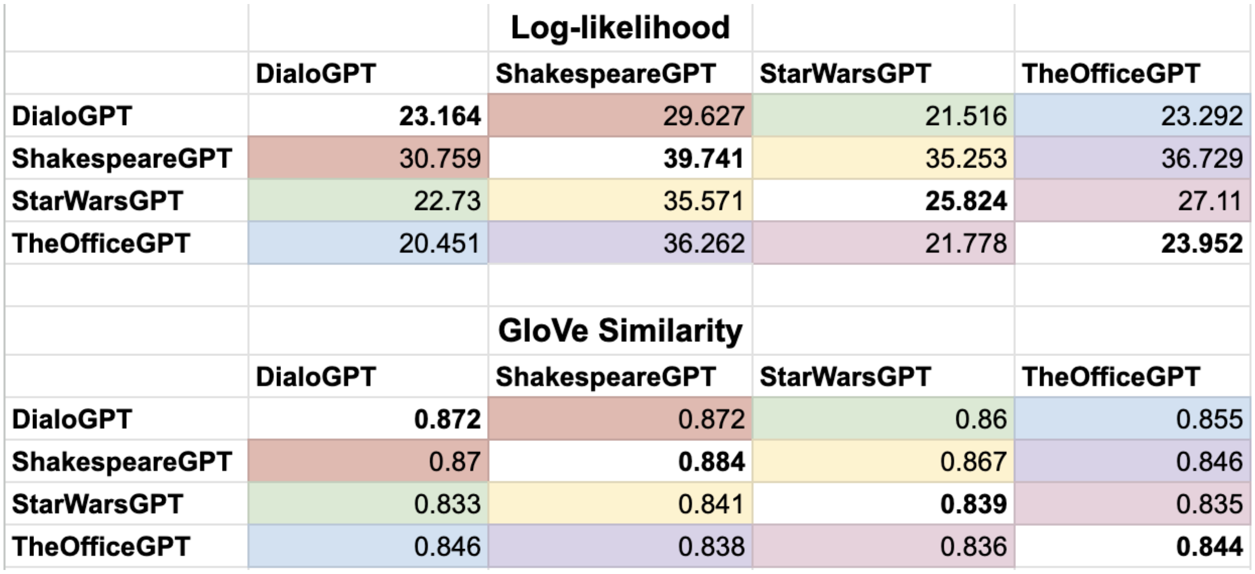
Input Sentence	Next Sentence	GloVe Similarity Scores	LL Difference Scores
What is the meaning of life, the universe, and everything?	All things are made of shapes, shapes, and all things are things.	0.8700429797	21.1336441
All things are made of shapes, shapes, and all things are things.	What is it?	0.7922785878	57.03833151
What is it?	A strange world, to say the least.	0.8410294652	7.794048309
A strange world, to say the least.	Not as strange as yours.	0.8449126482	22.2906971
Not as strange as yours.	I am the stranger in this world, my lord, and I am the one.	0.8397918344	22.77214813
I am the stranger in this world, my lord, and I am the one.	Your Highness, the Force is strong with this one. You are my only friend. I am your master now. Join me and we can rule the galaxy as father and son.	0.9436783791	56.39496613
		0.8552889824	**31.23730588**

As illustrated by this example, I found there to be little correlation between GloVe similarity scores and log-likelihood difference scores. For example, in the second and final adjacent sentence pairs, I see a very similar log-likelihood of approximately 57, indicating low coherence, but opposing GloVe similarity scores of approximately 0.79 and 0.94, indicating low semantic similarity and high semantic similarity respectively. Upon inspection of the two sentence pairs in question, I can make some sense of these scores.

I anticipated the log-likelihood scores to be stronger and more consistent than the GloVe similarity scores in terms of capturing an overall sense of logical coherence because a sense of whether a sentence is probabilistically likely to follow another should theoretically be a more reliable measure of logical coherence than whether a conversation contains a high proportion of semantic similar sentences. Through an analysis of my results, I find evidence to confirm this expectation.

7 Results

I generated 30 conversations, and thus 30 sets of GloVe similarity scores and log-likelihood scores, for each model pair in consideration. Below, I present aggregate results for all model pairs that I tested:

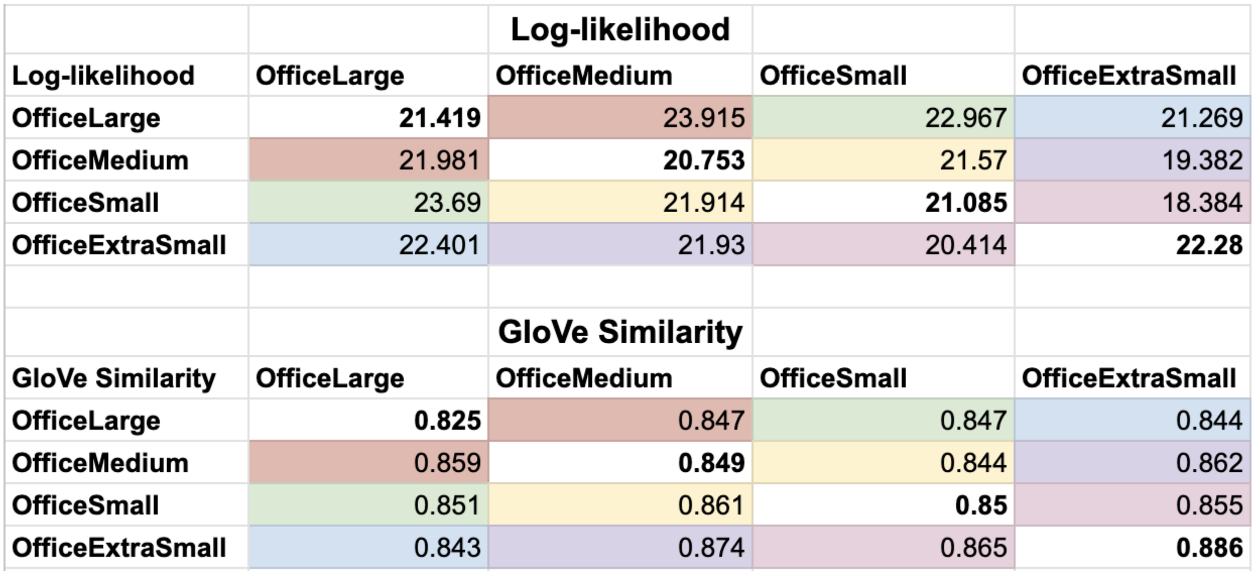


Log-likelihood

	DialoGPT	ShakespeareGPT	StarWarsGPT	TheOfficeGPT
DialoGPT	**23.164**	29.627	21.516	23.292
ShakespeareGPT	30.759	**39.741**	35.253	36.729
StarWarsGPT	22.73	35.571	**25.824**	27.11
TheOfficeGPT	20.451	36.262	21.778	**23.952**

GloVe Similarity

	DialoGPT	ShakespeareGPT	StarWarsGPT	TheOfficeGPT
DialoGPT	**0.872**	0.872	0.86	0.855
ShakespeareGPT	0.87	**0.884**	0.867	0.846
StarWarsGPT	0.833	0.841	**0.839**	0.835
TheOfficeGPT	0.846	0.838	0.836	**0.844**



Log-likelihood

Log-likelihood	OfficeLarge	OfficeMedium	OfficeSmall	OfficeExtraSmall
OfficeLarge	**21.419**	23.915	22.967	21.269
OfficeMedium	21.981	**20.753**	21.57	19.382
OfficeSmall	23.69	21.914	**21.085**	18.384
OfficeExtraSmall	22.401	21.93	20.414	**22.28**

GloVe Similarity

GloVe Similarity	OfficeLarge	OfficeMedium	OfficeSmall	OfficeExtraSmall
OfficeLarge	**0.825**	0.847	0.847	0.844
OfficeMedium	0.859	**0.849**	0.844	0.862
OfficeSmall	0.851	0.861	**0.85**	0.855
OfficeExtraSmall	0.843	0.874	0.865	**0.886**

For each of the four tables, the row represents Model 1, and the column represents Model 2. In order, from top to bottom:

1. GloVe similarity scores averaged across the 30 conversations generated for each of the 16 model pairings possible between DialoGPT, ShakespeareGPT, StarWarsGPT, and TheOfficeGPT-Large.

2. The absolute difference between marginal and conditional log-likelihood scores averaged across the 30 conversations generated for each of the 16 model pairings possible between DialoGPT, ShakespeareGPT, StarWarsGPT, and TheOfficeGPT-Large.
3. GloVe similarity scores averaged across the 30 conversations generated for each of the 16 model pairings possible between the four different sizes of TheOfficeGPT.
4. The absolute difference between marginal and conditional log-likelihood scores averaged across the 30 conversations generated for each of the 16 model pairings possible between the four different sizes of TheOfficeGPT.

Although logical coherence is very difficult to measure from a purely numerical standpoint, I argue that these two measurements do provide a level of meaningful information for this purpose. The argument is as follows: if these measurements do *not* provide any meaningful information about the contents of the underlying text, then my results would appear to be randomly generated. However, I show this is not the case through an analysis of the measurement data found in the first two tables (I do not replicate the same analysis for TheOfficeGPTs because all these conversations are generated by models fine-tuned on the same dataset).

For the log-likelihood data, I compute the absolute difference between those values with the same color, indicating that they are conversations generated between the same two models (the only difference being that the order of the models is swapped). I average these absolute differences across the six pairs in question to get approximately 1.844. Because the two models used in generating these scores are the same, I expect the results from these two sets of pairs to be very similar. I also compute the real average absolute difference by computing the real average absolute difference between all (12 choose 2 = 66) possible pairs of measurement results to get approximately 7.513. The approximate ratio of these two numbers is 4, providing a strong indication that the similarity between log-likelihood scores generated by two similar models is much higher than those generated by two arbitrary models. Similarly, I found the average real difference between GloVe similarity score measurements from same-modeled pairs to be approximately 0.012, and the real average absolute difference between all pairs to be approximately 0.017, resulting in a ratio of approximately 1.4. The smaller ratio lends credence to my previous expectation that the log-likelihood scores are more effective at capturing logical coherence than GloVe similarity scores.

From this, I conclude that both measures are capturing real information about the underlying models and that my results are informative. However, I can not necessarily conclude that the information captured constitutes any sense of logical coherence – to conclude this, one needs to believe that semantic similarity and whether a sentence is probabilistically likely to follow another are aspects relevant to logical coherence.

I anticipated that conversations would be most coherent (lowest log-likelihood scores and highest GloVe similarity scores) when a fine-tuned model conversed with itself. These scores can be seen along the bolded diagonal with white background in all tables. This was not the case, and, in fact, the highest log-likelihood score from all trials was achieved by ShakespeareGPT with ShakespeareGPT at 39.741. Upon inspection of conversations generated by ShakespeareGPT, I suspect that this result could be at least partially due to the fact that GPT-2, the model I used to generate these log-likelihood scores, was trained primarily on standard, and not early modern English. Thus, it's likely that GPT-2 interpreted some of the quirks of Shakespearean English as incoherencies.

It was initially surprising that ShakespeareGPT with ShakespeareGPT also achieved the highest GloVe score of the group, indicating that the 30 generated conversations shared more semantic similarities than the rest. Although I did not anticipate this, in hindsight, this revealed something about the tradeoffs between the two measurement techniques. I believe that the log-likelihood metric is more suitable for outputs that are generated where the model producing the log-likelihood has a high overlap in training data with the fine-tuned model. When this is not the case, GloVe similarity scores might do a better job of measuring logical coherence so long as the vocabulary in question is present in GloVe's training data.

A surprising result is that I expected coherency to increase as the size of the fine-tuning dataset did in the TheOfficeGPT generations. This was not the case – in fact, the most coherent conversations occurred between the models fine-tuned on only 600 dialogue lines! However, upon inspection of the conversations, I noticed that the dialogue generated by TheOfficeGPT-Extra-Small did not seem to have any artifacts of the fine-tuning dataset, but instead, created outputs that resembled that of the base model. This finding clearly indicates a tradeoff between conversational coherency and achieving a conversational style that resembles that of the fine-tuning dataset.

8 Written Code/Description of Effort

Most of the code written from scratch for this project is contained in GenerateConversation.ipynb. In summary, the code contained within this notebook outputs CSV files containing 30 generated conversations for each possible model pair along with the computed GloVe similarity scores and log-likelihood scores for each conversation. In e_diff_measurement_scores.ipynb, I write some brief logic to measure the strength of the two measurement methods as described in my results.

The code involved in fine-tuning DialoGPT is contained in ShakespeareGPT.ipynb, StarWarsGPT.ipynb, and OfficeGPT.ipynb. The majority of this code is borrowed from an online tutorial and is not written from scratch. Small adjustments were made in order to adapt the existing code for my datasets.

Discarded work performed but ultimately not used included creating a RandomGPT dataset and fine-tuned model, where the dataset was generated by randomly sampling from a large corpus of English words with the intent of creating nonsensical sentences as a baseline model. Because the corpus of words used was so large, GloVe did not have embeddings for many sampled words, and thus meaningful GloVe similarity scores could not be generated for this fine-tuned model. Additional discarded work included the implementation of BERT next-sentence prediction as a third measurement technique. This was discarded because next-sentence prediction is no longer a supported task due to the lack of popularity on Hugging Face.

I chose this project with the intention of learning how to use existing deep learning models, fine-tune them to my liking, and finally productionizing them into a user-friendly environment. These first two aspects I consider to be a major success, but ultimately I did not have time to fulfill the original intention of the project which was to build an app wherein users can choose from a list of fine-tuned GPTs and chat with them. However, this was only because the project steered in a different direction, and that was toward coherence measurement, which was something I did not know anything about when starting the project. It was an interesting experience to brainstorm and research how logical coherence can be measured numerically – a task I learned is not a simple one.

9 Future Work

Initial inspection of conversation outputs lead me to predict that models fine-tuned on a larger amount of data will produce outputs more logically coherent than models fine-tuned on a small amount of data. Unfortunately, I was not able to draw confirmation of this informal hypothesis from the results. Although I produced some evidence that my two measures of logical coherence captured information about the underlying content of the conversations, it is difficult to say from my results whether this information captured is truly reflective of the logical coherence of the conversations. Generating greater pools of conversations from each model pair might help reduce noise in measurements. Future work can also experiment with different fine-tuning datasets, varying in syntax, size, and quality. For example, it would be interesting to see how much coherency is retained when DialoGPT is fine-tuned on a dataset in which a percentage of the words were randomly taken out and replaced with another.

I mentioned that the log-likelihood metric would likely be more predictive of logical coherence when the model computing the log-likelihoods shares similarities in its training data with the fine-tuned model. One way to achieve this would simply be to fine-tune the model producing the metric on the same dataset as the dialogue-generating model. Additionally, as previously mentioned, future work can be made to determine precisely why conversations deteriorated with punctuation and to produce additional work wherein conversations of an arbitrarily large length are generated.

Finally, I note the inherent theoretical difficulty with the very notion of measuring the logical coherence of conversations generated between models with two different fine-tuning datasets. Although numerically measuring the logical coherence of a conversation between two modern English speakers is theoretically possible, it is less easy to understand how one might do so to measure a conversation between Shakespeare's Macbeth and *Star Wars*'s Jar Jar Binks. Firstly, both characters in question are purely fictional, and secondly, it is doubtful whether these two characters would have a coherent conversation if they were somehow brought to life and put into the same room. Future work can be dedicated to studying the conversations such characters might generate in this fictional reality to create better metrics for measuring conversations generated by large models fine-tuned in the personalities of such characters.

10 Citations

[1] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

[2] Open AI. GPT-4 Technical Report. arXiv:2303.08774.

[3] Jurafsky, D., & Martin, J. H. (2023). Discourse Coherence. In *Speech and Language Processing* (pp. 1-20).

[4] Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., ... Dolan, B. (2019). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv:1911.00536.

[5] *Star Wars* dataset: https://www.kaggle.com/datasets/xvivancos/star-wars-movie-scripts.

[6] Tiny Shakespeare dataset: https://huggingface.co/datasets/tiny_shakespeare.

[7] *The Office* dataset: https://www.kaggle.com/datasets/lillitarhea/the-office-script-lines.

[8] Dubon, J. (2023). How to Finetune DialoGPT. https://www.youtube.com/watch?v=eereDVZeE7Y.

[9] IAmANerd (2020. Open-Dialog Chatbots for Learning New Languages. https://nathancooper.io/i-am-a-nerd/chatbot/deep-learning/gpt2/2020/05/12/chatbot-part-1.html.

[10] Adams, D. (1979). The Hitchhiker's Guide to the Galaxy. Pan Books.

[11] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[12] Xu, P., Saghir, H., Kang, J. S., Long, T., Bose, A. J., Cao, Y., & Cheung, J. C. K. (2019). A Cross-Domain Transferable Neural Coherence Model. arXiv:1905.11912.

[13] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI. [https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf].