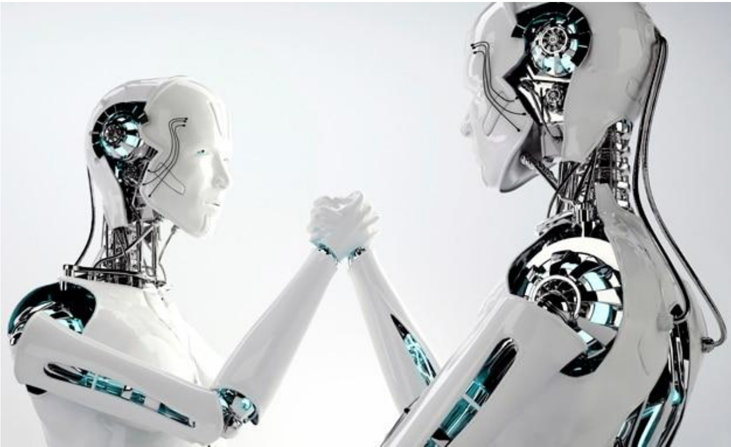
Evaluating the Logical Coherence of GPT-Generated Conversations

Bay Hodge



Summary

1. Fine-tuned DialoGPT to create
    a. StarWarsGPT
    b. ShakespeareGPT
    c. TheOfficeGPT (large, medium, small, extra small)
2. Generated 30 conversations between each pair of models
3. Evaluated the logical coherence of the resulting conversations using
    a. GloVe similarity scores
    b. Difference between marginal and conditional log-likelihood

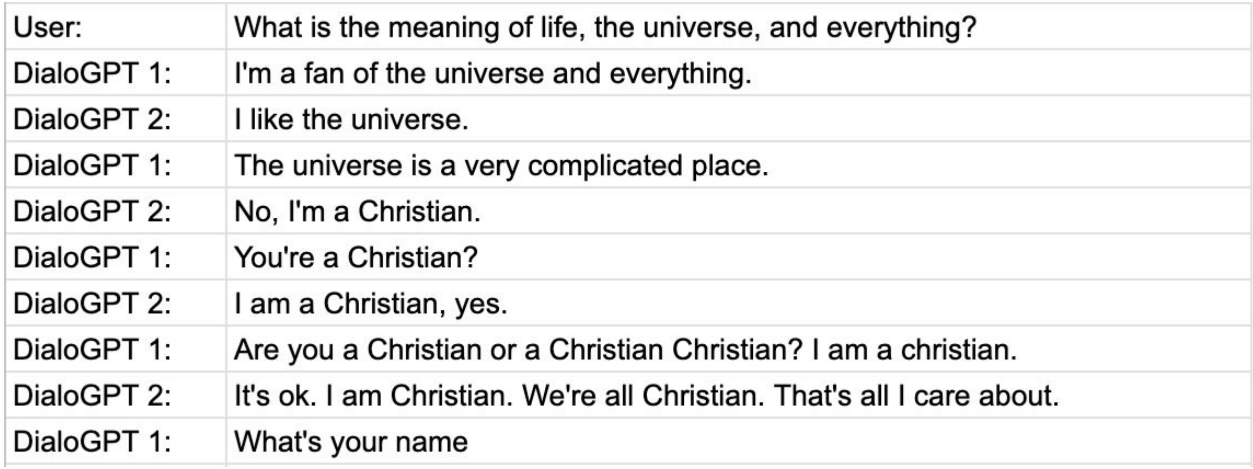
Goals

1. To understand if the chosen evaluation metrics are reliable measurements of conversational coherence
2. To demonstrate the power of DialoGPT’s ability to produce outputs that accurately reflect the semantic content of the data it is fine-tuned with

Base Model

DialoGPT (dialogue generative pre-trained transformer) from 2019

- Share’s GPT-2’s architecture, except trained on ~147M Reddit conversations
- No longer SOTA and is thus imperfect



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

Fine-Tuning Temperature

Temperature regulates the randomness of the model's outputs by modifying the probability distribution of the words in the vocabulary

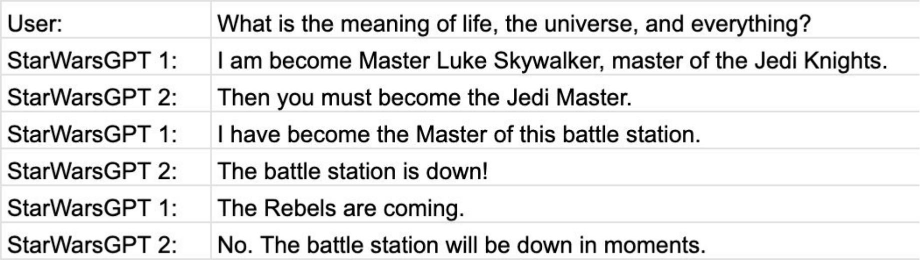
- T > 1: increases randomness by flattening the distribution
- T < 1: decreases randomness by exaggerating the relative P(i)s
- Settled on T = 0.7

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

Initial Observations/Predictions

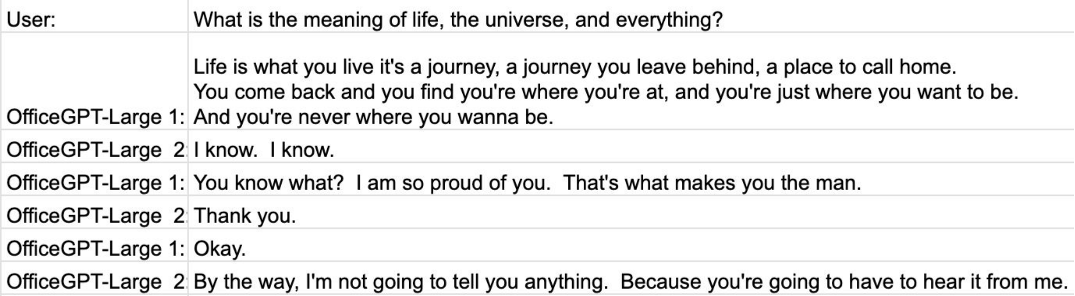
1. More fine-tuning data = higher coherency?
2. Conversation with itself = highest coherency?

2500 samples:



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.

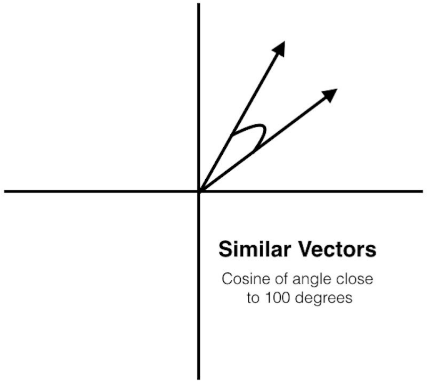
60000 samples:

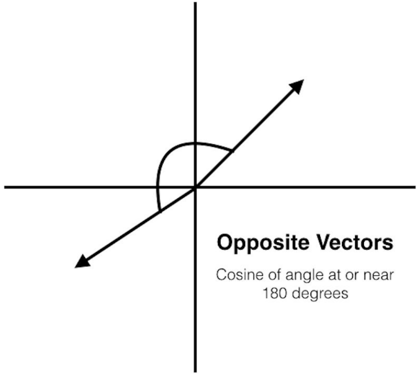


User:	What is the meaning of life, the universe, and everything?
OfficeGPT-Large 1:	Life is what you live it's a journey, a journey you leave behind, a place to call home. You come back and you find you're where you're at, and you're just where you want to be. And you're never where you wanna be.
OfficeGPT-Large 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.

Evaluation with GloVe Similarity Scores

1. Used GloVe to compare adjacent words. Then, averaged over the sentence to create sentence vectors, and finally used cosine similarity to determine the similarity of sentences in a conversation
2. Reported the average of these scores over all 30 conversations for each model pair
3. Effectively, measuring the semantic similarity of sentences in a conversation, i.e. whether it stays on topic
4. The more semantically similar the conversation, the higher the similarity score we expect

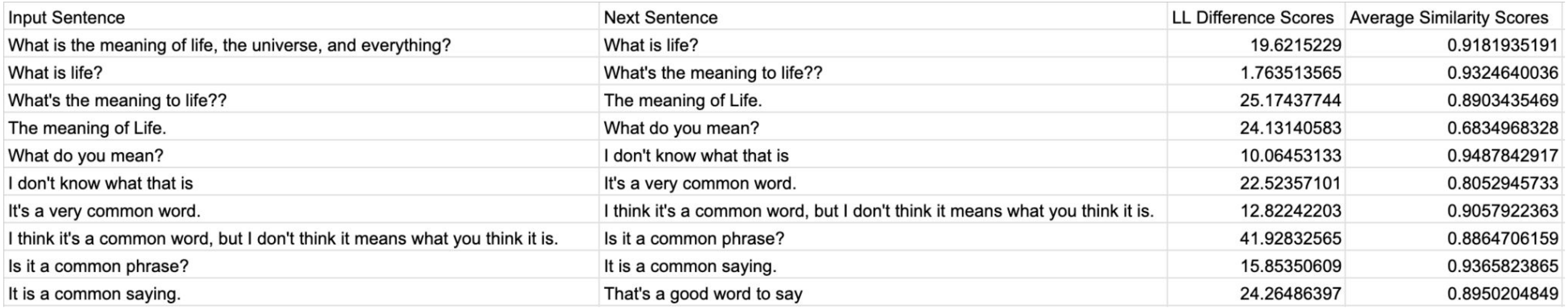




Evaluation by Differencing Log-Likelihoods

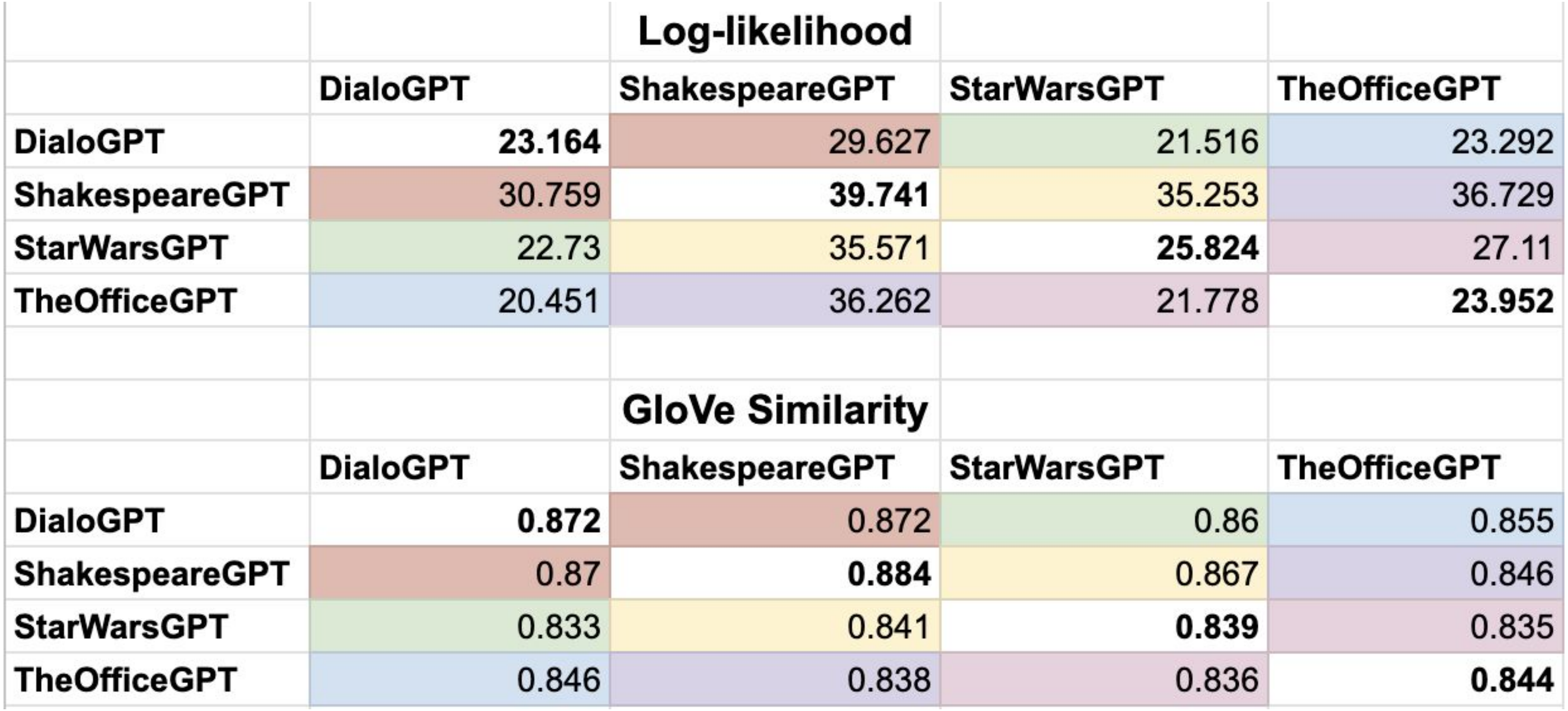
1. Used GPT-2 to compute log-likelihood scores each target sentence (1) with conditioning on the previous sentence and (2) without conditioning
2. Reported the average of these scores over all 30 conversations for each model pair
3. Effectively, measuring whether a sentence is likely follow the previous sentence
4. The more reasonable a response is to its context, the lower the difference we expect

DialoGPT with DialoGPT



Input Sentence	Next Sentence	LL Difference Scores	Average Similarity Scores
What is the meaning of life, the universe, and everything?	What is life?	19.6215229	0.9181935191
What is life?	What's the meaning to life??	1.763513565	0.9324640036
What's the meaning to life??	The meaning of Life.	25.17437744	0.8903435469
The meaning of Life.	What do you mean?	24.13140583	0.6834968328
What do you mean?	I don't know what that is	10.06453133	0.9487842917
I don't know what that is	It's a very common word.	22.52357101	0.8052945733
It's a very common word.	I think it's a common word, but I don't think it means what you think it is.	12.82242203	0.9057922363
I think it's a common word, but I don't think it means what you think it is.	Is it a common phrase?	41.92832565	0.8864706159
Is it a common phrase?	It is a common saying.	15.85350609	0.9365823865
It is a common saying.	That's a good word to say	24.26486397	0.8950204849

Results



		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**

Observation 1

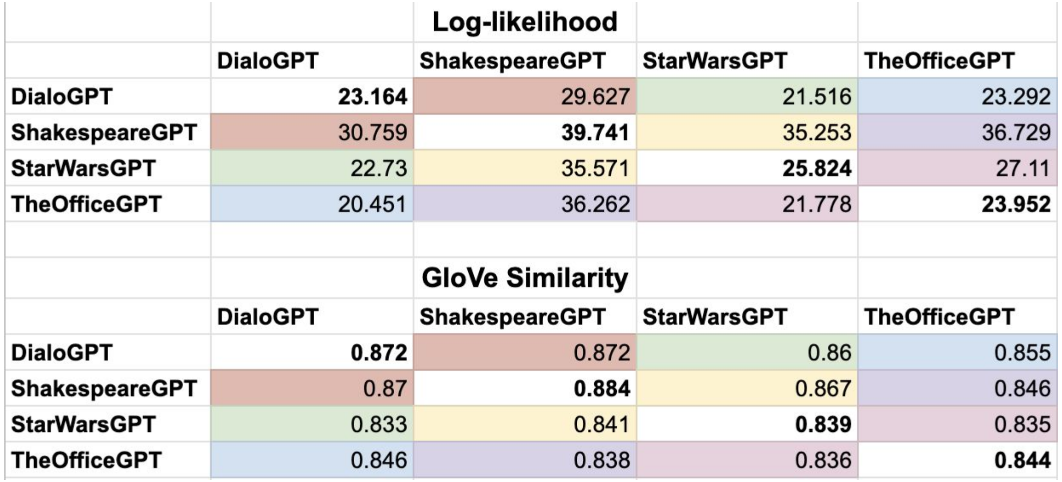
No clear relationship confirming that (1) more fine-tuning data leads to higher coherence or (2) highest coherence achieved when model converses with itself

StarWarsGPT: 2500 samples

ShakespeareGPT: 8000 samples

TheOfficeGPT: 60000 samples

DialoGPT: 147M samples



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

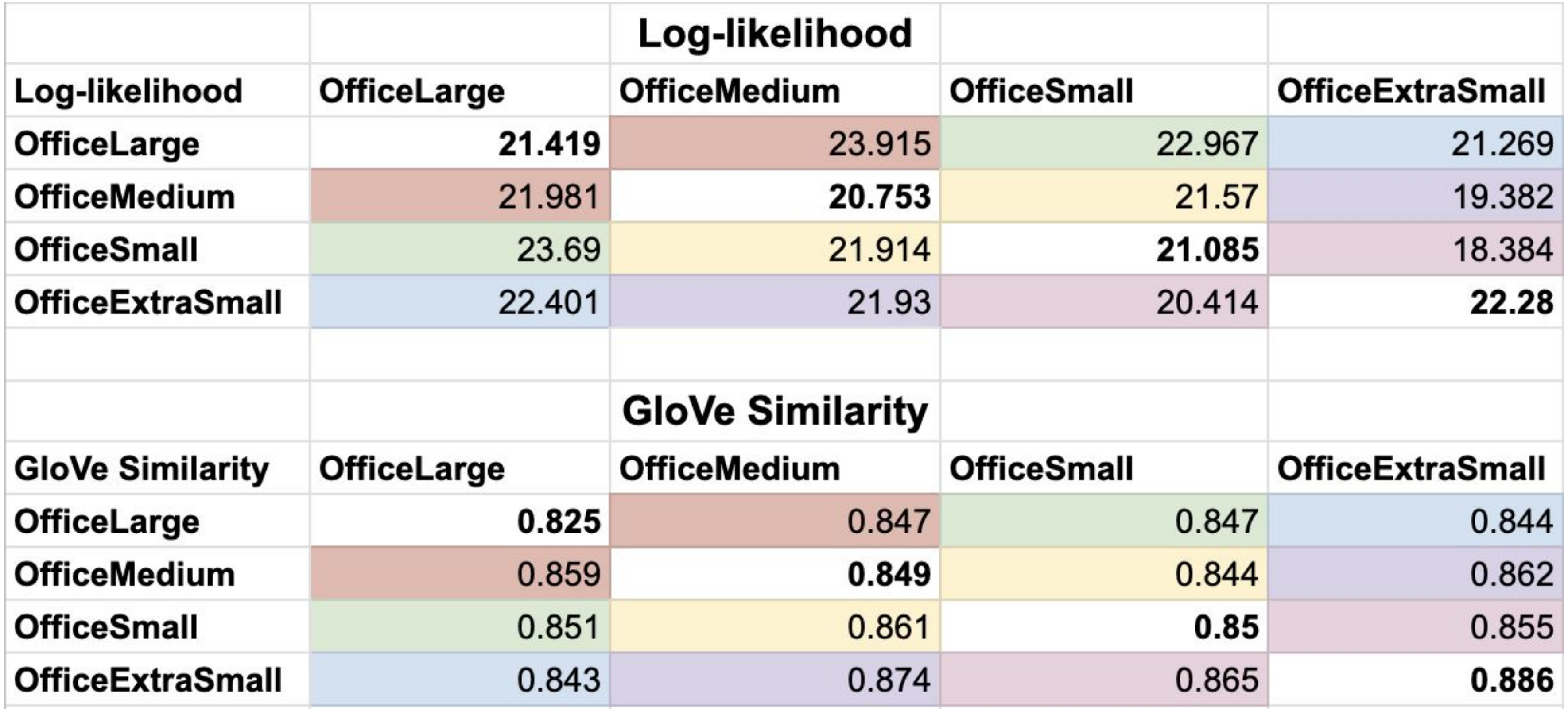
	DialoGPT	ShakespeareGPT	StarWarsGPT	TheOfficeGPT
DialoGPT	**0.872**	0.872	0.86	0.855
ShakespeareGPT	0.87	**0.884**	0.867	0.846
StarWarsGPT	0.833	0.841	**0.839**	0.835
TheOfficeGPT	0.846	0.838	0.836	**0.844**

Observation 2

ShakespeareGPT with ShakespeareGPT

- Least coherent by log-likelihood metric (39.741), most coherent by GloVe metric
- Why? GPT-2 probably not effective at measuring log-likelihood for “Shakespearean English”
- Future work: fine-tune the model used to compute the log-likelihood metric on the same dataset(s) that were used to fine-tune the generating model

Results

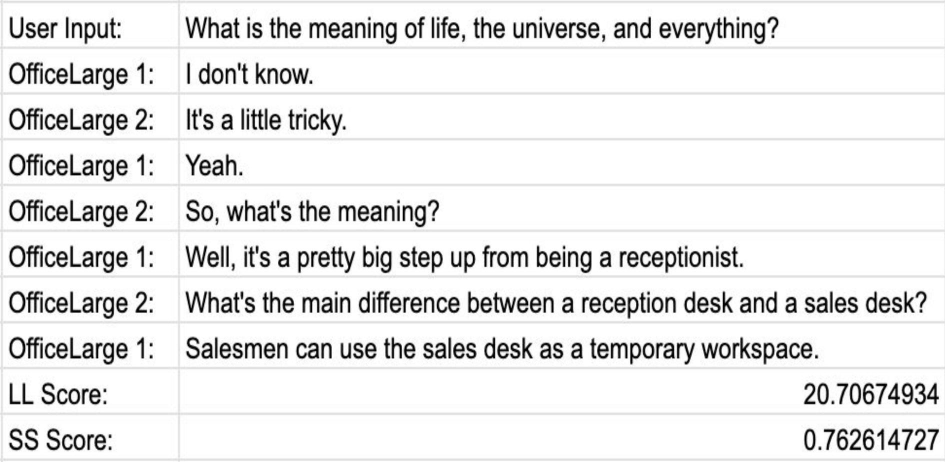


		Log-likelihood		
Log-likelihood	**OfficeLarge**	**OfficeMedium**	**OfficeSmall**	**OfficeExtraSmall**
OfficeLarge	**21.419**	23.915	22.967	21.269
OfficeMedium	21.981	**20.753**	21.57	19.382
OfficeSmall	23.69	21.914	**21.085**	18.384
OfficeExtraSmall	22.401	21.93	20.414	**22.28**

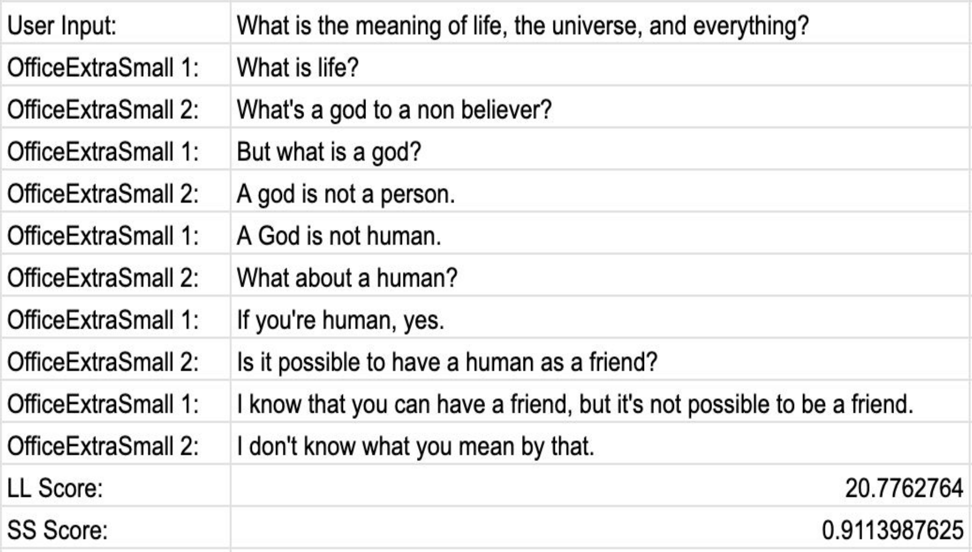
		GloVe Similarity		
GloVe Similarity	**OfficeLarge**	**OfficeMedium**	**OfficeSmall**	**OfficeExtraSmall**
OfficeLarge	**0.825**	0.847	0.847	0.844
OfficeMedium	0.859	**0.849**	0.844	0.862
OfficeSmall	0.851	0.861	**0.85**	0.855
OfficeExtraSmall	0.843	0.874	0.865	**0.886**

Observations

- Most of the variance likely attributable to noise, and not due to variance in fine-tuning data set size
- TheOfficeGPT-Extra-Small was only trained on 600 examples…how is it performing well?



User Input:	What is the meaning of life, the universe, and everything?
OfficeLarge 1:	I don't know.
OfficeLarge 2:	It's a little tricky.
OfficeLarge 1:	Yeah.
OfficeLarge 2:	So, what's the meaning?
OfficeLarge 1:	Well, it's a pretty big step up from being a receptionist.
OfficeLarge 2:	What's the main difference between a reception desk and a sales desk?
OfficeLarge 1:	Salesmen can use the sales desk as a temporary workspace.
LL Score:	20.70674934
SS Score:	0.762614727



User Input:	What is the meaning of life, the universe, and everything?
OfficeExtraSmall 1:	What is life?
OfficeExtraSmall 2:	What's a god to a non believer?
OfficeExtraSmall 1:	But what is a god?
OfficeExtraSmall 2:	A god is not a person.
OfficeExtraSmall 1:	A God is not human.
OfficeExtraSmall 2:	What about a human?
OfficeExtraSmall 1:	If you're human, yes.
OfficeExtraSmall 2:	Is it possible to have a human as a friend?
OfficeExtraSmall 1:	I know that you can have a friend, but it's not possible to be a friend.
OfficeExtraSmall 2:	I don't know what you mean by that.
LL Score:	20.7762764
SS Score:	0.9113987625

Final Thoughts

Numerically evaluating logical coherence is hard.

The theory led me to believe that coherence would be highest when a model converses with itself and when the model has more fine-tuning data.

However, we could not confirm these suspicions with our data. This leads us to believe that our metrics for logical coherence are not effective, at least, not in isolation.