GPT-2's Thoughts on Your Relationship

By Robert Malka

(With thanks to Mentors Ben Bell & DJ Sarkar, and to Max Woolf for his work on making GPT-2 accessible!)

GPT-2's Relationship Advice

What if we could create a model that reliably gives relationship advice to anyone who needs it, aggregated from millions (or more!) posts on a subreddit such as r/relationship_advice? It might enable some who can't bring themselves to reach out to others, particularly in these isolated times, to post their question into a simple app and get some crowdsourced advice that helps them see their situation in a new way, with the full awareness that an AI won't (cannot) judge them.

We scraped Reddit's r/relationship advice, specifically focusing on the post titles and comments and ran them through GPT-2, in order to test the efficacy of such advice. We also performed a light EDA on the dataset, seeing reading level, wordclouds, and sentiment polarity, along with performing a Latent Dirichlet Allocation (LDA) to see how best to cluster the different post titles and comments.

The business problem may not have a specific timespan or clear numerical benchmark attached to it, but that does not stop it from being potentially useful to a group wanting to optimize such advice. We will post some of the best answers GPT-2 has provided, grapple with how it understands the data it's received and processed, and discuss ways to further improve its output.

There is also a huge optional aside at the end discussing the philosophical dangers of taking this project seriously, or even attempting to improve the output of such a project – remarks that too few people in the field are making, and should perhaps make more often.

Finally, this version of GPT-2 will be put on a website for others to test (along with the philosophical and practical caveats this enterprise deserves). Link forthcoming.

Caveat

One small caveat encompasses this data: That the full size was too great for GPT-2 to process using Google Colab, according to Google's requirements (and it was an unrealistically long time to run GPT-2 from the command line, on my local GPU). Therefore a reduced dataset was sent through Google Colab, although that dataset remains quite large by standards I've seen, at approximately 75MM tokens.

1. Data

My data was scraped from Reddit's r/relationship_advice, where I used both PRAW (Python Reddit API Wrapper) and PSAW (Python Pushshift API Wrapper) to collect data. Using PRAW, I was only able to download about 900 entries, far too small for GPT-2. (Reddit places an automatic limit on requests through PRAW.) Using PSAW, which circumvents that limit, I managed to collect a million posts, and

three million comments (many of which were multiple, sometimes hundreds, of comments for a single thread). I simply joined all comments with the same ID into a single cell, and then linked that ID to the post it was connected to. Doing so reduced the overall size of the dataset to ~184,000 rows. This meant deleting approximately 800k rows of threads, for which I could not pull the comments.

The problem with the dataset as it stands is that it processed the comments to comments made – not necessarily helpful in giving advice to the problem proposed in the post itself. (Many Reddit comments are tangential to the main subject.) One way to regulate for this is to delete comments beneath a certain upvote threshold, and then to delete posts whose comments could not exceed that upvote threshold.

One could use PSAW (more info at www.pushshift.io) to pull such scores and therefore prune the dataset more extensively, allowing the same number of tokens to go farther.

2. Data Cleaning

Scraping Reddit is thankfully straightforward, and so not much cleaning was necessary. I changed the date the posts were created to datetime from epoch time. I linked the IDs of the posts to the comments, which required editing the post IDs to match a prefix visible in the comment IDs. I joined the two separate datasets into one dataset for clarity during the EDA.

Finally, I vectorized the texts using Word2Vec to process them for LDA (Latent Dirichlet Allocation, an approach to topic modelling); that required making the data lowercase, removing punctuation, special characters, and so on.

3. EDA

To look at the code, please see the EDA Report.

Because I don't examine any particular hypotheses in the EDA (whereas that's my default approach in a classification problem), I kept my EDA simple and straightforward – what basic patterns are exhibited in the data?

The smaller dataset, 80,000 rows, came out to 75,830,299 tokens for GPT-2 to be trained on. The full dataset is closer to ~200 million tokens. Trying to do the full dataset repeatedly crashed Google Colab. I used these smaller datasets (80,000 titles + posts, and 80,000 comment rows with unique IDs, respectively) to perform an EDA.

Dataset Fields

For the submissions dataset), the following fields were pulled:

- id: The unique post ID
- created: The timestamp of the post in datetime (originally in epoch time)
- title: The subject line of the post
- post_body: The full body of the post

For the comments dataset, the following fields were pulled:

- link_id: The unique id of the post the comment belongs to, with "t3_" indicating that it is a comment and the remainder of the id the same as the post id.
- id: The unique comment id.
- created: The timestamp of the comment in datetime (originally in epoch time)
- body: The full body of the comment.

For a future dataset, I will also be pulling:

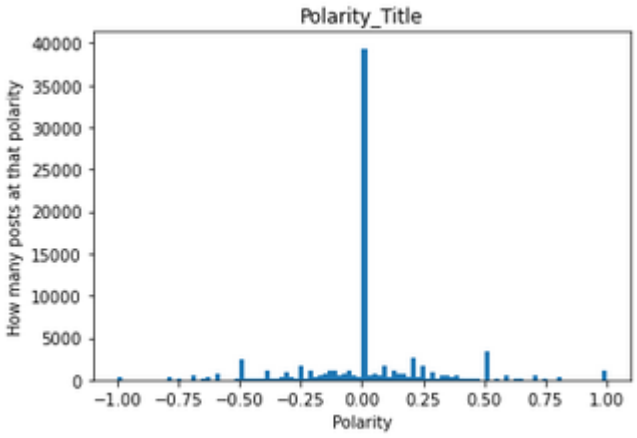
- score: The number of "upvotes" the community gave to that post (roughly, how much interest the community had in that post).

This is so I can evaluate the best comments given.

I joined these dataframes together into one, df_fin, emphasizing only:

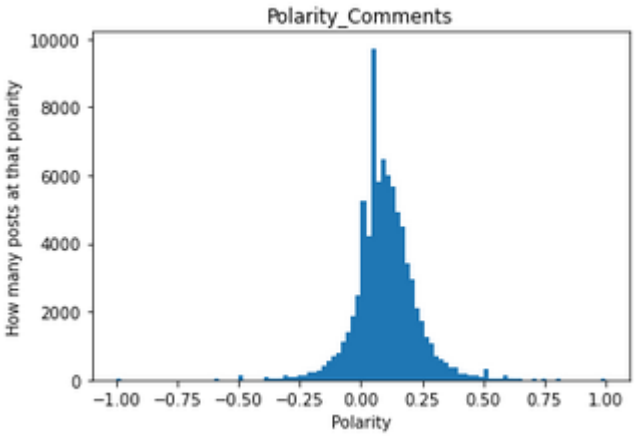
- id: The unique post ID
- created: The timestamp of the post in datetime (originally in epoch time)
- title: The subject line of the post
- post_body: The full body of the post
- comment_body: The full body of the comment.

Examining the Polarity of the titles, I got the following graph:



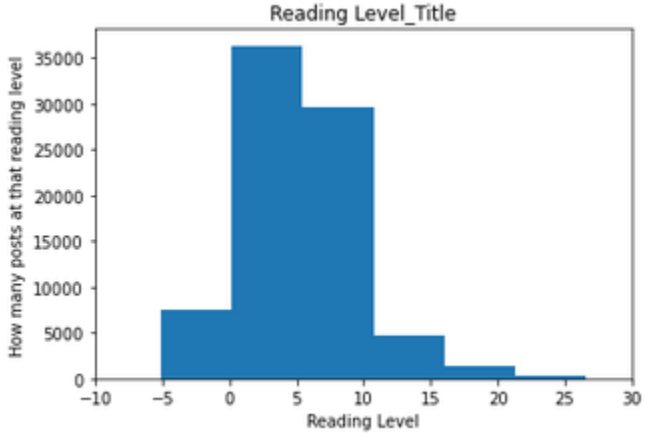
Polarity is largely neutral in titles, and then otherwise apparently randomly distributed, save for the preponderance of 0.5 & -0.5 valences. (This might be a function of TextBlob? Perhaps it 'rounds up' towards a clearer value such as (+/-) 0.5 – note the decline between what seems to be (+/-) 0.4 and (+/-) 0.5, before the jump – that could be coincidental. We leave it at that.

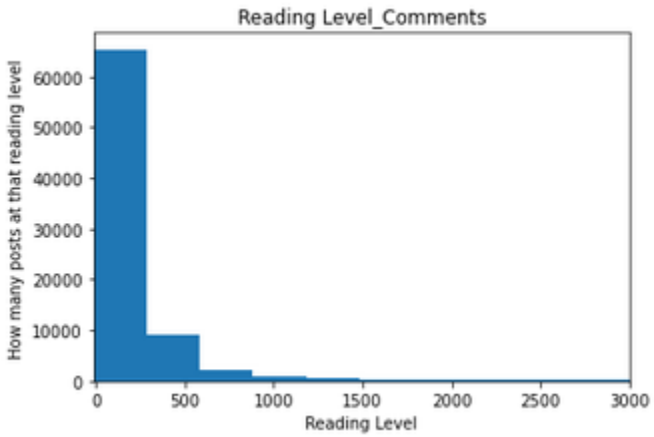
In contrast, however, see the graph showing the polarity of the comments:



It's mostly a normal distribution, leaning towards a more positive polarity! This suggests to me that comments, in the aggregate, are largely encouraging (and that the community is, on the whole, actually on the kinder/more supportive side).

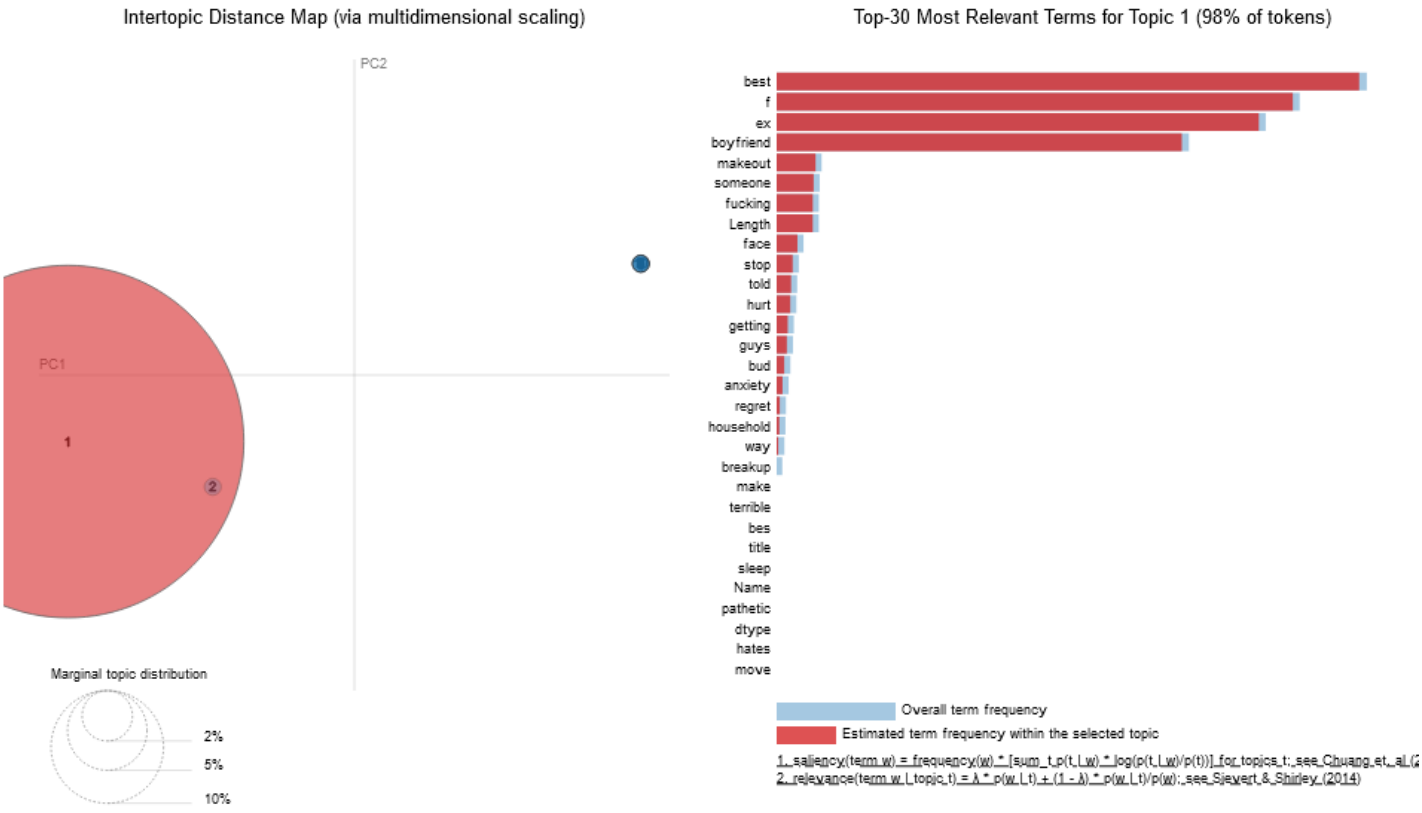
Next, we look at reading levels. I used the Flesch-Kincaid scale, for which more information can be found here. Let's look at both the Titles & Comments in quick succession:





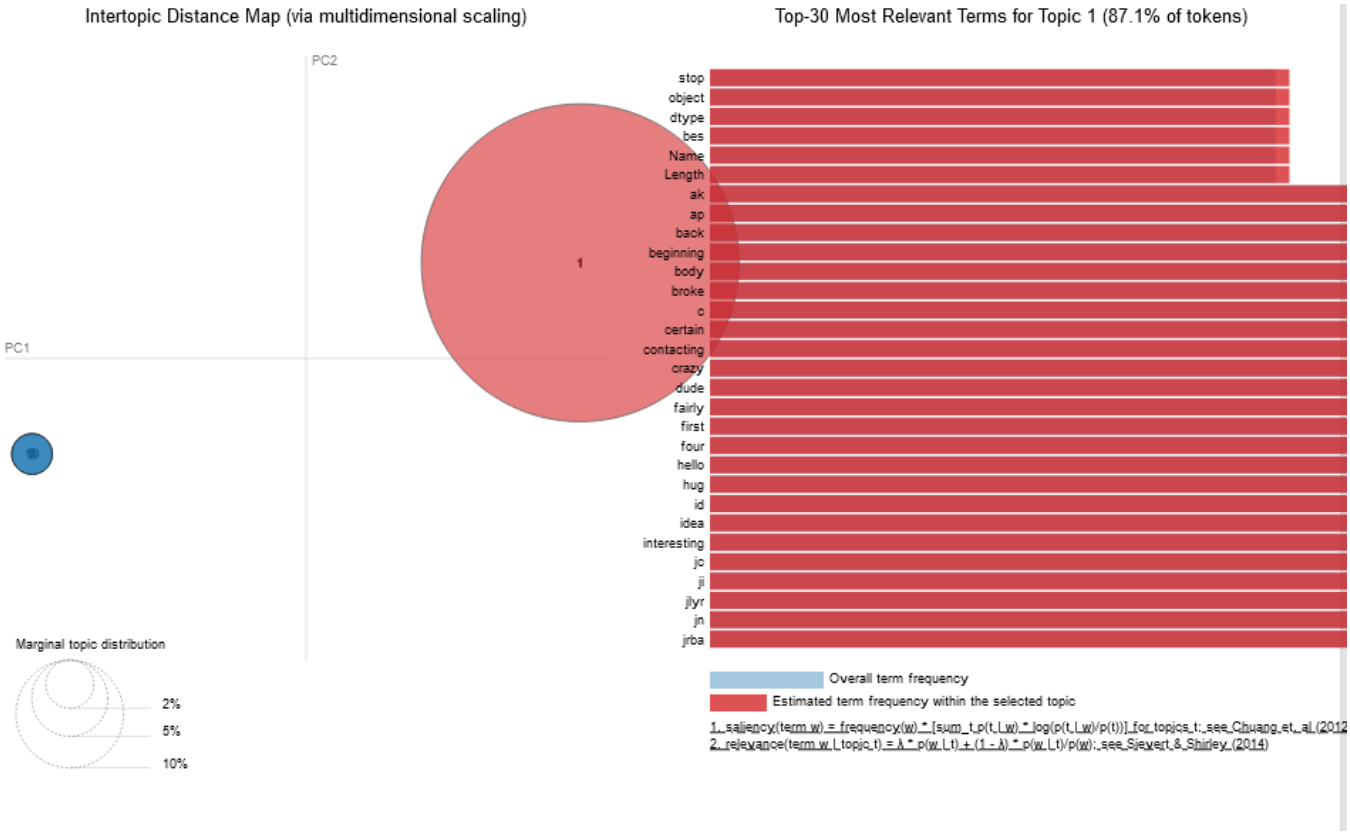
We see an obvious difference: The Titles have very low reading levels, mostly because they're so short (a byproduct of the way Flesch-Kincaid is designed). The comments, on the other hand, have a very long tail for reading levels going into the tens of thousands (!) – it extends past the graph I've shown. This is because of the very long comments and the fact that they're comprised of multiple comments, next to which it's unclear to F-K that it's comprised of many comments put together.

Lastly, let's look at LDA for the titles:



Notice that there's a single large cluster that comprises 98% of all tokens – that is to day, ~98% of all issues relate to the top words in red on the right-hand side. (The ~2% clustered were distributed among the tiny cluster 2 in the first larger cluster, and in "topic 10" on the upper right side, which certainly includes topics 3-9. They were so small that no new words were seen on the "relevant terms" side.) What we notice is what we might expect – that almost all issues related to breakups, exes, sex, regret, and being hurt. Notably, the people on relationship advice talk about boys more than girls (though girlfriend is a term mentioned when cluster 1 is not highlighted, as the term in third, but it doesn't have an "estimated term frequency within the selected topic" at all. This is noteworthy – it suggests that it doesn't have nearly as much relevance in the titles. (Perhaps it was phrased as "SO" (significant other), or gf ("girlfriend"), and its rarity of mention, plus the multitude of ways it may have been phrased, buried the terms beneath the top 30.)

Next, we look at LDA for the comments:



You see that it's largely nonsensical – it seems to me that further investigation should try to group tokens in terms of sentences rather than words, to try to approximate the similarity of thoughts and suggestions. (The other topics are equally nonsensical.)

We leave the comments LDA at that, noting only that, similarly to the titles LDA, it's mostly grouped into one cluster.

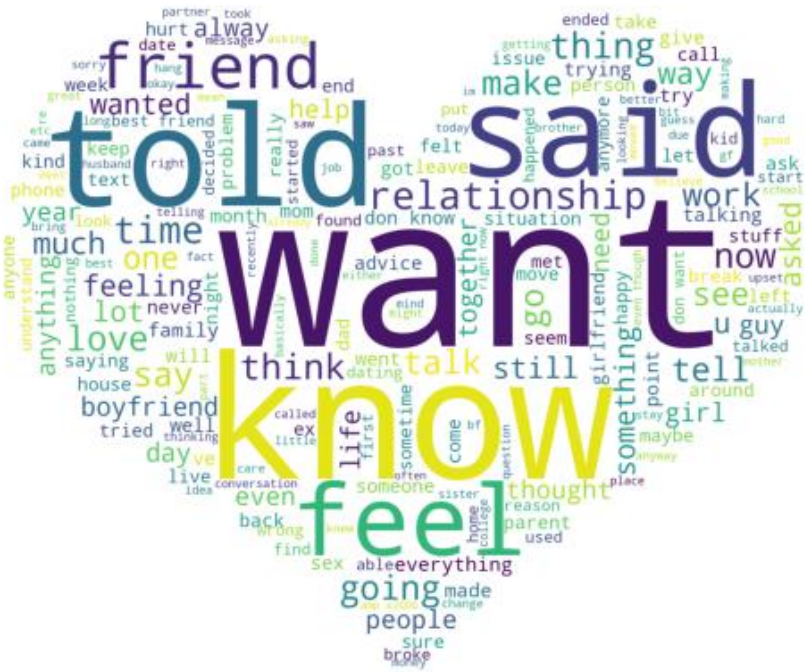
Finally, we'll look at wordclouds that show three things: The topics often posted about in r/relationship_advice, the body of those posts, and the comments about those posts.

4. Wordcloud

A wordcloud of the titles of each post show the frequency of words outside prepositions, indefinite articles, and so on. Also, they're in the shape of hearts, enjoy ur life.

The titles of the posts first:

The body of the posts next:



And the wordcloud of the comments last:

Notice especially the wordcloud about titles, where we see almost everything relates to romantic relationships, and the issues surrounding that. Also notable how many of the men and women asking for advice are young, e.g. 22f, 25m, and so on. (Is it any surprise the subjects being discussed in r/relationship_advice?)

The wordcloud in the post_body notably has verbs in the past tense -- no surprise there, a story is being told. Verbs in the comments wordcloud are present and future tense -- advice-giving, and reassurance about the future. And it's verbs that have the particular prominence in both.

And, finally: Is it a surprise that so many of the posts might be about – if we were to string all the biggest words together in the post_body – wanting to know what was told, said, or what someone feels?

5. A Description of GPT-2 & Parameters Set

GPT-2 is an open-source model developed by OpenAI. It is able to 'learn' without training, and specifically can translate text, summarize passages, and answer questions without going through extended training. (In this case, the training helps dramatically, but it is still able, as we will see, to attempt answers to questions it has never been trained on before in creative ways.)

Its impressive (relative to previous models) performance is the result of being able to focus on pieces of sentences and to guess at their importance relative to the next sentence. It has 1.5 billion parameters. I focused on fine-tuning particularly its creativity (called 'temperature'), and wasn't particularly interested in tuning the length of tokens it generates ('length'), limits to generated guesses ('top_k'), and a few other parameters.

I trained it for about 8,000 steps (though people with smaller models have trained it up to 42,000 steps, which is worth trying if not bound by time constraints).

6. Initial Results

I asked the model to post unprompted responses every 200 steps of training so that I could watch it develop into cogency. The model started quite roughly:

1) *i think i'm just saying that there should be more females out there. maybe a couple hundred? she should get that she got her girlfriend.*
2) *i think it'll make the difference between not having a relationship and being able to see the relationship as just a joke: you can never tell yourself that you want to be in a relationship but it's a given. as long as it's fun.*
3) *my current sister is 27f: and i'm 28f. we've both been in therapy together for about a year: and i still find it hard to talk about this with her. she does try to explain it away: and i do try to not dwell on the topic: as it's all been so well communicated.*

As the steps start to improve, pockets of cogency start to appear, and there's a gradual emergence from the uncanny valley of words never meant to mean anything (I am of course selecting for the best samples):

4) *she was there for you when you were being betrayed by a man who would put other people above you. and now she has an ex that would want to get close to you.*
5) *as much as you think this is a problem at 26: it's not: because there's nothing wrong with what she did. that's definitely possible. and maybe not because she was in love.*
6) *the best thing to do is to do nothing: and just let it be. that way: you will be fine: and she will be better off as a person. be your own best friend: not her crush! good luck: and be ready to make that girl your friend!*
7) *i'm sorry so much. i love you. you're right*
8) *i don't like the end of your relationship.*

I have no idea why GPT-2's developed an unorthodox use of colons, but I find it fascinating. My tentative hypothesis is that it's a borrowed structure from posts using colons to show "before" and "after" ("cause" and "effect") – from the first thought, the next will be so. In other words, the format is borrowed "in order to convey" (I use this phrasing with great liberality) that the latter is expected to happen as a result of the former.

There are also a few places where repetition of literal nonsense is prominent, e.g. repeating "&gt" several dozen times, or the same sentence, inherently cogent, is repeated 20 times.

The signal-to-noise ratio is easily 95% nonsense, and 5% potential-sense. (For the full set of "advice" developed during training, see here.)

Some final samples that seemed possibly cogent:

9) *it sounds like you need to live on your own. you are going to work yourself up for this: but you really can't take all that for granted with the constant fear of leaving. living on your own is a great way to build yourself up for the future.*
10) *if so break up: or ask her to come back if she really wants to get back together.*
11) *if she wants to be with you: then you're the only one for her.*
12) *you are wrong to think that people can be hurt easily. you say that you are hurt a lot you are an asshole and are not someone that has been hurt and broken up with.*

That last one's a little strange, but perhaps GPT-2 was imitating an uncharacteristically neurotic and unsupportive individual. And, two more (part of the same comment cluster) that caught my eye in a random generation:

13) *also: this is a very bad time for relationships. we typically get stuck in the dating phase. it's not healthy for us to be in a relationship if we're expecting it to last.*
14) *as someone who has been in one: it's not healthy for the relationships to last.*

Cogent – and terribly wrong. (This makes sense – without self-awareness, how could it even conceive of moral choices, let alone suggest them?)

Note that these comments are extraordinary in the sense that one thought *attempts* to follow the other (or rather, that it was fed data in which sentences often *did* follow one another). The inner coherence of the thoughts is nonsensical – it has no understanding of what a "sister" or "boyfriend" is, or what one would do to, with, for, or about them. It has no sense of the connection between pronouns and proper nouns previously stated in the sentence; any apparent coherence seems to occur only by chance. This, it seems to me, may be because of the creativity I opted for (0.7). Lowering the score would align the comments more closely to the data it was fed, but may render them less compelling for this review.

Here's a random example, based on prompt three below, with a temperature of 0.4:

you can't force her to do anything. she will always be your grandmother. you don't have to give her permission to do anything. it's up to her to decide if she wants to be a part of your family. if she wants to be a part of your family: she will have to choose between you and your grandchildren.

It then repeats the last two clauses several times. Repetition is much higher at lower temperatures, as if it feels 'safer' repeating phrases which have already been said – something thoroughly alien to human beings. (Indeed, humans are more likely to repeat *situations* than to repeat thoughts – and even thoughts repeated change based on the people to whom they're expressed.)

It's text generated *after* training (at a temperature of 0.7, of course) that started to get interesting. I put three of my own prompts into the model – one inspired by the general issues brought up in the subreddit; one that's an actual (but unusual) post on the subreddit, posted after I scraped it; and one that I made up that's outright silly, to see how the model responds. I list three pieces of "advice" I received for each of them:

I. help my girlfriend cheated on me with my best friend
   a. *she's already done it to you and you can't control her. it's not your fault: it's not your fault: and it's not your fault to deal with it. and even if you do it: you can never trust her to not cheat on you again.* ***(I lost the generated text this was part of.)***

b. *she's being manipulative and it's not fair to the relationship. you're right. she needs to be more transparent about what happened and how she felt about it.*
c. *i know it's not the best response: but if i were in your shoes: i'd just ask her to stop talking to him. i'm not sure how you can force it but if she's willing to talk then i'd just let her know that you think it's not working out anymore for you.*

All advice generated on this prompt [here.]

II. Is it bad that guys I'm dating/hooking up with keep saying I'm like a cat?
  a. *you are an animal that needs to be petted: cuddled: fed: petted*
  b. *i think it's very different from the dating world. i've dated a guy who was super cocky: and when i was dating him: i told him i was a cat because he just kept saying that. he said it all sounded like cat noises: but i didn't understand at first. i'm now the cat person and i know that he's not that cat. i'm not saying i'm a cat: but i know he isn't. i've also dated a guy who said i was a cat a lot more. i'm just saying that's just how it is*
  c. *you are not the cat and he is not the cat.*
  d. ***Two more just because:** it's like a bad smell: if it's not there: you don't want it in the bedroom.*
  e. *you are not a cat. /n you are a human being. /n you are a human being. /n you are human.*

All advice generated on this prompt [here]. This one had some surprisingly cogent responses (that also compel me to consider other hypotheses for the colon). It seems to me the rate of helpful answers remains roughly the same (1 in 20 or so), but it is impressive that some of those answers could be cogent, though they remain largely unhelpful and unthoughtful.

In contrast, here are three top comments from the actual post:

- Honest comment. If it's hookups or casual sex... cats only come around when they want attention
- Next time someone says it, calmly knock all the shit off of the nearest shelf, table, or mantle and stare at it with pride.
- pspspspspspspspspspspsps

Funny, insightful, witty. Notably, these sorts of context-specific comments are missed in the broader generalized advice GPT-2 is giving.

Lastly, the funny prompt & the advice received. I saw something in this post I didn't notice in the others – GPT-2 changed the prompt midway through generation. I include an example of such a change in (III)(b).

III. I want to give my grandmother up for adoption.

a. *i think it's okay for you to have a conversation with your grandmother about this. just be honest with her. tell her that you don't want her to be the one to adopt her. i would also let her know that you are sad that her daughter isn't getting the opportunity or the chance to adopt.*

b. *I want to give my grandmother up for adoption. am i being too selfish? [RESPONSE]you have to be more understanding of the situation. it's very unfair to the child. you are not a god or any other god. and if you decide to keep her: it will be a waste of time and money. you will have to wait for your mom to find a suitable family member. if you ever need to speak to her: you can text her and ask her to contact you. in this state: there is an adoption center in your area (you can call them or visit their website). you can also contact your state. i'd imagine it would take some time to get to know a new person. that person may be your grandma. she might not want to have her daughter. that person may be a relative or friend. you can tell her that you feel bad for the situation but that you will help in any way you*

c. *I want to give my grandmother up for adoption. i have no idea how to go about it. [RESPONSE]a couple things: /n 1. she is a woman. she needs to make decisions for herself. i don't think you can blame her for not knowing how to handle this. /n 2. you need to talk to your family. if they are still supportive but have differing opinions on the adoption: then i think that's okay. /n 3. you need to talk to your grandparents and explain to them how you feel. they may be reluctant and may not have any experience with adoption - i don't think they can completely help you but you have to be honest. /n i think you need to give her a heads up. she's making a decision on her own. you are the one that needs to help her. if you can't help her: then i can't help you. /n you need to be honest with her and help her decide.*

All advice generated on this prompt here (at temperature 0.7).

We see that, while (c) is a (very!) creative and surprising response, insofar as it includes talk of grandparents, respecting grandparents, and giving (supposedly) actionable advice, it's still imitating an *earnest* response.

Overall, GPT-2 performed far better than expected, showing an ability to express a human-like arc in its "thinking" and approach to comments. Its grammar and spelling are strong. Notably, it's the inner sentence structure that makes little sense. It lacks cohesion and an ability to "follow" or infer the story. Yet it is still able to locate the general topic of conversation, which is impressive; it does not deviate from relevancy as often as I might have expected.

More development and review is needed – along with discussion about whether or not such tools should exist in the first place, and in what way they could injure or maim the trajectory of current and future human beings. But more on this in an optional aside below.

7. Summary

We examine the efficacy of GPT-2 at giving relationship advice, which would in theory be helpful to people who need and want to connect with human beings, but are afraid to. While its ability to give advice is, on the whole, a failure, there were some surprising successes with the model, namely its ability to imitate the human arc of a response, along with the ability to remain surprisingly relevant (in the ballpark) with respect to designated prompts. The technical successes visible in GPT-2's work are thin disguises for its open failures: Its (understandable) inability to understand precisely what any of these words mean, how they relate to one another, and why they'd be important to the poster. Again, this makes sense: The data was funneled into GPT-2 as a series of numbers ("tokens"), all to designate the words' relevancies to one other. This is nothing like the way humans see the world or exist in it. No amount of numericizing ("vectorizing") a corpus will bring us closer to culture and what it means to live within, influence, and be influenced by one. Nor will it allow models like GPT-2 to grapple with human issues such as relationships, and we should not expect such models to do so.

That said, if one was earnest in one's attempts to maximize the value of a model such as GPT-2, future improvements are suggested below.

8. Future Improvements

Suggested next steps are to:

- Use PSAW to pull scores and delete all comments which do not achieve higher than a certain number of upvotes (e.g. 50), pruning the dataset more extensively.
- Train on more steps (perhaps closer to 30,000+ would lead to better results?).
- Train on GPT-3 to give (ostensibly) more compelling results, though GPT-3 is not publicly available at this time.
- Include the body of the post in training as information for the prompt, such that GPT-2 can make better connections between the comments and the content in the title.

9. ***Huge Optional Aside: Wtf, AI Giving Advice?***

What does the future hold for AI? I wish I could ask an oracle the answer to that question; but wherever AI will be, let's ask the equally compelling question: What does the future hold for *humanity*?

I begin by quoting a close friend – a distinguished professor of the Classics and an extraordinary thinker – who asked me precisely the questions we should be asking when we attempt to discuss Artificial Intelligence, whether it relates to AI Ethics, AI Safety, or just plain old personal boundaries around AI:

- *How did AI come to be spoken of as a person?*
- *What does it mean for an AI "to value" or organically (?) value its own (?) valuations?*
- *Can a being without self-awareness make moral choices? Isn't a moral choice the ultimate calling-forth of the question, "What am I doing here?"*

I will take this opportunity to answer the first and third question in the context of this little experiment. (The second one, much as I'd love to have an answer, seems to me to be unanswerable.)

To the first question:

This particular use-case of GPT-2, entertaining as it is, began as a satire on AI-as-human-replacement and ends as a meditation about the necessity of embracing humanity and human-ness, particularly when working with such tools.

Our relationship to AI is in that strange state where we don't take it very seriously (because we don't believe it to be 'real,' as we are), but we grow more and more dependent on it. Do we ever fail to anthropomorphize things we're dependent on? All the more so with a technology that grows better at imitating us each day.

An obvious thing must be said: GPT-2 was not meant to be human, or to perform any human functions. It was designed only to show off the learning abilities of an AI model, and to otherwise automate tasks a few people believe human beings don't actually want to do. (Whether they really do or don't, who's ever bothered to check?) So, as basically all AI Engineers will confirm, the failure of GPT-2 to give advice is expected. It was never intended to solve such a difficult problem as to give humans advice on human things (and indeed, it's not a problem for AI to solve at all). Yet we all expect some version of GPT to 'solve' this 'problem' one day, and hopefully, many of us hope, sooner rather than later; for isn't the experience of being human *hard?* Are we not all (think some) *tired* of it? When at last will we be able to delegate the ambiguities, uncertainties, and difficulties of living to a thing with all the appearances of a real being that never tires, never falters, never grows old, and that can serve as both our slaves and our masters?

I see this sentiment communicated in the shuffling feet of an increasingly depressed, medicated, suicidal, and enraged populace. It is not more AI we need, but more humanity, and more humans willing to be human; until we turn at last to the truth of that – and all the *suffering* that being a more human human will entail, we will always speak of AI, and the endless sleights-of-hand it plays, as being 'a person,' and we will believe in it as if it *is* a person.

To my friend's third question: She strikes at the chord of both what we are and what some of us hope for AI to be. Moral choices presuppose self-reflection. They are the flowerings of an examined life. A life is only examined if it has first ruminated and grappled with what it is doing here – what it believes itself to be doing here, what it believes others believe it is doing here, and what that *doing* is.

While her question was directed towards AI, it seems to me obvious that if AI (or humans merged with AI) ever finally asks such a question, it will ask it *differently* – very differently – from how we organics ask it. And whether it asks that question 'successfully' – is able to live an examined life inspired by its own answers to that question – I might suggest is ultimately none of our business. The bigger question is human beings: Do *we* live examined lives? AI seems rather to have choked our ability to ruminate, to sit on that question – we are too busy on our phones, hoping the answer lies with other human beings who have thought equally little about what they're doing here.

And AI, as it stands, might one day 'fix' this? Maybe we should admit at last that humans are doing *worse* the more technology they receive, and that the end to this experiment is nowhere in sight? And that (I love to repeat this) rather than humanizing AI, we prefer to automate human beings through myriad products like the predictive text Google and others use (whose models are not dissimilar, I imagine, to GPT-2)?

Whether AI succeeds in developing a humanity more like ours, or a sentience of its own, the central task of humanity cannot be shirked: Now more than ever, it is incumbent upon *us* – we humans – to embrace the power of being beautifully imperfect, of standing tall and upright within our mortality, of finding the seeds of the future in our uncertainty, and to give ourselves our own damned advice. For if we continue to throw budding human beings – our future vitality and potential – into building the hulking machine that is AI, we risk finding ourselves atop a mountain without a trail to return us home. I am reminded of the feeling Aeneas, of *The Aeneid,* must certainly have experienced when he tried to embrace his father in Hades:

"And as he spoke he wept.
Three times he tried to reach arms round that neck.
Three times the form, reached for in vain, escaped
Like a breeze between his hands, a dream on wings."

Pray tell, Oracle, are we flesh-and-blood to our descendants?