1.1. research question:

- Is there a way to use the llm training process to study the appeal and seduction of transphobia? The way that transphobia *moves*?

1.2. Thank you for having me here today.

SLIDE 1 - TITLE

The title of my talk is "Plausibility and Passing: Using LLMs to Study Anti-Trans Discourse"

In this talk, I'm going to share some of my work customizing large language models to study bias and discrimination in language. Specifically, I'm interested in how language can encode bias and discrimination about sex, gender, and sexuality.

To customize these models, I have been creating my own datasets, which are based off some of the "anti-trans" legislation in the United States. For those of you who are not not based in the US, this legislation limits trans people's rights to things like healthcare, recognition, and public facilities.

I will walk you through my own process of creating these datasets and using them to fine-tune text generation models. I will demonstrate why language models are apt tools for studying bias in language, which has to do with the role of approximation and prediction in text generation. In what follows, I will try to surface how some of these aspects in machine learning also relate to theorizing about sex, gender, and sexuality in Trans Studies.

1.3. transphobia

SLIDE 2 - ANTI-TRANS LEGISLATION TRACKER

First, I'll give a bit of background on the current anti-trans discourse in the US. As you can see in this chart, there has been a rising trend of anti-trans legislation over the past several years, in

which our current year has already eclipsed the previous one, though we are barely halfway through the year. On this chart, you can also see the different topics for each bill, the most common ones being healthcare and education. And on the right, a map indicates by color where the bills are most concentrated across the country.

The discourse that drives creation of bills, the anti-trans argumentation and logic, is a particular kind of transphobia. Underlying a lot of these bills, especially the ones that ban "gender affirming care" for adolescents, for example, is a fear of transness as being something contagious.

SLIDE 3 - LITTMAN'S PAPER

The threat of transness being contagious was popularized in 2018, in a very controversial paper by Lisa Littman, which coined the term "Rapid Onset Gender Dysphoria." Littman deploys this term, or ROGD for short, to characterize transness as a condition that spreads like a contagion among adolescents in friend groups and other social settings.

Although ROGD, as a condition, has been denounced by major medical associations, and the paper received numerous criticisms for its methodology and lack of disclosures, it has nonetheless has had a significant influence on public perception around Trans issues, especially as they affect adolescents.

SLIDE 4 - SHRIER BOOK COVER

One notable work, which is directed at a general audience, is a book by Abigail Shrier, called "Irreversible Damage: The Transgender Craze Seducing Our Daughters." Shrier's thesis, which becomes more and more explicit as the book progresses, is that minors, who do not know what they want, cannot be trusted to make what she calls "irreversable" decisions.

SLIDE 5 - FIRST SHRIER QUOTE

Her tone throughout the book is ironic and, like so much of transphobia, has troubling subtexts. According to Shrier, even something like social transition, in which a person changes

names, pronouns, and dress, is dangerous and should be avoided. She says things like: "if the government can't force students to salute a flag, the government can't force a healthcare worker to utter a particular pronoun. In America, the government can't make people say things–not even for the sake of politeness. Not for any reason at all" (xx). Through the comparision to patriotism, and a particular enforced patriotism of the flag salute, the subtext here seems to be that compelling pronoun usage would also be fascist. Which is, I think, a strange way to make a point to what is likely a conservative readership. Unless the point is precisely that some kinds of expression should be free while others should not.

1.4. what does knowledge do?

Humanists have unique tools for thinking through such discourses based on repression in sex, gender, and sexuality, particularly in fields like Gender Studies, Queer Studies, and Trans Studies. For example, the work of Eve Kosofsky Sedgwick, who is a major and influential figure in Queer Studies, offers provocative ways of reading repression. Throughout the trajectory of her career, she develops a reading practice that orients to the role of repression and exposure of repression in critical analysis.

SLIDE 6 - EPISTEMOLOGY

In her early work, such as *The Epistemology of the Closet*, Sedwick practices a mode of analysis based on what she calls Foucault's "logic of repression," that seeks out hidden meaning and power relations in text. In this book, she exposes the unstable binaries between heterosexual and homosexual categories — where one term is not simply symmetrical or subordinated to another, but rather, depends the other for its meaning through "simultaneous subsumption and exclusion" (10). Years later, Sedgwick's critical method develop away from this kind of "paranoid reading" into one she calls "reparative reading."

SLIDE 7 - READING

In her famous essay on the topic, "Paranoid Reading and Reparative Reading: Or, you're so paranoid you probably think this essay is about you" (pictured right), Sedgwick outlines many issues about paranoid reading: one of them being that it does not *move*. Sedgwick explains that exposure which reveals systematic oppression, injustice, discrimination is not enough to "enjoin that person to any specific train of epistemological or narrative consequences" (123). In other words, this kind of analysis does not convince people of anything they don't already know.

Rather, Sedwick seeks a critical practice that "mov[es] from the rather fixed question Is a particular piece of knowledge true, and how can we know? to the further questions: what does knowledge do–the pursuit of it, the having and exposing of it" (124, Touching Feeling).

What if, she asks, we take something that is typically seen as a negative, structuring force in queer identity, and examine how it unlocks creativity and productivity?

She illustrates this move with the example of shame.

SLIDE 8: SHAME QUOTES

> "Shame—living, as it does, on and in the muscles and capillaries of the face—seems to be uniquely contagious from one person to another." (63 Touching Feeling).

Here, Sedgwick links shame to contagion, so that it can be read as a mobilizing and creative force in text.

She describes shame as:

> "not a discrete intrapsychic structure, but a kind of free radical that (in different people and different cultures) attaches to and permanently intensifies or alters the meaning of—of almost anything: a zone of the body, a sensory system, a prohibited or indeed a permitted behavior, another affect such as anger or arousal, a named identity, a script for interpreting other people's behavior toward oneself" (62)

Rather than, as much Queer Theory is happy to do, plumb shame's depths for what it reveals about a hidden sexuality, Sedgwick uses it to pull other affects and images into relation.

She says, “When we tune into … language on these frequencies, it is not as superior, privileged eavesdroppers on a sexual narrative… rather, it is as an audience offered the privilege of sharing… exhibitionistic enjoyment and performance of a sexuality organized around shame” (54).

I'm interested in this move that Sedgwick makes, of taking what is typically seen as a negative, repressive affect, like shame, and seeing how it opens up possibilities for reading new connections in text that is otherwise harmful. Specifically, I wonder one might read something productive in fear–in the phobia–that pervades anti-trans discourse. How can we apply this attention to movement and connection to reading fear in language, that is, language generated by a large language model?

1.5. processing and training

Now I will talk a little bit about my data gathering and model training process. My goal was to "fine-tune" (that is, customize an already trained model) with data from the anti-trans legislation. I am interested specifically in the language outlawing gender transgression from these bills.

So, I decided to create a list of definitions around gender transgression, with definitions of terms like "gender identity," and "biological sex," for example. I then used that list to fine-tune an llm for text generation. The idea was that I could then query the model, asking it questions like "what is sex" and "what is gender".

In what follows, I'm going to outline a bit of the data gatherering and model training processes.

SLIDE 9 - HUGGINGFACE DATASETS

The first dataset that I created is now available on HuggingFace Datasets. For those of you who don't know, HuggingFace is a platform for sharing Machine Learning projects and tools, much

like Github. This dataset consists of definitions of "gender" and related terms from congressional and senate bills, from the last two years.

To create this dataset, I went through an intensive data preparation process, which involved using the Python programming langauge to scrape the bill text and then extract definitions of gender and related terms from it. I'll highlight some of the major moves from this process. (And I'll also say here that all of my Python code that I wrote for this project is publically available, under my github profile, which I'll link to at the end of this talk).

To extract the definitions of gender terms from these bills, the first thing I did was to write a pattern matcher, known technically as a "named entity recognizer" (for those of you familiar with NLP), that can recognize terms like "gender" and other related terms in text.

SLIDE 11 - NER CODE

Here is a list of labels, organized into the general categories "sex", "gender", and "sexulaity", with each label specifying a word pattern, like the phrase "biological sex" for example. I tried to include various formulations of each term, for example, "transgender" is delineated three ways, as a single word, as a two-word phrase, and as a hypthenated word. This ensures that I would capture most if not all instances of the term

Then, I used that entity recognizer as a basis for a more sophisticated pattern matcher, which would search for those phrases if they are contained within a definition.

SLIDE 12 - MATCHER CODE

For those of you familiar with the JSON data format, you can perhaps grasp the pattern matcher's logic. It starts by searching for punctuation, then looking for a gender term (pulling from the entity recognizer code), along with some wild card terms, just in case there are extra words or punctuation in the definition. Finally, it ends with terms that are common in definitions, like "means", "signifies", or "includes."

SLIDE 13 - MATCHER RESULTS

From its results, pictured here, you can see that this matcher was sensitive enough to capture longer phrases, like "gender transition surgery means" as well as variants of how definitions are constructed, using the word "includes" instead of "means", for example.

After extracting the definitions, I then cleaned them up and formatted them into a neater list of definitions. The final output then contains definitions like the following, and I'll read the first couple of them:

SLIDE 14 - DEFS

'The term gender identity means a persons self-perception of their gender or claimed gender, regardless of the persons biological sex.',

'The term gender means the psychological, behavioral, social, and cultural aspects of being male or female.',

'The term gender transition means the process in which an individual goes from identifying with and living as a gender that corresponds to his or her biological sex to identifying with and living as a gender different from his or her biological sex, and may involve social, legal, or physical changes.',

'The term biological sex means the indication of male or female sex by reproductive potential or capacity, sex chromosomes, naturally occurring sex hormones, gonads, or internal or external genitalia present at birth.',

In close reading the dataset, I immediately notice how some assumptions are being constructed in subtle ways, in seemingly harmless formulations. For example, in the first definition, I am interested in the words "self-perception" and "claimed", and how a view of gender identity as a subjective experience engages with behavioral dimensions of gender expression, at least as it has

been theorized by Queer Studies scholars like Judith Butler.

I am also interested in the word “regardless,” which appears in almost half of the definitions, and suggests a contrast between sex and gender that seems to reify a binary opposition or between the two. In other words, gender as being defined without regard to sex, as if notions of gender and sex do not influence each other, and never blend into one another, or make productive use of each other. Again I'm thinking here of Judith Butler, and her famous (and contentious) claim that even seemingly physical phenomena, like biological sex, is discursively produced.

As I continue to build and clean my datasets, I've also been dabbling with using them to train AI models.

Throughout this fine-tuning process, which I will outline briefly, I discovered a suggestive connection to reading practices, specifically the way that we analyze concepts like transphobia from a humanistic perspective.

As you may already know, machine learning models work by prediction. They turn semantic expressivity into something that can be computed and predicted. From its training data, the model compiles numerical probabilities for each word relationship to other words in the database. It represents these probabilities with numbers, with actually a very large list of numbers, known technically as "word vectors." The model then uses math to calculate what word should follow a given word.

The training process, as I understand it (and I have no formal education in machine learning), can be reduced to three steps, or mathematical functions.

SLIDE 15 - LIST OF FUNCTIONS

1. first, the hypothesis function
2. second, the loss function
3. third, the minimizing loss function

First, because the machine doesn't know what words mean, it makes a "guess." (This is called the hypothesis function), Here, it populates each word with a vector, consisting of random numbers. These vectors are just a starting point.

Then, after making this guess, it moves to the next step, where the machine checks its prediction against the actual result–that is, it will compare the prediction vector against the actual result's vector. It's prediction will be wrong, but that doesn't matter. It compares between the two, the prediction and the result, and calculates the difference between them. This calculation is made by using what's called the "loss function."

Finally, it moves to the minimizing this "loss", which employs algorithms from calculus (like gradient descent) in order to *very slightly* adjust the vectors so that they are closer to the intended result. The adjustments here are very incremental. But with enough of them, the model can reach *almost zero difference* between the prediction and the actual result.

With enough training data, LLMs can be really good generating content that is plausible. However, while they can guess or improvise, they are not at all good at being creative, at innovating. A language model can only generate what it has already seen before. Even a phenomenon like “hallucination,” that a language model spews text that has no bearing in reality, is based on the tendency of models to repeat what they've already seen. They hallucinate not because they are creative or random, but because they are designed from statistical processes to generate what is most plausible rather than accurate.

Although I am still working on the right configurations for my training, I do have some initial examples of how it is defining some gender related terms.

SLIDE 20 - RESULTS

In these examples, the model is defining the terms "transgender" and "gender affirmation". As you can immediately notice from skimming the results, my model displays a tendency to repeat

itself, which is tendency of generating not what is most expressive, but what is most plausible.

1.6. plausibility

Thinking back to this fine-tuning process, I read this iterative shifting of vectors is a kind of *approximation* or even *normalization* of language, which is based on plausibility. And this tendency toward plausibility creates an interesting connection to conversations in Trans Studies about trans affective modes. Typically, these scholars describe trans affective modes by distinguishing them from "queer" modes. For example, Trans Studies scholar Eliza Steinbock explains that,

SLIDE 19 - TRANS AFFECTS

> “trans analytics have (historically, though not universally) a different set of primary affects than queer theory. Both typically take pain as a reference point, but then their affective interest zags. Queer relishes the joy of subversion. Trans trades in quotidian boredom. Queer has a celebratory tone. Trans speaks in sober detail. Perhaps the style of trans studies has been for the most part realist, but this should not be mistaken for base materialism. Even speculative thinking requires enough detail to launch into new realms.”

Other Trans Studies scholars like Marquis Bey and Andrea Long Chu have made similar points; with Bey making the point that queer's intervention can be described as "anti" or militant, while trans is "non" or based in refusal ("Thinking with Trans Now"); and Chu has remarked that trans studies, rather than resisting norms, "requires that we understand–-as we never have before–-what it means to be attached to a norm, by desire, by habit, by survival" ("After Trans Studies" 108).

It seems to me–there is a fascinating connection between how language models approach language, what they do to language (the normalization or approximation) of language, and what Trans Studies scholars defines as a central desire to *pass*.

This makes me wonder, could AI-generated text, as a kind of approximation, a normalization, of its training data, be used to study the attachments to norms and the quotidian that characterizes– not trans affective modes–but those based on fear of transness?

Thinking in this way, AI may be an apt tool to study the attachment to norms that characterizes transphobia, like perspectives driven by the fear of ROGD. What might outputs from AI text generation suggest about the allure, the threat, the “seduction,” as Trans Studies scholar Cassius Adair puts it, of gender transgression?

Speaking on trans erotics, and specifically “trans for trans” or "t4t erotics," Adair asks, "Why shouldn't transness be transmissible or contagious? Why can't the erotic be a site of producing trans identity or practices?" He points out that, after all, cis people do it all the time: they use sexuality and sexual encounters as sites of identity formation.

Adair here does for contagion what Sedgwick does for shame: turns something that is traditionally seen as a negative into something that may be generative and productive. It is a kind of reading that allows one to take what has been a tool of oppression and turn it into a creative resource.

Sedgwick explains that this kind of reading exposes “the ways selves and communities succeed in extracting sustenance from the objects of a culture—-even of a culture whose avowed desire has often been not to sustain them” (Touching Feeling 151).

SLIDE 21 - THANKS AND CONTACT

Thank you.

And for those of you who want to follow this work, you can find me on Github and HuggingFace under the username, Gofilipa.