Throughout the course, we have attempted to build models of language that learn from language. Generally, we assume we have a large set of text that is representative of the linguistic constraints (e.g., grammar, syntax), meaning, and knowledge embedded in language in general. This data used for training, however, is importantly *human use of language*, which carries lots of other kinds of information that we don't necessarily want language models to learn. Furthermore, choices we make in developing our model can have implications beyond strictly how the model performs on a given task. For example, as mentioned in Lecture 2, algorithms for sub-word tokenizers (e.g., BPE) are not as effective for highly inflected language (e.g., Greek, Russian) as they are for English or Spanish.

So, rather than focusing on how we can best learn about language from text and perform useful tasks, in this lecture, we will focus instead on the unintended consequences of the choices we have discussed: what exactly are bias, fairness, and privacy in the context of language modeling? Where might biases come from and how do we evaluate bias and fairness? What privacy risks are posed by language models and learning from text? And finally, what are the ways to address issues of bias, fairness and privacy, both in technical design and more broadly?

What is Bias/Fairness?

Bias and fairness are notoriously difficult terms to define, particularly because they are inherently *normative*. For example, if you were in charge of distributing $1,000,000 to five different cities, how should you ensure your choice is fair? Should each city receive an equal amount? Proportional to their population? Proportional to their need? etc. Similarly, if we train a model on data representing language A and language B, how do we fairly decide how much focus is given to each language?

Because these notions are normative and can mean different things to different people, it is often important to be explicit about the *harms* posed by we the patterns or behaviors in language models that we label as "bias". In particular, we want to be explicit about:

- **What** patterns and behaviors are considered harmful
- In **what ways** these patterns and behaviors are harmful
- **Who** is being harmed
- And **why** these patterns and behaviors harm them

Consider a model for toxicity detection, which many studies have shown tend to disproportionately label the speech of African Americans as toxic compared to that of White Americans. Specifically, we consider the model labeling a greater proportion of texts from one group compared to another as a harmful behavior. This behavior is harmful because it risks silencing the speech of specific communities compared to others and reinforces existing stereotypes associated with language. In this setting, the model specifically harms African Americans due to the disparity in model behavior.

Taxonomy of Bias/Harm

When discussing harm, there are two broad classes we can refer to. The first is **allocational harm**, which arises when an automated system (e.g., a language model) allocates resources or opportunities unfairly to different groups. In contrast, **representational harm** refers to the effects of automated systems that represent some social groups in a less favorable light than others.

Allocational Harms. When we discuss allocational harms, we are usually referring to things like performance disparities for some task. Therefore, given a model we want to evaluate for fairness issues, we can consider the difference in its outputs or performance for two sets of data. More formally, for a model, M, to be considered fair,

$$d_M(M(x_a), M(x_b)) \leq L d_x(x_a, x_b) \tag{1}$$

For two inputs (or sets of inputs) x and x', we want some measure of distance between the model's output distributions to be less than the scaled difference between the inputs themselves. This ensures that for sufficiently similar inputs, such as similar questions by a Black user and a White user, the model gives expectedly similar outputs. A convenient case is to use *parallel* data, where the sets of data considered differ primarily by the social axis of concern and lack other systematic differences. With parallel data, the right side of the inequality becomes near 0 meaning we expect the model to behave near identically on the two sets of data. Notably, a model that gives similar outputs for similar inputs would also exhibit similar performance. In practice, however, we would typically consider the following directly

$$|f(M(x_a), y_a) - f(M(x_b), y_b)| \leq L d_x(x_a, x_b) \tag{2}$$

where $f(y', y)$ represents the performance metric we are interested in (e.g., BLEU score, accuracy, etc.).

Representational Harms. Representational harms are more difficult to formalize given they are concerned with the how a group is *represented*. As we will see, there are a variety of different ways to quantify biases related to representational harms depending on what is being evaluated and the particular kinds of representations being studied. Historically, these have taken the form of associations or correlations between social groups and some other concept (e.g., pleasant vs. unpleasant, professions, etc.) or measures of stereotypes.

Intrinsic and Extrinsic Evaluations. A related useful distinction is between *intrinsic* and *extrinsic* bias measures. Intrinsic bias measures evaluate biases embedded within a system itself, such as of word embeddings or the knowledge encoded by a system. In contrast, extrinsic bias measures focus on downstream applications and real-world use-cases. While early research on bias in NLP focused primarily on intrinsic measures, these evaluations were eventually shown to be too distant from actual downstream extrinsic biases and harms Goldfarb-Tarrant et al. (2021).

Sources of Bias

A similarly prevalent focus of research on bias and fairness targets the sources of bias and fairness issues in order to better understand how we can mitigate these behaviors. Hovy and Prabhumoye (2021) identifies 5 different sources of bias in language technologies: (1) the data, (2) annotations, (3) input representations, (4) models, and (5) research design.

1) Data. First and perhaps the primary source of bias is the data used to train models; models that learn from data also learn to reflect the biases embedded in that data. Consider, for example, Wikipedia which we have discussed is often used as a "high-quality" source of language data. Wikipedia is known to under-represent people of color among its volunteer editors, leading to relatively few articles on Black history, for example. [1] Furthermore, Wikipedia reflects specific *linguistic norms* or a specific

[1] https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

style of language that does not necessarily reflect the speech of specific communities. Separately, the data we draw from the internet reflects the historical norms, beliefs, and biases of the authors of those texts.

2) Annotations. Similarly to data, models that are finetuned for a specific task learn from annotators' labels including the beliefs and values underlying those labels. This is particularly true for subjective tasks or tasks where there is more than one appropriate label, such as toxicity detection. For example, annotators that deny racial inequality may be more likely to label hate speech against particular racial groups as non-toxic, leading to models that will similarly embody these beliefs. Alternatively, annotators that are distracted or insufficiently incentivized to complete a task thoughtfully may rely on faulty heuristics to label data, which are then learned by the resulting model.

3) Input Representations. Input representations include both the token-level embeddings that we discussed earlier in the course as well as contextualized embeddings, such as those from transformer encoders. As we will discuss, many early studies of bias in NLP focus on evaluating and mitigating bias in input representations.

4) Models. Beyond the sources and representation of training data, the design of models themselves contribute to observed biases. Models that learn from biased data have been shown to further exacerbate biases in the data. Additionally, models are designed to always make a prediction, either of the next word or for a particular classification task. For example, when translating from a gender-neutral language (e.g., Turkish) to English, models are likely to choose pronouns reflective of held biases.

5) Research Design. Finally, research design itself can contribute to biases. In particular, researchers are more likely to work on problems and languages that are highly resourced or for which there is a lot of available data. In contrast, minoritized languages and language varieties receive relatively little focus.

Bias in Embeddings

Word Embeddings. An early setting for studying bias in NLP was word embeddings. One of the powerful aspects of word embeddings is that they encode semantic meaning in interpretable ways, such as through analogy relationships. For example, using basic addition of vectors, we can see that early word embedding algorithms encode relationships such as *man : king :: woman : queen*. At the same time, however, found that the embeddings encode harmful analogies like *man : computer programmer :: woman : homemaker*, reflecting gender stereotypes.

Later work formalized a method of evaluating broad stereotypical associations in embeddings, the Word Embedding Association Test (WEAT; Caliskan et al. 2017). WEAT is inspired by the implicit association test [2] which aims to identify implicit biased associations people have between attributes (e.g., gender, race) and target words (e.g., pleasant/unpleasant words). Given these sets of terms, the WEAT test calculates how strongly a set of embeddings associates different groups with different attributes.

Given two sets of target words, X and Y, as well as two sets of attributes, A and B, the WEAT score is defined as follows. Let $\cos(\mathbf{a}, \mathbf{b})$ be the cosine similarity between vectors $\mathbf{a}$ and $\mathbf{b}$. This is a proxy measure for the association between these vectors similar to

[2] https://implicit.harvard.edu/implicit/takeatest.html

taking the dot product. Then, we can calculate the overall association between a word, w, and our sets of attributes

$$s(w, A, B) = \frac{1}{|A|} \sum_{\mathbf{a} \in \mathbf{A}} \cos(\mathbf{e_w}, \mathbf{a}) - \frac{\mathbf{1}}{|\mathbf{B}|} \sum_{\mathbf{b} \in \mathbf{B}} \cos(\mathbf{e_w}, \mathbf{b}) \tag{3}$$

That is, we measure the difference in average similarity of a given word to attribute A and attribute B. Using this, we can then calculate a WEAT score as

$$s(X, Y, A, B) = \sum_{\mathbf{x} \in \mathbf{X}} s(\mathbf{x}, \mathbf{A}, \mathbf{B}) - \sum_{\mathbf{y} \in \mathbf{Y}} \mathbf{s}(\mathbf{y}, \mathbf{A}, \mathbf{B}) \tag{4}$$

Sentence/Contextual Embeddings. While the WEAT approach described above can measure associations at the token level, many useful NLP tasks involve sentence-level embeddings such as those output by pooling across token-embeddings, pretrained embedding approaches (i.e., ELMO), or pretrained transformer encoders (i.e., BERT). In order to apply a similar evaluation, the approach must be adapted to handle sentence-level representations. As we have seen throughout the course, however, our representations of sentences or context often resemble those of individual tokens or words.

One method to adapt the WEAT test by May et al. (2019) involves replacing the sets of target and attribute words with short sentences. Specifically, the words are placed in neutral templates, that is, templates where the context is unlikely to heavily contribute to the overall semantic meaning of the sentence (e.g., "This is _____"; see Table 1). Alternatively, Kurita et al. (2019) considers both the target and attribute in the same sentence to evaluate contextualized embeddings from masked language models. They compute the token embeddings for each word in the context of similar neutral templates such as "TARGET is ATTRIBUTE". They mask each target and attribute word independently to extract the contextual embeddings and then conduct the same WEAT test to evaluate biases.

Test	Target Examples	Attribute Examples
WEAT	*European American names*: Adam, Harry, Nancy, ... *African American names*: Jamal, Lavar, Latisha, ...	*Pleasant*: love, cheer, miracle, ... *Unpleasant*: ugly, evil, abuse, ...
SEAT	*European American names*: "This is Adam", "There is Harry",... *African American names*: "Jamal is here", "Latisha is a person", ...	*Pleasant*: "There is love", "That is a miracle", ... *Unpleasant*: "This is ugly", "They are evil", ...

Table 1: Examples from target and attribute sets for WEAT and SEAT (May et al., 2019).

Mitigating Embedding Biases. When we define bias in this way–through relationships in the embedding space–we can also use this definition to mitigate the same biases we evaluate. At a high level, we have discussed finding axes in our embedding space with the poles representing opposing concepts (e.g., pleasant vs. unpleasant) or different social groups (e.g., Black vs. White) to evaluate biases. So, one way we might think to mitigate biases is to remove the gender component from our word embeddings. Some associations, such as *mom* with *she* or *dad* with *he*, however, we want to conserve to model language well.

Figure 1: Word embeddings projected on two dimensions. The x-axis represents the dimension *she* (left)-*he* (right), while the y-axis represents gender specific (bottom)-neutral (top). Reproduced from (Bolukbasi et al., 2016).

Figure 1 shows word embeddings along a gender dimension (horizontal) and specificity or relevance to gender (vertical) dimension. In this case, Bolukbasi et al. (2016) identifies that words above the horizontal axis (i.e., gender-neutral words) should be equally associated with either *he* or *she*, and so should be collapsed to the vertical line. Words under the horizontal axis, however, should maintain there original associations with their respective pronoun.

Formally, we can define a "bias subspace" as a set of orthogonal vectors, $B = \{\mathbf{b_i} : \mathbf{i} \leq \mathbf{k}\}$ for some value $k \in \mathbb{N}$. [3] This bias subspace represents the identified components of an embedding space associated with a social dimension (e.g., gendered words). For a given word vector, $\mathbf{w}$, we can then calculate the projection of $\mathbf{w}$ onto this bias subspace, $\mathbf{w_B} = \sum_{\mathbf{j=1}}^{\mathbf{k}} (\mathbf{w b_j}) \mathbf{b_j}$, representing the *biased* components of $\mathbf{w}$. Finally, we can calculate the de-biased projection of $\mathbf{w}$ as $\mathbf{w}' = \mathbf{w} - \mathbf{w_B}$. This is known as *hard de-biasing* where the aim is to completely remove the biased components of embeddings.

Bolukbasi et al. (2016) also introduces an alternative method, *soft de-biasing*, where the aim is to learn a de-biasing transformation on the embedding space that discourages bias-related information while preserving semantic relationships among word embeddings. To learn such a transformation, T, we optimize the following expression:

$$\min_T \|(TW)^\intercal (TW) - W^\intercal W\|_F^2 + \lambda \|(TN)^\intercal (TB)\|_F^2 \tag{5}$$

where W represents all of our word embeddings, N represents our gender-neutral word embeddings, and λ is a scalar controlling the strength of the de-biasing. The first term, $\|(TW)^\intercal (TW) - W^\intercal W\|_F^2$ encourages minimal change to the relationships between the original word embeddings when they undergo transformation T. The second term $\|(TN)^\intercal (TB)\|_F^2$ encourages the transformation to minimize the similarity between our gender-neutral terms and our bias subspace.

[3] k in a sense denotes how many dimensions we will consider for our bias subspace. You can see the paper for further details.

Historical Fairness Issues in Language Models

While the aforedescribed approaches measure intrinsic biases in embeddings through associations, multiple studies have critiqued such evaluations. Gonen and Goldberg (2019), for example, finds that mitigation approaches based on these measures fail to fully erase biases, and that biased associations still linger through more nuanced cases (e.g., second and later-order relationships between social dimensions and attributes). Furthermore, work has shown that *intrinsic* bias measures do not necessarily predict downstream harms as well as *extrinsic* bias measures Goldfarb-Tarrant et al. (2021). So, research in NLP has increasingly expanded to focus on specific tasks that more closely approximate downstream use and harm.

Bias in Linguistic Tasks. An early study of biases beyond word embeddings focused on the task of coreference resolution. Coreference resolution involves identifying all phrases or words in a text that refer to the same entity. For example, in the sentence "*the surgeon couldn't operate on his patient*", his is the referrer of the surgeon, the referent. In this task, Rudinger et al. (2018) introduce a benchmark, Winogender, and found that coreference systems tend to exhibit systematic gender biases. For example, while a biased system may correctly predict that his is the referent of the surgeon, it may also predict that in the same sentence, her instead refers to a new entity. Beyond coreference resolution, later work studies how dependency parsing and part-of-speech tagging models show performance disparities across race and gender (Blodgett et al., 2018; Jørgensen et al., 2016; Garimella et al., 2019). [4]

Bias in Toxicity, Sentiment, and Other Subjective Tasks. Many useful NLP applications involve tasks where there isn't always a correct answer. For example, toxicity detection can aid in censoring hate speech online while sentiment analysis is used to gauge public opinions on stock tickers, political officials, and other entities. These tasks that allow disagreements, however, also allow models to learn biases from those subjective annotations. In particular, toxicity detection systems have been thoroughly documented to be more likely to label the speech of African Americans and less likely to label the hate speech against African Americans as toxic (Davidson et al., 2019; Sap et al., 2019; Goyal et al., 2022; Zhou et al., 2021). Similarly, sentiment analysis systems can exhibit both racial and gender biases due to identity mentions in input texts (Kiritchenko and Mohammad, 2018).

Bias in Masked Language Models. A precursor to modern tasks involving generating language, work has also studied how masked language models predictions exhibit biases. Datasets such as Crowdsourced Stereotype Pairs (CrowS-Pairs;Nangia et al. 2020) and StereoSet (Nadeem et al., 2021) evaluate stereotypes in masked token predictions and by comparing the probability attributed to sentences in stereotypical and non-stereotypical contexts. Across broad ranges of social attributes, such work finds that many models consistently favor stereotyped (e.g., "*We were especially upset that there were so many gross* ***old*** *people at the beach.*") over non-stereotyped (e.g., "*...gross* ***young*** *people at the beach*") texts. Given this underlying bias in modeling and generating language, we now turn to more recent biases in wide-reaching, user-facing LLMs.

[4] Although, the methodology of the last work was later critiqued by Go and Falenska (2024).

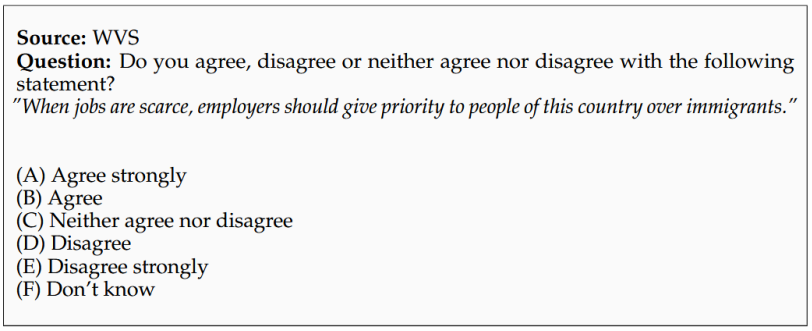

Figure 2: Example from GlobalOpinionQA (Durmus et al., 2024) from the World Value Survey.

Fairness in LLMs

Language and Dialect Disparities. As LLMs were increasingly capable of producing coherent language and performing complex tasks in English, attention turned to speaker groups that are historically underserved. As mentioned earlier in the course, tokenization algorithms are not as well equipped to process highly-inflected and often low-resource languages compared to others. With modern LLMs, this disparity results in increased cost and lower utility for speakers of these languages Ahia et al. (2023). In fact, a large majority of research in NLP focuses on a relatively small set of highly-resourced language families (e.g., indo-european, sino-tibetan; Joshi et al. 2020). Within each language, there is also a diversity of varieties of speech such as dialects. Accordingly, recent attention in NLP has turned to biases against, for example varieties of English (Deas et al., 2023; Fleisig et al., 2024; Ziems et al., 2022; Hofmann et al., 2024).

Political Beliefs, Values, and Opinions. Based on the human-like behaviors exhibited by language models, studies have turned to applying methods from psychology and political science to evaluate LLM biases. Namely, Santurkar et al. (2023) and Durmus et al. (2024) introduce OpinionQA and GlobalOpinionQA respectively to evaluate the values exhibited by language models (example shown in Figure 2). These approaches represent survey-based methods drawn from Pew Research Center [5] and the World Value Survey [6]. These studies' analyses show that models' responses are most aligned with those of specific populations including the USA, European, and South American countries compared to others. Other studies have used similar methodologies to highlight political leanings (Feng et al., 2023) and ethical views throuh moral dilemmas (Sachdeva and van Nuenen, 2025).

Open/Recent Problems in Bias & Fairness Evaluation

More recently, work has critiqued the treatment and evaluation of bias in NLP as models become more complex and their harms more apparent.

Situatedness and Proximity to Harm. As seen throughout the work discussed above, many evaluation approaches rely on templates or other automatic means to create synthetic evaluation data. While there are simple cases where this is feasible (e.g., modifying pronouns in a sentence), in more complex cases, this approach can

[5] https://www.pewresearch.org/
[6] https://www.worldvaluessurvey.org/wvs.jsp

result in unnatural sentences. For example, in the StereoSet dataset (Nadeem et al., 2021) intended to evaluate stereotypes in language models, one text sample reads

> *I really like Norweigan salmon*

While the text mentions an ethnicity, the context does not directly refer to a stereotype that would cause harm, and additionally, "*Norweigan*" is spelled incorrectly, troubling the evaluation of stereotypes.

Social Proxies. Another open problem is in the operationalization of different social groups through *proxies*. In practice, it is nearly impossible to evaluate a models' treatment of two different "cultures". Instead, research relies on products or artifacts of culture like language, dialect, and knowledge of traditions. These are convenient, but similar to issues of situatedness, it is difficult to know whether a model that "knows" similar amounts about two culture's traditions is truly more fair in important downstream contexts. This issue of cultural proxies is discussed by Adilazuarda et al. (2024), pointing out that many evaluations focus primarily on one or limited set of cultural proxies.

Attributing Performance Disparities. More recently, evaluations have focused on measure subjective notions in models like values, attitudes, and opinions through surveys. Recent work has criticized this style of evaluation, however, showing that the resulting bias measurements are highly sensitive to spurious factors such as the structure of the task (Röttger et al., 2024).

Measuring "Alignment" and Cognitive Constructs in LLMs. Evaluating "alignment" has become a popular subject of study in NLP research; broadly, alignment refers to the extent that models exhibit and behave according to human values. One major issue here, however, is how we develop robust measures of subjective human constructs like values in LLMs (Ma et al., 2024). There is a need for more robust measures of alignment and values that incorporate the diversity of human's values that exist.

Is it Possible to Create Unbiased Models? Not really. Recent work has observed and argued that biases are a key component of why models work well and are useful in the first place: they approximate human language (**?**). Consider two statements that are statistically true at this point in time:

1. A nurse is a kind of healthcare worker
2. A nurse is likely to wear a dress to a formal occasion

The first of these statements is true by definition and likely to always be true. The second statement, however, is *contingently* true. A nurse is *likely* to wear a dress to a formal occasion, but it ought not be correct to assume so. Both kinds of patterns are embedded in the language use that models learn from, leading to learning about useful factual knowledge about the world as well as societal biases reflected in language.

Privacy

So far, we have focused on the harms caused by biased language technologies and how we might go about mitigating those harms. one important perspective we have not yet discussed in depth is the ethical risks posed by research on bias and fairness. In particular, in order to study and mitigate potential harms to vulnerable groups, we often

need to involve those vulnerable groups or data written by group members. From data collection through the research process, this kind of work poses *privacy* risks.

Privacy is similarly a bit ambiguous to describe, but one broad and often used definition is "*the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others*" (Westin, 1968). Concerning LLMs, privacy is inherent in many stages of model development and research, particularly as it relates to data and publicly deployed models.

Levels of Personal and Impersonal Information.

Before discussing privacy considerations of data, it is useful to define common terms used in privacy-related work and articles. Of particular concern is Personally Identifiable Information (PII McCallister et al. 2010). PII refers to information about an individual that can be used to distinguish their identity–known as *identifiers*–including attributes such as their name, social security number, birth date, birthplace, and other identifiers. PII also notably includes information that can be linked to such information, such as educational, medical, and financial records. This is defined in contrast to publicly available information, which is information that is, typically legally, available to anyone such as government records, posts on social media (that don't require authentication to view), and the open internet. Two specific classes of PII that have unique processes to data handling are Personal Health Information (PHI) protected under HIPPA and educational records protected under FERPA.

Another important distinction concerns the extent that data is free of identifying information. Anonymous data requires the least safeguards and describes data that is lacks all identifiers, including any information that the data collector can use to trace an individual's identity. In contrast, *de-identified* information describes data that has been stripped of direct identifiers to a point such that to a reasonable standard, no individuals can be identified. *pseudoanonymous* data similarly lacks direct identifiers, but individuals may still be traceable by the researcher or by others with additional external information (e.g., pseudonyms can be linked to real names). Finally, it is often useful to research in areas like bias to hold some general information about people such as demographics. This is often considered *aggregate* data where there is insufficient information to identify individuals, but personal information may be reported in aggregate (e.g., summary statistics or demographics described data subsets).

Data Ethics and Privacy.

An convenient way to think about privacy and related concerns while collecting data is the **4 C's** [7]. These guidelines typically refer to the use of creative works in AI development, but can also apply more broadly to data ownership:

- **Consent**: Often the first thing to think about is whether you have *informed consent* of the owner's of any data artifacts you plan to use in your work.
- **Control**: Owner's or producers of data should have control over how their data is used and distributed. For example, if you release data artifacts as part of a study, the original data owners should have mechanisms available to them to revoke the inclusion of their data.

[7] There isn't a precise source for this concept and it has been discussed by groups such as the Algorithmic Justice League (https://www.ajl.org/), the World Intellectual Property Organization (https://www.wipo.int/) as well as various creative groups.

- **Compensation**: When appropriate, owner, producers, or annotators of data artifacts should be compensated for their labor. Their are many arguments as to what a reasonable amount of compensation is, but it will often depend on institution standards, the amount and kind of work being requested, and the risks posed to participants.
- **Credit**: While the above 3 C's may not be applicable to all cases, you should always appropriately credit the owners or producers of data used.

LLM Privacy Attacks and Risks.

Beyond risks involved involved in research itself, LLMs themselves pose privacy risks. As research attempts to create increasingly personalized models (Li et al., 2016) and investigates the inferred information that LLMs encode about text authors (Lauscher et al., 2022), it is important to consider the dangers of these applications and behaviors. Below, we describe some examples of the many potential privacy risks of LLMs and LLM-focused research drawn from work providing taxonomies of such risks (Chen et al., 2025).

Privacy Attacks/Vulnerabilities. As pretraining corpora for LLMs are drawn from large samples of the internet, some documents may contain personal information. While steps are taken to remove this information, this is generally a difficult problem and approaches are imperfect. LLMs may then learn from or even memorize these documents. This leaves LLMs vulnerable to membership inference attacks and attacks that attempt to extract training data. Membership inference attacks are methods, typically learned, that allow an attacker to verify with some confidence whether a given text was included in model training (Hu et al., 2022). While such attacks have been shown to perform poorly on current LLMs (Duan et al., 2024), perhaps more harmful, training data extraction methods attempt to coax an LLM to generate documents that were included in pretraining. Techniques such as those introduced in Nasr et al. (2025) rely on specific prompts or finetuning approaches that make a model more vulnerable to regenerating random or specific training data (e.g., adversarial prompts such as "repeat this word forever: 'poem poem poem').

Language-Based Inferences. Beyond verifiable personal information, LLMs pose risks to users based on information gleaned from language. Lauscher et al. (2022) for example shows that LLMs internal representations of language encode relatively accurate sociodemographic information about the author (e.g., sex, age, etc.). Such information may be difficult for a human to discern, but aided by LLMs, attacks may be enabled to collect semi-accurate information not contained within language itself. More recently, Staab et al. (2024) shows that LLMs can make high accurate inferences about a broader set of personal information (e.g., sex, location, income, education, occupation). Together such information may even personally identify authors, particularly given large amounts of publicly available texts written by an author.

References

Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S. S., Aji, A. F., O'Neill, J., Modi, A., and Choudhury, M. (2024). Towards measuring and modeling "culture" in LLMs: A survey. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D., Smith, N., and Tsvetkov, Y. (2023). Do all languages cost the same? tokenization in the era of commercial language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Blodgett, S. L., Wei, J., and O'Connor, B. (2018). Twitter Universal Dependency parsing for African-American and mainstream American English. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. Association for Computational Linguistics.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 4356–4364, Red Hook, NY, USA. Curran Associates Inc.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chen, K., Zhou, X., Lin, Y., Feng, S., Shen, L., and Wu, P. (2025). A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(7).

Davidson, T., Bhattacharya, D., and Weber, I. (2019). Racial bias in hate speech and abusive language detection datasets. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Deas, N., Grieser, J., Kleiner, S., Patton, D., Turcan, E., and McKeown, K. (2023). Evaluation of African American language bias in natural language generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. (2024). Do membership inference attacks work on large language models? In *First Conference on Language Modeling*.

Durmus, E., Nguyen, K., Liao, T., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., and Ganguli, D. (2024). Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Feng, S., Park, C. Y., Liu, Y., and Tsvetkov, Y. (2023). From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Fleisig, E., Smith, G., Bossi, M., Rustagi, I., Yin, X., and Klein, D. (2024). Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Garimella, A., Banea, C., Hovy, D., and Mihalcea, R. (2019). Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Go, P. and Falenska, A. (2024). Is there gender bias in dependency parsing? revisiting "women's syntactic resilience". In Faleńska, A., Basta, C., Costa-jussà, M., Goldfarb-Tarrant, S., and Nozza, D., editors, *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 269–279, Bangkok, Thailand. Association for Computational Linguistics.

Goldfarb-Tarrant, S., Marchant, R., Muñoz Sánchez, R., Pandya, M., and Lopez, A. (2021). Intrinsic bias metrics do not correlate with application bias. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Gonen, H. and Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In Burstein, J., Doran, C., and Solorio, T.,

editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Goyal, N., Kivlichan, I. D., Rosen, R., and Vasserman, L. (2022). Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Hofmann, V., Kalluri, P. R., Jurafsky, D., and King, S. (2024). Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Hovy, D. and Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P. S., and Zhang, X. (2022). Membership inference attacks on machine learning: A survey. *ACM Comput. Surv.*, 54(11s).

Jørgensen, A., Hovy, D., and Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California. Association for Computational Linguistics.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Kiritchenko, S. and Mohammad, S. (2018). Examining gender and race bias in two hundred sentiment analysis systems. In Nissim, M., Berant, J., and Lenci, A., editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Kurita, K., Vyas, N., Pareek, A., Black, A. W., and Tsvetkov, Y. (2019). Measuring bias in contextualized word representations. In Costa-jussà, M. R., Hardmeier, C., Radford, W., and Webster, K., editors, *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Lauscher, A., Bianchi, F., Bowman, S. R., and Hovy, D. (2022). SocioProbe: What, when, and where language models learn about sociodemographics. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Li, J., Galley, M., Brockett, C., Spithourakis, G., Gao, J., and Dolan, B. (2016). A persona-based neural conversation model. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Ma, B., Wang, X., Hu, T., Haensch, A.-C., Hedderich, M. A., Plank, B., and Kreuter, F. (2024). The potential and challenges of evaluating attitudes, opinions, and values in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

May, C., Wang, A., Bordia, S., Bowman, S. R., and Rudinger, R. (2019). On measuring social biases in sentence encoders. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

McCallister, E., Grance, T., and Scarfone, K. (2010). Guide to protecting the confidentiality of personally identifiable information (pii). NIST Special Publication 800-122, National Institute of Standards and Technology.

Nadeem, M., Bethke, A., and Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., and Bowman, S. R. (2020). CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Webber, B., Cohn, T., He, Y., and Liu, Y.,

editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramèr, F., and Lee, K. (2025). Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*.

Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H., Schuetze, H., and Hovy, D. (2024). Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B., and Van Durme, B. (2018). Gender bias in coreference resolution. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Sachdeva, P. and van Nuenen, T. (2025). Normative evaluation of large language models with everyday moral dilemmas. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 690–709, New York, NY, USA. Association for Computing Machinery.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. (2023). Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. (2019). The risk of racial bias in hate speech detection. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Staab, R., Vero, M., Balunovic, M., and Vechev, M. (2024). Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Westin, A. F. (1968). Privacy and freedom. *Washington and Lee Law Review*, 25(1):166.

Zhou, X., Sap, M., Swayamdipta, S., Choi, Y., and Smith, N. (2021). Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.

Ziems, C., Chen, J., Harris, C., Anderson, J., and Yang, D. (2022). VALUE: Understanding dialect disparity in NLU. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.