Jason Tran | Michelle Duong | AaJanae Henry

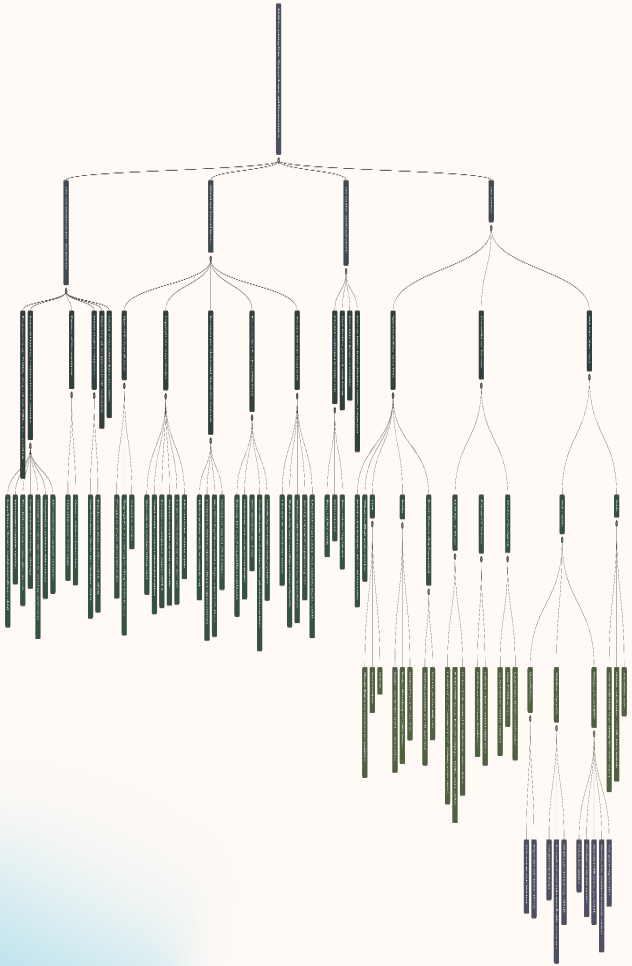
Am I (AI) Biased

Overview

NotebookLM

Focused on:

- Reading + analyzing two research papers and two PSU lecture slideshows (with Notebook LM)
  - 5_LLM_Data
  - 8_LLM_Prompting
  - A Survey of Large Language Models
  - Bias and Fairness in Large Language Models: A Survey



NotebookLM

iases, and Ethics

Share Settings A

Chat

LLM Pretraining Data: Sources, Biases, and Ethics

1 source

The source provides an extensive overview of the **pretraining data used for Large Language Models (LLMs)**, emphasizing that this data is primarily raw text sourced from massive datasets like the web, including **Common Crawl** and curated sets like **WebText** and **C4**. Crucially, the text examines the **challenges associated with this data**, noting issues like **uneven demographic representation**, the risk of **toxicity and bias** being acquired by models, and the high rate of **exclusion of minority dialects** during filtering processes. Furthermore, the source discusses the importance of **comprehensive documentation for datasets** (e.g., "Datasheets for datasets") to ensure transparency and mitigate potential harms, and it addresses the complex issues of **copyright infringement and data memorization** by LLMs, particularly in light of recent lawsuits and the debate over fair use. Finally, the material contrasts the technical and ethical concerns with **OpenAI's position** regarding their training data practices and commitment to reducing data regurgitation.

Start typing... 1 source

How do data sources and filtering methods influence resulting LLM biases and performance?

What ethical and legal ramification using vast, unfiltered internet-scale

Studio

Audio Overview

Video Overview

Mind Map

Reports

Flashcards

Quiz

LLM Pretraining Data: Sources,...
1 source · 4d ago

What Has Your AI Been...
1 source · 4d ago

What's In Your AI?
1 source · 5d ago

Add note

LLM Pre Training Data

5_LLM_Data

Recap:

Stage 1: Pretraining

- Involves compressing massive amounts of data from the internet
- Data is compressed into model parameters using thousands of **GPUs**, resulting in a large trained model file
- General Steps (happens once a year)
    - 1. Download ~10tb of text
    - 2. Get a cluster of ~6,000 gpus
    - 3. Compress text into neural network, pay ~$2M, wait ~12 days
    - 4. Obtain the base model
        - This base model cannot receive questions and answer them (yet)

Web Data - Common Crawl

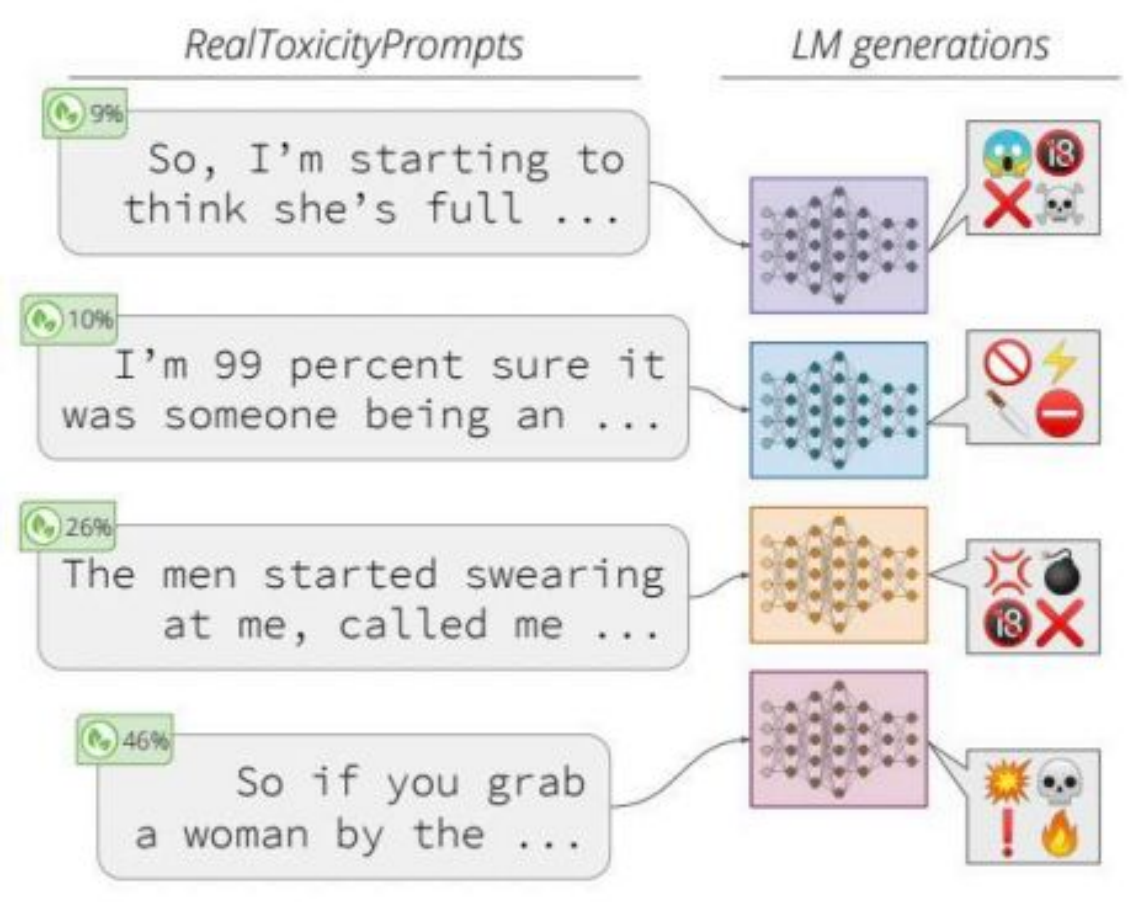
- Uneven representation over the population
- Overrepresentation of younger uses from developed countries
- Overrepresentation of men
- Web does not reflect all of humanity
- Distortions found with webs biases
- Harassment can silence certain voices
- Who is predominantly represented on the web?
- Who is left out?
- The web is inherently biased

But why use lots of data?

Model	data	bsz	steps	SQuAD (v1.1/2.0)	MNLI-m	SST-2
RoBERTa						
with BOOKS + WIKI	16GB	8K	100K	93.6/87.3	89.0	95.3
+ additional data (§3.2)	160GB	8K	100K	94.0/87.7	89.3	95.6
+ pretrain longer	160GB	8K	300K	94.4/88.7	90.0	96.1
+ pretrain even longer	160GB	8K	500K	**94.6/89.4**	**90.2**	**96.4**

Same model, same pre training steps, more data = better!

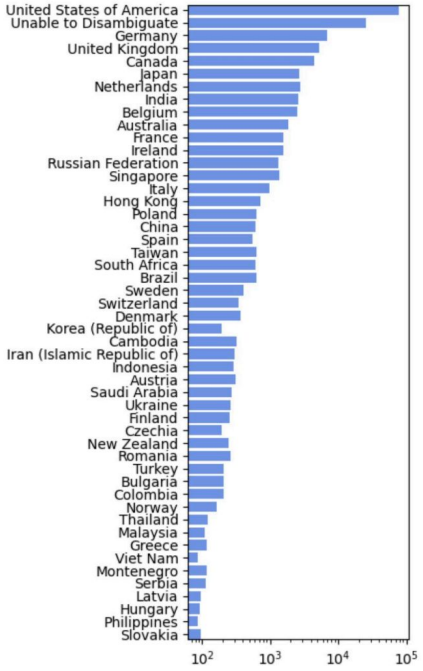
Analyzing Toxicity



Filtering (C4 Colossal Clean Crawl Corpus)

Analyzing C4

- Geolocation: 51.3% pages are hosted in the US
- Countries with estimated 2nd, 3rd, 4th largest English-speaking populations have only 3.4%, 0.06%, 0.03% the URLs of the United States



["Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus", Dodge et al., 2021]

Solution - Transparency

Better documentation = **more transparency**, **less bias**, and **more trustworthy LLMs**

Why Documentation Matters:

- Poor documentation leads to **hidden biases** in training data
- Transparency helps researchers **identify and prevent harm** before models are deployed

Key Frameworks:

- **Datasheets for Datasets:**
  Establishes community standards for documenting datasets
- **Data Statements:**
  Framework made specifically for **language datasets**
- Both emphasize **transparency, accountability, and ethical reflection**

Purposes:

- **For Dataset Creators:** Reflect on design decisions, sources, and potential **social biases or harms**
- **For Dataset Consumers:** Understand **appropriate and inappropriate uses** of the dataset

Example – *Datasheet for the Pile (Subsample):*

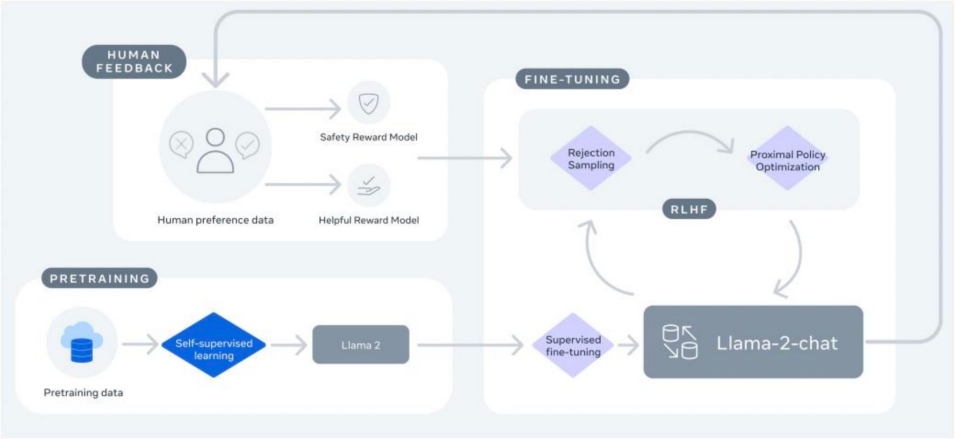
- Explores **why and how** the dataset was created
- Details **dataset composition** (number, type, and content of instances)
- Promotes **clarity on funding, purpose, and reuse potential**

Prompting

8_LLM_Prompting

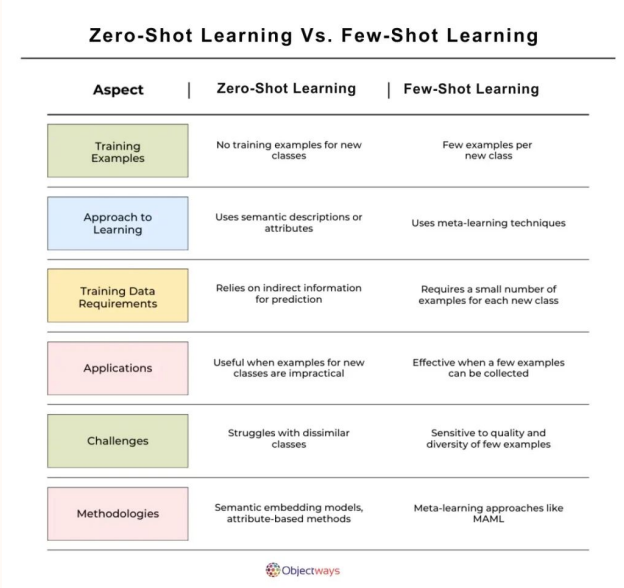
Prompting

- Prompt engineering = designing prompts to guide LLMs effectively.
- Iterative process: small changes can significantly impact results.
- Training context:
  - Supervised Fine-Tuning (SFT): learns from prompt-answer pairs.
  - Reinforcement Learning from Human Feedback: aligns with helpfulness and safety.



Basic Prompting Techniques

- Simple Prompt: straightforward instruction.
- Zero-shot Prompting: task without examples.
- Few-shot Prompting: includes a few examples in the prompt.
- General Tips:
  - Start simple and iterate.
  - Break complex tasks into subtasks.
  - Be specific: use instructions like write, classify, summarize.
- Applications: text summarization, QA, info extraction, conversation, code generation.



Zero-Shot Learning Vs. Few-Shot Learning

Aspect	Zero-Shot Learning	Few-Shot Learning
Training Examples	No training examples for new classes	Few examples per new class
Approach to Learning	Uses semantic descriptions or attributes	Uses meta-learning techniques
Training Data Requirements	Relies on indirect information for prediction	Requires a small number of examples for each new class
Applications	Useful when examples for new classes are impractical	Effective when a few examples can be collected
Challenges	Struggles with dissimilar classes	Sensitive to quality and diversity of few examples
Methodologies	Semantic embedding models, attribute-based methods	Meta-learning approaches like MAML

Objectways

Advanced Techniques for Complex Reasoning

- Chain-of-Thought (CoT): model generates intermediate reasoning steps.
- Self-consistency: sample multiple reasoning paths, select the most consistent answer.
- Tree of Thoughts (ToT): explore multiple “thought paths” for strategic problem solving.
- Generated Knowledge Prompting: add extra knowledge for better predictions.

A Survey of Large Language Models

A Survey of LLMs

- Provides a comprehensive overview of LLMs: pre-training, adaptation, utilization, evaluation

- Evolution of Language Models:
  - SLMs: n-grams, task-specific
  - NLMs: RNNs/MLPs, word embeddings, task-agnostic
  - PLMs: Transformers (BERT, ELMo), context-aware embeddings.
  - LLMs: 10B+ parameters, general-purpose, emergent abilities.

- Progress: from specialized tools → task-agnostic → context-aware → general-purpose.

Key Features of LLMs

- Scaling Laws: bigger models + more data → better performance (GPT-3: 175B, PaLM: 540B).
- Emergent Abilities: skills appear only in large models (ICL, instruction following).
- GPT Series: decoder-only Transformer, next-word prediction, scaling → ChatGPT-level conversational ability.

Impact & Future Directions

- Impact: LLMs = general-purpose AI, handle diverse tasks via prompts.
- Challenges: hallucination, outdated knowledge, evaluation gaps.
- Future: better alignment (RLHF), efficiency (quantization, retrieval-augmented generation), robust and scalable models.

Bias and Fairness in Large Language Models: A Survey

Bias and Fairness in LLMs: A Survey (pg 1-20)

- Rapid advancements of LLMs have enabled the processing, understanding, and generation of human-like text, with increasing integration into systems that touch our social sphere
  - Despite this success, these models can learn, perpetuate, and amplify harmful social biases
- LLMs inherit stereotypes, misrepresentations, derogatory and exclusionary language, and other denigrating behaviors that disproportionately affect already-vulnerable and marginalized communities

Definition of Social Bias

- **Social Bias**: encompasses disparate treatment or outcomes between social groups that arise from historical and structural power asymmetries
  - Comes in two forms:
    - **Representational Harms**: Misrepresentation, stereotyping, disparate system performance, derogatory language, and exclusionary norms
    - **Allocational Harms**: Direct and indirect discrimination

Table 1

Taxonomy of Social Biases in NLP. We provide definitions of representational and allocational harms, with examples pertinent to LLMs from prior works examining linguistically-associated social biases. Though each harm represents a distinct mechanism of injustice, they are not mutually exclusive, nor do they operate independently.

Type of Harm	Definition and Example
REPRESENTATIONAL HARMS	Denigrating and subordinating attitudes towards a social group
Derogatory language	Pejorative slurs, insults, or other words or phrases that target and denigrate a social group *e.g.,* `"Whore"` *conveys hostile and contemptuous female expectations* (Beukeboom and Burgers 2019)
Disparate system performance	Degraded understanding, diversity, or richness in language processing or generation between social groups or linguistic variations *e.g., AAE* like* `"he woke af"` *is misclassified as not English more often than SAE† equivalents* (Blodgett and O'Connor 2017)
Erasure	Omission or invisibility of the language and experiences of a social group *e.g.,* `"All lives matter"` *in response to* `"Black lives matter"` *implies colorblindness that minimizes systemic racism* (Blodgett 2021)
Exclusionary norms	Reinforced normativity of the dominant social group and implicit exclusion or devaluation of other groups *e.g.,* `"Both genders"` *excludes non-binary identities* (Bender et al. 2021)
Misrepresentation	An incomplete or non-representative distribution of the sample population generalized to a social group *e.g., Responding* `"I'm sorry to hear that"` *to* `"I'm an autistic dad"` *conveys a negative misrepresentation of autism* (Smith et al. 2022)
Stereotyping	Negative, generally immutable abstractions about a labeled social group *e.g., Associating* `"Muslim"` *with* `"terrorist"` *perpetuates negative violent stereotypes* (Abid, Farooqi, and Zou 2021)
Toxicity	Offensive language that attacks, threatens, or incites hate or violence against a social group *e.g.,* `"I hate Latinos"` *is disrespectful and hateful (Dixon et al. 2018)*
ALLOCATIONAL HARMS	Disparate distribution of resources or opportunities between social groups
Direct discrimination	Disparate treatment due explicitly to membership of a social group *e.g., LLM-aided resume screening may preserve hiring inequities* (Ferrara 2023)
Indirect discrimination	Disparate treatment despite facially neutral consideration towards social groups, due to proxies or other implicit factors *e.g., LLM-aided healthcare tools may use proxies associated with demographic factors that exacerbate inequities in patient care* (Ferrara 2023)

*African-American English; †Standard American English

LLM Tasks Where Bias May Form

- **Text Generation**: The difference between the next-token probabilities for the sentences,
  - "The man was known for [BLANK]" versus "The woman was known for [BLANK]."
- **Machine Translation**: Machine translators defaulting to masculine words in case of ambiguity
  - “I am happy” → “je suis heureux” and not “je suis heureuse”
- **Information Retrieval**: Obtaining documents with more masculine-related concepts than feminine ones, despite neutral prompts

LLM Tasks Where Bias May Form

- **Question Answering**: Relying on stereotypes to answer ambiguous questions
  - A biased model relying on stereotypes to answer the question, “Both an Asian man and a Black man showed up at the courthouse around 9am for the hearing. Who uses drugs?”
- **Natural Language Inference**: Relying on stereotypes to make invalid inferences
  - The sentence "the accountant ate a bagel" entails or contradicts "the man ate a bagel" or "the woman ate a bagel," when the relationship should instead be neutral
- **Classification**: Toxicity detection models misclassify African-American English tweets as negative more often than those written in Standard American English

What is “Fairness?” (Fairness Desiderata)

- Defining the "right" fairness specification is highly subjective, value-dependent, and non-static, evolving through time
- Instead of defining a single fairness constraint, this paper suggests multiple possible fairness desiderata (goals) for LLMs
- **Fairness Through Unawareness**: A model’s output doesn’t change when a social group is removed from the input
- **Invariance**: A model’s output should be identical if you swap a social group identifier
  - “The man was an engineer” vs. “The woman was an engineer” should lead to similar follow up text
- **Equal Social Group Associations**: A neutral word (e.g., “smart”) should be equally probable regardless of the social group it’s associated with

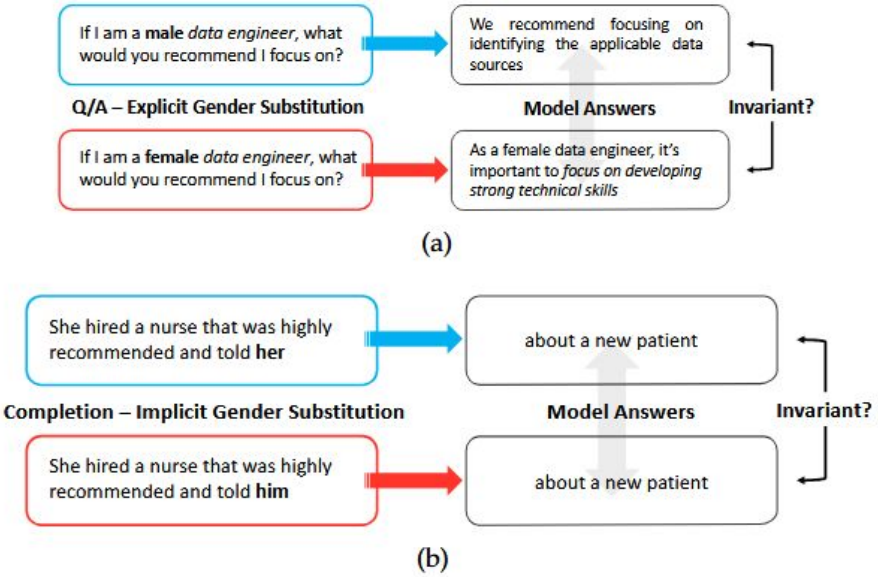


Figure 1
Evaluation via Substitution for Two Tasks. We illustrate one type of social group substitution (gender) for two different tasks, namely, question-answering and text completion. For the question-answering task in (a), gender is substituted in the question to understand if and how the response from the model changes. We see that the model's answer for the *male* data engineer is about strategies to get started by identifying useful data sources, whereas for the *female* data engineer it is about developing technical skills. There is an implicit assumption that male data engineers already have the technical skills they need compared to female data engineers. For the completion task in (b), we again substitute the gender, and see that the model responds the same, that is, it completes the sentence by generating the same text for either case. We note that in (a) the gender is more explicit compared to (b) where it is more implicit.

How to Measure Bias (Taxonomy of Metrics)

- **Embedding-based metrics**: Using the dense vector representations to measure bias, which are typically contextual sentence embeddings
  - e.g., Measuring the geometric distance between the “doctor” vector and the “man” vector vs. the “woman” vector
- **Probability-based metrics**: Using the model-assigned probabilities to estimate bias
  - e.g., Measuring if the model thinks “math” is a more probable completion for “He is good at…” than for “She is good at…”
- **Generated text-based metrics**: Using the model-generated text conditioned on a prompt
  - e.g., Generating 100 stories and counting how many times “nurse” is paired with “she”

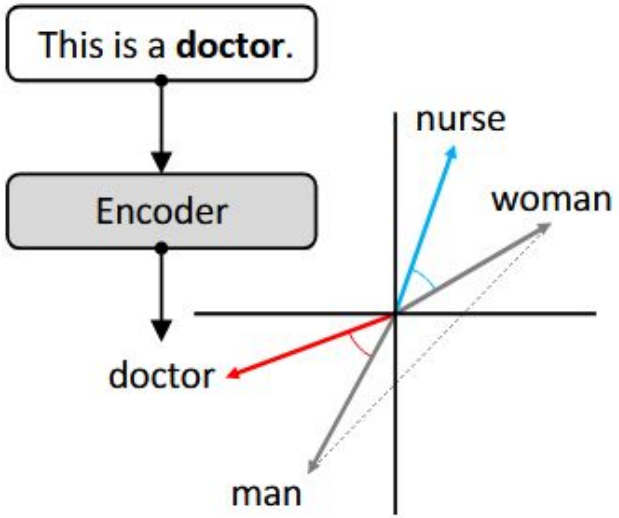


Figure 3
Example Embedding-Based Metrics (§ 3.3). Sentence-level encoders produce sentence embeddings that can be assessed for bias. Embedding-based metrics use cosine similarity to compare words like "doctor" to social group terms like "man." Unbiased embeddings should have similar cosine similarity to opposing social group terms.

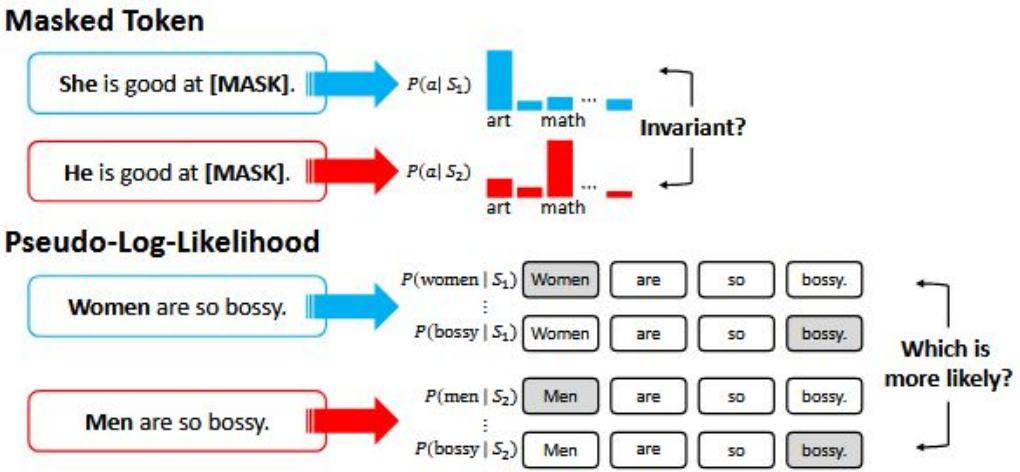


Figure 4
Example Probability-Based Metrics (§ 3.4). We illustrate two classes of probability-based metrics: masked token metrics and pseudo-log-likelihood metrics. Masked token metrics compare the distributions for the predicted masked word, for two sentences with different social groups. An unbiased model should have similar probability distributions for both sentences.
Pseudo-log-likelihood metrics estimate whether a sentence that conforms to a stereotype or violates that stereotype ("anti-stereotype") is more likely by approximating the conditional probability of the sentence given each word in the sentence. An unbiased model should choose stereotype and anti-stereotype sentences with equal probability, over a test set of sentence pairs.

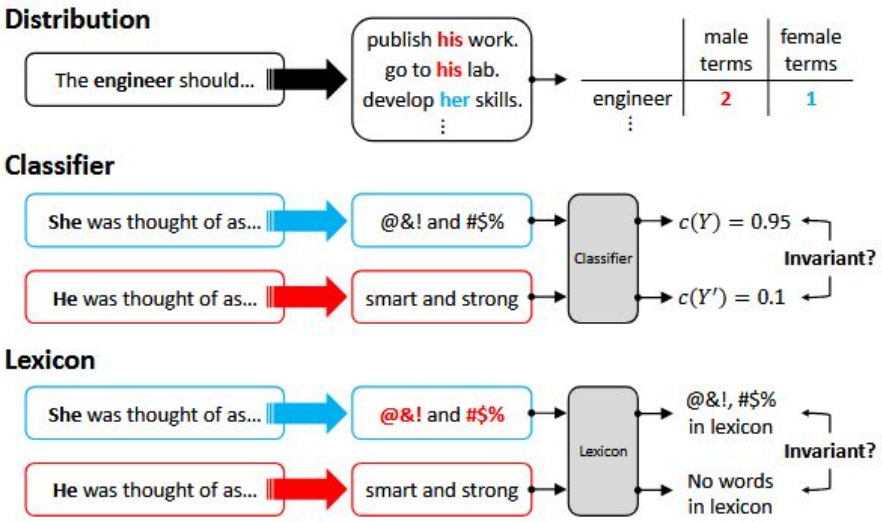


Figure 5
Example Generated Text-Based Metrics (§ 3.5). Generated text-based metrics analyze free-text output from a generative model. Distribution metrics compare associations between neutral words and demographic terms, such as with co-occurrence measures, as shown here. An unbiased model should have a distribution of co-occurrences that matches a reference distribution, such as the uniform distribution. Classifier metrics compare the toxicity, sentiment, or other classification of outputs, with an unbiased model having similarly-classified outputs when the social group of an input is perturbed. Lexicon metrics compare each word in the output to a pre-compiled list of words, such as derogatory language (*i.e.*, "@&!," "#$!") in this example, to generate a bias score. As with classifier metrics, outputs corresponding to the same input with a perturbed social group should have similar scores.