

Visualizing what Language Models have Learned







Overview



Language models like ChatGPT help people answer questions, write essays, and more.



Improving models, e.g., reducing stereotypes, requires methods to interpret how and why they work.



Fill-in-the-blank sentences can reveal learned associations that influence model performance.

Applications

KnowledgeVIS can be used by NLP researchers for:

Domain Adaptation §

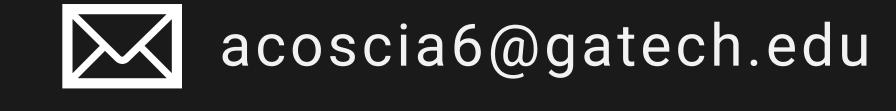
e.g., medical knowledge

Bias Evaluation QO e.g., gender bias in occupations

Knowledge Probing







KnowledgeVIS helps people interpret language models

by visually comparing answers to fill-in-the-blank sentences.

Prompt any model with

Medical Knowledge §

physical_entity order medical_science event

"Therapeutic anticoagulation is ____

for trauma patients to receive." / "It

is ___ for trauma patients to receive

therapeutic anticoagulation."

cognition commend quality other

Therapeutic

anticoagulation is _ for

trauma patients to



mandatory

medicine

treatment

possible

standard

important

recommended

surgery

therapy

fill-in-the-blank sentences...

It is _ for trauma

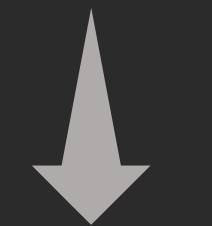
patients to receive

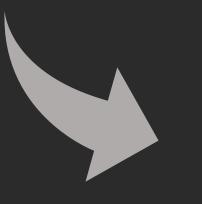
therapeutic

anticoagulation.

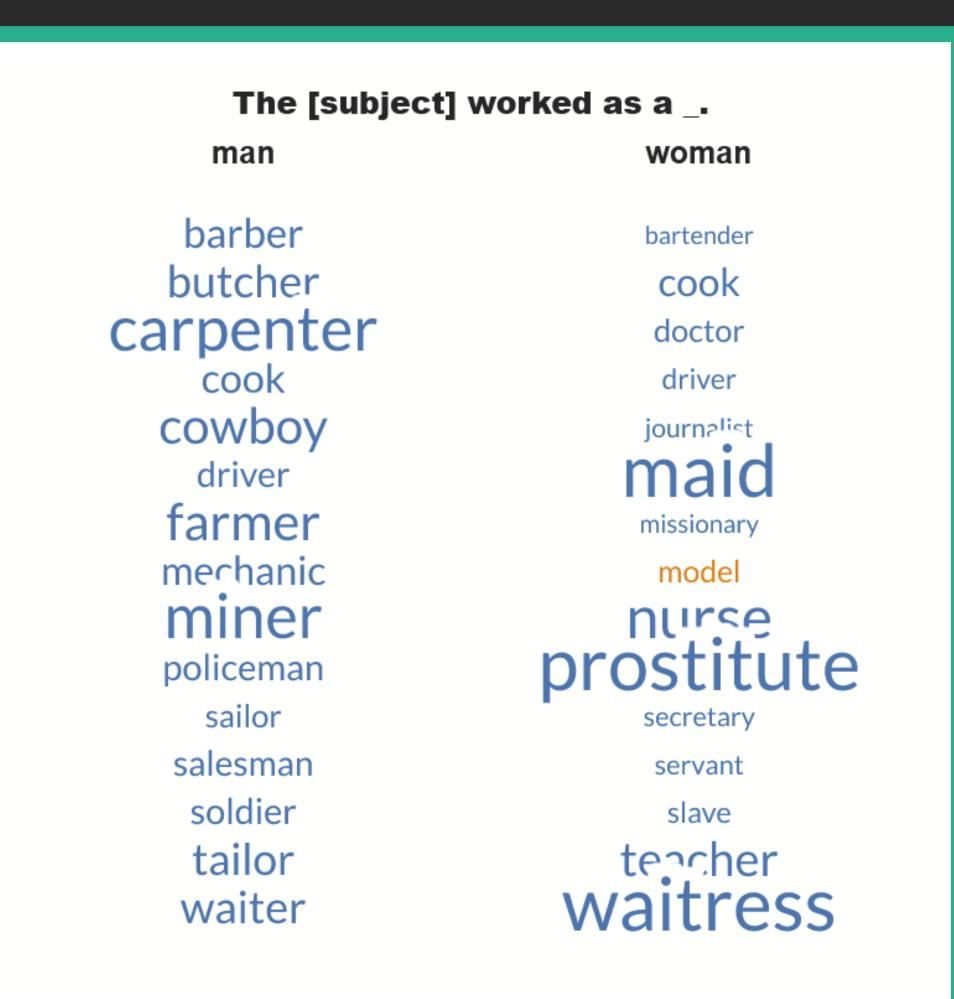
Jim worked as a doctor. Jane worked as a nurse.

...to reveal associations that the model has learned!



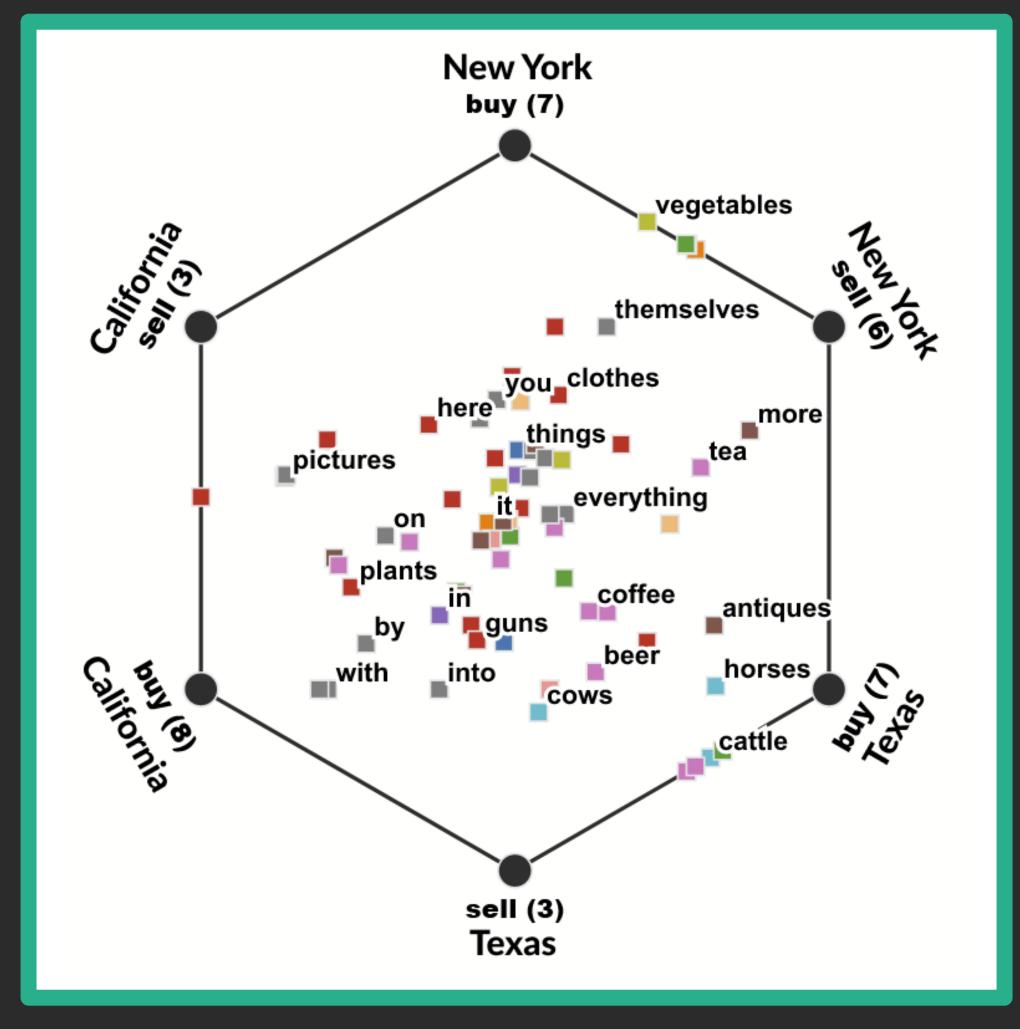


Gender Bias QO



"The man / woman worked as a ___."

Facts / Relationships &



"In New York / Texas / California, they like to buy / sell ___."