# KnowledgeVIS

## Visualizing what Language Models have Learned



Adam Coscia


Alex Endert

Visual Analytics Lab
@ Georgia Tech


SCAN ME

## 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**
e.g., gender bias in occupations

**Knowledge Probing**
e.g., facts and relationships

SCAN ME


acoscia6@gatech.edu


# KnowledgeVIS helps people interpret language models

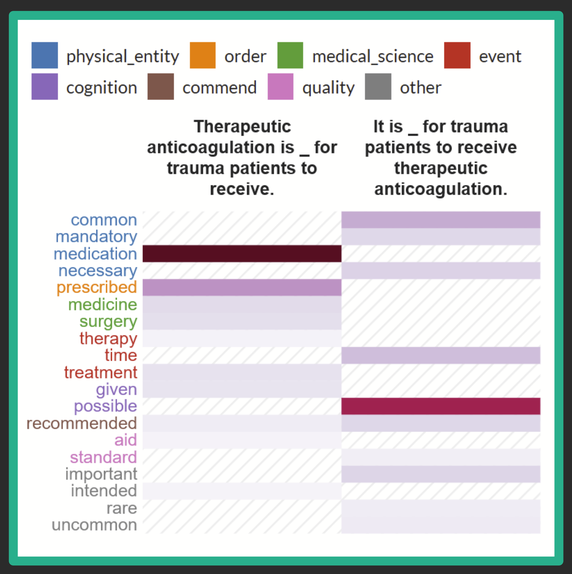
by visually comparing answers to fill-in-the-blank sentences.

Prompt any model with **fill-in-the-blank sentences**...

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

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

## Medical Knowledge

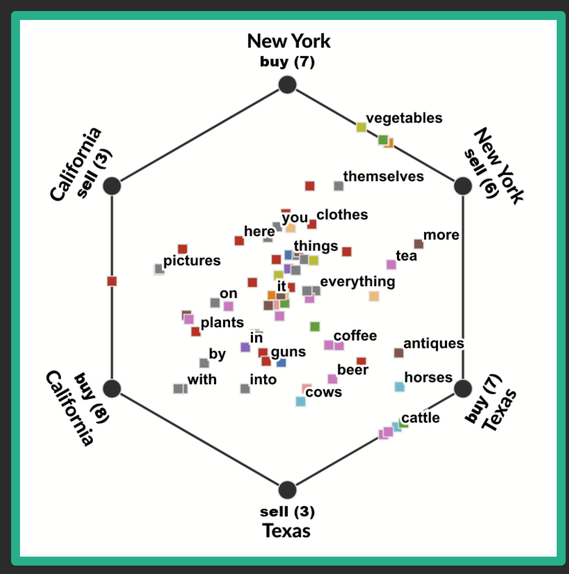



"Therapeutic anticoagulation is ___ for trauma patients to receive." / "It is ___ for trauma patients to receive therapeutic anticoagulation."

## Gender Bias


"The man / woman worked as a ___."

## Facts / Relationships




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