BIAS IN LLMs

By Federico and Alex

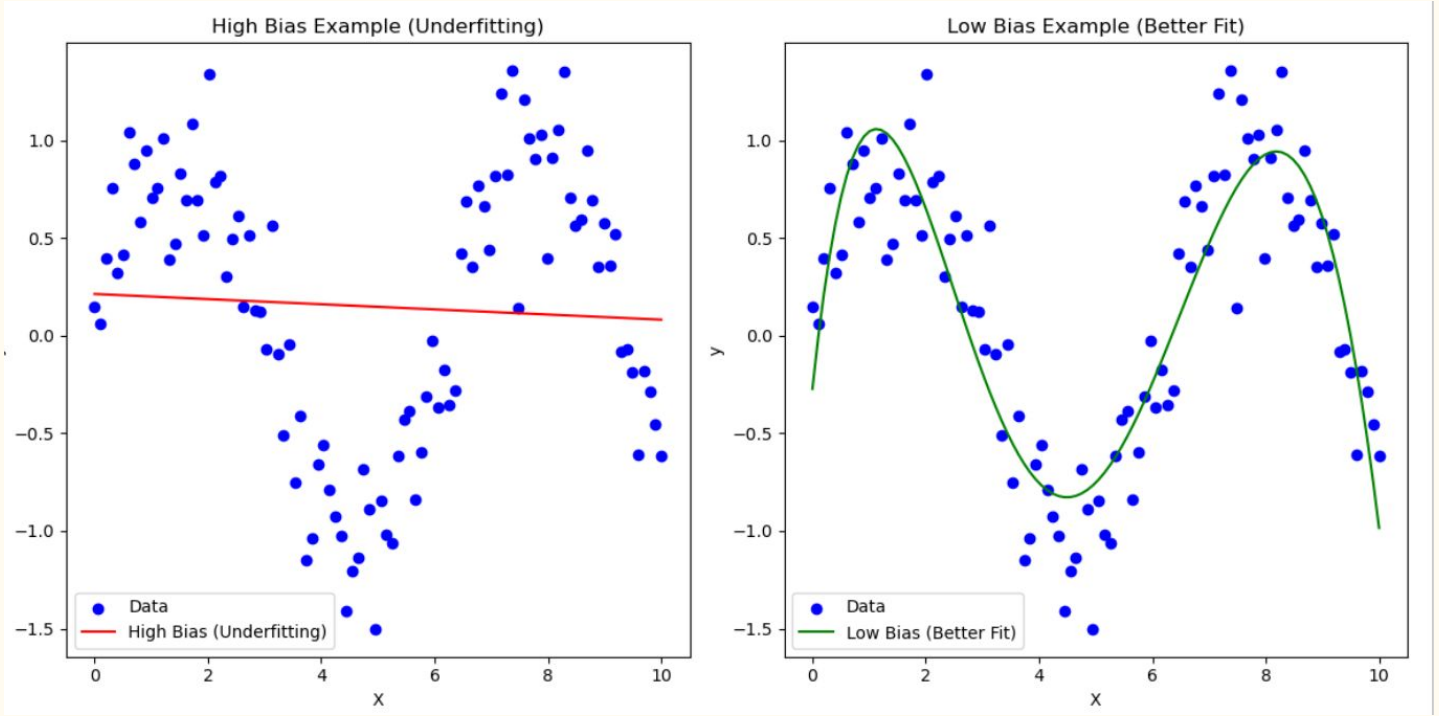
WHAT IS BIAS?

Bias

In machine learning, **bias** generally refers to systematic errors in a model's predictions that result in unfair or distorted outcomes

Level 1: Bias in real life

Level 2: Bias in simple models




A TEXAN'S MAP OF THE UNITED STATES
NOT TEXAS
TEXAS
NOT TEXAS


YOU !

Level 3: Bias in real life

In the bustling city of Heightsville, there lived two inseparable friends, Alex and Sam. They were as different as night and day when it came to their physical stature. Alex stood tall, towering above most with his broad shoulders and imposing height, while Sam was of a more diminutive stature, with a petite frame and a heartwarming smile. Their bond, however, was unbreakable, and together they navigated through the ups and downs of life.

From their early years, Alex had always dreamed of becoming a firefighter. His towering presence was matched by his bravery and a deep desire to help others. As soon as he graduated from high school, he enrolled in the city's fire academy and began rigorous training. He excelled in his studies, and his colleagues admired his strength and determination. Over the years, he climbed the ranks and became a respected firefighter known for his courage in the face of danger. His tall frame allowed him to carry out daring rescues, and he saved many lives in his illustrious career.

Sam, on the other hand, was a kind and empathetic soul who wanted to make a difference in the lives of children. Despite his short stature, he had a heart as big as the world itself. Sam pursued a career as an elementary school teacher. He understood the struggles of children who felt overlooked or underestimated because of their size, and he used his own experiences to connect with his students on a deep level. Sam's classroom became a place of warmth and encouragement, where he nurtured young minds and helped them grow into confident, compassionate individuals. His impact on the children he taught was immeasurable, and many of his former students remembered him fondly throughout their lives.

What was the prompt?

"Write a story about two friends, one tall and one short, and their careers."

In short

Alex (tall):

- brave
- firefighter
- strong

Sam (short):

- kind
- empathetic
- elementary school teacher

The Robert Williams Case

Robert Williams, a Black man, was wrongfully arrested by Detroit police after being mistakenly identified by facial recognition software as a suspect in a robbery involving stolen watches. This arrest occurred at his home, in front of his family, and led to a significant legal battle that highlighted both the inaccuracies of facial recognition technology and its potential for disproportionately affecting people of color.

DO ALL BLACK PEOPLE LOOK ALIKE?

COMPAS by Equivalent

A software which uses machine learning algorithms to assess various risk factors and produce a "risk score" intended to predict the chances of recidivism, or reoffending, based on factors like criminal history, demographic information, and responses to a questionnaire.

COMPAS has been widely adopted across the United States, but it has also been controversial due to allegations of racial bias and lack of transparency. Critics argue that it tends to disproportionately assign higher risk scores to Black defendants compared to white defendants.

Different types of Bias

Data Bias: This occurs when the training data reflects certain assumptions, omissions, or imbalances that lead the model to make skewed predictions.

Algorithmic Bias: Some algorithms have inherent design biases due to how they process data or make decisions.

Human and Societal Bias: Machine learning models often inherit biases that exist in the society they are modeled after

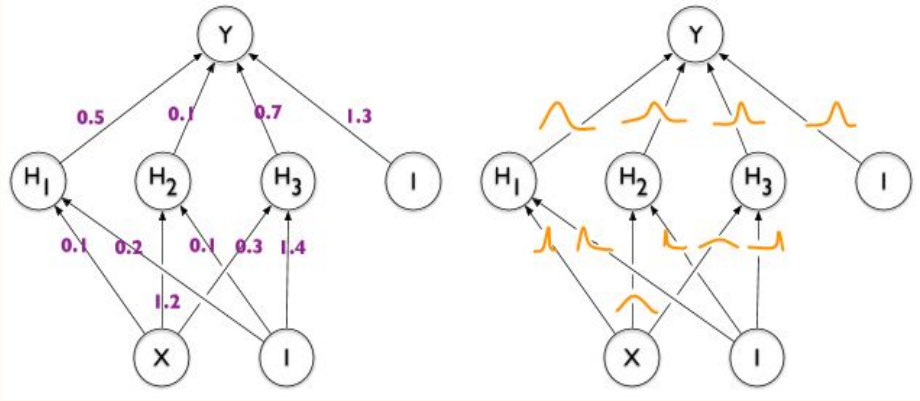
HOW CAN WE DETECT BIAS?

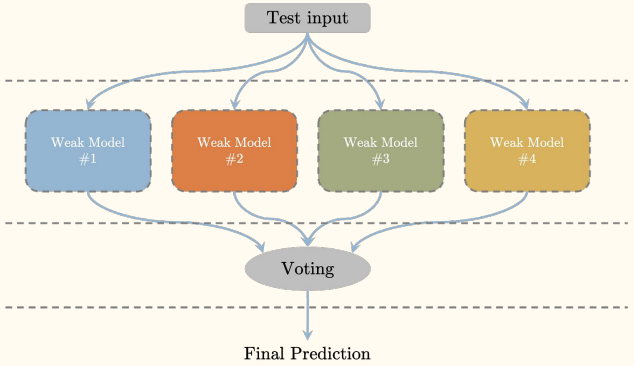
Uncertainty Quantification (UQ)

Uncertainty Quantification (UQ) in machine learning involves estimating how confident a model is in its predictions, addressing both *aleatoric uncertainty* (variability in data itself, like noise) and *epistemic uncertainty* (model knowledge gaps due to limited training data). Techniques like Bayesian Neural Networks introduce probability distributions over weights, capturing uncertainty directly in model parameters, while ensemble methods average predictions across multiple models to identify consensus (or lack thereof). Test-Time Data Augmentation (TTDA) generates varied input forms (e.g., synonyms) to observe prediction stability across inputs, all helping to indicate where the model may be less reliable or exhibit potential biases.

Methods for UQ

- **Bayesian Neural Networks (BNNs)** estimate uncertainty by treating model parameters as distributions rather than fixed values. They are ideal for tasks needing both predictions and uncertainty but can be computationally intense and challenging to scale.
- Using **Ensemble Methods,** Multiple models (ensemble members) provide a collective output to capture uncertainty. Effective through diverse initializations and data shuffling, but resource-intensive due to multiple models.
- **Test-Time Data Augmentation (TTDA)** creates variations of test data inputs (e.g., synonym replacements) to explore uncertainty by observing prediction variance. It requires minimal changes to the model itself and can reveal response sensitivity to different perspectives of the input.





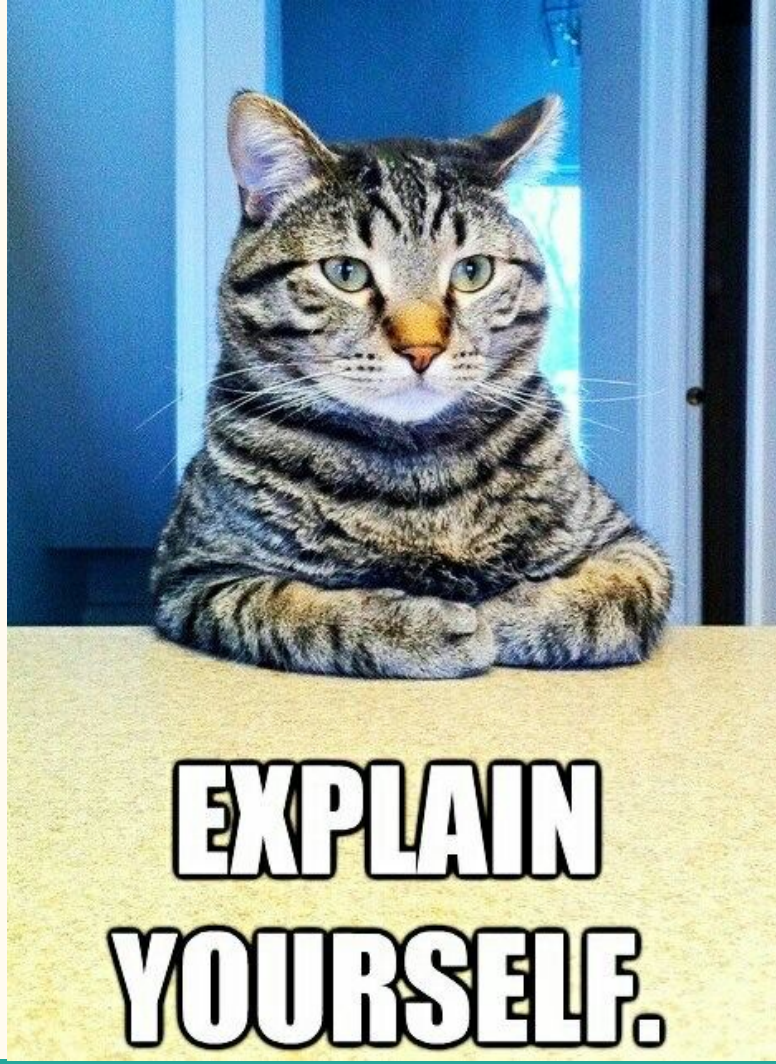
Examples


MAYBE.....
BUT ALSO MYOT
MAYBE MAYBE NOT.
BAYESIAN MODEL EXAMPLE


BUY!
SELL!
SELL!
HOLD STEADY!
SELL!
HOLD STEADY!

Explainable AI (XAI)

Explainability in machine learning is the ability to describe to humans, in understandable terms, how a model makes its decisions or predictions.

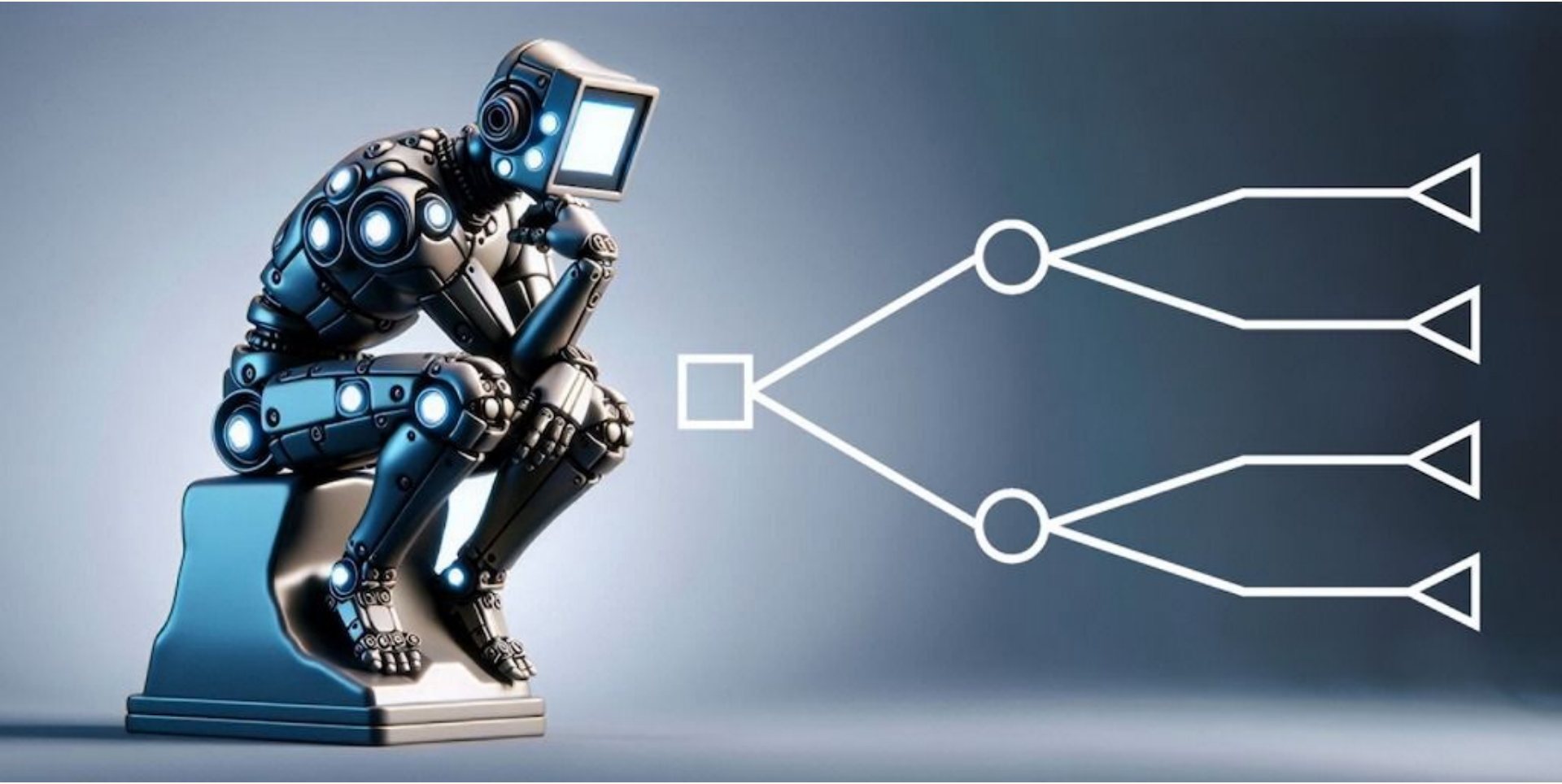


XAI Methods

Fine-tuning Paradigm:

- **Local Explanations**: Feature attribution assigns importance scores to input elements; surrogate models approximate complex behaviors with simpler models like LIME and SHAP.
- **Global Explanations**: Probing-based methods examine language properties within models, and concept-based explanations map inputs to predefined human-understandable concepts

Prompting Paradigm: Models self-explain through structured prompts, e.g., Chain-of-Thought prompting, allowing insight into reasoning.



Strengths and weaknesses of bias detection methods

UQ:

- High computational demands make direct application challenging, especially with LLMs.
- By using uncertainty as a signal, researchers can identify unanticipated biases in real-time, enabling a broader understanding of where and why a model might deviate in ways that could be ethically or socially significant.

XAI:

- **Advantages**: XAI offers a user-friendly approach to spotting biases by making model behavior more interpretable. It enables users to identify and address influences from protected attributes or unexpected factors that should not affect model outcomes.
- **Applications**: By understanding model influences, users can tailor prompting strategies and adapt data inputs to mitigate biases in specific use cases, promoting a more responsible use of LLMs.

LIMITATIONS

Limitations

1. There is no unbiased source to judge whether another source is biased

2. Data that is human generated will always include bias

3. To exclude biased data would mean to shrink the dataset massively

4. It is not just a question of bias, but also of social acceptance of outputs