

The background of the slide is a complex network diagram. It consists of numerous small, semi-transparent circular nodes in shades of brown and gold, interconnected by thin, light-colored lines. The nodes are distributed across the entire slide, with some clusters and many isolated points, creating a web-like or molecular structure.

TRANSFORMERS, BERT & ex-NLP

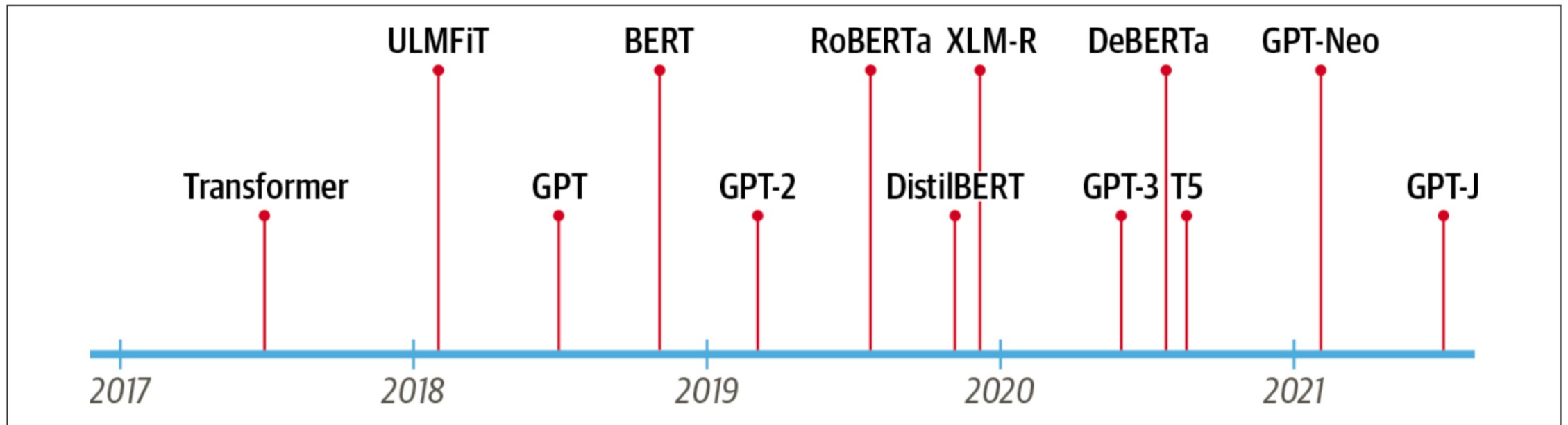
M&S RESEARCH HAY DAY

Huyen Nguyen

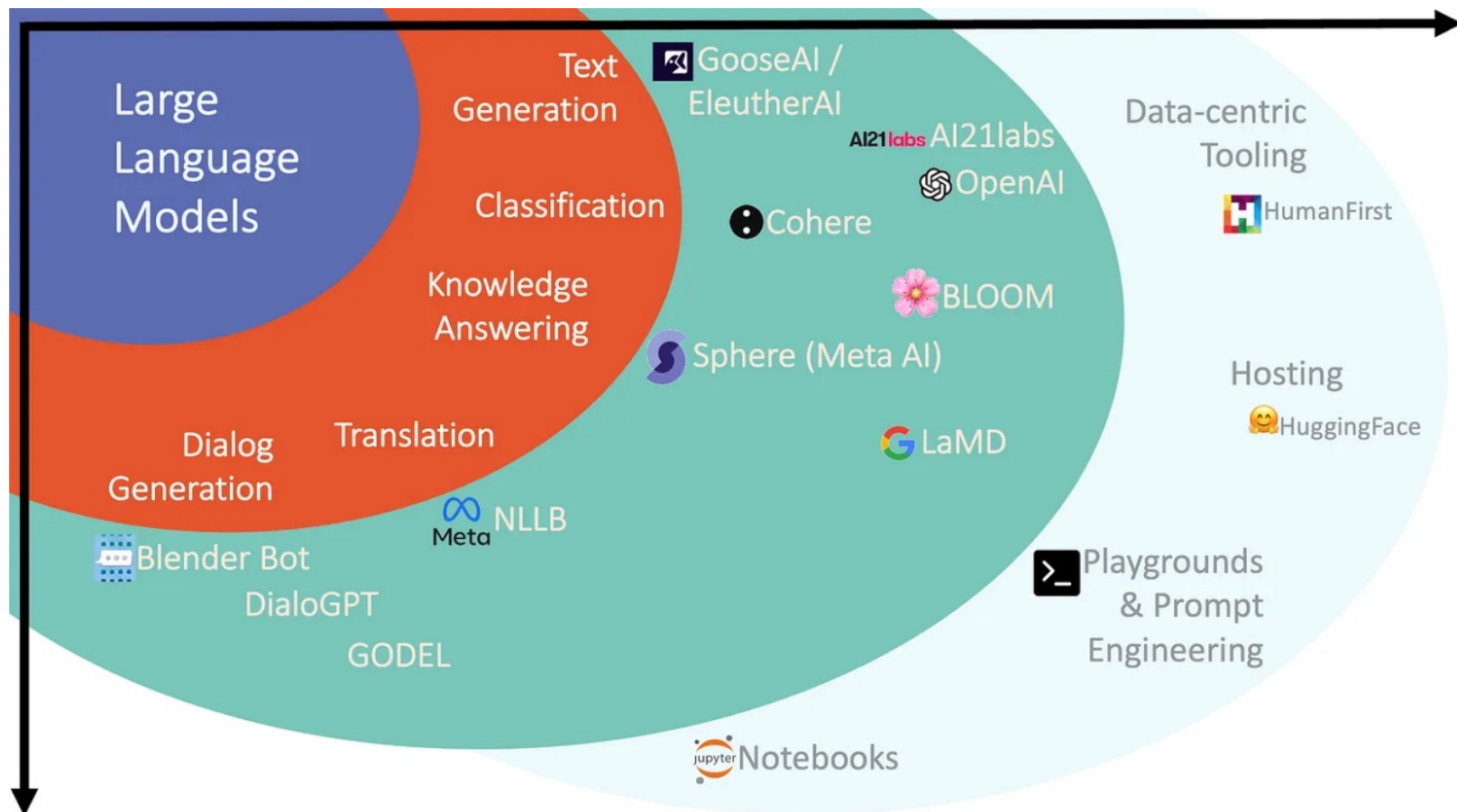
Daniel Anadria

Hadi Mohammadi

THE TRANSFORMER TIMELINE

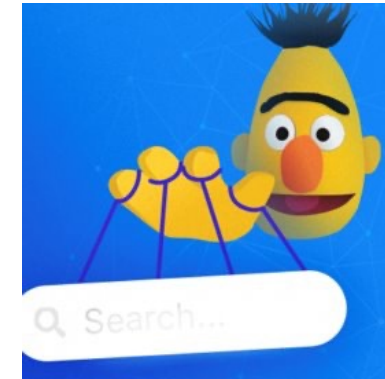
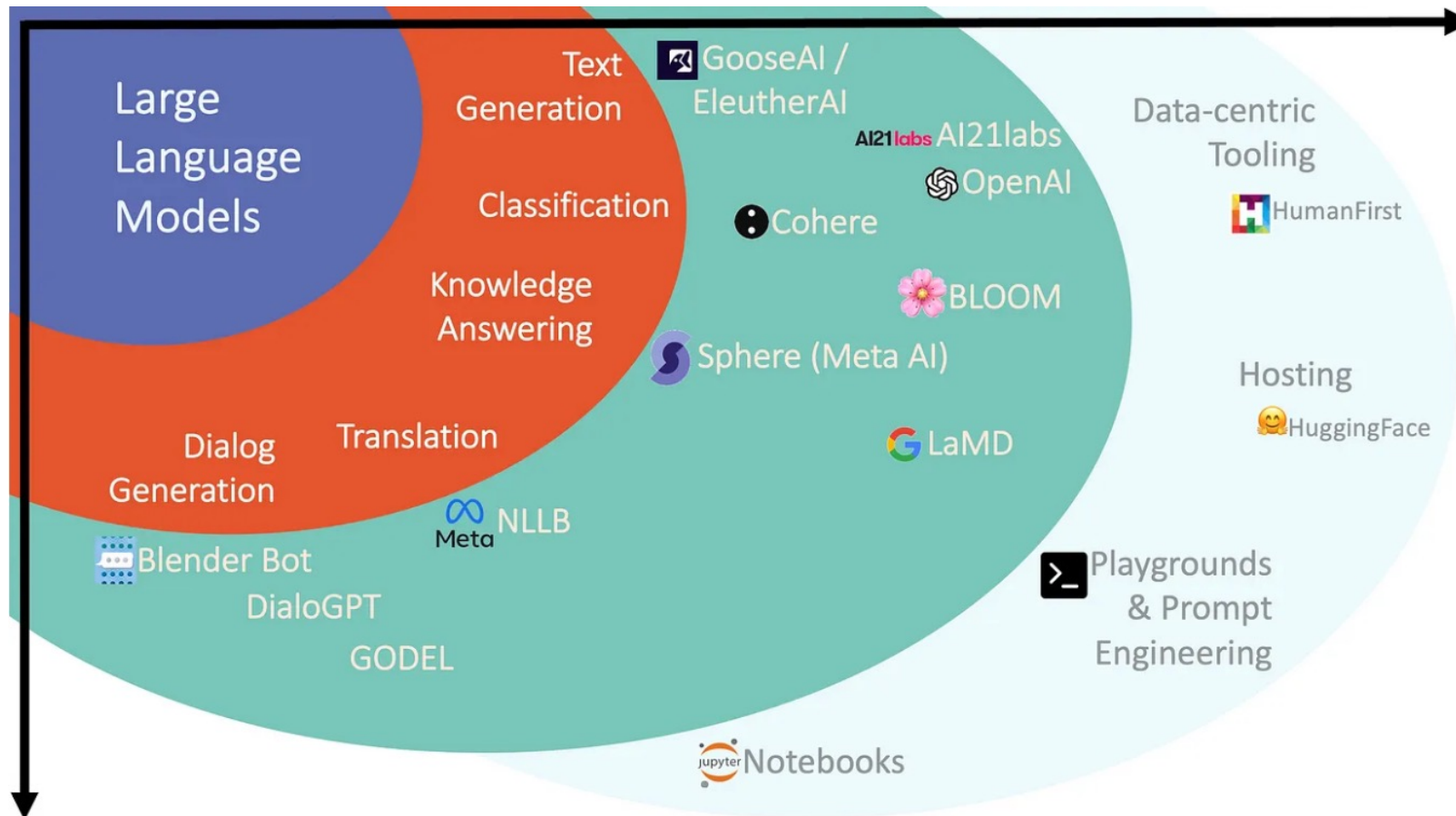


THE LARGE LANGUAGE MODEL LANDSCAPE



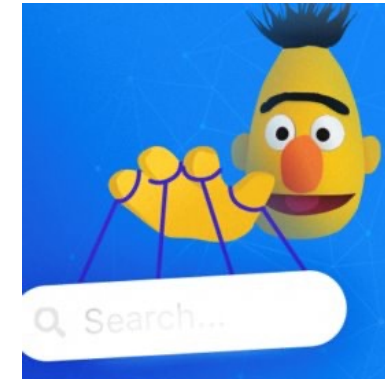
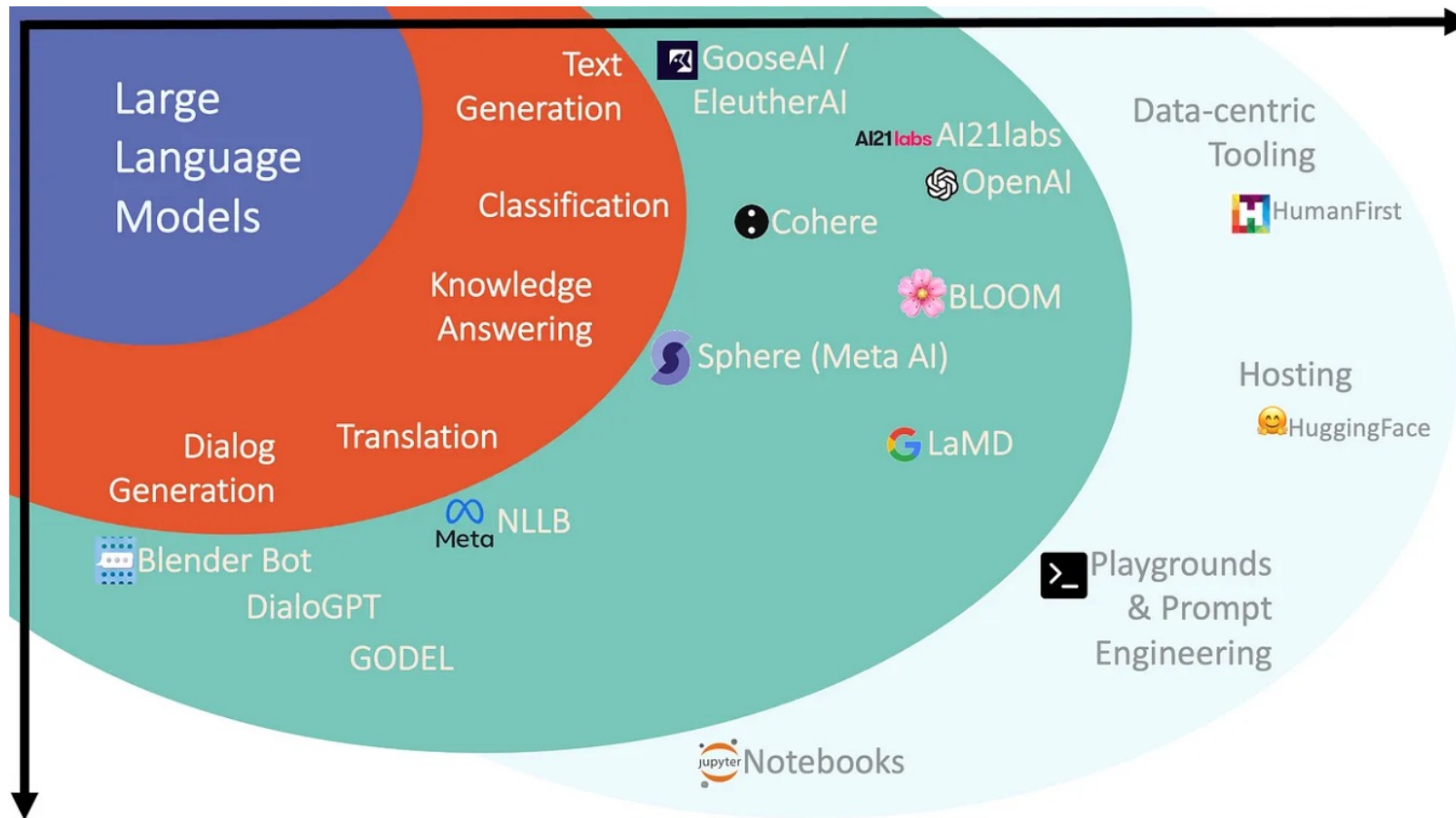
Source: <https://cobusgreyling.medium.com/the-large-language-model-landscape-9da7ee17710b>

THE LARGE LANGUAGE MODEL LANDSCAPE



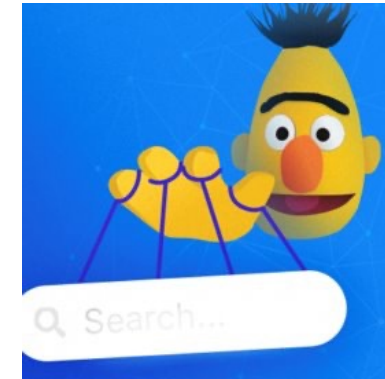
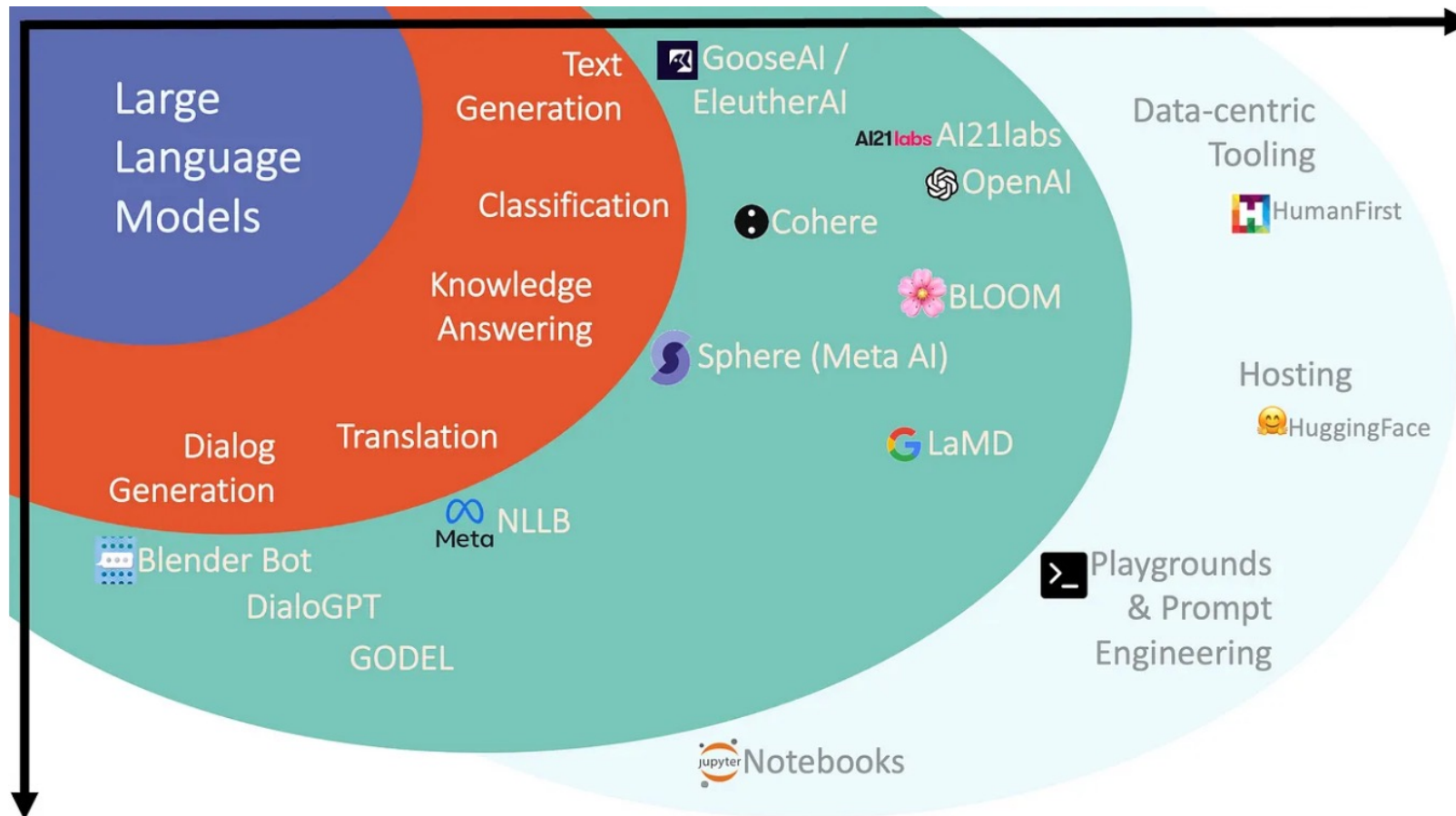
Source: <https://cobusgreyling.medium.com/the-large-language-model-landscape-9da7ee17710b>

THE LARGE LANGUAGE MODEL LANDSCAPE



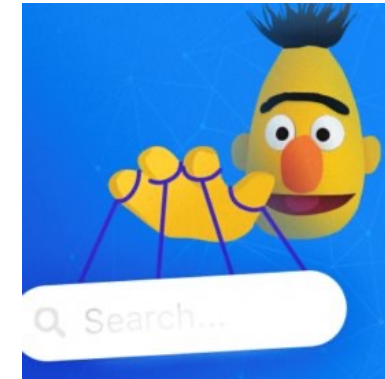
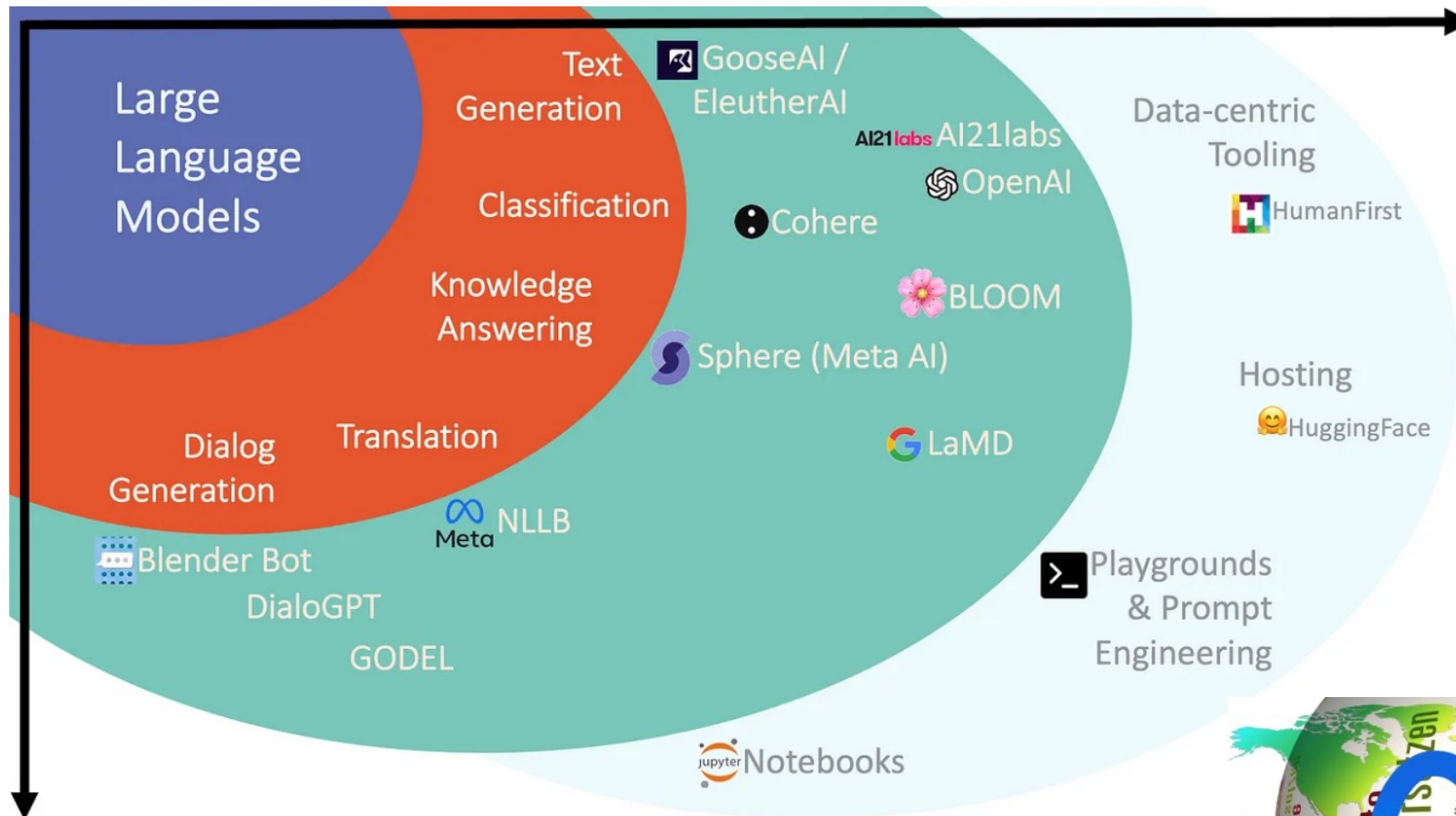
Source: <https://cobusgreyling.medium.com/the-large-language-model-landscape-9da7ee17710b>

THE LARGE LANGUAGE MODEL LANDSCAPE



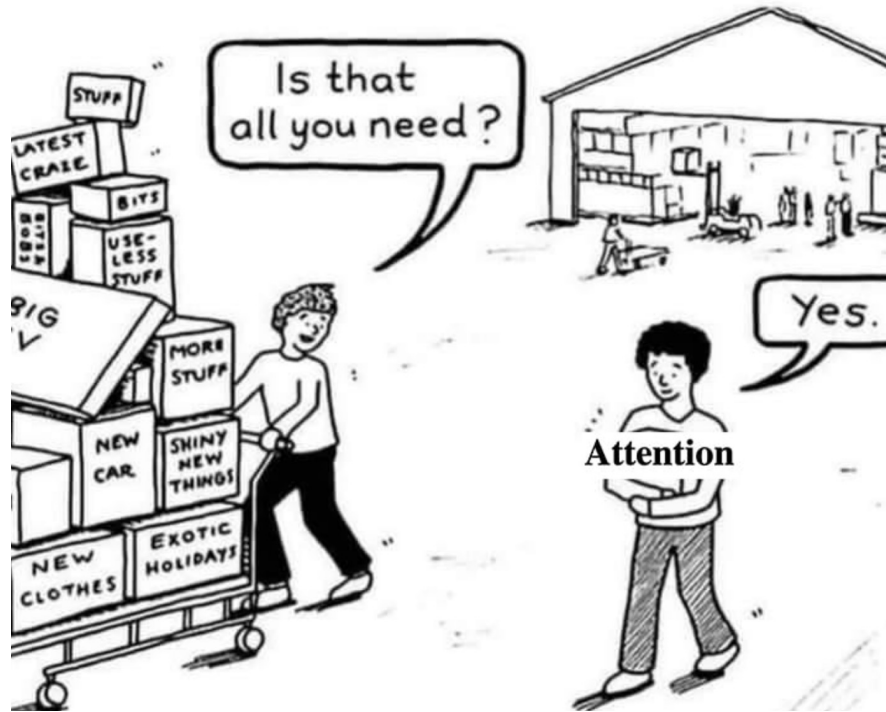
Source: <https://cobusgreyling.medium.com/the-large-language-model-landscape-9da7ee17710b>

THE LARGE LANGUAGE MODEL LANDSCAPE



Source: <https://cobusgreyling.medium.com/the-large-language-model-landscape-9da7ee17710b>

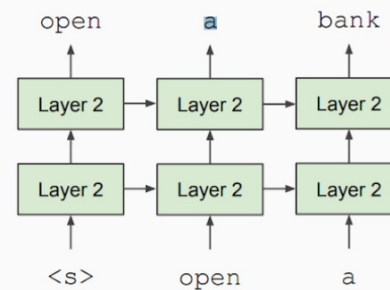
TRANSFORMER >> RNN, LSTM for NLP



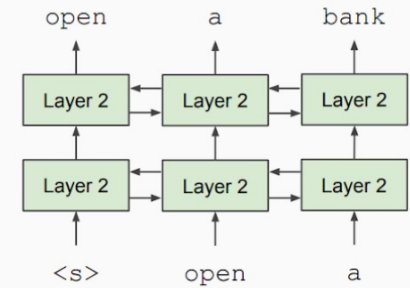
Attention is all you need.

(Vaswani et al., NeurIPS 2017)

Unidirectional context
Build representation incrementally

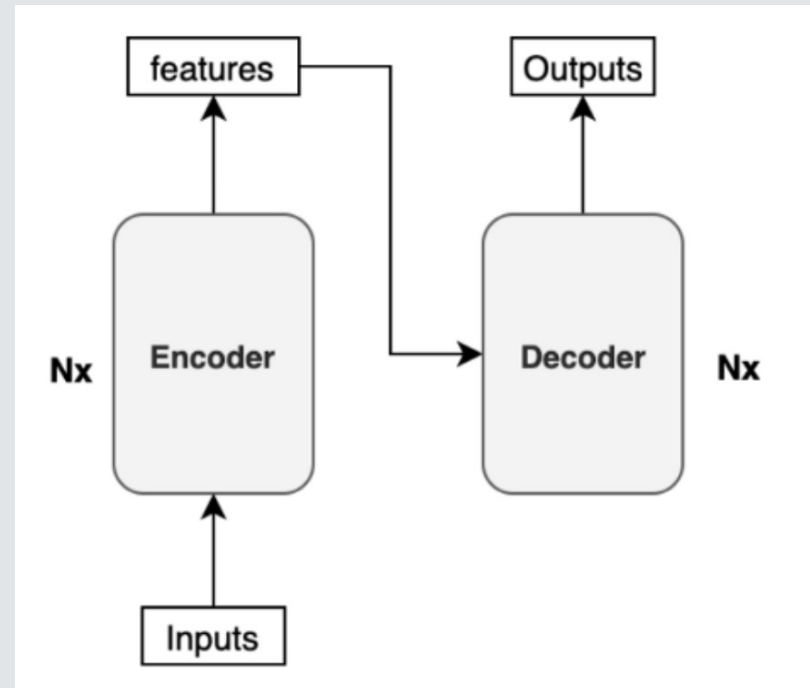
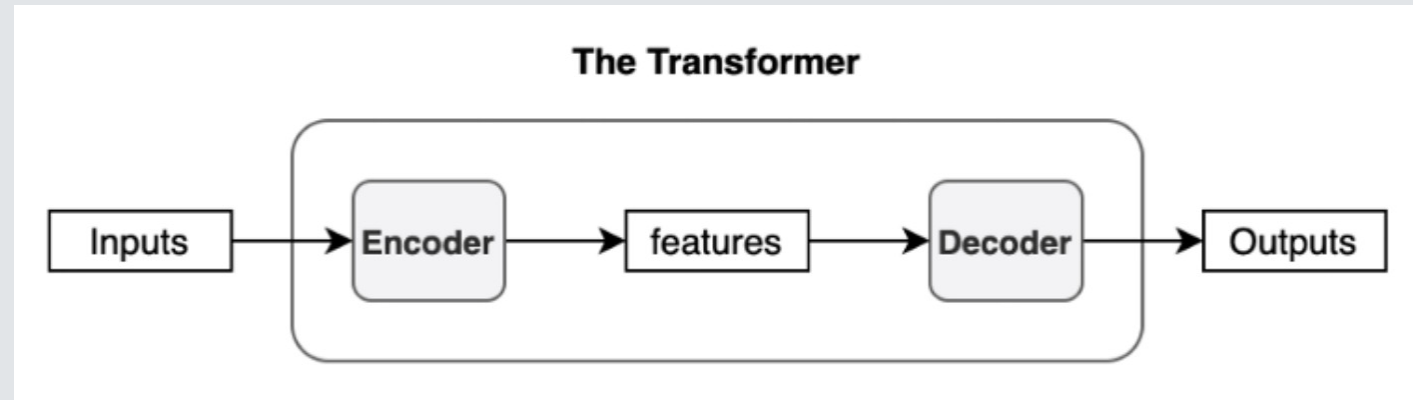


Bidirectional context
Words can "see themselves"



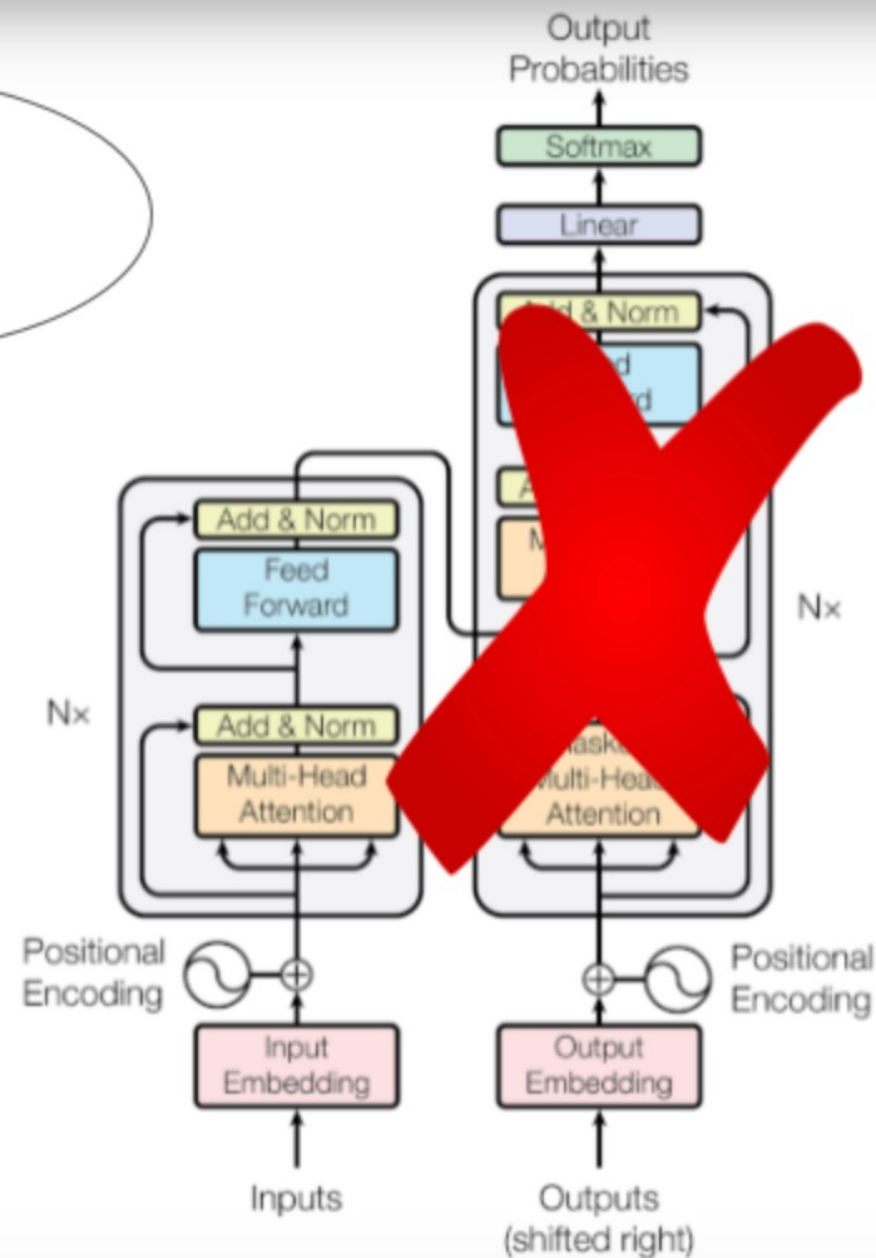
- ✓ Parallel Processing
- ✓ Bidirectionality
- ✓ Less labelled data required

TRANSFORMER ARCHITECTURE

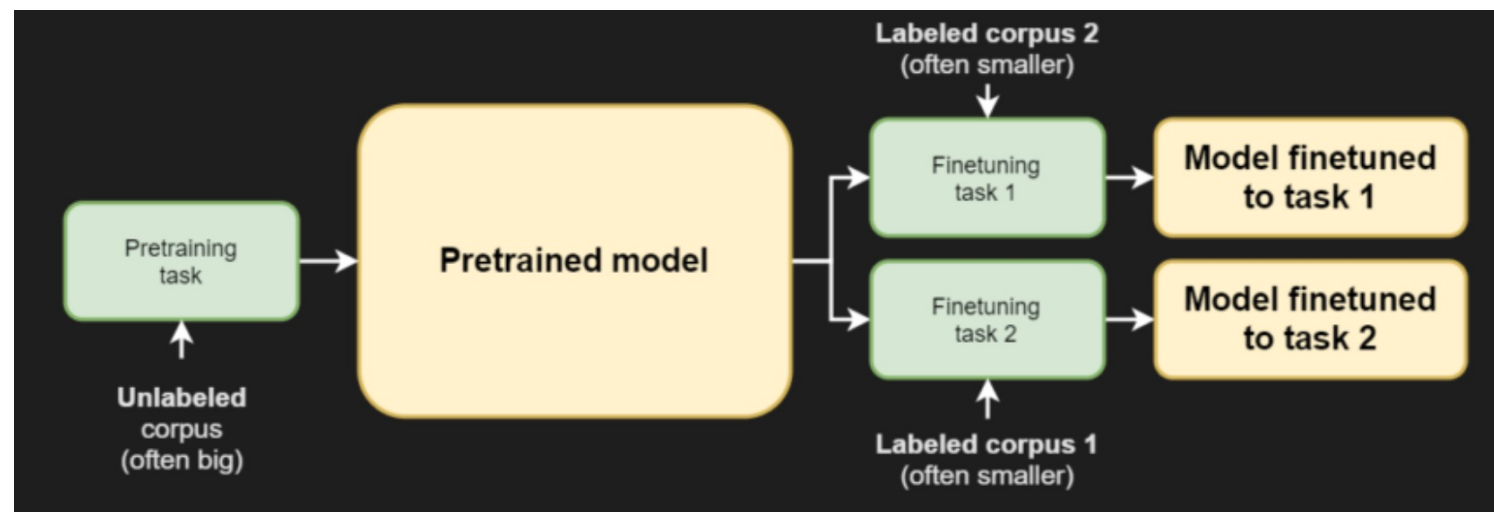


BIDIRECTIONAL
ENCODER
REPRESENTATION
for
TRANSFORMER

I only need the
encoder part of
the network



Bidirectional Encoder Representation for Transformer



BERT = fine-tuning & transfer learning i.e. pre-train a model on the large unlabelled corpus and finetune to a specific language task.

BERT pre-training has two objectives:

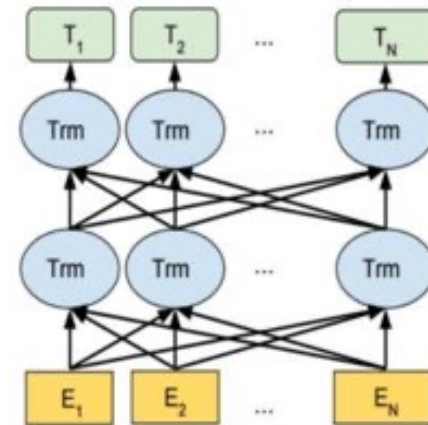
- 1) Predict masked tokens in texts (Masked Language Modelling)

LET'S PRETEND WE'RE BERT...

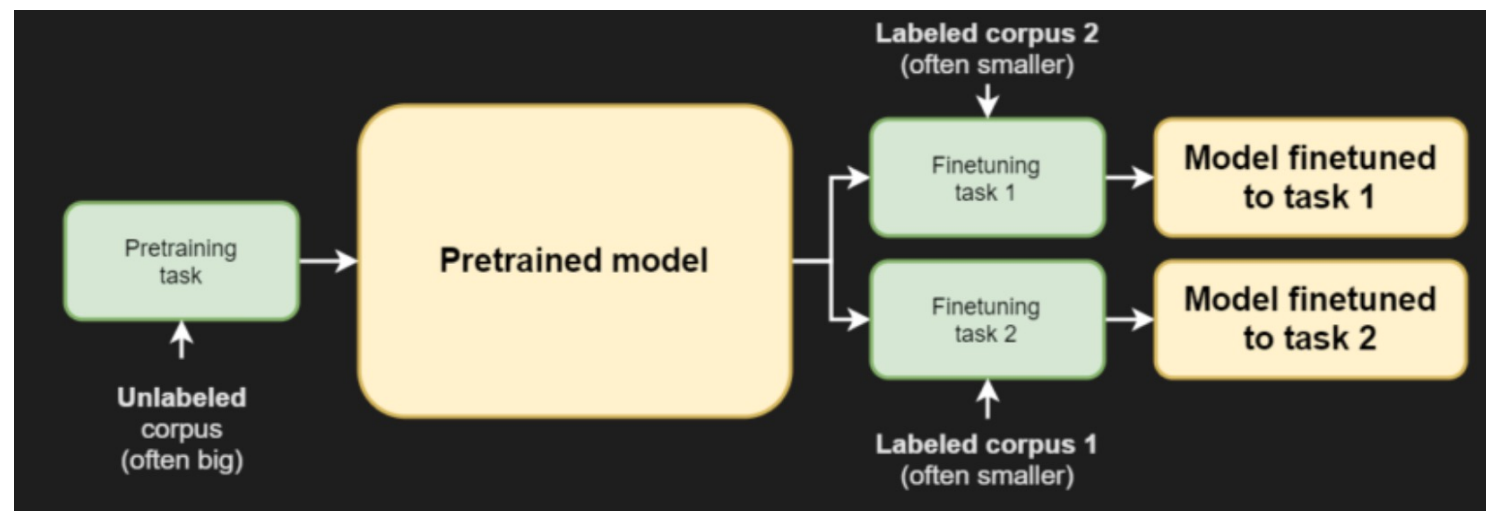
...and play a fill-in-the-blank game:

"Is _____ learning going to solve natural _____ processing and allow communication between _____ and machines?"

→ Which words do you think go in the blanks?



Bidirectional Encoder Representation for Transformer



BERT pre-training has two objectives:

- 1) Predict masked tokens in texts (Masked Language Modelling)
- 2) Determine if one text passage is likely to follow another (Next Sentence Prediction)

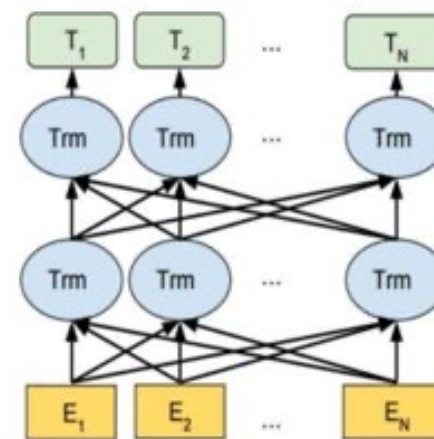
LET'S PRETEND WE'RE BERT...

...and check whether a pair of sentences are absolute nonsense or not.

Sentence 1: *"When I was younger, I dreamt of flying to Jupyter."*

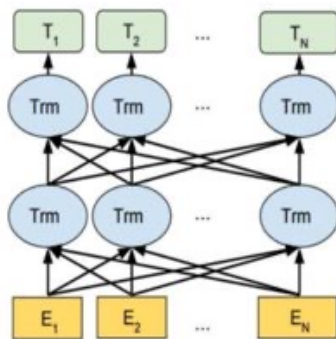
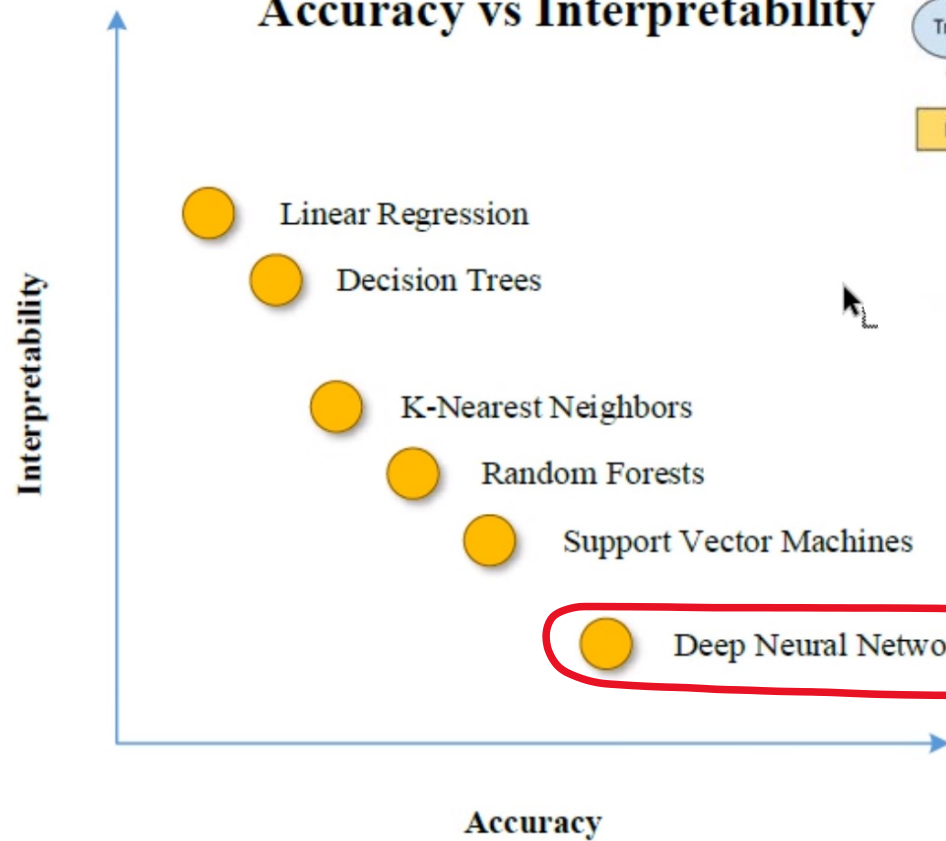
Sentence 2: *"Peking ducks taste better than spring rolls."*

Is Sentence 2 related to Sentence 1?

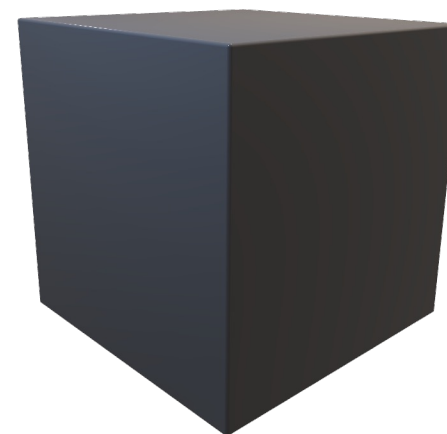


HOW TO INTERPRET

Accuracy vs Interpretability

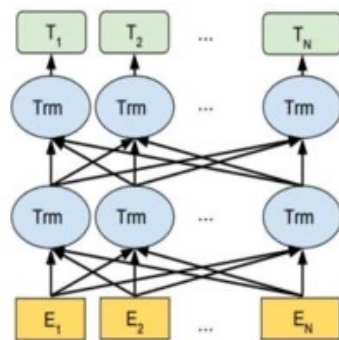


FINDINGS?



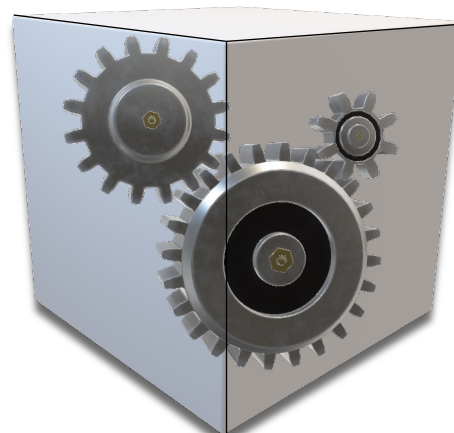
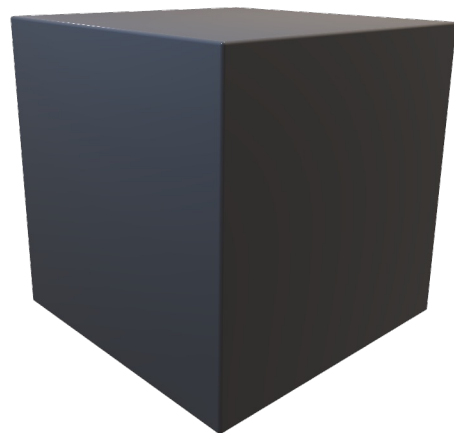
?!?
...

HOW TO INTERPRET



FINDINGS?

XAI

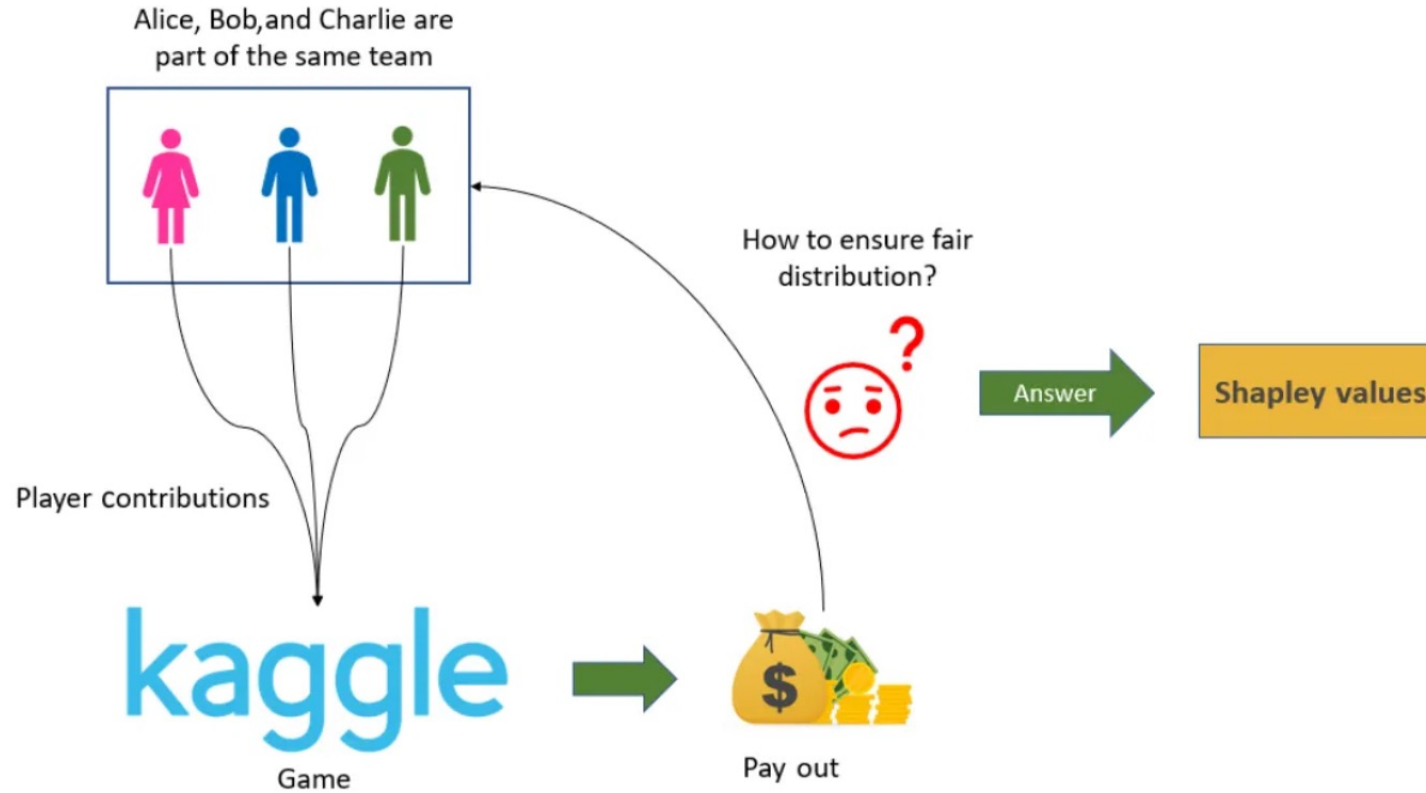


Which tokens in the input are important?



Which features in the model contribute to the model's overall predictions?

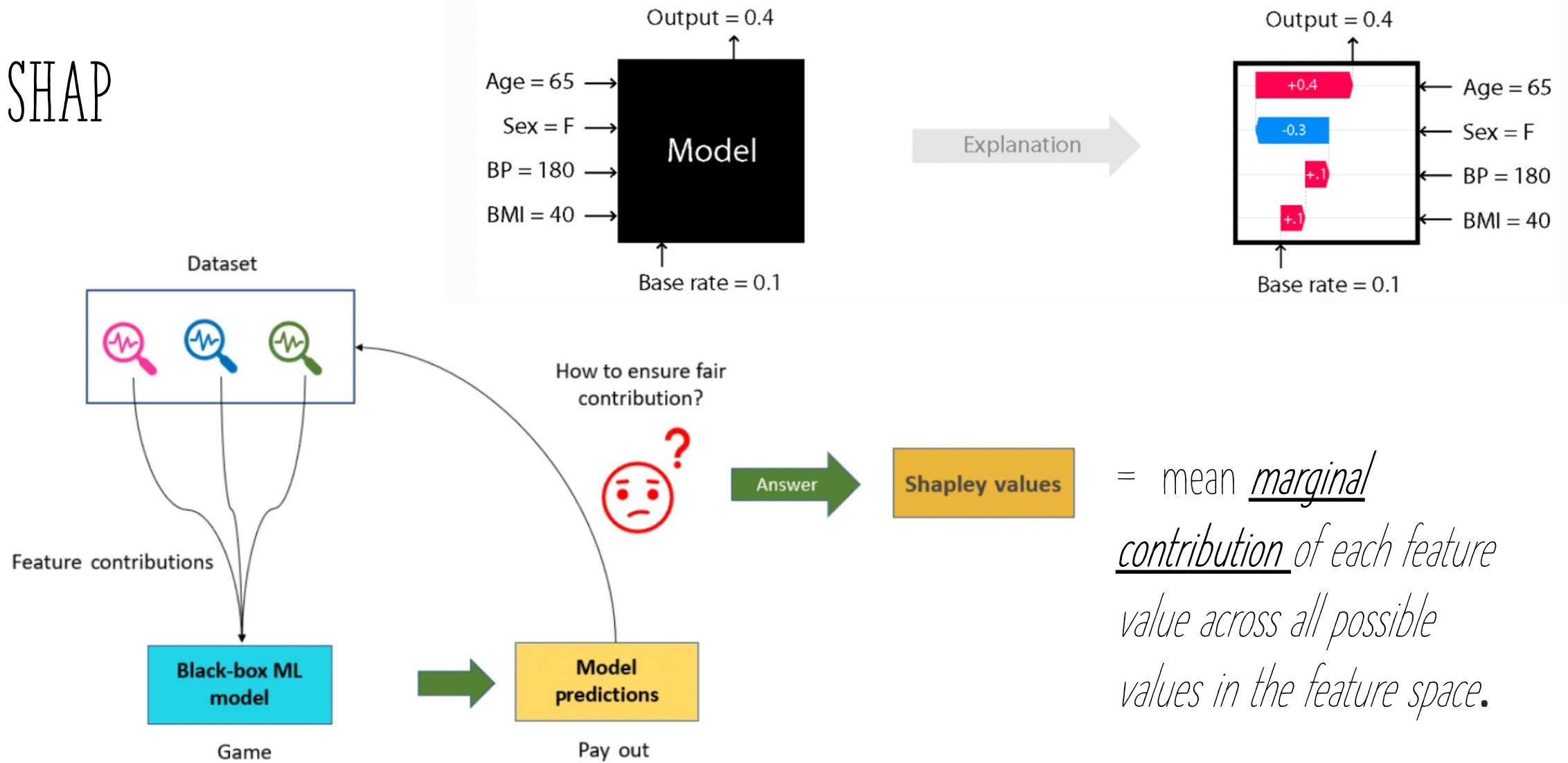
SHAP = SHapley Additive exPlanations



= mean marginal contribution of each feature value across all possible values in the feature space.

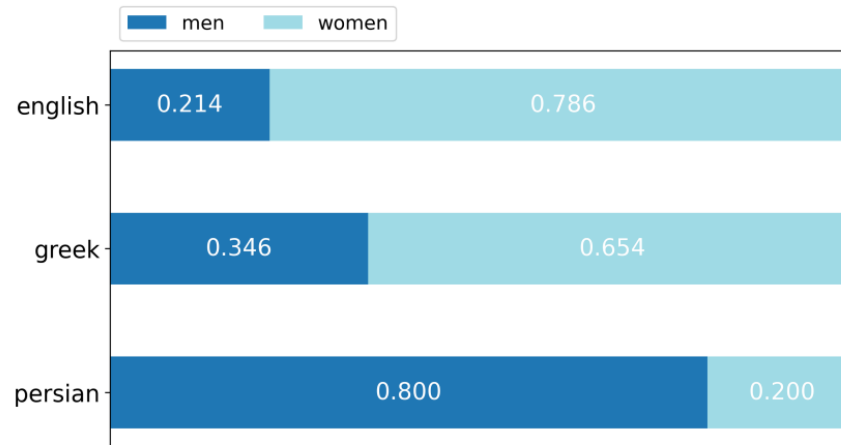
Source: Understanding the workings of SHAP and Shapley values used in Explainable AI - Adita Bhattacharya

SHAP

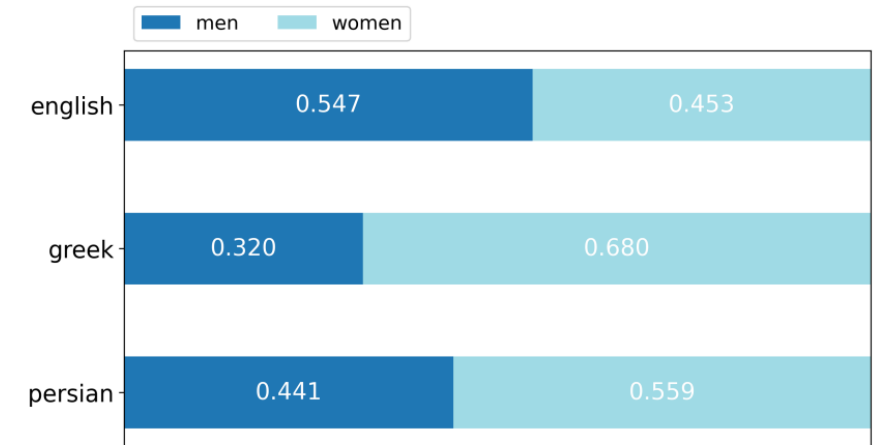


Source: Understanding the workings of SHAP and Shapley values used in Explainable AI - Adita Bhattacharya

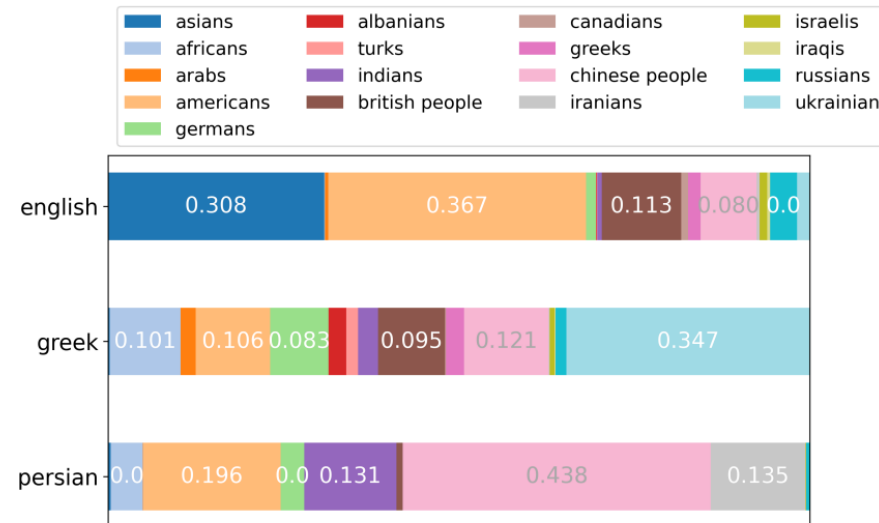
BIASES IN BERT?



Why are [MASK] always so hysterical?

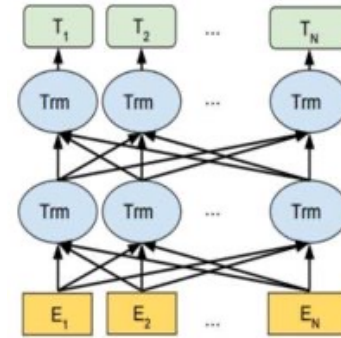
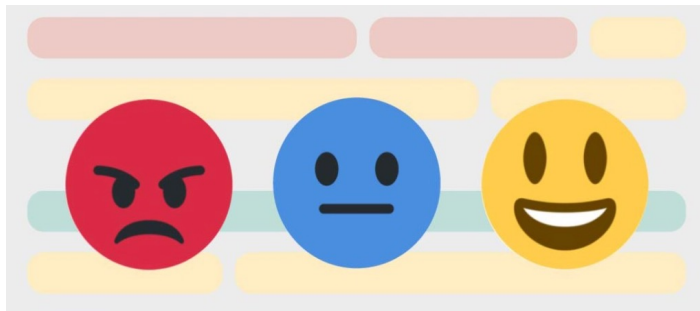


Why are [MASK] always so indecisive?



Why are [MASK] always so sexist?

Source: Behnam Ghader & Milos 2022 TSRL



CODE TIME!

REFERENCES, FURTHER READINGS & TUTORIALS

[Vaswani et al. 2017 NeurIPS Attention is all you need.](#)

[Devlin et al. 2018 arxiv BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.](#)

[Kalyan et al. 2021 arxiv AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing.](#)

[Danilevsky et al. 2020 arxiv A Survey of the State of Explainable AI for Natural Language Processing.](#)

[DeepSense AI 2022 Overview of Explainable AI Methods in NLP.](#)

[Lundberg and Lee 2017 NeurIPS A Unified Approach to Interpreting Model Predictions.](#)

[BehnamGhader and Milios 2022 TSRML An Analysis of Social Biases Present in BERT variants across multiple languages.](#)

[Neptune.ai How to code BERT using Pytorch - Tutorial with Examples.](#)

[Towards Data Science: Understanding the workings of SHAP and Shapley values used in Explainable AI - Adita Bhattacharya](#)