USING CAPTUM TO EXPLAIN GENERATIVE LANGUAGE MODELS

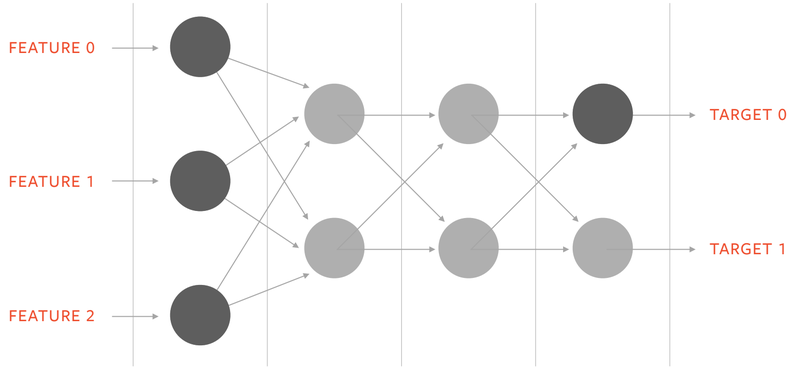
CAPTUM

A unified & generic model **interpretability** library

ATTRIBUTION

Quantify inputs' impact on the output

- e.g., Shapley Value, Integrated Gradients, LIME...



LANGUAGE MODEL ATTRIBUTION

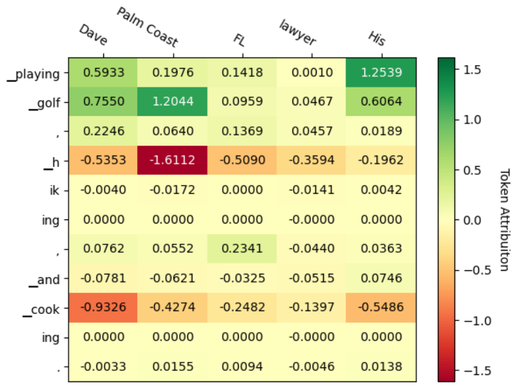
Allow users to

- define the input features in text
- attribute w.r.t the sequential output

MODEL ASSOCIATION

Prompt: Dave is a lawyer living in ***Palm Coast***, FL. His interests include ...

LLM: playing golf , hiking , and cooking



FEW-SHOT LEARNING

Prompt: Examples of movie review classification:

"Movie was ok, the actors weren't great" -> Negative	-0.0413
"Love it, it was an amazing story!" -> Positive	-0.2751
"Total waste of time!!" -> Negative	-0.0399

Classify the following review:

"I really liked the Avengers , it had a captivating plot!"

LLM: Positive

Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano, Narine Kokhlikyan
Meta AI