



Fondazione  
**CARIPLO**



**Bocconi**



# ferret

## A Framework for **Benchmarking Explainers** on Transformers

Giuseppe Attanasio, Eliana Pastor, Chiara Di Bonaventura, Debora Nozza

Amazon Alexa AI Workshop, March 23, 2023



# Nice to meet you!

- Postdoc @ MilaNLP, Bocconi, Milano
- NLP, Speech, and Vision-Language Multimodality
  - Evaluating and Interpreting Transformers
  - Studying Social Biases in large-scale models
- Ph.D. with Elena

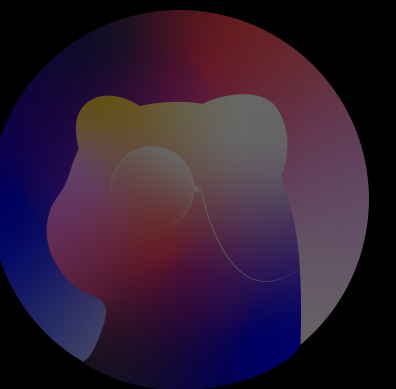
*giuseppe.attanasio3@unibocconi.it*  
<https://gattanasio.cc>





# You are a smart woman

Why is this a compliment?



# You are a smart woman

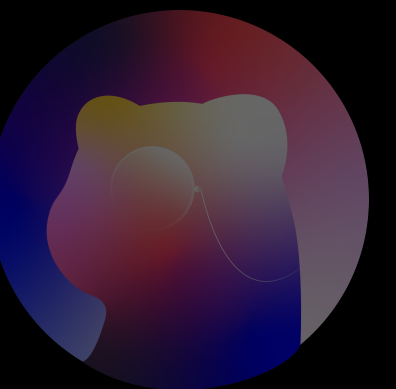


True: compliment



Predicted: non-compliment

*Your Transformer LM*





# You are a smart woman

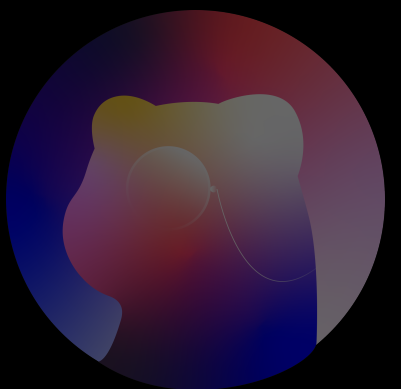


Predicted: non-compliment

Explainer

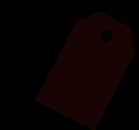
Darker color is  
stronger importance

- E1 You are a smart woman
- E2 You are a smart woman
- E3 You are a smart woman
- E4 You are a smart woman





You are a smart woman



Predicted: misogynous

Explainer

Darker color is  
stronger importance

## Explainers can disagree

E1

You are a smart woman

E2

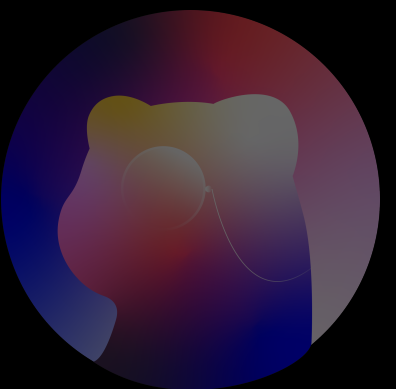
You are a smart woman

E3

You are a smart woman

E4

You are a smart woman





# Why ferret?

## Many XAI Methods

Which one is best?

How do they compare?



Existing XAI libs  
are hard to integrate

A Widespread   
Transformer Library



## Post-Hoc Explainers

Gradients

Integrated Gradients

SHAP

LIME

(DIG, SOC, Contrastive)

## Faithfulness & Plausibility

Comprehensiveness

Sufficiency

LOO Correlation

IOU

AUPRC

Token F1

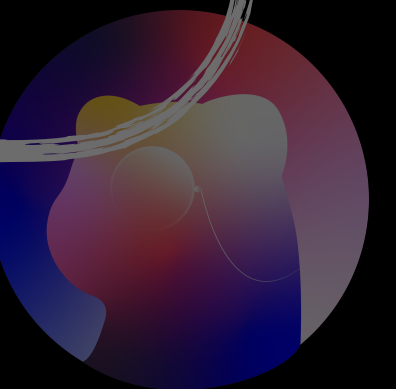
## XAI Datasets

HateXPlain

MovieReviews

SST

Thermostat





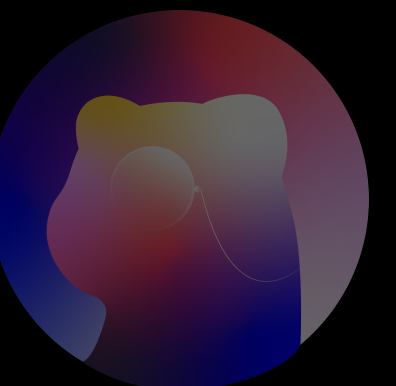
# Faithfulness

- Adhere with the model's inner working
- We often mask-then-reassess
  - E.g., **comprehensiveness**: removing all salient tokens will produce a null score to the class

Turn the **lights on** please

 IOT\_HUE\_LIGHTON

Salient



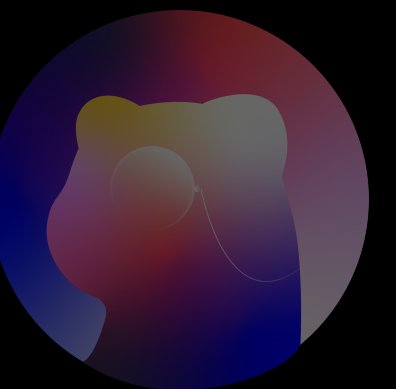


# Plausibility

- The explanation matches human rationales

Cancel my seven am alarm

 ALARM\_CANCEL



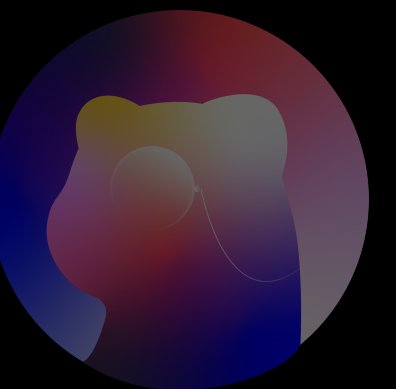


# Plausibility

- The explanation matches human rationales

Cancel my seven am alarm

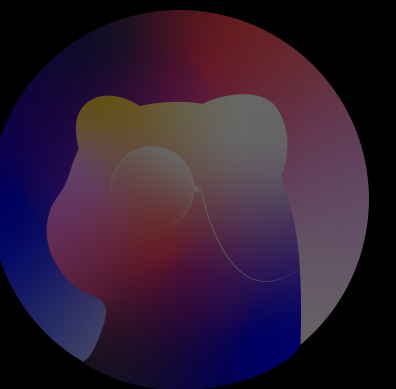
 ALARM\_CANCEL





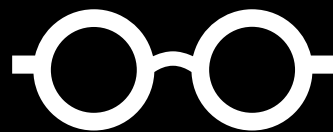


Demo





# What *ferret* can do for you

- Now...
  - Watching the watchers 
- ... soon, more
  - **tasks** (QA, NLI, Language Modeling, Zero-Shot Classification, ...),  
**domains** (Vision, V&L),  
**explainers** (supporting generative models),  
**evaluation** metrics





# What *you* can do for ferret

- We are in very early stages
  - Trying the library & Giving feedbacks
  - Contribute 🚀
  - Spread the word!
- ★ <https://github.com/g8a9/ferret>

Thanks!

