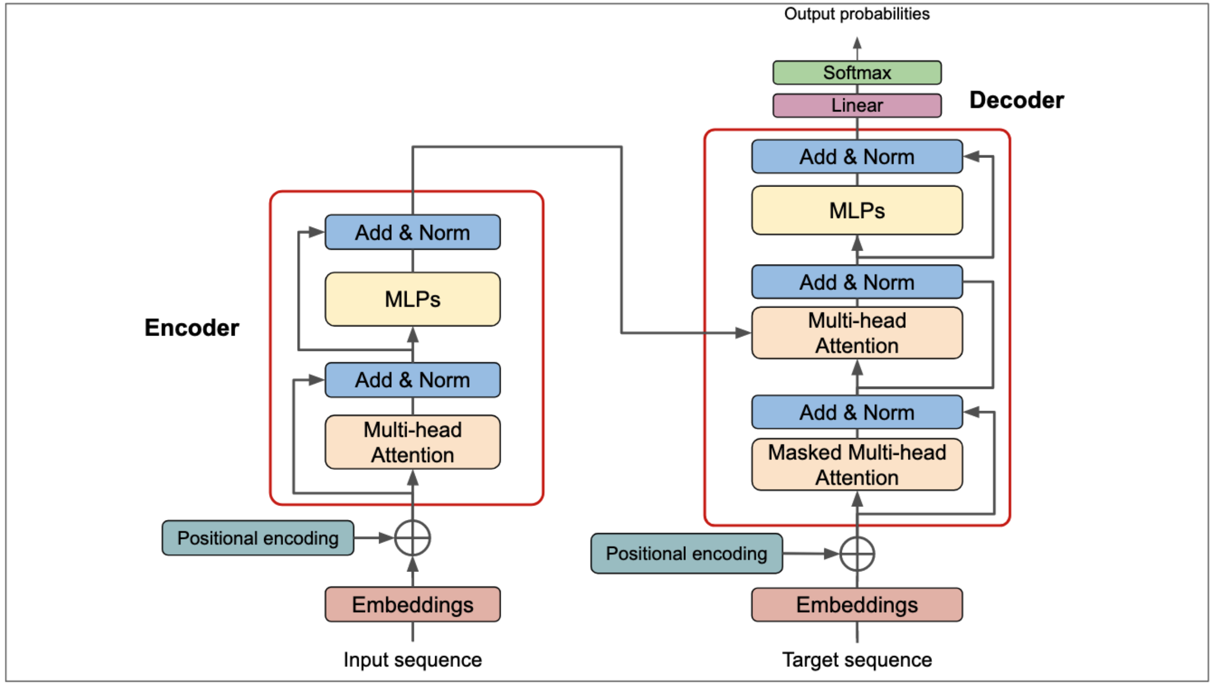
Large language model

PROGRESS REPORT
18/11/2024

LLM models

ENCODER BASED	DECODER BASED
BERT(RoBERTa, DistillBERT…), T5	GPT(GPT-4…), LLaMA
Sentiment analysis / Sentence classification	Chatbots / Text generation / Summaries



LLM models

- Input embedding

$X_i = P(T_i) + e(T_i)$, for token T_i

◦ *e is an embedding matrix learned by the model*

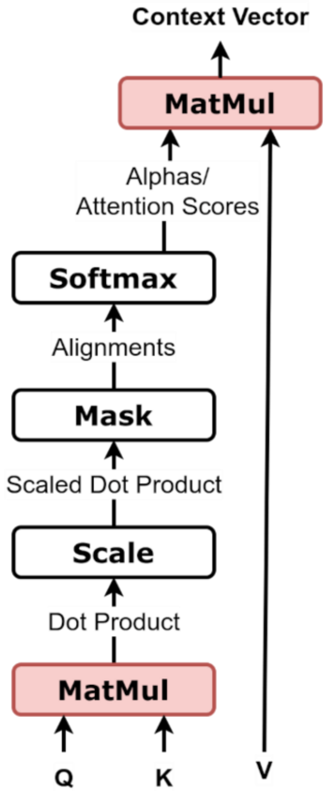
◦ *P is the positional encoding to keep track of order of tokens*

$Then, X = (X_i)_{i=1}^{n}$

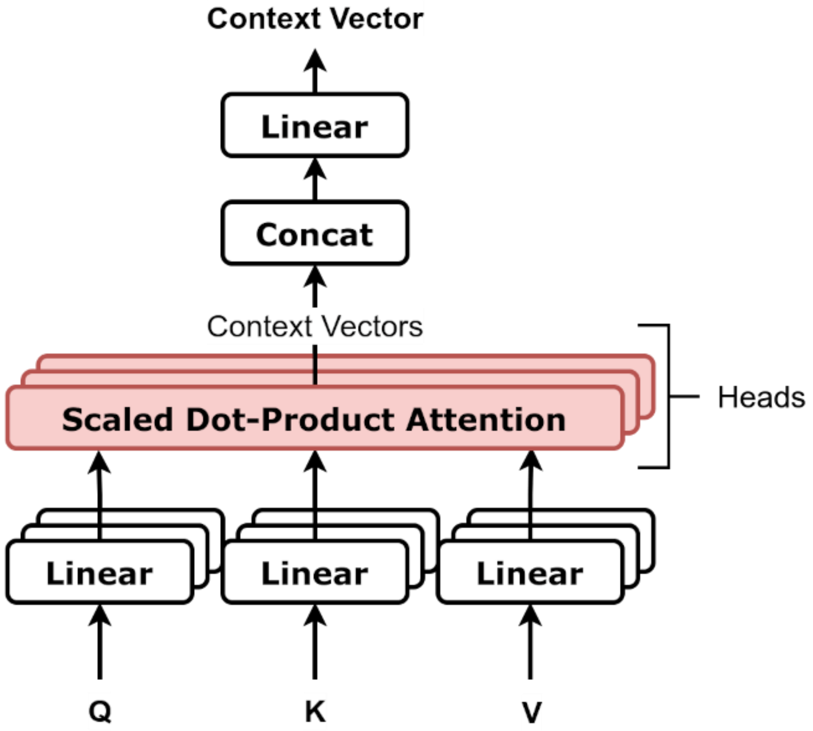
- Attention mechanism

$Q = XW_Q, K = XW_K \; V = XW_V$

$Attention(Q, K, V) = softmax\left(\frac{Q}{\sqrt{d}} K^T\right) V$



Scaled Dot-Product Attention



LLM models

- Input embedding:

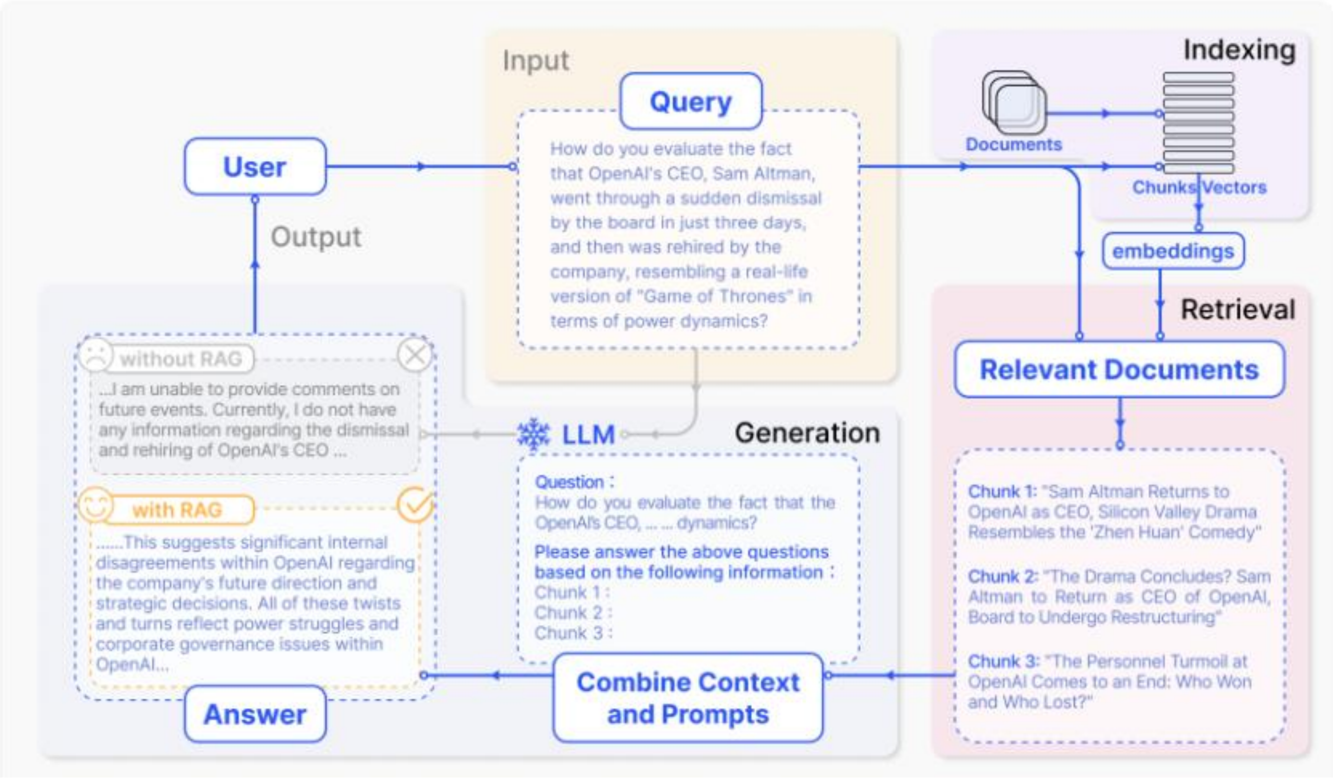
$X_i = P(T_i) + E(T_i)$ where $P(T_i)$ is the positional encoding of token T_i.

- Attention mechanism:

$- Q = \ XW_Q, K = XW_K \ V = XW_V$

$- softmax\left(K^T \frac{Q}{\sqrt{d}}\right) V^T$

RAG models



How do you evaluate the fact that OpenAI's CEO, Sam Altman, went through a sudden dismissal by the board in just three days, and then was rehired by the company, resembling a real-life version of "Game of Thrones" in terms of power dynamics?

...I am unable to provide comments on future events. Currently, I do not have any information regarding the dismissal and rehiring of OpenAI's CEO ...

......This suggests significant internal disagreements within OpenAI regarding the company's future direction and strategic decisions. All of these twists and turns reflect power struggles and corporate governance issues within OpenAI...

Question :
How do you evaluate the fact that the OpenAI's CEO, ... dynamics?
Please answer the above questions based on the following information :
Chunk 1 :
Chunk 2 :
Chunk 3 :

Chunk 1: "Sam Altman Returns to OpenAI as CEO, Silicon Valley Drama Resembles the 'Zhen Huan' Comedy"

Chunk 2: "The Drama Concludes? Sam Altman to Return as CEO of OpenAI, Board to Undergo Restructuring"

Chunk 3: "The Personnel Turmoil at OpenAI Comes to an End: Who Won and Who Lost?"

[https://huggingface.co/blog/hrishioa/retrieval-augmented-generation-1-basics]

Pre-retrieval:
- Query rewriting

Retrieval:
- Document chunker
- Embedding
- Similarity search

Post-retrieval:
- Reranking

Generation:
- Prompt engineer

Embedding models

- Sparse vocabulary representations : TF-IDF… $\mathrm{tf}-\mathrm{idf}_{i,j} = \mathrm{tf}_{i,j} log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$
- ELMo
  - Bidirectional LSTM to add context to the representation
- Transformer-based word embeddings
  - GTP (→), BERT (⇆) …
- Transformer-based sentence embeddings
  - SBERT, USE, GTE …

XAI techniques

1/ Feature attribution

Perturbation based technique

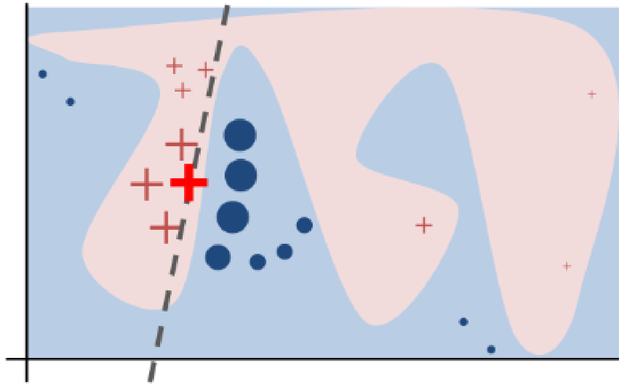
Gradient-based approaches

- Saliency maps
- Integrated gradients

Surrogate model

- Lime (Local Interpretable Model-agnostic Explanations)
- SHAP (SHapley Additive exPlanations)

Attention-based visualization



LIME [Ribeiro et al. 2016]

XAI techniques

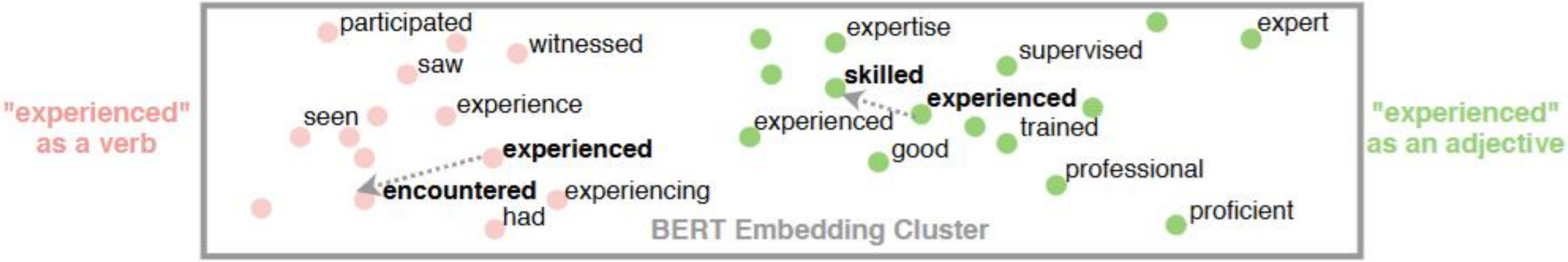
2/ Sample based

Adversarial samples

- input alterations due to small, hard-to-perceive changes for humans that lead to a change in outputs
- e.g. SemAttack

Counterfactual Explanations

seek to identify minimal changes to an input => output changes from a class y to y'



SemAttack [https://arxiv.org/pdf/2205.01287]

XAI techniques

3/ Probing based

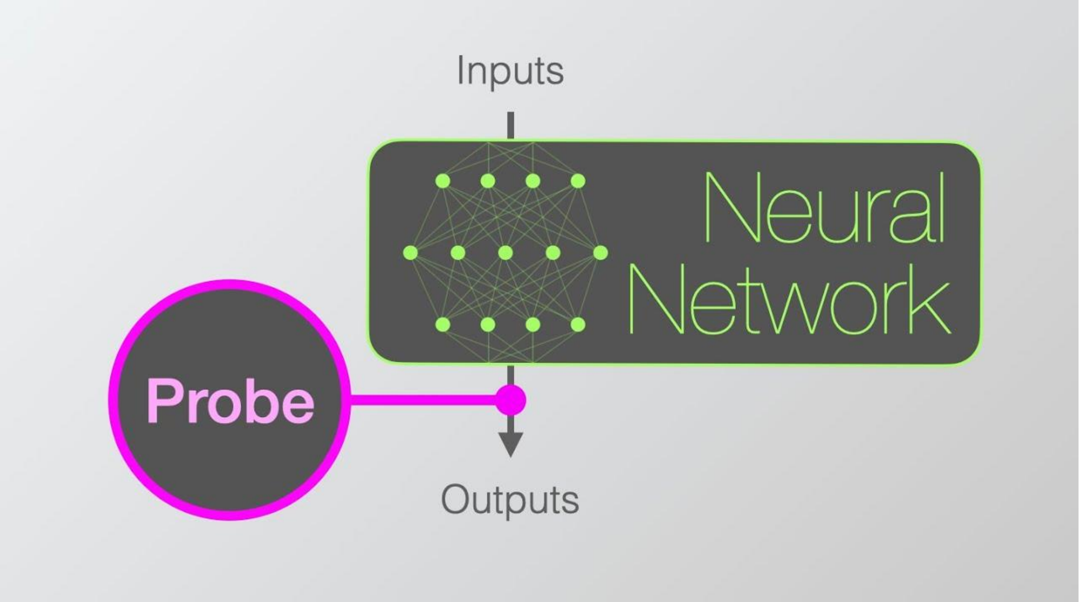
Understand internal representation of the model (information learned and encoded)

Knowledge based

Training classifier based on a layer

Concept based

Neuron activation explanation



XAI techniques

4/ Mechanistic interpretability

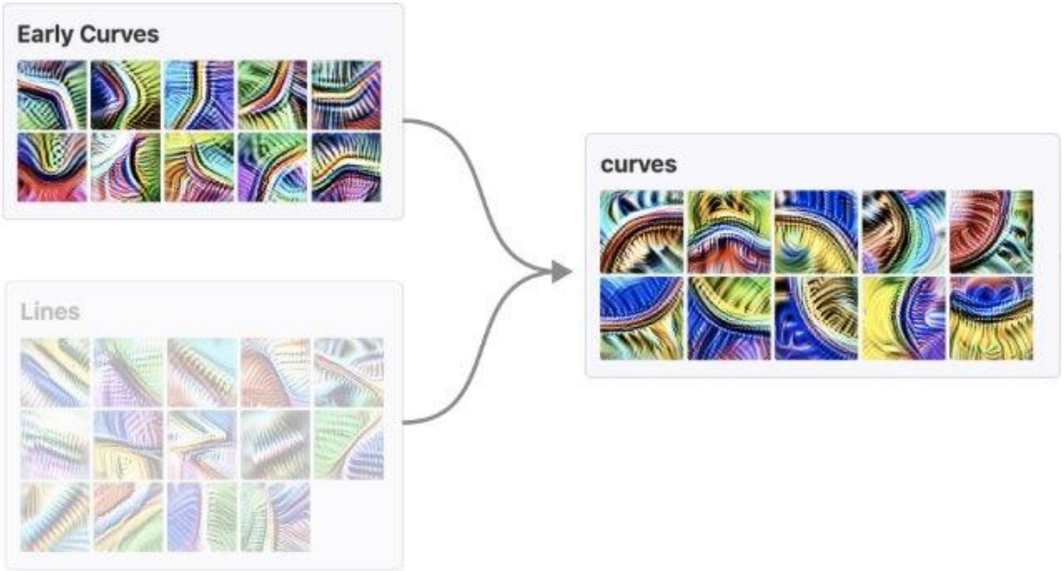
- investigates the causal structure of a model.

- seeks to identify how internal components (e.g., neurons, weights, or attention heads) interact

- Model can be viewed as a graph

Common approaches fall into three categories:

- circuit discovery
- causal tracing
- vocabulary lens



[https://distill.pub/2020/circuits/zoom-in/]

XAI techniques

5/ Structuring based on novel dimensions

It focuses on new perspectives that are not inherently part of the model's original design.

Novel dimensions external to the model's natural operational space (e.g., raw features, embeddings, or output probabilities)

Examples:

In **natural language processing (NLP)**, structuring representations by linguistic properties such as syntax, semantics, or sentiment.

In **computer vision**, structuring filters or layers based on the types of visual features they detect (e.g., edges, textures, objects).

Next steps

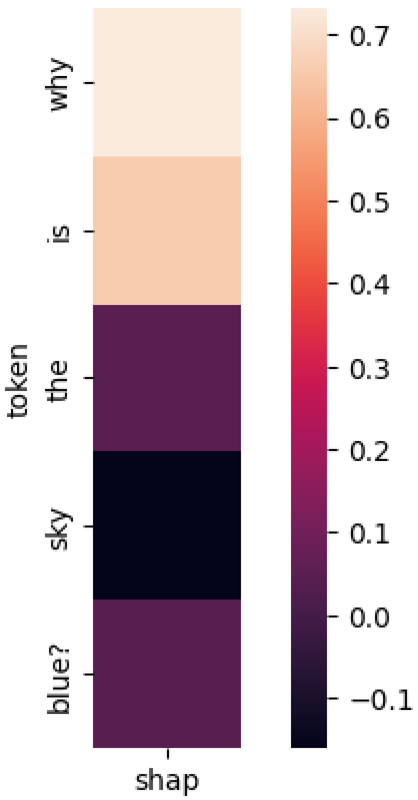
- Fix a context of study
  - Problem ?
  - Goal ?
  - Data ?
  - Model to explain ?
  - Approach ? (simpler to harder methods ..), Any preferences ?

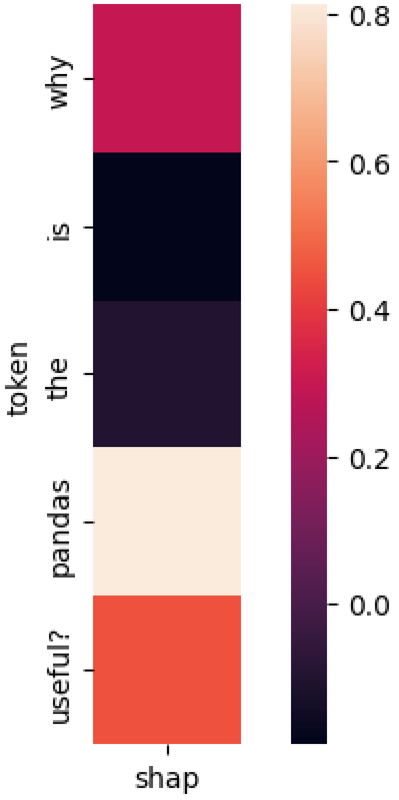
Example: TokenSHAP [arXiv:2407.10114]

- For $tokens = (x_1, \ldots, x_n)$ compute the baseline output b from LLM model

- Compute output for randomly sampled tokens b_C in $tokens$ and compare both methods $\mathrm{v}_C = cosine\_sim(b_C, b)$
    - For each x_i average each v_C in which x_i is and do the same for each v_C in which x_i is not

- $SHAP_i = with_i - without_i$

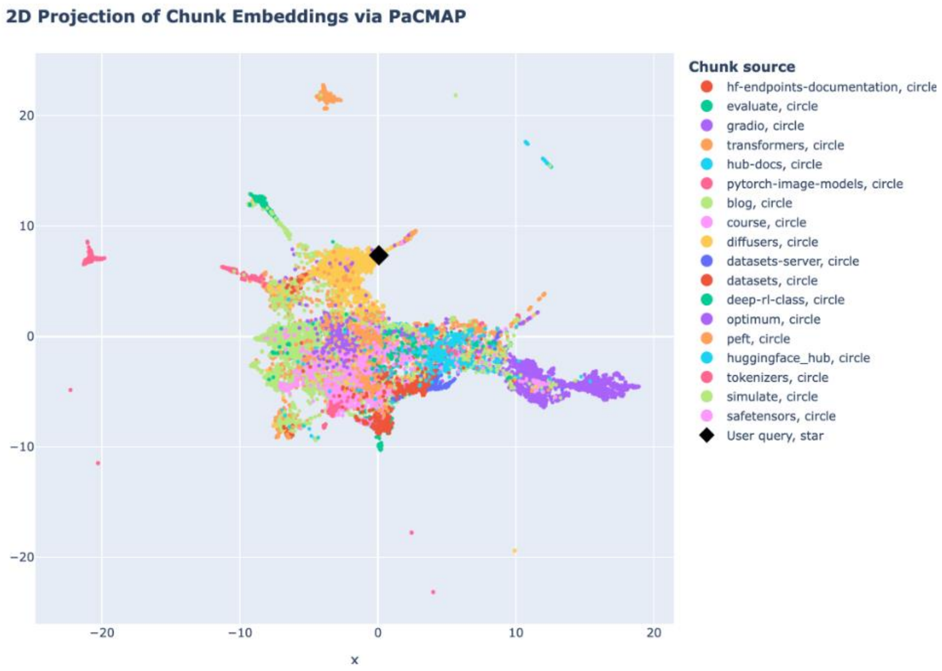




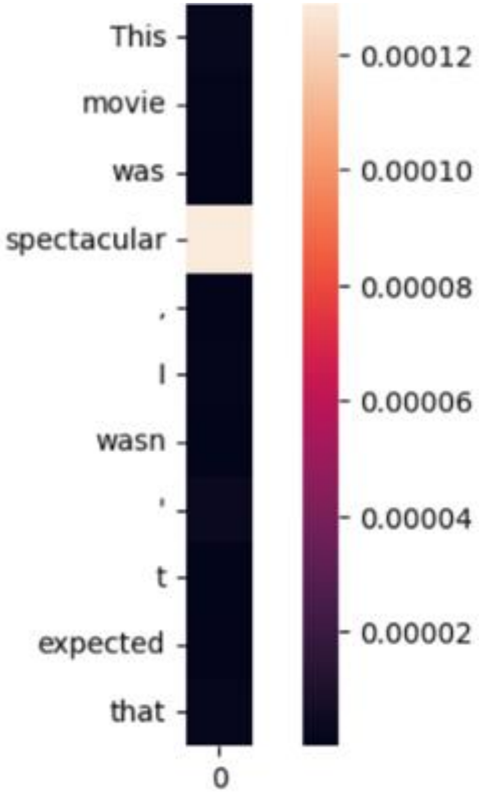
Other XAI techniques

Representation analysis

- UMAP, machine learning embeddings



Classifier SHAP [kokalj-etal-2021-bert]


This
movie
was
spectacular
I
wasn
'
t
expected
that