DIGITAL FINANCE

This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101119635

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

State Secretariat for Education,
Research and Innovation SERI

Funded by
the European Union

MARIE CURIE ACTIONS

DIGITAL

Reading

https://github.com/JShollaj/awesome-llm-interpretability?tab=readme-ov-file

LLMs Interpretability

DIGITAL

Funded by
the European Union

What Does BERT Look At? An Analysis of BERT's Attention

What Does BERT Look At?
An Analysis of BERT's Attention

Kevin Clark† Urvashi Khandelwal† Omer Levy‡ Christopher D. Manning†
†Computer Science Department, Stanford University
‡Facebook AI Research
{kevclark,urvashik,manning}@cs.stanford.edu
omerlevy@fb.com

Abstract

Large pre-trained neural networks such as BERT have had great recent success in NLP, motivating a growing body of research investigating what aspects of language they are able to learn from unlabeled data. Most recent analysis has focused on model outputs (e.g., language model surprisal) or internal vector representations (e.g., probing classifiers). Complementary to these works, we propose methods for analyzing the attention mechanisms of pre-trained models and apply them to BERT. BERT's attention heads exhibit patterns such as attending to delimiter tokens, specific positional offsets, or broadly attending over the whole sentence, with heads in the same layer often exhibiting similar behaviors. We further show that certain attention heads correspond well to linguistic notions of syntax and coreference. For example, we find heads that attend to the direct objects of verbs, determiners of nouns, objects of prepositions, and coreferent mentions with remarkably high accuracy. Lastly, we propose an attention-based probing classifier and use it to further demonstrate that substantial syntactic information is captured in BERT's attention.

1 Introduction

Large pre-trained language models achieve very high accuracy when fine-tuned on supervised tasks (Dai and Le, 2015; Peters et al., 2018; Radford et al., 2018), but it is not fully understood why. The strong results suggest pre-training teaches the models about the structure of language, but what specific linguistic features do they learn?

Recent work has investigated this question by examining the *outputs* of language models on carefully chosen input sentences (Linzen et al., 2016) or examining the internal *vector representations* of the model through methods such as probing classifiers (Adi et al., 2017; Belinkov et al., 2017). Complementary to these approaches, we study[1] the *attention maps* of a pre-trained model. Attention (Bahdanau et al., 2015) has been a highly successful neural network component. It is naturally interpretable because an attention weight has a clear meaning: how much a particular word will be weighted when computing the next representation for the current word. Our analysis focuses on the 144 attention heads in BERT[2] (Devlin et al., 2019), a large pre-trained Transformer (Vaswani et al., 2017) network that has demonstrated excellent performance on many tasks.

We first explore generally how BERT's attention heads behave. We find that there are common patterns in their behavior, such as attending to fixed positional offsets or attending broadly over the whole sentence. A surprisingly large amount of BERT's attention focuses on the deliminator token [SEP], which we argue is used by the model as a sort of no-op. Generally we find that attention heads in the same layer tend to behave similarly.

We next probe each attention head for linguistic phenomena. In particular, we treat each head as a simple no-training-required classifier that, given a word as input, outputs the most-attended-to other word. We then evaluate the ability of the heads to classify various syntactic relations. While no single head performs well at many relations, we find that particular heads correspond remarkably well to particular relations. For example, we find heads that find direct objects of verbs, determiners of nouns, objects of prepositions, and objects of possessive pronouns with >75% accuracy. We perform a similar analysis for coreference resolution, also finding a BERT head that performs quite well. These results are intriguing because the behavior of the attention heads emerges purely from self-supervised training on unlabeled data, without explicit supervision for syntax or coreference.

[1] Code will be released at https://github.com/clarkkev/attention-analysis.
[2] We use the English base-sized model.

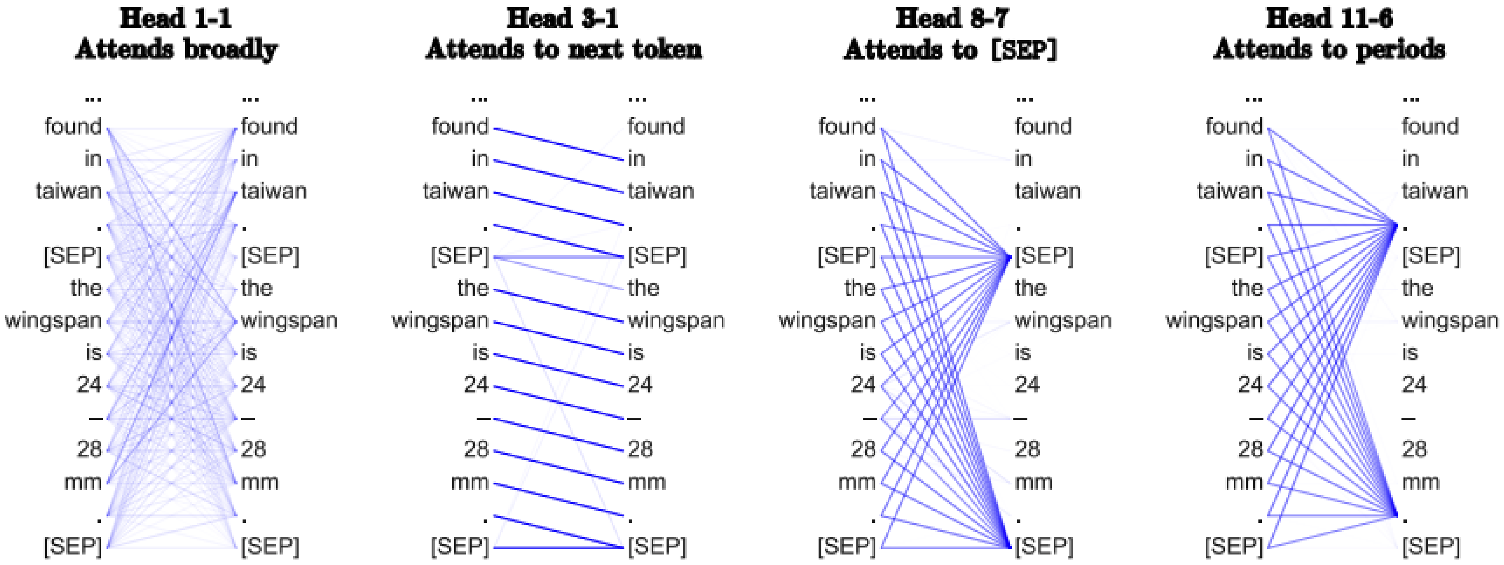
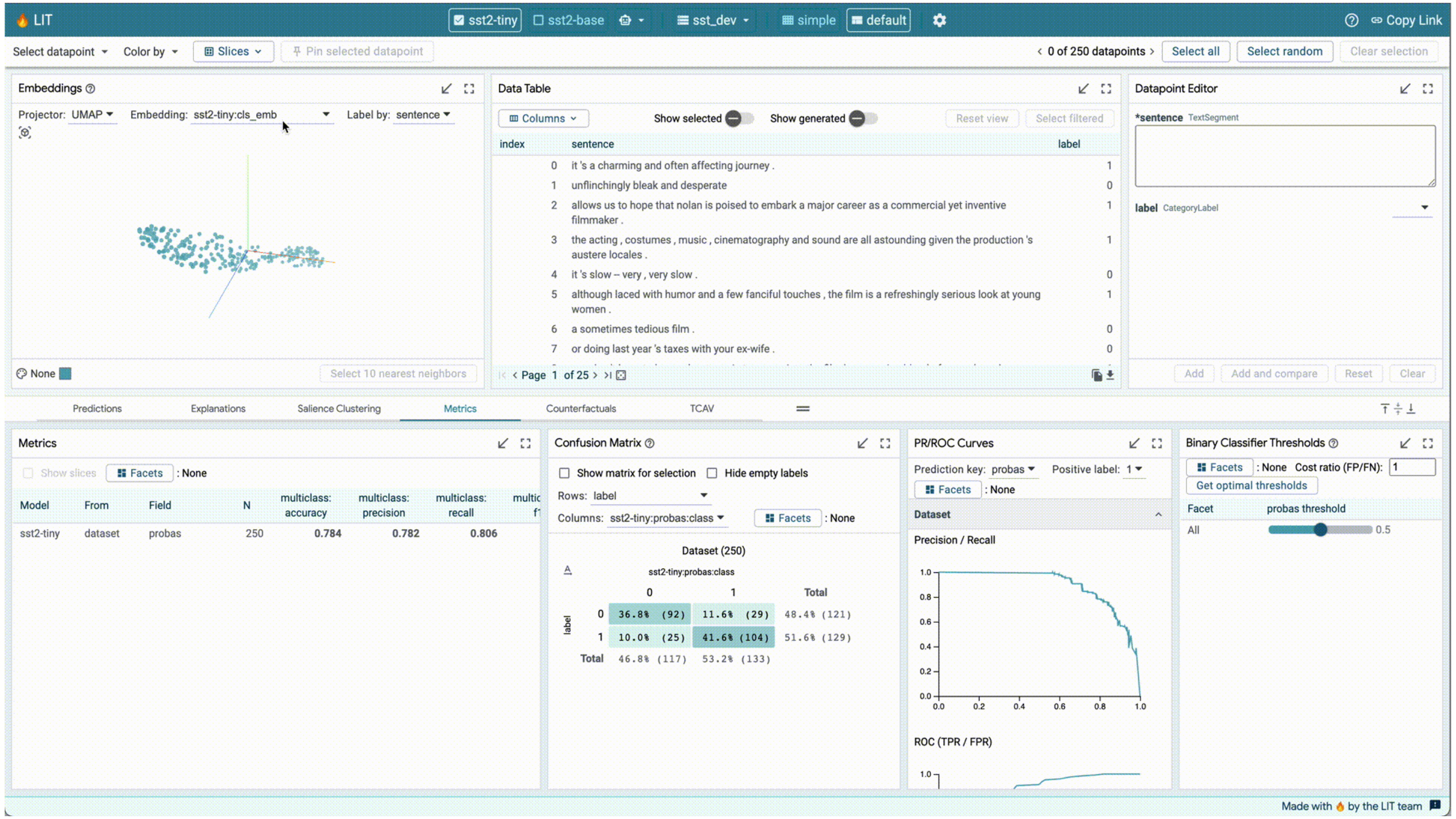


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

The Learning Interpretability Tool (LIT)

- is a **visual, interactive ML model-understanding** tool that supports text, image, and tabular data.



Projector: UMAP Embedding: sst2-tiny:cls_emb Label by: sentence

Show selected Show generated

index	sentence	label
0	it 's a charming and often affecting journey .	1
1	unflinchingly bleak and desperate	0
2	allows us to hope that nolan is poised to embark a major career as a commercial yet inventive filmmaker .	1
3	the acting , costumes , music , cinematography and sound are all astounding given the production 's austere locales .	1
4	it 's slow -- very , very slow .	0
5	although laced with humor and a few fanciful touches , the film is a refreshingly serious look at young women .	1
6	a sometimes tedious film .	0
7	or doing last year 's taxes with your ex-wife .	0

*sentence TextSegment

label CategoryLabel

Facets : None

Model	From	Field	N	multiclass: accuracy	multiclass: precision	multiclass: recall	multic f1
sst2-tiny	dataset	probas	250	0.784	0.782	0.806	

□ Show matrix for selection □ Hide empty labels

Rows: label

Columns: sst2-tiny:probas:class Facets : None

Dataset (250)

sst2-tiny:probas:class

label	0	1	Total
0	36.8% (92)	11.6% (29)	48.4% (121)
1	10.0% (25)	41.6% (104)	51.6% (129)
Total	46.8% (117)	53.2% (133)	

Prediction key: probas Positive label: 1

Facets : None

Dataset

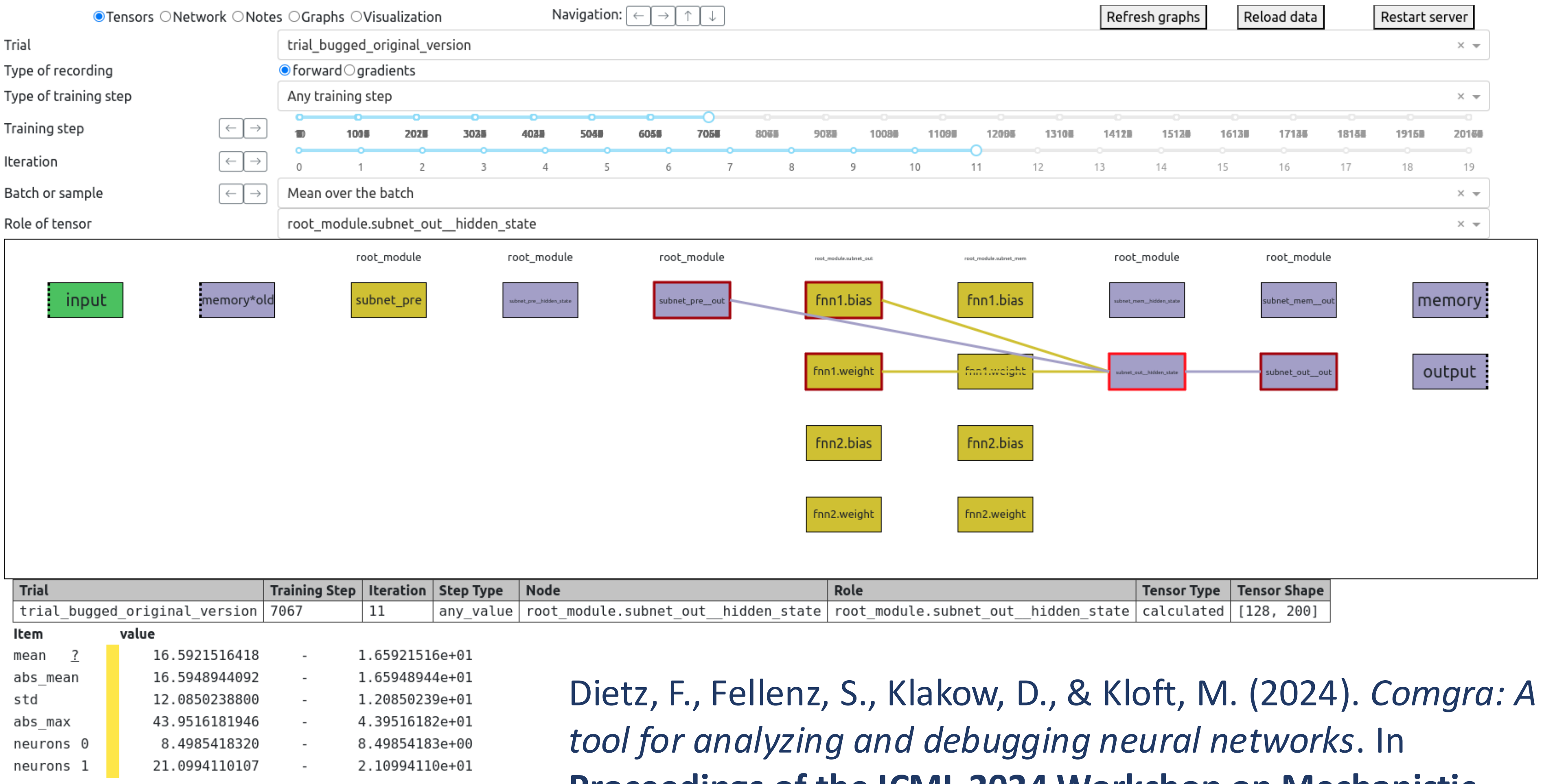
Facets : None Cost ratio (FP/FN): 1

Facet	probas threshold
All	0.5

https://www.youtube.com/watch?v=CuRI_VK83dU&t=3s

DIGITAL

Comgra: Computation Graph Analysis



Trial: trial_bugged_original_version

Type of recording: ● forward ○ gradients

Type of training step: Any training step

Training step

Iteration

Batch or sample: Mean over the batch

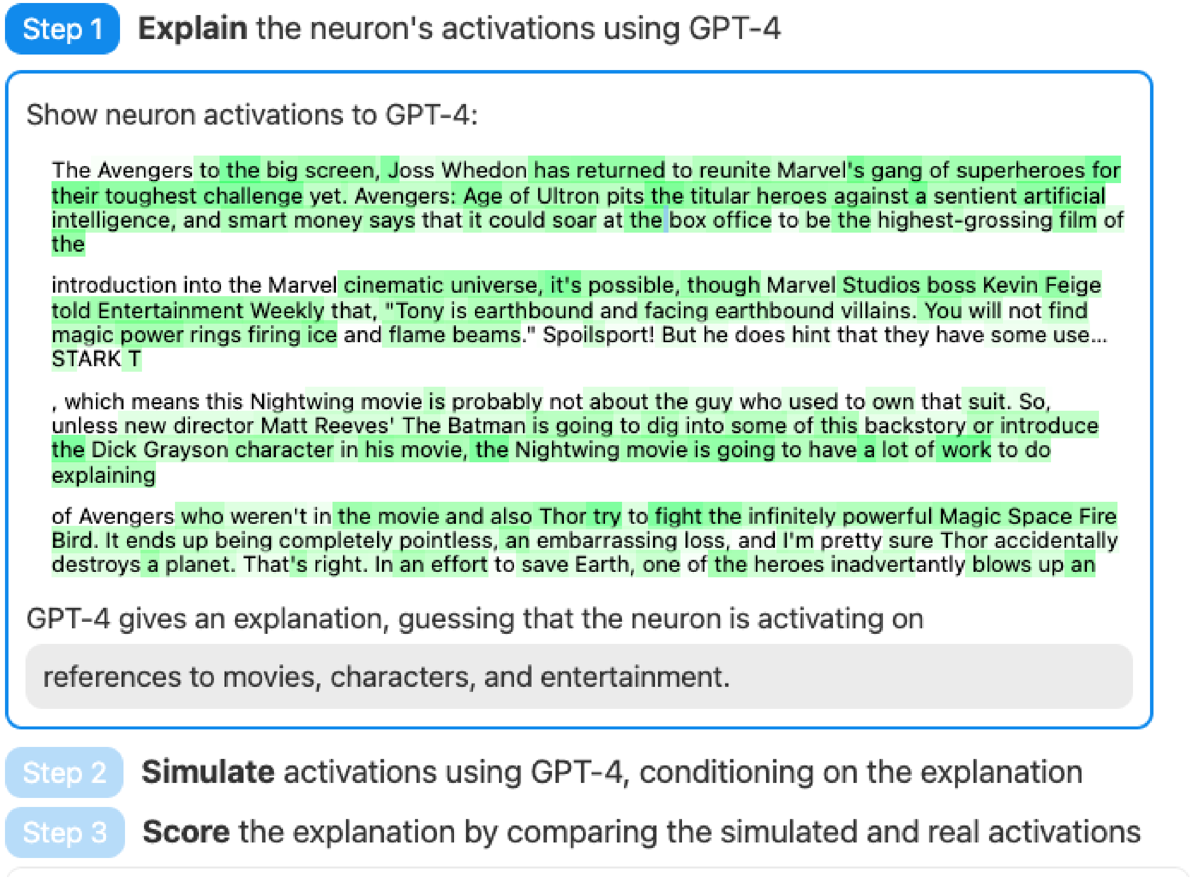
Role of tensor: root_module.subnet_out__hidden_state

Trial	Training Step	Iteration	Step Type	Node	Role	Tensor Type	Tensor Shape
trial_bugged_original_version	7067	11	any_value	root_module.subnet_out__hidden_state	root_module.subnet_out__hidden_state	calculated	[128, 200]

Item	value		
mean ?	16.5921516418	-	1.65921516e+01
abs_mean	16.5948944092	-	1.65948944e+01
std	12.0850238800	-	1.20850239e+01
abs_max	43.9516181946	-	4.39516182e+01
neurons 0	8.4985418320	-	8.49854183e+00
neurons 1	21.0994110107	-	2.10994110e+01

Dietz, F., Fellenz, S., Klakow, D., & Kloft, M. (2024). *Comgra: A tool for analyzing and debugging neural networks*. In **Proceedings of the ICML 2024 Workshop on Mechanistic Interpretability**.

Language models can explain neurons in language models



Step 1 **Explain** the neuron's activations using GPT-4

Show neuron activations to GPT-4:

The Avengers to the big screen, Joss Whedon has returned to reunite Marvel's gang of superheroes for their toughest challenge yet. Avengers: Age of Ultron pits the titular heroes against a sentient artificial intelligence, and smart money says that it could soar at the box office to be the highest-grossing film of the

introduction into the Marvel cinematic universe, it's possible, though Marvel Studios boss Kevin Feige told Entertainment Weekly that, "Tony is earthbound and facing earthbound villains. You will not find magic power rings firing ice and flame beams." Spoilsport! But he does hint that they have some use... STARK T

, which means this Nightwing movie is probably not about the guy who used to own that suit. So, unless new director Matt Reeves' The Batman is going to dig into some of this backstory or introduce the Dick Grayson character in his movie, the Nightwing movie is going to have a lot of work to do explaining

of Avengers who weren't in the movie and also Thor try to fight the infinitely powerful Magic Space Fire Bird. It ends up being completely pointless, an embarrassing loss, and I'm pretty sure Thor accidentally destroys a planet. That's right. In an effort to save Earth, one of the heroes inadvertantly blows up an

GPT-4 gives an explanation, guessing that the neuron is activating on

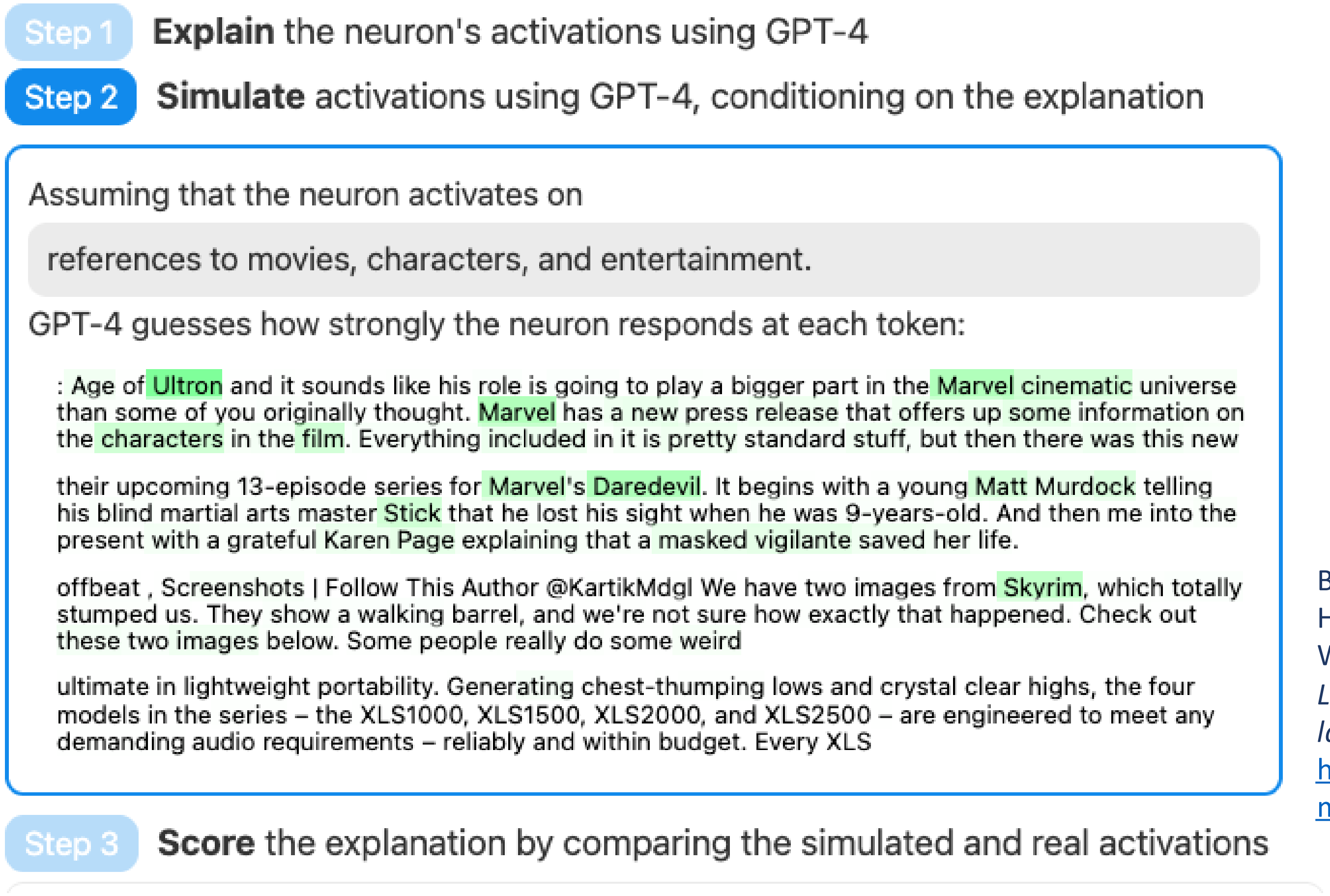
references to movies, characters, and entertainment.

Step 2 **Simulate** activations using GPT-4, conditioning on the explanation

Step 3 **Score** the explanation by comparing the simulated and real activations

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023, May 9). *Language models can explain neurons in language models*. OpenAI. https://openai.com/research/language-models-can-explain-neurons

Language models can explain neurons in language models



Step 1 **Explain** the neuron's activations using GPT-4

Step 2 **Simulate** activations using GPT-4, conditioning on the explanation

Assuming that the neuron activates on

references to movies, characters, and entertainment.

GPT-4 guesses how strongly the neuron responds at each token:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

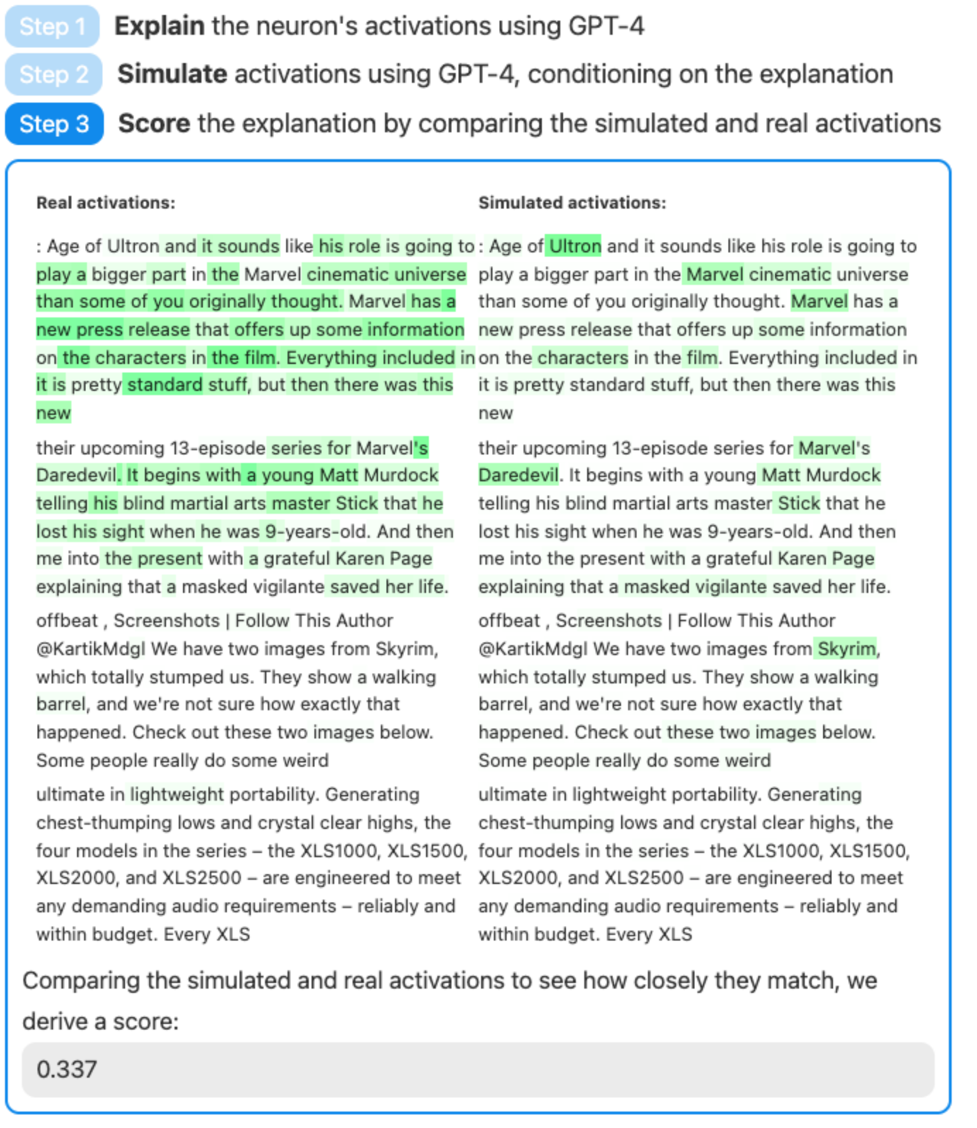
offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Step 3 **Score** the explanation by comparing the simulated and real activations

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023, May 9). *Language models can explain neurons in language models*. OpenAI. https://openai.com/research/language-models-can-explain-neurons

Language models can explain neurons in language models



Step 1 **Explain** the neuron's activations using GPT-4

Step 2 **Simulate** activations using GPT-4, conditioning on the explanation

Step 3 **Score** the explanation by comparing the simulated and real activations

Real activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

Simulated activations:

: Age of Ultron and it sounds like his role is going to play a bigger part in the Marvel cinematic universe than some of you originally thought. Marvel has a new press release that offers up some information on the characters in the film. Everything included in it is pretty standard stuff, but then there was this new

their upcoming 13-episode series for Marvel's Daredevil. It begins with a young Matt Murdock telling his blind martial arts master Stick that he lost his sight when he was 9-years-old. And then me into the present with a grateful Karen Page explaining that a masked vigilante saved her life.

offbeat , Screenshots | Follow This Author @KartikMdgl We have two images from Skyrim, which totally stumped us. They show a walking barrel, and we're not sure how exactly that happened. Check out these two images below. Some people really do some weird

ultimate in lightweight portability. Generating chest-thumping lows and crystal clear highs, the four models in the series – the XLS1000, XLS1500, XLS2000, and XLS2500 – are engineered to meet any demanding audio requirements – reliably and within budget. Every XLS

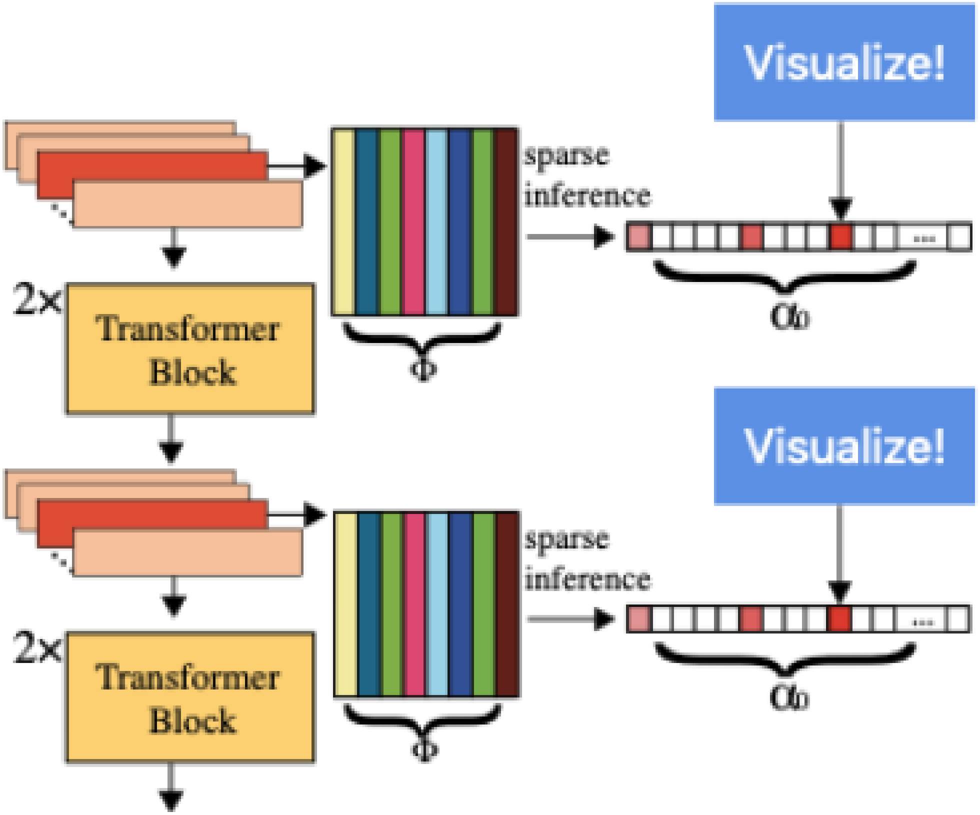
Comparing the simulated and real activations to see how closely they match, we derive a score:

0.337

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023, May 9). *Language models can explain neurons in language models*. OpenAI. https://openai.com/research/language-models-can-explain-neurons

DIGITAL

TransformVis



Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors by Zeyu Yun*, Yubei Chen*, Bruno A Olshausen, and Yann LeCun (DeeLIO Workshop@NAACL 2021).

https://transformervis.github.io/transformervis/

DIGITAL

DIGITAL

Funded by
the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe: Marie Skłodowska-Curie Actions. Neither the European Union nor the granting authority can be held responsible for them.

DIGITAL

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101119635