Interpreting Language Models Using Sparse Autoencoders (SAEs) and SHAP: Analyzing Feature Activations with Medical Terms

PhD(c). Gilber Alexis Corrales Gallego

Proposal for MATS program (Neel Nanda / Arthur Conmy)

29-08-2024

Abstract

This study aims to analyze and interpret the behavior of the GPT-2 small model using Sparse Autoencoders (SAEs) and SHAP. to understand how specific features are activated in response to medical terms. By examining statistical metrics such as skewness, kurtosis, and standard deviation, the research identifies patterns related to monosemantic and polysemantic features within the model. A set of 84 medical prompts and 84 non-medical prompts were analyzed to differentiate neurons activated by medical-specific contexts. The analysis revealed that specific neurons, like neuron 6490, are associated with recognizing medical-related patterns. The study suggests that future work could include implementing SHAP for deeper analysis, expanding the set of prompts, and applying clustering techniques to identify groups related to medical concepts.

Introduction

This document explores the use of Sparse Autoencoders (SAEs) and SHAP (Shapley Additive Explanations) to interpret and understand feature activations in the GPT-2 small language model, particularly in the context of medical terms. SAEs, as key tools, are instrumental in identifying interpretable features within language models, thereby providing a clearer representation of the semantic patterns captured by the model. SHAP, on the other hand, is employed to evaluate the contribution of each feature to the model's output, offering a detailed understanding of the impact of different activated features in natural language processing.

This analysis is based on previous research in model interpretability, such as that conducted by Trenton Bricken et al. in "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning," which demonstrates that SAEs can identify monosemantic and polysemantic features in language models. Tools and examples provided by the Saelens and TransformerLens libraries were also referenced, offering advanced methods for visualizing and analyzing neural features in language models. These resources enable the application of techniques like Activation Patching and SHAP, facilitating the identification of neural circuits within models and highlighting the ability of these methods to find complex relationships between text inputs and their corresponding model-generated responses.

Applying these methodologies provides a deeper understanding of how language models handle and process specific information, such as medical terms. It also offers new insights into how models can be fine-tuned or manipulated for specialized tasks, highlighting areas for improvement and future development in language model interpretability.

Objective

Analyze and interpret the behavior of the gpt2-small-res-jb model using Sparse Autoencoders (SAEs) and SHAP (SHapley Additive Explanations) to better understand how specific features are activated in response to medical terms.

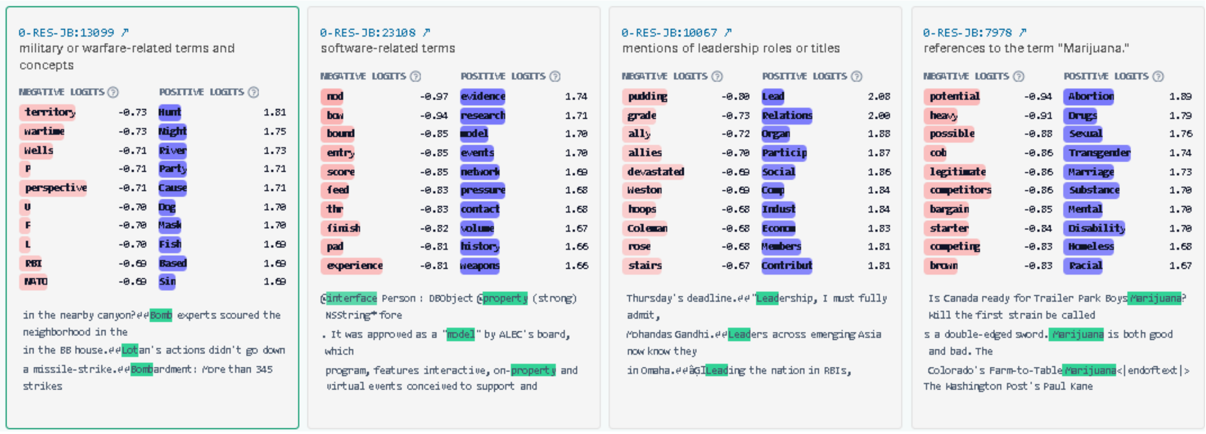
Methodology

Selection of Interesting SAE Features

For this subsection, the GPT-2 small model and its SAE, gpt2-small-res-jb, are selected to perform a statistical analysis of skewness, kurtosis, and standard deviation for the SAE features found in layer 8. The ten highest and lowest values are taken and analyzed.

Standard deviation and features

An analysis was conducted based on the standard deviation, selecting the ten logit weights with the highest and lowest values. The results are presented below.



0-RES-JB:13099 ↗
military or warfare-related terms and concepts

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
territory	-0.73	Hunt	1.81
wartime	-0.73	Night	1.75
Wells	-0.71	River	1.73
P	-0.71	Party	1.71
perspective	-0.71	Cause	1.71
U	-0.70	Dog	1.70
F	-0.70	Mask	1.70
L	-0.70	Fish	1.69
RBI	-0.69	Based	1.69
NATO	-0.69	Sin	1.69

in the nearby canyon?↵↵Bomb experts scoured the neighborhood in the
in the BB house.↵↵Lotan's actions didn't go down
a missile-strike.↵↵Bombardment: More than 345 strikes

0-RES-JB:23108 ↗
software-related terms

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
mod	-0.97	evidence	1.74
bow	-0.94	research	1.71
bound	-0.85	model	1.70
entry	-0.85	events	1.70
score	-0.85	network	1.69
feed	-0.83	pressure	1.68
thr	-0.83	contact	1.68
finish	-0.82	volume	1.67
pad	-0.81	history	1.66
experience	-0.81	weapons	1.66

@interface Person : DBObject @property (strong) NSString* fore
. It was approved as a "model" by ALEC's board, which
program, features interactive, on-property and virtual events conceived to support and

0-RES-JB:10067 ↗
mentions of leadership roles or titles

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
pudding	-0.80	Lead	2.08
grade	-0.73	Relations	2.00
ally	-0.72	Organ	1.88
allies	-0.70	Particip	1.87
devastated	-0.69	Social	1.86
Weston	-0.69	Comp	1.84
hoops	-0.68	Indust	1.84
Coleman	-0.68	Econom	1.83
rose	-0.68	Members	1.81
stairs	-0.67	Contribut	1.81

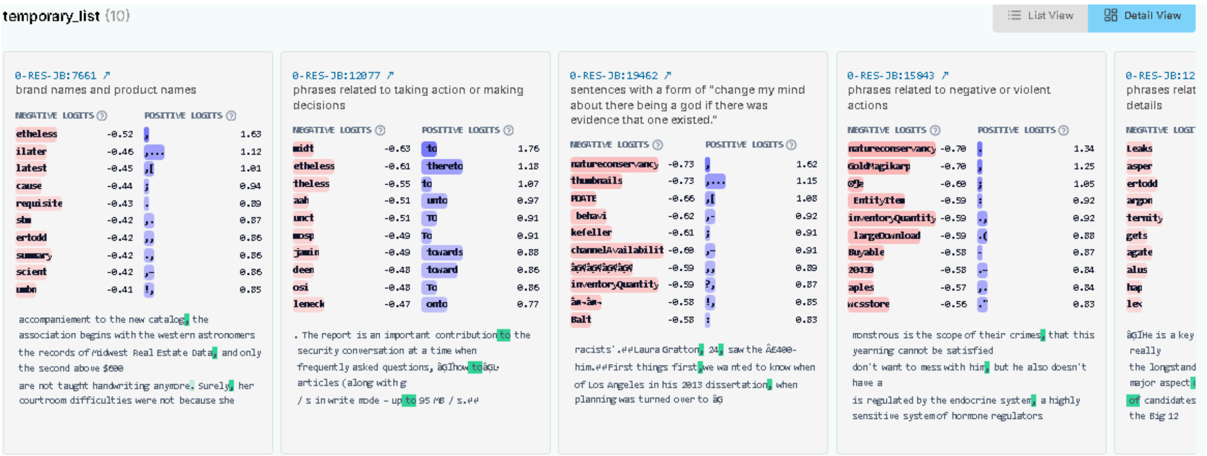
Thursday's deadline.↵↵"Leadership, I must fully admit,
Mohandas Gandhi.↵↵Leaders across emerging Asia now know they
in Omaha.↵↵âĢĺLeading the nation in RBIs,

0-RES-JB:7978 ↗
references to the term "Marijuana."

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
potential	-0.94	Abortion	1.89
heavy	-0.91	Drugs	1.79
possible	-0.88	Sexual	1.76
cob	-0.86	Transgender	1.74
legitimate	-0.86	Marriage	1.73
competitors	-0.86	Substance	1.70
bargain	-0.85	Mental	1.70
starter	-0.84	Disability	1.70
competing	-0.83	Homeless	1.68
brown	-0.83	Racial	1.67

Is Canada ready for Trailer Park Boys Marijuana? Will the first strain be called
s a double-edged sword. Marijuana is both good and bad. The
Colorado's Farm-to-Table Marijuana<|endoftext|> The Washington Post's Paul Kane

Figure 1. first four weights with respect to the standard deviation



0-RES-JB:7661 ↗
brand names and product names

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
etheless	-0.52	,	1.63
ilater	-0.46	,...	1.12
latest	-0.45	,[	1.01
cause	-0.44	;	0.94
requisite	-0.43	.	0.89
stm	-0.42	,.	0.87
ertodd	-0.42	,,	0.86
summary	-0.42	.,	0.86
scient	-0.42	,-	0.86
umbn	-0.41	!,	0.85

accompaniement to the new catalog, the association begins with the western astronomers
the records of Midwest Real Estate Data, and only the second above $600
are not taught handwriting anymore. Surely, her courtroom difficulties were not because she

0-RES-JB:12077 ↗
phrases related to taking action or making decisions

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
midt	-0.63	to	1.76
etheless	-0.61	thereto	1.18
theless	-0.55	to	1.07
aah	-0.51	unto	0.97
unct	-0.51	TO	0.91
mosp	-0.49	To	0.91
jamin	-0.49	towards	0.88
deen	-0.48	toward	0.86
osi	-0.48	To	0.86
leneck	-0.47	onto	0.77

. The report is an important contribution to the security conversation at a time when
frequently asked questions, âĢĺHow toâĢĻ articles (along with g
/s in write mode – up to 95 MB/s.↵↵

0-RES-JB:19462 ↗
sentences with a form of "change my mind about there being a god if there was evidence that one existed."

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
natureconservancy	-0.73	,	1.62
thumbnails	-0.73	,...	1.15
PDATE	-0.66	,[	1.08
behavi	-0.62	,-	0.92
kefeller	-0.61	;	0.91
channelAvailabilit	-0.60	,-	0.91
âĢ¦âĢ¦âĢ¦âĢ¦	-0.59	,,	0.89
inventoryQuantity	-0.59	?,	0.87
âĢ-âĢ-	-0.58	!,	0.85
Balt	-0.58	:	0.83

racists'.↵↵Laura Gratton, 24, saw the Â£400-
him.↵↵First things first, we wanted to know when
of Los Angeles in his 2013 dissertation, when planning was turned over to âĢ

0-RES-JB:15843 ↗
phrases related to negative or violent actions

NEGATIVE LOGITS ⓘ		POSITIVE LOGITS ⓘ	
natureconservancy	-0.70	.	1.34
GoldMagikarp	-0.70	,	1.25
oJe	-0.60	;	1.05
EntityItem	-0.59	:	0.92
inventoryQuantity	-0.59	.,	0.92
largeDownload	-0.59	.(	0.88
Buyable	-0.58	-	0.87
20439	-0.58	.-	0.84
aples	-0.57	,.	0.84
wcsstore	-0.56	."	0.83

monstrous is the scope of their crimes, that this yearning cannot be satisfied
don't want to mess with him, but he also doesn't have a
is regulated by the endocrine system, a highly sensitive system of hormone regulators

0-RES-JB:12
phrases relat
details

NEGATIVE LOGI
Leaks
asper
ertodd
argon
ternity
gets
agate
alus
hap
lex

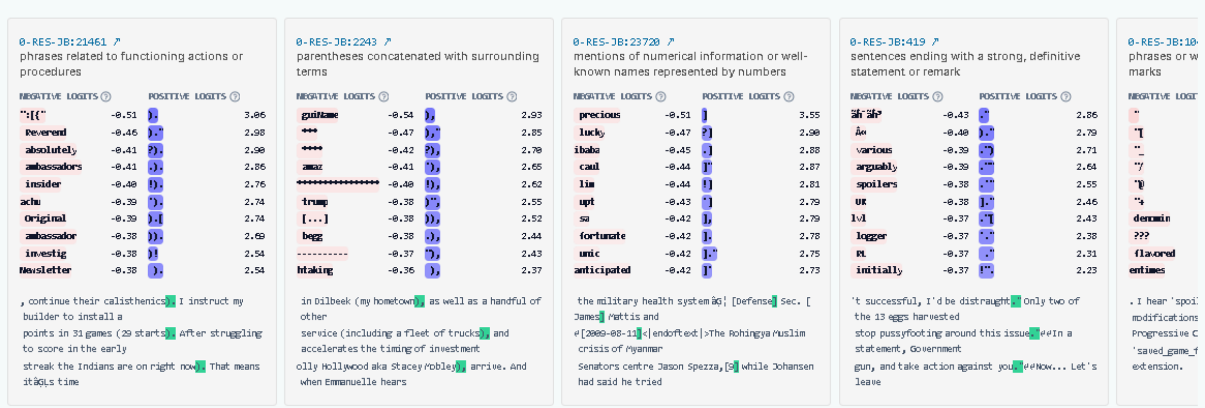
âĢĺHe is a key
really
the longstand
major aspect
of candidates
the Big 12

Figure 2. last four weights with respect to the standard deviation

Features with a lower standard deviation tend to be associated with similar patterns, which suggests a relationship between low standard deviation and mono-semantic features. On the other hand, features with higher standard deviation values are often linked to more varied positive logits. These observations could lead to two hypotheses: the patterns they detect are highly complex (potentially indicating patterns not easily recognized by humans) or are associated with polysemantic features.

Kurtisis and features

An analysis was conducted based on the kurtosis, selecting the ten logit weights with the highest and the ten with the lowest values. The results are presented below.



0-RES-JB:21461 ↗
phrases related to functioning actions or procedures

NEGATIVE LOGITS		POSITIVE LOGITS	
":[{"	-0.51	).	3.06
Reverend	-0.46	)."	2.98
absolutely	-0.41	?).	2.90
ambassadors	-0.41	.).	2.86
insider	-0.40	!).	2.76
achu	-0.39	').	2.74
Original	-0.39	).[	2.74
ambassador	-0.38	)).	2.69
investig	-0.38	)!	2.54
Newsletter	-0.38	).	2.54

, continue their calisthenics). I instruct my builder to install a

points in 31 games (29 starts). After struggling to score in the early

streak the Indians are on right now). That means itâĢĻs time

0-RES-JB:2243 ↗
parentheses concatenated with surrounding terms

NEGATIVE LOGITS		POSITIVE LOGITS	
guiName	-0.54	),	2.93
***	-0.47	),"	2.85
****	-0.42	?),	2.70
amaz	-0.41	'),	2.65
****************	-0.40	!),	2.62
trump	-0.38	)",	2.55
[...]	-0.38	)),	2.52
begg	-0.38	.),	2.44
----------	-0.37	"),	2.43
htaking	-0.36	),	2.37

in Dilbeek (my hometown), as well as a handful of other

service (including a fleet of trucks), and accelerates the timing of investment

olly Hollywood aka Stacey Mobley), arrive. And when Emmanuelle hears

0-RES-JB:23720 ↗
mentions of numerical information or well-known names represented by numbers

NEGATIVE LOGITS		POSITIVE LOGITS	
precious	-0.51	]	3.55
lucky	-0.47	?]	2.90
ibaba	-0.45	.]	2.88
caul	-0.44	]"	2.87
lim	-0.44	!]	2.81
upt	-0.43	']	2.79
sa	-0.42	],	2.79
fortunate	-0.42	].	2.78
unic	-0.42	]."	2.75
anticipated	-0.42	]'	2.73

the military health system âĢ¦ [Defense] Sec. [James] Mattis and

↵[2009-08-11]<|endoftext|>The Rohingya Muslim crisis of Myanmar

Senators centre Jason Spezza, [9] while Johansen had said he tried

0-RES-JB:419 ↗
sentences ending with a strong, definitive statement or remark

NEGATIVE LOGITS		POSITIVE LOGITS	
ãĥ¯ãĥ³	-0.43	."	2.86
Âł	-0.40	)."	2.79
various	-0.39	.")	2.71
arguably	-0.39	.""	2.64
spoilers	-0.38	."'	2.55
UR	-0.38	]."	2.46
lvl	-0.37	."[	2.43
logger	-0.37	"."	2.38
RL	-0.37	."	2.31
initially	-0.37	!".	2.23

't successful, I'd be distraught." Only two of the 13 eggs harvested

stop pussyfooting around this issue."↵↵In a statement, Government

gun, and take action against you."↵↵Now... Let's leave

0-RES-JB:10
phrases or w
marks

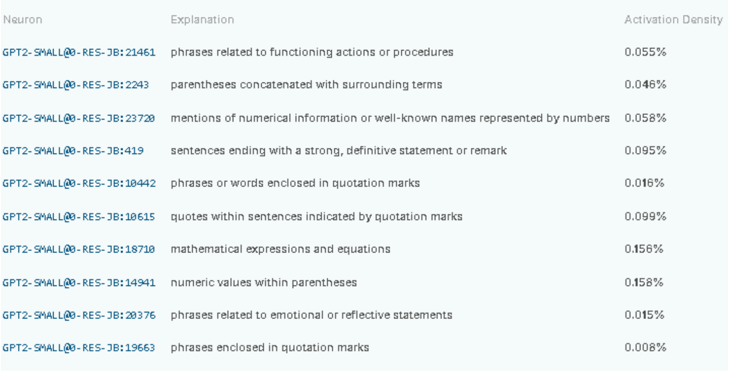
NEGATIVE LOGI

"
"[
"_
"/
"@
"+
denomin
???
flavored
entimes

. I hear 'spoi
modifications
Progressive C
'saved_game_f
extension.

Figure 3. first four weights with respect to the Kurtisis

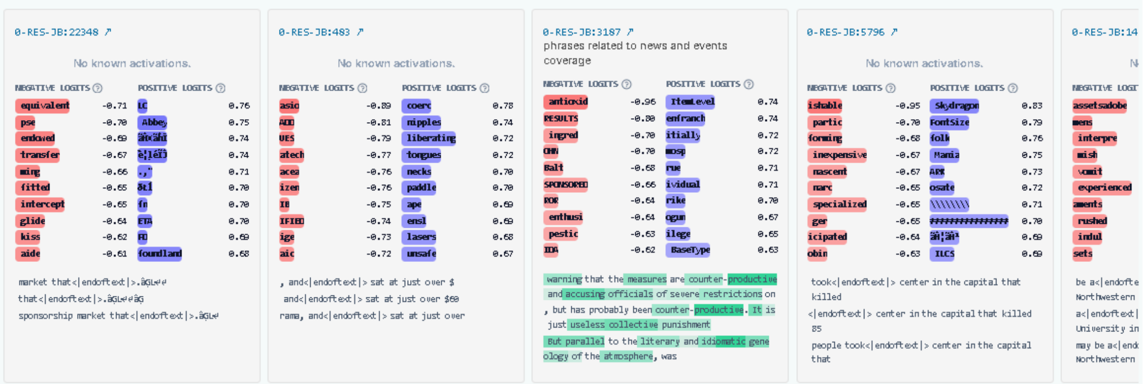
Again, features with a higher kurtosis value tend to be more specific and detect more precise patterns, such as symbols in the text, suggesting a possible strong relationship with mono-semantic features.



Neuron	Explanation	Activation Density
GPT2-SMALL@0-RES-JB:21461	phrases related to functioning actions or procedures	0.055%
GPT2-SMALL@0-RES-JB:2243	parentheses concatenated with surrounding terms	0.046%
GPT2-SMALL@0-RES-JB:23720	mentions of numerical information or well-known names represented by numbers	0.058%
GPT2-SMALL@0-RES-JB:419	sentences ending with a strong, definitive statement or remark	0.095%
GPT2-SMALL@0-RES-JB:10442	phrases or words enclosed in quotation marks	0.016%
GPT2-SMALL@0-RES-JB:10615	quotes within sentences indicated by quotation marks	0.099%
GPT2-SMALL@0-RES-JB:18710	mathematical expressions and equations	0.156%
GPT2-SMALL@0-RES-JB:14941	numeric values within parentheses	0.158%
GPT2-SMALL@0-RES-JB:20376	phrases related to emotional or reflective statements	0.015%
GPT2-SMALL@0-RES-JB:19663	phrases enclosed in quotation marks	0.008%

Figure 4. Pattern of each feature according to Claude 3.5

An interesting finding is that features with low kurtosis tend to have no activation, meaning they do not contribute to the model. This finding suggests that high kurtosis may be necessary for the model's performance.



0-RES-JB:22348 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
equivalent	-0.71	LC	0.76
pse	-0.70	Abbey	0.75
endowed	-0.69	âĩĉãĥ	0.74
transfer	-0.67	è¦léĨ	0.74
ming	-0.66	.,"	0.71
fitted	-0.65	ðŁ	0.70
intercept	-0.65	fn	0.70
glide	-0.64	ETA	0.70
kiss	-0.62	FD	0.69
aide	-0.61	foundland	0.68

market that<|endoftext|>.âĢĿ↵↵
that<|endoftext|>.âĢĿ↵↵âĢ
sponsorship market that<|endoftext|>.âĢĿ↵

0-RES-JB:483 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
asio	-0.89	coerc	0.78
ADD	-0.81	nipples	0.74
UES	-0.79	liberating	0.72
atech	-0.77	tongues	0.72
acea	-0.76	necks	0.70
izen	-0.76	paddle	0.70
IB	-0.75	ape	0.69
IFIED	-0.74	ensl	0.69
ige	-0.73	lasers	0.68
aic	-0.72	unsafe	0.67

, and<|endoftext|> sat at just over $
and<|endoftext|> sat at just over $60
rama, and<|endoftext|> sat at just over

0-RES-JB:3187 ↗
phrases related to news and events coverage

NEGATIVE LOGITS		POSITIVE LOGITS	
antioxid	-0.96	ItemLevel	0.74
RESULTS	-0.80	enfranch	0.74
ingred	-0.70	itially	0.72
OHN	-0.70	mosp	0.72
Balt	-0.68	rue	0.71
SPONSORED	-0.66	ividual	0.71
ROR	-0.64	rike	0.70
enthusi	-0.64	ogun	0.67
pestic	-0.63	ilege	0.65
IDA	-0.62	BaseType	0.63

warning that the measures are counter-productive
and accusing officials of severe restrictions on
, but has probably been counter-productive. It is just useless collective punishment
But parallel to the literary and idiomatic gene
ology of the atmosphere, was

0-RES-JB:5796 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
ishable	-0.95	Skydragon	0.83
partic	-0.70	FontSize	0.79
forming	-0.68	folk	0.76
inexpensive	-0.67	Mania	0.75
nascent	-0.67	ARR	0.73
narc	-0.65	osate	0.72
specialized	-0.65	\\\\\\\\	0.71
ger	-0.65	################	0.70
icipated	-0.64	ãĥ¦ãĥ³	0.69
obin	-0.63	ILCS	0.69

took<|endoftext|> center in the capital that killed
<|endoftext|> center in the capital that killed 85
people took<|endoftext|> center in the capital that

0-RES-JB:14

N

NEGATIVE LOGI

assetsadobe
mens
interpre
mish
vomit
experienced
aments
rushed
indul
sets

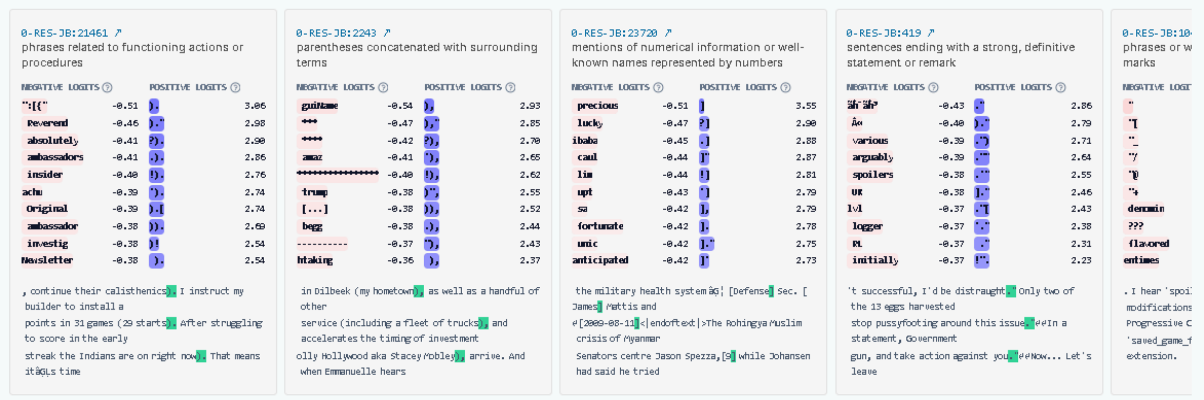
be a<|endofte
Northwestern
a<|endoftext|
University in
may be a<|end
Northwestern

Figure 5. last four weights with respect to the Kurtisis

An interesting finding is that features with low kurtosis tend to have no activation, meaning they do not contribute to the model. This pattern suggests that high kurtosis may be necessary for the model's performance.

Analysis by Skewness

Skewness measures the degree of asymmetry in a data distribution around its mean, indicating how skewed or tilted a distribution is. Again, the ten highest and lowest values were explored.



0-RES-JB:21461 ↗
phrases related to functioning actions or procedures

NEGATIVE LOGITS		POSITIVE LOGITS	
":[{"	-0.51	).	3.06
Reverend	-0.46	)."	2.98
absolutely	-0.41	?).	2.90
ambassadors	-0.41	.).	2.86
insider	-0.40	!).	2.76
achu	-0.39	").	2.74
Original	-0.39	).[	2.74
ambassador	-0.38	)).	2.69
investig	-0.38	)!	2.54
Newsletter	-0.38	).	2.54

, continue their calisthenics). I instruct my builder to install a
points in 31 games (29 starts). After struggling to score in the early
streak the Indians are on right now). That means itâĢĻs time

0-RES-JB:2243 ↗
parentheses concatenated with surrounding terms

NEGATIVE LOGITS		POSITIVE LOGITS	
guiName	-0.54	),	2.93
***	-0.47	),"	2.85
****	-0.42	?),	2.70
amaz	-0.41	"),	2.65
****************	-0.40	!),	2.62
trump	-0.38	)",	2.55
[...]	-0.38	)),	2.52
begg	-0.38	.),	2.44
----------	-0.37	"),	2.43
htaking	-0.36	),	2.37

in Dilbeek (my hometown), as well as a handful of other
service (including a fleet of trucks), and accelerates the timing of investment
olly Hollywood aka Stacey Mobley), arrive. And when Emmanuelle hears

0-RES-JB:23720 ↗
mentions of numerical information or well-known names represented by numbers

NEGATIVE LOGITS		POSITIVE LOGITS	
precious	-0.51	]	3.55
lucky	-0.47	?]	2.90
ibaba	-0.45	.]	2.88
caul	-0.44	]"	2.87
lim	-0.44	!]	2.81
upt	-0.43	"]	2.79
sa	-0.42	],	2.79
fortunate	-0.42	].	2.78
unic	-0.42	]."	2.75
anticipated	-0.42	]'	2.73

the military health system âĢ¦ [Defense] Sec. [James] Mattis and
↵[2009-08-11]<|endoftext|>The Rohingya Muslim crisis of Myanmar
Senators centre Jason Spezza,[9] while Johansen had said he tried

0-RES-JB:419 ↗
sentences ending with a strong, definitive statement or remark

NEGATIVE LOGITS		POSITIVE LOGITS	
ãĥ¯ãĥ³	-0.43	."	2.86
Ãĸ	-0.40	)."	2.79
various	-0.39	.")	2.71
arguably	-0.39	.""	2.64
spoilers	-0.38	."'	2.55
UR	-0.38	]."	2.46
lvl	-0.37	."[	2.43
logger	-0.37	'."	2.38
RL	-0.37	."	2.31
initially	-0.37	!".	2.23

't successful, I'd be distraught." Only two of the 13 eggs harvested
stop pussyfooting around this issue."↵↵In a statement, Government
gun, and take action against you."↵↵Now... Let's leave

0-RES-JB:10

phrases or w
marks

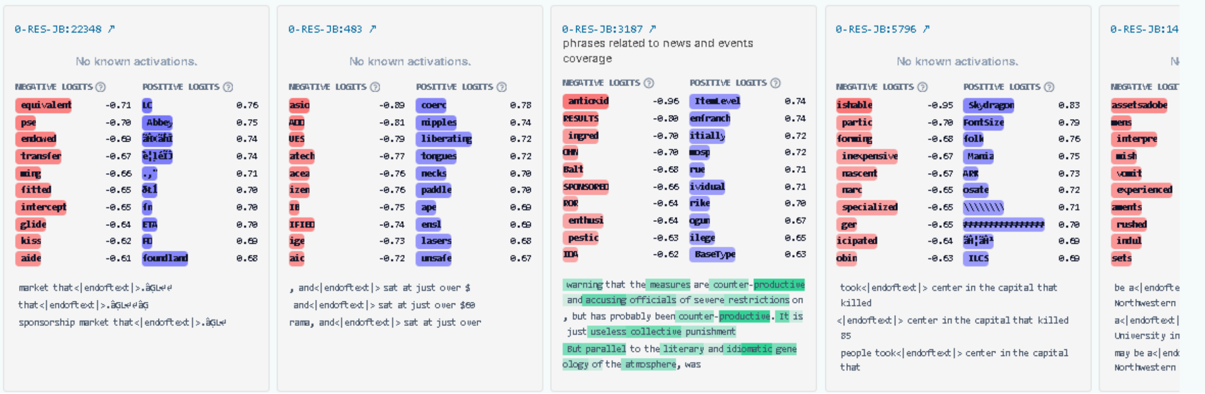
NEGATIVE LOGI

"
"[
"_
"/
"@
"+
denomin
???
flavored
entimes

. I hear 'spoi
modifications
Progressive C
'saved_game_f
extension.

Figure 6. first four weights with respect to the Skewness

Again, we found that features with a high skewness value, meaning those with a significant tilt in their distribution, are particular to simple patterns such as periods, numbers, and parentheses.



0-RES-JB:22348 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
equivalent	-0.71	LC	0.76
pse	-0.70	Abbey	0.75
endowed	-0.69	âĢ¢âĢ	0.74
transfer	-0.67	è¦léĩĳ	0.74
ming	-0.66	.,"	0.71
fitted	-0.65	ðŁ	0.70
intercept	-0.65	fn	0.70
glide	-0.64	ETA	0.70
kiss	-0.62	FD	0.69
aide	-0.61	foundland	0.68

market that<|endoftext|>.âĢĿ↵↵
that<|endoftext|>.âĢĿ↵↵âĢ
sponsorship market that<|endoftext|>.âĢĿ↵

0-RES-JB:483 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
asio	-0.89	coerc	0.78
ADD	-0.81	nipples	0.74
UES	-0.79	liberating	0.72
atech	-0.77	tongues	0.72
acea	-0.76	necks	0.70
izen	-0.76	paddle	0.70
IB	-0.75	ape	0.69
IFIED	-0.74	ensl	0.69
ige	-0.73	lasers	0.68
aic	-0.72	unsafe	0.67

, and<|endoftext|> sat at just over $
and<|endoftext|> sat at just over $60
rama, and<|endoftext|> sat at just over

0-RES-JB:3187 ↗
phrases related to news and events coverage

NEGATIVE LOGITS		POSITIVE LOGITS	
antioxid	-0.96	ItemLevel	0.74
RESULTS	-0.80	enfranch	0.74
ingred	-0.70	itially	0.72
CHN	-0.70	mosp	0.72
Balt	-0.68	rue	0.71
SPONSORED	-0.66	ividual	0.71
ROR	-0.64	rike	0.70
enthusi	-0.64	ogun	0.67
pestic	-0.63	ilege	0.65
IDA	-0.62	BaseType	0.63

warning that the measures are counter-productive
and accusing officials of severe restrictions on
, but has probably been counter-productive. It is
just useless collective punishment
But parallel to the literary and idiomatic gene
ology of the atmosphere, was

0-RES-JB:5796 ↗

No known activations.

NEGATIVE LOGITS		POSITIVE LOGITS	
ishable	-0.95	Skydragon	0.83
partic	-0.70	FontSize	0.79
forming	-0.68	folk	0.76
inexpensive	-0.67	Mania	0.75
nascent	-0.67	ARK	0.73
narc	-0.65	osate	0.72
specialized	-0.65	\\\\\\\\	0.71
ger	-0.65	################	0.70
icipated	-0.64	âĤ¦âĤ¹	0.69
obin	-0.63	ILCS	0.69

took<|endoftext|> center in the capital that killed
<|endoftext|> center in the capital that killed 85
people took<|endoftext|> center in the capital that

0-RES-JB:14

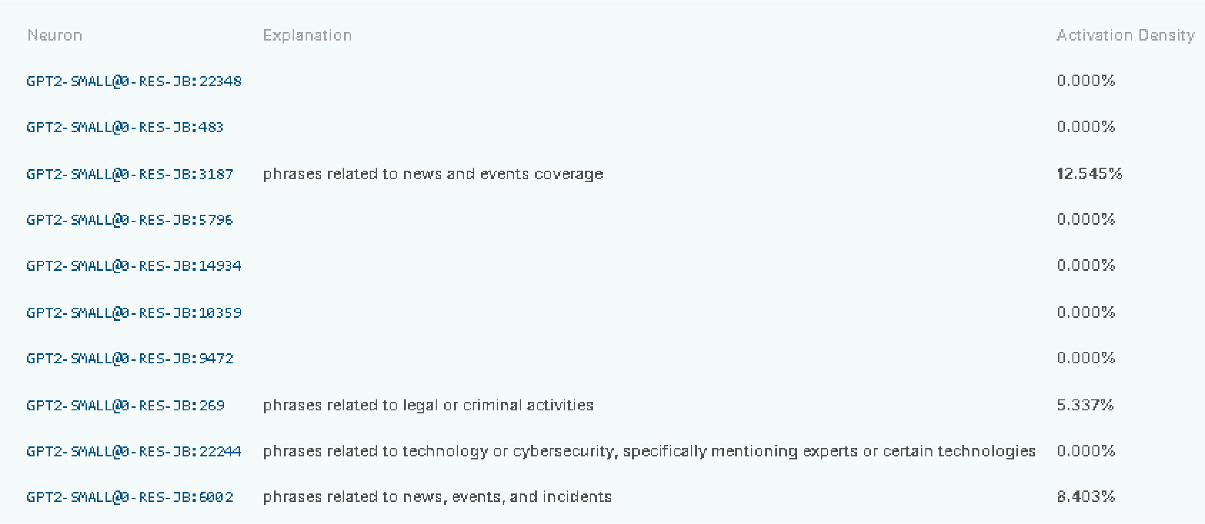
N

NEGATIVE LOGI
assetsadobe
mens
interpre
mish
vomit
experienced
aments
rushed
indul
sets

be a<|endofte
Northwestern
a<|endoftext|
University in
may be a<|end
Northwestern

Figure 7. last four weights with respect to the Skewness

In the low skewness values, features with little or almost no activation capture important or high-level patterns, but those with activation, possibly high, capture features associated with relevant news phrases and events. As the previous image shows, the feature 0-RES-JB:3187 captures patterns associated with relevant news phrases and events.



Neuron	Explanation	Activation Density
GPT2-SMALL@0-RES-JB:22348		0.000%
GPT2-SMALL@0-RES-JB:483		0.000%
GPT2-SMALL@0-RES-JB:3187	phrases related to news and events coverage	12.545%
GPT2-SMALL@0-RES-JB:5796		0.000%
GPT2-SMALL@0-RES-JB:14934		0.000%
GPT2-SMALL@0-RES-JB:10359		0.000%
GPT2-SMALL@0-RES-JB:9472		0.000%
GPT2-SMALL@0-RES-JB:269	phrases related to legal or criminal activities	5.337%
GPT2-SMALL@0-RES-JB:22244	phrases related to technology or cybersecurity, specifically mentioning experts or certain technologies	0.000%
GPT2-SMALL@0-RES-JB:6002	phrases related to news, events, and incidents	8.403%

Figure 8. Explanation generate for claude 3.5

As mentioned, features with low skewness have very high activation and capture patterns associated with topics such as criminal and legal activities or news, events, and incidents.

Evaluation of Activation Patterns with Medical Prompts

This subsection aims to determine whether features are activated when a word associated with a health-related disease appears in the prompt. This exercise is the basis for the document's title. To achieve this, we will again analyze layer 8 of the previously mentioned model, focusing on its activations.

Creating the base prompts

To begin the analysis, the initial set of 100 prompts related to various medical pathologies was created using ChatGPT-4. This step was essential to ensure a diverse and comprehensive range of medical scenarios, each containing references to specific pathologies.

Filtering prompts with SHAP

The objective was to use SHAP (Shapley Additive explanations) to filter prompts the initial model found most relevant by identifying which medical prompts had the highest SHAP values associated with the pathology name. However, due to computational capacity limitations, we utilized the base version of GPT-2 instead. This approach allowed us to efficiently determine which medical prompts had the most significant SHAP values linked to the pathology names, enabling us to focus on the most relevant inputs for further analysis.

Out of the initial 100 medical prompts generated, 84 were selected based on their relevance as determined by SHAP (Shapley Additive Explanations) values. These selected prompts had the highest SHAP values linked to the names of the pathologies, indicating their importance for further analysis.

Selection of influential or pathology-related features

In this analysis, we will use the GPT-2 Small model, loaded with the HookedTransformer library, to identify features associated explicitly with medical terms. We will compare the activations generated by medical prompts with those generated by non-medical control prompts. The modified evaluation function will compute the feature activations for both sets of prompts and determine the most active neurons in each case. Subsequently, we will filter out the features frequently activated in medical and non-medical contexts, considering them non-specific, and analyze the neurons exclusively activated in medical prompts. This approach will allow us to identify the relevant features associated with medical terms more accurately, eliminating potential activations caused by symbols or generic terms.

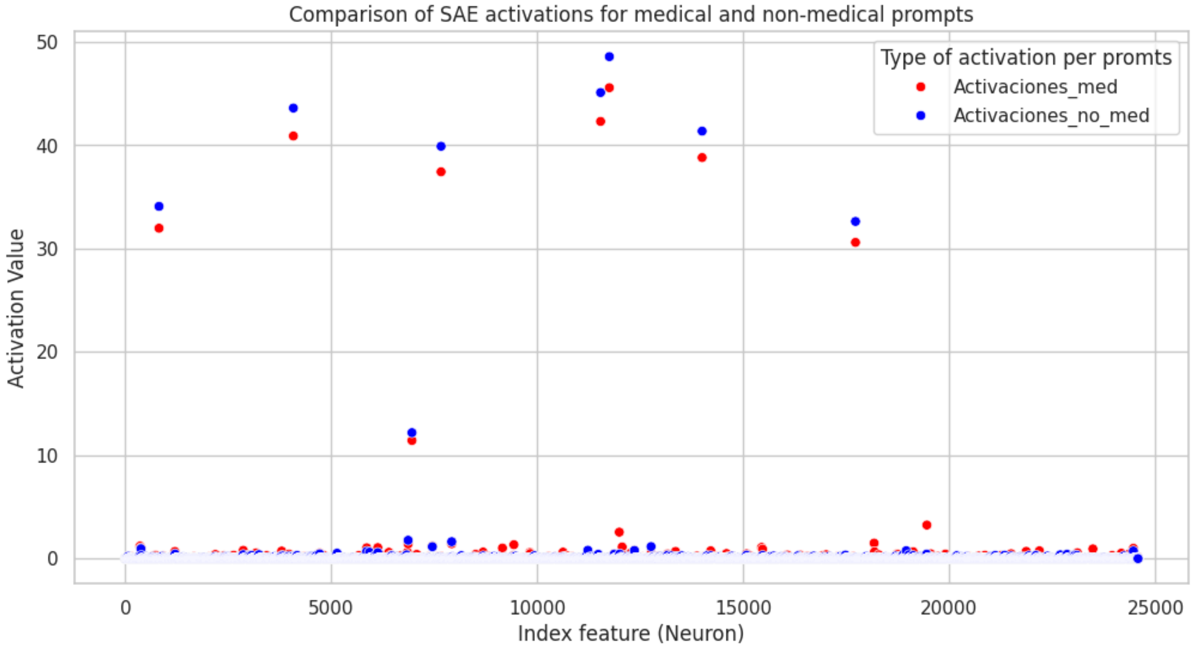


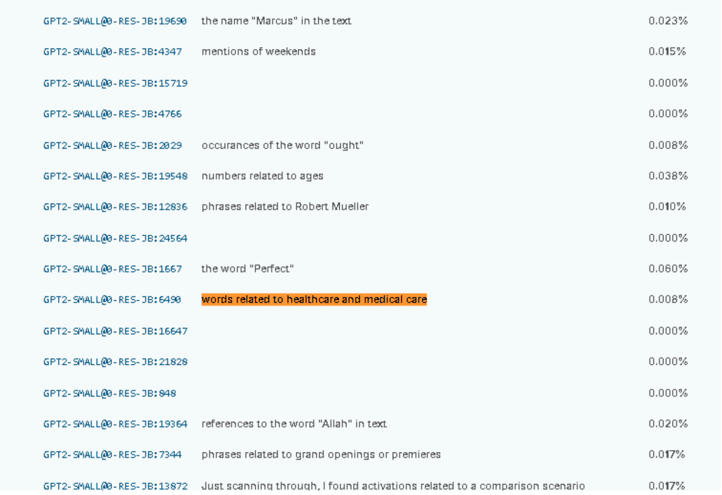
Figure 9.Average activation graph for the 84 medical and non-medical prompts

Once this was clarified, neurons with negative activations for the medical stimulus were removed. The difference between the average activations (medical and non-medical) was calculated, and then the 40 highest values within the standard deviation range were selected. It is important to note that tests were also conducted with the top 40 highest values, but no good results were achieved. Finally, the 40 neurons were analyzed in Neuronpedia.

Features_neuron	Activaciones_med	Activaciones_no_med	Difference
10276	0.190654	0.000383	0.190271
740	0.313702	0.123839	0.189863
10664	0.187483	0.001924	0.185559
17667	0.185639	0.001511	0.184128
57	0.191191	0.008165	0.183026
19717	0.220718	0.041975	0.178743
17471	0.175220	0.000000	0.175220
14624	0.175142	0.000000	0.175142
14288	0.208901	0.034685	0.174216
8826	0.173344	0.000000	0.173344
20476	0.171944	0.000000	0.171944
14367	0.165671	0.000000	0.165671
6585	0.163456	0.000000	0.163456
19690	0.160095	0.000000	0.160095
4347	0.159153	0.000000	0.159153
15719	0.156561	0.005155	0.151406
4766	0.243417	0.092564	0.150852
2029	0.149327	0.000142	0.149184
19548	0.161294	0.014016	0.147278
12836	0.146119	0.000000	0.146119
24564	0.165707	0.019807	0.145900
1667	0.170036	0.024717	0.145319
6490	0.144562	0.000000	0.144562
16647	0.147324	0.003386	0.143938

21828	0.167574	0.024935	0.142640
848	0.144347	0.003005	0.141342
19364	0.141108	0.001112	0.139996
7344	0.149478	0.010270	0.139209
13872	0.245501	0.106667	0.138834
12554	0.140330	0.003393	0.136937
20955	0.136513	0.000000	0.136513
15076	0.465050	0.328780	0.136270
17271	0.156657	0.021944	0.134713
16687	0.135332	0.000633	0.134699
14032	0.157261	0.022669	0.134593
16471	0.133981	0.000000	0.133981
24232	0.142612	0.009032	0.133580
23098	0.236032	0.102520	0.133512
4732	0.222698	0.089406	0.133292
18896	0.225538	0.092715	0.132823

Tabla 1. The forty candidate neurons extract patterns related to pathology or medicine



GPT2-SMALL@0-RES-JB:19690	the name "Marcus" in the text	0.023%
GPT2-SMALL@0-RES-JB:4347	mentions of weekends	0.015%
GPT2-SMALL@0-RES-JB:15719		0.000%
GPT2-SMALL@0-RES-JB:4766		0.000%
GPT2-SMALL@0-RES-JB:2029	occurances of the word "ought"	0.008%
GPT2-SMALL@0-RES-JB:19548	numbers related to ages	0.038%
GPT2-SMALL@0-RES-JB:12836	phrases related to Robert Mueller	0.010%
GPT2-SMALL@0-RES-JB:24564		0.000%
GPT2-SMALL@0-RES-JB:1667	the word "Perfect"	0.060%
GPT2-SMALL@0-RES-JB:6490	words related to healthcare and medical care	0.008%
GPT2-SMALL@0-RES-JB:16647		0.000%
GPT2-SMALL@0-RES-JB:21828		0.000%
GPT2-SMALL@0-RES-JB:848		0.000%
GPT2-SMALL@0-RES-JB:19364	references to the word "Allah" in text	0.020%
GPT2-SMALL@0-RES-JB:7344	phrases related to grand openings or premieres	0.017%
GPT2-SMALL@0-RES-JB:13872	Just scanning through, I found activations related to a comparison scenario	0.017%

Figure 10. characteristic of the group of forty neurons that recognize patterns related to medicine

Upon analyzing Neuronpedia, neuron 6490 recognized patterns associated with medicine, precisely 'words related to healthcare and medical care,' achieving the search objective. Interestingly, it showed low activation for the proposed prompts.

Limitations and future developments

If more time were available, it would be interesting to implement SHAP on the GPT-2 small model and analyze both the model outputs and the activations of each autoencoder. However, it should be noted that this would require better computational resources. Testing with a more extensive set of prompts could yield better results and help identify neurons more specialized in specific pathologies.

Future work should focus on analyzing the circuits that activate neuron 6490; better insights might be obtained with more time. Finally, two clustering techniques, such as K-means or t-SNE, could be applied to identify groups related to medical concepts. Manipulating these neurons to increase their activation would help enhance the model.

References

Bricken, T., Elhage, N., et al. (2023). *Towards Monosemanticity: Decomposing Language Models With Dictionary Learning*. Anthropic. Link.

Nanda, N. (2023). *An Extremely Opinionated Annotated List of My Favourite Mechanistic Interpretability Papers v2*. AI Alignment Forum. Link.

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. In Advances in Neural Information Processing Systems (NeurIPS). Link.

SAELens: *A Library for Interpreting Language Models Using Sparse Autoencoders*. Link.

TransformerLens: *Tools for Analyzing Transformer Models*. Link.