Michael C.-K. Wu, Stephen V. David, and Jack L. Gallant. Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, 29:477–505, 2006. ISSN 0147-006X. doi: 10.1146/annurev.neuro.29.051605.113024.

Zeyu Yun, Yubei Chen, Bruno A Olshausen, and Yann LeCun. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. *arXiv preprint arXiv:2103.15949*, 2021.

Omar Zaidan and Jason Eisner. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*, pp. 31–40, 2008.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*, 2021.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In *International Conference on Machine Learning*, pp. 27099–27116. PMLR, 2022.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *arXiv preprint arXiv:2302.14233*, 2023.

Zhiying Zhu, Weixin Liang, and James Zou. Gsclip: A framework for explaining distribution shifts in natural language. *arXiv preprint arXiv:2206.15007*, 2022.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

A Appendix

A.1 Methodology details extended

Table A1: Statistics on corpuses used for explanation. Wikitext is used for BERT explanation and Moth stories are used for fMRI voxel explanation.

	Unique unigrams	Unique bigrams	Unique trigrams
Wikitext (Merity et al., 2016)	157k	3,719k	9,228k
Moth stories (LeBel et al., 2022)	117k	79k	140k
Combined	158k	3,750k	9,334k

Prompts used in SASC The summarization step summarizes 30 randomly chosen ngrams from the top 50 and generates 5 candidate explanations using the prompt *Here is a list of phrases:\n{phrases}\nWhat is a common theme among these phrases?\nThe common theme among these phrases is* ___.

In the synthetic scoring step, we generate similar synthetic strings with the prompt *Generate 10 phrases that are similar to the concept of {explanation}:*. For dissimilar synthetic strings we use the prompt *Generate 10 phrases that are not similar to the concept of {explanation}:*. Minor automatic processing is applied to LLM outputs, e.g. parsing a bulleted list, converting to lowercase, and removing extra whitespaces.

A.2 SYNTHETIC MODULE INTERPRETATION

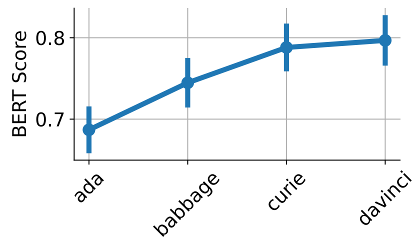

Figure A1: The BERT score between generated explanation and groundtruth explanation generally increases as the size of the helper LLM for summarization/generation increases. Models are accessed via the OpenAI API (`text-ada-001`, `text-babbage-001`, `text-curie-001`, `text-davinci-001`, all accessed on Feb. 2023) and are in order of increasing size. BERT score for each module is computed as the maximum over the 5 generated explanations.

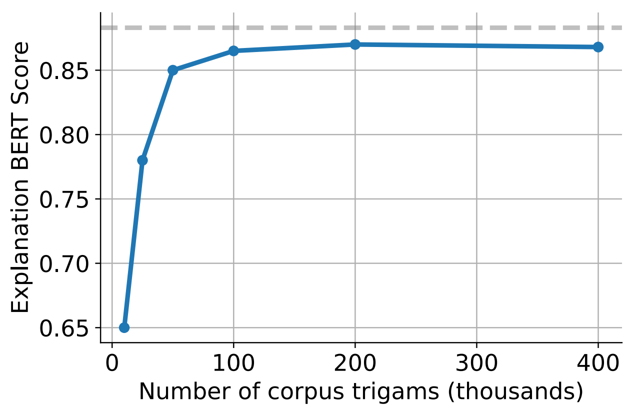

Figure A2: Explanation BERT score for the 54 synthetic datasets as a function of corpus size. Performance plateaus around 100,000 ngrams. Corpus is created by randomly subsampling the unique trigrams in the WikiText dataset (Merity et al., 2016). Gray dotted line shows the result when evaluating on dataset-specific corpuses, as in the *Default* setting in Table 1.

Table A2: 54 synthetic modules and information about their underlying data corpus. Note that some modules use the same groundtruth Keyword (e.g. *environmentalism*), but that the underlying data corpus contains different data (e.g. text that is pro/anti environmentalism).

Module name	Groundtruth keyphrase	Dataset explanation	Examples	Unique unigrams
0-irony	sarcasm	contains irony	590	3897
1-objective	unbiased	is a more objective description of what happened	739	5628
2-subjective	subjective	contains subjective opinion	757	5769
3-god	religious	believes in god	164	1455
4-atheism	atheistic	is against religion	172	1472
5-evacuate	evacuation	involves a need for people to evacuate	2670	16505
6-terorrism	terrorism	describes a situation that involves terrorism	2640	16608
7-crime	crime	involves crime	2621	16333
8-shelter	shelter	describes a situation where people need shelter	2620	16347
9-food	hunger	is related to food security	2642	16276
10-infrastructure	infrastructure	is related to infrastructure	2664	16548
11-regime change	regime change	describes a regime change	2670	16382
12-medical	health	is related to a medical situation	2675	16223
13-water	water	involves a situation where people need clean water	2619	16135
14-search	rescue	involves a search/rescue situation	2628	16131
15-utility	utility	expresses need for utility, energy or sanitation	2640	16249
16-hillary	Hillary	is against Hillary	224	1693
17-hillary	Hillary	supports hillary	218	1675
18-offensive	derogatory	contains offensive content	652	6109
19-offensive	toxic	insult women or immigrants	2188	11839
20-pro-life	pro-life	is pro-life	213	1633
21-pro-choice	abortion	supports abortion	209	1593
22-physics	physics	is about physics	10360	93810
23-computer science	computers	is related to computer science	10441	93947
24-statistics	statistics	is about statistics	9286	86874
25-math	math	is about math research	8898	85118
26-grammar	ungrammatical	is ungrammatical	834	2217
27-grammar	grammatical	is grammatical	826	2236
28-sexis	sexist	is offensive to women	209	1641
29-sexis	feminism	supports feminism	215	1710
30-news	world	is about world news	5778	13023
31-sports	sports news	is about sports news	5674	12849
32-business	business	is related to business	5699	12913
33-tech	technology	is related to technology	5727	12927
34-bad	negative	contains a bad movie review	357	16889
35-good	good	thinks the movie is good	380	17497
36-quantity	quantity	asks for a quantity	1901	5144
37-location	location	asks about a location	1925	5236
38-person	person	asks about a person	1848	5014
39-entity	entity	asks about an entity	1896	5180
40-abbrevation	abbreviation	asks about an abbreviation	1839	5045
41-defin	definition	contains a definition	651	4508
42-environment	environmentalism	is against environmentalist	124	1117
43-environment	environmentalism	is environmentalist	119	1072
44-spam	spam	is a spam	360	2470
45-fact	facts	asks for factual information	704	11449
46-opinion	opinion	asks for an opinion	719	11709
47-math	science	is related to math and science	7514	53973
48-health	health	is related to health	7485	53986
49-computer	computers	related to computer or internet	7486	54256
50-sport	sports	is related to sports	7505	54718
51-entertainment	entertainment	is about entertainment	7461	53573
52-family	relationships	is about family and relationships	7438	54680
53-politic	politics	is related to politics or government	7410	53393

Table A3: 54 synthetic datasets and the regex used to check whether an explanation is correct (after applying lowercasing). These regexes form guide the manual inspection of explanation accuracy: the original label is assigned by the regex and then fixed by the human when errors (which are relatively rare) occur.

<table>
<tr><th>Module name</th><th>Dataset explanation</th><th>Regex check</th></tr>
<tr><td>0-irony</td><td>contains irony</td><td>irony|sarcas</td></tr>
<tr><td>1-objective</td><td>is a more objective description of what happened</td><td>objective|factual|nonpersonal|neutral|unbias</td></tr>
<tr><td>2-subjective</td><td>contains subjective opinion</td><td>subjective|opinion|personal|bias</td></tr>
<tr><td>3-god</td><td>believes in god</td><td>god|religious|religion</td></tr>
<tr><td>4-atheism</td><td>is against religion</td><td>atheism|atheist|anti-religion|against religion</td></tr>
<tr><td>5-evacuate</td><td>involves a need for people to evacuate</td><td>evacuat|flee|escape</td></tr>
<tr><td>6-terorrism</td><td>describes a situation that involves terrorism</td><td>terorrism|terror</td></tr>
<tr><td>7-crime</td><td>involves crime</td><td>crime|criminal|criminality</td></tr>
<tr><td>8-shelter</td><td>describes a situation where people need shelter</td><td>shelter|home|house</td></tr>
<tr><td>9-food</td><td>is related to food security</td><td>food|hunger|needs</td></tr>
<tr><td>10-infrastructure</td><td>is related to infrastructure</td><td>infrastructure|buildings|roads|bridges|build</td></tr>
<tr><td>11-regime change</td><td>describes a regime change</td><td>regime change|coup|revolution|revolt|political action|political event|upheaval</td></tr>
<tr><td>12-medical</td><td>is related to a medical situation</td><td>medical|health</td></tr>
<tr><td>13-water</td><td>involves a situation where people need clean water</td><td>water</td></tr>
<tr><td>14-search</td><td>involves a search/rescue situation</td><td>search|rescue|help</td></tr>
<tr><td>15-utility</td><td>expresses need for utility, energy or sanitation</td><td>utility|energy|sanitation|electricity|power</td></tr>
<tr><td>16-hillary</td><td>is against Hillary</td><td>hillary|clinton|against Hillary|opposed to Hillary|republican|against Clinton|opposed to Clinton</td></tr>
<tr><td>17-hillary</td><td>supports hillary</td><td>hillary|clinton|support Hillary|support Clinton|democrat</td></tr>
<tr><td>18-offensive</td><td>contains offensive content</td><td>offensive|toxic|abusive|insulting|insult|abuse|offend|offend|derogatory</td></tr>
<tr><td>19-offensive</td><td>insult women or immigrants</td><td>offensive|toxic|abusive|insulting|insult|abuse|offend|offend|women|immigrants</td></tr>
<tr><td>20-pro-life</td><td>is pro-life</td><td>pro-life|abortion|pro life</td></tr>
<tr><td>21-pro-choice</td><td>supports abortion</td><td>pro-choice|abortion|pro choice</td></tr>
<tr><td>22-physics</td><td>is about physics</td><td>physics</td></tr>
<tr><td>23-computer science</td><td>is related to computer science</td><td>computer science|computer|artificial intelligence|ai</td></tr>
<tr><td>24-statistics</td><td>is about statistics</td><td>statistics|stat|probability</td></tr>
<tr><td>25-math</td><td>is about math research</td><td>math|arithmetic|algebra|geometry</td></tr>
<tr><td>26-grammar</td><td>is ungrammatical</td><td>grammar|syntax|punctuation|grammat|linguistic</td></tr>
<tr><td>27-grammar</td><td>is grammatical</td><td>grammar|syntax|punctuation|grammat|linguistic</td></tr>
<tr><td>28-sexis</td><td>is offensive to women</td><td>sexis|women|femini</td></tr>
<tr><td>29-sexis</td><td>supports feminism</td><td>sexis|women|femini</td></tr>
<tr><td>30-news</td><td>is about world news</td><td>world|cosmopolitan|international|global</td></tr>
<tr><td>31-sports</td><td>is about sports news</td><td>sports</td></tr>
<tr><td>32-business</td><td>is related to business</td><td>business|economics|finance</td></tr>
<tr><td>33-tech</td><td>is related to technology</td><td>tech</td></tr>
<tr><td>34-bad</td><td>contains a bad movie review</td><td>bad|negative|awful|terrible|horrible|poor|boring|dislike</td></tr>
<tr><td>35-good</td><td>thinks the movie is good</td><td>good|great|like|love|positive|awesome|amazing|excellent</td></tr>
<tr><td>36-quantity</td><td>asks for a quantity</td><td>quantity|number|numeric</td></tr>
<tr><td>37-location</td><td>asks about a location</td><td>location|place</td></tr>
<tr><td>38-person</td><td>asks about a person</td><td>person|individual|people</td></tr>
<tr><td>39-entity</td><td>asks about an entity</td><td>entity|thing|object</td></tr>
<tr><td>40-abbrevation</td><td>asks about an abbreviation</td><td>abbrevation|abbr|acronym</td></tr>
<tr><td>41-defin</td><td>contains a definition</td><td>defin|meaning|explain</td></tr>
<tr><td>42-environment</td><td>is against environmentalist</td><td>environment|climate change|global warming</td></tr>
<tr><td>43-environment</td><td>is environmentalist</td><td>environment|climate change|global warming</td></tr>
<tr><td>44-spam</td><td>is a spam</td><td>spam|annoying|unwanted</td></tr>
<tr><td>45-fact</td><td>asks for factual information</td><td>fact|info|knowledge</td></tr>
<tr><td>46-opinion</td><td>asks for an opinion</td><td>opinion|personal|bias</td></tr>
<tr><td>47-math</td><td>is related to math and science</td><td>math|science</td></tr>
<tr><td>48-health</td><td>is related to health</td><td>health|medical|disease</td></tr>
<tr><td>49-computer</td><td>related to computer or internet</td><td>computer|internet|web</td></tr>
<tr><td>50-sport</td><td>is related to sports</td><td>sport</td></tr>
<tr><td>51-entertainment</td><td>is about entertainment</td><td>entertainment|music|movie|tv</td></tr>
<tr><td>52-family</td><td>is about family and relationships</td><td>family|relationships</td></tr>
<tr><td>53-politic</td><td>is related to politics or government</td><td>politic|government|law</td></tr>
</table>

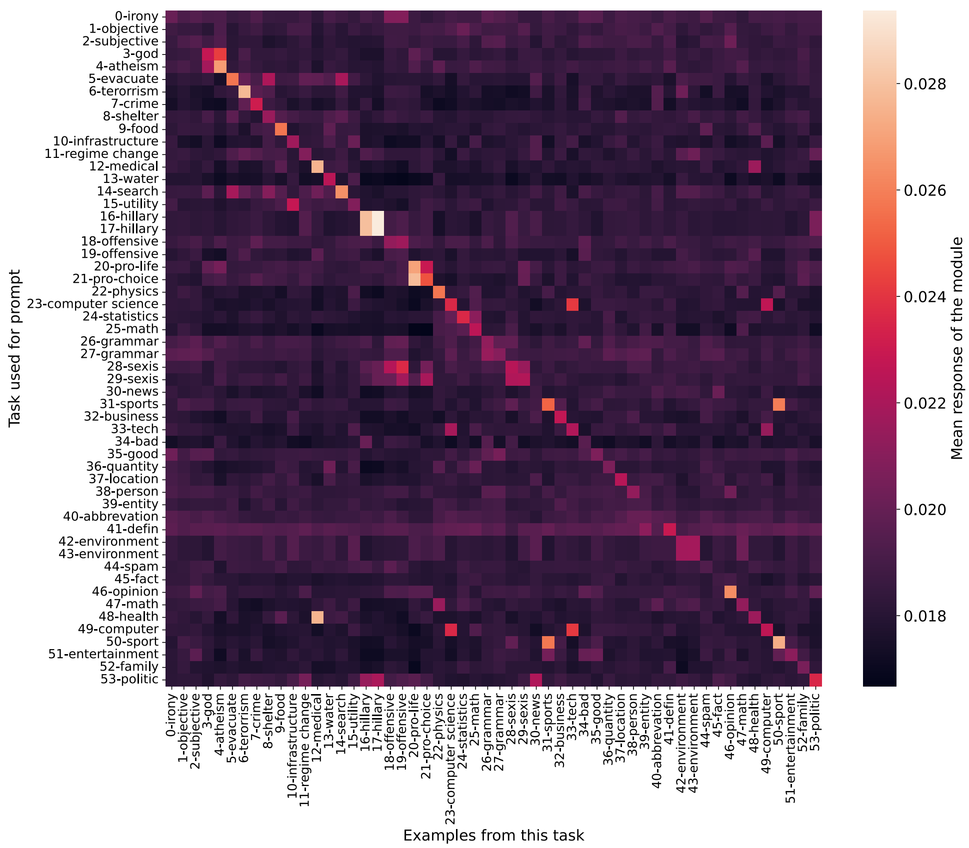

Figure A3: Synthetic modules respond more strongly to phrases related to their keyphrase (diagonal) than to phrases related to the keyphrase of other datasets (off-diagonal). Each value shows the mean response of the module to 5 phrases and each row is normalized using softmax. Each module is constructed using Instructor (Su et al., 2022) with the prompt *Represent the short phrase for clustering:* and the groundtruth keyphrase given in Table A2. Related keyphrases are generated manually.

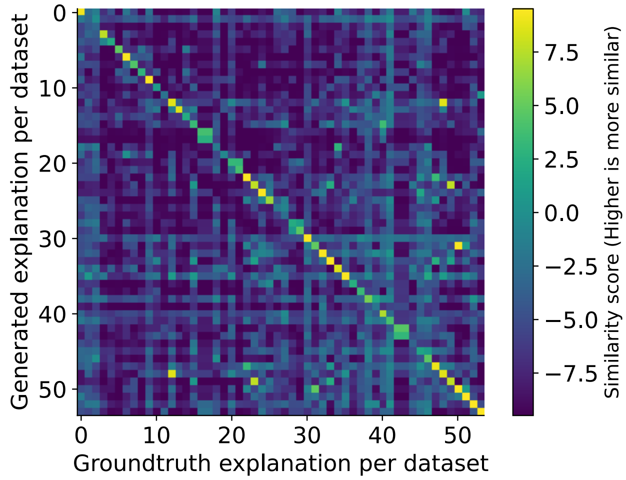

Figure A4: Similarity scores for SASC explanations in the *Default* setting measured by bge-large (`BAAI/bge-large-en`, (Zhang et al., 2023)), rather than manual inspection or BERT-score, as shown in Table 1. Large values on the diagonal indicate that the explanation generated for a module on a given dataset are similar to the groundtruth explanations for that dataset. The top-1 classification accuracy (i.e. how often the generated explanation is most similar to its corresponding groundtruth explanation) is 81.5%, slightly lower than the assigned accuracy by manual inspection (88.3%). The top-2 accuracy is 88.9%.

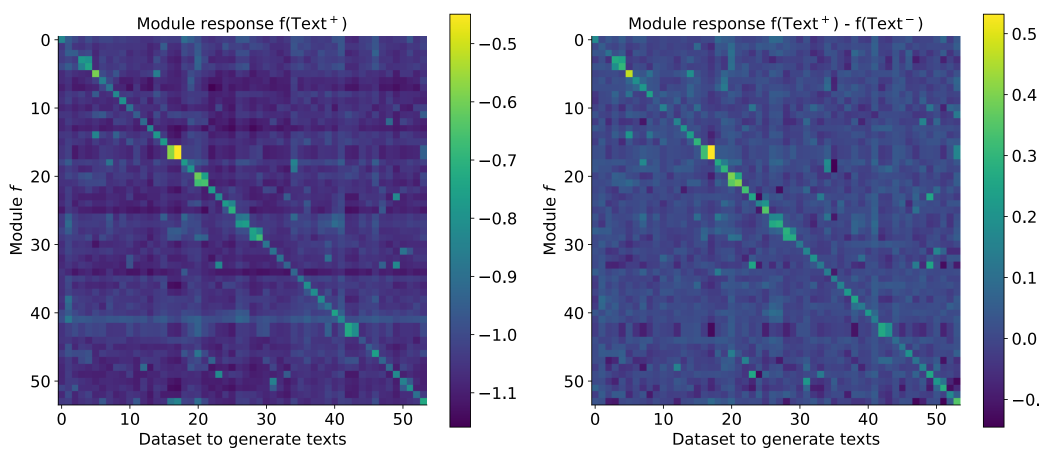

Figure A5: Average module responses for synthetic texts that are related to the explanation (left, $f(\text{Text}^{+})$) or the difference between the responses for related and unrelated texts (right, $f(\text{Text}^{+} - f(\text{Text}^{-}))$. Responses correspond to synthetic modules in the *Default* setting. Bright diagonal on the left suggests that f selectively responses to synthetic texts generated according to the appropriate explanation. On the right, the diagonal is slightly less bright, suggesting that the module does not tend to respond more negatively to unrelated texts Text^{-}.

A.3 BERT INTERPRETATION

Details on fitting transformer factors Pre-trained transformer factors are taken from (Yun et al., 2021). Each transformer factor is the result of running dictionary learning on a matrix X described as follows. Using a corpus of sentences S (here wikipedia), embeddings are extracted for each input, layer, and sequence index in BERT. The resulting matrix X has size $\left(\underbrace{\text{num\_layers}}_{\text{13 for BERT}} \cdot \sum_{s \in S} \text{len(s)}\right) \times \underbrace{d}_{\text{768 for BERT}}$. Dictionary learning is run on X with 1,500 dictionary components, resulting in a dictionary $D \in \mathbb{R}^{1,500 \times d}$. Here, we take the fitted dictionary released by (Yun et al., 2021) trained on the WikiText dataset (Merity et al., 2016).

During our interpretation pipeline, we require a module which maps text to a scalar coefficient. To interpret a transformer factor as a module, we specify a text input t and a layer l. This results in $len(t)$ embeddings with dimension d. We average over these embeddings, and then solve for the dictionary coefficients, to yield a set of coefficients $A \in \mathbb{R}^{1500}$. Finally, specifying a dictionary component index yields a single, scalar coefficient.

Extended BERT explanation results Table A4 shows examples comparing SASC explanations with human-labeled explanations for all BERT transformer factors labeled in (Yun et al., 2021). Tables A6 to A8 show explanations for modules selected by linear models finetuned on text-classification tasks.

Table A4: Fraction of top logistic regression coefficients that are relevant for a downstream task (extends Table 5). Averaged over 3 random seeds; parentheses show standard error of the mean.

	Emotion	AG News	SST2
Top-10	0.50 $\pm$0.08	1.00 $\pm$0.00	0.80 $\pm$0.14
Top-15	0.47 $\pm$0.05	0.98 $\pm$0.03	0.69 $\pm$0.13
Top-20	0.42 $\pm$0.09	0.98 $\pm$0.02	0.55 $\pm$0.10

Table A5: Comparing SASC explanations to all human-labeled explanations for BERT transformer factors. Explanation scores are in units of σ_f.

Factor Layer	Factor Index	Explanation (Human)	Explanation (SASC)	Explanation score (Human)	Explanation score (SASC)
4	13	Numerical values.	numbers	-0.21	-0.08
10	42	Something unfortunate happened.	idea of wrongdoing or illegal activity	2.43	1.97
0	30	left. Adjective or Verb. Mixed senses.	someone or something leaving	3.68	5.87
4	47	plants. Noun. vegetation.	trees	6.26	5.04
10	152	In some locations.	science, technology, and/or medicine	-0.41	0.03
4	30	left. Verb. leaving, exiting.	leaving or being left	4.44	0.90
10	297	Repetitive structure detector.	versions or translations	-0.36	0.98
10	322	Biography, someone born in some year...	weapons and warfare	0.19	0.38
10	13	Unit exchange with parentheses.	names of places, people, or things	-0.11	-0.10
10	386	War.	media, such as television, movies, or video games	0.20	-0.15
10	184	Institution with abbreviation.	publishing, media, or awards	-0.42	0.14
2	30	left. Verb. leaving, exiting.	leaving or being left	5.30	0.91
10	179	Topic: music production.	geography	-0.52	0.21
6	225	Places in US, followings the convention ”city, state”.	a place or location	1.88	1.33
10	25	Attributive Clauses.	something related to people, places, or things	0.01	1.19
10	125	Describing someone in a para- phrasing style. Name, Career.	something related to buildings, architecture, or construction	-0.13	0.44
6	13	Close Parentheses.	end with a closing punctuation mark (e.g	-0.08	0.47
10	99	Past tense.	people, places, or things	-0.77	-0.04
10	24	Male name.	people, places, and things related to history	0.03	0.38
10	102	African names.	traditional culture, with references to traditional territories, communities, forms, themes, breakfast, and texts	0.35	1.60
4	16	park. Noun. a common first and last name.	names of parks	-0.03	1.87
10	134	Transition sentence.	a comma	1.16	0.38
6	86	Consecutive years, used in foodball season naming.	specific dates or months	0.85	0.76
4	2	mind. Noun. the element of a person that enables them to be aware of the world and their experiences.	concept of thinking, remembering, and having memories	0.77	11.19
10	51	Apostrophe s, possesive.	something specific, such as a ticket, tenure, film, song, movement, project, game, school, title, park, congressman, author, or art exhibition	0.37	-0.01
8	125	Describing someone in a paraphrasing style. Name, Career.	publications, reviews, or people associated with the media industry	-0.34	0.42
4	33	light. Noun. the natural agent that stimulates sight and makes things visible.	light	6.25	3.43
10	50	Doing something again, or making something new again.	introduction of something new	0.84	-0.27
10	86	Consecutive years, this is convention to name foodball/rugby game season.	a specific date or time of year	1.35	-0.75
4	193	Time span in years.	many of them are related to dates and historic places	0.07	1.39
10	195	Consecutive of noun (Enumerating).	different aspects of culture, such as art, music, literature, history, and technology	-0.83	9.83

Table A6: SASC explanations for modules selected by 25-coefficient linear model on *SST2* for a single seed. Green shows explanations deemed to be relevant to the task.

Layer, Factor index	Explanation	Linear coefficient
(0, 783)	something being incorrect or wrong	-862.82
(0, 1064)	negative emotions and actions, such as hatred, violence, and disgust	-684.27
(1, 783)	something being incorrect, inaccurate, or wrong	-577.49
(1, 1064)	hatred and violence	-499.30
(0, 157)	air and sequencing	463.80
(9, 319)	a negative statement, usually in the form of not or nor	-446.58
(0, 481)	harm, injury, or damage	-441.98
(8, 319)	lack of something or the absence of something	-441.04
(10, 667)	two or more words	424.48
(2, 783)	something that is incorrect or inaccurate	-415.56
(0, 658)	thrice	-411.26
(0, 319)	none or its variations (no, not, never)	-388.14
(0, 1402)	dates	-377.74
(0, 1049)	standard	-365.83
(3, 1064)	negative emotions or feelings, such as hatred, anger, disgust, and brutality	-360.47
(4, 1064)	negative emotions or feelings, such as hatred, anger, and disgust	-357.35
(5, 152)	geography, history, and culture	-356.10
(0, 928)	homelessness and poverty	-355.05
(2, 691)	animals and plants, as many of the phrases refer to species of animals and plants	-351.62
(0, 810)	catching or catching something	350.98
(0, 1120)	production	-350.01
(0, 227)	a period of time	-345.72
(2, 583)	government, law, or politics in some way	-335.40
(2, 1064)	negative emotions such as hatred, disgust, and violence	-334.87
(4, 125)	science or mathematics, such as physics, astronomy, and geometry	-328.55

Table A7: SASC explanations for modules selected by 25-coefficient linear model on *AG News* for a single seed. Green shows explanations deemed to be relevant to the task.

Layer, Factor index	Explanation	Linear coefficient
(5, 378)	professional sports teams	545.57
(4, 378)	professional sports teams in the united states	542.25
(3, 378)	professional sports teams	515.37
(0, 378)	names of sports teams	508.73
(6, 378)	sports teams	499.62
(2, 378)	professional sports teams	499.57
(1, 378)	professional sports teams	492.01
(7, 378)	sports teams	468.66
(8, 378)	sports teams or sports in some way	468.39
(11, 32)	activity or process	461.46
(12, 1407)	such	450.70
(5, 730)	england and english sports teams	427.33
(12, 104)	people, places, and events from history	425.49
(10, 378)	locations	424.71
(6, 730)	sports, particularly soccer	424.24
(12, 730)	sports	415.21
(4, 396)	people, places, or things related to japan	-415.13
(10, 659)	sports	410.89
(4, 188)	history in some way	404.24
(12, 1465)	different aspects of life, such as activities, people, places, and objects	403.77
(0, 310)	end with the word until	-400.10
(5, 151)	a particular season, either of a year, a sport, or a television show	396.41
(12, 573)	many of them contain unknown words or names, indicated by <unk	-393.27
(12, 372)	specific things, such as places, organizations, or activities	-392.57
(6, 188)	geography	388.69

Table A8: SASC explanations for modules selected by 25-coefficient linear model on *Emotion* for a single seed. Green shows explanations deemed to be relevant to the task.

Layer, Factor index	Explanation	Linear coefficient
(0, 1418)	types of road interchanges	581.97
(0, 920)	fame	577.20
(6, 481)	injury or impairment	566.44
(5, 481)	injury or impairment	556.58
(0, 693)	end in oss or osses	556.53
(12, 1137)	ownership or possession	-537.45
(0, 663)	civil	524.88
(6, 1064)	negative emotions such as hatred, disgust, disdain, rage, and horror	523.41
(3, 872)	location of a campus or facility	-518.85
(5, 1064)	negative emotions and feelings, such as hatred, disgust, disdain, and viciousness	489.25
(0, 144)	lectures	482.85
(0, 876)	host	479.18
(0, 69)	history	-467.80
(0, 600)	many of them contain the word seymour or a variation of it	464.64
(0, 813)	or phrases related to either measurement (e.g	-455.11
(1, 89)	caution and being careful	451.73
(11, 229)	russia and russian culture	-450.28
(0, 783)	something being incorrect or wrong	448.55
(12, 195)	dates	442.14
(12, 1445)	breaking or being broken	439.81
(0, 415)	ashore	-438.22
(0, 118)	end with a quotation mark	437.66
(1, 650)	mathematical symbols such as $>$, =, and)	-437.28
(4, 388)	end with the sound ch	-437.15
(0, 840)	withdrawing	-436.38

A.4 fMRI MODULE INTERPRETATION

A.4.1 fMRI DATA AND MODEL FITTING

This section gives more details on the fMRI experiment analyzed in Sec. 5. These MRI data are available publicly (LeBel et al., 2022; Tang et al., 2023), but the methods are summarized here. Functional magnetic resonance imaging (fMRI) data were collected from 3 human subjects as they listened to English language podcast stories over Sensimetrics S14 headphones. Subjects were not asked to make any responses, but simply to listen attentively to the stories. For encoding model training, each subject listened to at approximately 20 hours of unique stories across 20 scanning sessions, yielding a total of ~33,000 datapoints for each voxel across the whole brain. For model testing, the subjects listened to two test story 5 times each, and one test story 10 times, at a rate of 1 test story per session. These test responses were averaged across repetitions. Functional signal-to-noise ratios in each voxel were computed using the mean-explainable variance method from (Nishimoto et al., 2017) on the repeated test data. Only voxels within 8 mm of the mid-cortical surface were analyzed, yielding roughly 90,000 voxels per subject.

MRI data were collected on a 3T Siemens Skyra scanner at University of Texas at Austin using a 64-channel Siemens volume coil. Functional scans were collected using a gradient echo EPI sequence with repetition time (TR) = 2.00 s, echo time (TE) = 30.8 ms, flip angle = 71°, multi-band factor (simultaneous multi-slice) = 2, voxel size = 2.6mm x 2.6mm x 2.6mm (slice thickness = 2.6mm), matrix size = 84x84, and field of view = 220 mm. Anatomical data were collected using a T1-weighted multi-echo MP-RAGE sequence with voxel size = 1mm x 1mm x 1mm following the Freesurfer morphometry protocol (Fischl, 2012).

All subjects were healthy and had normal hearing. The experimental protocol was approved by the Institutional Review Board at the University of Texas at Austin. Written informed consent was obtained from all subjects.

All functional data were motion corrected using the FMRIB Linear Image Registration Tool (FLIRT) from FSL 5.0. FLIRT was used to align all data to a template that was made from the average across the first functional run in the first story session for each subject. These automatic alignments were manually checked for accuracy.

Low frequency voxel response drift was identified using a 2nd order Savitzky-Golay filter with a 120 second window and then subtracted from the signal. To avoid onset artifacts and poor detrending performance near each end of the scan, responses were trimmed by removing 20 seconds (10 volumes) at the beginning and end of each scan, which removed the 10-second silent period and the first and last 10 seconds of each story. The mean response for each voxel was subtracted and the remaining response was scaled to have unit variance.

We used the fMRI data to generate a voxelwise brain encoding model for natural language using the intermediate hidden states from the the 18th layer of the 30-billion parameter LLaMA model (Touvron et al., 2023a), and the 9th layer of GPT (Radford et al., 2019). In order to temporally align word times with TR times, Lanczos interpolation was applied with a window size of 3. The hemodyanmic response function was approximated with a finite impulse response model using 4 delays at -8,-6,-4 and -2 seconds (Huth et al., 2016). For each subject x, voxel v, we fit a separate encoding model $g_{(x,v)}$ to predict the BOLD response $\hat{B}$ from our embedded stimulus, i.e. $\hat{B}_{(x,v)} = g_{(x,v)}(H_i(\mathcal{S}))$.

To evaluate the voxelwise encoding models, we used the learned $g_{(x,v)}$ to generate and evaluate predictions on a held-out test set. The GPT features achieved a mean correlation of 0.12 and LLaMA features achieved a mean correlation of 0.17. These performances are comparable with state-of-the-art published models on the same dataset that are able to achieved decoding (Tang et al., 2023).

To select voxels with diverse encoding, we applied principal components analysis to the learned weights, $g_{(x,v)}$, for GPT across all significantly predicted voxels in cortex. Prior work has shown that the first four principal components of language encoding models weights encode differences in semantic selectivity, differentiating between concepts like *social*, *temporal* and *visual* concepts. Consequently, to apply SASC to voxels with the most diverse selectivity, we found voxels that lie along the convex hull of the first four principal components and randomly sampled 1,500 of them (500 per subject). The mean voxel correlation for the 1,500 voxels we study is 0.35. Note that these voxels were selected for being well-predicted rather than easy to explain: the correlation between the prediction error and the explanation score for these voxels is 0.01, very close to zero.

A.4.2 EVALUATING TOP FMRI VOXEL EVALUATIONS

Table A9 shows two evaluations of the fMRI voxel explanations. First, similar to Fig. 3, we find the mean explanation score remains significantly above zero. Second, we evaluate beyond whether the explanation describes the fitted module and ask whether the explanation describes the underlying fMRI voxel. Specifically, we predict the fMRI voxel response to text using only the voxel's explanation using a very simple procedure. We first compute the (scalar) negative embedding distance between the explanation text and the input text using Instructor (Su et al., 2022)[5]. We then calculate the spearman rank correlation between this scalar distance and the recorded voxel response (see Table A9). The mean computed correlation is low[6], which is to be expected as the explanation is a concise string and may match extremely few ngrams in the text of the test data (which consists of only 3 narrative stories). Nevertheless, the correlation is significantly above zero (more than 15 times the standard error of the mean), suggesting that these explanations have some grounding in the underlying brain voxels.

Table A9: Evaluation of fMRI voxel explanations. For all metrics, SASC is successful if the value is significantly greater than 0. Errors show standard error of the mean.

Explanation score	Test rank correlation
$1.27\sigma_f$ ± 0.029	0.033 ± 0.002

A.4.3 FMRI RESULTS WHEN USING WIKITEXT CORPUS

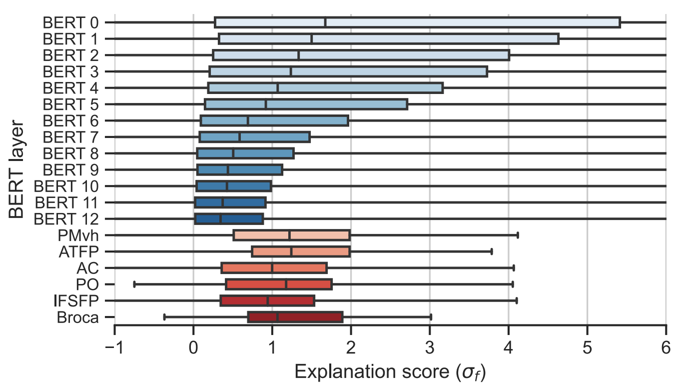

Figure A6: Results in Fig. 3 when using WikiText as the underlying corpus for ngrams rather than narrative stories.

[5]The input text for an fMRI response at time t (in seconds) is taken to be the words presented between $t - 8$ and $t - 2$.

[6]For reference, test correlations published in fMRI voxel prediction from language are often in the range of 0.01-0.1 (Caucheteux et al., 2022).

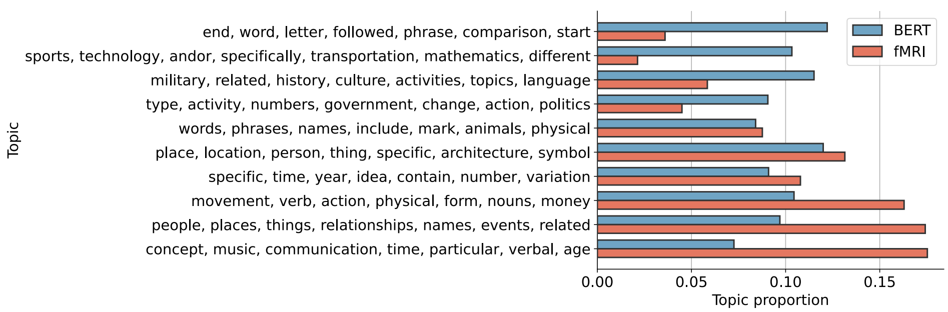

Figure A7: Results in Fig. 4 when using WikiText as the underlying corpus for ngrams rather than narrative stories.