Explaining black box text modules
in natural language with language models

Chandan Singh[*,1] **Aliyah R. Hsu**[*,1,2] **Richard Antonello**[3] **Shailee Jain**[3]

Alexander G. Huth[3] **Bin Yu**[1] **Jianfeng Gao**[1]

[1] Microsoft Research
[2] University of California, Berkeley
[3] The University of Texas at Austin
* Equal contribution

Abstract

Large language models (LLMs) have demonstrated remarkable prediction performance for a growing array of tasks. However, their rapid proliferation and increasing opaqueness have created a growing need for interpretability. Here, we ask whether we can automatically obtain natural language explanations for black box text modules. A *text module* is any function that maps text to a scalar continuous value, such as a submodule within an LLM or a fitted model of a brain region. *Black box* indicates that we only have access to the module's inputs/outputs. We introduce Summarize and Score (SASC), a method that takes in a text module and returns a natural language explanation of the module's selectivity along with a score for how reliable the explanation is. We study SASC in 2 contexts. First, we evaluate SASC on synthetic modules and find that it often recovers ground truth explanations. Second, we use SASC to explain modules found within a pre-trained BERT model, enabling inspection of the model's internals. All code for using SASC and reproducing results is made available on Github.
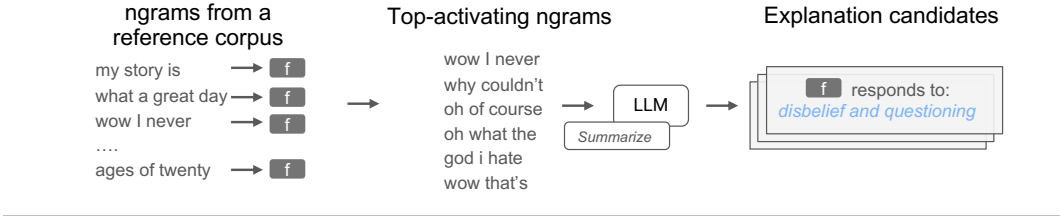
1 Introduction

Large language models (LLMs) have demonstrated remarkable predictive performance across a growing range of diverse tasks [1, 2]. However, the inability to effectively interpret these models has led them to be characterized as black boxes. This opaqueness has debilitated their use in high-stakes applications such as medicine [3], and raised issues related to regulatory pressure [4], safety [5], and alignment [6]. This lack of interpretability is particularly detrimental in scientific fields, where trustworthy interpretation itself is the end goal [7].

To ameliorate these issues, we propose Summarize and Score (SASC). SASC produces *natural language explanations for text modules*. We define a *text module f* as any function that maps text to a scalar continuous value, e.g. a neuron in a pre-trained LLM[1]. Given f, SASC returns a short natural language explanation describing what elicits the strongest response from f. SASC requires only black-box access to the module (it does not require access to the module internals) and no human intervention.

---

[1] Note that a neuron in an LLM typically returns a sequence-length vector rather than a scalar, so a transformation (e.g. averaging) is required before interpretation.

(1) Summarize ngrams into candidate explanations for module f

ngrams from a
reference corpus

my story is → f
what a great day → f
wow I never → f
....
ages of twenty → f

Top-activating ngrams

wow I never
why couldn't
oh of course
oh what the
god i hate
wow that's

LLM
Summarize

Explanation candidates

f responds to:
disbelief and questioning

(2) Score candidate explanations

Generate synthetic data

Explanation candidates

f responds to:
disbelief and questioning

LLM
Generate

Texts⁺
Related to explanation

How could you?
I Can't believe it
What do you mean!

Texts⁻
Not related

I agree
That sounds good
Nice idea

Explanation score

f → Maximize
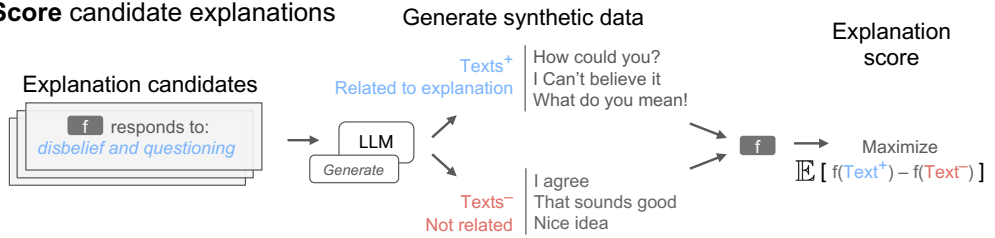$\mathbb{E}\,[\,f(\text{Text}^+) - f(\text{Text}^-)\,]$

Figure 1: SASC pipeline for obtaining a natural language explanation given a module f. **(i)** SASC first generates candidate explanations (using a pre-trained LLM) based on the ngrams that elicit the most positive response from f. **(ii)** SASC then evaluates each candidate explanation by generating synthetic data based on the explanation and testing the response of f to the data.

SASC uses two steps to ground explanations in the responses of f (Fig. 1). In the first step, SASC derives explanation candidates by sorting f's responses to ngrams and summarizing the top ngrams using a pre-trained LLM. In the second step, SASC evaluates each candidate explanation by generating synthetic text based on the explanation (again with a pre-trained LLM) and testing the response of f to the text; these responses to synthetic text are used to assign an *explanation score* to each explanation, that rates the reliability of the explanation. Decomposing explanation into these separate steps helps mitigate issues with LLM hallucination when generating and evaluating explanations.

We evaluate SASC in two contexts. In our main evaluation, we evaluate SASC on synthetic modules and find that it can often recover ground truth explanations under different experimental conditions (Sec. 3). In our second evaluation, we use SASC to explain modules found within a pre-trained BERT model after applying dictionary learning (details in Sec. 4), and find that SASC explanations are often of comparable quality to human-given explanations (without the need for manual annotation). Furthermore, we verify that BERT modules which are useful for downstream text-classification tasks often yield explanations related to the task.

2   Method

SASC aims to interpret a text module f, which maps text to a scalar continuous value. For example f could be the output probability for a single token in an LLM, or the output of a single neuron extracted from a vector of LLM activations. SASC returns a short explanation describing what elicits the strongest response from f, along with an *explanation score*, which rates how reliable the explanation is. In the process of explanation, SASC uses a pre-trained *helper LLM* to perform summarization and to generate synthetic text. To mitigate potential hallucination introduced by the helper LLM, SASC decomposes the explanation process into 2 steps (Fig. 1) that greatly simplify the task performed by the helper LLM:

Step 1: Summarization   The first step generates candidate explanations by summarizing ngrams. All unique ngrams are extracted from a pre-specified corpus of text and fed through the module f. The ngrams that elicit the largest positive response from f are then fed through the helper LLM for summarization. To avoid over-reliance on the very top ngrams, we select a random subset of the top ngrams in the summarization step. This step is similar to prior works which summarize ngrams using manual inspection/parse trees [8, 9], but the use of the helper LLM enables flexible, automated summarization.

The summarization step requires 2 choices: the corpus underlying the extracted ngrams, and the length of ngrams to extract. Using a larger corpus/higher order ngrams can make SASC more accurate, but the computational cost grows linearly with the unique number of ngrams in the corpus. The corpus should be large enough to include relevant ngrams, as the corpus limits what generated explanations are possible. (e.g. it is difficult to recover mathematical explanations from a corpus that contains no math).

Step 2: Synthetic scoring The second step aims to evaluate each candidate explanation and select the most reliable one. SASC generates synthetic data based on each candidate explanation, again using the helper LLM. Intuitively, if the explanation accurately describes f, then f should output large values for text related to the explanation ($Text^+$) compared to unrelated synthetic text ($Text^-$). We thus compute the explanation score as follows:

$$\text{Explanation score} = \mathbb{E}[f(Text^+) - f(Text^-)] \text{ with units } \sigma_f, \quad (1)$$

where a larger score corresponds to a more reliable explanation. We report the score in units of σ_f, the standard deviation of the module's response to the corpus. An explanation score of $1\sigma_f$ means that synthetic text related to the explanation increased the mean module response by one standard deviation compared to unrelated text. SASC returns the candidate explanation that maximizes this difference, along with the synthetic data score.

Limitations and hyperparameter settings While effective, the explanation pipeline described here has some clear limitations. First and foremost, SASC assumes that f can be concisely described in a natural language string. This excludes complex functions or modules that respond to a non-coherent set of inputs. Second, SASC only describes the inputs that elicit the largest responses from f, rather than its full behavior. Finally, SASC requires that the pre-trained LLM can faithfully perform its required tasks (summarization and generation). If an LLM is unable to perform these tasks sufficiently well, users may treat the output of SASC as candidate explanations to be vetted by a human, rather than final explanations to be used.

We use GPT-3 (`text-davinci-003`, Feb. 2023) [1] as the helper LLM (see LLM prompts in Appendix A.2). In the summarization step, we use word-level trigrams, choose 30 random ngrams from the top 50 and generate 5 candidate explanations. In the synthetic scoring step, we generate 20 synthetic strings (each is a sentence) for each candidate explanation, half of which are related to the explanation.

3 Recovering ground truth explanations for synthetic modules

Experimental setup for synthetic modules We construct 54 synthetic modules based on the pre-trained Instructor embedding model [10] (`hkunlp/instructor-xl`). Each module is based on a dataset from a recent diverse collection [11, 12] that admits a simple, verifiable keyphrase description describing each underlying dataset, e.g. *related to math* (full details in Table A3). Each module is constructed to return high values for text related to the module's groundtruth keyphrase and low values otherwise. Specifically, the module computes the Instructor embedding for an input text and for the groundtruth keyphrase; it then returns the negative Euclidean distance between the embeddings. The synthetic modules reliably produce large values for text related to the desired keyphrase (Fig. A2).

We test SASC's ability to recover accurate explanations for each of our 54 modules in 3 settings: (1) The *Default* setting extracts ngrams for summarization from the dataset corresponding to each module, which contains relevant ngrams for the ground truth explanation. (2) The *Restricted corpus* setting checks the impact of the underlying corpus on the performance of SASC. To do so, we restrict the ngrams we use for generating explanation candidates to a corpus from a random dataset among the 54, potentially containing less relevant ngrams. (3) The *Noisy module* setting adds Gaussian noise to all module responses in the summarization step; the standard deviation of the added noise is set to $3\sigma_f$.

SASC can recover ground truth descriptions Table 1 shows the performance of SASC at recovering ground truth explanations in terms of accuracy (calculated by verifying whether the ground

Table 1: Explanation recovery performance. For both metrics, higher is better. Each value is averaged over 54 modules and 3 random seeds; errors show standard error of the mean.

| | **SASC** | | Baseline (ngram summarization) | |
	Accuracy	BERT Score	Accuracy	BERT Score
Default	0.883 ±0.03	0.712 ±0.02	0.753 ±0.02	0.622 ±0.05
Restricted corpus	0.667 ±0.04	0.639 ±0.02	0.540 ±0.02	0.554 ±0.05
Noisy module	0.679 ±0.04	0.669 ±0.02	0.456 ±0.02	0.565 ±0.06

Table 2: Examples of recovered explanations for different modules in the *Default* setting.

	Groundtruth Explanation	SASC Explanation
Correct	atheistic	atheism and related topics, such as theism, religious beliefs, and atheists
	environmentalism	environmentalism and climate action
	crime	crime and criminal activity
	sports	sports
	definition	defining or explaining something
	facts	information or knowledge
Incorrect	derogatory	negative language and criticism
	ungrammatical	language
	subjective	art and expression

truth is essentially equivalent to the recovered explanation via manual inspection) and BERT-score [13][2]. In the *Default* setting, SASC successfully identifies 88% of the ground truth explanations. In the two noisy settings, SASC still manages to recover explanations 67% and 68% of the time for the *Restricted ngrams* and *Noisy module* settings, respectively. In all cases, SASC outperforms the baseline method which summarizes ngrams [8, 9] but does not use explanation scores to select among candidate explanations. Table A2 shows the results for the *Default* setting when varying different modeling choices and comparisons to additional baselines.

Table 2 shows examples of correct and incorrect recovered explanations along with the ground truth explanation. For some modules, SASC finds perfect keyword matches, e.g. *sports*, or slight paraphrases, e.g. *definition → defining or explaining something*. For the incorrect examples, the generated explanation is often similar to the ground truth explanation, e.g. *derogatory → negative language and criticism*, but occasionally, SASC fails to correctly identify the underlying pattern, e.g. *ungrammatical → language*. Some failures may be due to the inability of ngrams to capture the underlying explanation, whereas others may be due to the constructed module imperfectly representing the ground truth explanation.

Fig. 2 shows the cumulative accuracy at recovering the ground truth explanation as a function of the explanation score. Across all settings, accuracy increases as a function of explanation score, suggesting that higher explanation scores indicate more reliable explanations. This also helps validate that the helper LLM is able to sucessfully generate useful synthetic texts for evaluation.

4   Generating explanations for BERT transformer factors

Next, we evaluate SASC using explanations for modules within BERT [2] (`bert-base-uncased`). In the absence of ground truth explanations, we evaluate the explanations by (i) comparing them to human-given explanations and (ii) checking their relevance to downstream tasks.

BERT transformer factor modules   One can interpret any module within BERT, e.g. a single neuron or an expert in an MOE [15]; here, we choose to interpret *transformer factors*, following a previous study that suggests that they are amenable to interpretation [16]. Transformer factors learn a transformation of activations across layers via dictionary learning (details in Appendix A.3; corpus used is the WikiText dataset [17]). Each transformer factor is a module that takes as input

---

[2]BERT-score is calculated with the recommended base model `microsoft/deberta-xlarge-mnli` [14].
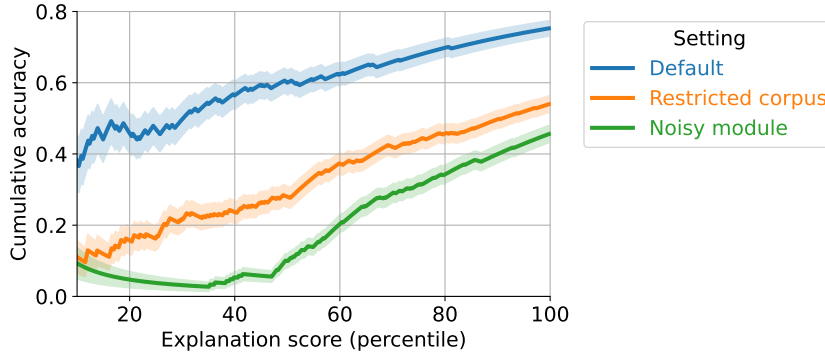
Figure 2: Cumulative accuracy at recovering the ground truth explanation increases as a function of explanation score. Error bars show standard error of the mean.

a text sequence and yields a scalar dictionary coefficient, after averaging over the input's sequence length. There are 1,500 factors, and their coefficients vary for each of BERT's 13 encoding layers.

Comparison to human-given explanations Table 3 compares SASC explanations to those given by humans in prior work (31 unique explanations from Table 1, Table 3 and Appendix in [16]). They are sometimes similar with different phrasings, e.g. *leaving or being left* versus *Word "left"*, and sometimes quite different, e.g. *publishing, media, or awards* versus *Institution with abbreviation*. For each transformer factor, we compare the explanation scores for SASC and the human-given explanations. The SASC explanation score is higher 61% of the time and SASC's mean explanation score is $1.6\sigma_f$ compared to $1.0\sigma_f$ for the human explanation. This evaluation suggests that the SASC explanations can be of similar quality to the human explanations, despite requiring no manual effort.

Table 3: Comparing sample SASC to human-labeled explanations for BERT transformer factors. See all explanations and scores in Table A5.

SASC Explanation	Human Explanation
names of parks	Word "park". Noun. a common first and last name.
leaving or being left	Word "left". Verb. leaving, exiting
specific dates or months	Consecutive years, used in football season naming.
idea of wrongdoing or illegal activity	something unfortunate happened.
introduction of something new	Doing something again, or making something new again.
versions or translations	repetitive structure detector.
SASC win percentage: **61%**	Human explanation win percentage: 39%
SASC mean synthetic score: **$1.6\sigma_f$**	Human explanation mean synthetic score: $1.0\sigma_f$

Mapping explained modules to text-classification tasks We now investigate whether the learned SASC explanations are useful for informing which downstream tasks a module is useful for. Given a classification dataset where the input X is a list of n strings and the output y is a list of n class labels, we first convert X to a matrix of transformer factor coefficients $X_{TF} \in \mathbb{R}^{n \times 19,500}$, where each row contains the concatenated factor coefficients across layers. We then fit a sparse logistic regression model to (X_{TF}, y), and analyze the explanations for the factors with the 25 largest coefficients across all classes. Ideally, these explanations would be relevant to the text-classification task; we evaluate what fraction of the 25 explanations are relevant for each task via manual inspection.

We study 3 widely used text-classification datasets: *emotion* [18] (classifying tweet emotion as sadness, joy, love, anger, fear or surprise), *ag-news* [19] (classifying news headlines as world, sports, business, or sci/tech), and *SST2* [20] (classifying movie review sentiment as positive or negative). Table 4 shows results evaluating the BERT transformer factor modules selected by a sparse linear model fit to these datasets. A large fraction of the explanations for selected modules are, in fact, relevant to their usage in downstream tasks, ranging from 0.35 for *Emotion* to 0.96 for *AG News*.

5

Table 4: BERT modules selected by a sparse linear model fit to text-classification tasks. First row shows the fraction of explanations for the selected modules which are relevant to the downstream task. Second row shows test accuracy for the fitted linear models. Bottom section shows sample explanations for modules selected by the linear model which are relevant to the downstream task. Values are averaged over 3 random linear model fits (error bars show the standard error of the mean).

|	Emotion	AG News	SST2
Fraction relevant	0.35 ± 0.082	0.96 ± 0.033	0.44 ± 0.086
Test accuracy	0.75 ± 0.001	0.81 ± 0.001	0.84 ± 0.001
Sample relevant explanations	negative emotions such as hatred, disgust, disdain, rage, and horror	people, places, or things related to japan	a negative statement, usually in the form of not or nor
	injury or impairment	professional sports teams	hatred and violence
	humor	geography	harm, injury, or damage
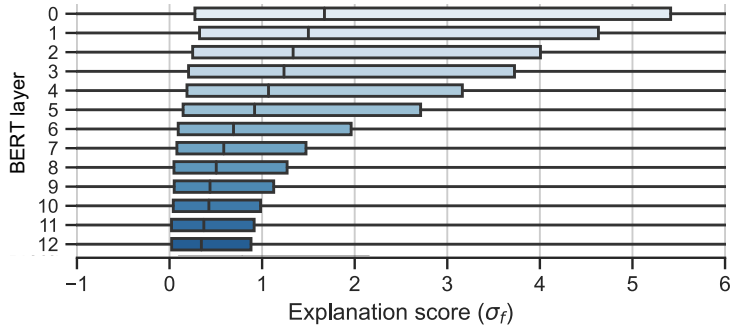
Figure 3: Explanation score for BERT modules. As the BERT layer increases, the explanation score tends to decrease, implying modules are harder to explain with SASC. Boxes show the median and interquartile range. The mean explanation score across all layers is $1.77\sigma_f$.

The *AG News* task has a particularly large fraction of relevant explanations, with many explanations corresponding very directly to class labels, e.g. *professional sports teams → sports* or *financial investments → business*. See the full set of generated explanations in Appendix A.3.

Explanation scores across layers SASC provides 1,500 explanations for transformer factors in 13 layers of BERT. Fig. 3 shows that the explanation score decreases with increasing layer depth, suggesting that SASC better explains factors at lower layers. Appendix B analyzes the characteristics of explanations for different BERT transformer factors and compares them to explanations generated for models fitted to fMRI brain responses to natural language.

5 Related work

A few related works study generating natural language explanations. MILAN [21] uses patch-level information of visual features to generate descriptions of neuron behavior in vision models. iPrompt [22] uses automated prompt engineering and D5 [12, 23]/GSClip [24] use LLMs to describe patterns in a dataset (as opposed to describing a module, as we study here). In concurrent work, [25] propose an algorithm similar to SASC that explains individual neurons in an LLM by predicting token-level neuron activations. A related line of work generates explanations that are approximately in natural language. [8] builds an explanation by manually inspecting the top ngrams eliciting the largest module responses from a corpus. [9] similarly extracts the top sentences from a corpus, but summarizes them using a parse tree. [26] use a gradient-based method to generate maximally activating text inputs and [27] builds a graph of ngrams to explain individual neurons.

Most prior works have focused on the problem of explaining a *single prediction* with natural language, rather than an entire module, e.g. for text classification [28–30], or computer vision [31, 32]. Besides natural language explanations, some works explain individual prediction via feature importances [33, 34], feature-interaction importances [35, 36], or extractive rationales [37, 38]. We build on a long line of recent work that explains neural-network *representations*, e.g. via probing [39, 40], via visualization [41, 42], by categorizing neurons [43–47], or localizing LLM knowledge [48, 49].

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Aaron E. Kornblith, Chandan Singh, Gabriel Devlin, Newton Addo, Christian J. Streck, James F. Holmes, Nathan Kuppermann, Jacqueline Grupp-Phelan, Jeffrey Fineman, Atul J. Butte, and Bin Yu. Predictability and stability testing to assess clinical decision instrument performance for children after blunt torso trauma. *PLOS Digital Health*, 2022.

[4] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[6] Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[7] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[8] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780, 2017.

[9] Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. Discovery of natural language concepts in individual units of cnns. *arXiv preprint arXiv:1902.07249*, 2019.

[10] Hongjin Su, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu, et al. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*, 2022.

[11] 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.

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

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

[14] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021.

[15] William Fedus, Jeff Dean, and Barret Zoph. A review of sparse expert models in deep learning. *arXiv preprint arXiv:2209.01667*, 2022.

[16] 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.

[17] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

[18] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.

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

[20] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[21] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*, 2022.

[22] Chandan Singh, John X Morris, Jyoti Aneja, Alexander M Rush, and Jianfeng Gao. Explaining patterns in data with language models via interpretable autoprompting. *arXiv preprint arXiv:2210.01848*, 2022.

[23] 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.

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

[25] Steven Bills, Nick Cammarata, Dan Mossing, William Saunders, Jeff Wu, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, and Jan Leike. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

[26] Nina Poerner, Benjamin Roth, and Hinrich Schütze. Interpretable textual neuron representations for nlp. *arXiv preprint arXiv:1809.07291*, 2018.

[27] Alex Foote, Neel Nanda, Esben Kran, Ionnis Konstas, and Fazl Barez. N2g: A scalable approach for quantifying interpretable neuron representations in large language models. *arXiv preprint arXiv:2304.12918*, 2023.

[28] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

[29] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361*, 2019.

[30] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. Wt5?! training text-to-text models to explain their predictions. *arXiv preprint arXiv:2004.14546*, 2020.

[31] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European conference on computer vision*, pages 3–19. Springer, 2016.

[32] 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*, pages 6720–6731, 2019.

[33] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

[35] Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *International Conference on Learning Representations*, page 26, 2019.

[36] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

[37] 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*, pages 31–40, 2008.

[38] Lei Sha, Oana-Maria Camburu, and Thomas Lukasiewicz. Learning from the best: Rationalizing predictions by adversarial information calibration. In *AAAI*, pages 13771–13779, 2021.

[39] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[40] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. *arXiv preprint arXiv:1906.08286*, 2019.

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

[42] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[43] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[44] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B Tenenbaum, William T Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*, 2018.

[45] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[46] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317, 2019.

[47] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding neurons in a haystack: Case studies with sparse probing, 2023.

[48] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*, 2022.

[49] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[50] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. A natural language fmri dataset for voxelwise encoding models. *bioRxiv*, pages 2022–09, 2022.

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[52] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020.

[53] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[54] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[55] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. *Advances in neural information processing systems*, 31, 2018.

[56] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9, 2023.

[57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[58] Doris Y. Tsao, Nicole Schweers, Sebastian Moeller, and Winrich A. Freiwald. Patches of face-selective cortex in the macaque frontal lobe. *Nature Neuroscience*, 11(8):877–879, August 2008.

[59] Liberty S Hamilton, Yulia Oganian, Jeffery Hall, and Edward F Chang. Parallel and distributed encoding of speech across human auditory cortex. *Cell*, 184(18):4626–4639, 2021.

[60] Jeffrey R. Binder, Rutvik H. Desai, William W. Graves, and Lisa L. Conant. Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, 19(12):2767–2796, December 2009.

[61] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, 320(5880):1191–1195, May 2008.

[62] Shinji Nishimoto, Alexander G Huth, Natalia Y Bilenko, and Jack L Gallant. Eye movement-invariant representations in the human visual system. *Journal of vision*, 17(1):11–11, 2017.

[63] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[64] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[65] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Deep language algorithms predict semantic comprehension from brain activity. *Scientific Reports*, 12(1):16327, 2022.

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 [17]	157k	3,719k	9,228k
Moth stories [50]	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

Table A2 shows the results for the *Default* setting when varying different modeling choices. Performance is similar across various choices, such as using bigrams or 4-grams rather than trigrams in the summarization step, or when using the LLaMA-2 13-billion parameter model [51] as the helper LLM rather than GPT-3. Additionally, we find that explanation performance increases with the capabilities of the helper LLM used for summarization/generation (Fig. A1).

Table A2 also shows comparisons to two additional baselines. Gradient-based explanations [26] use the gradients of f with respect to the input to generate maximally activating inputs, but fail to accurately identify the underlying groundtruth text, consistent with previous work in prompting [22, 52]. The topic modeling baseline generates explanations via topic modeling [53]: it learns a 100-component dictionary over ngrams. Then, explanation accuracy for a module is evaluated by taking the top-30 scoring ngrams for the module (as is done with SASC), finding the 5 topics with the highest scores for these ngrams, and manually checking whether there is a match between the groundtruth and any of the top-5 words in any of these topics. This method fails, largely because the topic model fails to construct topics relevant to each specific module, as the same input ngrams are shared across all modules.

Table A2: Explanation recovery accuracy when varying hyperparameters for the *Default* setting; averaged over 54 modules and 3 random seeds;

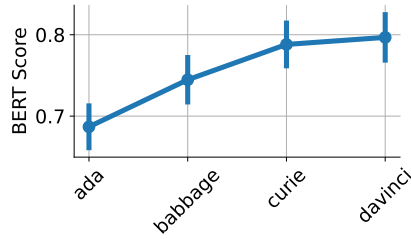| SASC (Original)	SASC (Bigrams)	SASC (4-grams)	SASC (LLaMA-2 summarizer)	SASC (LLaMA-2 generator)	Baseline (Gradient-based)	Baseline (Topic modeling)
0.883±0.03	0.815±0.04	0.889±0.03	0.870±0.03	0.852±0.04	0.093±0.01	0.111±0.01

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-003`) and are in order of increasing size. BERT score for each module is computed as the maximum over the 5 generated explanations.
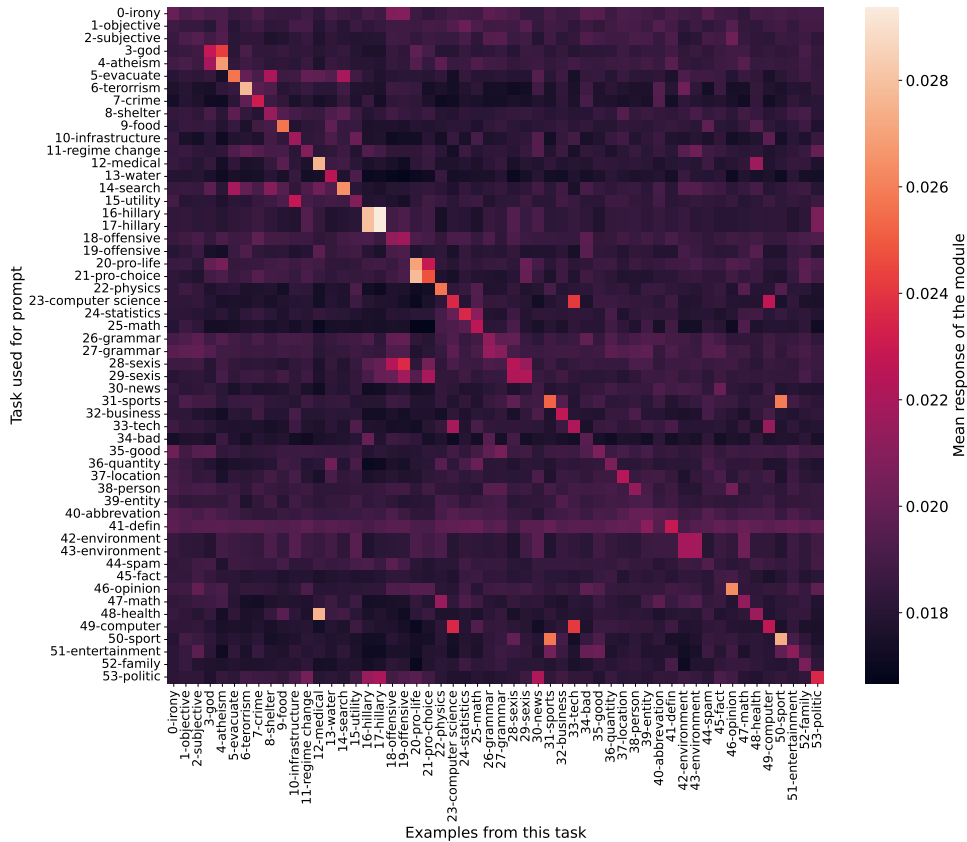
Figure A2: Synthetic modules respond more strongly to phrases related to their keyphrase (diagonal) than to phrases related to the keyphrase of other datasets (off-diagonal). Each value shows the mean response of the module to 5 phrases and each row is normalized using softmax. Each module is constructed using Instructor [10] with the prompt *Represent the short phrase for clustering:* and the groundtruth keyphrase given in Table A3. Related keyphrases are generated manually.

11

Table A3: 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

A.3 BERT interpretation

Details on fitting transformer factors Pre-trained transformer factors are taken from [16]. 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 [16] trained on the WikiText dataset [17].

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 A5 shows examples comparing SASC explanations with human-labeled explanations for all BERT transformer factors labeled in [16]. 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 4). Averaged over 3 random seeds; parentheses show standard error of the mean.

	Emotion	AG News	SST2
Top-10	0.50 ±0.08	1.00 ±0.00	0.80 ±0.14
Top-15	0.47 ±0.05	0.98 ±0.03	0.69 ±0.13
Top-20	0.42 ±0.09	0.98 ±0.02	0.55 ±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 (Ours)	Explanation score (Human)	Explanation score (Ours)
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

B Generating explanations for fMRI-voxel modules
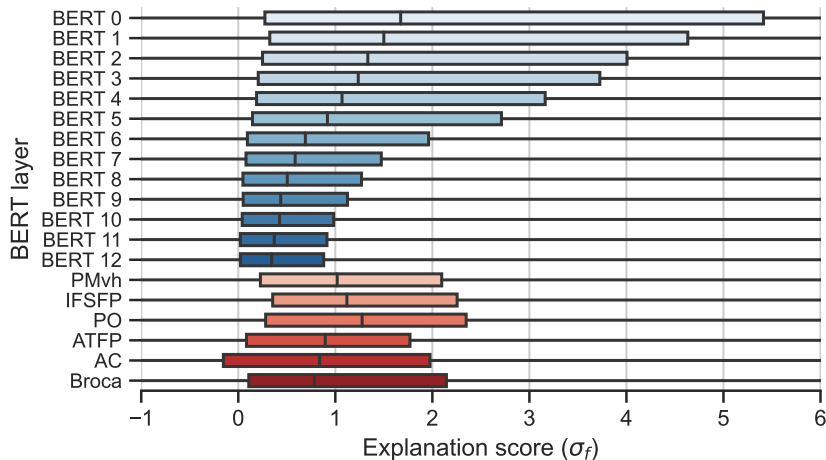
Figure B1: Explanation score for BERT modules. As the BERT layer increases, the explanation score tends to decrease, implying modules are harder to explain with SASC. Boxes show the median and interquartile range.
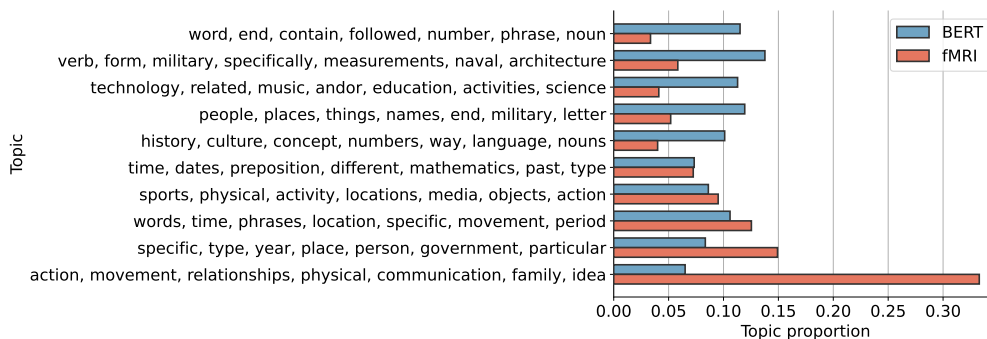

Figure B2: Topics found by LDA for explanations of BERT factors and fMRI voxels. Topic proportion is calculated by assigning each explanation to the topic with the largest coefficient. Topic proportions for BERT/fMRI explanations largely overlap, although the bottom topic consisting of physical/social words is much more prevalent in fMRI explanations.

fMRI voxel modules   A central challenge in neuroscience is understanding how and where semantic concepts are represented in the brain. To meet this challenge, one line of study predicts the response of different brain voxels (i.e. small regions in the brain) to natural language stimuli [54, 55]. We analyze data from [50] and [56], which consists of fMRI responses for 3 human subjects as they listen to 20+ hours of narrative stories from podcasts. We fit modules to predict the fMRI response in each voxel from the text that the subject was hearing by extracting text embeddings with a pre-trained LLaMA model (`decapoda-research/llama-30b-hf`) [57]. After fitting the modules on the training split and evaluating them on the test split using bootstrapped ridge regression, we generate SASC explanations for 1,500 well-predicted voxel modules, distributed evenly among the three human subjects and diverse cortical areas (see details on the fMRI experimental setup in Appendix B.1).

Voxel explanations   Table B1 shows examples of explanations for individual voxels, along with three top ngrams used to derive the explanation. Each explanation unifies fairly different ngrams under a common theme, e.g. *sliced cucumber, cut the apples, sauteed shiitake...* → *food preparation*. In some cases, the explanations recover language concepts similar to known selectivity in sensory modalities, e.g. face selectivity in IFSFP [58] and selectivity for non-speech sounds such as laughter in primary auditory cortex [59]. The ngrams also provide more fine-grained hypotheses

Table B1: Examples of recovered explanations for individual fMRI voxel modules. All achieve an fMRI predicted correlation greater than 0.3 and an explanation score of at least 1σ. The third column shows 3 of the ngrams used to derive the explanation in the SASC summarization step.

Explanation	ROI	Example top ngrams
looking or staring in some way	IFSFP	eyed her suspiciously, wink at, locks eyes with
relationships and loss	ATFP	girlfriend now ex, lost my husband, was a miscarriage
physical injury or pain	Broca	infections and gangrene, pulled a muscle, burned the skin
counting or measuring time	PMvh	count down and, weeks became months, three more seconds
food preparation	ATFP	sliced cucumber, cut the apples, sauteed shiitake
laughter or amusement	ATFP, AC	started to laugh, funny guy, chuckled and

for selectivity (e.g. *physical injury or pain*) compared to the coarse semantic categories proposed in earlier language studies (e.g. *emotion* [54, 60, 61]).

Fig. B2 shows the topics that fMRI explanations best fit into compared with BERT transformer factors. The proportions for many topics are similar, but the fMRI explanations yield a much greater proportion for the topic consisting of social words (e.g. *relationships*, *communication*, *family*) and perceptual words (e.g. *action*, *movement*, *physical*). This is consistent with prior knowledge, as the largest axis of variation for fMRI voxels is known to separate social concepts from physical concepts [54].

The selected 1,500 voxels often achieve explanation scores considerably greater than zero for their explanations (mean explanation score $1.27\sigma_f \pm 0.029$). Fig. B1 (bottom) shows the mean explanation score for the six most common fMRI regions of interest (ROIs) among the voxels we study here. Across regions, the fMRI voxel modules generally attain explanation scores that are slightly lower than BERT modules in early layers and slightly higher than BERT modules in the final layers. We also find some evidence that the generated fMRI voxel explanations can explain not just the fitted module, but also brain responses to unseen data (see Appendix B.2). This suggests that the voxel explanations here can serve as hypotheses for followup experiments to affirm the fine-grained selectivity of specific brain voxels. The trends in Fig. B2 and Fig. B1 are consistent whether the fMRI explanations are generated using ngrams from the original narrative stories or from the WikiText dataset (see Appendix B.3).

B.1 fMRI data and model fitting

This section gives more details on the fMRI experiment analyzed in Appendix B. These MRI data are available publicly [50, 56], 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 [62] 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 [63].

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 [57], and the 9th layer of GPT [64]. 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 [54]. 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 [56].

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.

B.2 Evaluating top fMRI voxel evaluations

Table B2 shows two evaluations of the fMRI voxel explanations. First, similar to Fig. B1, 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 [10][3]. We then calculate the spearman rank correlation between this scalar distance and the recorded voxel response (see Table B2). The mean computed correlation is low[4], 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 B2: 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

---

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

[4]For reference, test correlations published in fMRI voxel prediction from language are often in the range of 0.01-0.1 [65].

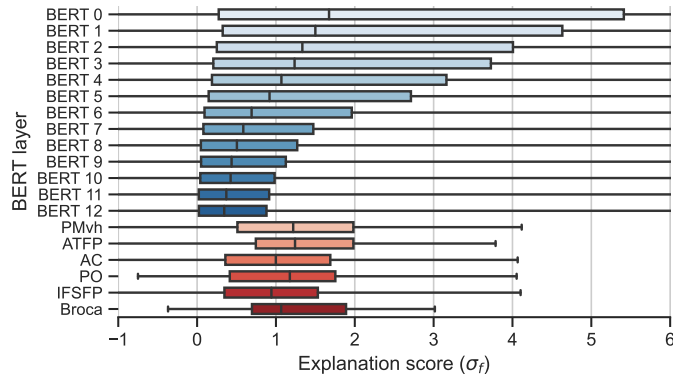## B.3 fMRI results when using WikiText Corpus

Figure B3: Results in Fig. B1 when using WikiText as the underlying corpus for ngrams rather than narrative stories.
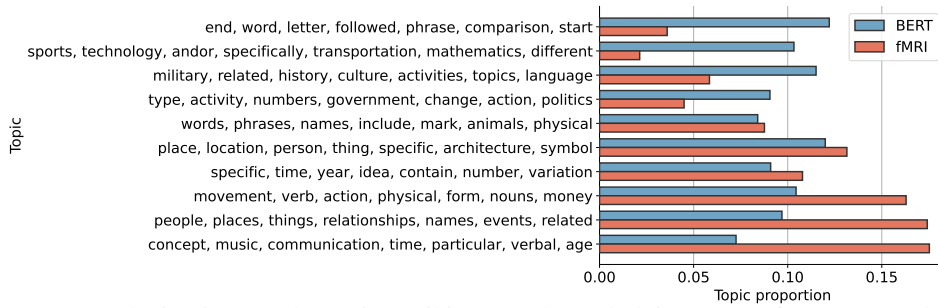

Figure B4: Results in Fig. B2 when using WikiText as the underlying corpus for ngrams rather than narrative stories.