Crafting Interpretable Embeddings for Language Neuroscience by Asking LLMs Questions

Vinamra Benara* UC Berkeley

Chandan Singh* Microsoft Research

John X. Morris Cornell University

Richard J. Antonello UT Austin

Ion Stoica UC Berkeley

Alexander G. Huth UT Austin

Jianfeng Gao Microsoft Research

*Equal contribution

Abstract

Large language models (LLMs) have rapidly improved text embeddings for a growing array of natural-language processing tasks. However, their opaqueness and proliferation into scientific domains such as neuroscience have created a growing need for interpretability. Here, we ask whether we can obtain interpretable embeddings through LLM prompting. We introduce question-answering embeddings (QA-Emb), embeddings where each feature represents an answer to a yes/no question asked to an LLM. Training QA-Emb reduces to selecting a set of underlying questions rather than learning model weights.

We use QA-Emb to flexibly generate interpretable models for predicting fMRI voxel responses to language stimuli. QA-Emb significantly outperforms an established interpretable baseline, and does so while requiring very few questions. This paves the way towards building flexible feature spaces that can concretize and evaluate our understanding of semantic brain representations. We additionally find that QA-Emb can be effectively approximated with an efficient model, and we explore broader applications in simple NLP tasks.[1]

1 Introduction

Text embeddings are critical to many applications, including information retrieval, semantic clustering, retrieval-augmented generation, and language neuroscience. Traditionally, text embeddings leveraged interpretable representations such as bag-of-words or BM-25 [1]. Modern methods often replace these embeddings with representations from large language models (LLMs), which may better capture nuanced contexts and interactions [2–7]. However, these embeddings are essentially black-box representations, making it difficult to understand the predictive models built on top of them (as well as why they judge different texts to be similar in a retrieval context). This opaqueness is detrimental in scientific fields, such as neuroscience [8] or social science [9], where trustworthy interpretation itself is the end goal. Moreover, this opaqueness has debilitated the use of LLM embeddings (for prediction or retrieval) in high-stakes applications such as medicine [10], and raised issues related to regulatory pressure, safety, and alignment [11–14].

To ameliorate these issues, we introduce question-answering embeddings (QA-Emb), a method that builds an interpretable embedding by repeatedly querying a pre-trained autoregressive LLM with a set of questions that are selected for a problem (Fig. 1). Each element of the embedding represents

[1] All code for QA-Emb is made available on Github at github.com/csinva/interpetable-embeddings.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

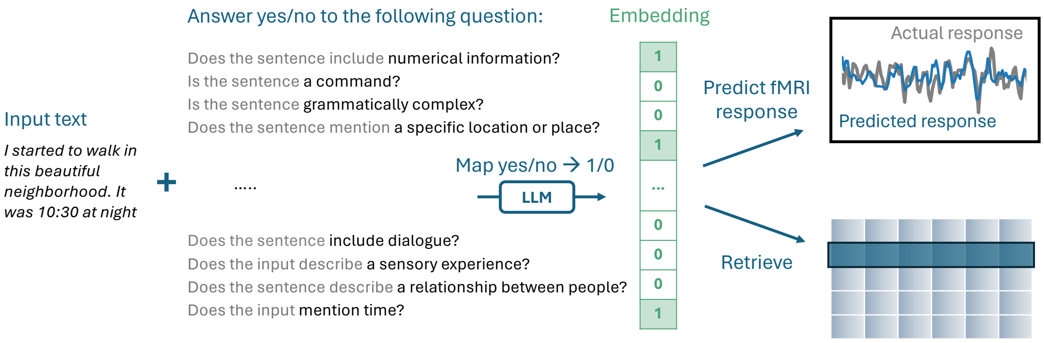
Answer yes/no to the following question:
Embedding
Does the sentence include numerical information?
Is the sentence a command?
Is the sentence grammatically complex?
Does the sentence mention a specific location or place?
Input text
I started to walk in this beautiful neighborhood. It was 10:30 at night
+
.....
Map yes/no → 1/0
LLM
Does the sentence include dialogue?
Does the input describe a sensory experience?
Does the sentence describe a relationship between people?
Does the input mention time?
1
0
0
1
...
0
0
0
1
Predict fMRI response
Actual response
Predicted response
Retrieve

Figure 1: QA-Emb produces an embedding for an input text by prompting an LLM with a series of yes/no questions. This embedding can then be used in downstream tasks such as fMRI response prediction or information retrieval.

the answer to a different question asked to an LLM, making the embedding human-inspectable. For example, the first element may be the answer to the question *Does the input mention time?* and the output would map *yes*/*no* to 1/0. Training QA-Emb requires only black-box access to the LLM (it does not require access to the LLM internals) and modifies only natural-language prompts, rather than LLM parameters. The learning problem is similar to the optimization faced in natural-language autoprompting [15, 16] or single-neuron explanation [17, 18], but seeks a set of questions rather than an individual prompt.

We focus on a single neuroscience problem in close collaboration with neuroscientists. Grounding in a neuroscience context allows us to avoid common pitfalls in evaluating interpretation methods [19, 20] that seek to test "interpretability" generally. Additionally, this focus allows to more realistically integrate domain knowledge to select and evaluate the questions needed for QA-Emb, one of its core strengths. Nevertheless, QA-Emb may be generally applicable in other domains where it is important to meaningfully interpret text embeddings.

In our neuroscience setting, we build QA-Emb representations from natural-language questions that can predict human brain responses measured by fMRI to natural-language stimuli. This allows for converting informal verbal hypotheses about the semantic selectivity of the brain into quantitative models, a pressing challenge in fields such as psychology [21]. We find that predictive models built on top of QA-Embs are quite accurate, providing a 26% improvement over an established interpretable baseline [22] and even slightly outperforming a black-box BERT baseline [23]. Additionally, QA-Emb yields concise embeddings, outperforming the interpretable baseline (that consists of 985 features) with only 29 questions.

We investigate two major limitations of QA-Emb in Sec. 5. First, with regards to computational efficiency, we find that we can drastically reduce the computational cost of QA-Emb by distilling it into a model that computes the answers to all selected questions in a single feedforward pass by using many classification heads. Second, we evaluate the accuracy of modern LLMs at reliably answering diverse yes/no questions. Finally, Sec. 6 explores broader applications for QA-Emb in a simple information retrieval setting and text-clustering setting.

2 Methods

QA-Emb is an intuitive method to generate text embeddings from a pre-trained autoregressive LLM (Fig. 1). Given a text input, QA-Emb builds an interpretable embedding by querying the LLM with a set of questions about the input. Each element of the embedding represents the answer to a different question asked to an LLM. This procedure allows QA-Emb to capture nuanced and relevant details in the input while staying interpretable.

Learning a set of yes/no questions QA-Emb requires specifying a set of yes/no questions $Q \in \mathcal{Q}_{\text{yes/no}}$ that yield a binary embedding $v_Q(x) \in \{0, 1\}^d$ for an input string x. The questions are chosen to yield embeddings that are suitable for a downstream task. In our fMRI prediction task, we optimize for supervised linear regression: given a list of n input strings X and a multi-dimensional continuous output $Y \in \mathbb{R}^{nxd}$, we seek embeddings that allow for learning effective ridge regression models:

$$Q = \underset{Q \in \mathcal{Q}_{\text{yes/no}}}{\text{argmin}} \left[\min_{\theta \in \mathbb{R}^d} \sum_i^n ||Y^{(i)} - \theta^T v_Q(X^{(i)})|| + \lambda ||\theta||_2 \right], \tag{1}$$

where θ is a learned coefficient vector for predicting the fMRI responses and λ is the ridge regularization parameter.

Directly optimizing over the space of yes/no questions is difficult, as it requires searching over a discrete space with a constraint set $\mathcal{Q}_{\text{yes/no}}$ that is hard to specify. Instead, we heuristically optimize the set of questions Q, by prompting a highly capable LLM (e.g. GPT-4 [24]) to generate questions relevant to our task, e.g. *Generate a bulleted list of questions with yes/no answers that is relevant for {{task description}}*. Customizing the task description helps yield relevant questions. The prompt can flexibly specify more prior information when available. For example, it can include examples from the input dataset to help the LLM identify data-relevant questions. Taking this a step further, questions can be generated sequentially (similar to gradient boosting) by having the LLM summarize input examples that incur high prediction error to generate new questions focused on those examples. While we focus on optimizing embeddings for fMRI ridge regression in Eq. (1), different downstream tasks may require different inner optimization procedures, e.g. maximizing the similarity of relevant documents for retrieval.

Post-hoc pruning of Q. The set of learned questions Q can be easily pruned to be made compact and useful in different settings. For example, in our fMRI regression setting, a feature-selection procedure such as Elastic net [25] can be used to remove redundant/uninformative questions from the specified set of questions Q. Alternatively, an LLM can be used to directly adapt Q to yield task-specific embeddings. Since the questions are all in natural language, they can be listed in a prompt, and an LLM can be asked to filter the task-relevant ones, e.g. *Here is a list of questions:{{question list}} List the subset of these questions that are relevant for {{task description}}*.

Limitations: computational cost and LLM inaccuracies. While effective, the QA-Emb pipeline described here has two major limitations. First, QA-Emb is computationally intensive, requiring d LLM calls to compute an embedding. This is often prohibitively expensive, but may be worthwhile in high-value applications (such as our fMRI setting) and will likely become more tenable as LLM inference costs continue to rapidly decrease. We find that we can dramatically reduce this cost by distilling the QA-Emb model into a single LLM model with many classification heads in Sec. 5.1. Otherwise, LLM inference costs are partially mitigated by the ability to reuse the KV-cache for each question and the need to only generate a single token for each question. While computing embeddings with QA-Emb is expensive, *searching* embeddings is made faster by the fact that the resulting embeddings are binary and often relatively compact.

Second, QA-Emb requires that the pre-trained LLM can faithfully answer the given yes-no questions. If an LLM is unable to accurately answer the questions, it hurts explanation's faithfulness. Thus, QA-Emb requires the use of fairly strong LLMs and the set of chosen questions should be accurately answered by these LLMs (Sec. 5.2 provides analysis on the question-answering accuracy of different LLMs).

Hyperparameter settings For answering questions, we average the answers from Mistral-7B [26] (`mistralai/Mistral-7B-Instruct-v0.2`) and LLaMA-3 8B [27] (`meta-llama/Meta-Llama-3-8B-Instruct`) with two prompts. All perform similarly and averaging their answers yields a small performance improvement (Table A2). For generating questions, we prompt GPT-4 [24] (`gpt-4-0125-preview`). Experiments were run using 64 AMD MI210 GPUs, each with 64 gigabytes of memory, and reproducing all experiments in the paper requires approximately 4 days (initial explorations required roughly 5 times this amount of compute). All prompts used and generated questions are given in the appendix or on Github.

3 Related work

Text embeddings Text embeddings models, which produce vector representations of document inputs, have been foundational to NLP. Recently, transformer-based models have been trained to yield embeddings in a variety of ways [2–7], including producing embeddings that are sparse [28] or have variable lengths [29]. Recent works have also leveraged autoregressive LLMs to build embeddings, e.g. by repeating embeddings [30], generating synthetic data [6, 31], or using the last-token distribution of an autoregressive LLM as an embedding [32]. Similar to QA-Emb, various works have used LLM answers to multiple prompts for different purposes, e.g. text classification [33–35], learning style embeddings [36], or data exploration [37].

Interpreting representations A few works have focused on building intrinsically interpretable text representations, e.g. word or ngram-based embeddings such as word2vec [38], Glove [39], and LLM word embeddings. Although their dimensions are not natively interpretable, for some tasks, such as classification, they can be projected into a space that is interpretable [40], i.e. a word-level representation. Note that it is difficult to learn a sparse interpretable model from these dense embeddings, as standard techniques (e.g. Elastic net) cannot be directly applied.

When instead using black-box representations, there are many post-hoc methods to interpret embeddings, e.g. probing [41, 42], categorizing elements into categories [43–46], categorizing directions in representation space [47–50], or connecting multimodal embeddings with text embeddings/text concepts [51–55]. For a single pair of text embeddings, prediction-level methods can be applied to approximately explain why the two embeddings are similar [56, 57].

Natural language representations in fMRI Using LLM representations to help predict brain responses to natural language has recently become popular among neuroscientists studying language processing [58–63] (see [64, 65] for reviews). This paradigm of using "encoding models" [66] to better understand how the brain processes language has been applied to help understand the cortical organization of language timescales [67, 68], examine the relationship between visual and semantic information in the brain [69], and explore to what extent syntax, semantics, or discourse drives brain activity [22, 70–77, 18]. The approach here extends these works to build an increasingly flexible, interpretable feature space for modeling fMRI responses to text data.

4 Main results: fMRI interpretation

A central challenge in neuroscience is understanding how and where semantic concepts are represented in the brain. To meet this challenge, we extend the line of study that fits models to predict the response of different brain voxels (i.e. small regions in the brain) to natural language stimuli. Using QA-Emb, we seek to bridge models that are interpretable [1, 22] with more recent LLM models that are accurate but opaque [58–60].

4.1 fMRI experimental setup

Dataset We analyze data from two recent studies [78, 79] (released under the MIT license), which contain fMRI responses for 3 human subjects listening to 20+ hours of narrative stories from podcasts. We extract text embeddings from the story that each subject hears and fit a ridge regression to predict the fMRI responses (Eq. (1)). Each subject listens to either 79 or 82 stories (consisting of 27,449 time points) and 2 test stories (639 time points); Each subject's fMRI data consists of approximately 100,000 voxels; we preprocess it by running principal component analysis (PCA) and extracting the coefficients of the top 100 components.

Regression modeling We fit ridge regression models to predict these 100 coefficients and evaluate the models in the original voxel space (by applying the inverse PCA mapping and measuring the correlation between the response and prediction for each voxel). We deal with temporal sampling following [22, 60]; an embedding is produced at the timepoint for each word in the input story and these embeddings are interpolated using Lanczos resampling. Embeddings at each timepoint are produced from the ngram consisting of the 10 words preceding the current timepoint. We select the best-performing hyperparameters via cross-validation on 5 time-stratified bootstrap samples of the training set. We select the best ridge parameters from 12 logarithmically spaced values between 10 and 10,000. To model temporal delays in the fMRI signal, we also select between adding 4, 8, or 12 time-lagged duplicates of the stimulus features.

Generating QA-Emb questions To generate the questions underlying QA-Emb, we prompt GPT-4 with 6 prompts that aim to elicit knowledge useful for predicting fMRI responses (precise prompts in Appendix A.3). This includes directly asking the LLM to use its knowledge of neuroscience, to brainstorm semantic properties of narrative sentences, to summarize examples from the input data, and to generate questions similar to single-voxel explanations found in a prior work [18]. This process yields 674 questions (Fig. 1 and Table A1 show examples, see all questions on Github). We perform feature selection by running multi-task Elastic net with 20 logarithmically spaced regularization parameters ranging from 10^{-3} to 1 and then fit a Ridge regression to the selected features.[2] See extended details on the fMRI experimental setup in Appendix A.1 and all prompts in Appendix A.3.

Baselines We compare QA-Emb to Eng1000, an interpretable baseline developed in the neuroscience literature specifically for the task of predicting fMRI responses from narrative stories [22]. Each element in an Eng1000 embedding corresponds to a cooccurence statistic with a different word, allowing full interpretation of the underlying representation in terms of related words. We additionally compare to embeddings from BERT [23] (`bert-base-uncased`) and LLaMA models [81, 27]. For each subject, we sweep over 5 layers from LLaMA-2 7B (`meta-llama/Llama-2-7b-hf`, layers 6, 12, 18, 24, 30), LLaMA-2 70B (`meta-llama/Llama-2-70b-hf`, layers 12, 24, 36, 48, 60), and LLaMA-3 8B (`meta-llama/Meta-Llama-3-8B`, layers 6, 12, 18, 24, 30), then report the test performance for the model that yields the best cross-validated accuracy (see breakdown in Table A3).

4.2 fMRI predictive performance

We find that QA-Emb predicts fMRI responses fairly well across subjects (Fig. 2A), achieving an average test correlation of 0.116. QA-Emb significantly outperforms the interpretable baseline Eng1000 (26% average improvement). Comparing to the two transformer-based baselines (which do not yield straightforward interpretations), we find that QA-Emb slightly outperforms BERT (5% improvement) and worse than the best cross-validated LLaMA-based model (7% decrease). Trends are consistent across all 3 subjects.

To yield a compact and interpretable model, Fig. 2B further investigates the compressibility of the two interpretable methods (through Elastic net regularization). Compared to Eng1000, QA-Emb improves performance very quickly as a function of the number of features included, even outperforming the final Eng1000 performance with only 29 questions (mean test correlation 0.122 versus 0.118). Table A1 shows the 29 selected questions, which constitute a human-readable description of the entire model.

Fig. 2C-D further break down the predictive performance across different brain regions for a particular subject (S03). The regions that are well-predicted by QA-Emb (Fig. 2C) align with language-specific areas that are seen in the literature [59, 82]. They do not show any major diversions from transformer-based encoding models (Fig. 2D), with the distribution of differences being inconsistent across subjects (see Fig. A1).

4.3 Interpreting the fitted representation from QA-Emb

The QA-Emb representation enables not only identifying which questions are important for fMRI prediction, but also mapping their selectivity across the cortex. We analyze the QA-Emb model which uses 29 questions and visualize the learned regression weights for different questions. Fig. 3 shows example flatmaps of the regression coefficients for 3 of the questions across the 2 best-predicted subjects (S02 and S03). Learned feature weights for the example questions capture known selectivity and are highly consistent across subjects. In particular, the weights for the question "*Does the sentence involve a description of a physical environment or setting*?" captures classical place areas including occipital place area [83] and retrosplenial complex [84], as well as intraparietal sulcus [85]. The weights for the question "*Is the sentence grammatically complex*?" bear striking similarity to the language network [82, 86], which is itself localized from a contrast between sentences and nonwords. Other questions, such as "*Does the sentence describe a physical action*?", which has strong right

[2] We run Elastic net using the MultiTaskElasticNet class from scikit-learn [80].

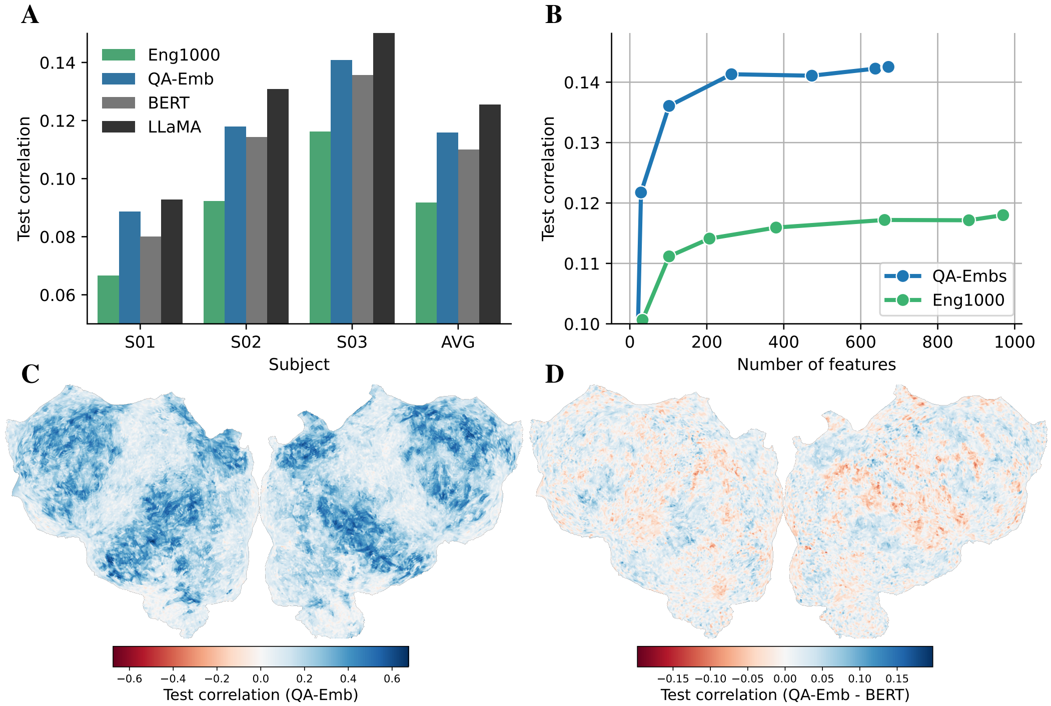

Figure 2: Predictive performance for QA-Emb compared to baselines. (A) Test correlation for QA-Emb outperforms the interpretable Eng1000 baseline, is on par with the black-box BERT baseline, and is worse than the best-performing LLaMA model. (B) Test correlation for method quickly grows as a function of the number of included questions. (C) Test correlation per voxel for QA-Emb. (D) Difference in the test correlation per voxel for subject between QA-Emb and BERT. Error bars for (A) and (B) (standard error of the mean) are within the points (all are below 0.001). (B), (C), and (D) show results for subject S03.

Table 1: Mean test correlation when comparing QA-Emb computed via many LLM calls to QA-Emb computed via a single distilled model. Distillation does not significantly degrade performance. All standard errors of the mean are below 10^{-3}.

	QA-Emb	QA-Emb (distill, binary)	QA-Emb (distill, probabilistic)	Eng1000
UTS01	0.081	0.083	0.080	0.077
UTS02	0.124	0.118	0.118	0.096
UTS03	0.136	0.132	0.142	0.117
AVG	**0.114**	**0.111**	**0.113**	**0.097**

laterality, do not have a strong basis in prior literature. These questions point to potentially new insights into poorly understood cortical regions.

5 Evaluating the limitations of QA-Emb

5.1 Improving computational efficiency via model distillation

To reduce the computational cost of running inference with QA-Emb, we explore distilling the many LLM calls needed to compute QA-Emb into a single model with many classification heads. Specifically, we finetune a RoBERTa model [87] (`roberta-base`) with 674 classification heads to predict all answers required for QA-Emb in a single feedforward pass. We finetune the model on answers from LLaMA-3 8B with a few-shot prompt for 80% of the 10-grams in the 82 fMRI training stories (123,203 examples), use the remaining 20% as a validation set for early stopping (30,801

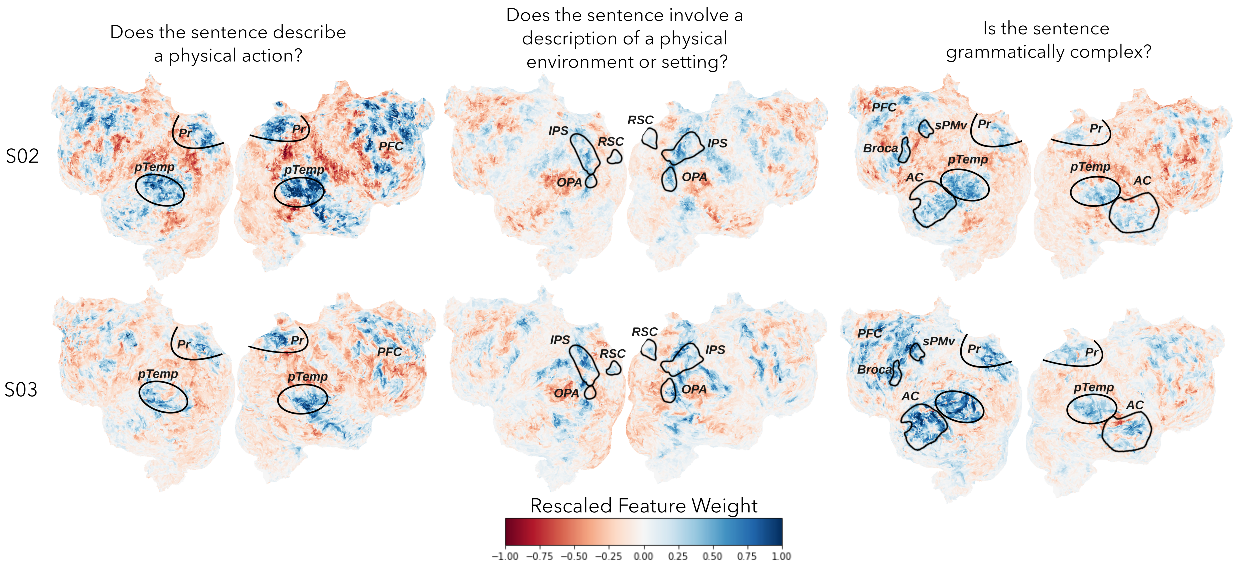
Does the sentence describe a physical action?
Does the sentence involve a description of a physical environment or setting?
Is the sentence grammatically complex?
S02
S03
Pr
pTemp
PFC
IPS
RSC
OPA
sPMv
Broca
AC
Rescaled Feature Weight
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

Figure 3: Learned feature weights for 3 example questions capture known selectivity and are consistent across subjects. All feature weights are jointly rescaled to the range (-1, 1) for visualization. Abbreviations: Pr = precuneus, pTemp = posterior temporal cortex, PFC = prefrontal cortex, IPS = intraparietal sulcus, RSC = retrosplenial complex, OPA = occipital place area, PPA = parahippocampal place area, Broca = Broca's area, sPMv = superior premotor ventral speech area, AC = auditory cortex.

examples), and evaluate on all 10-grams in the 2 testing stories (4,594 examples). We finetune using AdamW [88] with a learning rate of $5 \cdot 10^{-5}$.

When evaluated on the fMRI prediction task, the distilled model (*QA-Emb (distill, binary)* in Table 1) yields a performance only slightly below the original model. If we relax the restriction that the finetuned model yields binary embeddings and instead use the predicted probability for *yes*, the performance rises slightly to nearly match the original model (0.113 instead of 0.114 average test correlation) and maintains a significant improvement over the Eng1000 baseline. Note that the distilled model achieves an 88.5% match for yes/no answers on 10-grams for the test set. Nevertheless, the fMRI prediction for any given timepoint is computed from many questions and ngrams, mitigating the effect of individual errors in answering a question.

5.2 Evaluating question-answering faithfulness

We evaluate the faithfulness of our question-answering models on a recent diverse collection of 54 binary classification datasets [89, 90] (see data details in Table A4). These datasets are difficult, as they are intended to encompass a wider-ranging and more realistic list of questions than traditional NLP datasets.

Fig. 4 shows the classification accuracy for the 3 LLMs used previously along with GPT-3.5 (`gpt-3.5-turbo-0125`). On average, each of the LLMs answers these questions with fairly high accuracy, with GPT-4 slightly outperforming the other models. However, we observe poor performance on some tasks, which we attribute to the task difficulty and the lack of task-specific prompt engineering. For example, the dataset yielding the lowest accuracy asks the question *Is the input about math research?*. While this may seem like a fairly simple question for an LLM to answer, the examples in the negative class consist of texts from other quantitative fields (e.g. chemistry) that usually contain numbers, math notation, and statistical analysis. Thus the LLMs answer *yes* to most examples and achieve accuracy near chance (50%). Note that these tasks are more difficult than the relatively simple questions we answer in the fMRI experiments, especially since the fMRI input lengths are each 10 words, whereas the input lengths for these datasets are over 50 words on average (with some inputs spanning over 1,000 words).

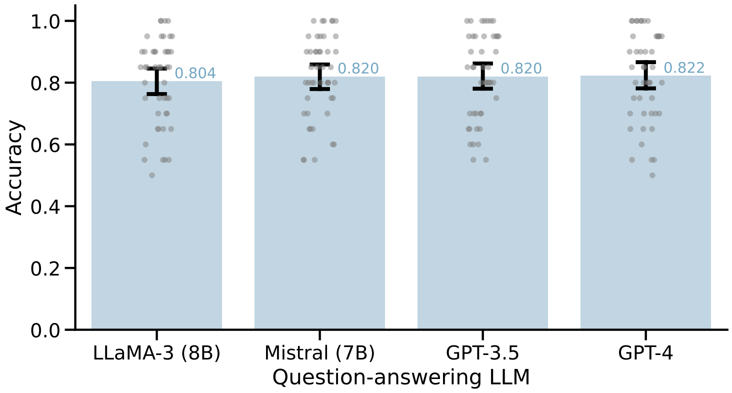

Figure 4: Performance of question-answering for underlying LLMs on the D3 collection of binary classification datasets. Each point shows an individual dataset and error bars show the 95% confidence interval.

Table 2: Information retrieval results for different interpretable embedding models. QA-Emb in combination with BM-25 achieves a slight improvement over the interpretable baselines. QA-Emb additionally yields reasonably strong performance compared to its embedding size. †Note that QA-Emb embeddings are binary, so the raw number of dimensions overrepresents the embedding's size relative to other methods. Error bars show standard error of the mean.

	Mean reciprocal rank	Recall@1	Recall@5	Size
Bag of words	0.37±.01	0.28±.02	0.42±.02	27,677
Bag of bigrams	0.39±.01	0.30±.02	0.44±.02	197,924
Bag of trigrams	0.39±.02	0.30±.02	0.44±.02	444,403
QA-Emb	0.45±.01	0.34±.01	0.50±.01	†**2,000**
BM-25	0.77±.01	0.69±.01	0.82±.01	27,677
BM-25 + QA-Emb	**0.80**±**.01**	**0.71**±**.01**	**0.84**±**.01**	29,677

6 Secondary results: evaluating QA-Emb in simple NLP tasks

6.1 Benchmarking QA-Emb for information retrieval

In this section, we investigate applying QA-Emb to a simplified information retrieval task. We take a random subset of 4,000 queries from the MSMarco dataset ([91], Creative Commons License) and their corresponding groundtruth documents, resulting in 5,210 documents. We use 25% of the queries to build a training set and keep the remaining 75% for testing. For evaluation, we calculate the cosine similarity match between the embeddings for each query and its groundtruth documents using mean reciprocal rank and recall.

To compute QA-Emb, we first generate 2,000 questions through prompting GPT-4 based on its knowledge of queries in information retrieval (see prompts in the Github). We use a regex to slightly rewrite the resulting questions for queries to apply to documents (e.g. *Is this query related to a specific timeframe?* → *Is this text related to a specific timeframe?*). We then answer the questions both for each query and for each corpus document, again using LLaMA-3 8B. Rather than fitting a ridge regression as in Eq. (1), we use the training set to learn a scalar for each question that multiplies its binary output to change both its sign and magnitude in the embedding (optimization details in Appendix A.4).

Table 2 shows the information retrieval results. Combining BM-25 with QA-Emb achieves a small but significant improvement over the interpretable baselines. QA-Emb on its own achieves modest performance, slightly improving slightly over a bag-of-words representation, but significantly underperforming BM-25. Nevertheless, its size is considerably smaller than the other interpretable baselines making it quicker to interpret and to use for retrieval.

Table 3: Clustering scores before and after zero-shot adaptation (higher is better). Errors give standard error of the mean.

	Rotten tomatoes	AG News	Emotion	Financial phrasebank	AVG	Embedding size (AVG)
Original	0.126±0.011	0.124±0.007	0.046±0.007	0.084±0.008	0.095	100
Adapted	**0.248±0.016**	**0.166±0.012**	**0.057±0.010**	**0.292±0.017**	**0.191**	**25.75±0.95**

6.2 Zero-shot adaptation in text clustering

We now investigate QA-Emb in a simplified text clustering setting. To do so, we study 4 text-classification datasets: Financial phrasebank ([92], creative commons license), Emotion [93] (CC BY-SA 4.0 license), AGNews [94], and Rotten tomatoes [95]. For each dataset, we treat each class as a cluster and evaluate the *clustering score*, defined as the difference between the average inter-class embedding distance and the average intra-class embedding distance (embedding distance is measured via Euclidean distance). A larger clustering score suggests that embeddings are well-clustered within each class.

In our experiment, we build a 100-dimensional embedding by prompting GPT-4 to generate 25 yes/no questions related to the semantic content of each dataset (e.g. for Rotten tomatoes, *Generate 25 yes/no questions related to movie reviews*). We then concatenate the answers for all 100 questions to form our embedding. These general embeddings do not yield particularly strong clustering scores (Table 3 top), as the questions are diverse and not particularly selective for each dataset.

However, simply through prompting, we can adapt these general embeddings to each individual dataset. We call GPT-4 with a prompt that includes the full list of questions and ask it to select a subset of questions that are relevant to each task. The result embeddings (Table 3 bottom) yield higher clustering scores, suggesting that QA-Emb can be adapted to each task in a zero-shot manner (in this simplified setting). Moreover, the resulting task-specific embeddings are now considerably smaller.

7 Discussion

We find that QA-Emb can effectively produce interpretable and high-performing text embeddings. While we focus on a language fMRI setting, QA-Emb may be able to help flexibly build an interpretable text feature space in a variety of domains, such as social science [9], medicine [10], or economics [96], where meaningful properties of text can help discover something about an underlying phenomenon or build trust in high-stakes settings. Alternatively, it could be used in mechanistic interpretability, to help improve post-hoc explanations of learned LLM representations.

As LLMs improve in both efficiency and capability, QA-Emb can be incorporated into a variety of common NLP applications as well, such as RAG or information retrieval. For example, in RAG systems such as RAPTOR [97] or Graph-RAG [98], explanations may help an LLM not only retrieve relevant texts, but also specify why they are relevant and how they may be helpful.

Learning text questions rather than model weights is a challenging research area, furthering work in automatic prompt engineering [15, 16]. Our approach takes a heuristic first step at solving this problem, but future work could explore more directly optimizing the set of learned questions Q in Eq. (1) via improved discrete optimization approaches and constraints. One possible approach may involve having LLMs themselves identify the errors the current model is making and improving based on these errors, similar to general trends in LLM self-improvement and autoprompting [99–102]. Another approach may involve improving the explanation capabilities of LLMs to help extract more questions more faithfully from data [103, 104].

Broader Impacts QA-Emb seeks to advance the field of LLM interpretation, a crucial step toward addressing the challenges posed by these often opaque models. Although LLMs have gained widespread use, their lack of transparency can lead to significant harm, underscoring the importance of interpretable AI. There are many potential positive societal consequences of this form of interpretability, e.g., facilitating a better understanding of scientific data and models, along with a better understanding of LLMs and how to use them safely. Nevertheless, as is the case with most ML research, the interpretations could be used to interpret and potentially improve an LLM or dataset that is being used for nefarious purposes. Moreover, QA-Emb requires substantial computational resources, contributing to increased concerns over sustainability.

References

[1] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[2] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[3] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[5] Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*, 2022.

[6] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

[7] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[8] Shailee Jain, Vy A Vo, Leila Wehbe, and Alexander G Huth. Computational language modeling and the promise of in silico experimentation. *Neurobiology of Language*, 5(1):80–106, 2024.

[9] 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.

[10] Xiao Zhang, Dejing Dou, and Ji Wu. Learning conceptual-contextual embeddings for medical text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9579–9586, 2020.

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

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

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

[14] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

[15] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

[16] 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.

[17] 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.

[18] Chandan Singh, Aliyah R Hsu, Richard Antonello, Shailee Jain, Alexander G Huth, Bin Yu, and Jianfeng Gao. Explaining black box text modules in natural language with language models. *arXiv preprint arXiv:2305.09863*, 2023.

[19] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[20] Finale Doshi-Velez and Been Kim. A roadmap for a rigorous science of interpretability. *arXiv preprint arXiv:1702.08608*, 2017.

[21] Tal Yarkoni. The generalizability crisis. *Behavioral and Brain Sciences*, 45:e1, 2022.

[22] 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.

[23] 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.

[24] OpenAI. GPT-4 technical report. *arXiv:2303.08774*, 2023.

[25] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

[26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, LÃ©lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, TimothÃ©e Lacroix, and William El Sayed. Mistral 7b, 2023.

[27] AI@Meta. Llama 3 model card. 2024.

[28] Kyoung-Rok Jang, Junmo Kang, Giwon Hong, Sung-Hyon Myaeng, Joohee Park, Taewon Yoon, and Heecheol Seo. Ultra-high dimensional sparse representations with binarization for efficient text retrieval. *arXiv preprint arXiv:2104.07198*, 2021.

[29] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

[30] Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*, 2024.

[31] Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. Gecko: Versatile text embeddings distilled from large language models, 2024.

[32] Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Promptreps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval, 2024.

[33] Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron C Wallace. Chill: zero-shot custom interpretable feature extraction from clinical notes with large language models. *arXiv preprint arXiv:2302.12343*, 2023.

[34] Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-design text classification with iteratively generated concept bottleneck. *arXiv preprint arXiv:2310.19660*, 2023.

[35] Chandan Singh, John Morris, Alexander M Rush, Jianfeng Gao, and Yuntian Deng. Tree prompting: Efficient task adaptation without fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6253–6267, 2023.

[36] Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, and Chris Callison-Burch. Learning interpretable style embeddings via prompting llms. *arXiv preprint arXiv:2305.12696*, 2023.

[37] Letian Peng, Yuwei Zhang, Zilong Wang, Jayanth Srinivasa, Gaowen Liu, Zihan Wang, and Jingbo Shang. Answer is all you need: Instruction-following text embedding via answering the question. *arXiv preprint arXiv:2402.09642*, 2024.

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[39] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[40] Chandan Singh, Armin Askari, Rich Caruana, and Jianfeng Gao. Augmenting interpretable models with large language models during training. *Nature Communications*, 14(1):7913, 2023.

[41] 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.

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

[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, 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.

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

[46] Nicholas Bai, Rahul A Iyer, Tuomas Oikarinen, and Tsui-Wei Weng. Describe-and-dissect: Interpreting neurons in vision networks with language models. *arXiv preprint arXiv:2403.13771*, 2024.

[47] Sarah Schwettmann, Evan Hernandez, David Bau, Samuel Klein, Jacob Andreas, and Antonio Torralba. Toward a visual concept vocabulary for gan latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6804–6812, 2021.

[48] Ruochen Zhao, Shafiq Joty, Yongjie Wang, and Tan Wang. Explaining language models' predictions with high-impact concepts. *arXiv preprint arXiv:2305.02160*, 2023.

[49] Yibo Jiang, Bryon Aragam, and Victor Veitch. Uncovering meanings of embeddings via partial orthogonality. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] Juri Opitz and Anette Frank. Sbert studies meaning representations: Decomposing sentence embeddings into explainable semantic features. *arXiv preprint arXiv:2206.07023*, 2022.

[51] Tuomas Oikarinen and Tsui-Wei Weng. Clip-dissect: Automatic description of neuron representations in deep vision networks. *arXiv preprint arXiv:2204.10965*, 2022.

[52] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

[53] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *arXiv preprint arXiv:2402.10376*, 2024.

[54] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197, 2023.

[55] Aya Abdelsalam Ismail, Julius Adebayo, Hector Corrada Bravo, Stephen Ra, and Kyunghyun Cho. Concept bottleneck generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

[56] Ruoyu Chen, Jingzhi Li, Hua Zhang, Changchong Sheng, Li Liu, and Xiaochun Cao. Sim2word: Explaining similarity with representative attribute words via counterfactual explanations. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(6), jul 2023.

[57] Karthikeyan Natesan Ramamurthy, Amit Dhurandhar, Dennis Wei, and Zaid Bin Tariq. Analogies and feature attributions for model agnostic explanation of similarity learners. *arXiv preprint arXiv:2202.01153*, 2022.

[58] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

[59] Richard Antonello, Aditya Vaidya, and Alexander Huth. Scaling laws for language encoding models in fmri. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[61] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014. Association for Computational Linguistics.

[62] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[63] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, March 2022. Number: 3 Publisher: Nature Publishing Group.

[64] John T. Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8(1):427–446, 2022.

[65] Shailee Jain, Vy A. Vo, Leila Wehbe, and Alexander G. Huth. Computational Language Modeling and the Promise of in Silico Experimentation. *Neurobiology of Language*, pages 1–27, March 2023.

[66] 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.

[67] Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13738–13749. Curran Associates, Inc., 2020.

[68] Catherine Chen, Tom Dupré la Tour, Jack Gallant, Daniel Klein, and Fatma Deniz. The cortical representation of language timescales is shared between reading and listening. *bioRxiv*, pages 2023–01, 2023.

[69] Sara F Popham, Alexander G Huth, Natalia Y Bilenko, Fatma Deniz, James S Gao, Anwar O Nunez-Elizalde, and Jack L Gallant. Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature neuroscience*, 24(11):1628–1636, 2021.

[70] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, pages 1336–1348. PMLR, July 2021. ISSN: 2640-3498.

[71] Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *bioRxiv*, pages 2023–05, 2023.

[72] Aniketh Janardhan Reddy and Leila Wehbe. Can fMRI reveal the representation of syntactic structure in the brain? preprint, Neuroscience, June 2020.

[73] Alexandre Pasquiou, Yair Lakretz, Bertrand Thirion, and Christophe Pallier. Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context, February 2023. arXiv:2302.14389 [cs].

[74] Khai Loong Aw and Mariya Toneva. Training language models for deeper understanding improves brain alignment, December 2022. arXiv:2212.10898 [cs, q-bio].

[75] Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. Technical report, bioRxiv, June 2022. Section: New Results Type: article.

[76] Subba Reddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models, December 2022. arXiv:2212.08094 [cs, q-bio].

[77] Richard Antonello, Chandan Singh, Shailee Jain, Aliyah Hsu, Jianfeng Gao, Bin Yu, and Alexander Huth. A generative framework to bridge data-driven models and scientific theories in language neuroscience, 2024.

[78] 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.

[79] 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.

[80] Fabian Pedregosa, Ga ë l Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12(Oct):2825–2830, 2011.

[81] 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.

[82] Evelina Fedorenko, Anna A Ivanova, and Tamar I Regev. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, pages 1–24, 2024.

[83] Joshua B Julian, Jack Ryan, Roy H Hamilton, and Russell A Epstein. The occipital place area is causally involved in representing environmental boundaries during navigation. *Current Biology*, 26(8):1104–1109, 2016.

[84] Anna S Mitchell, Rafal Czajkowski, Ningyu Zhang, Kate Jeffery, and Andrew JD Nelson. Retrosplenial cortex and its role in spatial cognition. *Brain and neuroscience advances*, 2:2398212818757098, 2018.

[85] Ilenia Salsano, Valerio Santangelo, and Emiliano Macaluso. The lateral intraparietal sulcus takes viewpoint changes into account during memory-guided attention in natural scenes. *Brain Structure and Function*, 226(4):989–1006, 2021.

[86] Saima Malik-Moraleda, Dima Ayyash, Jeanne Gallée, Josef Affourtit, Malte Hoffmann, Zachary Mineroff, Olessia Jouravlev, and Evelina Fedorenko. An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, 25(8):1014–1019, 2022.

[87] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[88] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[89] 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.

[90] 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.

[91] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human-generated machine reading comprehension dataset. 2016.

[92] P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65, 2014.

[93] 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, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[94] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[95] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[96] Anton Korinek. Language models and cognitive automation for economic research. Technical report, National Bureau of Economic Research, 2023.

[97] Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. *arXiv preprint arXiv:2401.18059*, 2024.

[98] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[99] 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.

[100] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*, 2023.

[101] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. Promptagent: Strategic planning with language models enables expert-level prompt optimization. *arXiv preprint arXiv:2310.16427*, 2023.

[102] Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

[103] Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. Deductive closure training of language models for coherence, accuracy, and updatability. *arXiv preprint arXiv:2401.08574*, 2024.

[104] Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. Towards consistent natural-language explanations via explanation-consistency finetuning. *arXiv preprint arXiv:2401.13986*, 2024.

[105] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

A Appendix

A.1 fMRI question details

Table A1: Questions list for model with 29 questions. Importance denotes the average absolute coefficient for each question (normalized by the importance of the top question).

Question	Importance
Is the sentence expressing skepticism or disbelief towards something or someone?	1.000
Does the sentence include dialogue?	0.983
Does the sentence describe a relationship between people?	0.924
Does the sentence involve the mention of a specific object or item?	0.900
Does the sentence include technical or specialized terminology?	0.882
Does the sentence contain a proper noun?	0.861
Does the input involve planning or organizing?	0.861
Does the sentence include numerical information?	0.850
Is time mentioned in the input?	0.844
Is the sentence grammatically complex?	0.815
Does the sentence include dialogue or thoughts directed towards another character?	0.811
Does the sentence describe a physical action?	0.809
Does the sentence include a conditional clause?	0.782
Does the sentence describe a visual experience or scene?	0.771
Does the input include a philosophical or reflective thought?	0.759
Is the sentence conveying the narrator's physical movement or action in detail?	0.749
Does the sentence describe a physical sensation?	0.744
Does the sentence involve a discussion about personal or social values?	0.739
Does the sentence reference a specific time or date?	0.719
Does the sentence express a philosophical or existential query or observation?	0.705
Does the sentence involve a description of physical environment or setting?	0.693
Does the input describe a sensory experience?	0.688
Does the sentence involve planning or decision-making?	0.684
Is the sentence a command?	0.682
Does the sentence describe a specific sensation or feeling?	0.672
Does the sentence contain a cultural reference?	0.667
Does the input include dialogue between characters?	0.594
Does the sentence mention a specific location or place?	0.547
Does the sentence reference a specific location or place?	0.545

A.2 fMRI prediction results extended

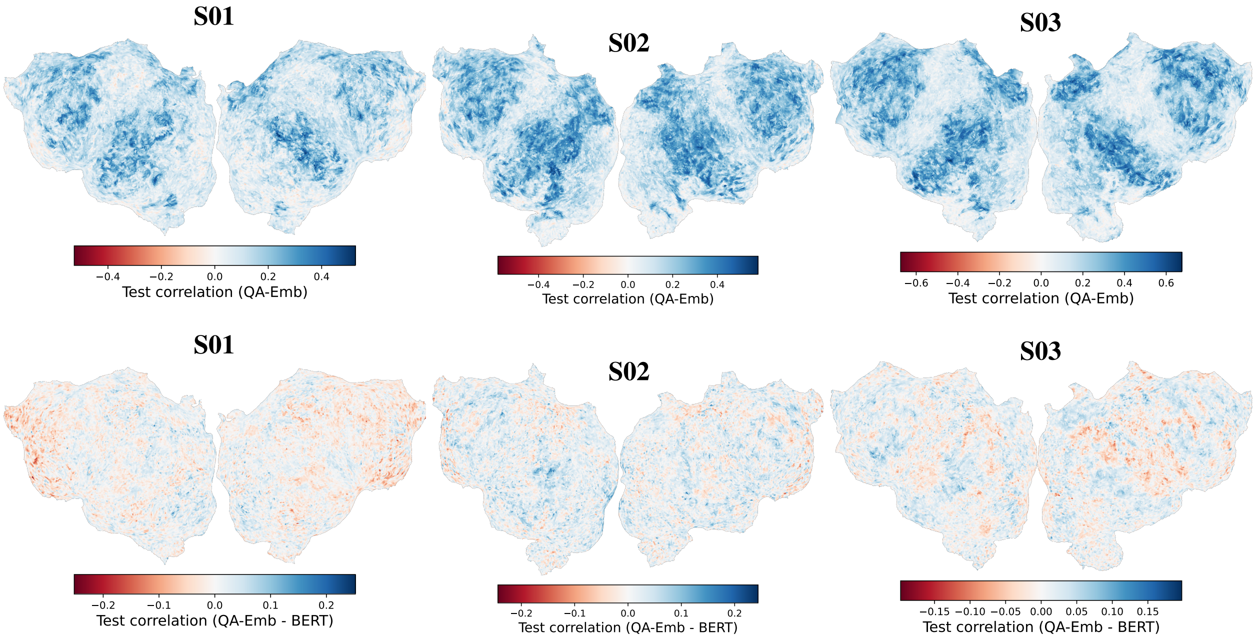
S01
S02
S03
−0.4 −0.2 0.0 0.2 0.4
Test correlation (QA-Emb)
−0.4 −0.2 0.0 0.2 0.4
Test correlation (QA-Emb)
−0.6 −0.4 −0.2 0.0 0.2 0.4 0.6
Test correlation (QA-Emb)
S01
S02
S03
−0.2 −0.1 0.0 0.1 0.2
Test correlation (QA-Emb - BERT)
−0.2 −0.1 0.0 0.1 0.2
Test correlation (QA-Emb - BERT)
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15
Test correlation (QA-Emb - BERT)

Figure A1: Predictive performance for QA-Emb (top row) and the difference between QA-Emb and BERT (bottom row).

Table A2: Mean test correlation for QA-Emb with different settings: varying the underlying prompts to source questions and the LLM used to answer the questions (fixing the number of time-lagged delays to 8). Ensemble generally provides a small boost over other models and Mistral slightly underperforms LLaMA-3 (8B).

		Ensemble	LLaMA-3 (8B)	LLaMA-3 (8B)-fewshot	Mistral (7B)
Subject	Questions				
S01	Prompts 1-3 (376 questions)	0.081	0.078	0.078	0.076
	Prompts 1-5 (518 questions)	0.089	0.085	0.085	0.082
	Prompts 1-6 (674 questions)	0.084	0.081	0.085	0.076
S02	Prompts 1-3 (376 questions)	0.120	0.112	0.119	0.112
	Prompts 1-5 (518 questions)	0.118	0.120	0.121	0.114
	Prompts 1-6 (674 questions)	0.124	0.119	0.121	0.108
S03	Prompts 1-3 (376 questions)	0.132	0.131	0.127	0.126
	Prompts 1-5 (518 questions)	0.137	0.136	0.135	0.129
	Prompts 1-6 (674 questions)	0.141	0.136	0.136	0.132
AVG	Prompts 1-3 (376 questions)	0.111	0.107	0.108	0.104
	Prompts 1-5 (518 questions)	0.115	0.114	0.114	0.108
	Prompts 1-6 (674 questions)	0.116	0.112	0.114	0.105

Table A3: Mean test correlation for different baseline models as a function of hyperparameters (number of time-lagged delays and layer for extracting embeddings)

Subject	S01			S02			S03			AVG		
Delays	4	8	12	4	8	12	4	8	12	4	8	12
BERT	0.084	0.080	0.075	0.114	0.108	0.107	0.136	0.139	0.136	0.111	0.109	0.106
Eng1000	0.079	0.067	0.077	0.096	0.092	0.082	0.110	0.117	0.116	0.095	0.092	0.092
LLaMA-2 (70B) (lay 12)	0.055	0.055	0.054	0.101	0.095	0.085	0.143	0.144	0.130	0.100	0.098	0.089
LLaMA-2 (70B) (lay 24)	0.075	0.059	0.049	0.097	0.104	0.092	0.149	0.153	0.152	0.107	0.105	0.098
LLaMA-2 (70B) (lay 36)	0.058	0.068	0.057	0.131	0.101	0.084	0.153	0.156	0.152	0.114	0.108	0.098
LLaMA-2 (70B) (lay 48)	0.093	0.060	0.052	0.114	0.094	0.091	0.148	0.151	0.149	0.118	0.102	0.098
LLaMA-2 (70B) (lay 60)	0.095	0.048	0.050	0.119	0.089	0.088	0.148	0.152	0.150	0.121	0.097	0.096
LLaMA-2 (7B) (lay 06)	0.074	0.067	0.039	0.120	0.088	0.084	0.138	0.144	0.133	0.111	0.100	0.085
LLaMA-2 (7B) (lay 12)	0.097	0.058	0.053	0.116	0.111	0.087	0.150	0.155	0.152	0.121	0.108	0.097
LLaMA-2 (7B) (lay 18)	0.079	0.076	0.042	0.123	0.103	0.090	0.143	0.153	0.150	0.115	0.111	0.094
LLaMA-2 (7B) (lay 24)	0.088	0.057	0.068	0.129	0.100	0.106	0.144	0.148	0.149	0.120	0.102	0.108
LLaMA-2 (7B) (lay 30)	0.057	0.045	0.045	0.130	0.098	0.099	0.139	0.149	0.148	0.109	0.097	0.097
LLaMA-3 (8B) (lay 06)	0.071	0.066	0.054	0.122	0.119	0.095	0.144	0.147	0.148	0.112	0.111	0.099
LLaMA-3 (8B) (lay 12)	0.089	0.073	0.050	0.110	0.099	0.095	0.146	0.151	0.153	0.115	0.108	0.099
LLaMA-3 (8B) (lay 18)	0.073	0.052	0.052	0.125	0.102	0.096	0.153	0.154	0.155	0.117	0.103	0.101
LLaMA-3 (8B) (lay 24)	0.090	0.053	0.047	0.106	0.113	0.095	0.146	0.149	0.148	0.114	0.105	0.097
LLaMA-3 (8B) (lay 30)	0.082	0.066	0.060	0.120	0.117	0.101	0.147	0.151	0.148	0.117	0.111	0.103

A.3 Prompts

A.3.1 Prompts for question generation

Prompt 1 *Generate a bulleted list of 500 diverse, non-overlapping questions that can be used to classify an input based on its semantic properties. Phrase the questions in diverse ways.*

Here are some example questions:
{{examples}}
Return only a bulleted list of questions and nothing else

Prompt 2 *Generate a bulleted list of 100 diverse, non-overlapping questions that can be used to classify sentences from a first-person story. Phrase the questions in diverse ways.*

Here are some example questions:
{{examples}}
Return only a bulleted list of questions and nothing else

Prompt 3 *Generate a bulleted list of 200 diverse, non-overlapping questions that can be used to classify sentences from a first-person story. Phrase the questions in diverse ways.*

Here are some example questions:
{{examples}}
Return only a bulleted list of questions and nothing else

Prompt 4 *Based on what you know from the neuroscience and psychology literature, generate a bulleted list of 100 diverse, non-overlapping yes/no questions that ask about properties of a sentence that might be important for predicting brain activity.*

Return only a bulleted list of questions and nothing else

Prompt 5 *# Example narrative sentences*
{{example sentences from dataset}}

# Example yes/no questions
{{example questions already asked}}

Generate a bulleted list of 100 specific, non-overlapping yes/no questions that ask about aspects of the example narrative sentences that are important for classifying them. Focus on the given narrative sentences and form questions that combine shared properties from multiple sentences above. Do not repeat information in the example questions that are already given above. Instead, generate complementary questions that are not covered by the example questions. Return only a bulleted list of questions and nothing else.

Prompt 6 *Generate more diverse questions that may occur for a single sentence in a first-person narrative story*

See exact prompts with examples in the Github repo.

A.3.2 Prompts for question answering

Standard prompt *<User>: Input text: {example}*
Question: {question}
Answer with yes or no, then give an explanation.

Few-shot prompt *<System>: You are a concise, helpful assistant.*
<User>: Input text: and i just kept on laughing because it was so
Question: Does the input mention laughter?
Answer with Yes or No.
<Assistant>: Yes
<User> Input text: what a crazy day things just kept on happening
Question: Is the sentence related to food preparation?
Answer with Yes or No.
<Assistant>: No
<User> Input text: i felt like a fly on the wall just waiting for
Question: Does the text use a metaphor or figurative language?
Answer with Yes or No.
<Assistant>: Yes
<User> Input text: he takes too long in there getting the pans from
Question: Is there a reference to sports?
Answer with Yes or No.
Answer with Yes or No.
<Assistant>: No
<User> Input text: was silent and lovely and there was no sound except
Question: Is the sentence expressing confusion or uncertainty?
Answer with Yes or No.
<Assistant>: No
<User> Input text: {example}
Question: {question}
Answer with Yes or No.
<Assistant>:

See exact prompts with examples in the Github repo.

A.4 Information retrieval details

Optimization details When fitting our QA-Emb model for information retrieval, we learn a single scalar per-question that is multiplied by each embedding before computing a similarity. To learn these scalars, we minimize a two-part loss. The first loss is the negative cosine similarity between each query and its similar documents. The second loss is the cosine similarity between each query and the remaining documents. We weight the first loss as 10 times higher than the second loss and optimize using Adam [105] with a learning rate of 10^{-4}. We run for 8 epochs, when the training loss seems to plateau.

Table A4: 54 binary classification datasets along with their underlying yes/no question and corpus statistics from a recent collection [89, 90].

Dataset name	Dataset topic	Underlying yes/no question	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.5 Details on question-answering evaluation datasets

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are clearly stated that QA-Emb can generate flexible embeddings using a pre-trained LLM (described in the Sec. 2) and that this experimentally improves performance primarily for an fMRI prediction problem (Sec. 4).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, limitations are discussed in the methods section (Sec. 2) as well as the entirety of Sec. 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper includes no theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, all details are fully provided in the paper (including extra details such as prompts in the supplementary material).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code and data are made openly available on Github at at https://anonymous.4open.science/r/interpretable-embeddings-70ED/readme.md.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes all specifications are given in the Methods and the experimental setup sections in the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes we give details about error bars in our results section. Note that Fig. 2 doesn't directly show error bars since they are small enough that they are within the points (see Fig. 2 caption).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Mentioned in the methods section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and found that the current paper conforms to it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Given in the discussion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks beyond those discussed in the Broader Impacts above, as it releases no new data or models.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We list the licenses in the main text for all datasets where they can be found (the main fMRI data is released under the MIT license).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects are involved in this study. The fMRI data analyzed here is collected in previous studies following the appropriate IRB protocols.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects are involved in this study. The fMRI data analyzed here is collected in previous studies following the appropriate IRB protocols.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.