

## A APPENDIX

### A.1 Dataset Information

Huth's dataset and the Narratives dataset both contain fMRI responses recorded while participants listened to English auditory language stimuli of spoken stories. Huth's dataset comprises data from 8 participants, with each participant listening to a total of 27 stories. As a result, each participant contributed approximately 6 hours of neural data, amounting to 9,244 time repetitions (TRs), i.e., the time frames for fMRI data acquisition. On the other hand, the Narratives dataset initially included a total of 365 participants. However, due to the significantly high computational demand, we selected a subset of 8 individuals who had engaged in at least 4 stories, with an average of 2,109 TRs collected from each participant. Pereira's dataset collects participants' fMRI signals while viewing English visual stimuli composed of Wikipedia-style sentences. In line with previous research by Luo et al. [31], we selected cognitive data from participants who took part in both experiments 2 and 3. This subset consists of 5 participants, each of whom watched 627 sentences selected from 177 passages. Each sentence corresponds to one TR, which represents one scan of fMRI data consisting of signals from approximately 10,000 to 100,000 voxels. The statistics of these datasets are provided in Table 5. All datasets received approval from ethics committees and are accessible for research purposes. We present the overall statistics of the above three fMRI datasets in Table 5.

### A.2 Dataset preprocessing

*Document corpus construction.* Pereira's dataset has a natural segmentation of documents, with approximately 3 to 4 sentences per document. Therefore, we utilized its inherent segmentation for our experiment. After defining the document corpus, we utilize the same protocol to select a query and the next token prediction task construction. So each query  $Q$  is either a piece of sentence in Pereira's dataset or a text span corresponding to a TR. For Huth's dataset and the Narratives dataset, the language stimuli are presented continuously without any natural document segmentation provided. Hence, we segment text spans presented in every 10 consecutive TRs as a document. This segmentation criterion results in an average document length similar to the passage length found in existing IR benchmarks, such as MS MARCO [5] (see Section A.1 for detailed statistics). According to the segmentation, the average document length is about 60, which is similar to the passage length of existing IR datasets, like MS MARCO [5], which was used to train our baseline RepLLaMA.

*Query construction.* Following existing research in language decoding from brain signals [50, 54], we split the text stimuli to construct the query according to the TR. For Pereira's dataset, we split each sentence into three parts with equal length. Two unique data samples are constructed by treating (i) the first third as the query and the second third as the ground truth continuation as well as (ii) combining the first two thirds as the query and using the last third as the ground truth continuation. For Huth's dataset and the Narratives dataset, we segmented the data by considering the perceived textual content during each TR as the ground truth continuation. We then truncated the preceding text and used it as the query. The truncation

is accomplished using a sliding window ranging from 1 to 3 TRs to pick the language stimuli. We detail the average length of the queries, the query continuations, and the length of documents in Section A.1. The statistics of the query generation task and the document ranking task are presented in Table 6.

### A.3 Query performance features

To study the effect of brain signals in query augmentation in queries with different features. We analyze the document ranking performance according to the original queries measured by the following features:

(1) Averaged ICTF (inverse collection term frequency) [8]: ICTF is a popular measure for the relative importance of the query terms and is usually measured by the following formulas:

$$ICTF(w) = \log\left(\frac{|D|}{TF(w, D)}\right) \quad (7)$$

where  $|D|$  is the number of all terms in collection  $D$ , and  $TF(w, D)$  is the term frequency (number of occurrences) of term  $w$  in  $D$ . Here we use the averaged ICTF of all terms  $w$  in the query.

(2) Averaged IDF (inverse document frequency) [17]: IDF is another widely used measure for the importance of the query terms and is typically measured by the following formulas:

$$IDF(w) = \log\left(\frac{N}{N_w}\right) \quad (8)$$

where  $N$  is the number of documents in the collection and  $N_w$  is the number of documents containing the term  $w$ . Here we use the averaged IDF of all terms  $w$  in the query.

(3) Specificity (or simplified clarity score) [11]: Specificity score measures the Kullback-Leibler divergence of the query's language model from the collection's language model, which can be formulated as:

$$Specificity = \sum_{w \in q} P(w | q) \log\left(\frac{P(w | q)}{P(w | D)}\right) \quad (9)$$

where  $P(w | q)$  and  $P(w | D)$  indicate the token possibility in the query and the document, respectively.

(4) Clarify [11]: Clarify score quantifies the ambiguity of a query w.r.t. a collection of documents. It measures the KL divergence between a relevance model induced from top-ranked documents retrieved by the original query.

$$Clarify(q, D_{q:M}^k) = \sum_{w \in V} P(w | D_{q:M}^k) \frac{P(w | D_{q:M}^k)}{P(w | D)} \quad (10)$$

where  $w$  and  $V$  denote a query term and the entire collection vocabulary, respectively,  $D_{q:M}^k$  indicates the top- $k$  document retrieved by model  $M$  using query  $q$ . The conjecture suggests that a larger KL divergence corresponds to a more clarified query and a better retrieval quality.

### A.4 Implementation Details

To efficiently manage and analyze the high-dimensional fMRI data, we employ two methods to reduce dimensionality. For Huth's dataset and Narratives dataset, we select features from brain regions identified by Musso et al. [40], which are known to be relevant to language

Table 5: Overall statistics of fMRI datasets.

Dataset	#Participants	#Total duration	#Duration per participant	#Total TRs	#TRs per participant	#Total words	#Words per participant
Pereira's	5	7.0 h	1.4 h	3,135	627	38,650	7,730
Huth's	8	3.5 days	10 h	122,992	15,374	427,296	53,412
Narratives	8	7.5h	56 min	16,868	2,109	80,160	10,020

Table 6: Overall statistics of the document corpus and query set constructed with the fMRI datasets.

Dataset	#Query	#Document	Query length	Continuation length	Doc length
Pereira's	1,254	168	5.8±2.5	4.5±1.5	46±6
Huth's	26,578	876	10.3±4.3	7.4±0.5	61.2±13
Narratives	4,979	162	9.5±4.7	6.0±1.9	60.0±23.5

Table 7: Examples of document ranking with BM25 using the original query or the augmented query in Huth's and Narratives dataset. Text in **blue** and in **purple** indicates content in the original query and generated by the query augmentation method, respectively.

Dataset	Method	Query Content	Top-ranked document	Relevance
Huth's	Original	with <b>one hand tied behind</b>	cup holder and gets ready to <b>hand</b> him some change and ... if he got a cellphone I gotta get <b>one</b> ...	0
	Unsup-Aug	with <b>one hand tied behind</b> <b>my eyes shut</b>	... like we're gonna hit and <b>I</b> just did the only thing <b>I</b> thought seemed right <b>I</b> just <b>shut my eyes</b> ...	0
	RS Brain	with <b>one hand tied behind</b> <b>thinking and what he's gonna</b>	... <b>he</b> just yells to me <b>his</b> like we're <b>gonna</b> hit and I just did the only thing I <b>thought</b> seemed right I just shut my eyes I took a deep	0
	Brain-Aug	with <b>one hand tied behind</b> <b>my back and I'm thinking</b>	<b>my back</b> which <b>I</b> only probably ever would have to do with ... they were a <b>handful</b> she was paying ten dollars an hour in nineteen eighty eight I kind of <b>thought</b> that all of <b>my</b>	1
Narratives	Original	<b>you get undressed and get into</b>	gentlemen <b>you</b> can't <b>get</b> away with this sooner or later somebody the or somebody is going to <b>get</b> wind of this madness ...	0
	Unsup-Aug	<b>you get undressed and get into</b> <b>somebody going away</b>	gentlemen <b>you</b> can't <b>get away</b> with this sooner or later somebody the or <b>somebody</b> is going to get wind of this madness ...	0
	RS Brain	<b>you get undressed and get into</b> <b>the bathtub and I'll wash</b>	<b>you</b> just come with <b>me</b> where <b>into</b> the tunnel <b>I'll</b> show you henry swanson led guy to a small hole on the ...	0
	Brain-Aug	<b>you get undressed and get into</b> <b>bed and I'll join you</b>	... now Arthur listen <b>I</b> say this in all sincerity will <b>bed</b> like a good guy and relax ...	1

Table 8: Comparison of different query augmentations methods on various datasets in terms of MAP.

Dataset	Query	RepLLaMa-Rocchio	BERT-QE	RepLLaMa-QE
Pereira's	original	0.855	0.765	0.848
Huth's	original	0.276	0.238	0.278
Narratives	original	0.356	0.341	0.352

processing in the human brain. For Pereira's dataset, we apply component analysis [1] on the original fMRI features to reduce the dimensionality to 1000. The 7B version of the Llama-2 model [52]

released in Huggingface <sup>2</sup> is adopted as the language model for generating the query continuation.

<sup>2</sup><https://huggingface.co/models>

We train Brain-Aug with the Adam optimizer [22] using a learning rate of  $1 \times 10^{-4}$  and a batch size of 8. The learning rate is selected from the set  $\{1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$  based on the experimental performance on Pereira’s dataset. The training of the warm-up step is stopped after ten epochs, while an early stop strategy was adopted in the training of the next token prediction task when no improvement was observed on the validation set for ten epochs. The entire training process was conducted on 16 A100 graphics processing units with 40 GB of memory and took approximately 12 hours to complete. During the inference stage, we utilize a beam search protocol with a width of 5.

When performing query generation for document ranking, we set the maximum number of words that can be expanded to 5. In Pereira’s dataset, the continuation will be 5 tokens unless the model generates a token indicating the end of the continuation. In the other two datasets, due to their higher perplexity, the model may generate content with lower quality. Therefore, during the generation process, we calculate the perplexity of the content generated up to the current step (note that this is the perplexity of the generated content, not the ground truth label). If the averaged perplexity at the current step exceeds a threshold of 1.5, the generation process is early stopped. Due to the fact that the queries constructed based on the above method can be quite long in Huth’s dataset and the Narratives dataset, in order to simulate real-world query submission scenarios, we randomly sampled 3 query terms from the original queries (when the query consists of more than 3 terms) when constructing the ranking tasks.

## A.5 Variants of Unsup-Aug

In addition to adopting RM3 and Rocchio as Unsup-Aug baselines, we include a recently proposed method BERT-QE [60] and its variant which replaces the BERT-based retrieval model with a more powerful retrieval model RepLLaMA (named RepLLaMA-QE). Table 8 is the result in terms of MAP. We observe that there is no significant performance difference between RepLLaMa-Rocchio and RepLLaMa-QE. However, the combined model Unsup+Brain-Aug shows significant improvement after the addition of brain signals.

On the other hand, we didn’t investigate the query augmentation method with user interaction because we focus on query augmentation in the query submission stage so we can facilitate the retrieval performance before a potentially bad user experience happens. However, investigating the combination of Brain-Aug and existing query augmentation methods with user interactions is a promising direction for future work.

## A.6 Example cases

We present the manually selected example cases in Huth’s and Narratives’s dataset in Table 7. In these cases, Brain-Aug leverages brain signals and ranks the relevant document as top-1. The selection of these examples was based on the higher NDCG@1 scores of the Brain-Aug compared to the baselines and controls. More cases can be found in the provided repository.

## A.7 Failures and Insights

In our research, we have also conducted two meaningful attempts, despite being unsuccessful, may provide insights for further research.

The first attempt was to explore whether EEG signals can be utilized for Brain-Aug, as EEG signals are easier to collect in real-world scenarios than fMRI. However, we found that in our experiment with two public EEG datasets, i.e., UERCM<sup>3</sup> and Zuco<sup>4</sup>, Brain-Aug did not outperform RS Brain. This implies that the existing quality of EEG data have limitations in their ability to decode semantics with Brain-Aug. The second attempt was to train a query augmentation model with brain signals to directly facilitate the document ranking task. We constructed the unified prompts using the same method of Brain-Aug and fed them into Repllama to obtain query representations. Then, we used a contrastive loss function to make these representations closer to the relevant documents. We found that training the model in this way makes it challenging to generalize the performance to the validation set. This could be potentially attributed to the label-inefficient issue in dense retrieval training settings. Future research can further explore this direction.

## A.8 AI assistants usage

After completing the paper, we employ ChatGPT<sup>5</sup> and Gemini<sup>6</sup> to identify writing typos. Subsequently, manual review and revision are performed to address these typos.

<sup>3</sup><https://github.com/YeZiyi1998/UERCM>

<sup>4</sup><https://osf.io/2urht/>

<sup>5</sup><https://chat.openai.com/>

<sup>6</sup><https://gemini.google.com/app>