Query Augmentation by Decoding Semantics from Brain Signals

Anonymous ACL submission

Abstract

Query augmentation is a crucial technique for refining semantically imprecise queries. Traditionally, query augmentation relies on extracting information from initially retrieved, potentially relevant documents. If the quality of the initially retrieved documents is low, then the effectiveness of query augmentation would be limited as well. We propose Brain-Aug, which enhances a query by incorporating semantic information decoded from brain signals. Brain-Aug generates the continuation of the original query with a prompt constructed with brain signal information and a ranking-oriented inference approach. Experimental results on fMRI (functional magnetic resonance imaging) datasets show that Brain-Aug produces semantically more accurate queries, leading to improved document ranking performance. Such improvement brought by brain signals is particularly notable for ambiguous queries.

1 Introduction

Understanding users' intentions is the key to the effectiveness of search engines. However, search engine users often struggle to precisely express their information needs, resulting in queries that are short (Kacprzak et al., 2017), vague (Yano et al., 2016; Cronen-Townsend et al., 2002), or inaccurately phrased, which compromise the retrieval effectiveness. To address this problem, query augmentation emerges as a crucial technique to refine the original queries into more effective expressions (Lavrenko and Croft, 2017; Mei et al., 2008). Traditionally, this reformulation process relies heavily on external document information such as expanding the query with contents from documents users have engaged with (Chen et al., 2021; Ahmad et al., 2019; Pereira et al., 2020).

The advent of neurophysiological interfaces offers a novel source of data to understand users' search intentions (Ye et al., 2022b; Michalkova et al., 2024). In information retrieval (IR) scenarios, several studies have revealed that brain signals can be used to predict users' relevance perception (Ye et al., 2022c; Eugster et al., 2014; Pinkosova et al., 2020) and cognitive state (Moshfeghi et al., 2016). These advances open new avenues in using brain signals as an alternative to conventional signals for query augmentation. Existing studies have investigated the use of brain signals to predict the relevance of perceived input (Eugster et al., 2016), which can be further used to extract relevant content for query augmentation (Ye et al., 2022a, 2024). The current process of query augmentation still relies on the quality of initially retrieved documents and cannot kick off before potentially unsatisfactory user interactions with those documents.

In this paper, we propose *query augmentation with brain signals* (Brain-Aug), which directly refines queries submitted by users through decoding semantics from their brain signals. With the help of computational language models, Brain-Aug proposes two techniques to effectively refine queries: (i) Prompt construction with brain signals: Brain signals corresponding to the query context are decoded into the language model's latent space to construct prompts accordingly; (ii) Training based on next token prediction and ranking-oriented inference: We teach the model to predict tokens in relevant documents as query continuation during training. Ranking-oriented features, i.e., inverse document frequency (IDF), are incorporated to generate effective query continuation that can distinguish different documents during inference.

We conduct experiments on three functional magnetic resonance imaging (fMRI) datasets. Results show that Brain-Aug can accurately generate query continuations for its augmentation and improve the ranking performance. Further investigation delves into different types of queries and shows that brain signals are particularly useful in enhancing the performance of ambiguous queries.

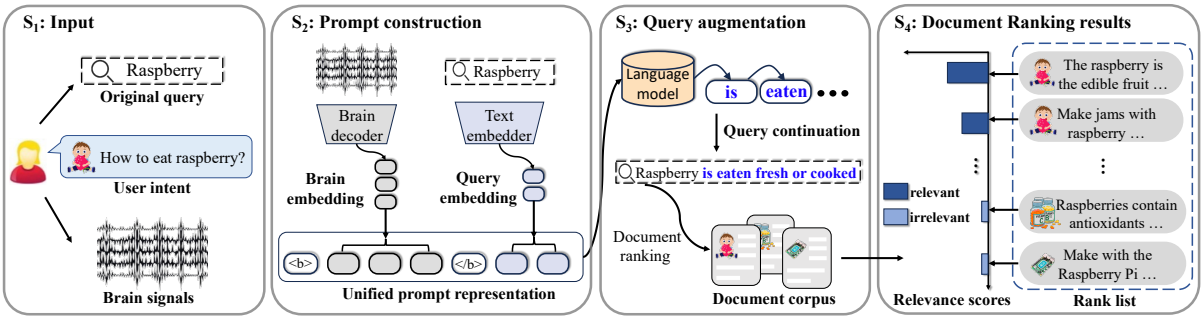


Figure 1: The procedure of query augmentation by decoding semantics from brain signals (Brain-Aug).

2 Related Work

Query augmentation. Traditionally, query augmentation can be categorized into two types: based on pseudo-relevance signals (Bi et al., 2019; Lavrenko and Croft, 2017) and based on user signals (Li et al., 2020). Approaches based on pseudo-relevance signals usually treat top-ranked documents in the initial retrieval step as relevant. Based on these relevant documents, Rocchio Jr (1971) and Lavrenko and Croft (2017) adopt a vector space model and a language model for refining the query representation to be closer to the top-ranked documents, respectively. In contrast, approaches based on user signals usually integrate information from documents the user has previously interacted with or queries they submitted historically. E.g., Chen et al. (2021) and Ahmad et al. (2019) build a sequence model to extract semantic representations from historical clicked documents to refine the query representation. Existing methods, either based on pseudo signals or user signals, are limited by their reliance on the quality of the documents and the accuracy of estimating their relevance.

Neuroscience & IR. There is increasing literature that adopts neuroscientific methods into IR scenarios (Chen et al., 2022; Gwizdka et al., 2017; Mostafa and Gwizdka, 2016). For example, Chen et al. (2022) built a prototype in which users can interact with the search systems with a brain-computer interface. Allegretti et al. (2015); Moshfeghi et al. (2016); Michalkova et al. (2024) conducts a series of work to study the cognitive mechanisms involved in the process of information retrieval. A common finding observed by existing literature is that(Allegretti et al., 2015; Eugster et al., 2014) brain signals can be utilized to as a relevance indicator. This indicator can be employed for query rewriting (Ye et al., 2022a; Eugster et al., 2016). Although this paradigm has been shown to be effective, it still relies on the quality of the retrieved documents. On the other hand, other studies have demonstrated that semantics could be decoded to some extent with brain signals (Wang and Ji, 2022) such as fMRI (Xi et al., 2023; Ye et al., 2023; Zou et al., 2021) and magnetoencephalogram (MEG) (Défossez et al., 2023). However, there is currently a lack of research investigating the utilization of the decoded semantics for query augmentation.

3 Method

We first formalize the query augmentation task and then present Brain-Aug.

3.1 Task formalization

The *input* to the task of augmenting queries with brain signals is a query submitted by a user plus the brain signals associated with the query context. We use Q to denote the query that is composed of n tokens, $Q = \{q_1, q_2, \ldots, q_n\}$. We use $B = \{b_1, \ldots, b_t\} \in \mathbb{R}^{t \times c}$ to represent the brain signal, which is a sequence of features extracted from fMRI data, where c is the number of fMRI features and t is the number of time frames in which brain recordings are collected.

Given the input query and brain signals, the *task* is to learn an autoregressive function F to refine the query based on the user's cognitive process. F generates a query continuation $M = \{m_1, ..., m_k\}$, which will be concatenated to the initial query Q as the augmentated query. Let m_i be the i-th token in M, the generation process is formalized as:

$$m_i = F(\{q_1, \ldots, q_n, m_1, \ldots, m_{i-1}\}, B; \Theta), \quad (1)$$

where Θ is the model parameters of F.

The effectiveness of query augmentation is measured *extrinsically* using the document ranking performance. Formally, let $\mathcal{D}$ be a document corpus and G be a ranking model (e.g., BM25 (Robertson et al., 2009), RepLLaMA (Ma et al., 2023)). The ranking model G estimates a ranking score $G(\{Q, M\}, d)$ for each document $d \in \mathcal{D}$ and the

document ranking performance can be measured by a ranking-based metric such as normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002) or mean average precision (MAP) (Järvelin and Kekäläinen, 2017).

3.2 Overall procedure

Fig. 1 provides an overview of the four-stage process of Brain-Aug: S_1 : Input to Brain-Aug consists of the original query and brain signals associated with the user's cognitive response within the query context. S_2 : Then a brain decoder is trained to align the representations of brain signals with the representation space of text embedding in the language model. This allows for creating a unified prompt representation that jointly models the brain responses and original queries. S_3 : A language model is adopted to generate the continuation of the original query by using a unified prompt representation. A ranking-oriented inference method is utilized to enhance the generation process to improve the ranking performance. S_4 : In this case, the original query "Raspberry" (sampled from Pereira's dataset in our experiment) is augmented to "Raspberry is eaten fresh or cooked". Consequently, documents with a focus on the subtopic of "eating raspberry" are ranked higher than those on "raspberry's nutrition" or "raspberry Pi".

3.3 Prompt construction

Motivated by existing literature that combines multimodal information as prompt (Ye et al., 2023; Liu et al., 2023a), the prompt for Brain-Aug is constructed by integrating the textual query with cognitive information derived from brain signals. First, the query's text Q is directly fed to the language model's embedding layer f_q to transform the tokens into latent vectors $V^Q = \{v_1^q, \ldots, v_i^q, \ldots, v_n^q\} \in \mathbb{R}^{n \times d}$, where n is the number of tokens, d is the embedding size of the language model.

Second, a brain decoder f_b is devised to embed each brain representation $b_i \in B$ into the same latent space $\mathbb{R}^d$, which can be formulated as $v_i^B = f_b(b_i)$. Based on preliminary empirical comparisons of transformers (Vaswani et al., 2017), linear layer, multilayer perceptron (MLP), and recurrent neural network (RNN), we decide to construct the brain decoder as a deep neural network f_b comprises (i) a MLP network f_m with ReLU (Fukushima, 1980) as the activation function, and (ii) a position embedding $P = \{p_1, \ldots, p_t\} \in \mathbb{R}^{t \times c}$. The position embedding is initialized using a uniform distribution. Element-wise addition is applied where each position embedding $p_i \in P$ is added to its corresponding fMRI features $b_i \in B$. The multi-layer perceptron network f_m is constructed with an input layer and two hidden layers that have the same dimensionality c as the input fMRI features, as well as the output layer with the dimensionality of d. In summary, the fMRI features corresponding to the i-th time frame, i.e., b_i, are fed into the brain decoder f_b, which can be expressed as:

$$v_i^B = f_b(b_i) = f_{mlp}(p_i + b_i). \quad (2)$$

Finally, the brain embedding V^B and the query embedding V^Q are concatenated with embeddings of two special tokens, i.e., $\langle b \rangle$ and $\langle /b \rangle$, marking the beginning and end of the brain embedding, respectively. The two special tokens are randomly initialized as one-dimensional vectors aligned with the dimensional structure of token embeddings in the language model. As a result, the prompt sequence S can be represented as:

$$S = \{\langle b \rangle, v_1^B, \ldots, v_t^B, \langle /b \rangle, v_1^W, \ldots, v_n^W\}. \quad (3)$$

This sequence, integrating both brain information and textual data, can be input to the language model for generating the query continuation.

Prior to the main training task detailed in Section 3.4, a warmup step (Huang et al., 2023) is adopted to align the distribution of the brain embedding with that of the text token's embeddings, ensuring that the brain embedding is primed for integration with the text prompt embedding. To streamline the process and enable training in an unsupervised manner, each $v_i^B \in V^B$ is mapped to the mean value of the corresponding query embeddings, i.e., $\frac{1}{n}\sum_{j=1}^{n} v_j^Q$. Mean square loss (MSE) loss is adopted for the warmup process:

$$L_{MSE} = \frac{1}{t}\sum_{i=1}^{t}\left(v_i^B - \frac{1}{n}\sum_{j=1}^{n} v_j^Q\right)^2. \quad (4)$$

3.4 Training objective

Given the unified prompt S, the training task is selected as the next token prediction task which predicts the continuation of S. The prompt sequence S is fed into a language model, e.g., the 7B version of LLaMA (Touvron et al., 2023) in our implementation. The language model then estimates the likelihood of the ground truth continuation $M^* = \{m_1^*, \ldots, m_k^*\}$ by using an autoregressive function $P_{\text{LM}}(m_i^* \mid \{m_1^*, \ldots, m_{i-1}^*\}, S)$ over the sequence S. The training objective is to maximize the likelihood of generating the ground

truth continuation:

$$\max_{\Theta} = \sum_{i=1}^{k} \log(P_{\text{LM}}(m_i^* \mid \{m_1^*, \ldots, m_{i-1}^*\}, S; \Theta)), \quad (5)$$

where $\Theta = \{\Theta^{LM}, \Theta^{f_b}, \Theta^{sp}\}$ is the model parameters, Θ^{LLM}, Θ^{f_b}, and Θ^{sp} are the parameters of the language model, the brain decoder, and the special tokens $\langle b \rangle$ and $\langle /b \rangle$, respectively.

Here, we propose to set the ground-truth label of the continuation as the content from the labeled relevant documents (see Section 4 for details). First, when a document is relevant, it must contain important information and tokens that can potentially be decoded from the brain signals (Pereira et al., 2018). Second, teaching models to expand queries with terms in potentially relevant documents could improve the performance of downstream retrieval models (Robertson et al., 2009). The training process follows the "prompt tuning" approach (Liu et al., 2023b) by keeping the parameters of the language model unchanged and fine-tuning only the prompt representation, i.e., Θ^{f_b}, and Θ^{sp}. In this way, we can train Brain-Aug efficiently with limited training data.

3.5 Ranking-oriented inference

During the inference stage, the generated continuations should also be able to distinguish between different documents. Therefore, we incorporate the IDF information (Robertson, 2004) of each token in the vocabulary when generating query continuation $\hat{M} = \{\hat{m}_1, \ldots, \hat{m}_k\}$. Let $\text{IDF}(\hat{m})$ be the IDF of token $\hat{m}$, then the generation likelihood of each token in $\hat{m}_i \in \hat{M}$ during the inference stage can be estimated as:

$$P_{\text{inf}}(\hat{m}_i) = \frac{P_{\text{LM}}(\hat{m}_i) + \alpha\,\text{IDF}(\hat{m}_i)}{\sum_{m \in \text{Vocab}}(P_{\text{LM}}(m) + \alpha\,\text{IDF}(m))}, \quad (6)$$

where $P_{LM}(m) = P_{LM}(m \mid \{\hat{m}_1, \ldots, \hat{m}_{i-1}\}, S; \Theta)$ represents the estimated likelihood of the next token m given the previously generated tokens $\{\hat{m}_1, \ldots, \hat{m}_{i-1}\}$, α is a hyperparameter, Vocab indicates the language model's vocabulary. This approach ensures that the query's continuation is not only contextually relevant but also effective in distinguishing documents in the retrieval process.

4 Experimental Setup

Next, we detail our experimental settings, which are designed to address three research questions: **(RQ1)** Is it possible to generate an augmented query with user's brain signals? **(RQ2)** Can we improve document ranking performance using the augmented query? **(RQ3)** How do brain signals improve different queries for document ranking? Together, these questions help us to understand the effectiveness of Brain-Aug to refine a query and improve ranking performance. Below, we describe the datasets and baselines. More implementation details are provided in Section A.4.

4.1 Datasets

Three publicly available fMRI datasets are adopted, namely Pereira's dataset (Pereira et al., 2018), Huth's dataset (LeBel et al., 2023), and the Narratives dataset (Nastase et al., 2021). We process the text stimuli in these datasets to transform them into ranking datasets consists of a document corpus and a set of queries. The dataset information is provided in Section A.1.

4.2 Data processing

Due to the lack of clear definitions for query and document parts in those existing fMRI datasets, we use the inverse cloze test (ICT) setting (Izacard et al., 2021; Lee et al., 2019) to test the query augmentation performance. The ICT setting selects a text span in the document as a pseudo query and the corresponding document is treated as relevant for this query. Formally, for a document $D = \{w_1, \ldots, w_m\}$, ICT extracts a span $Q = \{w_l, w_{l+1}, \ldots, w_r\}$ to form a relevant query-document pair $\{Q, D \backslash Q\}$, where $D \backslash Q = \{w_1, \ldots, w_{l-1}, w_{r+1}, \ldots, w_m\}$.

In Pereira's dataset, each document consists of 3-4 sentences, which are presented to the user as visual stimuli one by one. Due to the length of a sentence being too long as a query, we truncate the first one-third and two-thirds of the sentence to construct two queries for each sentence, resulting in 6-8 relevant query-document pair for each document. In Huth's and Narratives datasets, continuous contents are presented to the user as auditory stimuli. We utilize a fixed time interval of 20 seconds, which corresponds to 10 fMRI scans, to segment the stimuli into documents. Then, smaller time intervals of 2, 4, and 6 seconds are employed to segment queries of varying lengths from the document. We provide more details and statistical data for the document corpus and queries constructed in each dataset in Section A.2.

Due to the variability in brain data across participants, we trained separate models for each participant and evaluated Brain-Aug using a five-fold cross-validation on each participant's data. The data samples are randomly split into five folds ac-

cording to which document they belong to. Each fold of the cross-validation involves selecting one fold of the data as the test set, while the remaining four folds are split into training and validation sets. The sizes of the training, validation, and testing sets were roughly proportional to 3:1:1, respectively.

4.3 Training and evaluation setup

We train Brain-Aug with a next token prediction task. A data sample during this task consists of the query, its ground truth continuation, and corresponding brain signals. The ground truth continuation is selected as the textual content presented within a fixed period of time after the query (see Section A.2 for details). Taking into account the delayed effect of fMRI signals(Mitchell et al., 2008), we collect user's brain signals in a period of several seconds after the user perceives the textual content of the query. During this period, the user's brain representation has the potential to encode semantic information related to the query itself, as well as its continuation.

We first conduct *query generation analysis* to investigate the ability of Brain-Aug to generate query continuation that matches the ground truth label. The logarithm perplexity (Meister and Cotterell, 2021) is used to measure the likelihood of generating the ground truth continuation. The lower perplexity indicates the language model deems the ground truth continuation as more expected. We also investigate language similarity to demonstrate the extent to which the generated continuation is similar to the ground truth using the Rouge score (Lin, 2004).

Next, we augment the original query with its generated continuation and evaluate its performance in terms of *document ranking*. We employ document ranking metrics, including NDCG at different cutoffs (10 and 20) (Järvelin and Kekäläinen, 2002), Recall@20, and MAP (Järvelin and Kekäläinen, 2017).

4.4 Baselines and controls

Given the augmented query, we select two ranking models for document ranking, i.e., a sparse ranking model, **BM25** (Robertson et al., 2009), and a dense ranking model, **RepLLaMA** (Ma et al., 2023). To assess whether Brain-Aug helps document ranking, we compare its document ranking performance with several *baselines* and *controls*.

As *baselines* we select (i) **the original query**, and (ii) the query augmented with pseudo-relevance signals (denoted as **Unsup-Aug**). When using BM25 as the ranking model, we implemented RM3 (Lavrenko and Croft, 2017) as Unsup-Aug, which expands the query by selecting relevant terms from the top-ranked documents in the initial retrieval. When using RepLLaMA as the ranking model, we implement Rocchio (Bi et al., 2019) as Unsup-Aug, which refines the query vector to be closer to the top-ranked documents. (iii) We also reported the additional results by first using Brain-Aug, followed by Unsup-Aug, denoted as **Brain+Unsup**.

As *controls* we select variants or ablations of Brain-Aug. The first control is Brain-Aug without any brain input (denoted as **w/o Brain**), and thus the query continuation is generated solely depending on the original query and the language model. The second control is Brain-Aug with randomly sampled brain input (denoted as **RS Brain**). RS Brain involves sampling brain input that does not correspond to the query but is randomly selected from the same dataset. The last control is Brain-Aug without ranking-oriented generation in which the generation likelihood of each token is estimated without the IDF weight (denoted as **w/o IDF**).

5 Experiments and Results

We first analyze the performance of the generated query continuation by comparing it with the ground truth label. Then we investigate the document ranking performance with Brain-Aug and examine the relationship between query features and their ranking performance.

5.1 Query generation analysis

The query generation analysis results are presented in Table 1. From Table 1, we have the following observations.

(1) Brain-Aug exhibits lower perplexity and higher Rouge-L than its ablations without brain input (w/o Brain) and randomly sampled brain signals as input (RS Brain). This indicates that the semantic information decoded from brain signals can be integrated with a query to construct a more effective prompt for generating query continuation.

(2) The overall perplexity and Rouge-L on the Pereira dataset are lower and higher than on the other two datasets, respectively. This implies that the Pereira dataset, derived from Wikipedia data, exhibits superior performance in the task of query generation compared to the other two datasets,

Dataset	Query	log(PPL)(↓)	Rouge-L(↑)
Pereira's	w/o Brain	2.219*	0.213*
	RS Brain	1.967*	0.267*
	Brain-Aug	**1.946**	**0.272**
Huth's	w/o Brain	3.573*	0.148*
	RS Brain	3.111*	0.159*
	Brain-Aug	**2.997**	**0.167**
Narratives	w/o Brain	4.328*	0.083*
	RS Brain	3.532*	0.105*
	Brain-Aug	**3.471**	**0.109**

Table 1: Query generation performance averaged across participants in different datasets. Best results in boldface. * indicates $p \leq 0.05$ for the paired t-test of *Brain-Aug (Ours)* and the controls. PPL indicates perplexity.

which are based on spoken stories.

(3) The RS Brain outperforms w/o Brain across three datasets. Although RS Brain uses brain signals that do not correspond to the current query context, the unified prompt can enable generating content that aligns with the common data distribution of language usage in the dataset (e.g., all stimuli in Pereira's dataset are Wikipedia-style). On other other hand, w/o Brain is equivalent to a standard language model that generates continuations soly based on the query text. This difference explains RS Brain's superior performance compared to the w/o Brain. However, in the discussion in Section 5.2, we will show that this performance improvement in query generation does not necessarily lead to an improvement in document ranking.

Answer to RQ1. The results show that queries augmented with semantics decoded from brain signals are more aligned with the content of the relevant document with the help of brain signals.

5.2 Document ranking performance

Overall performance. Table 2 shows the document ranking performance with original queries, queries augmented with unsupervised signals (Unsup-Aug), and queries augmented with brain signals (Brain-Aug). We observe:

(1) Regardless of whether BM25 or RepLLaMa is used as the ranking model, Brain-Aug substantially outperforms the original query and Unsup-Aug. The only exception is observed when using RepLLaMa and metric MAP on Pereira's dataset. A possible explanation for this exception is the RepLLaMA's high performance on the Pereira dataset, which we discuss in observation (3).

(2) When considering various datasets and metrics, the Unsup-Aug query does not consistently outperform the original query. Significant differences between the performance achieved by the Unsup-Aug query and the original query emerge on the metric of Recall@20 when using BM25 as the ranking model. This observation suggests that Unsup-Aug, which improves query representation by tackling term mismatch issues, leads to an improvement in recall. When Brain-Aug is combined with Unsup-Aug (Brain+Unsup), we observe a performance gain when compared to Unsup-Aug. This highlights the effectiveness of brain signals in query augmentation and underscores the potential of combining them with traditional signals.

(3) We observe little difference in performance between RepLLaMa and BM25 on Huth's dataset and Narratives's dataset. This implies that in a zero-shot setting and cross-domain scenario (the datasets are derived from spoken stories, which differs from the training data of RepLLaMa), dense retrieval models like RepLLaMa are not necessarily better than traditional sparse retrieval models like BM25. This phenomenon is also observed in the BEIR dataset (Thakur et al., 2021). However, in Pereira's dataset, RepLLaMa shows significant improvement over BM25 with different query inputs. The impressive performance of RepLLaMa on Pereira's dataset can likely be attributed to the fact that the data in Pereira are likely to be used in the original construction of RepLLaMa.

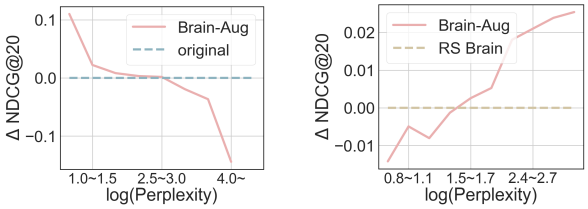
Decomposing Brain-Aug. Next, we investigate the contribution of brain signals and the ranking-oriented inference approach to Brain-Aug. Experimental results are presented in Table 3. First, we observe that removing (w/o Brain) or random sampling the brain inputs (RS Brain) leads to a decrease in performance. This indicates that semantic information decoded from brain signals within the query context enhances the query. Furthermore, while RS Brain consistently outperforms w/o Brain approach in terms of generation perplexity (see Section 5.1), it struggles to achieve better document ranking performance on the Huth's and Narratives datasets. This can be attributed to the fact that RS Brain, despite generating content that closely matches the token distribution of the whole dataset and reducing perplexity, fails to effectively differentiate between different documents within the dataset without semantics related to the query context. Last, we also observe a significant performance improvement when comparing Brain-Aug against its ablation without ranking-orient generation (w/o IDF). This suggests the importance of generating content that can be used to differentiate between documents.

Dataset	Query	BM25				RepLLaMA			
		N@10	N@20	R@20	MAP	N@10	N@20	R@20	MAP
Pereira's	original	$0.643^{*,\dagger}$	$0.664^{*,\dagger}$	$0.888^{*,\dagger}$	$0.594^{*,\dagger}$	0.878	$0.881^{*,\dagger}$	$0.964^{*,\dagger}$	0.858
	Unsup-Aug	$0.646^{*,\dagger}$	$0.655^{*,\dagger}$	$0.924^{*,\dagger}$	$0.590^{*,\dagger}$	$0.872^{*,\dagger}$	$0.877^{*,\dagger}$	$0.951^{*,\dagger}$	0.855
	Brain-Aug	0.671	**0.691**	**0.941**	**0.618**	**0.883**	**0.887**	**0.980**	**0.859**
	Brain+Unsup	**0.673**	0.686	0.936	0.615	0.878	0.882	0.975	0.853
Huth's	original	$0.297^{*,\dagger}$	$0.326^{*,\dagger}$	$0.536^{*,\dagger}$	$0.264^{*,\dagger}$	$0.299^{*,\dagger}$	$0.328^{*,\dagger}$	$0.520^{*,\dagger}$	$0.275^{*,\dagger}$
	Unsup-Aug	$0.291^{*,\dagger}$	$0.320^{*,\dagger}$	$0.575^{\dagger}$	$0.259^{*,\dagger}$	$0.302^{*,\dagger}$	$0.333^{*,\dagger}$	$0.537^{*,\dagger}$	$0.276^{*,\dagger}$
	Brain-Aug	0.306	0.340	$0.569^{\dagger}$	**0.273**	**0.310**	**0.342**	0.550	**0.281**
	Brain+Unsup	**0.309**	**0.342**	**0.580**	0.269	0.308	0.340	**0.552**	0.279
Narratives	original	$0.419^{*,\dagger}$	$0.434^{*,\dagger}$	$0.629^{*,\dagger}$	$0.355^{*,\dagger}$	$0.413^{*,\dagger}$	$0.426^{*,\dagger}$	$0.611^{*,\dagger}$	$0.351^{*,\dagger}$
	Unsup-Aug	0.440	$0.452^{\dagger}$	$0.670^{\dagger}$	$0.367^{*,\dagger}$	$0.416^{*,\dagger}$	$0.431^{*,\dagger}$	$0.629^{*,\dagger}$	$0.356^{*,\dagger}$
	Brain-Aug	0.441	0.458	0.669	**0.382**	0.430	**0.446**	0.641	**0.382**
	Brain+Unsup	**0.445**	**0.462**	**0.678**	**0.382**	**0.432**	**0.446**	**0.642**	0.380

Table 2: Document ranking performance averaged across participants. Best results in boldface. $*/\dagger$ indicates Brain-Aug / Brain+Unsup significantly outperforms the baseline ($p \leq 0.05$, paired t-test), respectively.

Dataset	Query	NDCG@20	MAP
Pereira's	w/o Brain	0.665^{*}	0.586^{*}
	RS Brain	0.678^{*}	0.604^{*}
	w/o IDF	0.684^{*}	0.609^{*}
	Brain-Aug	**0.691**	**0.618**
Huth's	w/o Brain	0.332^{*}	0.265^{*}
	RS Brain	0.321^{*}	0.256^{*}
	w/o IDF	0.332^{*}	0.266^{*}
	Brain-Aug	**0.340**	**0.273**
Narratives	w/o Brain	0.452^{*}	0.368^{*}
	RS Brain	0.448^{*}	0.367^{*}
	w/o IDF	0.450^{*}	0.373^{*}
	Brain-Aug	**0.458**	**0.382**

Table 3: Document ranking performance of *Brain-Aug (ours)* and its controls with ranking model BM25. Best results in boldface. * indicates $p \leq 0.05$ for the paired t-test of *Brain-Aug* and the baseline.



(a) Brain-Aug vs. original (b) Brain-Aug vs. RS B

Figure 2: Relationship between document ranking performance and perplexity of ground-truth query continuation in Pereira's dataset. "RS B" indicates the ablation of Brain-Aug that randomizes brain inputs. Δ NDCG@20 indicates performance gains of Brain-Aug.

Relationship between document ranking and query generation performance. Fig. 2 illustrates the relationship between the document ranking performance of Brain-Aug and RS Brain and the perplexity of query continuation measured using RS Brain. The lower perplexity of query generation indicates a higher likelihood of generating more accurate query continuation. This higher likelihood, as shown in Fig. 2a, further leads to an increase in document ranking performance. Conversely, Fig. 2b shows a different trend: when the perplexity is higher, the performance gain of Brain-Aug with its ablation RS Brain is higher. This implies that when generating accurate query continuations is difficult, semantics decoded from the query context with brain signals is more beneficial. This observation is consistent with findings by Ye et al. (2023) that the addition of brain signals lead to a more substantial performance improvement when generating continuations with higher uncertainty .

Example cases. Table 4 presents example cases with the original query "The shaking can" which is sampled from document d_{13} in Pereira's dataset. Brain-Aug leverages brain signals to expand the query with "be caused by an earthquake". As a result, the relevant document with the topic of the earthquake, d_{13}, is appropriately ranked at the top of the search results. Example cases for Huth's and Narratives dataset are provided in Section A.5.

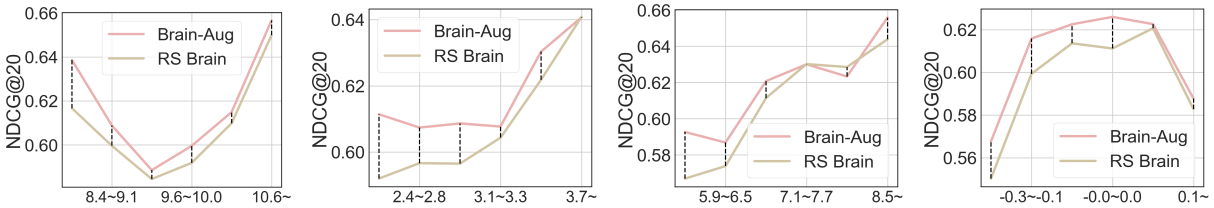
Answer to RQ2. We verified that a query augmented with semantics decoded from brain signals can significantly enhance document ranking performance. This performance enhancement is more pronounced when the generated query continuation is more accurately aligned with the query context.

5.3 Query performance analysis

Next, we investigate the performance improvement achieved by Brain-Aug for different queries by grouping queries according to their features. We select four query features: three pre-retrieval features (calculated based on query tokens), i.e., *ICTF*, *IDF*, and *specificity* score (Shtok et al., 2012), and one post-retrieval feature (calculated based on the information of retrieved documents), i.e., *clarify* score (Cronen-Townsend et al., 2002; Meng et al.,

Method	Query Content	Top-ranked document	Relevance
Original	The shaking can	d_{21}: The wind from the hurricane shook the house, shattering a window ... Later that night, with the wind shaking the house, ...	0
Unsup-Aug	The shaking can from house wind	d_{21}: The wind from the hurricane shook the house, shattering a window ... Later that night, with the wind shaking the house ...	0
RS Brain	The shaking can last anywhere from a few seconds to several minutes	d_{21}: The wind from the hurricane shook the house, shattering a window in the kitchen. ... Later that night, with the wind shaking the house, we fell asleep huddled on the sofa.	0
Brain-Aug	The shaking can be caused by an earthquake	d_{13}: Earthquakes shake the ground and can knock down buildings and other structures. *[MASK]* also trigger landslides and volcanic activity. Most earthquakes are caused by ...	1

Table 4: Examples of document ranking with BM25 using the original query or the augmented query in Pereira's dataset. Text in blue and in purple indicates content in the original query and generated by the query augmentation method, respectively. *[MASK]* indicates the position of the query "The shaking can" in the ICT setting.



(a) Avg ICTF (b) Avg IDF (c) Specificity (d) Clarify

Figure 3: Document ranking performance w.r.t. different query features in Pereira's dataset.

2023). For details on the query features, see Section A.3. We conjecture that larger feature values correspond to a more clarified query and usually result in better retrieval quality.

Fig. 3 depicts the document ranking performance w.r.t. different query features on Pereira's dataset. We have two key observations. (i) When the averaged IDF, specificity score, and clarity score increase, both Brain-Aug and the RS Brain show an improvement in retrieval performance. This indicates that a more specific query usually has a better retrieval performance. (ii) The performance gain of Brain-Aug compared to RS Brain is more pronounced when these features experience a decrease. This observation is supported by a significant negative Pearson's r between the improvement in NDCG@20 for Brain-Aug compared to RS Brain and the averaged ICTF, averaged IDF, specificity score, and clarity score, which are -0.14, -0.19, -0.17, and -0.32, respectively. This indicates that the performance improvement brought by brain signals is larger in queries prone to be vague or ambiguous.

Answer to RQ3. We have observed that queries prone to ambiguity (e.g., containing tokens with lower IDF scores or with low clarify scores) stand to gain more from Brain-Aug.

6 Discussion and Conclusion

Existing research incorporating physiological signals in IR tasks, whether based on eye-tracking (Bhattacharya et al., 2020) or brain signals (Ye et al., 2024; Eugster et al., 2014), has relied on predicting relevance of presented information. Here, we have investigated an alternative approach for directly augmenting queries based on the semantic information decoded from fMRI brain signals. Our findings revealed that decoding semantic representations from brain signals can enhance the generation of queries and subsequently improving document ranking. Moreover, we have observed that brain signals are more effective when the content to be generated has higher perplexity, indicating that decoded semantic information for unlikely query augmentations is more effective than it is for likely query augmentations. In conclusion, our findings open a horizon for new types of methods for understanding users by decoding semantics associated with information needs directly from brain signals. This process can kick off naturally as it happens as part of perceiving information and without requiring users to engage with any particular interaction technique or user interface.

7 Limitations

Our work has the following limitations pointing towards promising avenues for future research: (i) Our study utilized fMRI signals, which are not readily accessible in real-world human-computer interaction scenarios and have a significant delay of 2-8 seconds. More commonly used signals, such as electroencephalogram (EEG), have lower signal-to-noise ratios, which may limit their utility for semantic decoding. Currently, there is a lack of evidence that EEG can effectively decode semantics. In recent years, sensor technology like Functional near Infrared Spectroscopy (fNIRS) and MEG may become promising directions for future research. (ii) Our experiments simulate the document ranking with an ICT setting and show significant improvements over the baselines and carefully designed controls. Although ICT is commonly used to test retrieval performance, it is different from the most realistic search interaction. This simulation with ICT was driven by its advantage in building a sufficient number of queries and obtaining the corresponding query context to construct a substantial amount of training data. In the future, it would be worthwhile to explore settings that closely resemble real-world query interaction. This can be done through approaches such as training with ICT and testing with another corpus of queries, or by designing few-shot learning or cross-subject training models to enable query augmentation with a limited amount of data.

8 Ethical considerations

Recently, there has been a series of works attempting to utilize brain–computer interface (BCI) technology to enhance information accessing performance in various language-related applications, such as search (Eugster et al., 2016; Pinkosova et al., 2020; Allegretti et al., 2015) and communication (Pereira et al., 2018). Such technology is currently at a very early stage where such applications feel a long way off. However, it is important to discuss the associated concerns regarding privacy issues as the collection of brain signals is inherently susceptible to the actions of malicious third parties, which increases the risk of potential misuse or mishandling of sensitive information.

On the one hand, raw data collected via neurophysiological devices should be treated as private information, as such data can potentially be used to identify an individual (Alsunaidi et al., 2020) as well as their physiological disorders and thoughts (Yin et al., 2022). This technology may lead to risks such as influencing people's political opinions, and discrimination during recruiting based on their neural profiles. Therefore, the raw data should be avoided from being uploaded to the cloud for computation. It is necessary to filter sensitive information and decode only the information that helps the user accomplish their task with local computing. For publicly available datasets, ethical review and informed consent from each participant should be obtained, such as the dataset used in this paper (see Section A.1). Additionally, datasets should be used strictly for research purposes following their respective licenses.

On the other hand, there is a concern regarding the interaction log that might be recorded in applications like search engines. Although such interactions, such as clicks, comments, and submitted queries, are frequently recorded for improving individual user experience, the utilization of BCI can potentially pose greater risks. For example, it can be employed to capture users' genuine opinions on content within information systems, which can then be adopted in applications such as selective exposure and targeted advertising. Hence, users should have the right to decide whether they are willing to provide their interaction history to service providers. This is already specified in the legislation of many countries. In addition, the interaction history, even with users' permission, should undergo post-hoc filtering to remove any sensitive information before being utilized to train a model aimed at enhancing the commercial product.

9 Reproducibility

Our experiments use open-source datasets (Pereira's dataset (Pereira et al., 2018), Huth's dataset (LeBel et al., 2023), and the Narratives dataset (Nastase et al., 2021), which can be downloaded from the paper websites or OpenNeuro[1]). The data from Pereira et al. (2018) is available under the CC BY 4.0 license. The Huth's dataset and Narratives dataset are provided with a "CC0" license. Code is released using an anonymous link during the review process: https://anonymous.4open.science/r/Brain-Query-Augmentation-B6CC/. All code used in the paper are available under the MIT license after the review process.

[1] https://openneuro.org/

References

Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459.

Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 385–394.

Marco Allegretti, Yashar Moshfeghi, Maria Hadjigeorgieva, Frank E. Pollick, Joemon M. Jose, and Gabriella Pasi. 2015. When relevance judgement is happening? An EEG-based study. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–722.

Shikah J Alsunaidi, Nazar Abbas Saqib, and Khalid Adnan Alissa. 2020. A comparison of human brainwaves-based biometric authentication systems. *International Journal of Biometrics*, 12(4):411–429.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. 2016. MS MARCO: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Nilavra Bhattacharya, Somnath Rakshit, Jacek Gwizdka, and Paul Kogut. 2020. Relevance prediction from eye-movements using semi-interpretable convolutional neural networks. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 223–233.

Keping Bi, Qingyao Ai, and W Bruce Croft. 2019. Iterative relevance feedback for answer passage retrieval with passage-level semantic match. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 558–572. Springer.

David Carmel and Elad Yom-Tov. 2010. *Estimating the query difficulty for information retrieval*. Morgan & Claypool Publishers.

Jia Chen, Jiaxin Mao, Yiqun Liu, Ziyi Ye, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. 2021. A hybrid framework for session context modeling. *ACM Transactions on Information Systems*, 39(3):1–35.

Xuesong Chen, Ziyi Ye, Xiaohui Xie, Yiqun Liu, Xiaorong Gao, Weihang Su, Shuqi Zhu, Yike Sun, Min Zhang, and Shaoping Ma. 2022. Web search via an efficient and effective brain-machine interface. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1569–1572.

Steve Cronen-Townsend, Yun Zhou, and W Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, pages 1–11.

Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M Spapé, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2016. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports*, 6(1):38580.

Manuel J.A. Eugster, Tuukka Ruotsalo, Michiel M Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting term-relevance from brain signals. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 425–434.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Jacek Gwizdka, Rahilsadat Hosseini, Michael Cole, and Shouyi Wang. 2017. Temporal dynamics of eye-tracking and eeg during reading and relevance decisions. *Journal of the Association for Information Science and Technology*, 68(10):2299–2312.

Claudia Hauff, Vanessa Murdock, and Ricardo Baeza-Yates. 2008. Improved query difficulty prediction for the web. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 439–448.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA.

Emilia Kacprzak, Laura M Koesten, Luis-Daniel Ibáñez, Elena Simperl, and Jeni Tennison. 2017. A query log analysis of dataset search. In *Web Engineering: 17th International Conference, ICWE 2017, Rome, Italy, June 5-8, 2017, Proceedings 17*, pages 429–436. Springer.

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

Victor Lavrenko and W Bruce Croft. 2017. Relevance-based language models. In *ACM SIGIR Forum*, volume 51, pages 260–267. ACM New York, NY, USA.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. 2023. A natural language fMRI dataset for voxel-wise encoding models. *Scientific Data*, 10(1):555.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Hang Li, Harrisen Scells, and Guido Zuccon. 2020. Systematic review automation tools for end-to-end query formulation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2141–2144.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023b. Gpt understands, too. *AI Open*.

Yifei Luo, Minghui Xu, and Deyi Xiong. 2022. Cogtaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 904–920.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning LLaMA for multi-stage text retrieval. *arXiv preprint arXiv:2310.08319*.

Qiaozhu Mei, Dengyong Zhou, and Kenneth Church. 2008. Query suggestion using hitting time. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, pages 469–478.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

Chuan Meng, Negar Arabzadeh, Mohammad Aliannejadi, and Maarten de Rijke. 2023. Query performance prediction: From ad-hoc to conversational search. *arXiv preprint arXiv:2305.10923*.

Dominika Michalkova, Mario Parra Rodriguez, and Yashar Moshfeghi. 2024. Understanding feeling-of-knowing in information search: An EEG study. *ACM Transactions on Information Systems*, 42(3):1–30.

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

Yashar Moshfeghi, Peter Triantafillou, and Frank E Pollick. 2016. Understanding information need: An fMRI study. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 335–344.

Javed Mostafa and Jacek Gwizdka. 2016. Deepening the role of the user: Neuro-physiological evidence as a basis for studying and improving search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, pages 63–70.

Mariacristina Musso, Andrea Moro, Volkmar Glauche, Michel Rijntjes, Jürgen Reichenbach, Christian Büchel, and Cornelius Weiller. 2003. Broca's area and the language instinct. *Nature Neuroscience*, 6(7):774–781.

Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, et al. 2021. The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8(1):250.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Mateus Pereira, Elham Etemad, and Fernando Paulovich. 2020. Iterative learning to rank from explicit relevance feedback. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 698–705.

Zuzana Pinkosova, William J McGeown, and Yashar Moshfeghi. 2020. The cortical activity of graded relevance. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–308.

Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520.

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

Joseph John Rocchio Jr. 1971. Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*.

Anna Shtok, Oren Kurland, David Carmel, Fiana Raiber, and Gad Markovits. 2012. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems*, 30(2):1–35.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358.

Nuwa Xi, Sendong Zhao, Haochun Wang, Chi Liu, Bing Qin, and Ting Liu. 2023. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13277–13291.

Yuki Yano, Yukihiro Tagami, and Akira Tajima. 2016. Quantifying query ambiguity with topic distributions. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1877–1880.

Ziyi Ye, Qingyao Ai, Yiqun Liu, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. 2023. Language generation from human brain activities. *arXiv preprint arXiv:2311.09889*.

Ziyi Ye, Xiaohui Xie, Qingyao Ai, Yiqun Liu, Zhihong Wang, Weihang Su, and Min Zhang. 2024. Relevance feedback with brain signals. *ACM Transactions on Information Systems*, 42(4):Article No. 93.

Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022a. Brain topography adaptive network for satisfaction modeling in interactive information access system. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 90–100.

Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022b. Towards a better understanding of human reading comprehension with brain signals. In *Proceedings of the ACM Web Conference 2022*, pages 380–391.

Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuancheng Li, Jiaji Li, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022c. Why don't you click: Understanding non-click results in web search with brain signals. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 633–645.

Wutao Yin, Longhai Li, and Fang-Xiang Wu. 2022. Deep learning for brain disorder diagnosis based on fmri images. *Neurocomputing*, 469:332–345.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2021. Towards brain-to-text generation: Neural decoding with pre-trained encoder-decoder models. In *NeurIPS 2021 AI for Science Workshop*.

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. (2022), 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 in the ICT task 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 (Bajaj et al., 2016) (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 (Bajaj et al., 2016), which was used to train our baseline RepLLaMA.

Query construction Following existing research in language decoding from brain signals (Tang et al., 2023; Ye et al., 2023), 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) (Carmel and Yom-Tov, 2010): ICTF is a popular measure for the relative importance of the query terms and is usually measured by the following formulas:

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

where $\mid D \mid$ 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) (Hauff et al., 2008): IDF is another widely used measure for the importance of the query terms and is typically measured by the following formu-

Dataset	#Partic-ipants	#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 5: Overall statistics of 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 6: Overall statistics of the document corpus and query set constructed with the fMRI datasets.

las:

$$IDF(w) = log(\frac{N}{N_w}) \tag{8}$$

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) (Cronen-Townsend et al., 2002): Specificity score measures the Kullback-Leibler divergence of the query's language model from the collection's language model, which can be formulated as:

$$q = \sum_{w \in q} P(w \mid q) log(\frac{P(w \mid q)}{P(w \mid D)}) \tag{9}$$

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

(4) Clarify (Cronen-Townsend et al., 2002): 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.

$$Clarify(q, D^k_{q:M}) = \sum_{w \in V} P(w \mid D^k_{q:M}) \frac{P(w \mid D^k_{q:M})}{P(w \mid D)} \tag{10}$$

where w and V denote a query term and the entire collection vocabulary, respectively, $D^k_{q:M}$ 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. (2003), which are known to be relevant to language processing in the human brain. For Pereira's dataset, we apply component analysis (Abdi and Williams, 2010) on the original fMRI features to reduce the dimensionality to 1000. The 7B version of the Llama-2 model (Touvron et al., 2023) released in Huggingface [2] is adopted as the language model for generating the query continuation.

We train Brain-Aug with the Adam optimizer (Kingma and Ba, 2014) using a learning rate of 1×10^{-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.

[2] https://huggingface.co/models

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

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. *[MASK]* indicates the position of the selected query in the ICT setting.

A.5 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.6 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 [3] and Zuco [4], 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.7 AI assistants usage

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

[3] https://github.com/YeZiyi1998/UERCM
[4] https://osf.io/2urht/
[5] https://chat.openai.com/
[6] https://gemini.google.com/app