Query Augmentation with Brain Signals

Ziyi Ye
Department of Computer Science and Technology, Tsinghua University
Beijing, China
yeziyi1998@gmail.com

Jingtao Zhan
Department of Computer Science and Technology, Tsinghua University
Beijing, China
zhanjt20@mails.tsinghua.edu.cn

Qingyao Ai
Department of Computer Science and Technology, Tsinghua University, Zhongguancun Lab
Beijing, China
aiqy@tsinghua.edu.cn

Yiqun Liu*
Department of Computer Science and Technology, Tsinghua University, Zhongguancun Lab
Beijing, China
yiqunliu@tsinghua.edu.cn

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

Christina Lioma
University of Copenhagen
Copenhagen, Denmark
c.lioma@di.ku.dk

Tuukka Ruostalo
University of Copenhagen
Copenhagen, Denmark
LUT University
Lahti, Finland
tr@di.ku.dk

ABSTRACT

In the information retrieval scenario, query augmentation is an essential technique to refine semantically imprecise queries to align more closely with users' actual information needs. Traditional methods typically rely on extracting signals from user interactions such as browsing or clicking behaviors to augment the queries, which may not accurately reflect the actual user intent due to inherent noise and the dependency on initial user interactions. To overcome these limitations, we introduce **Brain-Aug**, a novel approach that decodes semantic information directly from brain signals of users to augment query representation. Brain-Aug builds on three techniques: (i) Structurally, an adapter network is utilized to project brain signals into the embedding space of a language model, allowing query augmentation conditioned on both the users' initial query and their brain signals. (ii) During training, we use a next token prediction task for query augmentation and adopt prompt tuning to efficiently train the brain adapter. (iii) At the inference stage, a ranking-oriented decoding strategy is implemented, enabling Brain-Aug to generate augmentations that improve ranking performance.

We evaluate our approach on multiple functional magnetic resonance imaging (fMRI) datasets, demonstrating that Brain-Aug not only produces semantically richer queries but also significantly improves document ranking accuracy, particularly for ambiguous queries. These results validate the effectiveness of our proposed Brain-Aug approach, and reveal the potential of using internal cognitive states to understand and augment text-based queries. Supplementary materials and code are available at https://github.com/YeZiyi1998/Brain-Query-Augmentation.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Users and interactive retrieval**.

KEYWORDS

Query augmentation, Prompt tuning, Brain-computer interface (BCI)

ACM Reference Format:
Ziyi Ye, Jingtao Zhan, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Christina Lioma, and Tuukka Ruostalo. 2024. Query Augmentation with Brain Signals. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, AustraliaProceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28-November 1, 2024, Melbourne, Australia.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3664647.3681658

1 INTRODUCTION

Understanding users' intentions is the key to effective search engines. In the interactions between users and search engines, queries play an important role in presenting the users' intentions and for search engines to retrieve relevant documents. However, search engine users often struggle to express their information needs precisely, resulting in queries that are short [20], vague [52], or

*Corresponding author.

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681658

inaccurately phrased [10], which compromises retrieval effectiveness.

To address this problem, query augmentation has emerged as a crucial technique to refine the original queries into more effective expressions [22, 32]. Traditionally, this reformulation process relies heavily on external document information such as expanding the query with contents from documents users have engaged with [2, 8, 42].

The advent of neurophysiological interfaces offers a novel source of data to understand users' search intentions [35, 56]. In information retrieval (IR) scenarios, several studies have revealed that brain signals can be used to predict users' relevance perception [13, 43, 57] and cognitive state [37]. 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 [12], which can be further used to extract relevant content for query augmentation [54, 55]. However, the existing process of query augmentation still relies on the quality of initially retrieved documents and cannot start before potentially unsatisfactory interactions with initial documents.

In this paper, we introduce **Brain-Aug**, a novel approach to query augmentation. It uses brain signals to directly refine user-submitted queries by decoding semantic information embedded in neural activity. Brain-Aug incorporates three core elements: (i) *Model architecture*: Brain-Aug employs a mapping network to transform brain signals into the input space of a transformer model. This allows the model to generate a query conditional on both the brain signals and the initial query simultaneously, effectively integrating neural data with computational language processing. (ii) *Training protocol*: We develop a specialized pre-training alignment task tailored for brain signals and a fine-tuning process specifically for query augmentation. This dual training strategy enhances the model's ability to decode users' intentions from their neural signals. (iii) *Ranking-oriented inference*: During inference, Brain-Aug implements a ranking-oriented decoding strategy that uses inverse document frequency (IDF) to generate query continuations. This method ensures that the augmented words not only fit the context but also possess distinctive characteristics to improve ranking performance.

We conduct comprehensive experiments to validate the effectiveness of Brain-Aug. Using a variety of functional magnetic resonance imaging (fMRI) datasets and different retrieval systems, our results robustly demonstrate that Brain-Aug can accurately interpret user intentions and enhance search engine performance. Our approach not only significantly outperforms traditional query augmentation methods but also enhances their efficacy when combined with these methods. Furthermore, we performed both quantitative and qualitative analyses to deeply analyze Brain-Aug's capabilities. Our analysis reveals that Brain-Aug can effectively augment ambiguous queries, creating clearer and more precise queries that significantly boost retrieval effectiveness.

In summary, our contributions are as follows: (i) We introduce Brain-Aug, a novel architecture that enhances query representations by incorporating brain signals as an additional input. We have devised training and inference protocols aimed at refining queries with greater semantic precision and enhanced discriminative capability across different documents. (ii) We demonstrate the effectiveness of Brain-Aug, showing that Brain-Aug refines queries to align semantically more closely with the search intent. We further show that the augmented query can be used to improve search performance in terms of document ranking. (iii) We analyze the performance gain achieved by Brain-Aug against its controls and unsupervised baselines. We observe that Brain-Aug is more effective in cases where the original query is probed to be ambiguous.

2 RELATED WORK

2.1 Query Augmentation

Traditionally, query augmentation can be categorized into two types: based on pseudo-relevance signals [6, 22] and based on user signals [25]. Approaches based on pseudo-relevance signals usually treat top-ranked documents in the initial retrieval step as relevant. Based on these relevant documents, Lavrenko and Croft [22] and Rocchio Jr [47] 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., Ahmad et al. [2] and Chen et al. [8] 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.

2.2 Neuroscience and Information Retrieval

There is increasing literature that adopts neuroscientific methods in IR scenarios [9, 15, 29, 38, 44]. For example, Chen et al. [9] built a prototype in which users can interact with search systems with a brain-computer interface. Allegretti et al. [3], Michalkova et al. [35] and Moshfeghi et al. [37] study the cognitive mechanisms involved in the process of information retrieval. A common finding observed in existing literature is that brain signals can be used to as a relevance indicator [3, 13]. This indicator can be employed for query rewriting [12, 55]. 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 can be decoded to some extent with brain signals such as fMRI [53, 60] and magnetoencephalogram (MEG) [11]. However, there is currently a lack of research investigating the use of the decoded semantics for query augmentation.

3 METHOD

We first formalize the query augmentation task and then present Brain-Aug, including its architecture, training objective, and inference process.

3.1 Task formalization

In search engines, queries submitted by users are often unclear, failing to accurately reflect their true intentions. As brain-computer

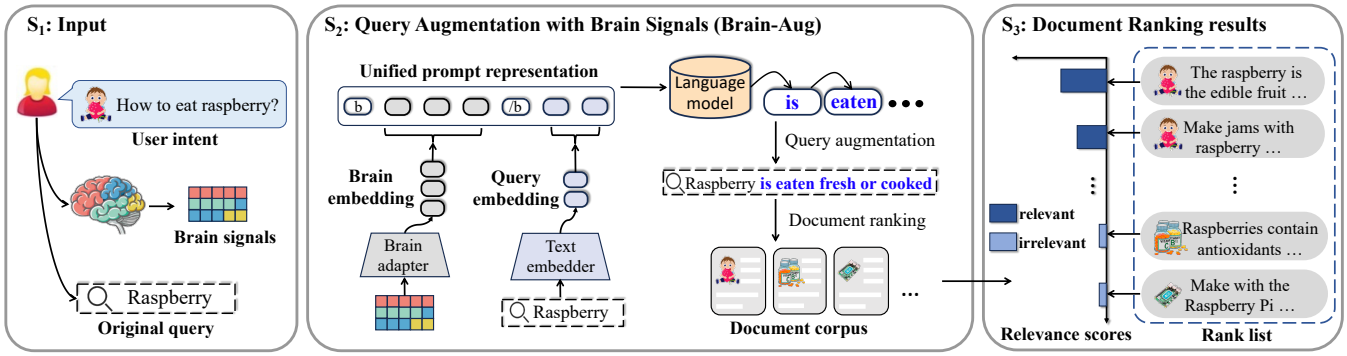


Figure 1: The procedure of query augmentation with brain signals (Brain-Aug). Brain-Aug constructs a unified prompt representation that jointly models the brain responses and original queries. With the unified prompt representation as input, a language model is adopted to generate the continuation of the original query for its augmentation.

interface techniques become increasingly cost-effective and wearable, this paper explores the potential of using brain signals to enhance the queries written by users. By incorporating the brain data, we aim to capture and reflect user intentions more precisely, augment queries, and thereby improve the accuracy of search results.

The *input* to the task of augmenting queries with brain signals is a query submitted by a user and the brain signals associated with the query context. Q is used to denote the query composed of n tokens: $Q = \{q_1, q_2, \ldots, q_n\}$. $B = \{b_1, \ldots, b_t\} \in \mathbb{R}^{t \times c}$ 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 that 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 $M = \{m_1, \ldots, m_k\}$, which will be concatenated to the initial query Q as an augmentated query. Let m_i be the i-th token of M, the generation process is formalized as:

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

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 [46], RepLLaMA [31]). 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) [18] or mean average precision (MAP) [19].

3.2 Overall Procedure

Fig. 1 provides an overview of the three-stage process of Brain-Aug: S_1: The 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 adapter 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. With the unified prompt representation as input, a language model is adopted to generate the continuation of the original query. A ranking-oriented inference method is used to enhance the generation process to improve the ranking performance. S_3: 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 Model Architecture

Brain-Aug integrates the textual query with cognitive information derived from brain signals and inputs them into a transformer model. The transformer is used to generate query augmentations based on the context of initial queries and user brain signals. In the following, we describe how to map the brain signals and input it to transformer.

First, since the brain signal extracted by fMRI cannot be directly processed by a pretrained language model, we devise a brain adapter f_b to embed each brain representation $b_i \in B$ into the same latent space $\mathbb{R}^d$, which can be formulated as $v_i^B = f_b(b_i)$. We implement it as a neural network f_b comprising (i) a MLP network f_m with ReLU [14] as the activation function, and (ii) a position embedding $P = \{p_1, \ldots, p_t\} \in \mathbb{R}^{t \times c}$. 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. Formally, the fMRI features b_i is mapped as:

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

where i denotes the i-th time frame.

Then, we acquire the embeddings of the initial query. We feed the query's text Q 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.

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:

$$S = \{\langle b\rangle, v_1^B, \ldots, v_t^B, \langle /b\rangle, v_1^Q, \ldots, v_n^Q\}. \tag{3}$$

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

3.4 Training

To effectively use brain signals for query augmentation, we design a two-stage training process. The first is an unsupervised training stage and is to warm-up the brain adapter for aligning the brain input to the latent space of the language model. The second is a supervised learning stage and is to guide the model to decode semantic information from brain signals for query augmentation.

3.4.1 Unsupervised training to warm-up the brain adapter. We design an unsupervised warm-up stage to align the distribution of the brain embedding with that of the text token's embeddings, ensuring that the brain embedding is suitable as the input of a language model. We construct training pairs in an unsupervised manner. Each pair consists of a series of brain signals and the associated text. Formally, let V^B be the mapped brain signals. Each $v_i^B \in V^B$ is trained to be close 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 training, which can be formulated as:

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

The inclusion of a warm-up stage in the adapter training phase of a language model with multimodal input is crucial, as illustrated by Liu et al. [27]. Our experimental results corroborate these findings. We observe that omitting the adapter training phase can result in problems like unstable training, exemplified by gradient explosion, and inferior performance compared to models that include a warm-up stage.

3.4.2 Query augmentation as next token prediction. Given the input S as formulated in Eq. (3), we train the model with the next token prediction task using a prompt tunning setup. Let $M^* = \{m_1^*, \ldots, m_k^*\}$ be the ground truth outputs. The language model is trained to predict M conditioned on S. The training objective is to maximize the likelihood of generating the ground truth, which can be formulated as:

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

where Θ is the model parameters. Constructing ground-truth labels presents a significant challenge, as our task assumes that users are not good at writing clear queries. Consequently, we cannot simply ask users to write a clearer ground-truth query. To address this issue, we hypothesize that an accurate representation of user intent corresponds to the documents they consider relevant. Therefore, we set the ground truth to be these relevant documents and train the model to reconstruct relevant documents based on user's brain signals and initial query. This approach effectively avoids the difficulty of having users directly annotate clear queries, as it is comparatively easier for them to identify relevant documents.

The training process follows the "prompt tuning" approach [28] by freezing the parameters of the language model and fine-tuning only the prompt representation S. This indicates that only the parameter of the brain adapter (Θ^{f_b}) and the parameter of the special tokens (Θ^{sp}) are updated. In this way, we can train Brain-Aug efficiently with limited training data constructed from brain imaging datasets.

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 [45] of each token in the vocabulary when generating query continuation $\hat{M} = \{\hat{m}_1, \ldots, \hat{m}_k\}$. Let $\mathrm{IDF}(\hat{m})$ 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:

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

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 Appendix A.4.

4.1 Datasets

Three publicly available fMRI datasets are adopted, namely Pereira's dataset [41], Huth's dataset [23], and the Narratives dataset [40]. 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 Appendix A.1.

4.2 Data Processing

We extract queries and documents from existing fMRI datasets following Izacard et al. [17] and Lee et al. [24]. Specifically, we select 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 $D = \{w_1, \ldots, w_m\}$, we extract 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 pairs for each document. In Huth's dataset and the Narratives datasets, continuous contents are presented to the user as auditory stimuli. We use 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 Appendix 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 according 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 Appendix A.2 for details). Taking into account the delayed effect of fMRI signals[36], 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 [33] 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 [26].

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) [18], Recall@20, and MAP [19].

4.4 Baselines and Controls

Given the augmented query, we select two ranking models for document ranking, i.e., a sparse ranking model, **BM25** [46], and a dense ranking model, **RepLLaMA** [31]. 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**. (ii) The query augmented with pseudo-relevance signals (denoted as **Unsup-Aug**). When using BM25 as the ranking model, we implement RM3 [22] 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 [6] 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 **Unsup+Brain-Aug**. (iv) To test whether the proposed method makes effective use of the brain signals, we also compare Brain-Aug with a baseline using fMRI signals for semantic classification [41], which is detailed in Appendix A.7.

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

Next, we evaluate the performance of Brain-Aug according to the similarity of the generated continuation and the ground truth label of continuation. The query generation analysis results are presented in Table 1. From Table 1, we have the following observations:

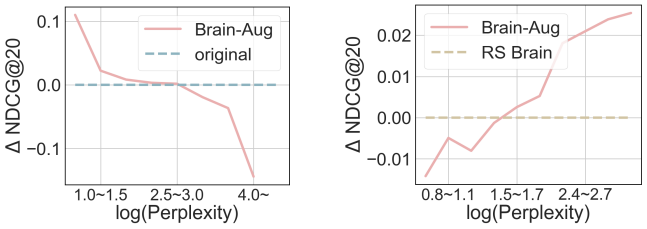
(1) Brain-Aug exhibits lower perplexity and higher Rouge-L than its ablations without brain input (w/o Brain) and randomly sampled brain signals as input (RS Brain). This indicates that the semantic information decoded from brain signals can be integrated with a query to construct a more effective prompt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

(3) 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 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.

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.

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**



(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.

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

5.2.1 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 the following:

(1) Regardless of whether BM25 or RepLLaMa is used as the ranking model, Brain-Aug substantially outperforms the original query and Unsup-Aug. According to NDCG@20 results, Brain-Aug improved the original query by 0.027 on Pereira's dataset, 0.014 on Huth's dataset, and 0.024 on the Narratives dataset. The only exception is observed when using RepLLaMa and the metric MAP on Pereira's dataset. A possible explanation for this exception is tRepLLaMA'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 (Unsup+Brain-Aug), 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 the Narratives 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 [50]. However, on 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.

5.2.2 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. Second, 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 Huth's dataset and the Narratives dataset. 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. Third, 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. Last, Brain-Aug significantly outperforms its variants that utilize a contrastive training target instead of a next-token prediction target [41]. This indicates that Brain-Aug is more effective in extracting semantic information from brain signals by using the inherent generalization capability of the language model (see Appendix A.7).

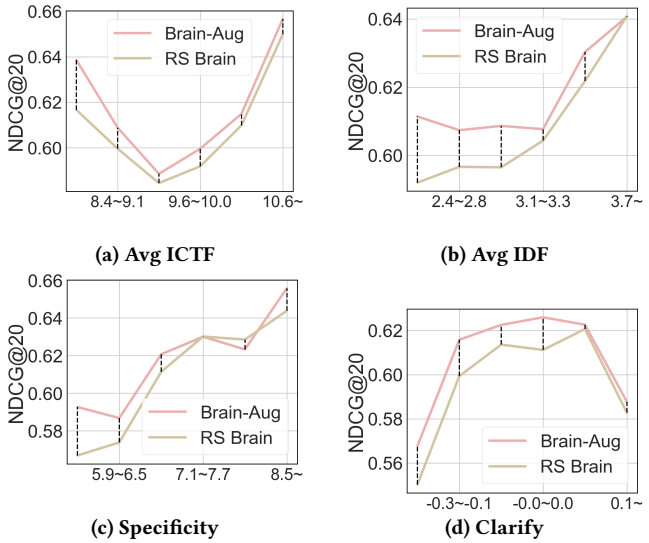
5.2.3 Relationship between document ranking and query generation performance. Fig. 2 illustrates the relationship between the document ranking performance of Brain-Aug and RS Brain and the perplexity of query continuation measured using RS Brain. The lower perplexity of query generation indicates a higher 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

Table 2: Document ranking performance averaged across participants, with our method (*Brain-Aug* & *Brain+Unsup*) marked by a ⋆. Best results are in boldface, and the second-best results are underlined. ∗/† indicates Brain-Aug / Brain+Unsup significantly outperforms the baseline ($p \leq 0.05$, paired t-test), respectively.

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

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.

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**



(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.

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. [53] that the addition of brain signals leads to a more substantial performance improvement when generating continuations with higher uncertainty.

5.2.4 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 uses 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. In contrast, when using the original query or augmenting it with unsupervised signals or randomly sampled brain signals, the document d_{21}, which discusses shaking wind, is erroneously ranked as the top result. This case study demonstrates the significant impact of incorporating brain signals into the query augmentation process. Example cases for Huth's dataset and the Narratives dataset are provided in Appendix A.6.

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

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.

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. also trigger landslides and volcanic activity. Most earthquakes are caused by ...	1

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 [48], and one post-retrieval feature (calculated based on the information of retrieved documents), i.e., *clarify* score [10, 34]. For details on the query features, see Appendix 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:

(1) 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.

(2) The performance gain of Brain-Aug over 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

6.1 Summary of Contributions

This paper investigates augmenting text-based queries based on the semantic information decoded from fMRI brain signals. Our findings revealed that representations decoded from brain signals can be used for query auto-regressive generation with transformer-based language models and subsequently improve 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.

6.2 Limitations and Future work

Our work has the following limitations pointing towards promising avenues for future research:

(1) Our study used fMRI signals, which may be not readily accessible in real-world human-computer interaction scenarios and have a significant delay of 2–8 seconds. At the preliminary stage of investigating brain signals in IR, we chose to use fMRI signals because fMRI has been extensively studied in semantic decoding among various physiological signals (e.g., EEG, ECoG, and fNIRS). The significant component of Brain-Aug is designed to be independent of the type of signal employed. With the advancement in brain-related devices, we believe that Brain-Aug has the potential to be applied in more lightweight devices including devices that share similar hemodynamic inputs (e.g., fNIRS), or utilize real-time electrical inputs (e.g., EEG). Additionally, we foresee its applicability beyond traditional IR, which includes virtual reality applications, and disabled services.

(2) Our experiments simulate the ranking task on fMRI datasets following [17, 24]. The simulation methodology is commonly used to test retrieval performance and retrieval model trained with this methodology has shown to be effective for real-world queries [17]. This simulation was driven by a more accurate query representation necessitates greater similarity to relevant documents, thereby enabling more straightforward retrieval. In the future, we plan to explore evaluations that closely resemble real-world tasks for search engine users. Despite the high cost of conducting these experiments, we believe it can help advance this promising field.

ACKNOWLEDGMENTS

This work is supported by Quan Cheng Laboratory (Grant No. QCLZD202301), by the Dutch Research Council (NWO, Grant No. 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006), and by the European Union's Horizon Europe program (Grant No. 101070212), by the Academy of Finland, and by the Horizon 2020 FET program of the EU through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Hervé Abdi and Lynne J. Williams. 2010. Principal Component Analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2, 4 (2010), 433–459.
[2] 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*. 385–394.
[3] 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*. 719–722.
[4] 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 (2020), 411–429.
[5] 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* (2016).
[6] Keping Bi, Qingyao Ai, and W Bruce Croft. 2019. Iterative Relevance Feedback for Answer Passage Retrieval with Pasage-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*. Springer, 558–572.
[7] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers.
[8] 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 (2021), 1–35.
[9] 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*. 1569–1572.
[10] 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*. 299–306.
[11] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding Speech Perception from Non-invasive Brain Recordings. *Nature Machine Intelligence* (2023), 1–11.
[12] 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 (2016), 38580.
[13] 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*. 425–434.
[14] 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 (1980), 193–202.
[15] 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 (2017), 2299–2312.
[16] 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*. 439–448.
[17] 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* (2021).
[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.
[19] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.
[20] 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*. Springer, 429–436.
[21] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* (2014).
[22] Victor Lavrenko and W Bruce Croft. 2017. Relevance-based Language Models. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 260–267.
[23] 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 voxelwise encoding models. *Scientific Data* 10, 1 (2023), 555.
[24] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300* (2019).
[25] 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*. 2141–2144.
[26] Chin-Yew Lin. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. 74–81.
[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485* (2023).
[28] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT Understands, Too. *AI Open* (2023).
[29] Yizhuo Lu, Changde Du, Qiongyi Zhou, Dianpeng Wang, and Huiguang He. 2023. MindDiffuser: Controlled Image Reconstruction from Human Brain Activity with Semantic and Structural Diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5899–5908.
[30] 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)*. 904–920.
[31] 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* (2023).
[32] 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*. 469–478.
[33] Clara Meister and Ryan Cotterell. 2021. Language Model Evaluation Beyond Perplexity. *arXiv preprint arXiv:2106.00085* (2021).
[34] 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* (2023).
[35] 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 (2024), 1–30.
[36] 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 (2008), 1191–1195.
[37] 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*. 335–344.
[38] Javed Mostafa and Jacek Gwizdka. 2016. Deepening the Role of the User: Neurophysiological Evidence as a Basis for Studying and Improving Search. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*. 63–70.
[39] 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 (2003), 774–781.
[40] 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 (2021), 250.
[41] 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 (2018), 963.
[42] 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*. 698–705.
[43] 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*. 299–308.

[44] Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. 2023. Modal-aware Visual Prompting for Incomplete Multi-modal Brain Tumor Segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3228–3239.

[45] Stephen Robertson. 2004. Understanding Inverse Document Frequency: On Theoretical Arguments for IDF. *Journal of Documentation* 60, 5 (2004), 503–520.

[46] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389.

[47] Joseph John Rocchio Jr. 1971. Relevance Feedback in Information Retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing* (1971).

[48] 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 (2012), 1–35.

[49] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2023. Semantic Reconstruction of Continuous Language from Non-invasive Brain Recordings. *Nature Neuroscience* (2023), 1–9.

[50] 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* (2021).

[51] 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* (2023).

[52] 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*. 1877–1880.

[53] 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* (2023).

[54] 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 (2024), Article No. 93.

[55] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Brain Topography Adaptive Network for Satisfaction Modeling in Interactive Information Access System. In *Proceedings of the 30th ACM International Conference on Multimedia*. 90–100.

[56] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Towards a Better Understanding of Human Reading Comprehension with Brain Signals. In *Proceedings of the ACM Web Conference 2022*. 380–391.

[57] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuancheng Li, Jiaji Li, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. 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*. 633–645.

[58] Wutao Yin, Longhai Li, and Fang-Xiang Wu. 2022. Deep Learning for Brain Disorder Diagnosis Based on fMRI Images. *Neurocomputing* 469 (2022), 332–345.

[59] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Re-ranking. *arXiv preprint arXiv:2009.07258* (2020).

[60] 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. [30], 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 used its inherent segmentation for our experiment. After defining the document corpus, we use 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 [49, 53], 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) [7]: ICTF is a popular measure for the relative importance of the query terms and is usually measured by the following formulas:

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

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) [16]: IDF is another widely used measure for the importance of the query terms and is typically measured by the following formulas:

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

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) [10]: The Specificity score measures the Kullback-Leibler divergence of the query's language model from the collection's language model, which can be formulated as:

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

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

(4) Clarify [10]: The 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 the Narratives dataset, we select features from brain regions identified by Musso et al. [39], which are known to be relevant to language processing in the human brain. For Pereira's dataset,

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 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	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 bed like a good guy and relax ...	1

we apply component analysis [1] on the original fMRI features to reduce the dimensionality to 1000. The 7B version of the Llama-2 model [51] released on Hugging Face[1] is adopted as the language model for generating the query continuation.

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

We train Brain-Aug with the Adam optimizer [21] using a learning rate of 1×10^{-4} and a batch size of 8. The learning rate is selected from the set $\{1\times10^{-3}, 1\times10^{-4}, 1\times10^{-5}\}$ 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

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 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 [59] 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 dataset and the Narratives dataset in Table 7. In these cases, Brain-Aug uses 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 A Baseline using fMRI signals for semantic classification

To test whether the proposed method better uses the brain signals, we test a method according to the principle illustrated by Pereira et al. [41]. First, we constructed the unified prompts using the same method of Brain-Aug and fed them into RepLLaMa to obtain augmented query representations. Then, the final hidden vectors of the query representations are used to map with the vectors generated from the documents with Repllama. To train the model, we adopt the contrastive loss proposed by Pereira et al. [41] to bring the representations of relevant documents and queries closer together. Finally, the documents are ranked according to their similarities with the query representations. Here are the experimental results on Pereira's dataset:

Method	N@20	R@20
Method inspired by Pereira et al. [41]	0.864	0.951
Brain-Aug	0.887	0.980

We found that the method inspired by Pereira et al. [41] significantly underperforms Brain-Aug and does not yield a significant performance difference from the original query. Indeed, we observe that while the training loss degrades using this method, the validation loss fails to decrease. This indicates that such a method lacks generalization ability. We speculated that this could be potentially attributed to the label-inefficient issue in their training settings with contrastive loss. Conversely, Brain-Aug based on the generative language modeling, gains more information by using the inherent generalization capability of the large model, thereby resulting in superior performance.

A.8 Brain-Aug with EEG inputs

As electroencephalogram (EEG) signals are easier to collect in real-world scenarios than fMRI, we also test whether EEG signals can be used for Brain-Aug. However, we found that in our experiment with two public EEG datasets, i.e., UERCM[2] and Zuco,[3] Brain-Aug did not outperform RS Brain. This implies that the existing quality of EEG data has limitations in their ability to decode semantics with Brain-Aug.

A.9 Ethical Considerations

Recently, there have been multiple publications aimed at using brain–computer interface (BCI) technology to enhance information accessing performance in various language-related applications, such as search [3, 12, 43] and communication [41]. Such technology is currently at a very early stage where related applications seem 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,

[2]https://github.com/YeZiyi1998/UERCM
[3]https://osf.io/2urht/

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 [4] as well as their physiological disorders and thoughts [58]. 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 Appendix 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 use 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 used to train a model aimed at enhancing the commercial product.

A.10 Reproducibility

Our experiments use open-source datasets (Pereira's dataset [41], Huth's dataset [23], and the Narratives dataset [40], which can be downloaded from the paper websites or OpenNeuro[4]). The data from Pereira et al. [41] is available under the CC BY 4.0 license. Huth's dataset and the Narratives dataset are provided with a "CC0" license. Our code is available at: https://github.com/YeZiyi1998/Brain-Query-Augmentation. All code used in this paper is available under the MIT license.

A.11 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.

[4]https://openneuro.org/
[5]https://chat.openai.com/
[6]https://gemini.google.com/app