On the Use of Electroencephalography in Query Performance Prediction

IAKOVOS TENEDIOS, NIALL MCGUIRE, and YASHAR MOSHFEGHI, NeuraSearch Laboratory, CIS, University of Strathclyde, Glasgow, UK

Query Performance Prediction (QPP) enables information retrieval systems to estimate search effectiveness without requiring explicit relevance judgements. While traditional QPP research has focused exclusively on textual features, we investigate on enhancing QPP through the multimodal integration of electroencephalography (EEG) and eye-tracking signals captured during listening and reading. We utilise a specialised dataset where queries are represented as text alongside corresponding neurophysiological recordings, with graded relevance judgements for multiple documents. Our methodology employs an ensemble architecture with dedicated models for each modality, followed by a meta-learner that produces final predictions. Experimental evaluation across reading and listening tasks demonstrates that generalised models trained across subjects achieve consistent improvements over text-only baselines, with the trimodal configuration (EEG+eye-tracking+text) reaching Pearson correlation of 0.458 for the ZuCo reading dataset and bimodal EEG+text models achieving 0.417 for the Narrative listening dataset, representing 50–80% higher performance than personalised single-subject models. Query-level analysis reveals that neurophysiological signals substantially improve predictions for fragmentary or semantically ambiguous query cases, while the majority show neutral effects. These findings establish the feasibility of neurophysiological QPP under specific conditions and provide design principles for integrating brain–computer interfaces in information retrieval systems. Our code can be found here.

CCS Concepts: • **Information systems** → **Query representation**;

Additional Key Words and Phrases: EEG, QPP, Information Retrieval, NeuraSearch, Machine Learning

ACM Reference format:
Iakovos Tenedios, Niall McGuire, and Yashar Moshfeghi. 2026. On the Use of Electroencephalography in Query Performance Prediction. *ACM Trans. Inf. Syst.* 44, 6, Article 147 (August 2026), 34 pages.
https://doi.org/10.1145/3816244

1 Introduction

Query performance prediction (QPP) represents a fundamental task in **information retrieval (IR)** systems, focusing on estimating retrieval effectiveness without requiring explicit relevance judgements or actual query execution [15, 23, 36]. This predictive capability offers significant practical benefits across numerous applications, including dynamic selection of optimal

The work was supported by Engineering and Physical Sciences Research Council (EP/W522260/1). Results were obtained using the ARCHIE-WeSt High Performance Computer based at the University of Strathclyde.
Authors' Contact Information: Iakovos Tenedios, NeuraSearch Laboratory, CIS, University of Strathclyde, Glasgow, UK; e-mail: iakovos.tenedios@strath.ac.uk; Niall McGuire, NeuraSearch Laboratory, CIS, University of Strathclyde, Glasgow, UK; e-mail: niall.mcguire@strath.ac.uk; Yashar Moshfeghi (corresponding author), NeuraSearch Laboratory, CIS, University of Strathclyde, Glasgow, UK; e-mail: yashar.moshfeghi@strath.ac.uk.



This work is licensed under Creative Commons Attribution International 4.0.

© 2026 Copyright held by the owner/author(s).
ACM 1558-2868/2026/8-ART147
https://doi.org/10.1145/3816244

ranking algorithms [98], efficient allocation of computational resources [24] and intelligent query reformulation for problematic queries [98]. By identifying potentially ineffective queries before users encounter unsatisfactory results, QPP enables search systems to implement pre-emptive interventions that enhance user experience and retrieval performance.

The exploration of QPP methods has progressed substantially within text-based IR, with researchers developing diverse approaches ranging from linguistic complexity analysis to statistical distribution modelling [23, 25, 35–37]. Advanced techniques have leveraged query robustness measures [77, 79], domain-specific knowledge sources [48] and score distribution patterns [84] to improve prediction accuracy. More recently, machine learning frameworks have demonstrated promising results by integrating multiple prediction signals through supervised approaches [21, 60, 81, 98] and neural architectures [4–6, 18, 27].

While QPP research has expanded beyond text to include content-based image retrieval [47, 67, 72, 75, 87, 92], neurophysiological data represent an unexplored opportunity for enhancing prediction capabilities by directly measuring cognitive processes during query formulation. The integration of **brain–computer interfaces (BCIs)** with IR presents a novel and promising direction that could have a notable impact on our understanding of query difficulty and user information needs. This opportunity emerges from growing evidence that neurophysiological signals captured during search interactions contain valuable information about cognitive processes underlying retrieval effectiveness. Recent research has demonstrated that key IR processes generate distinct and measurable neurophysiological patterns that can be captured and analysed throughout search sessions [42, 63]. These patterns emerge during critical phases, including the formation of information needs [58, 64], query formulation [46, 97] and relevance assessment [3, 29, 74]. The ability to decode these neurophysiological signals offers potential solutions to persistent IR challenges, such as the semantic gap between user intent and query formulation [9, 85] and vocabulary mismatches between queries and relevant documents [61, 83].

The integration of **electroencephalography (EEG)** data with QPP represents an unexplored research direction with potential to reveal cognitive aspects of query difficulty not evident from text alone. EEG offers millisecond-level temporal resolution that aligns with the rapid cognitive processes involved in query formulation, while its increasing affordability and portability make it the most practical neuroimaging modality for potential real-world applications. By capturing neural activity during query processing, EEG signals could potentially reveal aspects of query difficulty not evident from linguistic or statistical features alone. Whether this translates to improved performance prediction accuracy remains an open empirical question that this work investigates. This research is guided by the following **research questions (RQs)**:

- *RQ1*: Can neurophysiological signals captured during language processing effectively predict query performance in IR tasks?
- *RQ2*: Do multimodal models that incorporate neurophysiological and behavioural signals alongside text features outperform traditional text-only approaches to QPP?
- *RQ3*: How does the prediction performance vary between personalised (single-subject) models and generalised (multi-subject) models that leverage EEG signals?
- *RQ4*: Are there specific conditions or query types where neurophysiological signals provide particularly valuable predictive information beyond what is available in text features?

Our research introduces a novel approach to QPP that leverages both textual and neurophysiological data. We utilise an augmented EEG-text dataset specifically designed for IR tasks, where each query is represented both as text and as corresponding EEG signals recorded during query processing. Each query–document pair is associated with a relevance score, enabling the evaluation of retrieval effectiveness using standard IR metrics. To predict query performance, we employ an ensemble

modelling approach where separate machine learning models are trained on text and EEG features. The predictions from these individual models are then integrated through a meta-learner to produce the final performance prediction.

This dual-modality approach enables us to examine whether neurophysiological signals recorded during query processing can enhance traditional text-based features and improve prediction accuracy. We evaluate our models across multiple conditions: single-subject personalised models that measure adaptability to individual EEG signal variations and generalised models trained on averaged EEG representations to assess broader applicability. Through these experiments, we aim to demonstrate that EEG-captured information can, in specific contexts, improve traditional text-based QPP when integrated into combined approaches, outperforming single-modality methods that rely solely on either EEG or text. This research potentially opens new avenues for neurophysiologically informed IR systems.

The primary contributions of this work are as follows:

- A systematic investigation demonstrating that neurophysiological signals captured during text processing can predict the query performance, establishing preliminary evidence for the feasibility of EEG-enhanced QPP.
- A multimodal stacking architecture that enables systematic investigation of how neurophysiological and text-based features can be integrated for QPP, providing a controlled framework for evaluating the relative contribution of each modality.
- An empirical evaluation of EEG-enhanced QPP across both reading and listening modalities, demonstrating that generalised models trained across subjects can achieve improvements over text-only approaches in specific dataset and model configurations, while personalised models exhibit mixed results that vary substantially with task type and inter-subject EEG variability.
- A systematic evaluation of both personalised and generalised EEG-based prediction models, revealing insights into the relationship between individual cognitive processes and retrieval effectiveness.
- An exploratory analysis identifying task-level and dataset-level conditions under which neurophysiological signals tend to provide predictive value for QPP, supported by query-level breakdowns of cases where EEG improves, has a negligible effect, or degrades prediction accuracy (Tables 3 and 4), providing initial empirical grounding for the boundary conditions of neurophysiologically enhanced IR systems.

The remainder of this article is structured as follows. Section 2 provides background on QPP and the integration of neuroscience with IR. Section 3 details our experimental methodology, including dataset creation, metrics, training procedures, model implementations and feature importance analyses. Section 4 presents our experimental results, analysing the performance of both generalised and personalised models across different datasets and modalities along with feature importance and query-level analyses results. Section 5 discusses the implications of our findings, exploring the relationship between neurophysiological signals and query performance while acknowledging limitations and challenges. Finally, Section 6 concludes the article and outlines directions for future research.

2 Background

2.1 QPP

QPP seeks to estimate retrieval effectiveness before search execution, enabling systems to identify potentially problematic queries and take appropriate remedial actions. This capability addresses a

fundamental challenge in IR: determining whether a query will yield satisfactory results without requiring actual execution or relevance judgements [15].

The foundation of QPP research lies in identifying query characteristics and collection statistics that reliably correlate with retrieval performance. Pre-retrieval prediction methods analyse these characteristics without executing the search, examining features such as term specificity, query scope and term relationships. Specificity features, often implemented using **inverse document frequency (IDF)**, assess how discriminative query terms are within the collection [38]. Query scope measures estimate the fraction of the collection a query is likely to retrieve, with narrower scope generally indicating more focused queries [37]. Term relationship features evaluate semantic coherence among query terms, with the assumption that semantically related terms often yield better performance [100].

Traditional QPP approaches have evolved from simple statistical predictors to more sophisticated methods combining multiple features [36, 100]. However, these methods still rely exclusively on textual and collection-level features.

An innovation in pre-retrieval prediction has been the development of probabilistic prediction models that estimate the likelihood of term occurrence in relevant documents based on collection-wide statistics. The simplified clarity score [37] adapts post-retrieval concepts to the pre-retrieval setting by modelling query-collection divergence without requiring actual retrieval. Similarly, query difficulty predictors based on Jensen–Shannon divergence between query and collection language models provide efficient estimations of how distinctive a query's vocabulary is from the general corpus [36]. A comprehensive evaluation by Hauff et al. [36] demonstrated that combining these diverse pre-retrieval features typically outperforms any individual predictor, though effectiveness varies considerably across different collections and query types. In contrast, post-retrieval predictors analyse the retrieved document set to make informed assessments. These methods examine the distribution of relevance scores and the content characteristics of returned documents, focusing on patterns that typically correlate with query performance [16, 23]. Score distribution analysis examines statistical properties of relevance scores among retrieved documents. For instance, the standard deviation of scores can indicate whether the system clearly distinguishes between relevant and non-relevant documents, with higher variance often correlating with better performance. The normalised query commitment [84] measures the standard deviation of retrieval scores normalised by the mean score, providing a robust indicator of how committed the retrieval system is to a specific subset of documents. Clustering-based approaches analyse the topical coherence of retrieved documents, with the premise that effective queries retrieve topically focused result sets [99].

Query robustness predictors represent another important post-retrieval category. These methods assess performance stability under controlled perturbations, such as query reformulation through pseudo-relevance feedback [102] or application of different scoring functions [7]. The query feedback approach [102] constructs expanded queries using pseudo-relevance feedback and measures the similarity between the original results and those from the expanded query. The underlying assumption is that queries retrieving stable results despite these variations tend to perform better overall. Similarly, the query-perturbation approach introduced by Vinay et al. [93] assesses robustness by measuring result overlap when small modifications are made to the query terms or weights. While these post-retrieval approaches generally achieve higher prediction accuracy, their computational overhead limits applicability in time-sensitive applications, representing a fundamental tradeoff in the QPP domain.

To evaluate QPP methods, researchers employ a two-phase approach: measuring actual retrieval effectiveness and assessing prediction accuracy. Retrieval effectiveness is quantified using established metrics such as Average Precision, **normalised discounted cumulative gain (NDCG)** and Precision@k [21, 36, 100]. Prediction accuracy is primarily evaluated through correlation analysis

between predicted and actual performance, using Pearson correlation for linear relationships and Kendall's tau or Spearman's rho for potentially non-linear associations [36, 43, 80].

Machine learning approaches have significantly advanced QPP research by treating performance prediction as a supervised regression problem. Linear regression models combine multiple predictors with learned weights to optimise prediction accuracy [21, 101]. The support vector regression approach proposed by Cummins [25] demonstrated particular effectiveness in integrating diverse prediction signals while maintaining robustness across different collections. More sophisticated approaches employ genetic algorithms to identify optimal feature combinations [8], while recent work has explored neural network architectures that can capture complex, non-linear relationships between query characteristics and performance [27, 98].

Particularly notable is the shift toward contextual embedding-based QPP methods that leverage large pre-trained language models. Arabzadeh et al. [5] used BERT contextual embeddings to capture semantic relationships between queries and documents, demonstrating significant improvements over traditional statistical approaches. Similarly, Chen et al. [18] proposed a group-wise prediction framework using contextual embeddings that accounts for query similarity when making predictions, addressing the challenge of query variability. These learning-based methods typically outperform individual predictors, though they require careful cross-validation to ensure generalisability across different query types and collections [16].

Our research builds upon these foundations while introducing neurophysiological signals as a novel dimension in QPP. By incorporating EEG data captured during passive text comprehension, we explore whether cognitive measures can reveal aspects of query difficulty not detectable through traditional text analysis. This integration of neurophysiological signals with established QPP methodologies represents a promising direction for bridging the gap between linguistic query analysis and the cognitive processes that underlie effective query formulation and retrieval.

2.2 Neuroscience and IR

The integration of neuroscience with IR research provides valuable insights into users' cognitive processes during search interactions. This interdisciplinary field enables researchers to measure brain activity as users engage with information systems, revealing the aspects of search behaviour that traditional observation methods or self-reports cannot capture [42, 63]. Neurophysiological approaches have informed our understanding of key IR phenomena across the search interaction spectrum. These include the formation and development of information needs [64, 65], query formulation processes [42, 97], relevance assessment mechanisms [3, 73, 74] and the neurophysiological correlates of search satisfaction [70, 71]. By capturing the brain's electrical, hemodynamic or magnetic responses during these cognitive processes, researchers can develop more nuanced models of user behaviour that account for attention allocation, cognitive workload, semantic processing and emotional responses, all factors that influence search outcomes.

While various neuroimaging methods exist like **Functional Magnetic Resonance Imaging (fMRI)**, **Magnetoencephalography (MEG)**, EEG has emerged as the most practical for IR research due to its temporal resolution, affordability and portability [3, 29, 42].

EEG has emerged as the most practical neuroimaging technique in IR research due to its good temporal resolution (millisecond precision), relative affordability and increasing portability [3, 29, 34, 42]. Modern EEG systems can capture rapid cognitive processes, including the immediate neural responses to query terms or search results, while allowing for more naturalistic experimental conditions than fMRI or MEG. While EEG offers more limited spatial resolution (typically centimetre-level precision), recent advances in high-density arrays and source localisation algorithms have improved its spatial specificity.

The value of EEG for IR research stems from its ability to capture relevant brain activity patterns with temporal precision that aligns with the cognitive processes involved in search interactions.

These patterns include **event-related potentials (ERPs)** that reflect specific cognitive events, as well as oscillatory activity across different frequency bands associated with attention (alpha: 8–12 Hz), working memory (theta: 4–7 Hz) and cognitive processing (beta: 13–30 Hz and gamma: >30 Hz). For instance, the N400 component, a negative deflection occurring approximately 400 ms after stimulus presentation, has been linked to semantic processing difficulties and can potentially indicate when users encounter unexpected information during search [34].

Neurophysiological approaches offer practical benefits for IR system development. By providing objective measures of user cognitive states without requiring explicit feedback, these methods enable the development of adaptive systems that can respond to users' implicit needs. Early work by Eugster et al. [29] demonstrated that EEG signals could predict term relevance with above-chance accuracy, effectively capturing users' relevance assessments before they provided explicit judgements. Building on this foundation, Jacucci et al. [42] developed a system that incorporated EEG-based relevance feedback to enhance retrieval performance, demonstrating the feasibility of BCIs for IR.

Recent advances have expanded the potential applications of neurophysiological approaches in IR. Chen et al. [20] developed an EEG-based brain–machine interface for web search that could interpret users' neural responses to search results and adapt accordingly. Similarly, McGuire and Moshfeghi [58] demonstrated that EEG signals could predict the realisation of information needs with reasonable accuracy, potentially enabling systems to anticipate when users will initiate search activities. Ye et al. [97] further showed that brain signals could be used to augment queries, enhancing retrieval performance through implicit neurophysiological feedback.

Despite these developments, applying neurophysiological measures to QPP presents significant practical challenges related to equipment costs, signal processing complexity and potential intrusiveness. Additionally, most existing neuroscience-informed IR research has focused on relevance assessment or information need development, leaving unexplored the question of how neural signals during passive text comprehension might predict subsequent retrieval performance. Additionally, the practical constraints of neuroimaging, including equipment costs, signal processing complexity and potential intrusiveness, present barriers to widespread implementation in operational IR systems.

Nevertheless, the integration of neurophysiological measures with QPP offers advantages that justify addressing these challenges. Traditional QPP approaches rely on textual features and statistical patterns that may not fully capture the cognitive dimensions of query difficulty. Neural signals can potentially reveal aspects of semantic processing, cognitive load and attention allocation during query formulation that correlate with retrieval performance in ways not evident from query text alone. By measuring these cognitive processes directly, neural QPP approaches may complement existing methods and improve prediction accuracy, particularly for queries where statistical or linguistic features provide ambiguous signals. Our research addresses this opportunity by exploring whether EEG signals captured during language processing can enhance QPP models beyond what is possible with traditional text-based approaches. This direction leverages the temporal precision of EEG to align neural measurements with specific aspects of query formulation, potentially revealing cognitive markers of query difficulty that could inform more adaptive retrieval systems.

3 Experimental Methodology

3.1 Dataset Creation

The exploration of neurophysiological signals for QPP requires a specialised dataset that aligns EEG recordings with corresponding query–document pairs. While numerous datasets exist for text-based QPP evaluation [16, 35] and several collections capture EEG signals during text processing

[29, 49], there is a notable absence of resources that explicitly pair neurophysiological recordings with IR-specific constructs like queries and documents with multiple graded relevance judgements. This gap significantly hinders research at the intersection of neuroscience and IR, particularly for investigating how brain signals during query processing might predict retrieval effectiveness.

3.1.1 Dataset Candidates. To address this critical limitation, we carefully selected and aggregated data from three of the largest and most comprehensive EEG-text paired datasets available: ZuCo 1.0/2.0 [39, 40], the Alice Dataset [10] and the Narrative Speech Dataset [13]. Our selection criteria prioritised datasets with: (1) large-scale data collection (with more than 10 participants) with EEG recordings, (2) precise temporal alignment between neurophysiological signals and linguistic input, (3) naturalistic reading or listening conditions that approximate real-world scenarios, (4) comprehensive documentation of experimental procedures and data structures and (5) public availability with clear licensing for research purposes.

The ZuCo datasets provide simultaneous EEG and **eye-tracking (ET)** recordings from 30 participants engaged in natural reading tasks, capturing neural and ocular responses with word-level precision across 21,629 words and 128 electrodes. The Alice Dataset contains EEG signals from 49 participants listening to a narrated story (the first chapter of 'Alice's Adventures in Wonderland'), offering detailed temporal alignment between neural signals and linguistic stimuli across 2,129 words using 61 electrodes. Similarly, the Narrative Dataset records EEG responses from 19 participants during continuous speech processing under naturalistic conditions, maintaining precise alignment between brain activity and 12,742 words of spoken language using 128 electrodes. Together, these datasets offer a rich foundation of neurophysiological recordings during text comprehension, spanning both reading and listening modalities across 98 total participants and approximately 80 hours of EEG recordings.

To ensure consistent feature representation across these heterogeneous sources, we implemented a standardised pre-processing pipeline. For the ZuCo datasets, we utilised the existing pre-processed EEG data segmented into standard frequency bands (theta1: −6 Hz, theta2: 6.5–8 Hz, alpha1: 8.5–10 Hz, alpha2: 10.5–13 Hz, beta1: 13.5–18 Hz, beta2: 18.5–30 Hz, gamma1: 30.5–40 Hz, gamma2: 40–49.5 Hz) [39, 40], yielding 840 features per word (105 electrodes × 8 frequency bands). For the Alice and Narrative datasets, we applied: (1) fourth-order Butterworth bandpass filtering (0.5–100 Hz) to remove high-frequency noise while preserving relevant neural oscillations [22, 51]; (2) average referencing across all channels to mitigate reference-related distortions [96]; (3) downsampling from the original sampling rate (512 Hz for Alice and 500 Hz for Narrative) to 128 Hz to achieve computational efficiency while maintaining temporal resolution [14]; (4) zero-padding to achieve a consistent 128-channel architecture (from 61 channels in Alice and 128 in Narrative); (5) band-pass filtering was applied separately for each of the eight frequency bands (theta1, theta2, alpha1, alpha2, beta1, beta2, gamma1, gamma2) using their respective frequency ranges; (6) the Hilbert transform is computed on each filtered signal to obtain the representation of the analytic signal; (7) the amplitude envelope is extracted by taking the absolute value of the analytic signal, capturing the instantaneous power in each frequency band; (8) temporal averaging is performed by computing the mean amplitude across all time points for each channel, reducing the temporal dimension while preserving spatial and spectral information, yielding 1,024 features (128 electrodes × 8 frequency bands). Channel padding was used rather than channel selection to preserve all available spatial information across datasets, as the low spatial precision of EEG means no single subset of channels captures all relevant activity patterns [11]. All channels were treated as independent features in subsequent analyses. No epoch rejection was applied. All recorded data were retained to maintain ecological validity and maximise available training data. Each word-level EEG segment (epoch)

corresponds to the exact temporal boundaries of word presentation, determined by ET fixations for ZuCo (mean duration: 250 ms) and audio timestamps for Alice/Narrative (mean duration: 350 ms).

The temporal alignment between EEG signals and linguistic input represents a critical aspect of our dataset creation process. For the reading datasets (ZuCo), we utilised eye-tracker–derived onset and offset timestamps to precisely map EEG segments to individual words. This approach leverages the natural relationship between eye fixations and cognitive processing during reading [39]. For the listening datasets (Alice and Narrative), we mapped word-level timestamps from the audio recordings to the corresponding EEG segments using index matching through NumPy's searchsorted function, ensuring accurate temporal alignment between neurophysiological recordings and linguistic stimuli. These alignment approaches maintain the ecological validity of the neurophysiological measurements while providing the temporal precision necessary for investigating cognitive processes during text comprehension.

A critical limitation of existing EEG-text datasets is that they do not contain naturally occurring queries paired with neurophysiological recordings. Ideally, we would have EEG signals recorded during actual query formulation by users with specific information needs. However, no such dataset exists at the scale required for robust QPP evaluation. While query formulation additionally involves intent translation and meta-cognitive evaluation of one's information need [64], the semantic comprehension processes engaged during reading and listening share a common cognitive substrate with the relevance judgements that underlie retrieval effectiveness [3, 34]. This shared substrate is precisely what neural retrieval pre-training methods exploit when treating text spans as implicit queries for their surrounding context [17, 54], and recent empirical work has demonstrated that reading-elicited EEG signals can be directly mapped to passage representations in a shared semantic space, supporting passage retrieval without intermediate query translation [59], lending credibility to their use as a proxy in the absence of purpose-collected data. To address this fundamental constraint, we employ the **Inverse Cloze Task (ICT)** methodology [54]; full details of the ICT procedure, including the formal formulation and generation algorithm, are provided in the supplementary material.

As shown in Table 2, this approach yielded an average query length of 4.43 words for the reading dataset (ZuCo) and 7.57 words for the listening datasets (Alice and Narrative), with an overall average of 6.00 words across all datasets, closely matching our target range.

3.1.2 Dataset Language Model Augmentation. While the ICT approach (see Figure 1) provided us with binary relevance judgements (a document is either relevant or not to its source query), contemporary IR evaluation metrics like NDCG and contemporary QPP tasks require more nuanced, graded relevance assessments across multiple documents [16]. To address this limitation, we augmented our dataset using controlled document generation with the LLaMA 2 (7B) language model [90].

The limited size of EEG-text paired datasets presents a significant challenge for developing robust neurophysiological-enhanced IR systems [29, 42]. A critical consideration in addressing this limitation is maintaining the integrity of the neurophysiological-text relationships within the data. We specifically chose to augment passages rather than queries, as query augmentation would disrupt the fundamental semantic coherence between the EEG signals and their corresponding text segments [97]. This decision is rooted in the temporal and cognitive nature of EEG recordings; each neurophysiological signature represents the specific cognitive processes involved in formulating and understanding its paired query text [64].

Recent research has demonstrated the effectiveness of leveraging large language models for zero-shot IR tasks. Studies like HyDE [30] showed that language models could generate hypothetical relevant documents to enhance retrieval performance. Additionally, PromptReps [103]

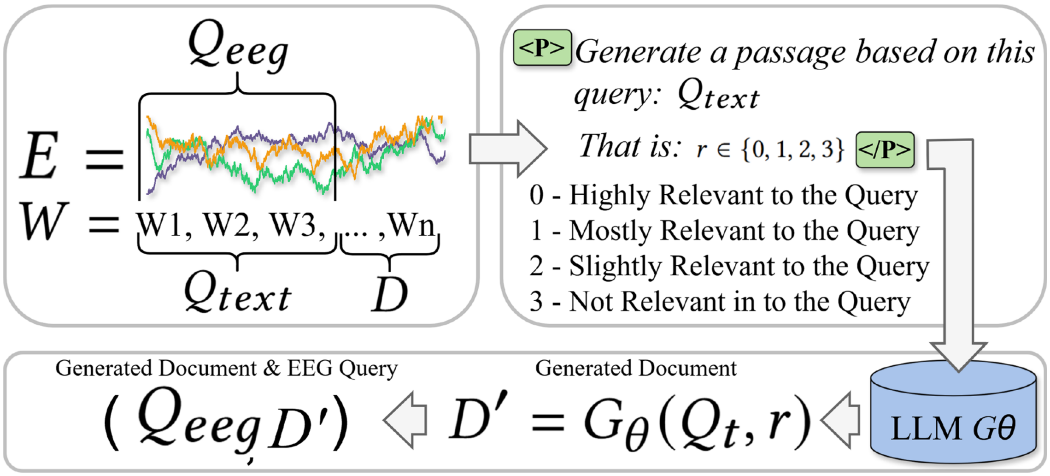


Fig. 1. ICT with EEG and text sequences, followed by document generation.

demonstrated that language models can directly produce effective query and document representations through careful prompt engineering. The recent introduction of ReDE-RF [44] further validated this direction by showing how language models can effectively judge document relevance to refine query representations. For our augmentation model, we utilise LLaMA 2 (7B) [91], chosen for its demonstrated capabilities in understanding and generating natural language while being computationally efficient. Recent work has demonstrated the effectiveness of LLMs in IR tasks, particularly through zero-shot and few-shot learning approaches [26]. These works, particularly on capturing nuanced relevance relationships, suggest that LLMs can effectively generate semantically controlled variations of documents while preserving meaningful relationships to queries [1].

For each query Q_t derived through ICT with its corresponding EEG signals Q_e, we defined a controlled generation function G_θ implemented using LLaMA 2 (7B), where θ represents the model parameters. This function takes a text query and relevance level $r \in \{0, 1, 2, 3\}$ as input:

$$G_\theta(Q_t, r) \rightarrow D'. \tag{1}$$

The generation is controlled through a prompt template:

$$P_r(Q_t) = [S; I_r; Q_t], \tag{2}$$

where S is the system context and I_r is the relevance-specific instruction. The passage is generated via nucleus sampling with nucleus probability $p = 0.9$ for quality control and temperature $T = 0.7$ for output diversity:

$$D' = \text{Sample}(G_\theta(P_r(Q_t)), p = 0.9, T = 0.7). \tag{3}$$

Our prompt engineering methodology emphasises precise control over generated content through a structured approach to instruction design. The system employs XML-tagged output requirements (`<passage></passage>`), building on established practices for controlling LLM outputs in IR tasks [30]. This design emerged from preliminary experiments where unconstrained generations often produced meta-commentary or preamble text that could contaminate relevance assessments. The prompting strategy employs templates for each relevance level, following approaches demonstrated effective in recent IR literature [45, 76]. For highly relevant passages (level 3), the prompt directs the model to 'Create a passage that conveys the same core information as the query but uses different vocabulary and phrasing'. For moderately relevant passages (level 2), the instruction becomes

Table 1. Example Query–Passage Pairs with Relevance Scores

Component	Content
Query	Kerouac got into trouble with the law following an incident involving a murder, which later formed the basis of a mystery novel…
Passage	**Original ICT Document** During Kerouac's time at Columbia University, he and Burroughs witnessed a murder that led to legal troubles. This incident directly inspired their 1945 collaboration 'And The Hippos Were Boiled in Their Tanks'…
Passage	**Generated Text (Relevance Score: 3)** The murder incident that entangled Kerouac with law enforcement became the primary source material for a mystery novel he would later write, drawing heavily from his firsthand experience…
Passage	**Generated Text (Relevance Score: 2)** Among Kerouac's many literary works, one novel drew inspiration from his encounters with crime and law enforcement, though it substantially fictionalised the events…
Passage	**Generated Text (Relevance Score: 1)** While primarily known for his spontaneous prose style, Kerouac occasionally wrote about crime, including one novel that touched on themes of murder and legal consequences…
Passage	**Generated Text (Relevance Score: 0)** In the depths of the Amazon rainforest, a group of scientists discovered a hidden underground cave system teeming with life…

The passages demonstrate varying degrees of relevance to the query, from highly relevant (score 4) directly addressing the specific context, to completely irrelevant (score 0).

Table 2. Statistics of the Generated Dataset (Split across Reading [ZuCo] and Listening [Alice+Narrative] Modalities)

Modality	#Subjects	#Pairs	#Queries	#Docs	Avg. Q Len	Avg. D Len
Reading	30	11,380	2,276	11,380	4.43 ± 2.43	73.71 ± 34.49
Listening	68	16,268	3,853	16,268	7.57 ± 4.41	81.26 ± 25.42
Total	98	27,648	6,129	27,648	6.00± 3.42	77.49± 29.96

'Create a passage that discusses the same general topic as the query but focuses on different aspects or details'. For slightly relevant passages (level 1), we instruct 'Create a passage that is only tangentially related to the query, sharing only the broadest thematic connections'. Finally, for non-relevant passages (level 0), the prompt specifies 'Create a passage that is completely unrelated to the query in topic and content, but maintains similar complexity'. Example query and passage pairs are illustrated in Table .1.

For each query, we generated passages with these four distinct relevance levels: highly relevant (level 3, preserving core semantic content with different vocabulary), moderately relevant (level 2, sharing the broader topic domain but focusing on different aspects), slightly relevant (level 1, preserving only broad domain connections) and non-relevant (level 0, maintaining similar linguistic complexity but discussing entirely different topics). This approach, inspired by recent work in synthetic IR dataset construction [6, 26], enabled us to create the fine-grained relevance gradations

necessary for robust IR evaluation while maintaining semantic coherence with the original text and its associated neurophysiological signals.

To ensure generation quality and consistency, we implemented several additional control mechanisms [1, 26]. A repetition penalty (1.1) prevented redundant phrase generation, while length constraints maintained consistent document sizes across relevance levels. The augmented dataset preserved all metadata from the original query–EEG pairs, including neurophysiological signatures and temporal alignment information, ensuring the ecological validity of the neurophysiological measurements was maintained throughout the augmentation process [29, 62].

3.2 Dataset Overview

To validate the quality of our augmentation process, we conducted a human evaluation where two experienced annotators assessed a stratified random sample of 300 query–document pairs. The results showed strong agreement both between annotators ($\kappa = 0.68$ for reading and $\kappa = 0.65$ for listening) and between human judgements and the generated relevance labels ($\kappa = 0.71$–0.73 for reading and $\kappa = 0.67$–0.69 for listening). These agreement levels are comparable to those reported in traditional IR evaluation campaigns [94], confirming that our controlled generation approach produces relevance gradations that align well with human judgements.

The resulting dataset comprises 27,648 query–document pairs across 6,129 unique queries from 98 participants, with each query represented by both text and temporally aligned EEG recordings. This resource enables the first systematic investigation of how neurophysiological signals can be applied to QPP, supporting research at the intersection of BCIs and QPP.

3.3 QPP Metrics

In line with common practice in QPP research, query performance is evaluated using NDCG@k. To obtain these performance metrics, we first indexed all documents in our collection using the BM25 algorithm [78], which ranks documents based on the frequency of query terms while accounting for term frequency saturation and document length normalisation:

$$\text{BM25}(q, d) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)}, \tag{4}$$

where $f(q_i, d)$ is the frequency of query term q_i in document d; $|d|$ is the document length; avgdl is the average document length across the collection; and k_1 and b are free parameters set to 1.2 and 0.75, respectively, following standard practice [78]. The IDF component is calculated as:

$$\text{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}, \tag{5}$$

where N is the total number of documents in the collection and $n(q_i)$ is the number of documents containing term q_i.

We selected BM25 as our retrieval algorithm for several important reasons. First, despite the emergence of neural retrieval models, BM25 remains a strong and reliable baseline in IR evaluation [88] and continues to be the standard first-stage retriever in many production systems. Second, its deterministic nature provides consistent and reproducible rankings, which is essential for reliable QPP evaluation. Third, BM25's efficiency allows for rapid indexing and retrieval across large document collections, making it particularly suitable for our extensive synthetic dataset. Finally, using BM25 maintains methodological consistency with prior QPP research [16, 35], enabling meaningful comparisons with established benchmarks.

For each query Q_t in the dataset, we retrieved the top-k documents and computed their performance scores based on the relevance judgements derived from our controlled document generation process, where relevance levels range from 0 (non-relevant) to 3 (highly relevant). This approach follows standard practices in QPP research [16, 35], where retrieval systems are evaluated against graded relevance assessments to produce performance metrics that serve as ground-truth labels for prediction models.

We use NDCG@k to evaluate graded relevance across the retrieved document list. NDCG@k assesses how well the ranking reflects the true relevance of the documents. It is computed as:

$$NDCG@k(Q_t) = \frac{DCG@k(Q_t)}{IDCG@k(Q_t)}, \tag{6}$$

where DCG@k is defined as:

$$DCG@k(Q_t) = \sum_{i=1}^{k} \frac{2^{rel(Q_t,D_i)} - 1}{\log_2(i+1)}. \tag{7}$$

IDCG@k represents the ideal DCG@k derived through ranking documents with respect to their actual relevance scores. This normalisation ensures that NDCG scores are bounded between 0 and 1, facilitating consistent comparisons between queries that have a varied number of associated relevant documents. This metric used has been found to be appropriate in the management of graded relevance judgements [16] and has been broadly used in QPP research [35].

For our metric, we set $k = 10$ following standard practice in passage retrieval evaluation [68, 88]. This value balances the need to capture sufficient ranking depth while focusing on the most relevant results typically presented in search result pages that users actually examine [95]. The choice of $k = 10$ is also consistent with previous QPP studies [16, 35], enabling meaningful comparisons with established benchmarks.

By computing these performance metrics for each query in our dataset using BM25 retrieval, we establish ground truth values that our QPP models aim to predict. This methodology is consistent with standard approaches in QPP research [35, 100], where the goal is to predict system performance prior to retrieving results. The use of BM25 as our retrieval algorithm is justified by its continued relevance as a strong baseline in IR evaluation [26, 88] and its widespread adoption in QPP studies [16, 35].

3.4 Training and Evaluation

Models were trained in both unimodal and multimodal settings. The unimodal setting involves training models on data from a single modality: text-only models simulate traditional pre-retrieval predictors, while EEG-only and ET-only (ZuCo only) models enable valid comparisons of each individual modality's performance. The multimodal setting involves training models on all possible combinations of modalities: EEG+Text, Text+ET, EEG+ET and the trimodal combination of all three modalities. This approach allows us to assess whether EEG data can complement text-based pre-retrieval predictors and enable the development of BCI-based QPP models.

Following conventional approaches, we treated QPP as a regression task, aiming to decrease the MSE between the predicted and actual performance scores [35]:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (M_i - \hat{M}_i)^2, \tag{8}$$

where M_i is the ground-truth score (NDCG@10) for query–document pair i and $\hat{M}_i$ is the score that our model with parameters θ predicts. Query text representations were tokenised using BERT's

WordPiece tokeniser, producing variable length sequences based on query length, which were padded to a length of 128 tokens.

EEG query representations were standardised across evaluation conditions, but their dimensionality differed slightly between the reading and listening datasets. For the reading dataset, EEG queries were represented as (number of query words, 840) for models requiring 1D inputs, where 840 corresponds to the mean frequency–amplitude values from 105 electrodes × 8 frequency bands. For models requiring 2D inputs, the same data were reshaped to (number of query words, 105, 8). For the listening dataset, EEG queries were represented as (1,024) for models requiring 1D inputs, where 1,024 corresponds to 128 electrodes × 8 frequency bands, and as (128, 8) for models requiring 2D inputs. To enable model training, all ZuCo EEG queries were padded to match the number of words in the longest query in the dataset.

Personalised Condition (Single-Subject Evaluation). In this condition, each model is trained on data from a single subject to evaluate whether BCI-based pre-retrieval predictors could be personalised for improved performance compared to text-only predictors. We evaluated models across all available modality combinations: unimodal (text-only, EEG-only, ET-only), bimodal (EEG+text, EEG+ET, ET+text) and trimodal (EEG+ET+text for ZuCo).

Generalised Condition (All-Subject Evaluation). In this condition, data from all subjects are pooled into a subject-agnostic dataset to evaluate whether a more generalised pre-retrieval predictor can be developed. Model architectures in this condition are identical as with the personalised condition.

In both conditions, models are trained using a train–validation–test split of 80/10/10. A fixed set of parameters was used to train each of the models. As is standard in QPP research [35], we report both difference and correlation metrics. For score differences, we use the MSE, where lower MSE indicates higher prediction accuracy. For correlation metrics, we report Pearson's ρ, Spearman's ρ and Kendall's τ, where higher positive correlations between predicted and actual scores suggest better ranking effectiveness. Statistical significance between models was assessed using paired t-tests comparing performance metrics across the 10 folds of cross-validation. To account for multiple comparisons, p-values were corrected using the Benjamini–Hochberg false discovery rate correction.

3.5 Model Implementations

3.5.1 Machine Learning Models. We implement stacking ensemble architectures using Scikit-Learn models (Random Forest and Gradient Boosting Regressors), selected for their proven effectiveness with EEG data [2, 12, 31, 55, 66]. Our ensembles support unimodal, bimodal and trimodal configurations.

Single-Modality Models. We train separate base models on individual modalities EEG, text and ET (only for ZuCo). An initial regression model is trained on the training data, which then produces predictions on the validation and test sets. The validation predictions are passed to a second regressor (meta-learner) for training. The meta-learner then generates predictions on the test data.

Multimodal Stacking. Our stacking architecture operates in two stages (see Figure 2) as follows:

(1) *Stage 1 (Base Models)*: We train separate base models on each modality's features. For bimodal stacking (e.g., EEG+ET), we train two base models. For trimodal stacking (EEG+ET+Text), we train three base models. Each base model is either a Random Forest or Gradient Boosting Regressor.
(2) *Stage 2 (Meta-Learner)*: The predictions from all base models are concatenated and used to train a meta-learner (stacker), which learns optimal weighting of individual modality

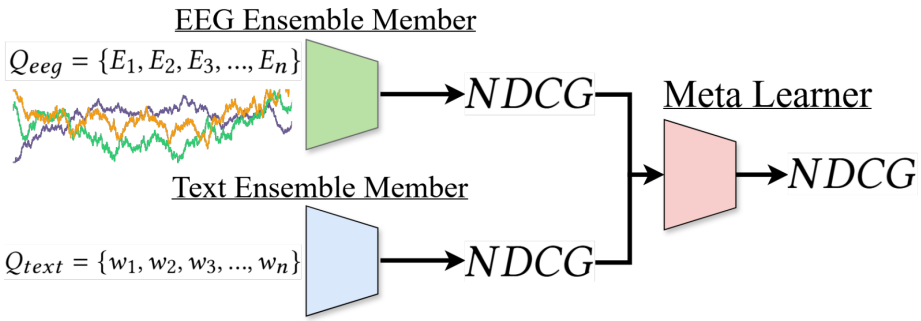


Fig. 2. Machine learning multi-modality ensemble overview.

predictions. The stacker is also a Random Forest or Gradient Boosting Regressor (same model type as the base models).

This stacking approach enables the meta-learner to discover non-linear combinations of modality-specific predictions, potentially capturing complementary information across the EEG signal, text and ET where available.

Model Hyperparameters. For Random Forest models, we use 250 estimators with unlimited tree depth. For Gradient Boosting models, we use 250 estimators with a learning rate of 0.5 and maximum depth of 3. All models use Scikit-Learn's random state (42) for reproducibility.

3.5.2 Deep Learning Models. We design and implement a range of deep learning models to process unimodal and multimodal signals, tailored for QPP. High dropout values were included in all models, as they were empirically found to reduce overfitting.

EEG Models. We implement the following three neural architectures for processing EEG signals:

- *LSTM*: A bidirectional LSTM with hidden size 256, 2 layers and mean pooling strategy. The model processes temporal EEG sequences and applies dropout (0.4) before the regression head.
- *CNN*: A 1D convolutional architecture with 3 convolutional blocks (channel sizes: 64, 128, 256) and kernel size 3. Each block includes batch normalisation, ReLU activation and dropout. Mean global pooling is applied before the regression layer.
- *BENDR*: We utilise a pre-trained BENDR model [50] consisting of a convolutional encoder and transformer-based contextualiser. The encoder and contextualiser weights are frozen during training, and we add a trainable regression head on top of the pooled contextual representations (hidden dimension: 512). An adaptive 1×1 convolutional projection layer is inserted when the input channel count differs from the pre-trained model's expected dimensions.

ET Models. For ET features, we employ the same LSTM and CNN architectures described above, with adjusted input dimensions to accommodate the ET feature space (default hidden size: 64 for LSTM). The architectures remain identical in terms of layer configuration and processing strategy.

Text Model (BERT-QPPcross). We use a pre-trained BERT model fine-tuned for QPP [5]. The BERT encoder weights are frozen, and we train only a dropout layer (0.4) followed by a linear regression head. The model uses BERT's pooler output (768-dimensional) as the query representation.

Multimodal Fusion. For bimodal and trimodal combinations, model embeddings are fused using simple concatenation of modality embeddings followed by an **Multi-Layer Perceptron (MLP)**.

3.6 Feature Importance and Model Explainability

To understand which EEG features drive QPP predictions, we conducted comprehensive feature importance analyses for the best-performing models in the generalised condition. This analysis provides insights into the neurophysiological mechanisms underlying query difficulty perception. Details on the explainability analysis procedure can be found in the supplementary material. All analyses were conducted on the held-out test sets from 10-fold cross-validation to avoid overfitting to training data patterns.

4 Results

Our results section is structured as follows. In Section 4.1, we present the results from the Generalised Condition, where models were trained and evaluated on pooled EEG data from all subjects. In Section 4.2, we present the results from the Personalised Condition, where models were trained and evaluated on data from a single subject at a time. Section 4.3 presents EEG frequency band importance analysis results, and Section 4.4 presents query-level performance analysis results.

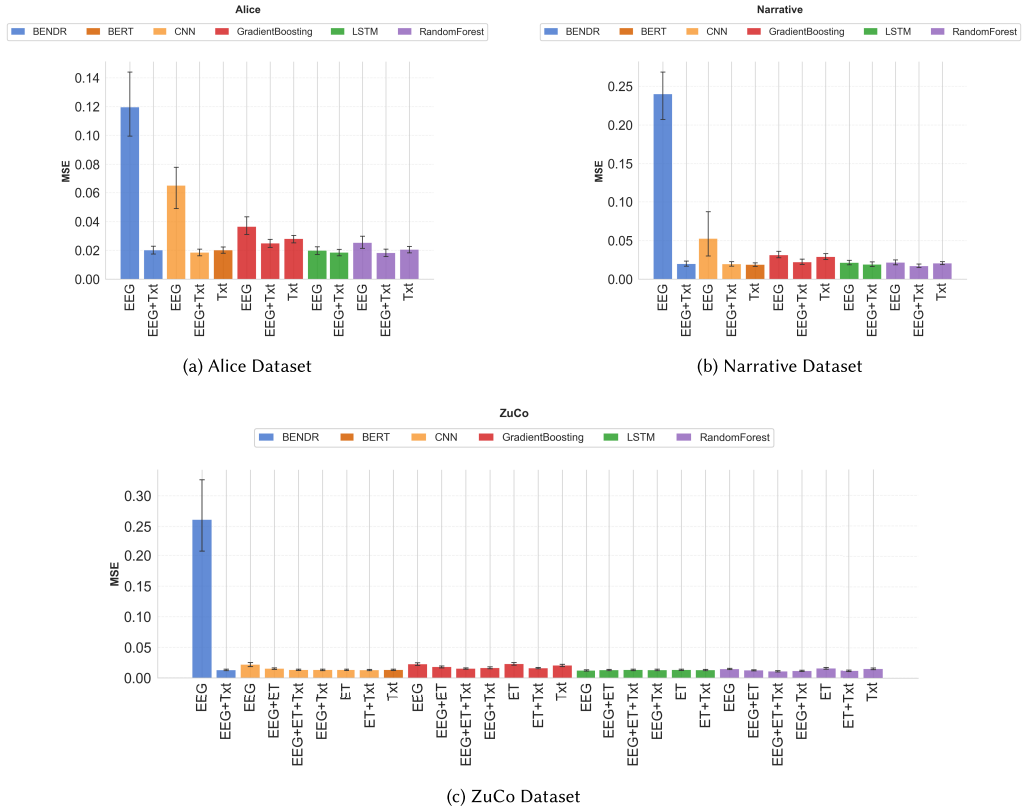
4.1 Generalised Results

Results for the Generalised Condition are presented in Figures 3 and 4, with full numerical results provided in Tables S1 and S2, and significance test results in Tables S5 and S6 in the supplementary material. Overall, the incorporation of neurophysiological and behavioural signals alongside text features consistently improved QPP performance, though the magnitude and statistical significance of these improvements varied across datasets and evaluation metrics. Reported correlation is consistent with previous research on query performance [5, 16].

The ZuCo reading dataset, which includes EEG, ET and text features, demonstrated the clearest benefits from multimodal fusion. Text-only models achieved correlations of Pearson: 0.29, Spearman: 0.29 and Kendall: 0.20, while the addition of EEG or ET individually produced significant improvements across all metrics. As shown in Figure 4(c), the trimodal model (EEG+ET+Text) yielded the strongest performance overall with Pearson: 0.46, Spearman: 0.47 and Kendall: 0.33, with statistically significant gains over all other configurations. These results suggest that EEG and ET capture distinct, non-overlapping aspects of cognitive processing that aid QPP. Notably, EEG-only and ET-only models showed moderate predictive power but substantially underperformed text-based models (see Figure 4(c)), reinforcing that neurophysiological signals are most valuable as complements to, rather than replacements for, textual features.

For the Alice listening dataset, EEG+Text models demonstrated the highest performance across correlation metrics with Pearson: 0.39, Spearman: 0.13, Kendall: 0.09 as displayed in Figure 4(a), with text-only models achieving the lowest MSE as made apparent in Figure 3(a). Figures 4(a) and 3(a) portray that EEG-only models yielded lower correlations and higher MSE compared to their text-based counterparts. The integration of EEG with text features (EEG+Text) produced improvements over EEG-only models across all metrics, with Random Forest achieving the best performance in this bimodal configuration. Statistical comparisons revealed that text models significantly outperformed EEG models, while EEG+Text models significantly outperformed EEG-only models and, for certain metrics, also exceeded text-only baselines. MSE differences between text and EEG+Text models were not statistically significant, suggesting that the primary benefit of EEG integration in this dataset lies in improving rank correlation rather than reducing prediction error.

As shown in Figure 4(b), the Narrative listening dataset exhibited higher overall correlations than Alice across all modality configurations, suggesting potentially stronger relationships between EEG activity and query performance for this dataset's queries. Consistent with Alice, text-only models outperformed EEG-only models across all correlation metrics. However, EEG+Text models



(a) Alice Dataset

(b) Narrative Dataset

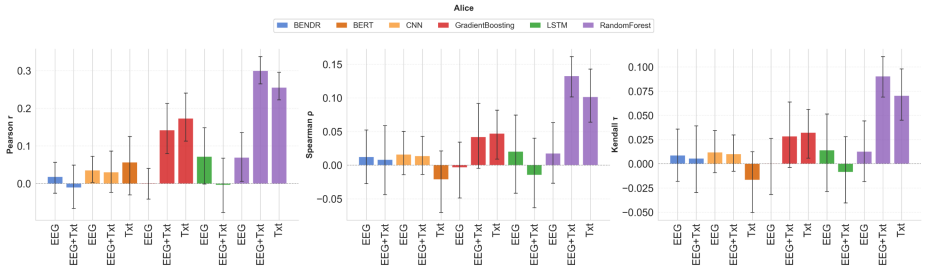
(c) ZuCo Dataset

Fig. 3. Query performance on the MSE *error* metric for generalised models across all three datasets. Bar plots show bootstrap confidence intervals for single-modality (Text, EEG, ET), bimodal (Text+EEG, Text+ET, EEG+ET) and trimodal (Text+EEG+ET) configurations, with colours indicating the model type. Note that ET data are only available in 3(c) of the ZuCo dataset.

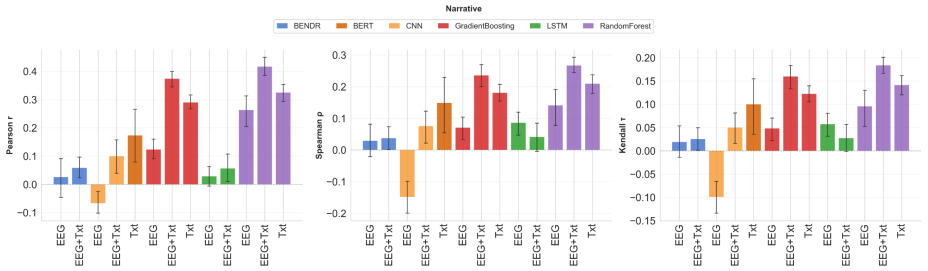
achieved the highest performance overall with Pearson: 0.42, Spearman: 0.27, Kendall: 0.18 with the lowest MSE value of 0.017. Random Forest emerged as the best-performing architecture across all modality conditions as evident in Figure 4(b). These results provide stronger evidence for the complementary value of EEG features in this dataset compared to Alice.

Figures 3 and 4 illustrate that across all three datasets, tree-based models (Random Forest and Gradient Boosting) consistently outperformed deep learning approaches (CNN, LSTM, BERT-QPP, BENDR) in both unimodal and multimodal settings. Random Forest in particular demonstrated the most substantial gains from multimodal fusion, with clear and consistent improvements when EEG and ET features were added to text baselines. Deep learning models showed weaker or inconsistent benefits from multimodal integration, suggesting they may require larger datasets, stronger regularisation, fine-tuning (in pre-trained cases) or alternative architectures to effectively leverage EEG signals for QPP. Between the two tree-based approaches, Random Forest generally achieved superior performance in multimodal configurations.

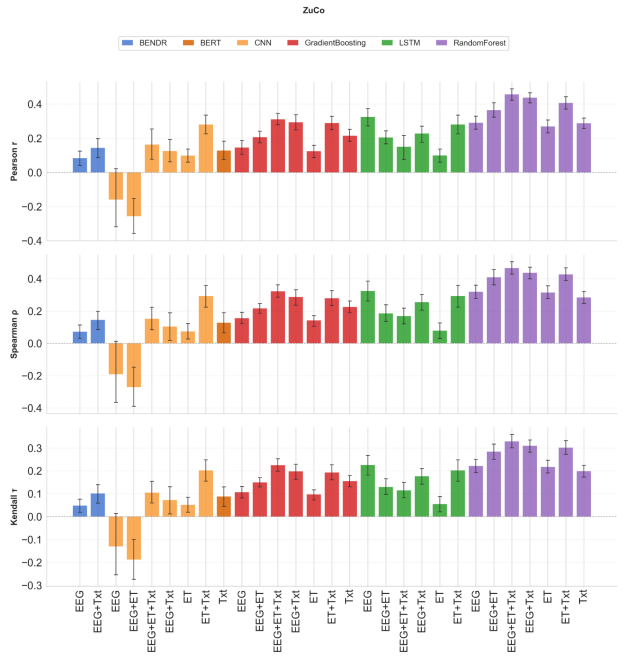
To verify that ensemble gains reflect genuine neurophysiological signal rather than ensemble mechanics alone, we conduct an inference-time noise ablation following methodology established in EEG-to-text research [57], where the Random Forest ensemble models trained on real EEG are evaluated with test-time EEG replaced by Gaussian noise $\mathcal{N}(\mu_{\text{EEG}}, \sigma_{\text{EEG}})$. As reported in Table S10 in



(a) Alice Dataset

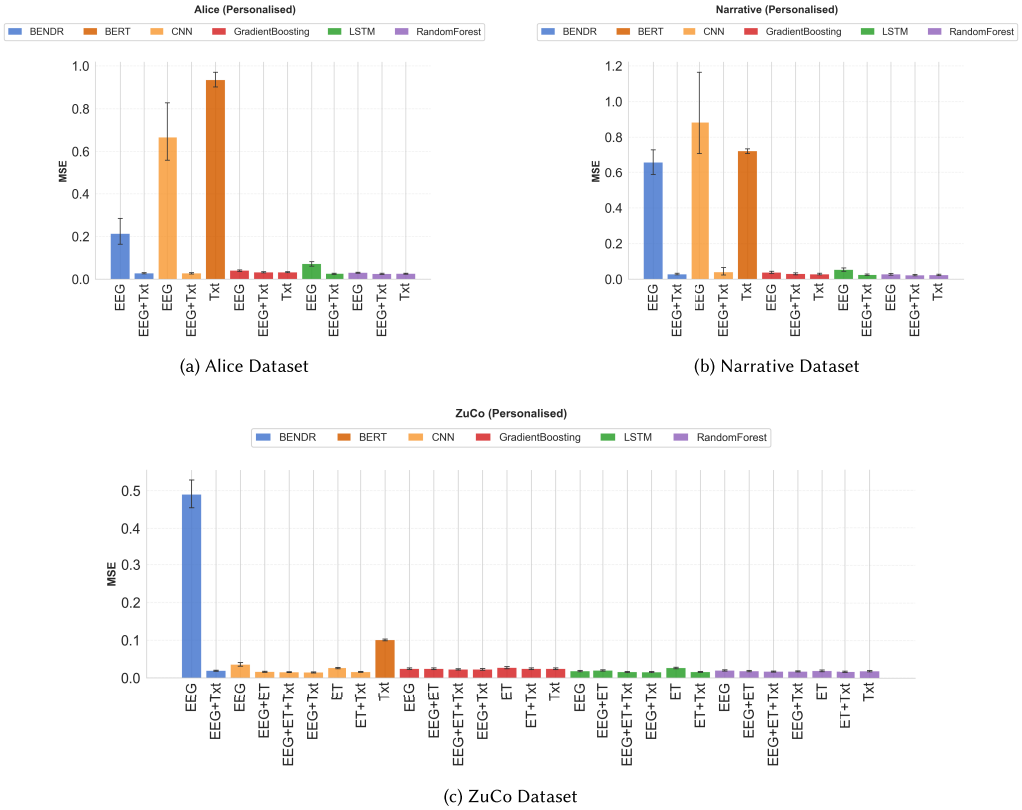


(b) Narrative Dataset



(c) ZuCo Dataset

Fig. 4. Query performance *correlation* metrics for generalised models across all three datasets. Bar plots show bootstrap confidence intervals for single-modality (Text, EEG, ET), bimodal (Text+EEG, Text+ET, EEG+ET) and trimodal (Text+EEG+ET) configurations, with colours indicating the model type. Note that ET data are only available in 4(c) of the ZuCo dataset.



(a) Alice Dataset

(b) Narrative Dataset

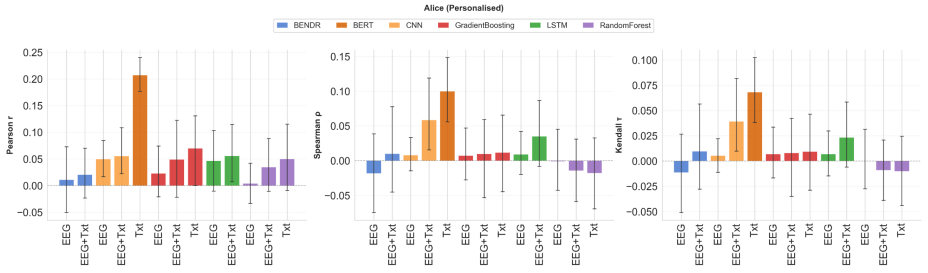
(c) ZuCo Dataset

Fig. 5. Query performance on the MSE *error* metric for personalised models across all three datasets. Bar plots show bootstrap confidence intervals for single-modality (Text, EEG, ET), bimodal (Text+EEG, Text+ET, EEG+ET) and trimodal (Text+EEG+ET) configurations, with colours indicating the model type. Note that ET data are only available in 5(c) of the ZuCo dataset.

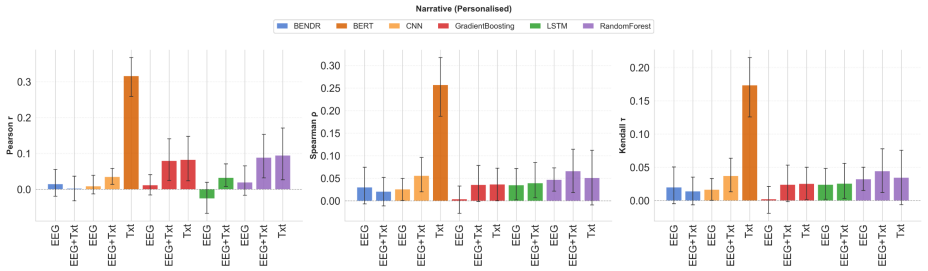
the supplemental material, real EEG consistently outperforms noise substitution across all metrics and datasets, with the exception of Pearson correlation on Alice. Beyond mean performance, the most noticeable difference is in prediction stability as real EEG yields lower standard deviations across all datasets as, e.g., in Spearman Correlation, ZuCo: 0.439 ± 0.061 vs. 0.264 ± 0.133, Narrative: 0.267 ± 0.042 vs. 0.188 ± 0.081; Alice: 0.132 ± 0.051 vs. 0.119 ± 0.079. Random Forest Ensemble scores displayed in Tables S1 and S2 in the supplemental material, for ZuCo and Narrative are even higher than the ensemble evaluated on random noise. Altogether, these indicate that especially in ZuCo and Narrative datasets, genuine EEG signal appears to improve mean performance and produces more reliable predictions across cross-validation folds.

4.2 Personalised Results

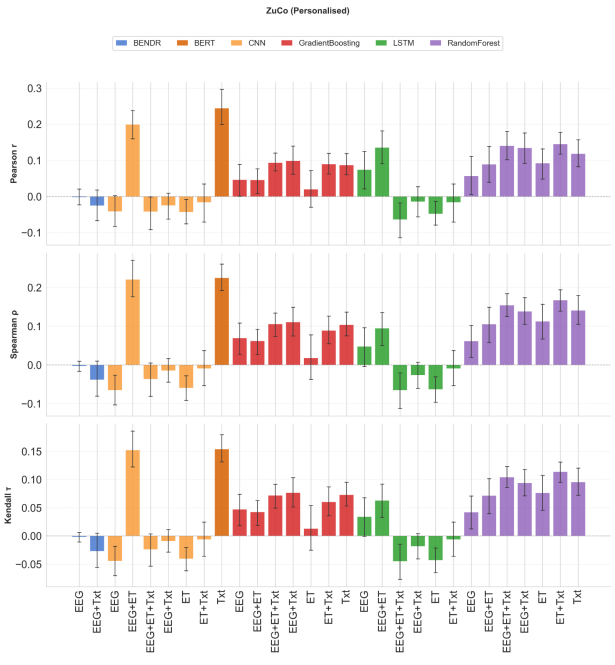
The results of personalised models, where each model is trained exclusively on data from a single subject, are shown in Figures 5 and 6, with full numerical results provided in Tables S3 and S4, and significance test results in Tables S7 and S8 in the supplementary material. In contrast to the generalised condition, personalised models exhibit greater performance variability, as shown in Figure 6 and reflected in the larger standard deviations across metrics reported in the supplementary tables. For instance, in the ZuCo dataset, the best personalised model (CNN with EEG+ET) achieves



(a) Alice Dataset



(b) Narrative Dataset



(c) ZuCo Dataset

Fig. 6. Query performance *correlation* metrics for personalised models across all three datasets. Bar plots show bootstrap confidence intervals for single-modality (Text, EEG, ET), bimodal (Text+EEG, Text+ET, EEG+ET) and trimodal (Text+EEG+ET) configurations, with colours indicating the model type. Note that ET data are only available in 6(c) of the ZuCo dataset.

a Pearson correlation of 0.368 ± 0.122, whereas the best generalised model (RF with EEG+ET+Text) achieves 0.458 ± 0.054 demonstrating both lower absolute performance and increased variability in the personalised setting. This pattern is observed across all three datasets, with personalised models showing standard deviations frequently exceeding ±0.10 for correlation metrics (Figure 6), compared to ±0.05–0.07 in generalised models (Figure 4). These results may reflect the inter-subject variability in EEG responses during language processing and possibly suggest a challenge for building effective subject-specific QPP models with limited per-subject training data (<70 samples per participant).

For the ZuCo reading dataset, personalised models demonstrate modest benefits from multimodal fusion, though these gains are smaller than those observed in the generalised condition. Text-only models (BERT) achieve Pearson correlation of 0.225 ± 0.113, significantly outperforming both EEG-only and ET-only models. The integration of EEG or ET with text yields improvements, with EEG+Text (RF) reaching 0.168 ± 0.144 and ET+Text (RF) achieving 0.202 ± 0.161. Notably, the bimodal EEG+ET combination using CNN achieves the highest correlations overall, though this performance is not significantly better than ET+Text for correlation metrics. As shown in Figure 6(c), the trimodal model does not significantly outperform the best bimodal models, in contrast to the generalised condition where the trimodal model achieved superior performance. Statistical comparisons reveal that while multimodal models significantly outperform EEG-only baselines, they show mixed results against text-only models, with some bimodal configurations (e.g., EEG+ET) achieving non-significant but slightly higher correlations while text alone remains competitive.

In the Alice listening dataset, personalised models show limited benefits from incorporating EEG signals alongside text features as illustrated in Figures 6(a) and 5(a). Text-only models demonstrate the strongest performance, with BERT achieving Pearson correlation of 0.097 ± 0.192, significantly outperforming EEG-only models across all correlation metrics as presented in Figure 6(a). The addition of EEG to text in bimodal models does not yield significant improvements over text alone and shows a significant reduction in Pearson correlation. While EEG+Text models achieve significantly better MSE than EEG-only models, they do not significantly improve upon text-only MSE. These results contrast with the generalised condition, where the best EEG+Text model (RF: 0.299 ± 0.060) outperformed the text-only equivalent and showed clearer benefits over text alone. Figure 6 illustrates higher variability in personalised text models (±0.192) compared to generalised models (±0.060, see Figure 4) suggesting individual differences in how text features predict query performance and that EEG signals may not provide consistent complementary information at the individual subject level for listening tasks.

The Narrative listening dataset exhibits strong text dominance in personalised models, with BERT achieving the highest correlations across all datasets and modality combinations as evident in Figure 6. As evident in Figure 6(b), EEG-only models demonstrate limited predictive power, with correlations near zero (best: LSTM, Pearson 0.004 ± 0.099). The contrast is notable: generalised EEG+Text achieves Pearson 0.417 ± 0.054, while personalised EEG+Text achieves 0.110 ± 0.107, a reduction of 0.307 in mean correlation with increased variability (clearly illustrated in Figure 6(b)). This pattern differs from the generalised condition, where EEG+Text models achieved the best overall performance and significantly outperformed text-only baselines. These results suggest that for the Narrative dataset, effective integration of EEG signals may require the cross-subject learning enabled by generalised models, and that subject-specific models may lack sufficient data to extract meaningful EEG-based features.

Across all three datasets, personalised models show task-dependent patterns in the utility of EEG signals for QPP. The ZuCo reading dataset shows multimodal benefits, with EEG+ET and ET+Text combinations approaching or matching text-only performance, suggesting that reading-related cognitive processes may generate more consistent individual-level neural signatures. In contrast,

Figure 6 suggests that both listening datasets show limited or negative contributions from EEG in personalised settings, with text-only models generally achieving stronger performance in terms of correlation. Architectural preferences differ from generalised models: while generalised RF trimodal models achieved the best ZuCo performance, personalised CNN models with EEG+ET achieve the highest correlations (0.368 ± 0.122), suggesting different optimal architectures depending on training paradigm. The performance gap between generalised and personalised models across all datasets, with generalised models achieving higher correlations with lower variance, evident between Figures 4 and 6, which indicates increased benefits for cross-subject learning in multimodal QPP.

4.3 Feature Importance Analysis

Feature importance analysis of the Random Forest models, presented in Figure 7, reveals that ZuCo and Narrative exhibit relatively uniform EEG frequency band importance distributions, where the full spectrum contributes meaningfully to QPP predictions. In contrast, the Alice dataset shows a more skewed pattern, with beta1 receiving substantially higher importance (~0.16) while theta bands contribute comparatively little. The balanced distributions observed in ZuCo and Narrative are consistent with the multifaceted nature of language processing, in which distinct frequency bands subserve complementary cognitive functions [32, 41, 56, 69, 86]. The skewed pattern in Alice may reflect the zero-padding applied to match electrode dimensionality across datasets, as *post hoc* validation confirmed that padded channels received non-negligible importance. A full breakdown of modality level and frequency band importance across all datasets and modality combinations is provided in the supplemental material.

4.4 Query-Level Performance Analysis

Whilst aggregate correlation metrics provide an insight into overall model performance, they obscure important query-level variation in the effectiveness of neurophysiological enhancement. To address this limitation, we conducted a comprehensive query-level analysis categorising individual queries based on whether the incorporation of EEG signals improved, degraded, or had a negligible impact on prediction accuracy. Queries were assigned to one of three groups based on the change in prediction error relative to the text-only baseline:

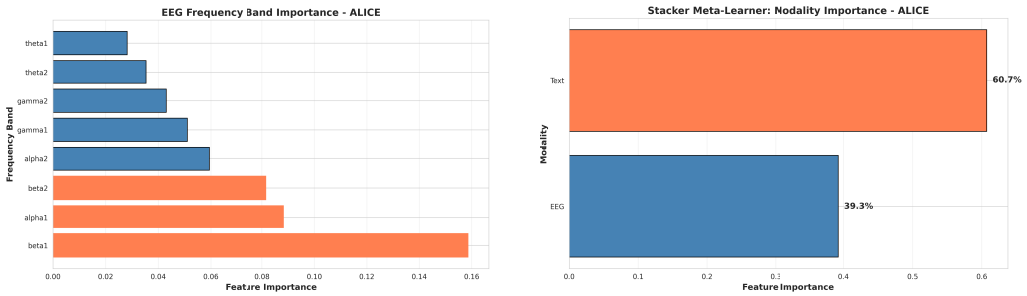
$$\text{Improvement} = \text{Error}_{\text{text}} - \text{Error}_{\text{EEG-enhanced}}, \tag{9}$$

yielding *EEG-helped* (Improvement > 0.1), *Neutral* ($-0.1 \leq$ Improvement ≤ 0.1) and *EEG-hurt* (Improvement < -0.1) categories. Full methodological details and per-configuration breakdowns are provided in the supplementary material.

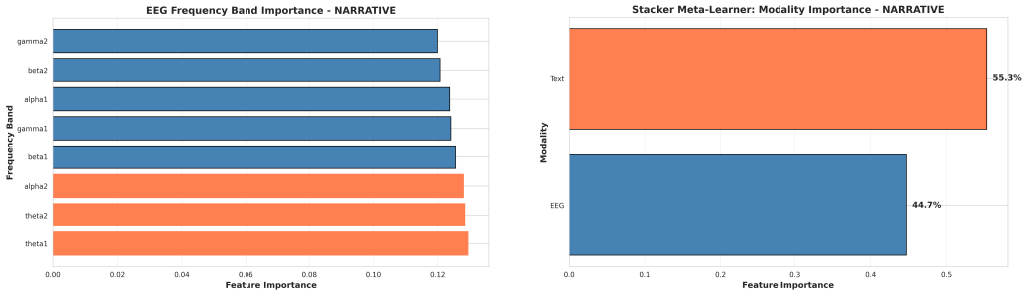
Distribution of Query-Level Outcomes. Table 3 summarises the distribution of query-level outcomes and mean prediction errors for the best-performing model in each dataset and modality configuration.

Across all datasets, the majority of queries (56–91%) fall into the neutral category, where text-only and EEG-enhanced models perform comparably. Of the remaining queries, the Narrative EEG+Text configuration shows the highest proportion of helped queries (25.0%), followed by ZuCo trimodal (19.4%) and Alice EEG+Text (16.7%), with hurt query proportions ranging from 8.7 to 19.0%. Text-only predictions drive the dominant signal in most cases, with EEG providing selective benefit rather than uniform improvement.

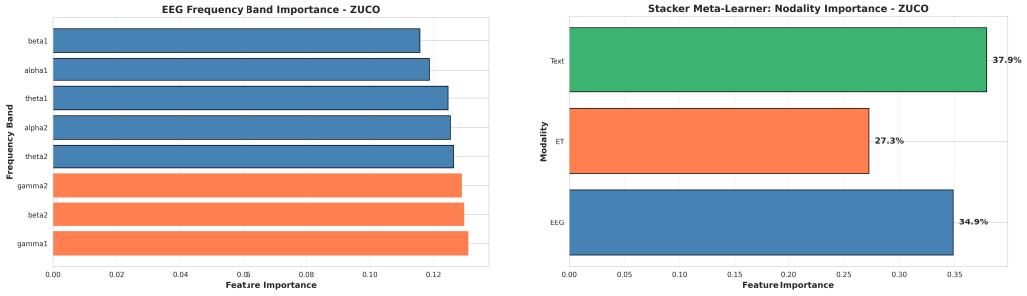
Exemplar Query Analysis. Table 4 presents representative queries from each outcome category. Helped queries consistently feature incomplete sentence fragments or semantically ambiguous content where text-only models exhibit high error (0.28–0.78); the incorporation of EEG signals reduces these errors substantially, often to near-zero. Hurt queries are characterised by very short, semantically underspecified phrases (e.g., 'not a', and 'all around') where text-only models already



(a) Alice



(b) Narrative



(c) ZuCo

Fig. 7. Feature importance analysis across datasets. For each dataset, left panels show Random Forest importance scores for EEG frequency bands (theta, alpha, beta, gamma), while right panels show stacker-level feature importance combining modalities. ZuCo includes ET features alongside EEG and text, while Alice and Narrative use EEG+Text combinations. In the EEG frequency band importance analysis figures, the top three bands are highlighted in orange.

achieve low error, and EEG signals introduce catastrophic noise (errors frequently exceeding 1.0). Neutral queries display moderate linguistic completeness, where both modalities capture similar aspects of query difficulty and their combination yields no additional predictive benefit. Full per-dataset discussion is provided in the supplementary material.

Signal Quality and Predictive Performance. We further examined the relationship between EEG **signal-to-noise ratio (SNR)** and query-level prediction benefit. Figure 8 reveals striking dataset-specific patterns. For ZuCo, the trimodal EEG+ET+Text configuration shows strong SNR dependency: high-SNR queries (>10 dB) yield approximately 38% helped queries vs. approximately 22%

Table 3. Query-Level Performance Summary across Datasets and Modality Configurations

Dataset	Modality	Model	Helped (%)	Neutral (%)	Hurt (%)	Mean Δ	Error
ZuCo (Reading with ET)							
	EEG+ET+Text	Random Forest	**19.4**	63.4	17.2	**0.0129**	**0.0768**
	EEG+Text	LSTM	5.3	**90.9**	3.8	0.0019	0.0878
	EEG	LSTM	6.1	90.4	**3.5**	0.0047	0.0850
Alice (Passive Listening)							
	EEG+Text	BENDR	16.7	**74.7**	**8.7**	**0.0066**	**0.1031**
	EEG	LSTM	**22.0**	65.1	13.0	0.0033	0.1065
Narrative (Passive Listening)							
	EEG+Text	Random Forest	**25.0**	56.0	19.0	**0.0045**	**0.0936**
	EEG	LSTM	9.7	**78.2**	**12.1**	−0.0077	0.1058

Values represent the percentage of queries in each category (helped: improvement > 0.1, neutral: $|\text{improvement}| \leq 0.1$, hurt: improvement < -0.1), mean improvement (Δ), and mean prediction error. Best-performing model per modality configuration is shown for each dataset. Bold values indicate the best performance within each dataset group.

at medium SNR (0–5 dB), a near-twofold difference attributable to the precise temporal alignment between eye fixations and EEG signals enabling cleaner capture of word-level cognitive processing. In contrast, EEG-only and EEG+Text configurations in ZuCo, and both configurations in Alice and Narrative, exhibit largely flat performance across SNR bins, suggesting that passive listening tasks face fundamental temporal alignment challenges that dominate over technical signal quality variations. Full mechanistic interpretation is provided in the supplementary material.

4.5 Summary of Findings

Our results provide evidence that neurophysiological signals can effectively predict query performance, though this effectiveness is conditional on the training paradigm, task type and query characteristics. In generalised settings, multimodal models consistently outperform text-only baselines across all datasets: the ZuCo trimodal configuration (EEG+ET+Text) achieves the highest performance overall (Pearson 0.458), Narrative's EEG+Text model significantly outperforms its text-only baseline (0.417 vs. 0.324), and Alice's EEG+Text model shows modest improvement (0.299 vs. 0.254). Feature importance analysis confirms that EEG contributes 34.9–44.7% of predictive information in these models, with balanced modality contributions in the best-performing configurations (ZuCo: 37.9% text, 34.9% EEG and 27.3% ET). These results are consistent with theoretical accounts of ensemble gains, which predict substantial improvements when component models rely on fundamentally different signal sources and thus produce complementary errors [89].

Generalised models substantially outperform personalised models, achieving higher correlations with 2–3 times lower variability (ZuCo: 0.458 ± 0.054 vs. 0.368 ± 0.122; Narrative: 0.417 ± 0.054 vs. 0.110 ± 0.107). In personalised settings, text-only models are often competitive or superior, suggesting that limited per-subject data relative to the EEG feature space may impede reliable pattern extraction [90]. Query-level analysis further reveals that neurophysiological benefits are not uniform: EEG signals improve predictions for 19.4–25% of queries, particularly fragmentary or semantically ambiguous ones where text-only models exhibit high error (0.28–0.78), while very short queries where text-only predictions are already accurate are actively harmed (8.7–19% of queries). Taken together, these findings establish that effective neurophysiological QPP requires

Table 4. Representative Exemplar Queries Illustrating When Neurophysiological Signals Help, Hurt, or Have Neutral Impact on Query Performance Prediction

Dataset	Category	Query Text	Text Error	Dual Error	Δ
ZuCo Reading Dataset					
	Helped	went Reprise again, shifting the next year to	0.6792	0.0000	**+0.6085**
		Courage, highlighting eight instances in which	0.4228	0.0000	+0.3492
		Senators risked their careers by standing			
		in 1953, divorced in 1961,	0.3157	0.0000	+0.2988
	Neutral	Brothers in 1931, became	0.3515	0.0000	+0.0268
		Phi Beta Kappa.	0.0483	0.0000	−0.0315
		Geiger	0.2381	0.0000	−0.0059
	Hurt	not a	0.0565	0.0000	**−1.4790**
		English Young December	0.0057	0.0000	−1.9916
		Batman, and a	0.0224	0.0000	−1.2263
Alice Listening Dataset					
	Helped	and the White	0.5586	0.0000	**+0.3805**
		she could	0.4510	0.0000	+0.3339
		and that if you	0.2779	0.0000	+0.2394
	Neutral	Down down	0.1117	0.0000	−0.0910
		Alice	0.3798	0.0000	−0.0457
		a candle is like after	0.0149	0.0000	−0.0173
	Hurt	me at home Why I would nt say anything about	0.0480	0.0000	**−0.6970**
		all around	0.0534	0.0000	−0.7143
		the neck of the bottle was a paper label with	0.1153	0.0000	−1.1968
Narrative Listening Dataset					
	Helped	TOO	0.7791	0.0000	**+0.4733**
		MAKE ME	0.4107	0.0000	+0.3484
		WOULD	0.7246	0.0000	+0.3312
	Neutral	ASKED	0.2217	0.0000	+0.0321
		AM	0.2703	0.0000	+0.0103
		BIRD CAME TOWARD SKIFF	0.0686	0.0000	+0.0009
	Hurt	COME WIND	0.1087	0.0000	**−1.0674**
		LONGER	0.0547	0.0000	−1.1914
		COME UP EASY LET	0.0078	0.0000	−1.2074

Queries are selected from each category to demonstrate characteristic patterns: *helped* queries show high text-only error substantially reduced by EEG signals; *Hurt* queries show low text-only error catastrophically increased by EEG; *Neutral* queries show minimal change between modalities. Δ represents improvement (Text Error − Dual Error). Bold indicates the highest improvement values in the example queries which either improved or hurt. Note the contrasting patterns: fragmentary queries benefit most from EEG, while very short phrases suffer most from EEG noise.

sufficient cross-subject training data, high signal quality and query characteristics that create textual uncertainty, which can be aided by high-quality neurophysiological data.

5 Discussion

In this research, we investigated whether neurophysiological signals captured during passive text comprehension can enhance QPP when combined with text-based features. Using a multimodal stacking architecture across three datasets—spanning, reading and listening tasks—we trained models integrating EEG, ET and text features in both generalised and personalised settings. The results reveal a nuanced picture: multimodal models yield consistent improvements over text-only baselines in generalised settings, but this benefit largely disappears in personalised settings, with query characteristics and signal quality acting as further moderating factors. We speculate that personalised performance degradation may reflect the curse of dimensionality [90], as each subject contributes approximately 70 query–score pairs against a feature space of hundreds of EEG features.

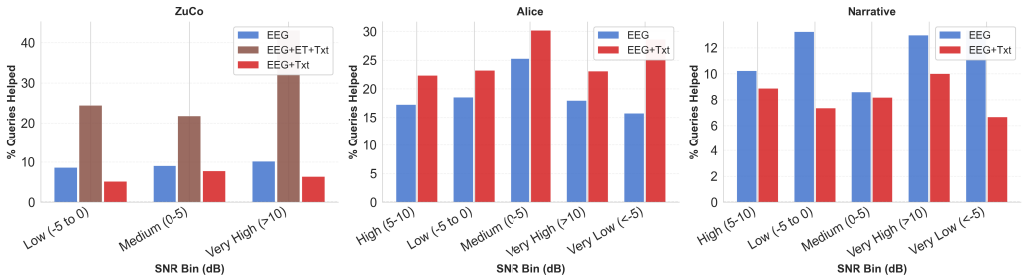


Fig. 8. QPP as a function of EEG signal quality. Bar plots show the percentage of queries helped by each modality configuration across SNR bins (Very Low: <−5 dB, Low: −5 to 0 dB, Medium: 0–5 dB, High: 5–10 dB and Very High: >10 dB). ZuCo includes three modality configurations: EEG+Text (orange), EEG-only (blue) and EEG+ET+Text (green). Alice and Narrative compare EEG-only (blue) and EEG+Text (orange) models. The ZuCo dataset shows strong SNR dependency, particularly for the trimodal configuration, whilst Alice and Narrative exhibit relatively flat performance across SNR bins.

The key contributions of our research are several fold. First, it provides a systematic investigation demonstrating that neurophysiological signals captured during text processing can predict query performance, establishing preliminary yet significant evidence for the feasibility of EEG-enhanced QPP, as apparent in Tables S5 and S6 in the supplemental material. Second, it employs a multimodal stacking architecture as a principled implementation choice for integrating neurophysiological, behavioural and text-based features, enabling controlled comparison of modality contributions for QPP. Third, the study provides empirical evaluation of EEG-enhanced QPP across both reading and listening modalities, demonstrating that generalised models trained across subjects can yield improvements over text-only approaches in specific configurations, though the magnitude and significance of these improvements vary across datasets and model types, and are absent or reversed in personalised settings where inter-subject EEG variability limits the extraction of neurophysiological features. Fourth, it provides a systematic evaluation of both personalised and generalised EEG-based prediction models, including a query-level analysis of the specific conditions under which EEG signals improve, degrade, or have negligible impact on prediction accuracy, as discussed in Section 4.4. Finally, the research provides an exploratory analysis suggesting task- and dataset-level conditions under which neurophysiological signals tend to provide predictive value, complemented by statistical breakdowns and representative exemplar queries illustrating when EEG improves, has a negligible effect, or degrades prediction accuracy (Tables 3 and 4). While a complete characterisation of which individual queries benefit from EEG enhancement remains an open challenge beyond the scope of this study, these findings offer initial empirical grounding for the design of neurophysiologically enhanced IR systems.

Our research presents certain limitations both in terms of IR methodology and EEG implementation. In terms of IR, our use of synthetic passage generation introduces notable constraints despite enabling controlled experimentation. While large language models have gained acceptance in the IR community for evaluation content generation [26, 76], synthetically generated passages inevitably lack some characteristics of naturally occurring text. Specifically, synthetic passages may not fully capture domain-specific terminology, contextual nuances and the linguistic complexities that emerge in real-world documents. Moreover, our query-level analysis demonstrates that EEG's predictive value depends critically on query characteristics, fragmentary or ambiguous queries benefit most, while very short phrases suffer most, suggesting that the distribution of query types in real search logs may substantially affect the practical utility of neurophysiological enhancement.

This limitation potentially affects how well our findings generalise to scenarios involving more diverse document collections with specialised terminology or complex discourse structures.

The ecological validity of our approach is further limited by the neurophysiological data collection conditions. The EEG signals in our dataset were recorded during passive reading and listening tasks, which represent a controlled but simplified version of information processing. In contrast, real-world search scenarios involve active engagement where users formulate queries in response to specific information needs, experience varying levels of cognitive load and adapt their strategies based on retrieved results. The cognitive processes involved in passive comprehension likely differ substantially from those in interactive search contexts where factors such as search intent, task complexity and user engagement dynamically influence behaviour and effectiveness. Reading and listening engage semantic comprehension and language integration processes that, while not identical to query formulation, represent the same fundamental cognitive operations that neural retrieval models exploit when learning query–document relevance [17, 54] and which have been empirically shown to encode sufficient semantic structure to support direct passage retrieval from brain signals [59]. Notably, our findings suggest that more active cognitive engagement may produce stronger neurophysiological signatures of query difficulty than passive tasks, though whether this extends to truly interactive search remains an open question.

The temporal alignment between EEG signals and textual content, while carefully implemented, represents another methodological limitation. Natural query formulation typically involves iterative refinement processes, meta-cognitive evaluations and decision-making stages that may not be fully captured in our controlled reading paradigm. The EEG signals associated with reading a span of text that was algorithmically selected as a query might not accurately represent the neural activity that would occur during intentional query generation by a user with a specific information need. From a modelling perspective, we evaluated both tree-based ensemble methods (Random Forest and Gradient Boosting) and deep learning approaches (CNN, LSTM, BERT-QPP, BENDR) to determine which architectures most effectively integrate neurophysiological signals for QPP. We prioritised establishing fair comparisons between unimodal and multimodal approaches using models of similar complexity within each paradigm, helping to isolate the impact of adding neurophysiological signals rather than confounding this effect with architectural differences. Our results revealed that tree-based models consistently outperformed deep learning approaches across all multimodal configurations in the generalised condition, despite mixed findings in the personalised condition. For instance, in the ZuCo dataset, Random Forest achieved 0.458 for the trimodal configuration compared to deep learning models which remained below 0.17, suggesting that tree-based ensemble methods may be more effective at capturing non-linear interactions between modalities without requiring the large training datasets that deep learning models typically need for multimodal fusion. This finding validates the use of ensemble methods for neurophysiological QPP and suggests that architectural complexity may be less critical than effective multimodal feature integration.

Regarding EEG methodology, several inherent challenges affect the applicability of our results. EEG data exhibit substantial between-subject and within-subject variability [82], which complicates the generalisation of findings. This variability stems from factors including differences in skull thickness, neural organisation, cognitive strategies and attentional states, all of which influence the electric fields measured at the scalp. The high individual variability we observed in personalised models likely reflects these underlying neurophysiological differences, even though as discussed before this may also be attributed to the limited sample points per subject compared to EEG features, which may make QPP challenging for certain subjects. However, our results demonstrate that generalised models can substantially overcome this individual variability through cross-subject learning, as generalised models achieved higher correlations with 2–3 times lower variance compared to personalised models. This suggests that pooling data across subjects enable extraction

of invariant neurophysiological signatures of query processing that transcend individual differences or personalised modelling complications, though this comes at the cost of requiring large-scale data collection. Furthermore, our SNR analysis revealed that signal quality critically determines EEG's utility, as in the ZuCo reading dataset, queries with SNR >10 dB showed nearly double the benefit compared to medium-quality signals, whereas listening datasets exhibited no SNR dependency, suggesting that technical improvements in signal quality may yield returns only for specific task modalities, where precise temporal alignment between neural activity and text processing is achievable. Furthermore, EEG signals are highly susceptible to environmental noise [28] and physiological artifacts from sources such as muscle movements, eye blinks and cardiac activity [19]. These artifacts introduce noise that can obscure the neural signals of interest, potentially limiting the reliable extraction of cognitive features related to query processing. Our query-level analysis provides empirical evidence of this challenge, with artifact contamination producing substantial errors particularly for very short queries. Additionally, methodological artifacts can affect interpretation, as in the Alice dataset, zero-padding of electrode channels to match dimensionality across datasets resulted in models assigning importance to padded channels, suggesting that models may exploit structural artifacts when genuine signal is weak or noisy. While we applied standard pre-processing procedures, residual artifacts may still affect our analysis, further complicating the practical application of EEG-based approaches in real-world IR systems where recording conditions are less controlled.

An important insight from our feature importance analysis concerns the relationship between feature utilisation patterns and predictive performance. The best-performing datasets demonstrated relatively uniform importance across EEG frequency bands, as ZuCo showed remarkably balanced utilisation with all bands contributing nearly equally, while Narrative exhibited similar patterns. In contrast, the weaker performing Alice dataset displayed highly skewed utilisation, with beta1 achieving substantially higher importance while theta bands contributed minimally, though this pattern should be interpreted cautiously given the padding artifacts discussed above. This relationship of distributed evenly features and performance suggests that successful neurophysiological QPP depends not merely on the presence of EEG signals, but on the model's ability to extract complementary information across the full frequency spectrum. Cross-spectrum utilisation observed in high-performing configurations indicates effective integration of diverse neural signatures associated with different cognitive processes that are encompassed within the scope of language processing, as briefly described in Section 4.3 and in more detail in Section S3.3 of the supplemental material. When models achieve this balanced extraction, they appear to capture a comprehensive neurophysiological profile of query terms rather than exploiting artifacts or noise in specific frequency ranges.

Future research in neurophysiologically enhanced QPP should address several key limitations identified in our study. To improve the ecological validity of the dataset, future work could focus on using more realistic, naturally occurring documents instead of synthetically generated passages, ensuring that the complexity and nuance of real-world evaluation content are captured [42, 46]. Additionally, future research should prioritise recording neurophysiological signals during active search tasks, such as query formulation [97] and interaction with retrieval systems, to better reflect the cognitive processes involved in actual IR scenarios [95]. Given our query-level findings showing differential benefits across query types, future work should investigate how the distribution of query characteristics in real search logs affects the utility of neurophysiological enhancement. Understanding the prevalence of queries that benefit from EEG signals vs. those that are harmed in naturalistic search contexts would help establish realistic expectations for practical deployment of EEG-enhanced QPP systems.

The development of more practical and unobtrusive EEG recording devices represents another critical direction for advancing this research. Recent advances in dry-electrode technology and

wireless wearable EEG headsets [53] offer promising opportunities for collecting neurophysiological data in more naturalistic settings, potentially addressing the ecological validity concerns while making such systems more practical for everyday use. Since our SNR analysis suggests that signal quality improvements may yield particularly strong returns for active cognitive tasks as reading tasks with ET showed strong SNR dependency, while passive listening tasks exhibited flat performance across SNR ranges. Given that query formulation in real-world search represents an active cognitive process more similar to reading than passive listening, these findings suggest that hardware improvements in signal quality could enhance the utility of neurophysiological signals for QPP in naturalistic search contexts, particularly when combined with behavioural signals that provide temporal alignment cues analogous to ET in reading tasks. As these technologies continue to improve in signal quality while becoming less intrusive, researchers should also carefully consider the ethical implications of BCIs in IR, including the issues of privacy, consent and the potential for implicit bias in neurophysiological measurements [33].

Investigating the cognitive processes that distinguish EEG signals capable of aiding QPP from those that cannot, both within-subjects and between listening and reading induced EEG data, e.g., through ERP analysis [74], would be valuable in refining the predictive power of EEG signals. Our query-level analysis reveals that distinct patterns with fragmentary or ambiguous queries benefit substantially while very short phrases are harmed, suggesting that understanding the underlying cognitive mechanisms that produce these differential effects could guide feature engineering and model design. Furthermore, our feature importance analysis indicates that balanced utilisation across EEG frequency bands correlates with better performance, suggesting that future work should focus on feature extraction methods that ensure comprehensive representation of all frequency bands rather than allowing models to over-rely on specific frequencies. Additionally, a future systematic query-level investigation could identify, how exactly prediction errors from the specific text-based query characteristics mentioned above, might be corrected by either the EEG or ET representations of those queries. More sophisticated EEG features derived from advanced signal processing techniques [52] could enhance performance across architectures. While our results show that machine learning models consistently outperformed deep learning approaches in the generalised condition, this comparison used the same feature representations for both paradigms, leaving open the question of whether improved feature engineering might benefit one approach more than the other. In terms of the degraded performance observed in personalised models, potential avenues for improvement include increasing the amount of per-subject data to mitigate possible curse of dimensionality effects, or investigating methods for adapting generalised models to individual users with minimal per-subject data, potentially through transfer learning or few-shot adaptation techniques, to combine the robustness of cross-subject learning with the potential benefits of personalisation. Lastly, it would be beneficial for future studies to train larger models on EEG-text data to enable meaningful comparisons with state-of-the-art models [50], possibly evaluating apart from late fusion done in our research early or joint modality fusion, to evaluate the true potential of neurophysiological signals in improving QPP. These efforts could pave the way for more practical and reliable BCI applications in IR systems.

6 Conclusion

In conclusion, this study was guided by several research questions aimed at exploring the role of neurophysiological signals in QPP. These questions focused on whether EEG data could effectively predict query performance, whether multimodal models incorporating neurophysiological and behavioural signals alongside text features would outperform traditional text-only approaches, how personalised models would perform compared to generalised models, and under which conditions neurophysiological signals could provide valuable predictive information beyond traditional

text-based features. We employed a multimodal stacking approach that combines textual, EEG and ET (where available) data to predict query performance, utilising an augmented dataset tailored for IR tasks. By training separate models on each modality's features, followed by integrating their outputs through a meta-learner, we explored the effectiveness of multimodal combinations for QPP. Our experiments systematically assessed both personalised models, which account for individual differences in EEG signals, and generalised models trained on pooled EEG representations, complemented by a query-level analysis identifying the specific conditions under which neurophysiological signals provide benefit, have a negligible effect, or actively degrade prediction accuracy.

The findings reveal that neurophysiological signals can contribute to QPP under specific conditions and dataset configurations. In generalised models trained across subjects, multimodal approaches demonstrated improvements over text-only baselines in several configurations, though the magnitude and consistency of these gains varied across datasets, modalities and model types, with some configurations showing no significant improvement or degradation. ZuCo's trimodal configuration (EEG+ET+Text) achieved Pearson correlation of 0.458, significantly outperforming all other configurations, while Narrative's EEG+Text model reached 0.417, substantially exceeding text-only performance (0.324). These generalised models achieved higher correlations with lower variance compared to personalised models, demonstrating that cross-subject learning enables extraction of neurophysiological signatures of query difficulty that transcend individual differences or personalised modelling complexities. Feature importance analysis revealed that successful integration requires cross-spectrum utilisation across EEG frequency bands, with the best-performing models showing uniform importance distributions rather than over-reliance on specific frequencies. Tree-based models consistently outperformed deep learning approaches in the generalised condition, suggesting that effective multimodal feature integration may be more critical than architectural complexity for neurophysiological QPP. The exploratory query-level analysis suggested potential boundary conditions: as shown in Tables 3 and 4, neurophysiological signals tend to improve predictions for fragmentary or semantically ambiguous queries where text-only models exhibit high error, while introducing noise for very short phrases where text-only predictions are already accurate, with the majority of queries showing neutral effects across all datasets. A complete characterisation of which individual queries benefit from EEG enhancement remains an important direction for future work. For reading tasks with ET, signal quality critically moderates effectiveness, with high SNR (>10 dB) nearly doubling the proportion of helped queries compared to medium-quality signals, while listening tasks showed no SNR dependency. These results indicate both the feasibility of EEG-enhanced QPP and guide possible design principles and boundary conditions necessary for effective EEG-enhanced IR systems.

References

[1] Zahra Abbasiantaeb, Chuan Meng, Leif Azzopardi, and Mohammad Aliannejadi. 2024. Can we use large language models to fill relevance judgment holes? arXiv:2405.05600. Retrieved from https://arxiv.org/abs/2405.05600

[2] Hezam Albaqami, Ghulam Mubashar Hassan, Abdulhamit Subasi, and Amitava Datta. 2021. Automatic detection of abnormal EEG signals using wavelet feature extraction and gradient boosting decision tree. *Biomedical Signal Processing and Control* 70 (2021), 102957.

[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] Negar Arabzadeh, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2021. Query performance prediction through retrieval coherency. In *Advances in Information Retrieval: Proceedings of the 43rd European Conference on IR Research (ECIR '21)*, Part II. Springer, 193–200.

[5] Negar Arabzadeh, Maryam Khodabakhsh, and Ebrahim Bagheri. 2021. BERT-QPP: Contextualized pre-trained transformers for query performance prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2857–2861.

[6] Negar Arabzadeh, Fattane Zarrinkalam, Jelena Jovanovic, Feras Al-Obeidat, and Ebrahim Bagheri. 2020. Neural embedding-based specificity metrics for pre-retrieval query performance prediction. *Information Processing & Management* 57, 4 (2020), 102248.

[7] Javed A. Aslam and Virgil Pavlu. 2007. Query hardness estimation using Jensen-Shannon divergence among multiple scoring functions. In *Proceedings of the European Conference on Information Retrieval*. Springer, 198–209.

[8] Shariq Bashir. 2014. Combining pre-retrieval query quality predictors using genetic programming. *Applied Intelligence* 40 (2014), 525–535.

[9] Nicholas J. Belkin. 1980. Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science* 5, 1 (1980), 133–143.

[10] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The Alice datasets: FMRI & EEG observations of natural language comprehension. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, et al. (Eds.), European Language Resources Association, 120–125. Retrieved from https://aclanthology.org/2020.lrec-1.15/

[11] Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. 2015. The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics* 9 (June 2015), 16. DOI: https://doi.org/10.3389/fninf.2015.00016

[12] Jordan J. Bird, Michael Pritchard, Antonio Fratini, Anikó Ekárt, and Diego R. Faria. 2021. Synthetic biological signals machine-generated by GPT-2 improve the classification of EEG and EMG through data augmentation. *IEEE Robotics and Automation Letters* 6, 2 (April 2021), 3498–3504. DOI: https://doi.org/10.1109/LRA.2021.3056355

[13] Michael P. Broderick, Andrew J. Anderson, Giovanni M. Di Liberto, Michael J. Crosse, and Edmund C. Lalor. 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology: CB* 28, 5 (March 2018), 803–809.e3. DOI: https://doi.org/10.1016/j.cub.2018.01.080

[14] Boris Burle, Laure Spieser, Clémence Roger, Laurence Casini, Thierry Hasbroucq, and Franck Vidal. 2015. Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology* 97, 3 (2015), 210–220.

[15] David Carmel and Elad Yom-Tov. 2010. *Estimating the Query Difficulty for Information Retrieval*. Morgan & Claypool Publishers.

[16] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. 2006. What makes a query difficult? In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 390–397.

[17] Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. arXiv:2002.03932. Retrieved from https://arxiv.org/abs/2002.03932

[18] Xiaoyang Chen, Ben He, and Le Sun. 2022. Groupwise query performance prediction with BERT. In *Proceedings of the European Conference on Information Retrieval*. Springer, 64–74.

[19] Xun Chen, Xueyuan Xu, Aiping Liu, Soojin Lee, Xiang Chen, Xu Zhang, Martin J. McKeown, and Z. Jane Wang. 2019. Removal of muscle artifacts from the EEG: A review and recommendations. *IEEE Sensors Journal* 19, 14 (2019), 5353–5368.

[20] 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 15th ACM International Conference on Web Search and Data Mining*, 1569–1572.

[21] Adrian-Gabriel Chifu, Léa Laporte, Josiane Mothe, and Md Zia Ullah. 2018. Query performance prediction focused on summarized Letor features. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1177–1180.

[22] Giulia Cisotto and Davide Chicco. 2024. Ten quick tips for clinical electroencephalographic (EEG) data acquisition and signal processing. *PeerJ Computer Science* 10 (Sept. 2024), e2256. DOI: https://doi.org/10.7717/peerj-cs.2256

[23] 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.

[24] J. Shane Culpepper, Charles L. A. Clarke, and Jimmy Lin. 2016. Dynamic cutoff prediction in multi-stage retrieval systems. In *Proceedings of the 21st Australasian Document Computing Symposium*, 17–24.

[25] Ronan Cummins. 2014. Document score distribution models for query performance inference and prediction. *ACM Transactions on Information Systems* 32, 1 (2014), 1–28.

[26] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. arXiv:2209.11755. Retrieved from https://arxiv.org/abs/2209.11755

[27] Suchana Datta, Debasis Ganguly, Derek Greene, and Mandar Mitra. 2022. Deep-QPP: A pairwise interaction-based deep learning model for supervised query performance prediction. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*, 201–209.

[28] Ehsan Eqlimi, Annelies Bockstael, Bert De Coensel, Marc Schönwiesner, Durk Talsma, and Dick Botteldooren. 2020. EEG correlates of learning from speech presented in environmental noise. *Frontiers in Psychology* 11 (2020), 1850.

[29] 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.

[30] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. arXiv:2212.10496. Retrieved from https://arxiv.org/abs/2212.10496

[31] Himaanshu Gauba, Pradeep Kumar, Partha Pratim Roy, Priyanka Singh, Debi Prosad Dogra, and Balasubramanian Raman. 2017. Prediction of advertisement preference by fusing EEG response and sentiment analysis. *Neural Networks: The Official Journal of the International Neural Network Society* 92 (Aug. 2017), 77–88. DOI: https://doi.org/10.1016/j.neunet.2017.01.013

[32] Anne-Lise Giraud and David Poeppel. 2012. Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience* 15, 4 (2012), 511–517.

[33] Gerd Grübler, Abdul Al-Khodairy, Robert Leeb, Iolanda Pisotta, Angela Riccio, Martin Rohm, and Elisabeth Hildt. 2014. Psychosocial and ethical aspects in non-invasive EEG-based BCI research—A survey among BCI users and BCI professionals. *Neuroethics* 7, 1 (April 2014), 29–41. DOI: https://doi.org/10.1007/s12152-013-9179-7

[34] 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.

[35] Claudia Hauff, Leif Azzopardi, and Djoerd Hiemstra. 2009. The combination and evaluation of query performance prediction methods. In *Proceedings of the European Conference on Information Retrieval*. Springer, 301–312.

[36] Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 1419–1420.

[37] Ben He and Iadh Ounis. 2004. Inferring query performance using pre-retrieval predictors. In *Proceedings of the International Symposium on String Processing and Information Retrieval*. Springer, 43–54.

[38] Ben He and Iadh Ounis. 2006. Query performance prediction. *Information Systems* 31, 7 (2006), 585–594.

[39] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data* 5, 11 (Dec. 2018), 180291. DOI: https://doi.org/10.1038/sdata.2018.291

[40] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 138–146.

[41] Alexandre Hyafil, Lorenzo Fontolan, Claire Kabdebon, Boris Gutkin, and Anne-Lise Giraud. 2015. Speech encoding by coupled cortical theta and gamma oscillations. *eLife* 4 (2015), e06213.

[42] Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wenzel, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, et al. 2019. Integrating neurophysiologic relevance feedback in intent modeling for information retrieval. *Journal of the Association for Information Science and Technology* 70, 9 (2019), 917–930.

[43] Parastoo Jafarzadeh and Faezeh Ensan. 2022. A semantic approach to post-retrieval query performance prediction. *Information Processing & Management* 59, 1 (2022), 102746.

[44] Nour Jedidi, Yung-Sung Chuang, Leslie Shing, and James Glass. 2024. Zero-shot dense retrieval with embeddings from relevance feedback. arXiv:2410.21242. Retrieved from https://arxiv.org/abs/2410.21242

[45] Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars-v2: Large language models as efficient dataset generators for information retrieval. arXiv:2301.01820. Retrieved from https://arxiv.org/abs/2301.01820

[46] Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing information seeking processes with multiple physiological signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1006–1017.

[47] Qianghuai Jia and Xinmei Tian. 2015. Query difficulty estimation via relevance prediction for image retrieval. *Signal Processing* 110 (2015), 232–243.

[48] Gilad Katz, Anna Shtock, Oren Kurland, Bracha Shapira, and Lior Rokach. 2014. Wikipedia-based query performance prediction. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1235–1238.

[49] Jukka-Pekka Kauppi, Melih Kandemir, Veli-Matti Saarinen, Lotta Hirvenkari, Lauri Parkkonen, Arto Klami, Riitta Hari, and Samuel Kaski. 2015. Towards brain-activity-controlled information retrieval: Decoding image relevance from MEG signals. *NeuroImage* 112 (2015), 288–298.

[50] Demetres Kostas, Stephane Aroca-Ouellette, and Frank Rudzicz. 2021. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience* 15 (2021), 653659.

[51] Shailaja Kotte and J. R. K. Kumar Dabbakuti. 2020. Methods for removal of artifacts from EEG signal: A review. *Journal of Physics: Conference Series* 1706 (2020), 012093.

[52] Daria La Rocca, Patrizio Campisi, Balazs Vegso, Peter Cserti, György Kozmann, Fabio Babiloni, and F. De Vico Fallani. 2014. Human brain distinctiveness based on EEG spectral coherence connectivity. *IEEE Transactions on Bio-Medical Engineering* 61, 9 (2014), 2406–2412.

[53] John LaRocco, Minh Dong Le, and Dong-Guk Paeng. 2020. A systemic review of available low-cost EEG headsets used for drowsiness detection. *Frontiers in Neuroinformatics* 14 (2020), 553352.

[54] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. arXiv:1906.00300. Retrieved from https://arxiv.org/abs/1906.00300

[55] Seung-Bo Lee, Yong-Jeong Kim, Sungeun Hwang, Hyoshin Son, Sang Kun Lee, Kyung-Il Park, and Young-Gon Kim. 2022. Predicting Parkinson's disease using gradient boosting decision tree models with electroencephalography signals. *Parkinsonism & Related Disorders* 95 (2022), 77–85.

[56] Ashley G. Lewis, Jan-Mathijs Schoffelen, Herbert Schriefers, and Marcel Bastiaansen. 2016. A predictive coding perspective on beta oscillations during sentence-level language comprehension. *Frontiers in Human Neuroscience* 10 (2016), 85.

[57] Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. 2025. BrainECHO: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of the Association for Computational Linguistics (ACL '25)*. Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.), Association for Computational Linguistics, 2762–2778. DOI: https://doi.org/10.18653/v1/2025.findings-acl.142

[58] Niall McGuire and Yashar Moshfeghi. 2024. Prediction of the realisation of an information need: An EEG study. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2584–2588.

[59] Niall Mcguire and Yashar Moshfeghi. 2025. Towards brain passage retrieval: An investigation of EEG query representations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*. ACM, New York, NY, 613–623. DOI: https://doi.org/10.1145/3726302.3730097

[60] Stefano Mizzaro, Josiane Mothe, Kevin Roitero, and Md Zia Ullah. 2018. Query performance prediction and effectiveness evaluation without relevance judgments: Two sides of the same coin. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 1233–1236.

[61] Saeedeh Momtazi and Dietrich Klakow. 2015. Bridging the vocabulary gap between questions and answer sentences. *Information Processing & Management* 51, 5 (2015), 595–615.

[62] Yashar Moshfeghi, Luisa R. Pinto, Frank E. Pollick, and Joemon M. Jose. 2013. Understanding relevance: An fMRI study. In *Advances in Information Retrieval: Proceedings of the 35th European Conference on IR Research (ECIR '13)*. Springer, 14–25.

[63] Yashar Moshfeghi and Frank E. Pollick. 2018. Search process as transitions between neural states. In *Proceedings of the 2018 World Wide Web Conference*, 1683–1692.

[64] Yashar Moshfeghi, Peter Triantafillou, and Frank Pollick. 2019. Towards predicting a realisation of an information need based on brain signals. In *Proceedings of the World Wide Web Conference*, 1300–1309.

[65] 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.

[66] Phantharach Natnithikarat, Theerawit Wilaiprasitporn, and Supavit Kongwudhikunakorn. 2023. Combining EEG and NLP features for predicting students' lecture comprehension using ensemble classification. In *Proceedings of the 2023 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 218–223. DOI: https://doi.org/10.1109/RIVF60135.2023.10471794

[67] Liqiang Nie, Meng Wang, Zheng-Jun Zha, and Tat-Seng Chua. 2012. Oracle in image search: A content-based approach to performance prediction. *ACM Transactions on Information Systems* 30, 2 (2012), 1–23.

[68] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. arXiv:1901.04085. Retrieved from https://arxiv.org/abs/1901.04085

[69] Jonas Obleser and Nathan Weisz. 2012. Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex (New York, N.Y.: 1991)* 22, 11 (2012), 2466–2477.

[70] Sakrapee Paisalnan, Frank Pollick, and Yashar Moshfeghi. 2021. Towards understanding neuroscience of realisation of information need in light of relevance and satisfaction judgement. In *Proceedings of the International Conference on Machine Learning, Optimization, and Data Science*. Springer, 41–56.
[71] Sakrapee Paisalnan, Frank Pollick, and Yashar Moshfeghi. 2022. Neural correlates of satisfaction of an information need. In *Proceedings of the International Conference on Machine Learning, Optimization, and Data Science*. Springer, 443–457.
[72] Daniel Carlos Guimarães Pedronette and Ricardo da S. Torres. 2015. Unsupervised effectiveness estimation for image retrieval using reciprocal rank information. In *Proceedings of the 2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 321–328.
[73] 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.
[74] Zuzana Pinkosova, William J. McGeown, and Yashar Moshfeghi. 2022. Revisiting neurological aspects of relevance: An EEG study. In *Proceedings of International Conference on Machine Learning, Optimization, and Data Science*. Springer, 549–563.
[75] Eduard Poesina, Radu Tudor Ionescu, and Josiane Mothe. 2023. IQPP: A benchmark for image query performance prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2953–2963.
[76] Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! arXiv:2312.02724. Retrieved from https://arxiv.org/abs/2312.02724
[77] Fiana Raiber and Oren Kurland. 2014. Query-performance prediction: Setting the expectations straight. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, 13–22.
[78] Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.
[79] Haggai Roitman, Shai Erera, and Bar Weiner. 2017. Robust standard deviation estimation for query performance prediction. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, 245–248.
[80] Haggai Roitman and Oren Kurland. 2019. Query performance prediction for pseudo-feedback-based retrieval. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1261–1264.
[81] Dwaipayan Roy, Debasis Ganguly, Mandar Mitra, and Gareth J. F. Jones. 2019. Estimating Gaussian mixture models in the local neighbourhood of embedded word vectors for query performance prediction. *Information Processing & Management* 56, 3 (2019), 1026–1045.
[82] Simanto Saha and Mathias Baumert. 2020. Intra-and inter-subject variability in EEG-based sensorimotor brain computer interface: A review. *Frontiers in Computational Neuroscience* 13 (2020), 87.
[83] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 595–604.
[84] Anna Shtok, Oren Kurland, and David Carmel. 2016. Query performance prediction using reference lists. *ACM Transactions on Information Systems* 34, 4 (2016), 1–34.
[85] Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. 2007. Identifying ambiguous queries in web search. In *Proceedings of the 16th International Conference on World Wide Web*, 1169–1170.
[86] Antje Strauß, Malte Wöstmann, and Jonas Obleser. 2014. Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience* 8 (2014), 350.
[87] Shaoyan Sun, Wengang Zhou, Qi Tian, Ming Yang, and Houqiang Li. 2018. Assessing image retrieval quality at the first glance. *IEEE Transactions on Image Processing* 27, 12 (2018), 6124–6134.
[88] 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:2104.08663. Retrieved from https://arxiv.org/abs/2104.08663.
[89] Ryan Theisen, Hyunsuk Kim, Yaoqing Yang, Liam Hodgkinson, and Michael W. Mahoney. 2023. When are ensembles really effective? In *Advances in Neural Information Processing Systems*, Vol. 36, 15015–15026.
[90] Sergios Theodoridis and Konstantinos Koutroumbas. 2006. *Pattern Recognition*. Elsevier.
[91] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288. Retrieved from https://arxiv.org/abs/2307.09288
[92] Lucas Pascotti Valem and Daniel Carlos Guimarães Pedronette. 2021. A denoising convolutional neural network for self-supervised rank effectiveness estimation on image retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 294–302.

[93] Vishwa Vinay, Ingemar J. Cox, Natasa Milic-Frayling, and Ken Wood. 2006. On ranking the effectiveness of searches. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 398–404.

[94] Ellen M. Voorhees. 2001. The philosophy of information retrieval evaluation. In *Proceedings of the Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 355–370.

[95] Ryen W. White. 2016. *Interactions with Search Systems*. Cambridge University Press.

[96] Dezhong Yao, Yun Qin, Shiang Hu, Li Dong, Maria L. Bringas Vega, and Pedro A. Valdés Sosa. 2019. Which reference should we use for EEG and ERP practice? *Brain Topography* 32, 4 (July 2019), 530–549. DOI: https://doi.org/10.1007/s10548-019-00707-x

[97] Ziyi Ye, Jingtao Zhan, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Christina Lioma, and Tuukka Ruotsalo. 2024. Query augmentation with brain signals. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7561–7570.

[98] Hamed Zamani, W. Bruce Croft, and J. Shane Culpepper. 2018. Neural query performance prediction using weak supervision from multiple signals. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 105–114.

[99] Zhongmin Zhang, Jiawei Chen, and Shengli Wu. 2018. Query performance prediction and classification for information search systems. In *Proceedings of the 2nd International Joint Conference on Web and Big Data (APWeb-WAIM '18)*, Part I. Springer, 277–285.

[100] Ying Zhao, Falk Scholer, and Yohannes Tsegay. 2008. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval: Proceedings of the 30th European Conference on IR Research (ECIR '08)*. Springer, 52–64.

[101] Yun Zhou and W. Bruce Croft. 2006. Ranking robustness: A novel framework to predict query performance. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 567–574.

[102] Yun Zhou and W. Bruce Croft. 2007. Query performance prediction in web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 543–550.

[103] Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and G. Zuccon. 2024. PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Retrieved from https://api.semanticscholar.org/CorpusID:269449435

Received 16 May 2025; revised 13 March 2026; accepted 26 April 2026