Beyond One-Size-Fits-All: A Study of Neural and Behavioural Variability Across Different Recommendation Categories

Georgios Koutroumpas
Telefónica Scientific Research
Barcelona, Spain
georgios.koutroumpas@upc.edu

Sebastian Idesis
Telefónica Scientific Research
Barcelona, Spain
sebastianariel.idesis@telefonica.com

Mireia Masias Bruns
Telefónica Scientific Research
Barcelona, Spain
mireia.masiasbruns@telefonica.com

Carlos Segura
Telefónica Scientific Research
Barcelona, Spain
carlos.segura@telefonica.com

Joemon M. Jose
University of Glasgow
Glasgow, United Kingdom
joemon.jose@glasgow.ac.uk

Sergi Abadal
Universitat Politècnica de Catalunya
Barcelona, Spain
abadal@ac.upc.edu

Ioannis Arapakis
Telefónica Scientific Research
Barcelona, Spain
ioannis.arapakis@telefonica.com

Abstract

Traditionally, Recommender Systems (RS) have primarily measured performance based on the accuracy and relevance of their recommendations. However, this algorithmic-centric approach overlooks how different types of recommendations impact user engagement and shape the overall quality of experience. In this paper, we shift the focus to the user and address for the first time the challenge of decoding the neural and behavioural variability across *distinct* recommendation categories, considering more than just relevance. Specifically, we conducted a controlled study using a comprehensive e-commerce dataset containing various recommendation types, and collected Electroencephalography and behavioural data. We analysed both neural and behavioural responses to recommendations that were categorised as **E**xact, **S**ubstitute, **C**omplement, or **I**rrelevant products within search query results. Our findings offer novel insights into user preferences and decision-making processes, revealing meaningful relationships between behavioural and neural patterns for each category, but also indicate inter-subject variability.

CCS Concepts

• **Information systems → Users and interactive retrieval**; Recommender systems; • **Computing methodologies** → *Cognitive science*; • **Human-centered computing → User studies**.

Keywords

Recommender Systems; Electroencephalography; Behavioural Analysis; User Study; E-commerce

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICTIR '25, Padua, Italy
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744616

1 Introduction

The rapid growth of e-commerce has fundamentally changed how users discover and interact with products and services online. While this digital transformation has created more revenue opportunities for service providers, it has also introduced significant challenges for users who must navigate an overwhelming number of choices. Specifically, when presented with large array situations (e.g., limitless products to purchase from), users are at higher risk of experiencing choice overload [28], which can degrade their quality of experience with an online service or platform.

Recommender Systems (RS) have been shown to alleviate this "paradox of choice" [58] by facilitating access to relevant content and improving the browsing experience [25, 73]. In settings where the abundance of options can result in unsatisfying choices or abandonment, the user experience is ultimately determined by the RS capacity to filter irrelevant content and recommend items regarded as desirable. While significant advances have been made in improving recommendation accuracy [15, 18, 50], these approaches often optimise mainstream metrics at the expense of other content-derived qualitative aspects, such as diversity and novelty of recommendations [4, 13, 18, 50, 60]. More importantly, current approaches primarily focus on algorithmic performance, potentially overlooking the complex cognitive and behavioural responses that different recommendation categories may elicit from users.

Previous research has explored neural correlates of relevance detection [17, 54] and search satisfaction [44, 54], mainly in binary relevance tasks involving search terms and documents. In contrast, we explore how users process and respond to *different* recommendation categories, by leveraging a real-life, e-commerce query-product dataset ("ESCI" [56]) , where the relationship between user intent and recommended products is more nuanced than simple relevance. Our research addresses the following questions:

- **RQ1:** Do different recommendation categories evoke differentiable neural signatures in brain activity as measured through Electroencephalography (EEG), and can these patterns be reliably identified?

- **RQ2:** How do different recommendation categories influence users' perception (and subsequent observable behaviour) regarding item relevance, purchaseability, and recommendation diversity?
- **RQ3:** Can we identify neurophysiological markers of engagement (e.g., Frontal Alpha Asymmetry (FAA) [9]) that systematically vary across recommendation categories, and what do these variations reveal about the effectiveness of different recommendation strategies?

Specifically, by combining EEG measurements with behavioural data, our study offers a fresh perspective on recommendation effectiveness that goes beyond traditional accuracy metrics. This approach allows us to examine whether different recommendation categories evoke distinct neural signatures, how they influence user perception of relevance and purchase intent, and their impact on overall engagement. Our findings could help inform the development of more sophisticated RS in the future so that they better align with users' intent and true consumer needs.

In summary, our **contributions** are the following:

- We present the first comprehensive study examining neural and behavioural responses across different recommendation categories, leveraging a real-life, e-commerce dataset, and reveal distinct patterns of brain activity for **E**xact, **S**ubstitute, **C**omplement, or **I**rrelevant product recommendations.
- We introduce novel methodological approaches for analysing the relationship between recommendation types and user cognitive processes, combining EEG measurements with behavioural metrics in a controlled experimental setting.
- We provide empirical evidence demonstrating how different recommendation categories influence user perception of relevance (and subsequent observable behaviour), with implications for balancing accuracy and diversity in RS.
- We analyse neural markers of engagement (e.g. FAA, increased theta/beta power) and their association to different recommendation types.

2 Related Work

Despite progress, key challenges remain in RS, including the cold-start problem [13, 18, 50], bias in sparse ratings [18, 32, 50], and limitations of accuracy-based evaluation [18, 50, 60]. While Deep Learning (DL) and Large Language Models (LLMs) offer potential [15], they require significant resources and still struggle with concepts like diversity and novelty. Understanding user cognition and integrating neural signals into RS is a promising direction [7, 14, 37, 54, 63], but little research has explored how different recommendation types shape neural and behavioural responses. In what follows, we review explicit, implicit, and neuroscience-informed feedback methods for relevance prediction.

2.1 Explicit Feedback Methods

Explicit feedback involves direct user input, such as ratings, reviews, or preference selections. While widely studied in RS research, it has key limitations [47]. A main challenge is the need for active user participation, which many users avoid [26], resulting in data sparsity that hampers recommender system performance [2]. Moreover, explicit feedback often reflects biases and reliability issues, as users may struggle to express preferences, especially for unconscious or hard-to-quantify experiences [3, 53]. These drawbacks have motivated the investigation of alternative methods for capturing user preferences [33].

2.2 Implicit Feedback Methods

Implicit feedback refers to user interactions passively collected without direct input, offering an alternative to explicit ratings in RS [31]. Such signals include clicks, dwell time, scrolling, and purchase history [68]. Its main advantage is non-intrusiveness, reducing user effort and cognitive load [30]. However, these signals can be ambiguous; interactions may not reliably indicate true interest, introducing noise [33]. For example, long dwell time could reflect either engagement or confusion [12]. To overcome these issues, researchers are investigating neurophysiological measures like EEG to directly assess cognitive states and engagement.

2.3 Neuroimaging Methods for Relevance Prediction

Traditional RS often overlook the subjective emotional and cognitive responses that are crucial for understanding user engagement and satisfaction [37]. Recent studies have leveraged EEG to investigate the neural underpinnings of relevance judgments. For example, Pinkosova et al. [54] recorded EEG data during a binary relevance assessment task, demonstrating that neural signals can provide valuable insights into how users perceive relevant content. Similarly, Ye et al. [72] compared brain signals of users examining non-clicked search results of varying usefulness, providing insights into the neural correlates of relevance judgments.

Expanding on this work, Jacucci et al. [29] developed a fully integrated Information Retrieval (IR) system that utilizes online implicit relevance feedback based on brain activity and eye movements, enabling real-time adaptation of search results. In a similar vein, Ruotsalo et al. [57] proposed decoding emotional responses from EEG as an alternative dimension of relevance, highlighting the role of affect in user preference modelling. More recently, Ye et al. [71] reviewed the application of Brain-Computer Interface (BCI) in IR research, identifying several opportunities for integrating active or passive BCIs into RS to provide real-time insights into user cognitive states, thereby improving the relevance prediction.

These studies demonstrate that neuroimaging techniques, particularly EEG, offer promising new avenues for relevance prediction by directly tapping into user cognition, emotion, and engagement. Integrating neurophysiological signals into RS models could enable systems to move beyond conventional behavioural tracking toward a more comprehensive understanding of the user experience and inform the design of RS algorithms, particularly those incorporating multi-objective criteria.

3 Methodology

Our experimental methodology is based on the acquisition of EEG signals from participants as they evaluate query-product pairs. Next, we describe the neurophysiological experiment, including participant recruitment and experimental design. Then, we outline the data curation and pre-processing steps. Finally, we detail the signal processing techniques and methods used to extract relevant neural and behavioural features in response to different recommendations.

3.1 Design

The study used a within-subjects design with one independent variables: (1) recommendation category (with four levels: **E**xact, **S**ubstitute, **C**omplement, or **I**rrelevant). The dependent variables were (1) user neural activity, as registered by the EEG and (2) user behaviour, as measured via self-reports.

3.2 Participants

Twenty-one participants enrolled in this study (8 females; 3 left-handed), aged 18 to 59 years ($M = 28.02$, $SD = 8.62$) and of mixed nationality. The participants were recruited via Telefónica's Testers platform and mailing lists, and provided written consent prior to the experiment. All participants had normal or corrected-to-normal vision, did not suffer from any type of disability, nor sensory or neurological difficulty. Moreover, participants were not under any kind of neurological medication or treatment that could interfere with the study (e.g., antiepileptics). Participants self-reported at least a C1 level of English proficiency and prior experience with e-commerce platforms (e.g., Amazon). Upon completion of the study, they were compensated with 40 EUR in Amazon vouchers.

3.3 Apparatus

3.3.1 Equipment. For the study, stimuli were presented on a 27-inch monitor (23.53 × 13.24 inches; 1920 × 1080 pixels; 60 Hz), positioned 60 cm from participants. Responses were collected via a standard mouse and keyboard. Stimulus delivery, hardware synchronisation, and timing optimisation were controlled using PsychoPy[1] [51]). EEG data were recorded with an actiCHamp Plus system (Brain Products GmbH, Germany) using 32 active Ag/AgCl electrodes arranged according to the 10/20 system. Signals were sampled at 256 Hz, referenced online to Cz, and grounded at AFz. Electrode impedances were kept below 10 kΩ and monitored via BrainVision Recorder[2] (Brain Products GmbH, Germany) to ensure reliable Signal-to-Noise Ratio (SNR). Precise synchronization of stimulus events and EEG data was achieved by transmitting millisecond-accurate hardware triggers from PsychoPy to BrainVision Recorder via a parallel port.

3.3.2 Dataset. We used the Amazon dataset [45], a widely adopted real-world e-commerce corpus in RS research [5, 6, 22, 34–36, 55], containing product details such as image URLs, titles, prices, and categories. Additionally, we incorporated the Shopping Queries dataset [56], comprising 130k unique queries and 2.6 million manually labeled query-product relevance judgments. Each query lists up to 40 results with ESCI labels (**E**xact, **S**ubstitute, **C**omplement, or **I**rrelevant) [43], along with product and query metadata (IDs, titles, descriptions, texts, labels). This multilingual dataset includes English, Japanese, and Spanish queries. After joining both datasets via product IDs and removing inconsistencies or incomplete entries, we excluded non-English queries with fewer than two **E**xact products. The final dataset (Table 1) included product identifiers, titles, descriptions, prices, images, queries, and recommendation types.

Table 1: Dataset statistics on Joint ESCI/Amazon Review Datasets, including item counts and percentage.

Dataset Content	Joint Dataset	Used in Study
Dataset length in rows	63,261	974
Unique queries	13,214	120
Unique items	41,781	353
ESCI Label Distribution		
Exact	52,104 (82.36%)	263 (74.50%)
Substitute	7,519 (11.89%)	30 (8.5%)
Irrelevant	2,507 (3.96%)	30 (8.5%)
Complement	1,131 (1.79%)	30 (8.5%)
Top 12 Main Categories by Unique Items		
Amazon Home	10,005 (23.95%)	69 (19.55%)
Tools & Home Improvement	4,811 (11.51%)	33 (9.35%)
Toys & Games	4,437 (10.62%)	57 (16.15%)
Sports & Outdoors	4,138 (9.90%)	61 (17.28%)
Grocery	3,844 (9.20%)	27 (7.65%)
Automotive	3,807 (9.11%)	7 (1.98%)
Amazon Fashion	2,789 (6.68%)	23 (6.52%)
Office Products	2,571 (6.15%)	42 (11.90%)
Pet Supplies	2,292 (5.49%)	18 (5.10%)
Industrial & Scientific	1,075 (2.57%)	10 (2.83%)
Arts, Crafts & Sewing	1,059 (2.53%)	2 (0.57%)
Computers	953 (2.28%)	4 (1.13%)

3.3.3 Ground-truth Labels. To determine the recommendation category of each product with respect to a query, we applied the ESCI relevance judgements, as specified by McAuley et al. [43]:

- **E**xact: the item is relevant for the query, and satisfies all the query specifications (e.g., a water bottle matching all attributes of a query "plastic water bottle 24oz", such as material and size).
- **S**ubstitute: the item is somewhat relevant, i.e., it fails to fulfil some aspects of the query but the item can be used as a functional substitute (e.g., fleece for a "sweater" query)
- **C**omplement: the item does not fulfill the query, but could be used in combination with an exact item (e.g., track pants for "running shoes" query)
- **I**rrelevant: the item is irrelevant, or it fails to fulfill a central aspect of the query (e.g., socks for a "telescope" query, or a wheat flour bread for a "gluten–free bread" query)

3.3.4 Stimuli. From the joint dataset (Table 1: "Used in Study"), we sampled 120 queries in a uniform way across product categories (e.g., books, sports, electronics, clothing, etc.) and paired them with 120 product recommendations, 30 per recommendation category (e.g., **E**xact, **S**ubstitute). The use of search queries allows us to capture user preferences and simulate interaction history—signals commonly used in RS. This information can subsequently inform the ranking of items in a recommendation setting, thereby aligning our experimental findings with real-world applications.

Between the query and the product recommendation, we introduced a sequence of $k \in \{1, 2, 3\}$ **E**xact products, thus varying the position of the recommended product between second, third, or fourth in the sequence. This afforded a more ecologically valid

[1] https://www.psychopy.org/
[2] https://www.brainproducts.com/

(a) Query (b) Exact Product (c) Substitute Product (d) Complement Product (e) Irrelevant Product

Figure 1: Stimuli examples: (a) Query "Avengers night light"; (b) Exact recommendation; (c) Substitute recommendation (Marvel night light); (d) Complementary recommendation (light bulb that can be used with a night light); (e) Irrelevant recommendation.

setting—similar to the oddball paradigm [62]—where the participant is not conditioned to expect the recommendation at a fixed position. Each product description was shortened to 10–15 words using Meta's Llama 3 model via the LM Studio API. A custom prompt was used to *consistently* instruct the model to retain only the most useful and descriptive information relevant to the query, removing unnecessary details (e.g., product codes). We further used the EB Garamond font to optimise for reading speed [66] and reserve more time for cognitive processing of the stimuli. When more than one product image was available, we opted for the most representative using a majority voting based on three human annotators. For the stimuli design we applied a template that resembled a simplified version of Amazon e-commerce platform, showing the product image on the left side of the visual box and the product description and price to the top right (Figure 1). A subtle "Recommended" logo was added for recommended products, to suggest their condition.

3.4 Procedure

Prior to the experiment, participants were asked to complete an entry questionnaire that collected information related to exclusion criteria (e.g., vision issues or neurological difficulties) and prior experience with e-commerce platforms. Then, a short tutorial and Q/A session was held that covered the study in detail, during which participants were informed about their right to withdraw from the experiment at any moment without facing any adverse consequences. Upon signing a consent form, a preparation of the equipment was made (i.e. fitting the EEG cap, applying the gel to the electrodes, checking the impedance level for each channel).

Our experiment consisted of 120 trials that implemented the oddball paradigm [62]. In each trial, participants were presented with a search query followed by a sequence of one to three **E**xact products for context, followed by a product recommendation that belonged to one of four recommendation categories (Figure 3). The presentation duration was set to 3 seconds for context products and 4 seconds for the recommendation. At the end of each trial, participants had to self-label the recommended product and rate it according to the following dimensions: (1) Relevance, (2) Likelihood of purchase, and (3) Diversity. Ratings were assessed using a 5-point Likert scale (higher values indicate a stronger agreement).

To control for order effects, all participants viewed the same set of samples in a randomized order, following a Latin-squares design. Also, both query reading and product ratings were self-paced (to ensure task comprehension), and there was no time limit for completing each trial. Considering EEG equipment constraints and potential participant fatigue, we limited the experiment duration

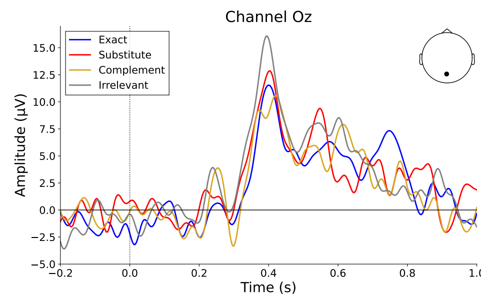


Figure 2: Example Event Related Potential (ERP) showing average neural responses to four recommendation categories

to 90-120 min and incorporated breaks at the one-third and two-thirds marks. Upon completing the study, participants were asked to complete an exit questionnaire that inquired about demographics.

3.5 Behavioural analysis

To assess participants' decision-making and product categorization accuracy (i.e., how well self-labels matched the original dataset labels) we performed a thorough analysis of behavioural data, focusing on classification accuracy, relevance ratings, purchase likelihood, and diversity. Additionally, we investigated the statistical relationships between these measures to understand behavioural patterns in the task. Specifically, we conducted multiple statistical tests: the Shapiro-Wilk test was used to assess normality, Levene's test examined homogeneity of variance, while classification accuracy was determined using Chi-square tests. Depending on the aforementioned tests, the Friedman test was applied for comparing multiple related conditions, and the Wilcoxon signed-rank test was used for pairwise comparisons of behavioural measures

3.6 EEG pre-processing

We excluded two participants (out of 21) from subsequent analyses due to poor task comprehension and low SNR. EEG data were processed using the MNE-Python library [20]. We extracted 4.5-second epochs around each event, including a 0.5-second pre-stimulus baseline for correction, per trial. Data were band-pass filtered between 0.5-20 Hz, as our initial spectral analysis showed no stimulus-specific information above 20 Hz. Artifact rejection was performed using FASTER [46], an automated de-noising pipeline that handles channel rejection, epoch removal, Independent Component Analysis (ICA), and channel interpolation. The outcome of this preprocessing is the Single-trial ERPs computed per epoch (Figure 2).

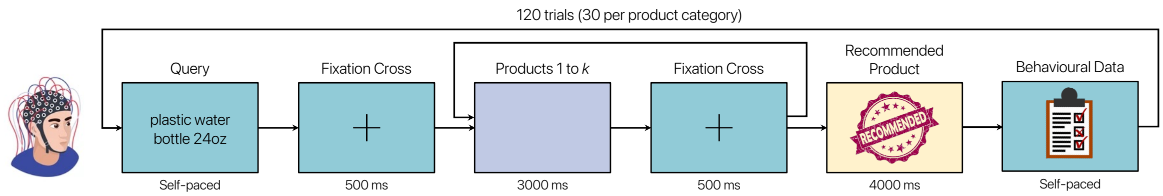


Figure 3: Overview of single-trial experimental protocol. Each participant performs 120 trials. In each trial, the participant is shown a query, followed by a sequence of $k \in \{1, 2, 3\}$ exact products, and concluded by a product recommendation that can be any of the four categories (Exact, Substitute, Complement, or Irrelevant).

3.7 EEG analysis

In what follows, we present a comprehensive analysis of the EEG data through three complementary approaches. First, we conduct within-subjects classification to examine individual-specific neural patterns and their variability across recommendation categories. Second, we perform between-subjects analysis to identify common neural signatures and evaluate the generalisability of our findings across participants. Finally, we analyse established neurophysiological markers of engagement to quantify how different recommendation categories modulate user responses.

3.7.1 Within-subjects classification. We investigated whether distinct neural patterns emerge across recommendation categories at the participant level. In line with prior work [17, 40, 59, 61, 67, 75], we trained a Support Vector Machine (SVM) classifier to perform pairwise classification between recommendation categories, using EEG features supported by prior research [10, 11, 21, 64]. Specifically, the classifier's input consisted of several features derived from each epoch: (1) Single-trial ERPs, representing the denoised temporal response; (2) Power Spectral Density (PSD) within relevant frequency bands, using the Welch's method; (3) Kullback-Leibler (KL) relative to the mean training-evoked responses [21, 74]; (4) signal complexity measures. All features were computed over a one-second window following stimulus presentation. To reduce input dimensionality, Principal Component Analysis (PCA) was applied, retaining 99% of the variance. We omit multi-class classification due to a well-known issue of performance degradation [1, 39, 70].

ERPs capture the early neural response to a stimulus, identifiable by a distinct voltage peak. Their latency, topography, and amplitude correlate with specific cognitive functions, making ERPs a useful metric for studying neural processes. Complementing this, our frequency-based features target well-established EEG bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and low-beta (13–20 Hz) [10]. Each band is associated with distinct cognitive processes: delta with attention and cognitive effort, theta with memory encoding and cognitive control, alpha with attentional suppression and workload regulation, and low-beta with active thinking and decision-making. Our preliminary spectral analysis indicated that frequencies above 20 Hz exhibited minimal task-related variation, prompting the exclusion of high-beta and gamma bands to reduce muscle artifact influence [11].

Additional features were derived through data-driven methods. For example, KL divergence was used to quantify the similarity between signals within the same condition, while complexity measures like approximate entropy [49, 76], Higuchi's fractal dimension [27, 64, 76], Katz's fractal dimension [16, 64], and Detrended Fluctuation Analysis (DFA) [52, 64], helped differentiate experimental conditions based on their inherent signal complexity [64].

To ensure stable performance estimates, we created five different train-test splits (70% training, 30% testing) from the initial dataset. For each split, the classes were balanced via random down-sampling. The down-sampling was done five times per split to avoid biases, yielding 25 dataset variations per participant for independent performance evaluation and accounting for data variability.

3.7.2 Between-subject classification. We expanded our experiments to explore and characterise consistent neural patterns across participants. Given the inherent noise in raw EEG signals, group-level analysis is particularly advantageous as it enables the use of ERPs for such a purpose.

The typical approach to ERP computation involves averaging all epochs per condition and participant, which reduces our dataset from (N = 19 participants × 30 epochs × 4 conditions) to (N = 19 subjects × 1 ERPs × 4 conditions). Moreover, given that each participant contributed a moderate number of epochs per condition, this can lead to suboptimal ERPs. To address these limitations, we implemented a data augmentation strategy that generates multiple synthetic ERPs (N = 19 participants × 50 ERPs × 4 categories) per participant and condition through a pseudo-random linear combination of epochs. Specifically, the variability of each epoch was used to weight its contribution to the synthetic ERPs, with larger variability assumed to indicate broader coverage of brain activity. To further increase diversity, these weights were adjusted by introducing noise, multiplying them by a random factor within the range [0.5, 1]. This approach was designed to model uncontrolled variability in attention and brain activity during each task. The adjusted weights were then used in a weighted averaging process to generate synthetic ERPs, which were subsequently used for training.

Considering the potential noise in the derived ERPs, we implemented a data-driven classification approach for each pair of conditions, using a two-stage cross-validation approach. First, an outer leave-one-out cross-validation was employed, splitting the participants and all their ERP variations into training and test sets. Within each training set, an inner 5-fold cross-validation was used to train a base model for each EEG channel. Each model consisted of a Multi-Layer Perceptron (MLP) with a single hidden layer of 50

Table 2: Classification accuracy (Acc.) and % of difference (Diff) from baseline; $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$ (Bonferroni-corrected).

Feature set	Exact vs. Complement		Exact vs. Substitute		Exact vs. Irrelevant		Complement vs. Substitute		Complement vs. Irrelevant		Substitute vs. Irrelevant	
	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)
EEG Theta Delta	0.545***	9.0	0.524***	4.8	0.545***	9.0	0.519*	3.8	0.531***	6.2	0.537***	7.4
EEG KL	0.545***	9.0	0.524***	4.8	0.544***	8.8	0.519*	3.8	0.530***	6.0	0.537***	7.4
EEG DFA HFD	**0.546*****	**9.2**	0.524***	4.8	0.544***	8.8	0.520*	4.0	0.529***	5.8	0.537***	7.4
EEG KL Delta	0.544***	8.8	**0.525*****	**5.0**	0.544***	8.8	0.518*	3.6	0.530***	6.0	0.537***	7.4
EEG Theta HFD	**0.546*****	**9.2**	0.524***	4.8	**0.546*****	**9.2**	0.519*	3.8	0.530***	6.0	0.537***	7.4
KL Beta	0.496	-0.8	0.495	-1.0	0.511	2.2	**0.533*****	**6.6**	0.509	1.8	0.509	1.8
Entropy Theta Delta	0.505	1.0	0.483	-3.4	0.539***	7.8	0.483	-3.4	**0.549*****	**9.8**	0.543***	8.6
KL Theta Delta	0.503	0.6	0.492	-1.6	0.539***	7.8	0.495	-1.0	0.535***	7.0	**0.549*****	**9.8**

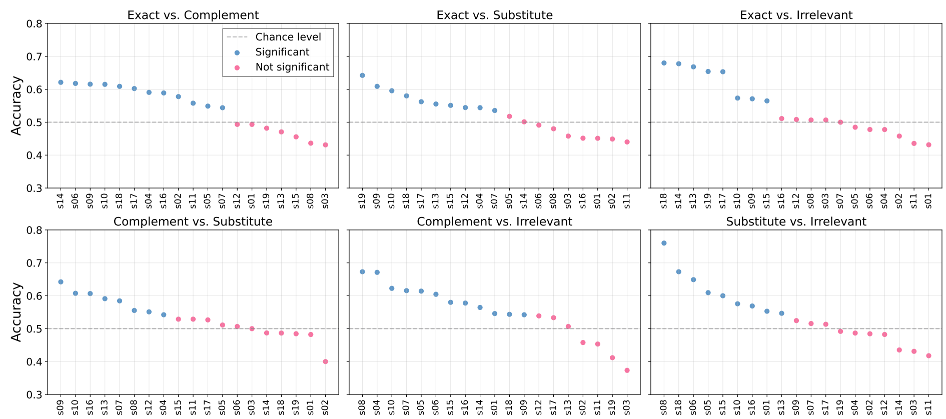


Figure 4: Accuracy plots of the best performing feature combination per condition, according to Table 3.

units. Training was limited to 500 iterations, with a regularisation parameter $\alpha = 0.05$ to prevent overfitting.

The mean validation accuracy of each base model was then used to generate a topomap (Figure 5), illustrating the predictive power of each electrode's recorded activity. Additionally, the probabilistic predictions from these models served as features for training and testing a logistic regression classifier, at the outer cross-validation level. This meta-model predicted the final accuracy for each subject based on its predictions. Finally, the mean accuracy across all participants was computed. Given the stochastic nature of ERP generation, the entire process was repeated five times, and the mean and standard deviation of the resulting accuracies were computed.

3.7.3 Between-subjects engagement analysis. EEG data were analysed to examine FAA, a proxy of engagement, approach-avoidance motivation [65], and emotional regulation [24]. Specifically, FAA is calculated as the difference in alpha power between left (F3) and right (F4) frontal electrodes, which has been associated with emotional processing, cognitive control, and decision-making [19]:

$$\mathrm{FAA} = \log(P_{F4}) - \log(P_{F3}) \tag{1}$$

Higher FAA values indicate greater relative right-frontal activity, which has been linked to withdrawal-related emotions, while lower (or negative) values indicate greater left-frontal dominance, often associated with approach motivation [9]. Additionally, theta and beta frequency bands' power was analysed following prior work [23]. In a similar way, higher powers indicate a higher engagement of the participant with the task, providing complementary insight into neural processing during the experiment.

To extract the different frequency bands' power, we applied Welch's method to compute PSD within the specific ranges as described previously. Welch's method is commonly used in EEG research to estimate power in different frequency bands, as it improves the robustness of spectral estimates by averaging over multiple overlapping windows [8]. The resulting PSD values for the electrodes of interest were averaged across the epoch to obtain a single mean power value per electrode (C3 for theta/beta bands and F3/F4 for FAA calculation). For this analysis, six subjects were excluded due to poor SNR for the corresponding calculations.

4 Results and Discussion

4.1 Neural variation across recommendation categories

This section addresses **RQ1:** *Do different recommendation categories evoke distinct neural signatures in brain activity as measured through EEG, and can these patterns be reliably identified?* Initially, we report results on within-subjects classification that examines individual-specific neural patterns and their variability across recommendation categories. Then, we discuss our between-subjects analysis with aims to identify common neural signatures and evaluate the generalisability of our findings across participants.

4.1.1 Within-subject experiments. We begin our analysis by focusing on the within-subject classifier performance to examine the consistency of neural responses to each recommendation category at the individual level. Table 2 shows the best performing configurations in terms of mean accuracy, for each pairwise comparison

Table 3: Mean classification accuracy (Acc.) and % of difference (Diff) from random baseline (only above-chance results from Table 2 as reported by Eugster et al. [17]) for each feature set and recommendation category comparison.

Feature set	Exact vs. Complement		Exact vs. Substitute		Exact vs. Irrelevant		Complement vs. Substitute		Complement vs. Irrelevant		Substitute vs. Irrelevant	
	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)	Acc.	Diff (%)
EEG Theta Delta	**0.591**	**18.2**	**0.567**	**13.4**	0.583	16.6	**0.570**	**14.0**	0.580	16.0	0.562	12.4
EEG KL	**0.591**	**18.2**	**0.567**	**13.4**	0.590	18.0	0.563	12.6	**0.581**	**16.2**	0.558	11.6
EEG DFA HFD	**0.591**	**18.2**	0.566	13.2	0.590	18.0	0.569	13.8	0.579	15.8	0.558	11.6
EEG KL Delta	**0.591**	**18.2**	**0.567**	**13.4**	0.597	19.4	0.568	13.6	0.580	16.0	0.558	11.6
EEG Theta HFD	0.590	18.0	0.565	13.0	**0.599**	**19.8**	**0.570**	**14.0**	0.579	15.8	0.559	11.8
KL Beta	0.562	12.4	0.549	9.8	0.556	11.2	0.560	12.0	0.558	11.6	0.547	9.4
Entropy Theta Delta	0.568	13.6	0.546	9.2	0.577	15.4	0.559	11.8	**0.582**	**16.4**	**0.591**	**18.2**
KL Theta Delta	0.572	14.4	0.540	8.0	0.577	15.4	0.547	9.4	0.576	15.2	0.579	15.8

and across all participants. Our results suggest that, in most cases, the classifier's average performance exceeds chance levels, indicating the presence of meaningful neurophysiological differences between conditions [17]. These findings are further supported by a Mann-Whitney significance test, which confirms that our predictions significantly differ from chance ($p < .01$, Bonferroni corrected).

Specifically, the "E vs. S" comparison is associated with the lowest mean accuracy, a trend that remains even after excluding individual models (42%) with near-chance performance, as shown in Table 3 and Figure 4. These results suggest that, while neural responses to **E**xact and **S**ubstitute items might still be distinguishable in some cases, the classifier's reduced performance in this pair suggests that these two categories consistently evoke more similar neural responses than other comparisons. This overlap highlights a key insight: from a neurological perspective, the brain may process substitute alternatives and exact matches in a closely related manner.

On the other hand, the "C vs. I" (0.549), "S vs. I" (0.549), and "E vs. I" (0.546) comparisons yielded the highest mean accuracies, demonstrating that the brain differentiates between product recommendations relevant to the search query (**E**xact, **C**omplement, or **S**ubstitute) and those that are not. This pattern holds after removing low-performing individual models, as shown in Table 2. However, the ordering of accuracies shows minor variations across models, with "E vs. I" emerging as the top performer. This comparison also had the highest number of excluded subjects (Figure 4), suggesting that although 42% of individuals seem to display a clearer neural differentiation between exact and irrelevant items, a substantial subset (58%) may not, reflecting a slightly larger inter-individual variability. Similar accuracies were observed for the "E vs. C" comparison (0.546 for the whole sample, and 0.591 after excluding 37% of individual models, based on Figure 4), indicating the presence of relevant brain activity differences. While speculative, these differences could be attributed to the associative and semantic processing involved in identifying complementary products, as opposed to the more straightforward processing of exact product matches.

Regarding the best-performing features in individual models, we observe that the raw EEG signal consistently dominates over other features, leading to similar outcomes across all experiments that incorporate it. This suggests that the raw EEG signal captures a substantial amount of information, making it a strong baseline for classification. However, for the top-performing comparisons, such as "C vs. I" and "S vs. I", combinations of a few derived parameters (most notably theta frequency power, which is also among the top-performing features in the "E vs. I" model) yielded better results. This suggests that these derived features are particularly well-suited

Table 4: Meta-classifier accuracy across different comparisons over multiple runs, for the between-subject classification strategy. Bold denotes best run.

Comparison	1	2	3	4	5	AVG ± STD
E vs. I	**0.642**	0.600	0.537	0.563	0.500	0.568 ± 0.055
E vs. C	**0.595**	0.568	0.558	0.505	0.563	0.558 ± 0.033
E vs. S	0.432	0.442	0.421	0.374	**0.453**	0.424 ± 0.031
I vs. C	0.500	**0.584**	0.537	0.537	0.421	0.516 ± 0.061
I vs. S	0.459	0.459	0.463	0.489	**0.605**	0.495 ± 0.063
C vs. S	0.647	**0.674**	0.663	0.611	0.626	0.644 ± 0.026

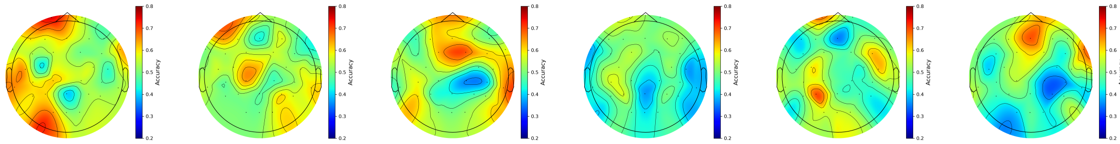
to capturing distinctions between these specific recommendation categories. These findings align with those of Liang et al. [38], who demonstrated that frontal theta activity supports the detection of mismatched information in visual working memory tasks.

In other comparisons, models using derived features underperformed those based on raw EEG signals, likely due to a loss of neural complexity and unmodelled data variations. Overall, within-subject results align with observed user behaviour (Section 4.2). Above-chance classification across comparisons suggests distinct neural correlates, though expectedly overlapping categories (e.g., "E vs. S" and "C vs. S") show weaker differentiation.

4.1.2 Between-subject experiments. The results from the meta-models trained on all subjects show that, for certain comparisons, accuracies are comparable to or even higher than those of individual models after excluding chance-level performers. This suggests that using ERPs may help reduce noise and leverage data from all participants to identify more robust neural patterns. However, for some comparisons, the mean meta-accuracies remain around chance level, indicating that the advantages of this approach may not be uniform across all conditions. The average accuracies of the meta-models for each comparison are presented in Table 4.

Consistent with the previous results, the "E vs. S" comparison yielded the lowest mean accuracy. In contrast, comparisons involving products relevant to the query (**E**xact or **C**omplement) and irrelevant ones remain among the top performers. Unexpectedly, the highest-performing classifier is the "C vs. S" comparison, which was associated with one of the lowest performances in the previous section. These results might suggest that the distinction between **C**omplement and **S**ubstitute products are linked to subtle neural patterns that become more apparent when working with ERPs.

The topographic accuracy distribution of channel-based models (Figure 5) shows that the lowest-performing meta-model ("E vs. S") lacked electrodes with accuracy above 0.6, reinforcing previous



(a) C vs. S **(b) E vs. C** **(c) E vs. I** **(d) E vs. S** **(e) I vs. C** **(f) I vs. S**

Figure 5: Mean base-models' accuracy per electrode and category comparison (discussed in Section 3.7.2).

Table 5: Significance tests for Relevance, Likelihood of Purchase, and Diversity across recommendation categories. Values in parentheses are medians. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (Bonferroni-corrected).

Test	Comparison	Relevance	Likelihood of Purchase	Diversity
Friedman		$\chi^2(3) = 56.37^{***}$	$\chi^2(3) = 50.74^{***}$	$\chi^2(3) = 31.67^{***}$
Paired Wilcoxon	**E**xact vs **S**ubstitute	$W = 3^{**}$ (5 vs. 4)	$W = 7.5^{**}$ (4 vs. 4)	$W = 15.5^{**}$ (2 vs. 3)
	Exact vs **C**omplement	$W = 0^{***}$ (5 vs. 4)	$W = 0^{***}$ (4 vs. 4)	$W = 10^{***}$ (2 vs. 3)
	Exact vs **I**rrelevant	$W = 0^{***}$ (5 vs. 1)	$W = 0^{***}$ (4 vs. 1)	$W = 13^{***}$ (2 vs. 5)
	Substitute vs **C**omplement	$W = 36.5^{*}$ (4 vs. 4)	$W = 18^{**}$ (3 vs. 3)	$W = 17.5^{**}$ (2 vs. 3)
	Substitute vs **I**rrelevant	$W = 0^{***}$ (4 vs. 1)	$W = 0^{***}$ (3 vs. 1)	$W = 12^{***}$ (2 vs. 5)
	Complement vs **I**rrelevant	$W = 0^{***}$ (4 vs. 1)	$W = 12^{***}$ (3 vs. 1)	$W = 14^{**}$ (3 vs. 5)

findings. In contrast, the best-performing models consistently captured distinct activity patterns in specific brain regions. Notably, variations in the mid and left prefrontal cortex correlated with higher predictive power, suggesting its role in relevance assessment [38] ("E", "C", and "S" vs. "I") and concept learning [38, 41] ("E vs C"). The prefrontal cortex, key to executive functions such as planning, decision-making, and social behavior, integrates information broadly, while its left ventrolateral region is particularly specialized for language and verbal memory [38, 42].

Moreover, the "C vs. S" comparison shows a pronounced and extensive left-lateralization difference, which may reflect distinct involvement of a range of cognitive tasks. These tasks could include differences in semantic processing (e.g., complement products might be more closely linked to this type of processing), as well as visual attention and top-down modulation [48] (e.g., substitute products, which are closer to the search, may require more attention to detail). However, the precise nature of this difference remains uncertain, and further investigation is needed to determine its significance.

In summary, neural responses to **E**xact and **S**ubstitute product recommendations are largely similar, with greater differentiation for more distinct product types. The "C vs. S" comparison ranks among the best-performing models, indicating that subtle neural patterns emerge with ERPs. Topographic analysis highlights the prefrontal cortex as central to these distinctions. Our findings also suggest that evaluating **S**ubstitute and **C**omplement products engages broader cognitive processes.

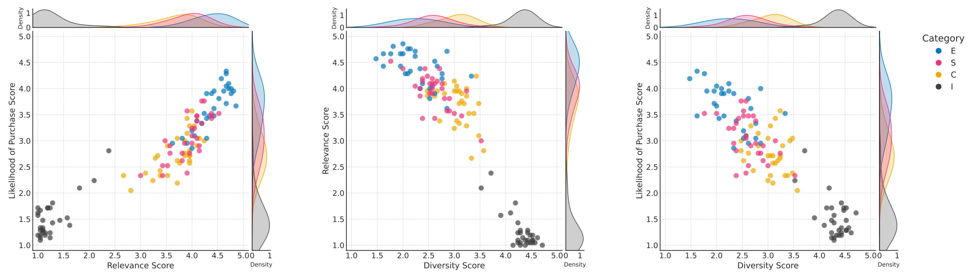
4.2 Behavioural variation across recommendation categories

This section addresses **RQ2:** *How do different recommendation categories influence users' perception (and subsequent observable behaviour) regarding item relevance, purchaseability, and recommendation diversity?* We find that, on average, users correctly classify 71% of product categories, with most confusion occurring between the **E**xact and **S**ubstitute categories. A chi-square test of independence ($X^2(16) = 9.96$, $p = .86$) revealed no significant differences between expected and observed confusion matrices, indicating that perceived recommendation categories align with predefined labels.

Considering all product categories, our analysis (Table 5 and Figure 6) reveals that **E**xact and **I**rrelevant recommendations lie at the extremes across Relevance, Likelihood of Purchase, and Diversity scales. In contrast, **S**ubstitute and **C**omplement recommendations score in the mid-range. The Diversity vs. Likelihood and Diversity vs. Relevance plots show that while **E**xact and **I**rrelevant recommendations are distinctly separated at the high and low ends, the clusters for **S**ubstitute and **C**omplement tend to overlap—**S**ubstitute with **E**xact at the upper end and with **C**omplement at the lower end. The Relevance vs. Likelihood plot indicates that **E**xact, **S**ubstitute, and **C**omplement recommendations achieve above-average scores.This suggests that **C**omplement and **S**ubstitute recommendations can improve product diversity without sacrificing relevance.

Furthermore, Figure 6 shows correlation trends of the behavioural scales, regardless of categories. As expected, Relevance correlates positively with Likelihood of Purchase (LoP) but negatively with Diversity. Interestingly, while LoP also shows a negative correlation with Diversity, the effect is weaker and not statistically significant. This suggests that intermediate recommendations, such as **S**ubstitute and **C**omplement, strike a balance—maintaining relevance while enhancing diversity. Thus, RS could improve user experience by incorporating these intermediate categories rather than relying solely on highly relevant or novel recommendations.

Last, we analysed behaviour across classification types. We applied the Shapiro-Wilk and Levene tests, which revealed significant deviations from normality and homogeneity. Then, we used repeated-measures Friedman and paired Wilcoxon signed-rank tests. Our results demonstrate the importance of designing RS that account for behavioural variability (as shown in Table 5), suggesting that integrating intermediate recommendation types enhance perceived relevance and foster a more diverse, engaging experience.



(a) Relevance vs. Likelihood of purchase (p=.02, r=.68).

(b) Diversity vs. Relevance (p=.01, r=-.56).

(c) Diversity vs Likelihood of purchase (p=.1, r=-.42).

Figure 6: Scatter plots comparing average rating scores including Pearson Correlation values between ratings (each dot represents mean score per stimulus, per category).

4.3 Engagement profiling

This section addresses **RQ3:** *Can we identify neurophysiological markers of engagement that systematically vary across recommendation categories, and what do these variations reveal about the effectiveness of different recommendation strategies?* To investigate differences in engagement metrics (e.g., FAA, Theta and Beta power; Section 3.7.3) across recommendation categories, we conducted a repeated measures ANOVA with category (Exact, Irrelevant, Substitute, Complement) as the within-subject factor (Figure 7). For FAA, we observed a significant main effect of category ($F(3, 56) = 2.91$, $p < .05$), indicating that FAA varied across recommendation categories. The **E**xact condition showed the highest and only positive mean value (Mean = 9.91e-14). This finding indicates that this condition elicited greater left frontal activity, which has been associated with approach motivation and positive affect.

Similarly, for Theta power, there was a significant main effect of Category ($F(3, 56) = 7.63$, $p < .01$). Also, the **E**xact condition showed to elicit the highest value (Mean = 8.9). We noted the same with Beta power ($F(3, 56) = 3.5$, $p < .05$) where the **E**xact condition showed the highest value (Mean = 4.22). These findings reinforce that **E**xact recommendations are associated with more positive cognitive and affective responses, as reflected in neural activity.

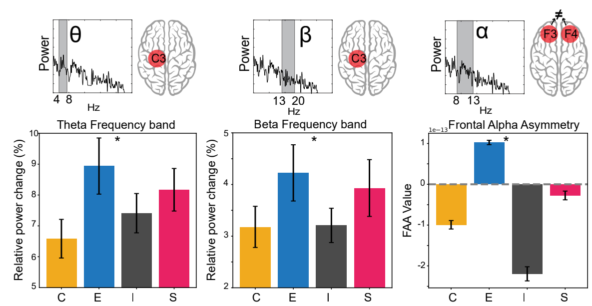


Figure 7: Engagement across recommendation categories.

5 Conclusions

Recommender systems have traditionally prioritised relevance, often neglecting how different types of recommendations shape not only user ratings but also underlying neural processing and decision strategies. Specifically, our study demonstrated that **E**xact and **S**ubstitute items evoke comparable cognitive responses and user evaluations, while **C**omplement recommendations strike a balance between high relevance, purchasability and diversity, outperforming **I**rrelevant items without sacrificing novelty. By leveraging item-to-item relationships from historical interactions, RS could improve recommendation quality and even alleviate cold-start issues for new users or products.

Moreover, we observed notable individual differences—some users clearly distinguish between recommendation categories (Figure 4; subsection 4.2), whereas others react more uniformly, and perceptions of **C**omplement items range from "irrelevant clutter" to "valuable extras." Such variability highlights the shortcomings of one-size-fits-all approaches and points toward multi-objective, personalised strategies that integrate richer behavioural—and where possible, neural—signals. For instance, neural and behavioural user signals could be utilized to personalize complementary item recommendations in a manner similar to Yan et al. [69], offering an advantage through the incorporation of user signals. Looking beyond e-commerce, these principles of contextualisation and adaptive, user-specific tuning are likely to improve recommendation effectiveness in other domains like music, video and gaming.

Despite its insights, this study has some limitations. First, while EEG captures neural dynamics well, its low spatial resolution limits source localisation compared to fMRI or MEG. It is also prone to motion and electromagnetic artifacts; although preprocessing was applied, some noise remains. Future work could use more robust artifact rejection or combine EEG with other physiological measures. Second, the lab setting reduces confounds but may not reflect real-world cognition and emotion. Ecologically valid setups, e.g., real-time, adaptive interactions, could improve generalisability. Finally, categorisation is subjective and shaped by individual experience; personalised models could help capture this variability.

Acknowledgments

This work was supported by Horizon Europe's European Innovation Council under Pathfinder project SYMBIOTIK (grant 101071147) and by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR), Generalitat de Catalunya (grant 2023 DI 00080).

References

[1] Haider Abdulkarim and Mohammed Al-Faiz. 2021. Online multiclass EEG feature extraction and recognition using modified convolutional neural network method. *International Journal of Electrical and Computer Engineering* 11 (05 2021), 4016–4026. doi:10.11591/ijece.v11i5.pp4016-4026

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on* 17 (07 2005), 734–749. doi:10.1109/TKDE.2005.99

[3] Mashael Aldayel, Mourad Ykhlef, and Abeer Al-Nafjan. 2020. Deep learning for EEG-based preference classification in neuromarketing. *Applied Sciences* 10, 4 (2020), 1525.

[4] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic Effects on the Diversity of Consumption on Spotify. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2155–2165. doi:10.1145/3366423.3380281

[5] Diego Antognini, Claudiu Musat, and Boi Faltings. 2021. Interacting with Explanations through Critiquing. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 515–521. doi:10.24963/ijcai.2021/72 Main Track.

[6] Diego Antognini, Claudiu Cristian Musat, and Boi Faltings. 2020. T-RECS: a Transformer-based Recommender Generating Textual Explanations and Integrating Unsupervised Language-based Critiquing. *ArXiv* abs/2005.11067 (2020). https://api.semanticscholar.org/CorpusID:218863076

[7] Sathsarani K. Bandara, Uvini C. Wijesinghe, Badra P. Jayalath, Saumya K. Bandara, Prasanna S. Haddela, and Lumini M. Wickramasinghe. 2021. EEG Based Neuromarketing Recommender System for Video Commercials. In *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. doi:10.1109/ICIIS53135.2021.9660742

[8] Kurt Barbe, Rik Pintelon, and Johan Schoukens. 2009. Welch method revisited: nonparametric power spectrum estimation via circular overlap. *IEEE Transactions on signal processing* 58, 2 (2009), 553–565.

[9] Catarina Barros, Ana Rita Pereira, Adriana Sampaio, Ana Buján, and Diego Pinal. 2022. Frontal alpha asymmetry and negative Mood: a cross-sectional study in older and younger adults. *Symmetry* 14, 8 (2022), 1579.

[10] Gyorgy Buzsaki. 2006. *Rhythms of the Brain*. Oxford university press.

[11] György Buzsáki and Fernando Lopes da Silva. 2012. High frequency oscillations in the intact brain. *Progress in neurobiology* 98, 3 (2012), 241–249.

[12] Mark Claypool, Phong Le, Makoto Wased, and David Brown. 2001. Implicit interest indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*. 33–40.

[13] Aminu Da'u and Naomie Salim. 2020. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review* 53, 4 (2020), 2709–2748. doi:10.1007/s10462-019-09744-1

[14] Keith M. Davis, Michiel Spape, and Tuukka Ruotsalo. 2023. Contradicted by the Brain: Predicting Individual and Group Preferences via Brain-Computer Interfacing. *IEEE Transactions on Affective Computing* 14, 4 (2023), 3094–3105. doi:10.1109/TAFFC.2022.3225885

[15] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys). arXiv:2404.00579 [cs.IR] https://arxiv.org/abs/2404.00579

[16] R. Esteller, G. Vachtsevanos, J. Echauz, and B. Litt. 2001. A comparison of waveform fractal dimension algorithms. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* 48, 2 (2001), 177–183. doi:10.1109/81.904882

[17] Manuel Eugster, Tuukka Ruotsalo, Michiel Spapé, Ilkka Kosunen, Oswald Barral, Niklas Ravaja, Giulio Jacucci, and Samuel Kaski. 2014. Predicting term-relevance from brain signals. (07 2014). doi:10.1145/2600428.2609594

[18] Chen Gao. 2022. Causal Inference in Recommender Systems: A Survey and Future Directions. *arXiv preprint arXiv:2208.12397* (Dec 2022). Version 2.

[19] Ian H Gotlib. 1998. EEG alpha asymmetry, depression, and cognitive functioning. *Cognition & Emotion* 12, 3 (1998), 449–478.

[20] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. 2013. MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience* 7, 267 (2013), 1–13. doi:10.3389/fnins.2013.00267

[21] Anjum Gupta, Shibin Parameswaran, and Cheng-Han Lee. 2009. Classification of electroencephalography (EEG) signals for different mental activities using Kullback Leibler (KL) divergence, In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan. *IEEE Int. Conf. Acoust. Speech Signal Process*, 1697–1700. doi:10.1109/ICASSP.2009.4959929

[22] Deepesh V. Hada, Vijaikumar M., and Shirish K. Shevade. 2021. ReXPlug: Explainable Recommendation using Plug-and-Play Language Model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 81–91. doi:10.1145/3404835.3462939

[23] Laura K Halderman, Bridgid Finn, JR Lockwood, Nicole M Long, and Michael J Kahana. 2021. EEG correlates of engagement during assessment. *ETS Research Report Series* 2021, 1 (2021), 1–17.

[24] Eddie Harmon-Jones, Philip A Gable, and Carly K Peterson. 2010. The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological psychology* 84, 3 (2010), 451–462.

[25] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 368–377.

[26] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *2008 Eighth IEEE International Conference on Data Mining*. doi:10.1109/ICDM.2008.22

[27] A.J. Ibáñez-Molina and S. Iglesias-Parro. 2014. Fractal characterization of internally and externally generated conscious experiences. *Brain and Cognition* 87 (2014), 69–75. doi:10.1016/j.bandc.2014.03.002

[28] Sheena S Iyengar and Mark R Lepper. 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. *Journal of personality and social psychology* 76, 3 (1999), 349.

[29] Giulio Jacucci, Oswald Barral, Pedram Daee, Markus Wendt, Baris Serim, Tuukka Ruotsalo, Patrik Pluchino, Jonathan Freeman, Luciano Gamberini, Samuel Kaski, and Benjamin Blankertz. 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. doi:10.1002/asi.24161

[30] Gawesh Jawaheer, Martin Szomszor, and Patty Kostkova. 2010. Comparison of implicit and explicit feedback from an online music recommendation service. In *proceedings of the 1st international workshop on information heterogeneity and fusion in recommender systems*. 47–51.

[31] Gawesh Jawaheer, Peter Weller, and Patty Kostkova. 2014. Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 4, 2 (2014), 1–26.

[32] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. 2022. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics* 11, 1 (2022). doi:10.3390/electronics11010141

[33] Lukas Lerche. 2016. Using implicit feedback for recommender systems: characteristics, applications, and challenges. (2016).

[34] Jiacheng Li, Zhankui He, Jingbo Shang, and Julian McAuley. 2023. UCEpic: Unifying Aspect Planning and Lexical Constraints for Generating Explanations in Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 1248–1257. doi:10.1145/3580305.3599535

[35] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. *ArXiv* abs/2105.11601 (2021). https://api.semanticscholar.org/CorpusID:235187032

[36] Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized Prompt Learning for Explainable Recommendation. *ACM Trans. Inf. Syst.* 41, 4, Article 103 (March 2023), 26 pages. doi:10.1145/3580488

[37] Sukun Li and Xiaoxing Liu. 2022. Toward a BCI-Based Personalized Recommender System Using Deep Learning. In *2022 IEEE 8th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. doi:10.1109/BigDataSecurityHPSCIDS54978.2022.00042

[38] Tengfei Liang, Zhonghua Hu, and Qiang Liu. 2017. Frontal Theta Activity Supports Detecting Mismatched Information in Visual Working Memory. *Frontiers in Psychology* 8 (10 2017). doi:10.3389/fpsyg.2017.01821

[39] Claudia Lindig-León, Sylvain Rimbert, and Laurent Bougrain. 2020. Multiclass Classification Based on Combined Motor Imageries. *Frontiers in Neuroscience* 14 (2020), 559858. doi:10.3389/fnins.2020.559858 Published 2020 Nov 19.

[40] Fabien Lotte, Laurent Bougrain, Andrzej Cichocki, Maureen Clerc, Marco Congedo, Alain Rakotomamonjy, and Florian Yger. 2018. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering* 15, 3 (2018), 031005.

[41] Michael L Mack, Alison R Preston, and Bradley C Love. 2020. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* 11, 1 (Jan. 2020), 46.

[42] Michael L. Mack, Alison R. Preston, and Bradley C. Love. 2020. Ventromedial prefrontal cortex compression during concept learning. *Nature Communications* 11, 1 (2020), 46. doi:10.1038/s41467-019-13930-8

[43] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring Networks of Substitutable and Complementary Products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 785–794. doi:10.1145/2783258.2783381

- [44] Yashar Moshfeghi and Frank Pollick. 2021. Neural Correlates of Realisation of Satisfaction in a Successful Search Process. *Proceedings of the Association for Information Science and Technology* 58 (10 2021), 282–291. doi:10.1002/pra2.456
- [45] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China. doi:10.18653/v1/D19-1018
- [46] H. Nolan, R. Whelan, and R. B. Reilly. 2010. FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods* 192, 1 (2010), 152–162. doi:10.1016/j.jneumeth.2010.07.015
- [47] Douglas W. Oard and Jinmook Kim. 1998. Implicit Feedback for Recommender Systems. https://api.semanticscholar.org/CorpusID:8987583
- [48] Spiro P Pantazatos, Ted K Yanagihara, Xian Zhang, Thomas Meitzler, and Joy Hirsch. 2012. Frontal-occipital connectivity during visual search. *Brain Connect.* 2, 3 (July 2012), 164–175.
- [49] C. Pappalettera, A. Cacciotti, L. Nucci, F. Miraglia, P. M. Rossini, and F. Vecchio. 2023. Approximate entropy analysis across electroencephalographic rhythmic frequency bands during physiological aging of human brain. *Geroscience* 45, 2 (April 2023), 1131–1145. doi:10.1007/s11357-022-00710-4
- [50] Krupa Patel and Hiren B. Patel. 2020. A state-of-the-art survey on recommendation system and prospective extensions. *Computers and Electronics in Agriculture* 178 (2020), 105779. doi:10.1016/j.compag.2020.105779
- [51] Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior research methods* 51 (2019), 195–203.
- [52] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. 1994. Mosaic organization of DNA nucleotides. *Phys. Rev. E* 49 (Feb 1994), 1685–1689. Issue 2. doi:10.1103/PhysRevE.49.1685
- [53] Alexandre Pereda-Baños, Ioannis Arapakis, and Miguel Barreda-Ángeles. 2015. On Human Information Processing in Information Retrieval (Position Paper). In *SIGIR Workshop on Neuro-Physiological Methods in IR Research (NeuroIR)*. ACM.
- [54] Zuzana Pinkosova, William J. McGeown, and Yashar Moshfeghi. 2023. Revisiting Neurological Aspects of Relevance: An EEG Study. In *Machine Learning, Optimization, and Data Science*. Springer Nature Switzerland.
- [55] Yuanpeng Qu and Hajime Nobuhara. 2024. Generating Explanations for Explainable Recommendations Using Filter-Enhanced Time-Series Information. *IEEE Access* 12 (2024), 78480–78495. doi:10.1109/ACCESS.2024.3408252
- [56] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. 2022. Shopping Queries Dataset: A Large-Scale ESCI Benchmark for Improving Product Search. arXiv:2206.06588 [cs.IR] https://arxiv.org/abs/2206.06588
- [57] Tuukka Ruotsalo, Kalle Mäkelä, Michiel M. Spapé, and Luis A. Leiva. 2023. Affective Relevance: Inferring Emotional Responses via fNIRS Neuroimaging. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3539618.3591946
- [58] Barry Schwartz. 2004. The paradox of choice: Why less is more. *New York: Ecco* (2004).
- [59] Tanu Srivastava, Shilpi Singh, and Sanjeev Bhardwaj. 2013. Support vector machine technique for EEG signals. *International Journal of Engineering Research and Applications* 3, 1 (2013), 533–538.
- [60] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2021. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. *CoRR* abs/2110.15097 (2021). https://arxiv.org/abs/2110.15097
- [61] Abdulhamit Subasi, Ahmet Alkaya, Rahime Keskin, and Etem Koklukaya. 2017. Classification of EEG signals based on pattern recognition approach. *International Journal of Imaging Systems and Technology* 27, 3 (2017), 213–221.
- [62] Samuel Sutton, Margery Braren, Joseph Zubin, and ER John. 1965. Evoked-potential correlates of stimulus uncertainty. *Science* 150, 3700 (1965), 1187–1188.
- [63] Bella Tadson, Jared Boasen, François Courtemanche, Noémie Beauchemin, Alexander-John Karran, Pierre-Majorique Léger, and Sylvain Sénécal. 2023. Neuro-Adaptive Interface System to Evaluate Product Recommendations in the Context of E-Commerce. In *Design Science Research for a New Society: Society 5.0*, Aurona Gerber and Richard Baskerville (Eds.).
- [64] Shaohua Tang and Zheng Li. 2024. EEG complexity measures for detecting mind wandering during video-based learning. *Scientific Reports* 14, 1 (8 4 2024), 8209. doi:10.1038/s41598-024-58889-9
- [65] Ryan Thibodeau, Randall S. Jorgensen, and Sanghag Kim. 2006. Depression, anxiety, and resting frontal EEG asymmetry: A meta-analytic review. *Journal of Abnormal Psychology* 115, 4 (2006), 715–729. doi:10.1037/0021-843X.115.4.715
- [66] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B. Miller, Jeff Huang, and Ben D. Sawyer. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 38 (March 2022), 56 pages. doi:10.1145/3502222
- [67] Chun-Shu Wei, Yuan-Pin Lin, Yu-Te Wang, Tzyy-Ping Jung, Nima Bigdely-Shamlo, and Chin-Teng Lin. 2016. Classification of motor imagery EEG signals with support vector machines and particle swarm optimization. *Computational and mathematical methods in medicine* 2016 (2016).
- [68] Ryen W White, Joemon M Jose, Ian Ruthven, EM Voorhees, and DK Harman. 2002. Comparing explicit and implicit feedback techniques for web retrieval: Trec-10 interactive track report. In *Proceedings of the tenth text retrieval conference (TREC-10)*.
- [69] An Yan, Chaosheng Dong, Yan Gao, Jinmiao Fu, Tong Zhao, Yi Sun, and Julian McAuley. 2022. Personalized complementary product recommendation. (2022). https://www.amazon.science/publications/personalized-complementary-product-recommendation
- [70] Banghua Yang, Fenqi Rong, Yunlong Xie, Du Li, Jiayang Zhang, Fu Li, Guangming Shi, and Xiaorong Gao. 2025. A multi-day and high-quality EEG dataset for motor imagery brain-computer interface. *Scientific Data* 12, 1 (2025), 488. doi:10.1038/s41597-025-04826-y
- [71] Ziyi Ye, Qingyao Ai, and Yiqun Liu. 2024. Brain-Computer Interface Meets Information Retrieval: Perspective on Next-generation Information System. In *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding*. 61–65.
- [72] Ziyi Ye, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuancheng Li, Jiaji Li, Xuesong Chen, Min Zhang, and Shaoping Ma. 2022. Why Don't You Click: Understanding Non-Click Results in Web Search with Brain Signals. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 633–645.
- [73] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 582–590.
- [74] Yingchun Zhang, Xin Wang, Yiwen Jiang, and Xiaowei Xu. 2024. EEGGAN-Net: enhancing EEG signal classification through data augmentation framework. *Frontiers in Human Neuroscience* 18 (2024), 1430086. doi:10.3389/fnhum.2024.1430086
- [75] Yong Zhang, Yong Wang, Jianting Jin, and Xingyu Wang. 2020. EEG signal classification based on SVM with improved squirrel search algorithm. *Computational intelligence and neuroscience* 2020 (2020).
- [76] M. A. Zúñiga, Á. Acero-González, J. C. Restrepo-Castro, M. Á. Uribe-Laverde, D. A. Botero-Rosas, B. I. Ferreras, M. C. Molina-Borda, and M. P. Villa-Reyes. 2024. Is EEG Entropy a Useful Measure for Alzheimer's Disease? *Actas Esp Psiquiatr* (2024). doi:10.62641/aep.v52i3.1632