基于脑机接口的信息检索技术研究

(申请清华大学工学博士学位论文)

培 养 单 位： 计算机科学与技术系

学　　科： 计算机科学与技术

研　究　生： 叶 子 逸

指 导 教 师： 刘 奕 群　教　授

二〇二五年六月

基于脑机接口的信息检索技术研究

叶子逸

Brain-Computer Interface for Information Retrieval

Dissertation submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Engineering

in

Computer Science and Technology

by

Ye Ziyi

Dissertation Supervisor: Professor Liu Yiqun

June, 2025

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

黄萱菁	教授	复旦大学
崔鹏	副教授	清华大学

答辩委员会名单

主席	张民	教授	哈尔滨工业大学（深圳）
委员	杜军平	教授	北京邮电大学
	马少平	教授	清华大学
	刘奕群	教授	清华大学
	艾清遥	副教授	清华大学
	崔鹏	副教授	清华大学
秘书	郭志强	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》及上级教育主管部门具体要求，向国家图书馆报送相应的学位论文。

本人保证遵守上述规定。

作者签名：叶子逸　　导师签名：文峰祥

日　　期：2025.5.20　　日　　期：2025.5.20

摘 要

在信息泛滥的大数据时代，互联网和信息检索系统成为了我们获取多模态信息的主要途径。这些信息在我们的大脑深处通过神经元间的化学和电信号传递，进而形成思维与感知。然而，当前的信息检索交互范式主要依赖于用户的显式输入（例如文本查询）和隐式反馈（例如点击行为）等传统媒介，这使得在信息检索系统难以充分理解用户大脑中真实的认知过程、信息需求以及反馈信号。

为了应对上述问题，本文研究了基于脑机接口的信息检索系统技术作为一种新的信息交互范式，通过直接解码用户的脑信号来实现更高效的信息交互。本文从用户认知过程理解、用户信息需求解码和用户反馈建模三个方面探讨了脑机接口为信息检索系统带来的新机遇，并开展了以下研究：（1）在用户认知过程理解方面，本文利用脑信号能够直接反映用户心理活动和能捕捉细粒度认知状态的特性，分别构建了基于脑拓扑结构的用户满意度检测模型和细粒度用户阅读过程理解模型。这些模型能够从脑信号中提取与用户认知状态相关的信息，从而实现引入脑机信号的用户认知模型构建，并助力信息检索系统的界面设计和模型构建。（2）在用户信息需求解码方面，传统的脑信号解码方法需要预定义解码的语义目标的集合，不能在开放的信息检索场景下直接应用。因此，本文提出了一种生成式的脑信号解码方法。该方法通过解码脑信号，驱动大语言模型生成开放词表下的语言内容，以表征用户大脑中的语义信息。本文进一步将该方法应用到了信息检索场景下的查询扩展任务当中。实验发现该方法能提高文档排序的准确性，并且在用户意图不明确的歧义查询中表现更为优异。（3）在用户反馈建模方面，本文针对事实性/非事实性搜索任务的不同特性，构建了基于脑信号的用户反馈建模技术。在事实性搜索场景中，本文设计了一种用户意图建模方法，该方法在缺乏传统反馈信号（如点击）的情况下，实现了更高效的相关性反馈和文档重排序。在非事实性搜索场景下，本文开发了在不同搜索场景下融合脑信号和传统反馈信号的算法，并在多个相关性反馈任务上取得了更好的性能。

为支撑以上研究，本文结合用户研究、系统与算法设计以及实验验证的方法，构建了一系列相关的数据集、原型系统和开源算法。这些成果展示了脑机接口技术在信息检索系统中的应用潜力，为进一步探索脑机接口在信息检索技术中的应用以及发展下一代人与信息系统的交互范式奠定了基础。

关键词：信息检索；脑机接口；用户建模；神经信号模型；语言模型

Abstract

In the era of big data and information overload, the internet and information retrieval systems have become our primary means of accessing multi-modal information. This information is transmitted through chemical and electrical signals between neurons, generating thoughts and perceptions in our brains. However, existing information retrieval interaction paradigms rely on users' explicit inputs (such as text queries) and implicit feedback (such as click behaviors), making it challenging to fully capture users' cognitive processes, information needs, and feedback signals related to the brain's cognitive functions during the information retrieval process.

To address these challenges, this dissertation investigates the potential of applying Brain-Computer Interface (BCI) to information retrieval systems. It explores the new opportunities BCI brings to information retrieval systems from three aspects: understanding users' cognitive processes, decoding users' information needs, and modeling user feedback. The research includes the following: (1) For understanding users' cognitive processes, brain signals can inherently reflect users' mental activities and be used to capture users' fine-grained cognitive states. This dissertation constructs a user satisfaction detection model based on brain topology and a framework to detect fine-grained user states in the reading process. These models facilitate the development of user cognitive models by integrating brain signals and enhance the interaction capabilities of information retrieval systems, thereby guiding interface design and model construction. (2) For decoding users' information needs, traditional methods treat brain decoding as a classification problem from a set of predefined semantic candidates, which cannot be directly applied to open-world information retrieval scenarios. Therefore, we propose a generative brain signal decoding method. This method decodes brain signals to drive a large language model to generate open-vocabulary language content aligned with the semantic information in the user's brain. This dissertation further applies this method to the query expansion task in information retrieval scenarios, significantly improving the accuracy of document ranking, especially in ambiguous queries where user intent is unclear. (3) For modeling user feedback, this dissertation addresses the limitations of traditional signals in complex search scenarios and the differing characteristics of factual and non-factual search tasks, and develops a user feedback modeling technology based on brain signals.

In factual search scenarios, the dissertation designs a user intent modeling method that achieves more efficient relevance feedback and document re-ranking in the absence of traditional feedback signals like clicks. In non-factual search scenarios, the dissertation develops algorithms that integrate brain signals with traditional feedback signals across different search contexts, achieving better performance in multiple relevance feedback tasks.

To support the above research, this dissertation employs a combination of user studies, system/algorithm design, and experimental validation to construct a series of related datasets, prototype systems, and open-source algorithms. These results demonstrate the potential of BCI technology in information retrieval systems, laying the foundation for further exploration of BCI applications in information technology and the development of next-generation human-information system interaction paradigms.

Keywords: Information Retrieval; Brain-Computer Interface; User Modeling; Neural Signal Model; Language Model

目 录

插图和附表清单

符号和缩略语说明

BCI	脑机接口，英文全称为 Brain-Computer Interface
fMRI	功能核磁共振成像，英文全称为 functional Magnetic Resonance Imaging
fNIRS	功能近红外光谱，英文全称为 functional Near Infrared Spectroscopy
IR	信息检索，英文全称为 Information Retrieval
LLM	大语言模型，英文全称为 Large Language Model
ECoG	脑皮层电图，英文为 Electrocorticogram
EEG	脑电图，简称脑电，英文为 Electroencephalogram
SERP	搜索引擎结果页面，英文全称为 Search Engine Result Page
ROI	感兴趣区域，英文全称为 Region of Interest
ERP	事件相关电位，英文全称为 Event Related Potential
SVM	支持向量机，英文全称为 Support Vector Machine
BTA	脑拓扑结构自适应网络，英文全称为 Brain Topography Adaptive Network
DT	决策树，英文全称为 Decision Tree
CNN	卷积神经网络，英文全称为 Convolutional Neural Network
GCN	图卷积网络，英文全称为 Graph Convolutional Network
SHAP	SHapley 特征解释，英文全称为 SHapley Additive exPlanations
RNN	循环神经网络，英文全称为 Recurrent Neural Network
MLP	多层感知机，英文全称为 Multi-Layer Perceptron
GBDT	梯度提升决策树，英文全称为 Gradient Boosting Decision Tree
CRF	条件随机场，英文全称为 Conditional Random Field
AUC	曲线下面积，英文全称为 Area Under Curve
MRR	平均倒数排名，英文全称为 Mean Reciprocal Rank
ECG	心电图，英文为 Electrocardiogram
NDCG	归一化折损累积增益，英文全称为 Normalized Discounted Cumulative Gain
MAP	平均精度均值，英文全称为 Mean Average Precision
LM	语言模型，英文全称为 Language Model
LR	逻辑回归，英文全称为 Logistic Regression

FM	因子分解机，英文全称为 Factorization Machine
BLEU	双语评估替补，英文全称为 Bilingual Evaluation Understudy
ROUGE	面向召回的评估替补，英文全称为 Recall-Oriented Understudy for Gisting Evaluation
WER	词错误率，英文全称为 Word Error Rate
FDR	错误发现率，英文全称为 False Discovery Rate
IDF	逆文档频率，英文全称为 Inverse Document Frequency
ICTF	逆语料库词频，英文全称为 Inverse Corpus Term Frequency
DE	微分熵，英文全称为 Differential Entropy
SSVEP	稳态视觉诱发电位，英文全称为 Steady-State Visual Evoked Potential
BOLD	血氧水平依赖，英文全称为 Blood Oxygen Level Dependent
TMS	经颅磁刺激，英文全称为 Transcranial Magnetic Stimulation
AG	角回，英文全称为 Angular Gyrus
AC	听觉皮层，英文全称为 Auditory Cortex
PFC	前额叶皮层，英文全称为 Prefrontal Cortex
PrCu	楔前叶，英文为 Precuneus
MEG	脑磁图，英文为 Magnetoencephalography
IRF	交互式相关性反馈，英文全称为 Interactive Relevance Feedback
RRF	回顾式相关性反馈，英文全称为 Retrospective Relevance Feedback

第 1 章　引言

1.1　研究背景与研究问题

21 世纪被认为是一个信息时代，信息技术的飞速发展正深刻改变着人类社会的方方面面。到 2025 年，互联网的数据增长量达到了 180 泽比特①，这些数据不仅包括文本和图像，还涵盖了视频、音频、传感器数据以及各种复杂结构化和非结构化数据。然而，与之相比，人脑的记忆容量仅为 0.0025 泽比特[1]，这种数据的爆炸式增长已远超人类个体的信息处理能力上限（见图 1.1）。为应对这一挑战，信息检索 (Information Retrieval，IR) 技术应运而生，其核心任务是从海量信息中高效筛选或生成满足用户需求的内容。例如，谷歌 (Google)、百度 (Baidu)、亚马逊 (Amazon) 等公司开发的搜索引擎利用复杂的算法从庞大的数据集中提取出与用户查询最相关的信息，并通过分析用户的搜索行为、点击历史以及社交媒体活动等多维度的数据，实现个性化信息分发与内容优化。Siri 等智能助手则结合信息检索与自然语言处理技术，理解并响应用户指令，提供从信息查询到事务管理的多样化服务。

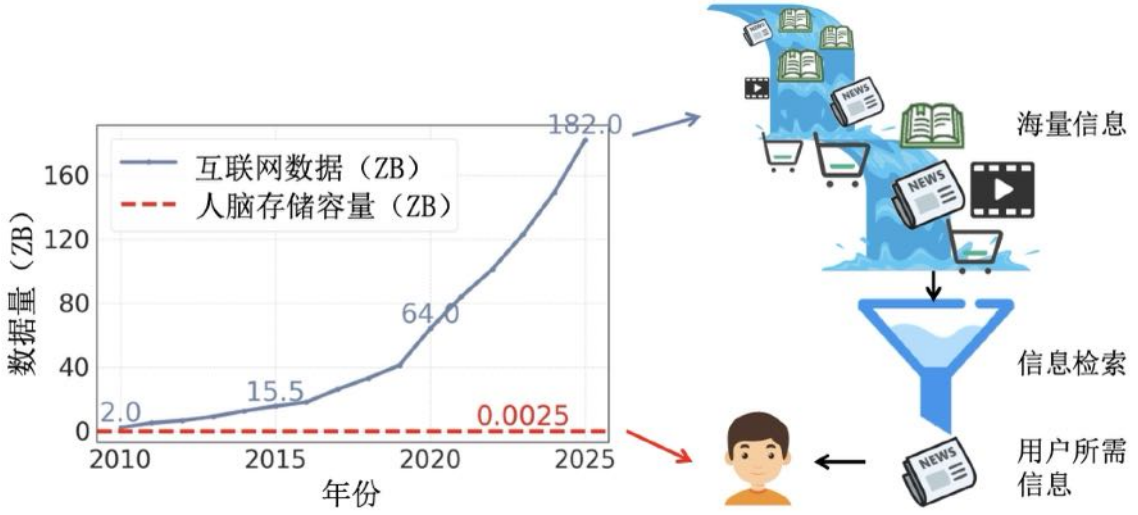
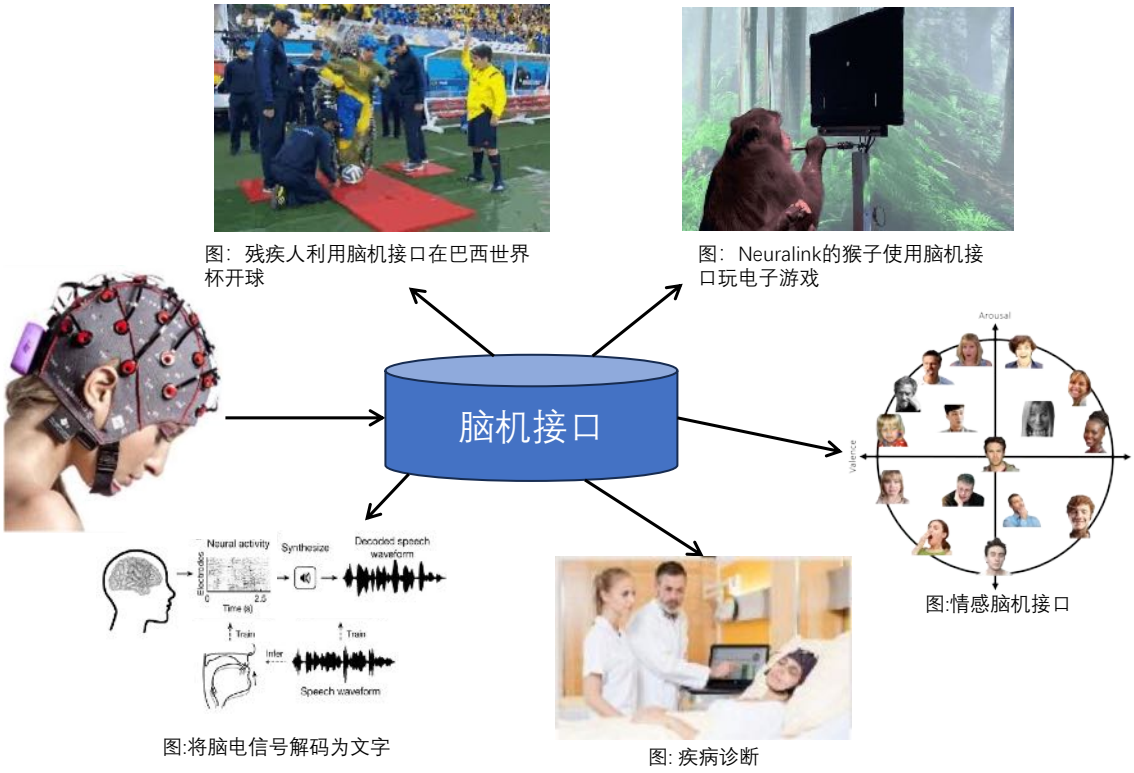


图 1.1　信息检索承担了从海量信息中筛选或生成用户所需要的内容的角色

随着这些信息检索系统日趋智能，信息检索系统亦逐步成为辅助乃至部分替代人类思考、分析、评估和记忆功能的工具[2]。Sparrow et al.[3] 的研究进一步指出，信息检索系统的普及甚至改变了人类的记忆模式：人们倾向于记住如何获取信息（即信息索引），而非信息本身。信息检索技术的进步模糊了人类内部知识与外部信息存储的界限，促使我们重新审视信息传递的本质。在物理层面上，信息是光

① https://explodingtopics.com/blog/data-generated-per-day



图：残疾人利用脑机接口在巴西世界杯开球

图：Neuralink的猴子使用脑机接口玩电子游戏

图:情感脑机接口

图:将脑电信号解码为文字

图: 疾病诊断

图 1.2　脑机接口的常见应用场景示意图

纤、电缆中传输的电信号；而在生物层面，信息则是大脑中神经元之间传递的化学和电信号。文字和图像只是信息传递的媒介，它们作为符号系统，通过人类的解读成为有用的“信息”。在这个过程中，人与信息检索系统交互的效率很大程度上受限于这些媒介的表达能力。然而，这些传统媒介的表达能力存在固有局限。例如，用户常难以用精确的查询词描述其复杂信息需求[4]，其对系统反馈的真实体验也难以被准确捕捉与表达[5]。实际上，信息的核心功能并不依赖于这些媒介的形式。例如，在科幻故事《三体》中，三体人通过脑电波来直接进行思维的交流。基于脑电波的信息交流虽属畅想，却也启发我们：在信息技术飞速发展的今日，从信息检索的根本使命出发，探索超越传统媒介的、更直接高效的信息交互途径，已成为一项亟待探索的重要课题。具体而言，若能直接采集用户大脑中的信息，我们能否更精确地洞察用户需求、捕捉用户反馈，进而实现人与信息系统之间更为高效的沟通？

这一设想引向了脑机接口 (Brain-Computer Interface, BCI) 技术这一前沿领域[6-7]。脑机接口作为一种在人脑与外部设备间直接建立信息通路的革命性技术（如图 1.2所示），为实现人与信息系统之间更直接、高效的交互提供了全新可能[8]。这种技术将采集的人脑的神经活动信号解码成计算机能理解的信号和指令，从而实现人与机器系统的交互。近年来，随着神经科学、人工智能等领域的进步，各类

脑机接口设备在技术成熟度和应用范围上都取得了显著进展。尽管通过脑电波直接进行思维交流的场景与现阶段的脑机接口技术还有很大差距，但将脑机接口技术应用于信息检索系统，已成为一个极具潜力的研究方向。这不仅有望帮助信息系统更深入地理解和建模用户认知过程，还可能辅助系统根据用户实时心理活动动态分发和推荐内容，为构建更高效的人机交互范式开辟新径。

在最典型的信息检索场景中，用户向信息检索系统输入查询词，信息检索系统则会根据该查询词进行匹配，以返回相关的信息。在这样的基本交互范式之下，信息检索系统有着三个重要的研究目标：第一个研究目标是**用户认知过程理解**，用户在信息检索系统中的认知过程是复杂的，包括信息需求产生、信息需求表达、信息需求匹配等多个阶段。信息检索系统需要对用户认知过程进行理解，以建模用户行为并动态地适应用户需求。第二个研究目标是**用户信息需求解码**，传统上，用户只能以文本的形式来表达信息需求。然而，用户在信息检索系统中的文本输入经常是简短、不明确、甚至有歧义的。实际上，用户通常都不是信息检索系统的专家，构建适合于信息检索系统进行精准信息匹配的查询词对于用户而言难度很大。因此，如何从不精确的查询词和隐式的用户行为中解码出真实的用户的信息需求也是一个重要的研究问题。第三个研究目标是**用户反馈建模**，用户在信息检索系统中的反馈通常是隐式且包含噪声的，信息系统需要在这样的复杂环境中去理解用户偏好和构建用户反馈模型。然而，受限于交互范式，现有的信息检索系统一般是客观且静态地根据用户的行为返回交互结果，缺乏对于用户的认知过程和真实意图的理解，从而很难动态地去适应用户的需求。尽管信息检索技术的算法不断进步，其应用范围也从搜索引擎扩展到短视频推荐系统、聊天机器人、智能助手等多个场景，但这些系统仍无法完全摆脱现有的交互范式，无法实现与用户心理活动的直接交互。因此，信息检索系统距离真正成为高效的认知辅助工具乃至“人脑的外接设备”[2]还有很长的路要走。

图 1.3展示了一个传统的信息检索系统交互通路和基于脑机接口的潜在增强方式。在传统交互通路（蓝线）中，用户的信息需求和真实状态都很难为信息检索系统所获取。脑机接口（红线）则提供了一种新的交互通路，通过采集和分析人脑的神经活动信号，脑机接口有潜力去直接理解用户在信息检索系统中的认知过程，乃至解码用户的意图和信息需求，并对用户的反馈和状态进行更高效精准的建模，从而实现全新的交互环路构建。在引入脑机接口的信息检索范式中，本文认为存在以下几个方面的挑战：

第一个挑战是**引入脑机信号的用户认知模型构建**。脑机信号与传统的信息系统信号不同，它是一种难以解释的多变量时序信号，其中每个变量对应大脑不同空

图 1.3　基于脑机接口的信息检索系统研究范式

间位置的数据采集点，并可能涉及不同的脑功能区。传统信息检索系统缺乏利用此类信号的成熟模式，因此如何发挥脑机信号的优势，基于该信号构建理解用户认知过程的模型是首要问题。本文认为，脑机信号相较于传统用户行为信号，主要具备两大优势：一方面，脑机信号是一种和人类大脑的认知活动直接相关的反馈，而非基于行为的反馈。信息检索系统中的传统用户信号，包括用户主动输入的查询词、点击行为、在交互内容上的驻留时长等信息，都是用户认知过程的一种间接的表达。在很多场景下，这些信号都是有偏的，例如，用户点击一个文档，可能是因为该文档的标题吸引了他，而并非因为文档内容本身能够匹配用户的信息需求，即所谓的“坏点击”或者“标题党”[9]。脑机信号能够直接反映用户的认知状态，从而在某些传统信号存在偏差或噪声的情况下展现出潜在优势。然而，由于脑机信号采集自不同功能的脑区，如何有效地建模其空间和功能连接关系，以充分发挥其在反映用户满意度等认知状态方面的潜力，仍然是一个重要的挑战。另一方面，脑机信号可以提供更细粒度的反馈。脑机信号的时序特性可以区分用户在不同时间点的状态。举例而言，当用户点击了一个文档，通常会认为用户对这个文档是感兴趣的，其文档内容很可能是相关的。脑机信号作为一个时序信号，可以通过时序拆分对应到文档的不同部分。因此，脑机信号可以被用于估计一篇被点击的文档的哪些部分是相关的，哪些部分则是不相关的，甚至可以对文档相关的内在原因做出更深入的分析，从而为理解用户细粒度的认知过程提供更好的指导。然而，如何利用这样的时序特性来设计用户实验，并构建细粒度的用户认知模型也是一个重要的挑战。

第二个挑战是**信息获取场景下的脑信号解码技术**。脑信号本身信噪比较低，如

何从复杂的信号中提取与当前信息获取任务相关的有用信息至关重要。传统上，脑信号的分析涉及滤波、特征提取、模式匹配等步骤[10-11]，但这些方法通常需要大量的先验知识，并且难以处理复杂的多维度的脑信号。近年来，随着深度学习的快速发展，基于深度学习的脑信号解码方法取得了显著进展，这些方法通过端到端的方式学习从脑信号中提取特征的模型。但受限于脑信号中广泛存在的噪声，从复杂的、低信噪比的脑信号中提取出有用信息仍是一个悬而未决的挑战。此外，传统上的脑信号解码一般是作为一个分类任务，即给定一些预定义的标签，通过脑信号来解码和匹配其中一个或者多个标签。例如，在情绪识别任务中，会定义唤起（arousal）和效价（valence）两个维度，通过脑信号来解码和匹配这两个维度；在动作解码任务中，脑机接口的目标是辨别用户当前是使用左手还是右手进行动作。而真实信息检索系统中的用户信息需求和用户反馈是多维度的，因此很难预定义一个有限的集合来框定用户的所有信息需求和行为。因此，如何设计开放式的、不受限于有限类别标签的脑信号解码模型是一个重要的挑战。

第三个挑战是**信息交互环路中的信息系统反馈机制**。信息系统如何有效利用脑机接口解码的用户状态信息进行动态反馈，包括反馈时机决策与反馈机制构建，是另一项关键挑战。传统上，信息检索系统中的反馈机制基于用户的显式标注或者隐式行为，例如，用户对某个文档表示满意或者点击了某个文档，则认为该文档是相关的，那么信息系统就会倾向于推荐类似的文档。然而，显式标注往往是比较稀疏的，而隐式行为只是用户意图的间接表达，在建模用户反馈方面并不准确。结合第一个挑战中提到的脑机信号的特性，本文认为在不同的信息检索情境下甄别不同信号的有用性，并将脑机接口信号作为一种补充，对于信息系统反馈机制的构建至关重要。然而，基于脑机接口的反馈具有特异性和一般性[12-13]。例如，有些场景下用户反馈是关于某个查询的通用反馈，可以作为信息系统为其他用户分发内容的参考；而有些场景下用户反馈是个性化的，一个用户的反馈可能只适用于该用户。因此，在此复杂交互情境下，如何最大化脑信号的效用仍有待深入探索。另一方面，如何判断是否基于脑信号反馈进行动态响应，并更好地整合不同信号所提供的反馈，也是一个亟待解决的挑战。

1.2　研究框架与内容

针对上文阐述的研究目标与核心挑战，本文聚焦于基于脑机接口的新型信息检索范式，系统性地开展了以下三个模块的研究工作。

研究整体框架如图 1.4所示。具体地，本文的研究包含以下三个方面：

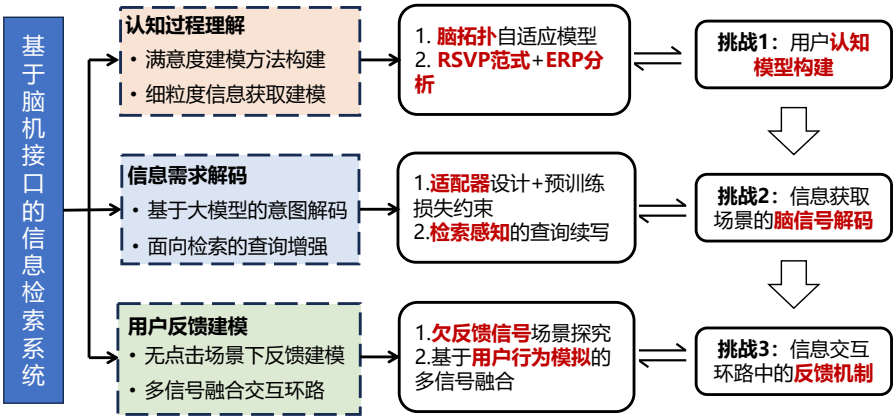


图 1.4　本文研究工作的整体架构

（1）用户认知过程理解　脑信号与认知活动直接相关和能提供细粒度反馈信息的特性为构建引入脑信号的用户认知模型和实现更有效的用户认知过程理解提供了可能。本文从两个方面开展研究。利用脑信号与认知活动直接相关的特性，本文探究了其在用户对信息系统满意度检测中的应用。用户满意度是信息检索系统的最重要指标之一，通过对用户的满意度进行检测，可以更好地理解用户，并为信息检索系统设计提供指导。然而，满意度作为一种复杂认知状态，涉及情感、注意力和认知负荷等多维度脑功能及其协同作用，从脑信号中准确提取满意度相关特征极具挑战性，尤其在建模脑功能区动态连接方面，现有方法仍显不足。为解决这一问题，本文提出了脑拓扑结构自适应的网络（Brain Topography Adaptive Network，BTA），用于从脑信号中提取用户的满意度状态。实验结果表明，BTA 在满意度建模任务上超越了现有的拓扑不变网络和一系列现有的拓扑感知的网络。同时，基于该模型预测的满意度信号反馈能够有效提高系统的交互能力。

另一方面，针对脑信号能提供细粒度时序反馈的特点，本文开展了用户阅读理解过程中的细粒度脑信号分析。本文设计了一个用户实验，利用脑电图（Electroencephalogram，EEG）设备记录了用户在阅读理解任务中的词和句子级别脑信号，对用户在信息检索过程中的认知负荷、语义主题理解和推理加工等认知活动进行了深入的分析。进一步，本文设计了基于脑电的阅读理解模型框架（Unified framework for EEG-based Reading Comprehension Modeling，UERCM），用于两个用户阅读理解状态检测的任务：答案句分类和答案词提取。实验结果表明，能够满足用户信息需求和不能满足用户信息需求的内容，会引发不同的神经反应，而大脑信号作为这样的反馈信号，可以被有效地利用于词、句级别的用户认知状态检测。

通过以上两方面的研究，本文验证了脑信号在用户认知过程理解和交互式信息系统中的潜力。这不仅为改进现有的信息检索系统提供了新思路，也为未来构建基于脑机接口的智能信息系统奠定了基础。

（2）用户信息需求解码　用户的信息需求的理解是信息检索系统中最重要的研究目标之一。传统上，信息需求基于用户的查询词来体现。然而，由用户提交的查询词却常常由于其过于简短、存在歧义等问题而表现不佳。本文通过直接解码用户的脑信号，来实现更细致地捕捉用户在查询过程中的认知状态和意图，从而筛选和生成更符合用户需求的信息结果。为突破传统脑信号解码模型在开放词表能力上的局限，本文提出了一种新的自回归生成方法，直接将从功能性磁共振成像（functional Magnetic Resonance Imaging，fMRI）解码的表示作为输入，驱动大语言模型（Large Language Model，LLM）生成语言内容来表征用户大脑中包含的语义信息。通过端到端的训练和无监督损失约束的技术，本文实现了从低信噪比的脑信号中提取信息并生成开放候选集合的语义内容的效果，其性能甚至超越已有的一些基于有限候选集合做选择的方法，使得在复杂的信息检索场景下实现用户语义解码成为可能。

进一步，本文将这样的方法应用于信息检索系统中的查询扩展任务，即基于原始查询文本和用户的大脑信号来生成新的查询文本。实验结果表明，该方法生成的查询文本不仅语义上更准确，还显著提高了文档排序的准确性，尤其在处理模糊查询时，性能提升更加明显。

（3）用户反馈建模　在信息检索系统中，准确建模用户反馈至关重要，而传统上依赖点击行为等隐式反馈信号的方式在复杂的搜索场景中存在局限性。为解决这些问题，本文引入了脑电信号作为新的反馈信号来源。首先，本文探讨了事实性搜索的场景。随着搜索技术的进步，用户在执行事实性搜索任务时，一个越来越普遍的现象是无需点击任何结果，而是直接通过搜索引擎提供的总结性内容或文档摘要来完成检索任务。因此，一个没有被用户点击的文档（无点击文档）很多时候并不代表该文档不好。传统依赖点击等显式行为建模搜索结果相关性的方法，在无点击或复杂搜索场景下存在明显局限。通过采集和建模用户的脑信号，本文能够在用户无需点击搜索结果的情况下获取搜索结果的相关性反馈。本文设计了一个用户实验，采集用户在无点击场景下的脑信号，并利用脑信号估计用户对搜索结果的相关性判断。实验结果表明，不同相关程度的无点击搜索结果会引发显著不同的脑电信号。考虑到用户反馈的特异性和一般性，本文进一步设计了两种重排序方法：个性化方法和通用意图建模方法。这两种方法分别利用个体和群体估计的相关性对搜索结果重排序，其中通用意图建模方法相比不依赖用户反馈建模的文档排序基线提升了 45.2%（基于评价指标 NDCG@1）。

进一步，本文研究了在非事实性搜索的场景下如何融合传统的点击等信号和脑机信号来实现更高效的用户反馈建模。相比事实性问题，非事实性搜索问题通

常没有明确的答案，因此用户在搜索过程中可能需要和多个搜索结果进行包括点击在内的交互，并从中综合出自己的答案。本文开展了用户实验和采集了相关数据，并设计了一个通用的融合多种信号和面向不同任务的相关性反馈框架。实验结果显示，本文所提出的框架可以高效地结合脑机接口信号、伪相关信号和隐式信号的相关性反馈。其中，引入脑机信号的相关性反馈显著提升了文档重排序的性能。与未使用脑机信号的相关性反馈相比，基于 NDCG@1 的结果提升了 17%。实验分析还发现，在用户点击信号缺失或存在噪音的情况下，脑信号带来的性能增益更加显著。这表明，在不同的搜索场景中，脑信号的实际效用可能存在差异。因此，本文进一步提出了一种自适应融合不同反馈信号的方法，通过协助信息系统更好地理解反馈信号在不同搜索场景下的重要性提升了其在文档重排序任务上的性能。

1.3　本文结构安排

本文共分为六章，具体安排如下：

第一章为引言，介绍本文的研究背景和研究问题，阐述了传统信息检索系统在“用户认知过程理解”、“用户信息需求解码”和“用户反馈建模”三个方面的研究目标，提出了基于脑机接口的信息检索系统研究范式，并概况了该范式面临的主要挑战和本文的主要研究内容和贡献。

第二章为相关工作，系统综述信息检索领域的发展现状和通用范式，并指出了传统信息检索范式的研究现状与不足，以及脑机接口技术的基本原理和在不同领域的应用进展。

第三章重点介绍基于脑电信号的用户认知过程理解方法。本章内容包括构建基于脑拓扑结构的用户满意度检测模型和细粒度用户认知过程理解模型，以及一个针对用户在信息检索过程中的阅读理解进行的用户研究。本章相关成果发表于 CCF A 类会议 TheWebConf（旧称 WWW）2022 长文[14]、CCF A 类会议 Multimedia 2022 长文[15]和 CCF A 类会议 Multimedia 2024 的 Workshop 论文[16]。

第四章详细阐述融合脑机信号的信息需求解码方法，尤其是如何设计在开放环境下，利用 fMRI 信号构建基于语言的信息需求解码模型并应用于下游的查询改写和文档排序任务。本章相关成果发表于 SCI 一区期刊 Nature Communications Biology[17]和 CCF A 类会议 Multimedia 2024 长文[18]。

第五章介绍基于用户神经反馈建模的检索结果优化方法，本章分别在事实性问题和非事实性问题的场景下探究了脑信号作为用户反馈的潜力，研究内容包括反馈信号处理、优化策略设计和实验验证等。本章相关成果发表于 CCF A 类会议

SIGIR 2022 长文[19]和 CCF A 类期刊 TOIS[20]。

第六章总结全文工作，并对未来研究方向进行展望。

本文作者为全部所提到工作的唯一第一作者。所有工作的代码和用户实验数据均已开源①：包括三个数据集（第3.4节的 UERCM 数据集②，第5.3节的事实性搜索相关性反馈数据集③和第5.4节的非事实性搜索相关性反馈数据集④），以及所有方法的代码（如第3.3节的 BTA 模型⑤、第3.4节的 UERCM 模型⑥，第4.3节的 BrainLLM 模型⑦和第4.4节的 Brain-Aug 模型⑧等）。

① 参见https://yeziyi1998.github.io/。
② https://github.com/yeziyi1998/UERCM
③ https://github.com/yeziyi1998/Search-Brainwave
④ https://github.com/THUIR/Brain-Relevance-Feedback
⑤ https://github.com/yeziyi1998/DL4EEG
⑥ https://github.com/yeziyi1998/UERCM
⑦ https://github.com/YeZiyi1998/Brain-language-generation
⑧ https://github.com/YeZiyi1998/Brain-Query-Augmentation

第 2 章　国内外研究现状

信息检索系统和脑机接口都是学术界和产业界的重要研究及应用课题。信息检索指的是从大量非结构化或半结构化的数据中寻找相关信息的过程，具体包括信息的搜索、生成、推荐等。在学术研究方面，信息检索系统自从谷歌（Google）诞生以来一直是研究的热点，成果广泛发表在人工智能和数据挖掘的国际顶级学术会议，包括国际互联网大会（The Web Conference），数据挖掘大会（Knowledge Discovery and Data Mining，KDD），信息检索大会（Special Interest Group on Information Retrieval，SIGIR）和期刊《信息系统汇刊》（Transactions on Information Systems，TOIS），《IEEE 知识与数据工程汇刊》（IEEE Transactions on Knowledge and Data Engineering，TKDE）等。近年来，随着以短视频为代表的富媒体推荐系统，以及以 ChatGPT 为代表的对话式信息获取系统的崛起，信息检索系统也面临着新的挑战和机遇。

而脑机接口一词则是 1973 年由加州大学洛杉矶分校的计算机科学家雅克·维达尔（Jacques Vidal）首次提出，他认为有朝一日，大脑中的信号可以被直接用来控制假肢等外部设备。此后，研究人员对这样的想法进行了大量的研究，并取得了显著的进展，相关成果广泛发表在《自然》（Nature）及其子刊,《IEEE 神经系统与康复工程汇刊》（IEEE Transactions on Neural Systems and Rehabilitation Engineering），《神经工程期刊》（Journal of Neural Engineering）等学术期刊上。

在产业应用上，信息检索系统是当前人工智能商业化最成功的技术之一，相关技术也都应用到了谷歌、百度、微软、阿里巴巴、字节跳动等科技公司的产品中。而脑机接口技术则主要应用在医疗、康复、游戏、教育等领域，包括 Neuralink、BrainCo、DeepMind、脑虎科技等公司都在脑机接口领域有大量的投入。然而，在商业化方面，脑机接口目前主要在癫痫、脊髓损伤等相关疾病的治疗和康复中发挥一定作用。在更广泛的现实场景中，脑机接口的应用仍处于相对早期的阶段。

本章将首先介绍信息检索系统的交互范式，尤其是交互范式的更迭和传统的交互信号可能存在的不足。然后介绍神经科学与脑机接口的基本概念和相关研究。最后将介绍基于脑机接口设备开展的，与信息检索系统具有潜在关联的已有研究。后续章节（第 3 至第 5 章）在讨论具体技术方案时，会针对性地对相关文献进行更细致的阐述。

2.1　信息检索的交互范式

信息检索（Information Retrieval，IR）系统在人类的信息获取过程中扮演着至关重要的角色。随着互联网和大数据的快速发展，用户与信息检索系统的交互方式也在不断演进，交互范式的更迭对系统性能和用户体验产生了深远的影响。交互范式的更迭主要有三个不同的出发点[21]：“技术”、“用户”和“技术 + 用户”。

从技术角度出发，随着机器学习、深度学习和自然语言处理等技术的发展，信息检索系统中的语义分析流程在不断更新。这使得系统可以加深对被检索文档和用户提问的理解和知识表示，从而也更迭了用户在信息检索系统中的交互方式。具体而言，信息检索系统从仅基于关键词匹配的布尔检索模型发展到基于向量空间模型的检索模型，再到基于深度学习的检索模型，这些模型在检索效果和用户体验上都有了显著的提升。布尔检索模型作为最早期的信息检索模型，主要是基于查询的关键词匹配（与、或、非）结果的布尔表达式来进行文档检索。这样的范式简单直观，但要求用户具备一定的逻辑能力和领域知识，且查询结果的相关性和完整性难以保证。为了克服布尔检索模型的不足，向量空间模型被提出。这些模型（例如 BM25[22], TF-IDF[23] 等）将文档和查询表示为多维语义向量，通过计算向量之间的相似度（如余弦相似度）来衡量文档与查询的相关性。例如，BM25 算法将文档表征为在词向量空间中的向量，并平衡了词频和文档长度对相关性评分的影响，其简单有效的计算公式和良好的性能使其在许多信息检索系统中得到了广泛应用。向量空间模型使得用户在设计检索词时，只需要考虑检索词是否会出现在目标文档中，而系统会自动计算每个词的权重和相关性评分，这大大降低了用户的使用门槛。近年来，随着深度学习的发展和预训练语言模型的兴起，基于神经网络的检索模型将向量空间表征从稀疏向量推广到稠密向量，并逐渐成为主流。这些模型通过学习大规模语料库中的数据分布，能够更好地捕捉文档和查询之间的复杂关系，在这样的检索范式下，用户输入的检索词甚至不需要和目标文档的词语完全匹配，系统也能够通过语义理解来返回相关性较高的结果。例如，BERT 模型[24]通过预训练学习到丰富的语言知识和语言表征能力，从而在信息检索任务中取得了显著的性能提升。自 2022 年以来，随着 ChatGPT 的兴起，基于大模型的信息检索系统逐渐成为研究热点[25]。通过在信息检索任务的各个阶段（如文档召回、文档精排、查询重写和文档理解）中应用大模型的语言和推理能力，这些系统的性能显著提升。ChatGPT 的普及也推动了检索范式的演化，使用户能够通过自然语言交互而非精心设计的查询词来获取信息。

从用户角度出发，信息检索系统需要从用户侧获取更多有用信息，理解用户的点击、查询等行为和情感、认知等用户因素。在此基础上，研究人员要关注系统

用户界面的优化与技术模型的设计。用户在信息检索系统中的浏览模式将影响检索结果的展示和检索页面的布局。通过用户实验等方法，研究者揭示了用户在信息检索过程中的位置偏置[26]，认知偏差[27]，信任偏差[28]，标题党偏差[9]等现象。其中，位置偏差是指用户在浏览搜索结果时，通常会首先注意到页面顶部的内容，因此上方位置的结果获得更多的点击，即使它们并不一定是最相关的。而认知偏差是指用户在处理信息时，由于其认知特性和经验而产生的系统性偏差。例如，用户可能会更倾向于选择与其已有知识和信念一致的搜索结果，而忽视与其相悖的信息。而信任偏差是指用户对某些信息源或搜索结果的信任度不同，导致其点击行为受到影响。例如，用户可能会更倾向于点击来自知名网站或品牌的搜索结果，而忽视其他同样相关但不太知名的结果。标题党偏差指的是某些文档使用夸张或误导性的标题来吸引用户点击，而这些标题往往与实际内容的相关性较低。这种偏差会影响用户的点击行为，并可能导致搜索引擎误判内容的实际相关性。通过系统的用户研究，信息系统可以更好地理解用户的真实偏好和用户行为中存在的偏差效应，从而优化系统设计。例如，搜索引擎中经典的点击模型[26]通过分析用户点击行为来评估检索结果的质量，并且引入了位置偏差来分析用户点击行为并进行文档重排。Chen et al.[29]通过用户实验研究了信息系统中的“诱饵效应”，并构建了一个模型来评估现有评价指标受诱饵效应影响的脆弱性。Xie et al.[30]通过实验研究了用户在图像搜索中不同于文本搜索的浏览行为，从而构建了图像搜索下更合理的检索结果质量评价模型。Li et al.[31]通过眼动实验分析了用户在搜索引擎访达页阶段的浏览行为，并基于眼动实验中观测到的位置偏差、认知偏差等现象构建了一个用于计算文档与查询语义相似度的模型。

从“技术 + 用户”角度出发，强调耦合用户意图理解和信息检索系统模型设计，融入用户研究以优化系统。“技术 + 用户”范式的主要研究目标是通过优化技术，弥合算法和系统设计与用户需求和期望之间的差距，进而优化用户体验。从最初的简单查询和关键词匹配进化到复杂的多轮对话、交互式信息获取、推荐系统、智能助手等，信息检索系统既在不断地改变交互范式来提升用户获取信息的效率，也在不断地衍生对应的新兴技术来优化相关范式。例如，Liu et al.[32]从“技术 + 用户”的角度出发探究了交互式信息检索和多轮对话系统的设计。这些系统能够与用户进行连续的交互，逐步澄清用户的意图，并在此过程中不断优化检索结果。在这些系统中，一种经典的技术是相关性反馈[33]：用户在初次查询后对检索结果进行标注（如相关或不相关），系统根据这些标注信息调整检索模型和结果排序。例如，Chen et al.[34]探索了在多轮检索场景中，如何利用用户历史查询和点击行为来优化检索结果。此外，还有一些研究探讨了搜索系统作为一个智能助手

ECoG　　fMRI　　EEG (湿电极)　　EEG (干电极)

图 2.1　常见的脑成像设备示意图

的形态，通过澄清、反问、多轮对话等方式了解用户的具体需求和背景，从而提供更具有针对性的信息和建议。例如，Bi et al.[35]通过在搜索系统中基于用户的负向反馈提出澄清式问题，从而更好地建模用户意图。Liu et al.[36]探索了在法律场景下如何利用机器模型构建代理来实现对话式法律案例检索。另一方面，随着互联网时代的发展，文本之外的场景信息（包括图像、视频、音频等）也逐渐成为信息检索系统的重要组成部分，这些场景下全新的用户需求也为系统设计带来了全新的挑战。例如，Xie et al.[37]通过在图像搜索中引入上下文信息来提升检索效果。Datta et al.[38]探索并利用文本信息和图像信息的交互效应来提升检索效果。Knees et al.[39]概述了音乐检索和推荐的场景下如何利用音乐索引、音乐描述、音乐相似度等特征来提升检索效果，并对音乐检索的下一代交互形态进行了展望。

尽管信息检索系统在交互范式上取得了显著进展，但仍存在一些不足之处。例如，传统交互范式难以深入理解用户意图，缺乏对用户情感、认知等因素的考量，无法有效建模用户行为，以及对用户反馈的响应不够及时。这些不足的根源在于许多场景下用户与信息检索系统的交互信号质量不高[40-41]。例如，用户构建的查询词可能不够有效，用户的点击、查询、浏览等行为背后的原因非常多样，系统难以理解。此外，用户在信息检索系统中的情感、注意力、满意度和认知等心理状态未被系统感知。因此，除了从技术角度推进交互范式的更新外，还需要对交互范式本身和交互信号的获取进行更深入的探索。

2.2　神经科学、脑成像技术与脑机接口

神经科学是研究神经系统及其功能的多学科领域，涵盖从分子和细胞水平到系统和行为层面的广泛研究。其中，大脑作为神经系统的最重要组成部分，其功能和机制一直是研究的热点。在神经科学的研究当中，使用脑成像技术，包括 fMRI、EEG、脑磁图（Magnetoencephalogram，MEG）、正电子发射断层扫描（Positron Emission Tomography，PET）等，来研究大脑的结构和功能是重要的研究手段（图 2.1提供了部分脑成像技术的实景图）。一方面，脑成像技术被广泛应用于研究包括语言在内的不同模态下的人类的视觉、听觉等感知过程。Price[42]总结了 100 篇 fMRI 与人类

语言研究相关的论文，分析了大脑在语音产生、语音理解等语言处理过程中涉及的不同脑区以及激活模式。Beres[43]指出了脑电图高时间分辨率的特点尤其适合研究大脑如何连续处理语言理解，同时介绍了如何基于事件相关电位（Event-Related Potential，ERP）方法研究语言理解与产生的过程。该方法在本文的研究中也有所应用（参见第三章）。另一方面，脑成像技术还被应用于研究决策、注意力和记忆等认知过程[44]。例如，Rissman et al.[45]基于 fMRI 数据的分析提出了记忆编码的分布式模式，并揭示了影响信息编码、维护或检索方式的过程，从而为记忆理论提供了支持。Li et al.[46]和 Cabeza et al.[47]分别用 EEG 和 PET 探索了个体的情景记忆编码和记忆检索。

脑成像技术在神经科学中的应用不仅协助了医学和心理学领域的研究人员开展对大脑的研究，也在工程领域得到延伸并发展了一个新的学科，即脑机接口。脑机接口指的是通过大脑活动的实时监测和解码，建立起大脑与外部设备之间的直接通信通道。根据脑成像技术的不同，脑机接口可以分为侵入式和非侵入式两种方式。表 2.1总结了部分侵入式和非侵入式的脑机接口类型及特性。侵入式技术的脑机接口需要通过开颅手术，将电极植入大脑内的特定区域从而采集相应的电信号。例如，皮层脑电（Electrocorticography，ECoG）和立体定向脑电（Stereoelectroencephalography，SEEG）通过穿透大脑膜中的微电极来获取颅内信号。此外，也有一些方法将电极放置在头皮下方但不在灰质内，这些方法被称为硬膜外 ECoG（也被称为半侵入式脑电）。由于脑脊液的阻隔，这些方法采集的信号的信噪比弱于硬膜内脑电，但对大脑环境的破坏较小。侵入式方法能够提供较高精度的神经信号，但需要手术植入，具有一定的风险[48]。相比侵入式脑机接口，非侵入式脑机接口不会对脑组织造成任何损伤，并由于其安全、经济、易操作等特点，在脑机接口领域得到了更广泛的应用，本文中所有的研究内容都基于非侵入式脑机接口。非侵入式方法涉及的脑信号除了电信号（即 EEG）之外，还包括磁信号（即 MEG），代谢信号（如功能性近红外光谱（functional Near-Infrared Spectroscopy，fNIRS）和 fMRI）等。这些信号各自具有不同的优缺点。例如，脑电信号具有毫秒级的时间分辨，能够实时捕捉脑电活动的快速变化，设备便携且成本较低。但其空间分辨率较低，难以精确定位信号源的具体脑区，且信号容易受到电磁干扰和肌电噪声的影响，信噪比较低。MEG 通过测量神经活动产生的磁场来监测脑功能，具有与 EEG 相似的毫秒级时间分辨率和更高的空间分辨率，但其设备价格昂贵，体积庞大，阻碍了其普及和应用。代谢信号（如 fMRI 和 fNIRS）通过测量与脑活动相关的血氧水平变化来反映神经活动。基于代谢信号的脑机接口能够提供高分辨率的脑部结构图像，但时间分辨率较低（秒级），且由于血氧水平变化具有显著的延迟

效应，难以对快速的神经活动进行及时监测和响应。

表 2.1　部分脑机接口类型及特性总结

类型	技术名称	信号特征	优点	缺点
侵入式	硬膜内脑电（ECoG）	硬膜内电极，高频神经信号	高时空分辨率，高信噪比	需开颅手术，存在感染风险
	立体定向脑电（SEEG）	硬膜内电极，立体定向信号	深部脑区监测，高信噪比	需开颅手术，存在感染风险
	硬膜外脑电（ECoG）	硬膜外电极，有脑脊液阻隔	创伤较小，较高信噪比	微创手术，感染风险较小
非侵入式	脑电图（EEG）[a]	头皮电极	完全无创，毫秒级响应，低成本便携	空间分辨率低，易受干扰
	脑磁图（MEG）	磁场检测，神经电流源定位	高时间分辨率，毫米级空间精度	成本高，需磁屏蔽环境
	功能性核磁共振（fMRI）[b]	血氧代谢监测	毫米级空间分辨率，全脑覆盖	时间分辨率低，成本高，设备体积大
	功能性近红外光谱（fNIRS）	近红外光谱，皮层血流监测	较高空间分辨率，可移动式设备	深度受限，信号质量略差于 fMRI

[a] 本文中第三、第五章所采用的技术。
[b] 本文中第四章所采用的技术。

脑机接口的工作流程通常始于信号采集阶段，通过上述的脑成像技术获取大脑的神经信号。采集到的数据通常是复杂的时间序列，其时空分辨率和不同的脑成像技术有关。例如 EEG 数据的时间分辨率一般为毫秒级，每毫秒的数据通常包含几个到数百个通道的脑电信号；而 fMRI 数据的时间分辨率一般为秒级，每秒的数据可能包含数万个体素的脑血氧变化信号。接下来，这些信号会通过一些预处理来去除不必要的噪声，然后进行特征提取和选择。例如，通过时域、频域或时频域分析（如计算功率谱密度或小波变换）获得关键特征，并可能采用降维技术优化数据表示。随后，这些特征被用于模式匹配或输入到机器学习模型中进行分类任务，从而识别用户的意图或状态。这些解码信息会转化为具体的指令，使得计算机系统可以根据用户反馈进行调整，以提高性能和用户体验。例如本文的第五章将 EEG 信号解码为用户对信息内容的相关性估计，并用来优化信息检索系统的文档排序。本文的第四章将具有更高空间分辨率的 fMRI 信号解码为文本形式的用户意图表征，并基于这些解码内容来进行查询词的改写。

脑机接口技术不仅为神经科学的基础研究提供了新的工具和视角，也为医疗、教育、游戏、娱乐等各类应用场景带来了革命性的改变。在医疗领域，脑机接口可以为协助截肢者或脊髓损伤的患者进行假肢的控制。包括匹兹堡大学[49]，清华大

学[50]等研究机构都成功实现了让患者使用侵入式脑信号直接控制机械假肢的脑机接口，从而恢复一定的自主行动能力。此外，脑机接口也可以被应用于癫痫的检测与控制[6]，抑郁症、焦虑症等精神疾病的诊断与治疗[51]等。近年来，脑机接口被应用于“思想翻译”也是非常引人瞩目的方向，包括语音发声的解码[52]、语言感知的解码[53]、视觉感知的解码[54]，为失语症患者提供了一种全新的交流方式。但受限于现有的脑机接口信号精度，“思想翻译”仍然是一个非常有挑战性的任务。目前，直接将脑信号解码为自然语言的研究尚不成熟，尤其是在使用非侵入式设备时面临更多困难。而在医疗场景之外，脑机接口的研究主要是基于非侵入式脑机接口信号。一方面，非侵入式脑机接口可以被应用于用户状态的检测。例如，脑机接口可以被用于情感识别[55]，从而为游戏、电影等娱乐场景提供更加个性化的体验。此外，在教育领域，脑机接口可以用于监测学生的注意力水平，帮助教师实时调整教学策略，提高学习效果[56]。类似的注意力检测还可以应用于军事领域对士兵的注意力状态进行监控，从而为战场指挥提供决策支持（参见美国兰德公司白皮书[57]）。另一方面，非侵入式脑机接口在一定程度上能够解码用户的意图和指令。尽管实现直接的“思想翻译”仍然具有挑战性，但可以通过某些特定的范式将脑信号转换为各种指令和操作。例如，稳态视觉诱发电位（Steady-State Visual Evoked Potential，SSVEP）范式是一种常用的方法。该范式通过将用户指令集合绑定到屏幕上不同频率和相位的闪烁块上，并利用视觉皮层激发的脑电信号的频率和相位来对应不同的按键操作[11]。当用户注视某个闪烁块时，脑电信号中的特定频率和相位会增强，从而使系统能够识别用户的选择。基于这样的范式，非侵入式脑机接口也被广泛地应用于残障人士的无障碍系统和游戏等场景。

2.3　信息检索系统与脑机接口

尽管信息检索和脑机接口的研究目标不同，但它们都旨在实现用户和信息之间更有效的交互。为了实现这个目标，信息检索系统从信息匹配的算法设计，用户交互的范式设计等方面都进行了大量研究；而脑机接口的研究则侧重于如何从大脑中获取有效的信息并转化为相应的指令或者反馈。值得一提的是，ACM 前主席 Vinton 曾经提出过信息检索系统的一个终极理想：“或许在 2064 年，我可以直接在思考的同时获取信息”[2]。这一愿景为信息检索和脑机接口的结合提供了一个宏伟的目标。目前，结合脑机接口与信息检索的研究相对较少，但有一些相关工作能够为基于脑机接口的信息检索系统研究提供一些启发。这些研究主要可以分为三个方面，基于脑机接口的计算机系统交互，基于脑机接口的信息获取过程的用户研究，以及基于脑机接口的主动式系统构建。

基于脑机接口的计算机系统交互方面的研究主要致力于解决基于手、眼、语音的交互方式不可行的场景，服务于特殊人群或虚拟现实游戏和军事行动等特殊场景。现有研究表明，脑机接口可以用于电脑屏幕上的光标控制[58-59]，这表明其有潜力被应用于控制基于屏幕交互的信息系统。最近，有一项研究[60]首次构建了一个基于脑电输入和 SSVEP 方法实现的信息检索系统原型，该系统能够查询并浏览网页。除了通过脑机接口进行信息检索系统的主动控制之外，脑机接口还可以用于将信息从信息检索系统直接传递给用户，例如利用经颅磁刺激（Transcranial Magnetic Stimulation，TMS）[61]等相关的基于脑机接口的信息输入技术。尽管相关研究仍处于初级阶段，但这些研究为脑机接口在信息检索系统中协助用户交互提供了新的思路。

基于脑机接口的信息获取过程的用户研究主要关注于理解用户在信息系统中的行为的认知基础。近年来，有一些研究者利用脑机接口探索了用户在信息获取过程中的相关性判断[62-63]、信息需求的产生[10]和情感[64]等心理活动在神经系统方面的表现。这为更好地理解用户在使用信息检索系统时的动机和体验提供了见解。例如，Moshfeghi et al.[65,66]利用了 fMRI 高空间分辨率的优势，探索了信息需求产生的过程中的脑拓扑分布。他们揭示了一个标志性区域——后扣带皮层，在信息需求产生的时刻下活动更为显著。该区域被认为是协调大规模脑拓扑网络的关键枢纽。这一系列研究最终提出了“信息需求实现的神经心理学模型”，包括三个相互关联的组成部分：记忆检索、信息流调节和感知。此外，基于 EEG 的脑机接口技术由于其高时间分辨率，广泛应用于捕捉信息检索过程中不同时间尺度上的脑活动表现，尤其是基于 ERP 的方法[67]。例如，Kangassalo et al.[68]研究了在阅读不同特异性单词时激发的神经活动，发现单词的特异性与增强的脑活动有关。这些神经相关性通常在单词呈现后 200 至 800 毫秒之间出现。因此，用户能够在较短的时间窗口内自然地识别单词之间的区别，这有助于他们区分特定文档或构建查询。在另一项研究中，Pinkosova et al.[69]通过 EEG 发现用户在进行相关性判断时表现出分级现象。研究报告了与常见 ERP 成分（P300、N400、P600，即 300ms、400ms、600ms 处的正波、负波和正波）相关的脑电活动在三种相关性等级之间的显著差异。这些差异表明，在分级相关性反馈过程中，用户的多种认知过程以不同的程度参与其中。Ji et al.[70]在上述研究的基础上，将传统网页搜索场景中的过程按照用户认知脑电的差异分为信息需求实现、查询构建、查询提交和相关性判断。这一分类基于用户在信息检索过程中的不同认知阶段，为理解用户的搜索行为提供了新的视角。

基于脑机接口的主动式系统构建主要探讨如何将脑机接口信号作为计算机系

统的额外输入，从而驱动或优化整个信息交互流程。在现有的研究中，脑机接口主要用作相关性预测的信号，以预测用户对查询结果的偏好[62,71]。这些研究共同表明，脑信号可以作为有效的相关性判定指标。例如，Eugster et al.[62]利用脑信号预测词的相关性，并进一步应用于相关词语推荐[72]。而 Allegretti et al.[71]和 Golenia et al.[73]则利用脑信号预测用户对图像刺激的偏好。在和本文最相关的一项研究中，Gwizdka[74]发现，仅使用眼动追踪技术和单通道的 EEG 就可以预测用户在浏览网页时的相关性判断，从而为更轻量级的脑机接口在信息检索系统中的应用提供了可能。除了用于相关性预测之外，脑机接口还被应用于用户的知识水平预测[75]、行为内容预测[70]以及信息需求的产生预测[65]等领域。然而，这些研究大多还处于早期阶段，尚未与真实的信息检索系统的设计充分融合，以实现更佳的检索环路设计。随着脑机接口变得越来越经济、轻量化，一些研究者认为，基于脑机接口的主动式信息检索范式可以首先在若干垂直应用中实现。例如，将脑机接口与虚拟现实设备结合[76]，在虚拟现实游戏中解放双手，同时仍然能够进行查询和与系统交互。然而，现有研究仍然面临一些挑战：例如尚未构建引入脑机信号的信息系统环路设计，缺乏有效的脑信号解码技术和信息系统反馈机制的设计等。尽管如此，基于脑机接口的主动式信息检索技术潜力巨大，未来有望在多个领域实现更高效和自然的人机交互。

第 3 章　基于脑机接口的用户认知过程理解与建模

3.1　本章引言

近年来，神经成像技术（如 EEG 和 fMRI）的快速发展，使得在信息检索情境下通过探索大脑活动来理解用户的认知过程成为可能。例如，Moshfeghi et al.[66] 和 Allegretti et al.[71] 分别利用 fMRI 和 EEG 设备，探索了用户在使用信息检索系统时信息需求的出现和相关性判断过程。这些研究构成了理解用户在信息获取情境下认知过程的重要一步，并提供了以前的用户研究技术（如眼动追踪、鼠标移动和显式标注等）无法获得的神经层面的发现。尽管现有研究在用户相关性判断的预测和机理分析方面取得了一定进展，但这些研究仍不足以全面理解用户在信息检索过程中的情感体验和细粒度认知过程。本章基于神经成像技术的两个特点，即其能够解码用户的真实满意度和情感状态以及细粒度认知过程，开展了相关研究：

首先，本章设计了信息检索过程中基于脑信号的**用户满意度建模**方法。用户满意度不仅反映用户是否找到了相关信息，还和他们对整个信息获取过程的情感反应和状态有着密切联系。通过捕捉和分析用户在信息检索过程中的脑信号，可以更准确地估计他们的满意度，从而改进搜索结果排序和推荐系统的性能。这将有助于构建更加智能和人性化的信息检索系统，提供更加个性化和高效的服务。其次，本章深入探讨了用户在**文本阅读理解过程中的细粒度认知机制**。用户在信息获取过程中不仅需要理解和提取文本中的信息，还需定位和处理与查询相关的关键信息。这种复杂的认知过程无法通过对文章级别的粒度的相关性判断来完全解释。通过研究用户在文本阅读理解过程中的大脑神经反应，本章揭示了他们在信息处理中的注意力分配和认知负荷。这对于优化信息呈现方式和提升信息检索系统的用户体验具有重要意义。综合这两个方面，本章在现有研究的基础上，融合了基于满意度的情感反应和阅读理解过程的细粒度认知行为，以更深入地理解和建模用户的认知过程。同时，本章讨论了其对未来信息检索系统设计的潜在影响，并设计了相关的下游应用。

在用户满意度建模方面，本章设计了一种基于 EEG 的满意度建模方法，该方法利用脑信号来估计用户在搜索和推荐系统中的满意度。目前，鲜有研究深入分析和设计信息检索过程中利用脑信号进行满意度建模的方法。与现有的 EEG 分类任务不同，用户满意度建模具有挑战性，因为满意度涉及认知过程（如原型理论[77] 和评估理论[78]）和情感状态（如唤醒[79]）。这一综合过程始于枕叶区域，并涉及到额顶注意回路、背外侧额叶区域和内侧颞叶区域的大脑功能[80]。为了连接

和聚合不同大脑区域的信号，本研究认为在满意度建模中捕捉脑信号的拓扑关系是必要的。现有研究通过在 EEG 分类模型中引入拓扑信息已经取得了一系列成果。在这些模型中，卷积神经网络（Convolutional Neural Network，CNN）[81-82]或图卷积网络（Graph Convolutional Network，GCN）[83-86]是最广泛使用的。它们在运动想象[81]和情感识别任务[86]中取得了较高的分类性能。然而，这些方法将脑拓扑连接视为固定的，未能捕捉和数据相关的脑拓扑连接关系，因此无法在满意度建模中取得令人满意的效果。为了解决上述问题，本章设计了一种新的用于 EEG 的满意度建模架构，名为大脑拓扑自适应网络（Brain Topology Adaptive Network，BTA）。该网络应用多中心性编码模块生成具有三维拓扑信息的编码。然后，受到注意力机制的启发[87]，本章采用空间注意力模块以数据依赖的方式捕捉大脑的认知连接。BTA 的有效性在两个典型的交互式信息检索场景中（搜索和推荐）得到验证，其效果要优于现有的最先进的 EEG 分类模型。此外，本章探讨了在搜索和推荐场景中可以利用 BTA 估计的满意度在多大程度上改善搜索结果重排序和视频评分预测性能，揭示了脑信号推断的用户满意度为交互信息系统带来的潜在收益（如图 3.1）。

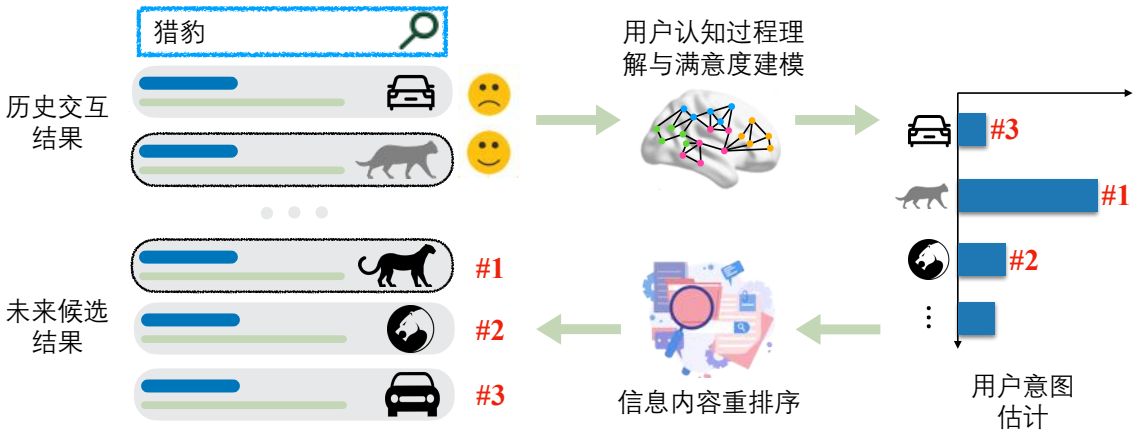


图 3.1　一个用户满意度感知的搜索引擎的示例

在文本阅读理解过程中的细粒度认知机制探究方面，本章借助 EEG 设备，开展了一个用户实验，以探究用户在阅读理解过程中，对于关键信息（答案内容和语义相关内容）和其他信息（其他支撑性内容）的大脑反应是否存在可检测的差异。在这项研究中，被试者需要通过阅读一段文本来回答给定的问题，同时会用 EEG 设备收集他们的脑活动。本研究将用户阅读的文本信息以词为粒度，根据其与问题的关联性分为三类，即答案内容、语义相关内容和其他支撑性内容，并通过 ERP 分析这一神经科学的典型方法[88]来分析用户在阅读这三类内容时的神经活动差异。基于分析，本章发现用户的神经活动会因内容类型不同而有所变化。值得注意的是，与认知负荷相关的特定 ERP 成分 N100-P200[89]（即 100ms-200ms 处

的负波和正波）在不同内容类型之间存在显著差异。此外，用户在阅读答案词会观察到较大的 P600 的 ERP 成分（即 600ms 处的正波），该成分的出现通常被认为和知识推理的认知活动相关。基于这一神经基础，本章为信息检索系统的设计提出了若干见解。例如，（1）排序模型的构建应在考虑文档相关性的基础上，也关注其细粒度结构，以减少用户的认知负荷，（2）文档结果摘要不仅应提供语义相关内容，还应提供更丰富的支撑性信息以便用户理解。此外，本章探索了利用脑信号检测阅读状态的可能性，并提出了一种基于脑电图的阅读理解用户建模框架，该框架可以利用脑信号定位用户在阅读理解过程中对答案句和答案词的判断。实验结果表明，与未训练模型相比，该建模方法在答案句分类任务中提高了 0.179（基于平均精确率的平均（Mean Average Precision，MAP）），在答案词提取任务中基于曲线下面积（Area Under the Curve，AUC）提高了 0.157，并且优于包括支持向量机（Support Vector Machine，SVM）、条件随机向量场（Conditional Random Field，CRF）在内的基线方法。

本章的结构如下：第 3.2节介绍相关工作，包括用户满意度、阅读理解和基于脑电图的分类模型设计方法等。第 3.3节介绍了基于 EEG 的满意度建模方法和下游应用。第 3.4节介绍了基于脑电图的阅读理解用户实验和建模方法。第 3.5节对本章进行小结。

3.2　相关工作

3.2.1　用户满意度

满意度衡量用户对系统的主观感受，反映了用户信息需求的满足程度[90]。用户满意度建模在信息系统的性能提升和评估中具有重要价值[91-92]。过去，研究者们集中于通过用户的隐式反馈信号（点击、停留时间、滚动等）来建模用户满意度[93-94]。近年来，许多研究引入了新的用户信号和相应策略来估计用户满意度，例如鼠标移动[95]和眼动追踪[96]。然而，隐式反馈信号只是对真实用户满意度的间接表现，因此常常不准确[97]。例如，在搜索场景中，Liu et al.[98]发现大比例（45.8%）的眼动停留与相关性估计无关。在新闻推荐中，Wang et al.[9]指出，由于“标题党”效应，用户的点击行为不能简单地被视为正面信号。为了解决上述弊端，本章探索了利用脑信号进行满意度建模的效果，并提出了 BTA 模型。此外，本章将基于脑信号推断的满意度应用于下游的搜索和推荐任务，从而展示了脑信号作为有效用户反馈在设计交互信息获取系统中的优势。

3.2.2　阅读理解

阅读理解是在基于文本的搜索情境中获取信息的认知过程之一，涉及视觉处理、语义理解和信息获取[99]。已有一些研究利用鼠标移动[100]和眼动追踪[101]探讨了用户在文档阅读过程中的行为模式和注意力分配机制。例如，Gwizdka[102]通过眼动研究阅读行为，指出文本文档的阅读的过程和用户的感知相关性相关。Li et al.[101]研究了段落级阅读理解中的注意力分布，结合用户的眼动和显性反馈，进一步提出了一个两阶段阅读模型。此外，现有的研究还探索了阅读理解过程中的隐性反馈。例如，Liu et al.[100]利用鼠标移动研究搜索引擎结果页面（Search Engine Result Page，SERP）的检验过程并预测用户对网页的满意度。Cole et al.[103]发现，阅读理解过程中的眼动模式可以推断用户的先验知识，以更好地建模搜索上下文。然而，这些方法无法直接揭示大脑中的实际认知活动以及阅读理解过程中的潜在心理因素。尽管神经科学技术已被用于研究一般领域的阅读行为，例如词汇识别[104]和句法分析[105]，但很少有研究文献专注于研究涉及信息查找过程中的用户阅读理解。因此，“在信息检索的情境下，阅读理解的本质（从认知神经科学的角度）是什么?”这一问题仍然是一个未解之谜。

3.2.3　基于脑电图（EEG）的分类模型

EEG 具有高时间分辨率、非侵入式和相对较低的成本优势。它广泛应用于涉及运动想象[81-82]、情感识别[84-86]和睡眠阶段预测[106-107]的研究中。最近基于 EEG 的分类模型的趋势已从拓扑不变算法（如 SVM[12]、决策树（Decision Tree，DT）[108]和循环神经网络（Recurrent Neural Network，RNN）[109]）发展到考虑拓扑结构的方法，如 CNN 和 GCN。拓扑感知模型在学习表示时考虑了 EEG 特征的拓扑结构。例如，Lawhern et al.[81]和 Kostas et al.[82]将 3D EEG 拓扑信息映射到 2D 表示，并采用基于 CNN 的架构聚合通道信息。此外，还有一些研究[83-84,110]应用基于图卷积网络的方法，采用公共邻接矩阵自动学习 EEG 通道之间的聚合权重。例如，Li et al.[86]利用多域信息为其基于 GCN 的模型构建邻接矩阵，实现基于 EEG 的情感识别。

然而，由于满意度与多种认知过程和情感状态相关[77-79]，在不同的信息检索任务下，大脑连接可能会有所不同[111-112]。以前的研究一般把大脑的功能连接视作固定的，很少以完全数据依赖的方式自适应学习大脑连接。一个例外是 HetEmotionNet[85]，它利用互信息[113]为每个数据样本构建邻接矩阵，然后应用图卷积。然而，互信息仅代表 EEG 特征的相似性。简单地聚合具有相似特征的 EEG 通道并不总是合理的。例如，先前研究表明左右脑之间神经活动的不对称性是有意义

的[114]。本章的研究在现有工作的基础上，结合多中心球坐标编码和空间注意力机制来实现完全数据驱动的聚合策略，提升了基于 EEG 的分类模型在满意度建模上的效果。

3.3　基于脑拓扑结构的用户满意度建模方法及其应用

3.3.1　常用符号定义

本节和后续章节将使用一些和基于脑机接口的信息检索任务常用的符号来表示不同的概念和变量，见表 3.1。

表 3.1　基于脑机接口的信息检索任务的常用符号

符号	含义
X	脑机信号特征（EEG 或 fMRI 等），X^s 表示 EEG 的频域特征，X^t 表示 EEG 的时域特征
Y	预测目标，如用户满意度标签（第 3.3节），答案词与非答案词（第 3.4节）等，$\hat{Y}$ 表示模型预测值
R	相关性得分，$R^{cn}, R^{cx}, R^{c}, R^{bs}, R^{p}$ 分别表示基于内容（content）、上下文（context）、点击信号（click）、脑信号（brain signal）和伪相关信号（pseudo relevance）的得分，$\hat{R}$ 表示模型预测值
Q	查询，Q^0 表示初始查询，Q' 表示扩展或增强后的查询
$\mathcal{D}$	文档列表，$\mathcal{D}_h$ 表示用户交互的历史文档，$\mathcal{D}_u$ 表示未交互的文档
F	表示计算模型，本文中包括从 EEG 特征到满意度的映射模型 F_s，文档排序模型 F_{search}（BM25[22]等）或物品推荐模型 F_{rec}（Wide&Deep[115]等）
Π	表示排序相关的评估指标，本文中常见的评估指标包括归一化折损累积增益（Normalized Discounted Cumulative Gain，NDCG）[116]，基于平均精确率的平均（Mean Average Precision，MAP）[117]等

3.3.2　问题定义

本节将 EEG 信号的时域特征定义为 $X^t = \{x_1^t, x_2^t, ..., x_N^t\} \in \mathbb{R}^{N\times E}$，其中 N 是时域特征的长度，E 是 EEG 通道的数量。EEG 信号的频域特征则表示为 $X^s = \{x_1^s, x_2^s, ..., x_B^s\} \in \mathbb{R}^{B\times E}$，其中 B 是频域特征的长度，E 是 EEG 通道的数量。频域特征是从 B 个频带（例如，$\delta, \theta, \alpha, \beta$ 和 γ）中提取的微分熵（Differential Entropy，DE）特征[118]。基于 EEG 的满意度建模问题是要学习一个从 EEG 特征到用户满

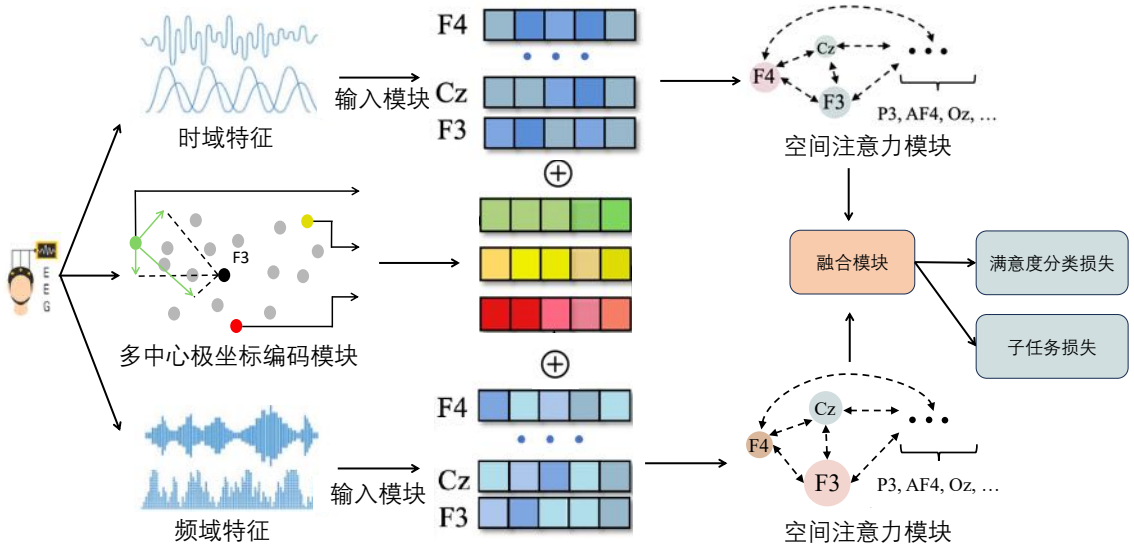


图 3.2　大脑拓扑自适应网络（BTA）架构

意度的映射函数 F_s，其公式如下：

$$\hat{y} = F_s(X^t, X^s) \tag{3.1}$$

其中 $\hat{y}$ 表示估计的满意度得分。通过预测用户对交互内容的满意度得分，信息系统可以更好地理解用户，并提供更多令人满意的内容。例如，本节探索了两种常见的信息获取场景，即搜索和推荐，详细内容见第 3.3.4节。

3.3.3　大脑拓扑自适应网络

3.3.3.1　模型概述

所提出的满意度建模方法的模型架构如图 3.2所示。该方法名为大脑拓扑自适应网络（Brain Topology Adaptive Network，BTA），它由时域数据流和频域数据流组成，这两个数据流具有相似的结构。该模型由五个组件组成：（1）输入编码模块，（2）多中心球坐标编码模块，（3）空间注意力模块，（4）融合与分类模块和（5）子任务模块。首先，通过输入模块将频域特征或时域特征线性投影到隐空间中。然后，多中心球坐标编码模块根据每个通道与不同空间中心点的拓扑关系生成空间编码。之后，空间注意力模块使用注意力机制聚合通道信息，以捕捉大脑的认知连接。最后，融合与分类模块融合时域数据流和频域数据流，以在信息获取任务中预测用户满意度。此外，在模型训练之前，我们用子任务模块替换融合与分类模块以初始化多中心球坐标编码模块中的空间中心点的嵌入向量。

3.3.3.2　输入编码模块

在输入编码模块，频域特征或时域特征先被线性投影到隐空间中，公式如下：

$$H^t = W^t \cdot X^t + B^t, H^s = W^s \cdot X^s + B^s \tag{3.2}$$

其中 $X^t \in \mathbb{R}^{N\times E}, X^s \in \mathbb{R}^{B\times E}$ 分别是时域和频域 EEG 特征，$W^t \in \mathbb{R}^{H\times N}$，$W^s \in \mathbb{R}^{H\times B}$，$B^t \in \mathbb{R}^{H\times E}, B^s \in \mathbb{R}^{H\times E}$ 是可学习的参数，$H^t \in \mathbb{R}^{H\times E}, H^s \in \mathbb{R}^{H\times E}$ 是输入向量。本章将 X^t 和 X^s 转换为相同维度的隐空间向量，这是为了与多中心球坐标编码模块中编码的拓扑信息的向量维度对齐（见第 3.3.3.3节）。时域和频域数据流共享相同的架构，因此为简便起见，在第 3.3.3.3节和第 3.3.3.4节中，本章省略了上标 t 和 s（例如，$H \in \{H^t, H^s\}$）。

3.3.3.3　多中心球坐标编码模块

为了对每个脑电通道编码其空间信息，本章提出了一种多中心球坐标编码的策略。具体来说，我们在大脑拓扑空间中选择 M 个空间点 $\{C_1, C_2, ..., C_M\}$ 作为空间中心点。对于每个空间中心点 C_j，我们构建一个以 C_j 为原点的球坐标系 $\mathcal{F}_j$。$\mathcal{F}_j$ 的 z 轴方向和 x 轴方向被定义为人脑的正上方和正前方。然后，我们构建一个编码 $p_{i,j} \in \mathbb{R}^H$，它根据其在 $\mathcal{F}_j$ 中的球坐标为每个 EEG 通道 E_i 分配三个可学习的嵌入向量，公式如下：

$$p_{i,j} = \rho_{i,j} \cdot c_{j,\rho} + \theta_{i,j} \cdot c_{j,\theta} + \varphi_{i,j} \cdot c_{j,\varphi} \tag{3.3}$$

其中 $(\rho_{i,j}, \theta_{i,j}, \varphi_{i,j}) \in \mathbb{R}^3$ 是 E_i 在 $\mathcal{F}_j$ 中的球坐标，$(c_{j,\rho}, c_{j,\theta}, c_{j,\varphi}) \in \mathbb{R}^{3\times H}$ 是与 C_j 相关的嵌入向量。为了丰富 EEG 通道的空间表示，每一个 EEG 通道 E_i 获得 M 个多中心球坐标编码，即 $p_{i,j}, j \in \{1, ..., M\}$。然后通过以下方式将多中心球坐标编码与时域或频域输入的隐向量组合：

$$z_i = h_i \oplus \sum_{1\leqslant j\leqslant M} p_{i,j} \tag{3.4}$$

其中 $h_i \in \mathbb{R}^H$ 是 $H^\intercal = \{h_1, ..., h_e\}$ 的子向量，$H^\intercal$ 是 $H \in \{H^t, H^s\}$ 的转置矩阵，$\oplus$ 是一个逐元素操作符。从而，引入多中心球坐标编码后的输入向量可以表示为 $Z = \{z_1, ..., z_e\}^\intercal \in \mathbb{R}^{H\times E}$。本章采用简单的逐位相加作为 $\oplus$，因为它对训练模型的开销最小，并且更复杂的交互操作符在实验中并没有带来额外增益。通过在输入中使用多中心球坐标编码，输入编码向量获得额外的脑拓扑信息。因此，模型可以同时捕获脑拓扑相关的频域或时域信息。

3.3.3.4　空间注意力模块

空间注意力模块旨在自适应地学习 EEG 通道之间的大脑认知功能连接关系。该模块首先应用一个多头注意力层来计算交互序列：

$$\begin{aligned} Z_1 &= \text{MultiHead}(Z^\intercal, Z^\intercal, Z^\intercal) \\ &= \text{Concat}(head_1, head_2, ..., head_{N_h})W^O \end{aligned}$$

其中 $head_i = \text{Attention}(Z^\intercal W^Q, Z^\intercal W^K, Z^\intercal W^V)$, $W^Q, W^K, W^V \in \mathbb{R}^{H\times H/N_h}$ 是可训练参数，N_h 是注意力头的数量，$\text{Attention}(Q, K, V)$ 是缩放点积注意力机制[87]。$\text{Concat}(\cdot)$ 表示在第二个维度上的连接操作，连接后的向量被输入到一个线性矩阵 $W^O \in \mathbb{R}^{H\times H}$ 中以获得空间交互向量 $Z_1 \in \mathbb{R}^{E\times H}$。

接下来，本章应用一个批量归一化层 BN 来加速训练过程，并获得输出向量 $Z_2 = \text{BN}(Z_1) \in \mathbb{R}^{E\times H}$。尽管 Transformer 等在自然语言处理中流行的架构一般使用层归一化[87]，但本研究中发现批量归一化在基于脑电的满意度建模任务中有更好的效果。其原因可能是批量归一化可以缓解 EEG 信号的不稳定性和异常值的影响[119]。

3.3.3.5　融合与分类模块

通过上述模块，输入的时域特征 X_t 和频域特征 X_s 分别转换为输出向量 Z_2^t 和 Z_2^s。这些输出向量被展开得到 $Z_3 = \text{Concat}(z_{2,1}, ..., z_{2,e}) \in \mathbb{R}^{EH}$，其中 $Z_2 = \{z_{2,1}, ..., z_{2,e}\} \in \mathbb{R}^{H\times E}$，$Z_2 \in \{Z_2^t, Z_2^s\}$，$Z_3 \in \{Z_3^t, Z_3^s\}$。然后，为了融合 Z_3^t 和 Z_3^s，我们采用一个全连接层、一个激活函数和一个 softmax 函数，公式如下：

$$\hat{y} = \text{softmax}(W\sigma(\text{Concat}(Z_3^t, Z_3^s)) + B) \tag{3.5}$$

其中 σ 表示激活函数（在本章的实验中为 Gelu），$W \in \mathbb{R}^{EH\times 1}$ 和 $B \in \mathbb{R}^1$ 是可训练参数，$\hat{y}$ 是估计的满意度分数。最后，分类交叉熵被用作损失函数，定义如下：

$$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{3.6}$$

其中 y 是真实标签。

3.3.3.6　子任务模块

此外，为了初始化与每个 $j \in \{1, 2, ...M\}$ 相关的嵌入向量 $c_{j,\rho}$、$c_{j,\theta}$ 和 $c_{j,\varphi}$，我们在满意度建模任务之前应用了一个无监督的子任务。无监督子任务被设计为一个预测一些掩码 EEG 特征的重构任务，除分类层外，该任务与前述模型的架构完全共享。具体而言，我们首先为每个数据样本生成随机的二进制噪声掩码 $W_{t,\text{mask}} \in \mathbb{R}^{H\times N}$ 和 $W_{s,\text{mask}} \in \mathbb{R}^{H\times N}$，输入 X_t 和 X_s 被掩码为 $\widetilde{X} = W_{\text{mask}} \odot X, X \in \{X_t, X_s\}$。

然后，我们用 $\hat{X}$ 替换 X，并用上述的模块生成输出的隐向量 $\hat{Z}_2$。接下来，我们使用一个可训练的线性连接层重构 X'：$X' = \text{linear}(\hat{Z}_2)$。最后，均方误差损失被用来计算重构损失，公式如下：

$$L_{\text{MSE}} = \sum_{W_{\text{mask}}(i,j)=0} (X'(i,j) - X(i,j))^2 \tag{3.7}$$

其中我们计算与掩码值预测相关的预测损失用于模型训练。在子任务之后，与每个 $j \in \{1, 2, ...M\}$ 相关的嵌入向量 $c_{j,\rho}$、$c_{j,\theta}$ 和 $c_{j,\varphi}$ 被用做于满意度建模任务的初始化参数。我们提出的 BTA 模型的完整训练过程总结在算法 3.1中。其中，无监督子任务旨在初始化一个更好的与每个球坐标系相关的嵌入向量，该向量的编码信息能够反映 EEG 通道之间的空间关系。本研究发现，当监督标签有限时，该初始化过程能有效提升模型的性能，而当使用更大的数据集时该步骤则不必要，详细的实验结果可参见第 3.3.5.1节。

算法 3.1 BTA 的训练过程。

Input: 用户对搜索结果的脑信号 $\langle X^1, ..., X^N \rangle$；搜索结果的相关性 $\langle Y^1, ..., Y^N \rangle$；初始化的 BTA 模型 $\boldsymbol{\Phi}$。
计算所有 $X \in \langle X^1, ..., X^N \rangle$ 的时域和频域特征 X^t 和 X^s。
$\boldsymbol{\Phi}' = \text{Copy}(\boldsymbol{\Phi})$。
for iteration=1,2,... **do**
    for 所有 $X \in \langle X^1, ..., X^N \rangle$ **do**
        生成随机二进制噪声掩码 $W_{t,\text{mask}}$ 和 $W_{s,\text{mask}}$。
        $\widetilde{X} = W_{\text{mask}} \odot X, X \in \{X_t, X_s\}$。
        计算重构损失 $L_{\text{MSE}} = L_{\text{MSE}}(\boldsymbol{\Phi}', \widetilde{X}, X)$。
        用 L_{MSE} 更新 $\boldsymbol{\Phi}'$。
    end for
end for
用 $\boldsymbol{\Phi}'$ 替换 $\boldsymbol{\Phi}$ 与每个球坐标相关的嵌入向量。
for iteration=1,2,... **do**
    for 所有 $X \in \langle X^1, ..., X^N \rangle$ 和 $Y \in \langle Y^1, ..., Y^N \rangle$ **do**
        计算分类损失 $L = L(\boldsymbol{\Phi}, X, Y)$。
        用 L 更新 $\boldsymbol{\Phi}$。
    end for
end for
返回 $\boldsymbol{\Phi}$

3.3.4　实验设置

本节详细介绍了数据集和实验设置。本章在 Search-Brainwave[19]数据集和 AMIGOS[64]数据集上进行满意度建模实验（详见3.3.4.1）。接着，本章在 Search-Brainwave 数据集上通过搜索结果重排序任务，以及在 AMIGOS 数据集上通过评分预测任务，以探索引入 EEG 信号进行满意度建模的交互

式搜索和推荐系统的性能。本章的实验实现代码基于 PyTorch ①，并公开于https://github.com/YeZiyi1998/DL4EEG-Classification。

3.3.4.1　数据集

Search-Brainwave 数据集　Search-Brainwave 数据集[19]记录了 18 名参与者在执行搜索任务时产生的脑信号。每个参与者平均完成 69.6 个搜索任务，并在每个搜索任务中查看 3.6 个相应的搜索结果。在此过程中，收集了他们对搜索结果的显式反馈（即满意或不满意）。数据集还提供了每个搜索结果的真实相关性标签以用于性能评估。本章选择此数据集是因为它是专为交互搜索任务设计的基准数据集。对于特征提取，本章使用了官方预处理的时域和频域特征（即原始信号和 DE[118]特征）。

AMIGOS 数据集　AMIGOS 数据集[64]包含 40 名参与者在视频刺激下产生的 EEG、心电图（Electrocardiogram，ECG）和其他生理信号。每位参与者在两个实验中分别观看了 16 段短视频和 4 段长视频。他们对每个视频在“效价”、“唤醒”、“优势”、“熟悉度”和“喜好”上进行 1 到 9 的评分。AMIGOS 数据集并不是专为推荐场景设计的，但该数据集包含丰富的参与者信息（匿名化的参与者人口统计信息、个性问卷信息和情绪问卷信息）。这些信息可以用于个性化评分预测任务中的用户嵌入训练（参见3.3.4.4节）并用于推荐场景的评分预测任务。

本节的实验只采用了短视频实验中的数据。本章将用户的“喜好”标注以 5 为阈值划分为满意（$\geqslant 5$）和不满意（< 5），以用于本节的二元满意度估计任务。对于特征提取，本章利用官方预处理的 EEG 信号，并应用一个步长为一秒的无重叠窗口将每个视频分成多个片段。窗口长度不同于原始论文[64]，但和绝大多数现有的 EEG 分类任务[83,85]一致。本节中使用了不同的时间窗口的原因是原始论文还考虑了其他生理信号，如 ECG，其需要比 EEG 更长的时间窗口[120]。对于每个视频片段的脑信号，本章用原始信号用作时域特征，并使用傅里叶变换在四个频带（即 θ，α，β 和 γ）上提取 DE[118]特征作为频域特征。

3.3.4.2　满意度预测

基线　本节采用三类 EEG 分类模型作为基线：拓扑不变模型、基于 CNN 的模型和基于 GCN 的模型。

拓扑不变模型包括 SVM[121]、决策树[122]和多层感知机（Multi-Layer Perceptron，MLP）[123]。本章使用 scikit-learn 库[124]来实现这些模型。

① https://pytorch.org/

基于 CNN 的模型包括 EEGNet[81]和 BENDR[82]。EEGNet 通过堆叠了多个 CNN 层以提升其性能。BENDR 应用 CNN 来提取 EEG 特征，然后使用 Transformers 捕捉 EEG 信号的时域特征模式，该方法在各种 BCI 任务中实现了最先进的性能。它还采用跨数据集预训练任务来初始化参数。本章用其开源代码实现 EEGNet 和 BENDR，并使用公开的预训练权重来初始化 BENDR 模型。对于 EEGNet，由于输入数据长度的不同，原始参数设置不适用于 AMIGOS 数据集。因此，本章将第二卷积层的内核大小减少到 7，最后一个池化层的内核大小减少到 $(2,1)$ 以进行 AMIGOS 数据集的实验。

基于 GCN 的模型包括 DGCNN[83]、RGNN[84]和 Het-EmotionNet[85]。DGCNN 应用一个公共邻接矩阵以动态聚合多通道 EEG 信息。RGNN 在 DGCNN 的基础上，使用两个正则化器来提高鲁棒性。Het-EmotionNet 利用互信息来建模拓扑信息，并将时域和频域信息融合在一起。本章用开源代码实现 RGNN 和 Het-EmotionNet，而 DGCNN 由我们自己实现（DGCNN 没有开源代码）。对于 RGNN[84]，由于其发布的代码中没有公开正则化参数，本章经验性地将其设置为 0.001。

参数设置　本章使用 Adam 优化器[125]训练 BTA。中心数 M 从 $\{1,3,5,7,9,14\}$ 中选择。结果表明，$M=3$ 能达到最佳性能。此外，本章发现选择不同球坐标系的中心点 C_1、C_2、C_3（例如，随机选择三个通道作为中心点）导致的差异很小。因此，本章选择国际 10-20 EEG 系统中的中心点作为 C_1。C_2 和 C_3 则被选择为左右乳突点，因为它们在现有 EEG 研究中被广泛用作参考电极[126]。初始学习率、批量大小和隐藏维度 H（如第 3.3.3节所述）分别从 $\{0.01,0.05\}$、$\{8,32\}$ 和 $\{8,16,32\}$ 中选择。多头注意力机制中的头数设置为 8。此外，在无监督子任务中，本研究将随机掩码比例设置为 15%，与时间序列预测中的相关工作一致[119]。

评估　本章在实验中应用 AUC 和 F1-score 作为评估指标。对于 Search-Brainwave 数据集，本章基于每个被试者的数据分别进行实验，并使用任务独立的十折交叉验证来评估模型：将搜索任务分为十折，每次在其余九折上进行训练，然后对剩下的一折进行评估。对于 AMIGOS 数据集，按照现有研究[127-128]，本章同样基于每个被试者的数据分别进行实验，并应用十折交叉验证来随机划分数据。

3.3.4.3　下游任务 1：搜索结果重排序

任务定义　在搜索结果重排序任务中，未交互的候选搜索结果将根据基于历史交互的搜索结果估计的满意度进行重排序。对于一个查询 Q，其搜索结果为列表 $\mathcal{D}=\langle d_1,...,d_n\rangle$。本研究假定用户交互了前几个结果 $\mathcal{D}_h=\langle d_1,...,d_h\rangle$，因此对第 i

个搜索结果的交互表示为 $I_i = \{\hat{y}_i, d_i\}$，其中 $\hat{y}_i$ 是对搜索结果 d_i 的满意度估计得分。然后，本研究希望在未交互的候选搜索结果列表 $\mathcal{D}_u = \langle d_{h+1}, ..., d_n \rangle$ 中将相关的搜索结果排在前面。因此，搜索结果重排序任务的目标是：

$$\max \sum_{Q} \Pi(\hat{\mathcal{D}}_u, R_u), \hat{\mathcal{D}}_u = F_{\text{search}}(Q, \mathcal{D}_u, I_1, ..., I_h) \tag{3.8}$$

其中 Π 表示评估指标，例如 NDCG[116] 或 MAP[117]，R_u 是 $\mathcal{D}_u$ 的真实相关性标签，$\hat{\mathcal{D}}_u$ 是返回的重排序列表，F_{search} 表示根据历史交互重新排序未交互的候选搜索结果的策略，详见下文的方法部分。

方法　为了测试脑信号预测的满意度在搜索结果重排序任务中的有效性，本章构建了一个统计语言模型[129]来重写查询 Q 以更好地表示用户意图。语言模型估计历史搜索结果 $d_i = \{w_{i,1}, w_{i,2}, ...\} \in \mathcal{D}_h$ 中的词与查询 $Q = \{q_1, ..., q_k\}$ 的相关性，以找到与查询最相关的词。然后，查询被重写为包含 l 个最相关词的 $Q' = \{q_1, ..., q_k, w_1, ..., w_l\}$。在统计语言模型中，词的相关性 $R(w)$ 可以表示为：

$$R(w) = \text{LM}(\mathcal{D}_h, Q, P_{\mathcal{D}_h}) \tag{3.9}$$

其中 $P_{\mathcal{D}_h}$ 是 $\mathcal{D}_h$ 的概率分布，LM 表示一个统计语言模型。通常，如果一个词出现在具有较高分布概率的搜索结果中，其相关性会更高。一般在没有先验知识的情况下，统计语言模型将 $P_{\mathcal{D}_h}$ 设为一个均匀分布，可以表示为*均匀语言模型*（Uniform Language Model, ULM）。通过从脑信号中预测的满意度，我们调整 $P_{\mathcal{D}_h}$，为满意的搜索结果分配更高的概率：

$$P(d|\mathcal{D}_h) = (\lambda + \hat{y}(d))/(\lambda|\mathcal{D}_h| + \sum_{d \in \mathcal{D}_h} \hat{y}(d)) \tag{3.10}$$

其中 λ 是一个超参数，用于平滑估计的满意度得分 $\hat{y}(d)$。本节的这个方法被记为*满意度增强语言模型*（Satisfaction-enhanced Language Model, SLM）。在使用语言模型将查询 Q 扩展为 Q' 后，我们采用排序算法基于 Q' 来生成重新排序的搜索结果列表 $\hat{\mathcal{D}}_u$。

本节的实验使用 BM25[22] 作为排序算法。对于参数设置，本节将重写词的长度 l 和超参数 λ 分别经验性地设置为 5 和 2。

评估　对于 Search-Brainwave 数据集中的每个搜索任务，如果用户检查了 h 个搜索结果，我们将重新排序剩余的 $n - h$ 个搜索结果，并使用搜索结果的真实相关性标签评估重排序列表。注意 h 因搜索任务而异，这是因为用户可以随时停止搜索。最后，评价指标 NDCG 和 MAP 被用于评估搜索任务的平均性能。

3.3.4.4　下游任务 2：评分预测

任务定义　评分预测是一个经典的交互式推荐任务，该任务根据历史用户-物品交互来预测未交互的用户-物品对的评分。特别地，本节将用户和物品集合分别表示为 U 和 V。历史用户-物品交互表示为 $I_h = \{(u, v, \hat{y}_{u,v}) | u \in U, v \in V\}$，其中 $\hat{y}_{u,v}$ 是通过脑信号估计的用户 u 对物品 v 的满意度得分。这里基于脑信号估计的满意度得分 $\hat{y}_{u,v}$ 替代了在传统推荐任务中的显式评分或隐式反馈（例如，点赞、点击）。评分预测任务的目标是更好地估计未交互的用户-物品对 $I_{\text{un}} = \{(u, v, y_{u,v}) | u \in U, v \in V, (u, v, *) \notin I_h\}$ 的真实标签 $y_{u,v}$。综上，评分预测任务可以被形式化为：

$$\max \Pi(I_{\text{un}}, F_{\text{rec}}(I_h, \{(u, v) | (u, v, y_{u,v}) \in I_{\text{un}}\})) \tag{3.11}$$

其中 Π 表示评估指标，即 AUC，F_{rec} 是评分预测策略，详见方法部分，本节使用用户的显式标注作为真实标签 $y_{u,v}$。

方法　个性化推荐中的评分预测任务旨在学习一个映射函数 F，基于用户和物品的嵌入向量学习用户-物品对的评分：

$$y_{(u,v)} = F(e_u, e_v) \tag{3.12}$$

其中 e_u, e_v 分别表示用户嵌入和物品嵌入，$y_{(u,v)}$ 是用户-物品对 (u, v) 的评分。在推荐系统中，真实标签 $y_{(u,v)}$ 往往难以获取[9]。因此，在训练过程中，本研究假设可以获得 α 比例的用户-物品对的真实标签 T。同时，对于剩下的 $1-\alpha$ 比例的用户-物品对，本研究使用估计的满意度得分 $\hat{y}_{(u,v)}$ 作为满意度标签来训练模型。最终，真实标签 $y_{(u,v)}$ 被用于评估模型性能。

本节使用 $F^{T(\alpha)}$ 来表示使用 α 比例真实标签 T 训练得到的推荐模型，使用 $F^{T(\alpha),S}$ 来表示使用 α 比例真实标签 T 和 $1-\alpha$ 比例的满意度标签 S 训练得到的推荐模型。本节选取了一些最常用的个性化推荐方法作为 F，包括逻辑回归 (Logistic Regression，LR)[130]、因子分解机 (Factorization Machine，FM)[131] 和 Wide&Deep[115]。所有推荐方法均使用 Recbole 的开源代码实现[132]，并应用其默认的参数设置。

评估　本章将 AMIGOS 数据集中的每个视频段视为一个物品，并将其分配为物品嵌入的独热编码。对于用户嵌入，本章基于用户的人口统计问卷、个性问卷和情绪问卷中用户在各个选项中的选择构建了一个 71 维嵌入向量。本章将用户-物品对随机分为训练、验证和测试集，比例为 8:1:1。在训练集中，本章将从脑信号推断的满意度替换一定比例的真实标签，并在验证和测试集上用真实标签评估评分预测性能。

3.3.5　实验结果与讨论

本章的实验包含两个部分。首先，本章通过实验证实 BTA 在满意度预测任务中的有效性，并将 BTA 与流行的 EEG 分类基线进行比较，深入研究 BTA 的各个组件的效果，详见3.3.5.1节。其次，本章研究了利用 BTA 估计的满意度提升交互式信息获取的性能，即将 BTA 预计的满意度得分用于搜索结果重排序任务和评分预测任务，详见3.3.5.2节和3.3.5.3节。

3.3.5.1　满意度预测

表 3.2　不同模型的满意度估计性能[a]

模型		Search-Brainwave		AMIGOS	
		F1	AUC	F1	AUC
拓扑不变	DT	0.5642*	0.5205*	0.5608*	0.6245*
	MLP	0.6196*	0.5204*	0.5629*	0.6123*
	SVM	0.6227*	0.5189*	0.5580*	0.5892*
基于 CNN	BENDR	0.7118*	0.7291*	0.5580*	0.5869*
	EEGNet	0.7254*	0.7614*	0.6025*	0.6920*
基于 GCN	DGCNN	0.7170*	0.7374*	0.6630*	0.7663*
	HetEmotionNet	0.7362*	0.7717*	0.6428*	0.7405*
	RGNN	0.7440*	0.7663*	0.6694*	0.7782*
拓扑自适应	BTA（本方法）	**0.7837**	**0.8278**	**0.7143**	**0.8353**

[a] * 表示与 BTA 相比的差异在 $p < 0.01$ 的水平上显著。

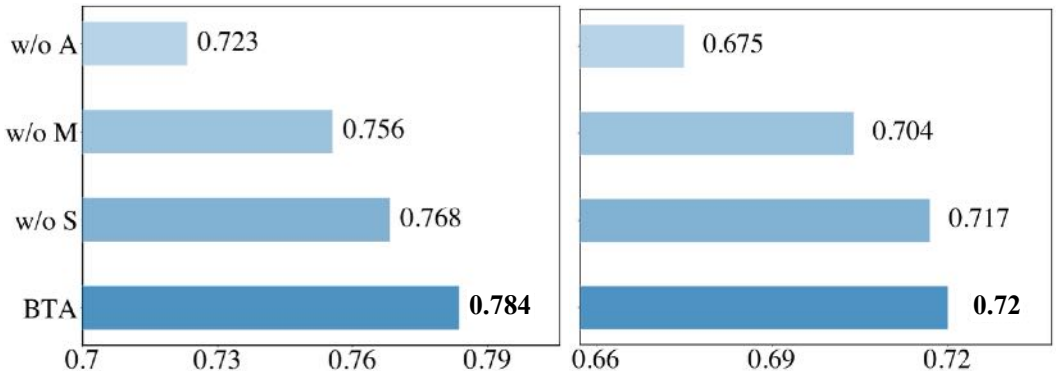
整体表现　表 3.2展示了不同模型在 F1-score 和 AUC[133]指标下的满意度预测性能。从表中可以观察到：（1）所有基于 CNN 和 GCN 的模型在大多数评估指标上都优于拓扑不变模型。拓扑不变模型直接将所有 EEG 通道的特征连接在一起，因此忽略了拓扑信息，导致其性能受限。（2）基于 CNN 的模型整体表现不如基于 GCN 的模型。基于 CNN 的架构将 3D 拓扑信息压缩为 2D 表示，并在相邻的 EEG 通道中聚合信息。相反，基于 GCN 的模型利用可学习的邻接矩阵来学习更灵活的聚合策略，比基于 CNN 的模型更好地利用了拓扑信息。（3）提出的 BTA 在所有模型中表现最好。一方面，BTA 引入了多中心球坐标编码以利用 3D 的空间关系，而先前的模型没有考虑这种 3D 信息。另一方面，BTA 中的空间注意力机制被用于

自适应地捕捉拓扑信息，因此我们可以在不同的数据样本中学习不同的聚合策略。相反，基于 CNN 和 GCN 的模型则是利用固定的聚合权重在所有数据样本中进行聚合。一个例外是 HetEmotionNet，它使用了 EEG 通道之间的互信息来学习聚合权重。下面的分析将进一步比较 BTA 和 HetEmotionNet 在拓扑信息的利用方面的差异。

消融研究　为了探索 BTA 中不同组件的有效性，本章进行了关于模型结构或训练策略的消融实验。图 3.3显示了 BTA 及其变体在 F1-score 指标下的结果。w/o A、w/o M 和 w/o S 分别表示消融空间注意力模块、多中心球坐标编码模块和无监督子任务模块的 BTA 模型。如图 3.3所示，在消融三个组件之一时，两个任务的表现都有不同程度的下降。这表明这些组件都促进了模型性能。在这些组件中，空间注意力模块起到了最重要的作用，这表明自适应聚合通道信息的有效性。此外，无监督子任务模块带来的性能提升最小。特别是在 AMIGOS 数据集中，与 BTA w/o S 相比，BTA 的性能提升仅为 0.003 的 F1 值。其原因可能是 AMIGOS 数据集包含更多的训练样本和监督标签，因此无监督的参数初始化过程不太必要。

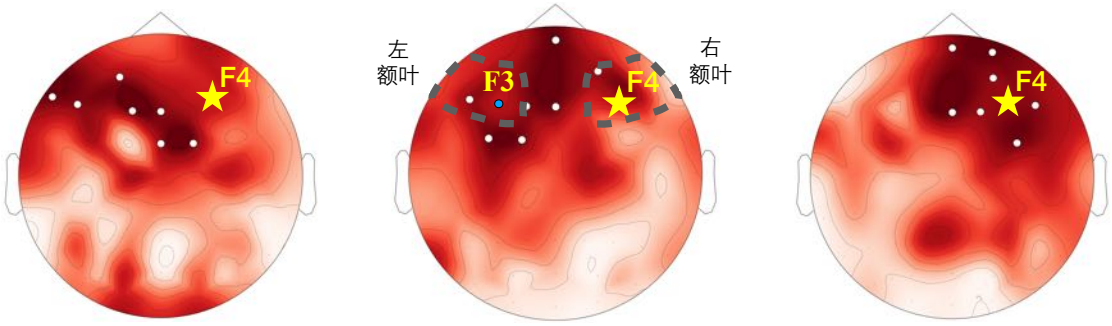
脑拓扑分析　本章通过可视化聚合权重来探索 BTA 和基线 HetEmotionNet 中的大脑拓扑信息。虽然 BTA 和 HetEmotionNet 有不同的架构，但两者都学习了一个聚合权重来从 EEG 通道中聚合信息。对于 BTA，本节定义聚合权重为所有数据样本中 EEG 通道之间两两的平均注意力权重。对于 HetEmotionNet，聚合权重则被表示为所有数据样本中 EEG 通道之间做聚合时的平均边缘权重[85]。

图 3.4展示了 Search-Brainwave 数据集中，其余通道关于 F4 通道的聚合权重可视化。本章选择通道 F4，是因为先前的神经科学研究[114,134]表明额叶 α 频段的不对称性（即左额叶和右额叶之间的 α 波段差异）与人的动机、愿望和积极/消极的



(a) F1（Search-Brainwave）　　(b) F1（AMIGOS）

图 3.3　BTA 与其变体的性能比较。A: 空间注意力机制；M: 多中心球坐标编码；S: 无监督子任务。



(a) BTA（满意）　(b) BTA（不满意）　(c) HetEmotionNet（满意）

图 3.4　BTA/HetEmotionNet 在满意/不满意数据样本中的聚合权重可视化结果。颜色越深表示对通道 F4 的聚合权重越高，突出显示的通道在该数据样本中权重最高的 7 个通道。

情感相关。从图 3.4可以观察到，HetEmotionNet 和 BTA 关于 F4 通道的聚合权重分别在右额叶和左额叶出现了更多的高聚合权重。HetEmotionNet 采用互信息[113]来获得边缘权重。由于相邻通道通常共享更高的互信息分数，因此其聚合过程更倾向于聚合一些最相邻的通道。相反，BTA 通过注意力机制和多中心性编码自适应地捕捉 EEG 通道之间的拓扑关系。因此，它以更灵活的方式聚合信息。有趣的是，BTA 的聚合权重在左额叶更高。这意味着 BTA 利用更高的权重来捕捉 F4 与左额叶通道（如 F3）之间的关系，这与神经学研究中额叶 α 频段的不对称性[134]具有一致性。此外，BTA 中的满意和不满意样本具有不同的拓扑关系，这也表明数据依赖的建模策略可以适应不同的用户状态。

3.3.5.2　搜索结果重排序性能

表 3.3　不同模型的搜索结果重排序性能[a]

模型	NDCG@1	NDCG@5	NDCG@10	MAP@10
BM25	0.6881*	0.7397*	0.8164*	0.7333*
ULM	0.7237*	0.7620*	0.8309*	0.7687*
SLM	**0.7351**	**0.7767**	**0.8337**	**0.7741**

[a] * 表示与 SLM 的差异在配对 T 检验中 p-value< 0.01。

表 3.4　查询改写的案例研究，其中加粗的词语与用户意图相关

查询	搜索结果/满意度	ULM 改写	SLM 改写
恒牙	（1）...，在线医疗建议：宋医生 /☹ （2）孩子**多大**长恒牙，... /☺	恒牙，牙医，知道，在线，**孩子**，**儿童**	恒牙，**孩子**，**多大**，**几岁**，知道，**儿童**

表 3.3中汇报了搜索结果重序任务的性能。可以观察到，所提出的 SLM 优于

表 3.5　各模型的个性化评分预测的 AUC 性能[a]

模型	LR	FM	Wide&Deep
$F^{T(0.1)}$	0.6798	0.5010	0.7027
$F^{T(0.1),S}$	**0.7537**	**0.8056**	**0.8207**

[a] ∗ 表示与 SLM 的差异在配对 T 检验中 *p*-value< 0.01。

BM25 和 ULM。这表明了 SLM 的有效性，即通过从具有较高估计用户满意度的搜索结果中推断相关词来重写查询能够提升重排序性能。

为了进一步探索 SLM 和 ULM 的查询重写性能，本节进行了一项案例研究以分析它们的差异。表 3.4显示了一个搜索任务，该任务需要用户探索有关恒牙生长时期的信息。ULM 同时使用了出现在搜索结果 1（例如，“在线”，“牙医”）和搜索结果 2（例如，“儿童”，“孩子”）中的词重写查询。另一方面，借助预测的满意度，SLM 则倾向于用出现在令用户更满意的搜索结果（搜索结果 2）中的词来重写查询。例如，搜索结果 2 中的“孩子”，“几岁”和“多大”等词被用于 SLM 的重写。该重写的查询可以更好地呈现用户意图，并进一步提高重排序性能。

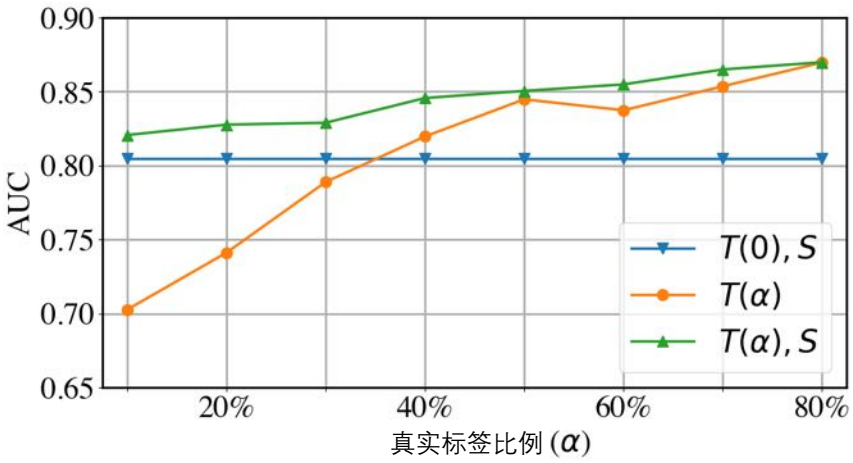


图 3.5　Wide&Deep 模型的性能：仅使用预测满意度作为标签（$F^{T(0),S}$）、使用不同比例的真实标签（$F^{T(\alpha)}$）和组合使用一定比例的真实标签和预测满意度标签（$F^{T(\alpha),S}$）

3.3.5.3　评分预测性能

表 3.5总结了真实标签比例取值为 0.1 的情况下的不同模型的评分预测性能。如表 3.5所示，在引入预测满意度 S 的情况下，所有推荐模型的性能都大幅提升。这表明了使用基于脑信号预测的满意度来训练推荐模型的有效性。

此外，为了更好地了解脑信号在多大程度上可以改善交互推荐系统，本章探

(a) 侧面视角　　(b) 顶部视角

图 3.6　实验环境和设备实景图

索了 Wide&Deep 在使用不同比例的真实标签 $\boldsymbol{T}$ 及是否使用预测满意度 $\boldsymbol{S}$ 时的性能。图 3.5展示了它们的性能比较。实验观察到，仅使用脑信号估计的满意度 $\boldsymbol{S}$（即 $\boldsymbol{F}^{T(0),S}$）的 Wide&Deep 与使用约 30%-40% 真实标签 $\boldsymbol{T}$（即 $\boldsymbol{F}^{T(\alpha)}$）的 Wide&Deep 同样有效。此外，$\boldsymbol{F}^{T(\alpha),S}$ 在相同的真实标签比例 $\boldsymbol{\alpha}$ 下优于 $\boldsymbol{F}^{T(\alpha)}$，这表明引入预测的满意度 $\boldsymbol{S}$ 可以稳定地提升性能。需要注意的是，在推荐场景中，真实的用户满意度标注往往是稀缺的[135]。因此，基于脑信号预测的满意度在提高推荐模型的性能方面具有重要意义。

3.4　用户阅读理解的细粒度认知过程理解与建模

3.4.1　用户研究

本章应用脑机接口来探索用户阅读理解任务中的细粒度认知过程，并开展了一个基于人类参与者的用户研究。在该研究中，参与者被招募来进行若干阅读理解任务。每个任务包括一个事实性问题和对应的一个句子，句子具有不同的相关性等级，参与者需要判断句子是否能够回答问题。本研究的代码和数据集已在以下地址开源：https://github.com/YeZiyi1998/UERCM。本用户研究遵循保护人类参与者研究的伦理程序，并获得清华大学心理学系伦理委员会的批准（2021 伦审第 04 号）。

3.4.1.1　参与者

本研究招募了 21 名年龄在 18 至 27 岁之间的大学生（平均值 =22.10，标准差 =2.07）。其中有 11 名男性和 10 名女性，他们主要专业为计算机科学、物理、艺术和工程。完成整个任务大约需要两小时，其中包括 40 分钟的准备时间，用于佩戴

表 3.6　用户阅读理解研究的任务示例[a]

问题	世界上最大的哺乳动物是什么？
完全相关	**蓝鲸**是世界上最大的动物，成年的体积达到了 33 米。
相关	世界上 表面面积最大的动物 是北极雪耳水母。
不相关	据估计，人体内有大约 100 亿个毛细血管。

[a] **粗体**和下划线 分别表示标注的答案词和语义相关词。

脑电帽和进行实验任务指导。每位参与者在完成所有任务后可获得 240 元人民币的报酬。

3.4.1.2　任务准备

数据集　在用户研究中，本章选择了 WebQA[136] 作为实验语料，WebQA 是一个事实性问答数据集，所有的问题都有明确的事实性答案。本章使用这个数据集的理由如下：（1）它是最大的中文问答数据集之一。（2）它提供了正确答案及其对应原因的人工注释。首先，本研究从数据集中手动采样了 155 个涵盖科学、历史、体育和艺术等主题的问题。然后，针对采样的每个问题，本研究又从数据集中再选择三个对应的句子，并手动为每个句子标注相关性标签。具体来说，会选择正确答案对应的句子、由 BM25 检索但不包含答案片段的句子以及随机选择的句子分别作为完全相关、相关和不相关的候选句子。接下来，本研究进行了进一步的标注来人工验证并纠正它们的相关性标签，标注的定义见第 3.4.1.5节。

接下来，本研究手动修正了一些句子的语法问题并精简了一些太长的句子。最终，问题的平均长度为 8.7 词（标准差 =4.0），句子的平均长度为 9.8 词（标准差 =3.0）。不同相关性水平句子的示例见表 3.6。

按照上述步骤，本研究获得了一个包含 155 个问题和 465 个句子（每个问题有三个对应句子）的数据集。在用户研究中，每个参与者在阅读完问题后，将接着阅读从三个候选句子中随机抽取的一个句子。

标注　在构建阅读理解数据集后，本研究招募了三名外部评估员来标注句子级的相关性，并标注答案词和语义相关词。标注数据的示例见表 3.6。作为每个单词和句子的分类任务，三位标注者之间的 Fleiss's kappa 在句子级相关性评估上为 0.9542（几乎完美的一致性），在答案词识别上为 0.9343（几乎完美的一致性），在语义相关词识别上为 0.7848（强一致性）。

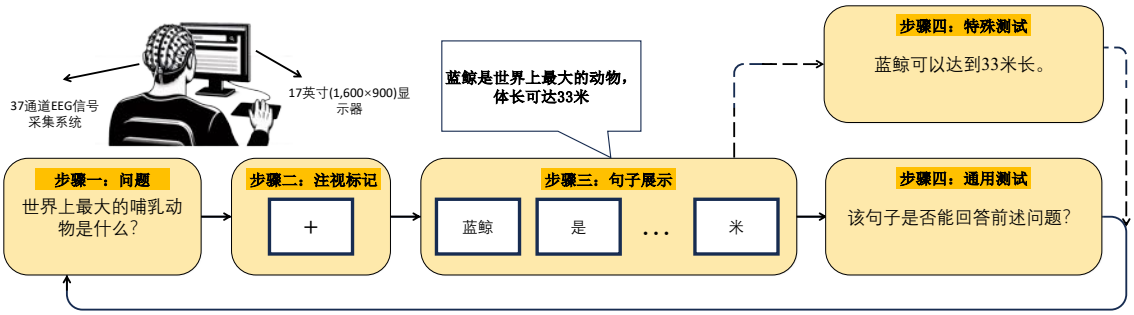


图 3.7　用户阅读理解研究的主任务流程

3.4.1.3　用户研究流程

用户研究的流程包括 3 个阶段，实验准备、主实验和实验后调研，具体如下。

阶段 1：实验准备　开始时，参与者填写一个实验前问卷，问卷内容主要包含参与者的人口信息，并签署关于安全和隐私保护的知情同意书（参见图 6.1）。然后，他们将阅读一份用户研究的主要流程说明和主实验要求。在主实验开始之前，参与者将进行一个包含五个任务的训练阶段，该任务内容和主实验阶段中的任务完全一致。该训练步骤主要是确保参与者能够熟悉主实验的流程和要求。

阶段 2：主实验　图 3.7展示了主实验中每个任务的流程。主实验总共包含 150 个任务，并分为六组，每组包含 25 个任务。任务按照图 3.7中显示的步骤顺序进行（步骤 1 到步骤 4）：（步骤 1）参与者观看从数据集中随机选择的事实性问题。当他们认为对问题内容完全理解后，可以按空格键进入第二步。（步骤 2）屏幕中心显示一个注视十字以吸引参与者的注意力，并指示接下来的句子内容的展示位置。注视十字将展示 1,000 毫秒。（步骤 3）一个从三个候选句子中随机选择的句子将逐词呈现，每个词将展示 750 毫秒。词语的连续呈现是在自然句子阅读任务中开展 ERP 研究的典型方法[137]。750 毫秒的阅读速度的设置参考了关于刺激诱发的脑电反应延续时长的研究[138]。（步骤 4）参与者进行一个二元的测试，该测试随机从普通测试和特殊测试中选择。普通测试的问题是“该句子能否回答先前的问题?”而特殊测试是一个涉及该句子内容的事实性问题，例如“蓝鲸的长度可以达到 33 米”。参与者需要对此进行是或否的二元判断。普通测试是为了确认参与者仔细阅读了问题并能够判断给定问题与句子之间的关系。而特殊测试是为了确保即使参与者可以提前做出普通测试的判断，他们也认真阅读完了整个句子。最后，在参与者按下按键（J 键表示“是”，F 键表示“否”）完成测试后，可以开始下一个任务。在每组中，测试的准确性将被主试监控，以确保参与者认真完成任务。在整个过程中，参与者的 EEG 数据被记录。

阶段 3：实验后调研　完成主实验后，他们需填写关于给定问题熟悉度的实验后问卷。

3.4.1.4　预实验

在开展用户研究之前，本章进行了一项预实验以确保 EEG 记录系统和用户研究程序正常运行。预实验的参与者为 21 名参与者之外的四人。预实验参与者提供的详细反馈用于调整用户研究的参数设置，包括字体大小、试验数量、休息时间等。此外，遵循先前的研究[139]，本章设计了普通测试和特殊测试以确保参与者能够认真执行问答任务。为了调整每种测试的概率，本章将特殊测试的比例设置选为 10% 或 20%，发现 10% 的比例下，特殊测试的准确率超过了 90%，而 20% 的比例下，特殊测试的准确率没有显著提高。因此，普通测试和特殊测试被采用的概率分别被设置为 90% 和 10%。

3.4.1.5　相关性水平的定义

相关性水平的定义如下：

- 完全相关：该句子直接回答了问题，因此我们可以获得问题的确切答案。它值得在搜索引擎中作为头部结果。
- 相关：该句子提供了一些与问题相关的信息。它在语义上相关，但对解决问题的贡献可能有限。
- 不相关：该句子未提供任何关于问题的有用信息，并且在语义上不相关。

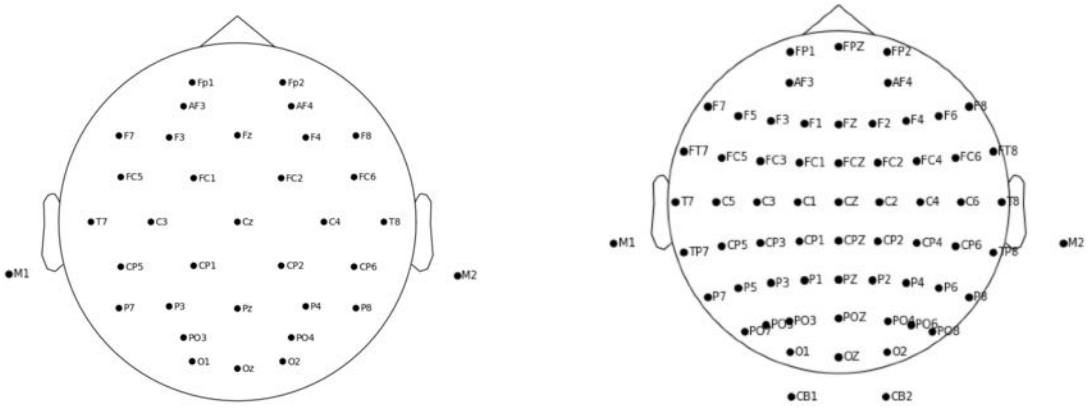
这些定义修改自 TREC 2019 深度学习赛道中的四个相关性级别定义[140]。本章将高度相关和相关的相关性级别合并为该定义中的相关，以简化任务设置。

3.4.1.6　设备

本章的用户研究基于一台 17 英寸显示器的笔记本电脑，分辨率为 1,600×900。一个 40 电极的 Scan NuAmps Express 系统（Compumedics Ltd.，VIC，澳大利亚）和一个 37 通道的 Quik-Cap（Compumedics NeuroScan）被用于采集参与者的 EEG 数据。EEG 设备的通道/电极分布可参见图 3.8(a)。

3.4.1.7　ERP 分析方法

事件相关电位（Event-related potential，ERP）是指大脑结构在响应特定事件或刺激时产生的和时间相关的电压波动[141]。它通常指的是在实验刺激发生后不到 1,000 毫秒的短暂时期内的电压波动。ERP 成分是在不同时间窗口中激发的波幅，包括 N100、N400（100 毫秒、400 毫秒的负波）和 P200、P600（200 毫秒、600 毫



(a) 32 导（第 3.4节的用户研究）　　(b) 64 导（第 5.3-5.4节的用户研究）

图 3.8　32 导/64 导的 EEG 通道示意图

秒的正波）。先前的研究表明，ERP 成分的激活与感官和认知过程相关的神经活动有关。ERP 成分之间的波形变化也被广泛研究，例如从 N100 成分到 P200 成分的变化[89]。为了从神经科学的角度深入理解人类阅读理解，本章应用了标准的 ERP 分析方法，包括数据预处理、时间窗口和感兴趣区域（Region of interest，ROI）的划分以及统计方法。

数据预处理　EEG 数据根据标准程序进行预处理。首先，EEG 数据被重新参考到平均乳突（M1 和 M2）。其次，对每个通道应用基线校正以去除信号中的波动。第三，对 EEG 数据在 0.5-30.0Hz 的频率范围内进行滤波，以保留 EEG 频率带，去除市电等其他电信号的干扰。第四，根据触发时间点（定位感兴趣的 EEG 数据的时间点）提取任务相关的 Epoch（即简短的 EEG 片段，在本研究的实验设置中为 1,000ms），并使用刺激前-200 至 0ms 进行基线校正。第五，执行参数化噪声协方差模型[142]以去除与眼球、心脏和肌肉伪影相关的成分。此外，绝对最大电压超过阈值 100μV 的 Epoch 被标记为坏段，不用于后续分析。最后，EEG 数据被降采样到 500 Hz，ERP 数据在相同类型的词中平均，以供后续分析。

时间窗口　本章将 ERP 数据划分为若干时间窗口以区分不同 ERP 成分。数据划分参考了 Lehmann et al.[143]提出的方法，计算了 0-750ms 之间的全局场功率（Global field power，GFP），并根据功率分布确定 ERP 的时间段。最终，确定的时间段分别为 N100：60-120ms，P200：120-320ms，N400：320-520ms，以及 P600：520-750ms。作为早期成分，N100 是一种无意识的用户反应，还不涉及文本内容的语义理解[88,144]。因此，本节只讨论 P200、N400 和 P600 成分上的发现。

表 3.7　在所有时间窗口及其 ROI 中，答案词（A）、语义相关词（S）和普通词（O）之间的 ERP 波幅的统计显著性差异

时间窗口	ROI	事后检验	ANOVA p[a]
120-320ms	额叶	A>S*	*
	顶叶	A>O*	*
320-520ms	中央脑区	A>S*,A>O**	**
	右侧颞叶	A>O**	**
	顶叶	A>S*,A>O**	**
520-750ms	中央脑区	A>S**, A>O**	**
	左侧颞叶	A>S**, A>O*, S<O*	**
	顶叶	A>S*, A>O**	**

[a] */** 表示使用事后成对 Bonferroni 检验和重复测量方差分析测试的统计显著性水平分别为 $p < 0.05/0.001$。

ROI　不同的大脑区域具有不同的功能，例如，顶叶与逻辑和数学思维相关。在信息检索领域，额叶、顶叶和右颞叶被认为与相关性判断有关[145]。因此，有必要选取 ROI 来开展研究。本章对每个 ERP 成分的固定时间窗口中的数据应用置换 T 检验。然后根据显著的电极及其所处的大脑区域确定 ROI。根据空间分布，将电极分配到七个大脑区域：前额、额叶、中央、顶叶、左颞叶、右颞叶和枕叶。每个时间窗口下，选定的 ROI 如表 3.7所示。

统计学方法　为了测试不同类型词之间 ERP 成分的差异，本章应用了重复测量方差分析。自变量是三种类型的词：答案词、语义相关词和普通词。不同类型词的例子可参见表 3.6。因变量是给定时间窗口和 ROI 中的平均信号。本章也尝试了在不同句子相关性下的 ERP 效应，发现与整体的 ERP 分析的结果相似，因此并未讨论相关结果。在多组比较之前，本章使用了 Shapiro-Wilk's 测试检查数据的正态性。同时，为了检查重复测量方差分析的可行性，使用了 Mauchly's 测试验证每个条件的球形假设。当球形假设不满足时，应用 Greenhouse-Geisser 方法进行校正。最后，本章应用 Bonferroni 事后检验对有显著差异的组进行组间的成对比较。

3.4.2　统计分析

3.4.2.1　问卷和行为反应

本章使用了一个实验后问卷收集用户对所有问题的熟悉程度，采用五级 Likert 量表（非常熟悉、有点熟悉、既不熟悉也不陌生、有点陌生、完全陌生）。约三分之一的问题被报告为用户熟悉的（非常熟悉：21.07%，有点熟悉：16.85%），而另三分之一的问题被报告为用户不熟悉的（有点陌生：26.9%，完全陌生：3.78%）。其余的问题被报告为既不熟悉也不陌生（31.4%）。在不同熟悉程度之间的 ERP 分析没有显示出显著差异。结果表明，无论用户的熟悉程度如何，阅读过程都会在大脑中引发相似的模式。

行为反应通过二元测试的准确率和反应时间进行分析。在普通测试中，完全相关、相关和不相关的句子的判断准确率分别是 97.93%、92.03% 和 89.98%。反应时间分别为：完全相关为 1.00 秒，相关为 1.29 秒，不相关为 1.39 秒。这些结果表明，考虑到句子的分级相关性，行为反应相应地有所不同。因此，我们可以推测这些差异背后存在一些神经活动相关的因素。

3.4.2.2　ERP 成分

ERP 分析观察到的差异的显著性水平如表 3.7所示。此外，图 3.9提供了不同类型词（答案词、语义相关词和普通词）在中央脑区的 ERP 波形的总体平均值。在不同时间窗口上，本章有以下发现：

120-320 毫秒　在 120-320 毫秒的时间窗口中，由不同类型词生成的 P200 波形的差异略微显著（$p < 0.05$，位于额叶和顶叶）。N100-P200 幅度（从 N100 到 P200 的平均波形变化）在额叶（$F[2, 40] = 19.51, p < 0.001$）、中央脑区（$F[2, 40] = 20.94, p < 0.001$）和顶叶（$F[2, 40] = 29.14, p < 0.001$）中存在高度显著的差异。Bonferroni 事后检验表明，答案词的 N100-P200 幅度显著高于语义相关词（$p < 0.001$）和普通词（$p < 0.001$）。

先前的研究表明，阅读中的较低认知负荷与 N100-P200 幅度的增加相关[89]。答案词中 N100-P200 幅度的增加可能表明，当用户定位到答案时，认知资源需求减少。

320-520 毫秒　在词刺激开始后 320-520 毫秒时间窗口内 N400 成分波形的总体平均值在不同脑区上差异显著，包括中央脑区（$F[2,40] = 12.57, p < 0.001$）、右侧颞叶（$F[2,40] = 17.34, p < 0.001$）和顶叶（$F[2, 40] = 15.59, p < 0.001$）。Bonferroni 事

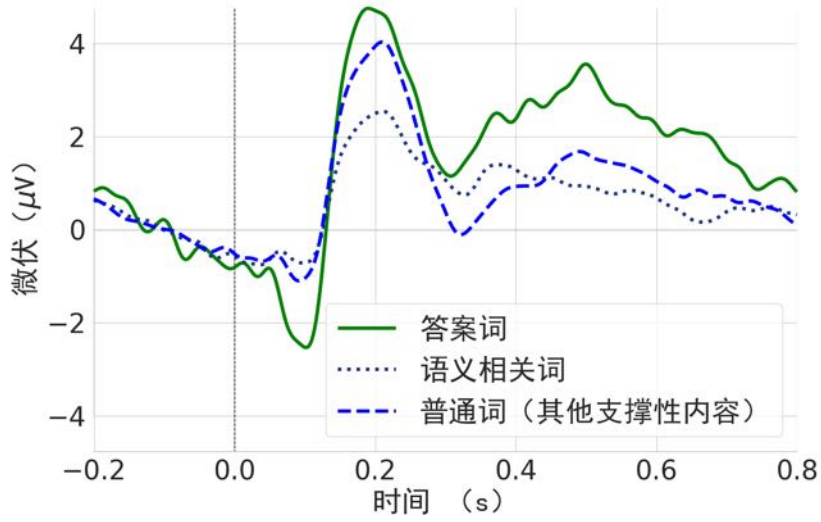


图 3.9　不同词类型在中央脑区（Cz + FCz + C3 + C4 + FC3 + FC4）的 ERP 波形总体平均值

后检验发现，答案词在 N400 中的平均负波显著小于语义相关词（$p < 0.05$）和普通词（$p < 0.001$）。此外，在电极 T4 和 T6 上，语义相关词的平均负波显著小于普通词（$p < 0.05$）。

N400 与处理即将到来的词的语义上下文关联性密切相关[146-147]。在当前语义上下文中词的 “期望值” 越高，通常会导致较小的 N400 负波。本章的统计分析表明，答案词的 N400 负波小于语义相关词。语义相关词的 N400 负波则小于普通词。“期望值” 的发现与 P200 成分中关于认知负荷的发现有一定的关联性，因为较高 “期望值” 的词可能需要较少的认知资源。此外，本节还发现语义相关词比普通词具有更高的 “期望值”。

520-750 毫秒　在 520-750 毫秒时间窗口内刺激诱发的平均 P600 波形显示出在中央脑区（$F[2, 40] = 17.45$, $p < 0.001$）、左侧颞叶（$F[2, 40] = 15.87$, $p < 0.001$）和顶叶（$F[2, 40] = 20.27$, $p < 0.001$）的显著效应。Bonferroni 事后检验表明，在中央脑区，答案词在 P600 中的平均正波显著大于语义相关词（$p < 0.001$）和普通词（$p < 0.001$）。此外，在左侧颞叶，语义相关词的平均正波显著小于普通词（$p<0.01$）。

传统上，P600 被认为与句法异常相关①。最近的研究表明，P600 可能与语义主题异常[148]和认知推理[149]也相关。在信息检索场景中，Eugster et al.[72]发现相关词会引发更高的 P600 幅度。Pinkosova et al.[69]指出，更高相关性与 P600 幅度之间的联系可能来自大脑中的语篇记忆（Discourse memory）。在本节的研究中，我们人为检查后发现实验中使用的句子在句法层面没有问题。因此，不同内容之间的差异

① https://en.wikipedia.org/wiki/P600_(neuroscience)

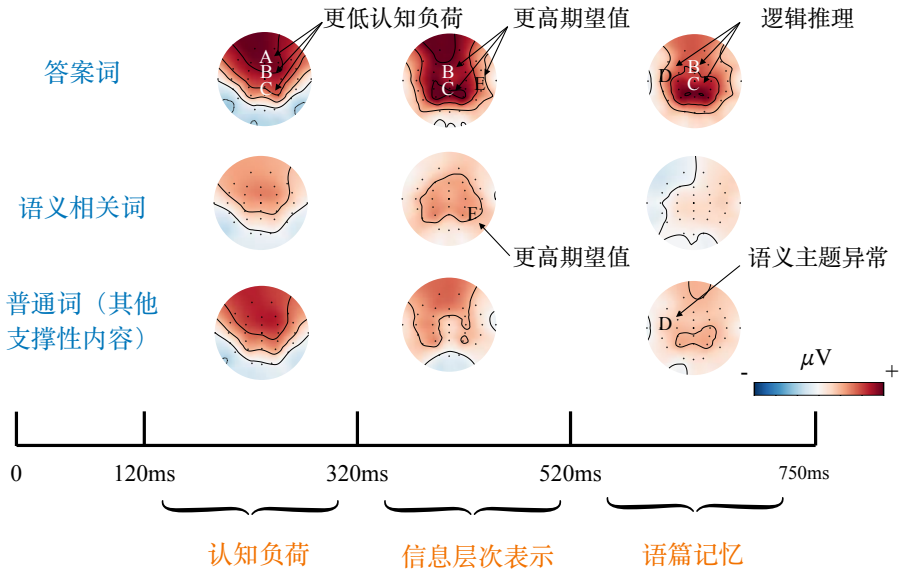


图 3.10　三种词类型在各时间窗口平均的 ERP 幅值的拓扑图及相关心理现象。A、B、C、D、E 和 F 分别指额叶、中央、顶叶、左侧颞叶、右侧颞叶和电极 T6。120-320 毫秒时间窗口的上下界为 ±5 μV，其他时间窗口为 ±3 μV。

更可能是由语义主题异常和逻辑推理引起的。这两个方面被 Pinkosova et al.[69] 指出也与语篇记忆有关。

更具体地说，P600 在答案词中最高，其次是普通词，而在语义相关词中最低，尤其是在与语言识别相关的左侧颞叶。对于答案词，逻辑推理相关的思维活动会在人的大脑中启动，导致显著更高的 P600。类似地，语义相关词也可能与推理处理有关，但程度较小。这两种词在语义上是相关的，因此与语义主题异常的关系最小。然而，对于普通词，语义主题异常相较于语义相关词更为显著，因为它可能涉及当前语义主题之外的信息。因此普通词也会有相对较高的 P600 幅度。总的来说，语义相关词可能会消耗相对较少的语篇记忆，因此其 P600 幅度同时低于普通词和答案词。然而，这一有趣的现象有可能需要进一步探索以验证其潜在的神经机制。

3.4.2.3　讨论

本章的研究为阅读理解过程中处理关键信息和普通信息的神经差异提供了见解。一方面，跨时间窗口的 ERP 分析显示，在阅读理解过程中处理关键信息和普通信息的神经差异是存在的。另一方面，我们认为各种认知活动，如认知负荷、语义主题理解、逻辑推理，支撑了这些神经反应（总结于图 3.10）。与以往使用眼动数据的研究不同[101]，本章的发现建立在涉及人类如何处理文本信息的更深层次的认知水平上。这些认知差异可以帮助我们理解阅读理解过程，进而启发更加主动和人性化的搜索系统，为信息检索系统设计提供若干见解：

（1）**文档排序：** N100-P200 幅度上的发现表明，当参与者定位到答案时，认

知资源需求减少。认知资源的需求，即认知容量[150]，会影响用户与外部系统的体验。此外，Jiang et al.[150]指出，认知容量的减少可能导致阅读理解的任务成功率降低，并与阅读理解情境中的误解现象有关。因此，从认知容量的角度来看，我们认为文档结构中包含易于访问的潜在答案内容是重要的。一种更好的文档结构形式是结合简明扼要的关键信息内容和详尽的补充内容。在实践中，搜索引擎在构建排序模型时应在整体相关性的基础上，也考虑细粒度的文档结构，特别是潜在答案的位置和显示样式。

（2）**搜索结果摘要构建：**当用户定位到答案时，我们推测他们会将额外的认知资源转移到其他神经功能上（例如，扩展工作记忆的容量以用于信息回忆和管理）。其中，答案词相关的 P600 效应暗示了逻辑推理的过程在这一阶段于大脑中发生。此外，本章还发现语义相关内容所需的逻辑推理功能较少，因此其 P600 效应甚至小于其他支撑性内容。在当前的搜索界面设计中，我们通常会在搜索引擎结果页上看到包含大量语义相关内容但忽略其他支撑性内容的搜索结果摘要。虽然提供大量语义相关内容使搜索结果更具吸引力（如本章的 ERP 分析中所示，会给用户带来更高的期望值），但在某些情况下可能导致用户实际点击后的不满意。因此，我们建议搜索引擎在提取摘要时应做更加公平的考虑，不仅要展示和语义相关的内容，还要考虑展示内容能否为更好的用户理解提供证据和背景。

（3）**BCI 增强的搜索系统：**随着脑机接口设备变得低成本和便携 ①，研究人员认为在不久的将来，BCI 有可能被应用于在线教育、网络浏览和搜索[151]。由于阅读理解在这些场景中是一个常见任务，利用 BCI 更好地理解阅读状态是可能的，我们相信这将有利于更好的人机交互。例如，通过 BCI，搜索引擎可以理解哪些内容更能够满足用户，并进一步提供更多有用信息，尤其是在用户意图不明确的情况下。基于在用户阅读不同内容时在神经活动上存在可检测差异的发现，本章将进一步去探索使用脑电信号作为阅读理解任务隐性反馈的有效性，并在第 3.4.3节中详细阐述。

3.4.3　实验与讨论

为了探索阅读过程，本章基于用户研究中收集的 EEG 数据进行了两个实验任务，即答案句分类和答案词提取。答案词提取任务和答案句分类任务都是一个二分类问题，分别用于估计用户阅读的某个词是否是答案和某个句子为完全相关的概率。这两个任务在机器阅读理解和信息检索的研究中都至关重要。传统上，这两个任务的输入一般是基于阅读理解的文本。然而，与以往的研究不同，本章使用 EEG 信号作为输入，而不是基于阅读理解的文本，前者是一种更加个性化的用

① https://the-unwinder.com/reviews/best-eeg-headset/

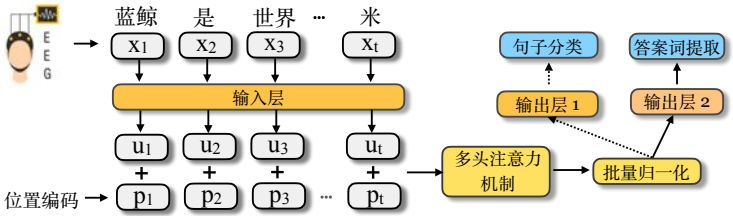


图 3.11　基于 EEG 的阅读理解状态检测框架结构

户反馈信号。

3.4.3.1　框架设计

为了用统一框架解决这些问题并展示 EEG 信号的有效性，本章提出了一个基于 EEG 的阅读理解状态检测框架（Unified framework for EEG-based Reading Comprehension Modeling，UERCM）。UERCM 为这两个任务提供了一个通用框架，同时利用可学习的位置编码和注意力机制，以捕捉句子中 EEG 特征的局部交互作用。

如图 3.11所示，针对一个特定的词级别 EEG 序列 $X \in \mathbb{R}^{t\times d} = [x_1, x_2, ..., x_t]$，其中 t 是句子长度，d 是词级别 EEG 特征的长度，我们首先应用一个输入层将其线性投影到 h 维向量空间，其中 h 是隐藏层维度：

$$U = W_h X + b_h \tag{3.13}$$

其中 $W_h \in \mathbb{R}^{d\times h}$ 和 $b_h \in \mathbb{R}^{t\times h}$ 是可学习参数，$U \in \mathbb{R}^{t\times h}$ 是隐藏层向量，随后将作为多头注意力层的输入。然后，我们在向量 U 中添加位置编码 $P \in \mathbb{R}^{t\times h} = [p_1, p_2, ..., p_n]$，得到 $U' \in \mathbb{R}^{t\times h} = U + P$。本节采用可学习的位置编码代替基于正弦函数的位置编码，因为前者在实验中表现更好。之后，一个多头注意力层被用于计算局部交互序列：

$$Z = \text{MultiHead}(U', U', U') \tag{3.14}$$

其中 $Z \in \mathbb{R}^{t\times h}$ 是输出向量。接下来，我们应用一个批量归一化层加速和稳定训练过程，得到 $Z' = \text{BatchNormalization}(Z)$。Vaswani et al.[87] 建议在多头注意力层后使用层归一化，在各种自然语言处理任务中获得了性能提升。尽管如此，本章发现批量归一化在实验中表现更好。其原因可能是归量归一化可以缓解 EEG 特征稳定性差的影响，这是自然语言处理中的预训练词嵌入所不存在的问题。除了批量归一化的设计，本章提出的框架在注意力层的数量上也不同于其在自然语言处理任务中的应用。例如，UERCM 框架只使用了一个注意力层，这是因为本章的实验发现单个注意力层效果要优于更多的注意力层堆叠。

在此之后，给定表示 Z，本章针对两个任务采用不同策略进行聚合以获得答

案句分类（$\hat{y}_s \in \mathbb{R}^1$）和答案词提取（$\hat{Y}_o \in \mathbb{R}^t = [\hat{y}_{o,1}, \hat{y}_{o,2}, ..., \hat{y}_{o,t}]$）的预测：

$$\hat{y}_s = \text{softmax}(W_s \text{ReLU}(\text{Concat}(z_1, z_2, ..., z_t)) + b_s) \tag{3.15}$$

$$\hat{y}_{o,i} = \text{softmax}(W_o \text{ReLU}(z_i) + b_o), i = 1, 2, ..., t \tag{3.16}$$

其中 $W_s \in \mathbb{R}^{th \times 1}$, $b_s \in \mathbb{R}^1$, $W_o \in \mathbb{R}^{h \times 1}$, 和 $b_o \in \mathbb{R}^1$ 是线性输出层的参数。

最后，本章采用交叉熵函数作为学习目标，一个样本句子的损失 L_s（在答案句分类任务中）和 L_o（在答案词提取任务中）为：

$$L_s = -y_s log\hat{y}_s + (1 - y_s) log((1 - \hat{y}_s)) \tag{3.17}$$

$$L_o = -\sum_i (y_{o,i} log\hat{y}_{o,i} + (1 - y_{o,i}) log((1 - \hat{y}_{o,i}))) \tag{3.18}$$

其中 $y_s \in \mathbb{R}^1$, $y_{o,i} \in \mathbb{R}^1$ 分别为句子标签和词标签的真实值。

在训练过程中，本章分别尝试了独立或联合优化这两个任务，它们的表现相似。因此，本章仅报告独立训练这两个子任务的实验结果。

3.4.3.2　EEG 特征

和第3.3节一致，本章提取了频域特征和时域特征来作为模型输入。在阅读理解场景下，EEG 的不同波段的频域特征与注意力（δ 和 β[152-153]）、认知表现（θ 和 α[153]）和语义违背（γ[154]）相关。先前的研究也发现过频域特征在相关性预测中的有效性[74,155]。另一方面，用户接收到刺激时特定短时间窗口内的时域信息也是常用的 EEG 特征。在神经科学的先有研究中，时域特征中的 ERP 成分（如 N170 和 P300）在目标检测中的有效性已被展示[156]。在第 3.4.2.2节中的分析也显示了 ERP 与阅读理解过程中用户心理活动的潜在相关性。基于上述原因，本章的研究中提取了频域特征和时域特征作为 EEG 特征。

具体来说，本章选择了来自三个大脑区域（中央、右颞叶和顶叶）的 EEG 特征。原因是这些区域在不同类型词的认知反应中具有显著差异（参见第 3.4.2.2节）。对于频域特征，本章从 δ（0.5-4Hz）、θ（4-8Hz）、α（8-13Hz）和 β（13-30Hz）的频率带计算平均带功率和微分熵。对于时域特征，本章分别从 P200（120-320ms）、N400（320-520ms）和 P600（520-750ms）中均匀采样五个时间点。因此，对应于每个词的 EEG 特征是一个 69 维的向量（频域：2 个特征 ×4 个频段 ×3 个区域，时域：5 个时间点 ×3 个 ERP×3 个区域）。

特征分析　为了探索不同 ROI 和时间窗口中时域特征的有效性，本章使用 SHapley 特征解释（SHapley Additive exPlanations，SHAP）[157] 来研究特征重要性。SHAP

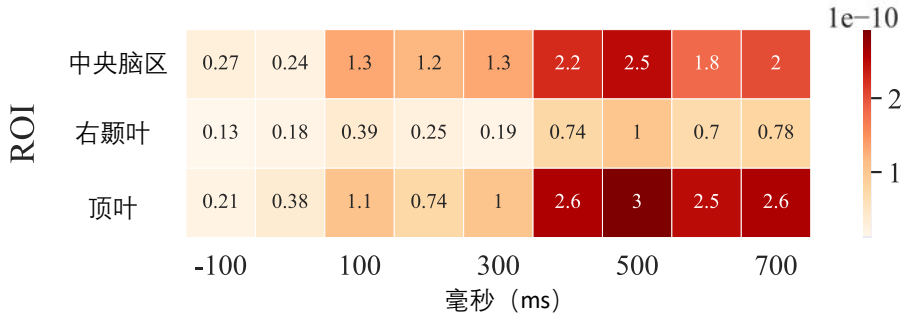


图 3.12　用户阅读理解任务中脑电特征的平均 SHAP 值。较高的 SHAP 值表示较高的特征重要性。

是一种可以解释机器学习模型中特征重要性的方法。图 3.12展示了在一个线性回归模型中，答案词提取任务中特征的平均 SHAP 值。颜色越暖表示 SHAP 值越高，表明该特征在答案词提取中更关键。结果显示，顶叶和中央脑区中特征重要性的贡献超过右颞叶。并且，400ms 之后的 EEG 特征具有更高的重要性，这可能是因为视觉信息的认知编码需要一定时间，同时 320-750ms 中的 N400 和 P600 成分与语义理解相关。此外，N100，即脑电信号在 100ms 附近的负波，相比相邻时间窗口中的其他成分中具有更高的特征重要性。考虑到 N100 尚未涉及语义加工，本章在前述分析中并未详细介绍。作为一个早期成分，N100 被认为与语义理解发生之前的视觉辨别过程有关[88]。它和定向反应或与先前经验的刺激进行“匹配过程”[158]有关。因此，N100 的特征重要性很可能和用户在阅读理解过程中对即将出现的答案匹配过程的心理预期有关。

3.4.3.3　实验基线与数据划分

本章采用四个监督学习模型 SVM、MLP、梯度提升决策树（Gradient Boosting Decision Tree，GBDT）和 RNN 作为基线。其中答案词提取任务作为一个二分类问题，可以通过二分类器 SVM、MLP 和 GBDT 直接解决。对于 RNN，由于其迭代结构，本章结合其与随机向量场的方法将二分类问题视为序列标记任务，以预测其句子级上下文中的每个词的标签。此外，本章也汇报一个未训练模型的结果作为基线，其中所有预测基于随机选择，因此其 AUC 为 0.5。在答案句分类任务中，本章将完全相关的句子视为正样本，其他句子都视为负样本。此外，如果将该任务视作是对一个问题下对应的三个不同相关性的句子进行排序时，它也可以被视为一个排序问题。对于 SVM、MLP 和 GBDT，一个句子被预测为正样本的概率是基于答案词提取任务中每个词的预测答案概率综合计算的。更具体地说，句子的得分 S 可以表示为：

$$S = \frac{max(W_1, ..., W_n) + mean(W_1, ..., W_n) + median(W_1, ..., W_n)}{3} \tag{3.19}$$

其中 W_i 表示给定句子中第 i 个词的得分（预测概率），n 是该句子中的词数。根据公式 3.19，句子的得分是句子中词的得分的最大值/中值/均值的平均值。此方法将词级别的预测信息集成到句子级别，参考了现有工作中基于眼动特征的注意力估计任务中的方法[159]。对于 RNN 模型，本章将句子中每个词的 EEG 特征输入网络，再将最后一个隐藏层连接到一个全连接层以获得概率分布。其中，句子级相关性标签用于计算损失。注意，答案句分类任务中的 RNN 不需要结合 CRF 模块。最后，本章也对比一个未训练模型，其在该任务下的 AUC 为 0.5，MAP 为 0.615。本章不比较其他交互特征输入的基线，因为（1）本章的用户研究不涉及鼠标移动或眼动等其他交互行为记录。（2）本章所提出的框架需要适应各种互联网浏览和离线阅读的情况，在这些情况下，并没有用户提交的查询等文本内容特征可用。

为了验证在不同设置下的有效性，本章执行两种数据划分策略来进行实验：任务独立和用户独立。任务独立策略将每个问题及其对应的句子划分为十折，然后在验证每一折时使用其余折进行训练。用户独立策略在验证模型在每个参与者上的性能时使用其余参与者的数据来训练监督模型。在评价指标上，AUC 同时被用于答案词提取和答案句分类任务，而 MAP 被用于答案句分类任务，因为答案句分类任务也可以被视为一个排序问题。本章计算其他模型相对于未训练模型的提升（ΔAUC 和 ΔMAP）以说明脑信号的有效性。

3.4.3.4　参数设置

本节详细说明 UERCM 和其他基线的参数设置。对于所有的模型，参数的设置分别根据两种数据划分策略（即用户独立和任务独立）中的平均 AUC 进行调整。

UERCM 使用 Adam 优化器以 8 的批量大小端到端训练。对于超参数，注意力头的数量、隐藏维度和学习率分别从 $\{4, 8\}$、$\{16, 32\}$ 和 $\{10^{-4}, 10^{-3}, 10^{-2}\}$ 中选择。此外，为了加速训练过程，本章在 NVIDIA TITAN XP 12G GPU 上训练 UERCM，并在验证性能在五次迭代后不再提高时采用早停策略。UERCM 的实现代码基于 PyTorch ①。

对于其他基线模型，参数设置的详细信息如下：（1）对于 SVM，本章使用径向基函数核，并从 $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$ 中选择正则化参数，根据数据分布自动选择核系数 ②。（2）对于 MLP，学习率和隐藏维度分别从 $\{10^{-4}, 10^{-3}, 10^{-2}\}$ 和 $\{16, 32, 64\}$ 中选择。（3）对于 GBDT，参数包括学习率、估计器数量、叶节点和最大树深度，其超参数分别从 $\{10^{-4}, 10^{-3}, 10^{-2}\}$、$\{100, 200, 400\}$、$\{4, 8\}$ 和 $\{4, 8\}$ 中选择。

① https://pytorch.org/
② https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html

表 3.8　答案词提取和答案句分类的实验结果[a]

模型	答案词提取		答案句分类			
	$\Delta AUC_{任务独立}$	$\Delta AUC_{用户独立}$	$\Delta AUC_{任务独立}$	$\Delta MAP_{任务独立}$	$\Delta AUC_{用户独立}$	$\Delta MAP_{用户独立}$
SVM	0.072*†	0.069*†	0.092*†	0.065*†	0.103*†	0.078*†
MLP	0.079*†	0.084*†	0.141*†	0.077*†	0.122*†	0.086*†
GBDT	0.086*†	0.077*†	0.097*†	0.079*†	0.125*†	0.074*†
RNN（+CRF）	0.146*†	0.151*	0.132*	0.089*†	0.165*†	0.101*†
UERCM	**0.152***	**0.157***	**0.173***	**0.147***	**0.236***	**0.179***

[a] ∗/† 分别表示与未训练模型/UERCM 的差异在 p 值 < 0.05 下显著。

3.4.3.5　结果与讨论

表 3.8展示了答案词提取任务和答案句分类任务在两种数据集划分策略（任务独立和用户独立，见第 3.4.3.3节）下的实验结果。从表 3.8可以看出，所有基于 EEG 特征训练的模型均显著优于未训练模型。这些结果证明了使用 EEG 数据来定位答案词和监控用户寻找答案过程的可行性。此外，在不同的数据划分策略中，UERCM 都取得了显著的性能提升，尤其是在答案句分类任务上。这表明注意力机制策略和脑信号序列建模使 UERCM 能够大幅度超越传统机器学习基线（SVM 和 GBDT）和深度神经网络基线（MLP 和 RNN（+CRF））。

进一步，本章分别深入探讨不同模型在这两个任务中的表现，观察到：

（1）对于答案词提取，基线模型 SVM、MLP 和 GBDT 表现显著差于 UERCM。原因可能是它们将任务视为对每个词的二元分类问题，而忽略了序列信息。相反，采用序列建模策略的模型，即考虑序列交互信息的 RNN（+CRF）和 UERCM，显著优于其他基线。尽管 UERCM 在该任务中没有显著优于 RNN（+CRF），但其也有一定的优势。例如，UERCM 可以实现并行计算，而 RNN 由于其迭代性质无法并行。对于实时 BCI 设备，UERCM 可以加速推理过程并利用实时的隐式反馈来提升系统性能。

（2）对于答案句分类，UERCM 在 ΔMAP 上取得了显著的改进，并在两种数据划分策略中显著优于其他基线。尤其是，本章发现 RNN（+CRF）尽管也考虑了序列建模，但表现显著差于 UERCM。这种现象可能是由脑信号的特性引起的。脑信号通常包含如眨眼和心跳引起的电位波动。尽管本章应用了标准的预处理方法，但脑信号的数据质量仍不稳定。对于 RNN 模型，其性能可能会受在迭代过程中遇到的质量较差的信号的影响。但本章提出的 UERCM 是相对稳定的，因为它可以自动学习不同质量的信号对于序列学习的权重，并从序列中的其他数据中提取有

价值的信息。

3.5　本章小结

本章基于脑机接口，开展了对用户在信息检索过程中的神经认知过程理解与建模的研究。一方面，脑机接口能够提供与人类大脑认知活动直接相关的反馈，因此在表征用户满意度方面，相比基于文本相关性的满意度预估要更为准确和个性化。基于此，本章设计了一个深度模型，利用多中心球坐标编码和空间注意力机制，对信息检索过程中的脑拓扑结构进行建模，从而更准确地预估用户满意度。实验结果表明，该方法在搜索和推荐场景中都取得了优于现有的最先进 EEG 分类模型的效果。同时，本章也探讨了基于脑信号推断的满意度在搜索和推荐场景中的应用潜力。另一方面，脑机接口可以提供更细粒度的用户信号，而传统上的用户认知过程研究粒度较粗，一般是在文档、段落等粒度上进行研究。本章通过基于 EEG 的用户研究，深入探讨了用户在文本阅读理解过程中的细粒度认知机制，研究了用户在阅读理解过程中在处理关键信息（如答案内容和语义相关内容）和其他信息的大脑活动差异。该研究揭示了与认知负荷、逻辑推理、语义主题匹配等相关的特定 ERP 成分在用户阅读不同内容时的差异。基于这些发现，本章提出了若干信息检索系统设计的见解，如在排序模型中应考虑文档的细粒度结构和在文档摘要构建中应提供丰富的支撑性信息。同时，本章也提出了一个基于 EEG 的用户阅读状态检测模型，通过采集和分析用户在阅读过程中的脑信号，可以更准确地对文档内容与用户需求之间的匹配程度进行判断。综合以上两个方面，本章在现有的关于信息检索场景下的用户认知过程研究的基础上，引入脑机接口从更细粒度、更真实的角度对用户认知过程进行建模，提供了以前的用户研究技术（如眼动追踪、鼠标移动和显式标注等）无法获得的神经层面的发现。本章同时设计了相关的用户状态检测模型并开展了在下游应用上的实验，为后续构建基于脑机接口的信息需求解码和用户反馈相关算法提供了有力支撑。

本章和**用户满意度建模**方法相关的工作发表于 CCF A 类会议 Multimedia 2022 的长文[15]，和**文本阅读理解过程中的细粒度认知机制**相关的工作发表于 CCF A 类会议 TheWebConf（旧称 WWW）2022 的长文[14]，概述脑机接口在信息检索系统中协助用户认知过程与理解的过程的工作发表于 CCF A 类会议 Multimedia 2024 的 Workshop 论文[16]。

第 4 章　基于脑机接口的信息需求解码

4.1　本章引言

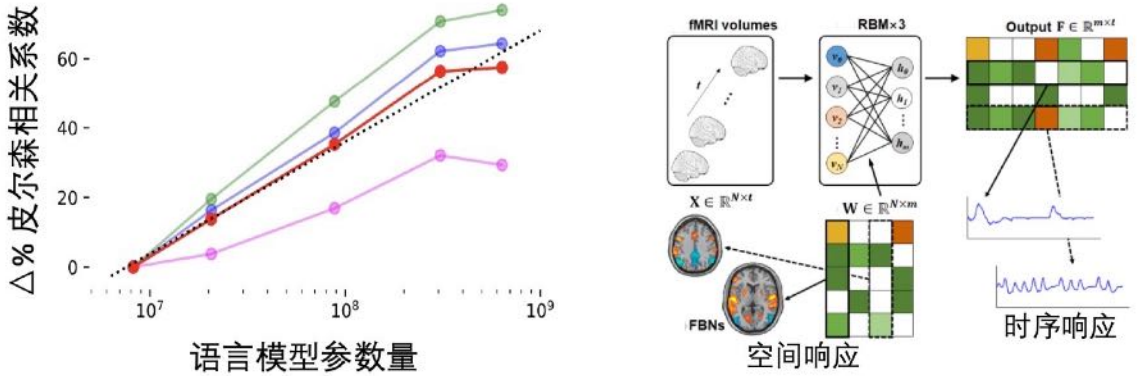
理解用户信息需求对于任何信息系统的有效性至关重要。传统上，用户信息需求的表达一般是基于文本、图片、语音等形式，并且信息系统的表现很大程度上依赖于信息需求表达的准确程度。例如，在搜索引擎当中，用户输入的查询词的准确性很大程度上影响了相关文档检索的准确率。在智能助手应用（如 ChatGPT）当中，系统生成准确的反馈往往依赖于用户精确的提问和需求说明。然而，信息系统的用户常常难以准确表达他们的信息需求。例如，由用户构建的文本输入往往简短[4]、模糊[160]、或措辞不准确[161]，从而影响了信息检索的效果（参见图 4.1）。因此，在信息系统中，如何更好地进行用户信息需求理解仍然是悬而未决的挑战。

脑机接口作为一种新兴的交互媒介，提供了从大脑信号直接解码用户信息需求的可能性，并为理解和增强用户的信息获取过程提供了全新的视角和技术手段。现有的研究表明，脑机接口能够以分类任务的形式从大脑记录中解码出有意义的语义信息，例如从一组词[162-163]、句子[53,164]和主题[165]中选择和脑信号更为匹配的一项作为解码内容。在最近的研究中，Moses et al.[166]使用植入感觉运动皮层的电极，从一位失语症患者的大脑记录中成功解码了 50 个词汇中的目标词。Pereira et al.[164]利用非侵入式的 fMRI 数据，从一对或一组以视觉刺激呈现的句子中解码目标句子。在信息检索场景中，也有若干研究表明大脑信号可以用来预测用户的相关性感知[19,62,69]和认知状态[66]。这些进展为使用大脑信号作为用户输入的补充信号开辟了新的途径。

然而，构建基于脑机接口的用户需求理解系统还存在两个明显的挑战。一是如何从脑信号中更准确地提取复杂、多变、难以预先定义的用户意图。现有的脑语言解码方法中通常都将解码任务视作一个分类任务，从多个预定义的项目（词、句子等）中选择和脑信号更为匹配的一项。这种基于分类的脑语言解码方法难以直接在解空间广泛的用户意图解码任务下得到应用。二是如何将脑机接口的用户信息需求解码和下游的信息系统服务更有机地结合起来，从而为用户提供有效的反馈内容。该挑战的关键在于解码的信息内容需要有优化信息系统的相关功能实现的潜力，同时需要设计反馈环路来利用解码出来的信息。

针对第一个挑战，近年来预训练的大语言模型，尤其是基于生成设置的模型[167-169]，已成为计算语言建模的主导方法，也为我们构建更高效的基于文本的脑信号解码算法提供了契机。这些大语言模型将语言构建过程视为一个生成问题。

图 4.1　信息系统的用户常常难以准确表达他们的信息需求



(a) 大脑信号与语言模型表征相关，且该相关性和语言模型的参数量的对数成正比[174]

(b) 语言模型表征与从大脑信号中提取的时序和空间信息显著相关[175]

图 4.2　现有工作中关于语言模型与大脑信号的相关性分析

给定一个文本提示，预训练模型会根据从大量文本中学到的统计语义知识生成最有可能的续写内容。通过以自回归的方式解决语言生成问题，语言模型可以构建语义和句法连贯的连续文本[169]。利用语言模型的强大能力，最近有一些工作开始尝试基于语言模型提升语言脑机接口[53,170]的能力。例如，Tang et al.[53]使用预训练语言模型预构建一组可能的语言候选，然后根据它们与从 fMRI 数据解码的语义表示的相似性选择最佳的一个。

然而，这些现有方法将大脑解码和语言生成视为两个独立阶段，从大脑信号中提取的语义信息仅在语义候选生成之后才被利用起来。尽管预训练语言模型已经有很好的模拟人类语言的能力，但它们仅基于训练材料生成最可能的内容[167-168]。换句话说，预训练语言模型生成的语言未必反映从大脑记录中解码的语义。因此，将大脑信号直接整合到语言生成过程仍是一个开放且未解决的挑战。同时，越来越多的研究强调语言模型和人脑在计算原则之间的相似性[171-173]（参见图 4.2）。这些研究同时也表明大脑表征和大语言模型的表征具有一定的关联性。

在此背景下，本章提出了基于生成式大语言模型的脑语言解码的方法（Brain

Language Reconstruction with Generative Large Language Models，BrainLLM）。Brain-LLM 将从大脑信号中解码的语义表征直接用于连续语言的生成阶段。与现有工作[53,170]不同，BrainLLM 在语言生成阶段直接整合大脑信号，从而更好地在语言生成阶段就利用了脑信号的信息。本章在三个 fMRI 数据集中测试了 BrainLLM 的性能，并发现其显著强于已有的先生成后分类的方法[53]和其对照模型。此外，该方法为神经科学和生物医学工程的相关研究提供了新的潜在应用。例如，BrainLLM 可以用于计算任何语言内容的生成概率，而不再局限于有限数量的预定义候选，从而能够促进语言解码系统的开发、人脑语言编码方式的研究和应用于用户信息需求解码场景。

为了解决第二个挑战，本章进一步研究了从大脑中解码出来的语义信息如何能被信息系统利用。本章在搜索系统的查询扩展（也称为查询增强）场景中研究了该问题。查询扩展是信息检索领域的一种技术，通过在用户原始查询中添加相关词语或语义信息，以提高检索结果的准确性和覆盖范围[129,176]。其目的在于弥补用户查询中可能存在的模糊性和歧义性，帮助检索系统更好地理解用户意图，从而返回更符合用户需求的结果。传统上，查询扩展依赖于外部文档信息，例如一个常见方法是从用户已交互的历史文档内容来丰富查询表征[34,177-178]。然而，现有的查询扩展过程仍依赖于最初检索到的文档的质量，并且在获取用户与初始返回的文档的交互信号之前无法启动。

在这样的背景下，本章将 BrainLLM 和查询扩展技术结合，将大脑信号作为额外输入来增强查询表示，该方法命名为基于脑信号的查询增强（Query Augmentation with Brain Signals，Brain-Aug）。Brain-Aug 采用了和 BrainLLM 一致的策略进行大脑语义的解码，但使用了不同的推理方法，旨在查询扩展的内容既需要符合用户语义，也需要能够增强文档的区分能力来优化查询效果。在多个 fMRI 数据集和不同检索系统上，本章发现 Brain-Aug 能够准确解读用户意图并增强搜索引擎性能。并且，本章发现 Brain-Aug 不仅显著优于传统的查询扩展方法，并且也能与已有方法结合。进一步的分析显示，Brain-Aug 在模糊查询上表现更好，这暗示了基于脑机接口的信息需求解码方法的潜在应用场景。

本章的结构如下：第 4.2节介绍相关工作，包括基于脑机接口的语言解码和查询扩展。第 4.3节介绍了生成式大语言模型的脑语言解码的方法。第 4.4节介绍了该方法在查询扩展场景下的应用。第 4.5节对本章进行小结。

4.2　相关工作

4.2.1　基于脑机接口的语言解码

语言脑机接口在语言解码中主要致力于从大脑信号中提取和重构语言信息。从采集设备而言，语言脑机接口的研究通常使用侵入式和非侵入式的脑机接口技术。侵入式技术如皮层脑电图（ECoG），在解码精度和实时性方面表现优异，已被用于从失语症患者的脑信号中解码出基本的词汇[166]。另一方面，非侵入式技术包括功能性磁共振成像（fMRI）和脑电图（EEG）等，由于其安全性和易用性，得到了比侵入式技术更广泛的应用。例如，Pereira et al.[164]利用 fMRI 数据成功解码了语义信息，从而实现了从视觉刺激中选择目标句子的任务。

从实验范式和算法技术而言，语言解码的实验设置包含两大类，基于动作指令的解码和基于语义感知的解码。基于动作指令的解码方法是将语言解码任务简化为可以通过一系列动作执行来实现的任务。例如，常见的方法包括通过脑机接口在键盘上选择按键以实现单个字母的拼写[11]、或用脑机接口控制光标实现语言内容的手写[179]等。这类方法的神经学基础实际上与大脑的语言区功能关系不大，而是从运动、视觉等相关脑区信号解码出预定义的动作指令，再将其转化为文本内容。其技术优势在于可以和运动想象脑机接口[8]、稳态视觉诱发电位脑机接口[11]等成熟的技术结合，从而实现较高的解码准确率。然而，这样的方法需要用户通过学习和训练相关动作指令来熟悉脑机接口的使用，并且需要用户主观付出较高的精力来进行内容输入。随着神经科学和机器学习技术的进步，近年来基于语义感知的解码也得到了越来越多的关注。基于语义感知的解码侧重于理解和提取大脑中与语言理解和语义处理相关的活动。通过分析与语言理解、语义整合以及概念形成相关的脑区活动，语言脑机接口可以在大脑处理特定语言输入、想象相关语言内容时提取相关特征。相比基于动作指令的解码方法，语言脑机接口仍面临诸如信号噪声、个体差异以及解码效率等挑战[180]。不过，基于语义感知的解码方法的优势在于其可以在用户与信息系统交互的同时，不影响对用户想法的解码和理解。这意味着在用户专注于自然交流或信息处理的过程中，系统可以实时解码用户的语义意图，为实时语义检索、智能助手或沉浸式虚拟现实系统等应用提供支持。

4.2.2　查询扩展

传统上，查询扩展可以分为两类：基于伪相关信号的方法[129,181]和基于用户信号的方法[182]。基于伪相关信号的方法通常将初始检索步骤中排名靠前的文档视为相关文档。在这些相关文档的基础上，Lavrenko et al.[129]和 Rocchio Jr[33]分别采用向量空间模型和语言模型来优化查询表示，使其更接近排名靠前的文档。例

如，Rocchio 的方法基于以下公式扩展原有查询：

$$\vec{Q}' = \vec{Q}_0 + \alpha \frac{1}{k} \sum_{\vec{D}_j \in D_{top\_k}} \vec{D}_j \tag{4.1}$$

其中，$\vec{Q}_0$ 是原始查询的语义向量，$\vec{Q}'$ 是增强后的查询向量，D_{top_k} 是初次查询后排名前 k 位的文档，$\vec{D}_j$ 是第 j 个文档的语义向量。

与伪相关信号不同，基于用户信号的方法通常整合用户之前交互过的文档或历史提交的查询中的信息。例如，Ahmad et al.[177]和 Chen et al.[34]构建了一个序列模型，从历史点击文档中提取语义表示以优化查询表示。然而，现有的方法，无论是基于伪相关信号还是用户信号，都受限于对文档质量的依赖以及对其相关性估计的准确性。例如，Li et al.[183]指出当反馈信号估计的文档相关性不准确时，扩展后的查询效果还不如原有的查询效果。

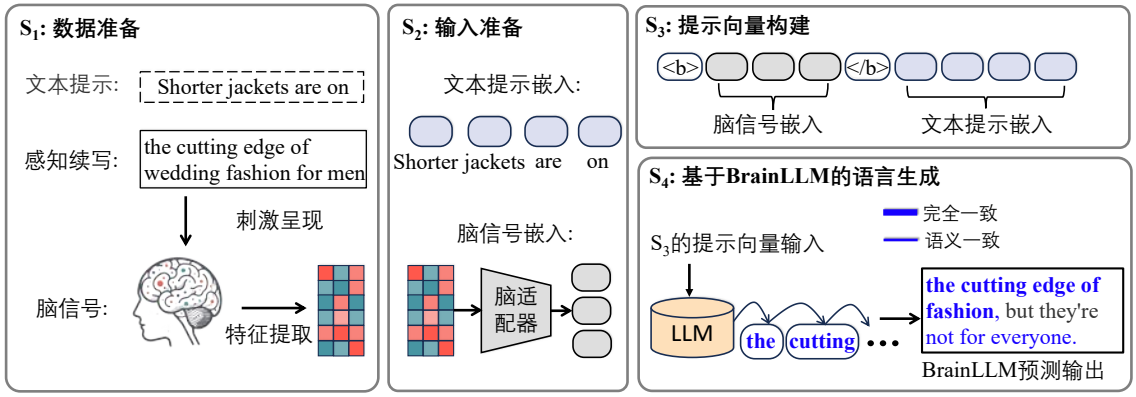
4.3　基于生成式模型的大脑语义解码

4.3.1　问题定义

给定一个由一系列词元 $\{w_1, w_2, w_3, \cdots, w_n\}$ 组成的文本提示 W，其续写内容是 $M = \{m_1, m_2, \cdots, m_k\}$。大脑语义解码的任务目标是通过参与者在感知到续写内容 M 时的脑信号和文本提示 W 来预测其续写 M。本节中将 M 称为“感知续写”。定义脑部记录 $X = \{x_1, \cdots, x_t\} \in \mathbb{R}^{t \times c}$ 是从血氧水平依赖（Blood Oxygen Level Dependent Signal，BOLD，一种利用 fMRI 设备检测的大脑活动）信号中提取的一系列特征，其中 c 表示神经特征的数量，t 表示收集脑信号的时间帧数。本研究采集感知续写的刺激内容呈现后的 t 个时间帧的脑信号作为输入，在每个时间帧内，BOLD 信号会被采集一次。这样的做法考虑了 BOLD 信号的延迟效应[162]（t 设置为 4，与现有工作一致[53,184]）。大脑语义解码的任务旨在学习一个自回归函数 F，该函数可以利用文本提示 W 和脑部记录 X 作为输入，一次生成一个感知续写 M 的词元。这个过程可以形式化为 $\hat{m}_i = F(\{w_1, \cdots, w_n, \hat{m}_1, \cdots, \hat{m}_{i-1}\}, X; \Theta)$，其中 $\hat{m}_i$ 是模型生成的第 i 个词元，Θ 是模型参数。

4.3.2　模型方法

BrainLLM 包括图 4.3(a)所示的四个关键步骤：（1）收集大脑数据并提取特征（参见第 4.3.3.1节）；（2）基于大脑适配器从大脑记录中学习嵌入；（3）从大脑和文本两个模态构建提示；（4）基于提示向量和预训练语言模型以自回归的方式生成语言。大脑适配器学习将大脑表征的空间映射到与预训练语言模型中的文本嵌入



(a) 基于脑信号的语言生成流程（BrainLLM）。生成过程包括四个主要阶段：S_1：采集用户在感知续写内容时引发的脑信号。S_2：通过脑信号适配器提取脑信号特征，并将其转换为与标准语言模型中的文本嵌入形状匹配的隐向量。S_3：将脑信号嵌入与文本提示嵌入拼接为提示输入。S_4：将提示输入传入语言模型以实现语言生成。BrainLLM 生成的内容可以完全匹配（例如“the cutting edge of”）或语义相似/核心内容匹配（例如“not for everyone”）于感知续写。

文本提示	感知续写	BrainLLM预测输出	对照模型预测输出
Shorter jackets are on	the cutting edge of wedding fashion for men	**the cutting edge of fashion**, but they're not for everyone	their way out **of** style, but they're still popular.
A wall is a	solid structure that defines and sometimes protects an area	**structure that defines and sometimes protects an area**	vertical **structure** made of stone, brick or concrete
I'm just stand- -ing there like	the proverbial deer in headlights	a **deer in the headlights**	an idiot
she was like petite I could have	folded her up and put her my pocket	picked **her up** with one hand	driven her to work every day

(b) BrainLLM 及其对照模型（PerBrainLLM）的语言生成示例。蓝色文本和**加粗蓝色**文本分别表示生成内容与真实值（感知续写）被人工标注为语义相似或完全匹配。

图 4.3　BrainLLM 的方法流程和生成示例

相同维度的空间，实现融合大脑模态和文本模态的提示表示的生成。同时本研究采用了一种称为“提示微调”[185]的方法和基于下一个词元预测的损失函数训练脑信号适配器。该方法在训练期间冻结预训练语言模型中的参数，仅更新脑信号适配器。为此，解码器的模型参数可以通过相对于通常用于训练大语言模型的数据量小得多的有限神经数据进行充分训练。以下我们先对工作中使用的预训练语言模型进行介绍，然后分别对步骤（2）的脑信号适配器、步骤（3）中涉及的提示向量构建和步骤（4）涉及的训练方法进行详细介绍。

4.3.2.1　预训练语言模型

本章的研究使用了在 Huggingface 上发布的预训练语言模型（https://huggingface.co/models），即 Llama-2（https://huggingface.co/meta-llama/Llama-2-7b）和 GPT-2 系列（https://huggingface.co/gpt2）。这些语言模型的功能类似。通常，它们首先通过嵌入层将输入词元转换为一系列隐向量。然后，这些向量被输

表 4.1　本实验中使用的大语言模型的统计数据

模型	参数数量	Transformer 层数	嵌入大小	词汇表大小	量化	最大输入词元
Llama-2	70 亿	32	4,096	32,000	float16	4,096
GPT-2-xl	15 亿	48	1,600	50,257	float32	1,024
GPT-2-large	7.74 亿	36	1,280	50,257	float32	1,024
GPT-2-medium	3.45 亿	24	1,024	50,257	float32	1,024
GPT-2	1.17 亿	12	768	50,257	float32	1,024

入到一个多层神经网络中，该网络使用多头自注意力机制来聚合序列中每个向量的表示[87]。基于这种架构，对于任意长度为 n 的输入词元序列 $S = \{s_1, s_2, \cdots, s_n\}$，语言模型可以估计下一个词元 s_{n+1} 在给定序列 S 上的先验概率分布 $P(s_{n+1} \mid S)$。传统上，输入词元 S 是基于文本的。而在本章的方法中，脑信号被纳入序列 S 的构建中，从而实现对脑输入感知的语言生成。

4.3.2.2　脑信号适配器

脑信号适配器是一个深度神经网络 f_x，输入为脑信号特征 $X = \{x_1, \cdots, x_t\} \in \mathbb{R}^{t \times c}$，输出为脑信号嵌入 $V^X = \{v_1^X, \cdots, v_t^X\} \in \mathbb{R}^{t \times d}$，其中 d 等同于大语言模型的嵌入维度。与已有的其他跨模态连接的模型不同[186-189]，BrainLLM 中的脑信号适配器通过非线性方式建模脑信号表示，考虑了 BOLD 信号的延迟效应，并采用了位置嵌入进行序列建模。

具体来说，f_x 包括：（1）一个位置嵌入 $P = \{p_1, \cdots, p_t\} \in \mathbb{R}^{t \times c}$，用于表示 BOLD 信号采集过程中的时间顺序编码；（2）一个多层感知机网络 f_m，用于将脑信号表示转换为与文本模态共享的隐空间向量。位置嵌入通过均匀分布初始化，并被设置为可训练的向量。每个位置嵌入 $p_i \in P$ 通过逐元素加法与其对应的脑信号特征 $x_i \in X$ 相加。多层感知机网络 f_m 由一个输入层、两个与输入的脑信号特征维度 c 相同的隐藏层，以及一个维度为 d 的输出层构成。ReLU[190]被用作激活函数。

形式化而言，第 i 个时间帧的脑信号特征 x_i 被输入到脑信号适配器 f_x，即 $v_i^X = f_x(x_i) = f_m(p_i + x_i)$。输出向量嵌入 v_i^X 的维度与大语言模型的嵌入维度一致，可进一步用于与文本模态共同构建输入。

4.3.2.3　提示向量构建

在 BrainLLM 中，文本提示直接输入到大语言模型的嵌入层 f_w，将词元转换为隐向量 $V^W = \{v_1^W, \cdots, v_n^W\} \in \mathbb{R}^{n \times d}$，其中 n 为词元数量，d 为嵌入维度。不同大语言模型对应的 d 值可参见表 4.1，所有统计数据基于原始论文[167,169]和开源平台（https://huggingface.co/meta-llama/Llama-2-7b和https://huggingface.co/gpt2）的报告。同时，脑信号适配器 f_x 将脑信号嵌入到与文本相同的潜在空间，维度为 d。具体来说，对于每个 $x_i \in X$，脑信号适配器将其嵌入到空间 $\mathbb{R}^d$，形式化为 $v_i^X = f_x(x_i)$。提示向量构建的目标是将脑信号嵌入 V^X 与文本嵌入 V^W 拼接在一起，使得语言模型能够在统一的表示空间中同时感知来自脑信号和文本的模态信息。

为了有效地区分这两种模态，本章引入了两个特殊词元，即 $\langle brain \rangle$ 和 $\langle /brain \rangle$，分别用来表示基于脑信号的嵌入的开始和结束。这些特殊词元被随机初始化为一维向量 $v^{\langle brain \rangle}$ 和 $v^{\langle /brain \rangle}$，其维度 d 与预训练语言模型中的词元嵌入维度一致。因此，输入序列 I 可以形式化为：$I = \{v^{\langle brain \rangle}, v_1^X, \cdots, v_t^X, v^{\langle /brain \rangle}, v_1^W, \cdots, v_n^W\}$。

4.3.2.4　训练方法

本节介绍了 BrainLLM 的两阶段的训练过程。第一阶段是无监督训练阶段，用于预热脑信号适配器，使其能够将基于脑信号的输入映射到语言模型词元嵌入所处的隐空间。第二阶段是监督学习阶段，用于指导模型从脑信号中解码语义信息以实现语义解码。

预热脑信号适配器的无监督训练　无监督的预热阶段的目标是使脑信号嵌入的分布在整体上与文本词元的嵌入分布对齐，确保脑信号嵌入适合作为语言模型的输入。训练对以无监督方式构建，每对由一系列脑信号及其关联的文本组成。形式化地，设 V^X 为映射后的脑信号，每个 $v_i^X \in V^X$ 被训练为接近相应查询嵌入的均值，即 $\frac{1}{n}\sum_{j=1}^{n} v_j^W$。本章采用均方误差损失函数进行训练，其公式为：

$$L_{\text{MSE}} = \frac{1}{t}\sum_{i=1}^{t}\left(v_i^X - \frac{1}{n}\sum_{j=1}^{n} v_j^Q\right)^2 \tag{4.2}$$

Liu et al.[191]认为，具有多模态输入的语言模型的适配器训练有必要加入预热步骤。本研究设计了基于脑信号的语义解码的预热方法，并通过实验结果证实了加入预热步骤的有效性。实验中观察到，如果省去适配器的预热阶段，可能导致训练不稳定（如梯度爆炸）以及性能差于包含预热阶段的模型。

基于下一个词元预测的监督学习训练　基于脑信号和文本模态的输入序列 I，本章利用语言模型的 Transformer 架构进行自回归生成。主要的训练目标是最大化感

知续写的生成概率：

$$\max_{\Theta} \sum_{i=1,2,\cdots,k} \log(P(m_i \mid I, \{m_1, \cdots, m_{i-1}\}; \Theta)) \tag{4.3}$$

其中 $\Theta = \{\Theta^{\mathrm{LLM}}, \Theta^{f_x}, \Theta^{sp}\}$ 为模型参数，Θ^{LLM}、Θ^{f_x} 和 Θ^{sp} 分别为语言模型参数、脑信号适配器参数以及特殊词元 $\langle brain \rangle$ 和 $\langle /brain \rangle$ 的参数。本研究通过“提示调优”（prompt tuning）技术[185]，在保留语言模型原始参数不变的同时，从有限数量的数据样本中学习有用的信息用于微调涉及输入表示构建的模型参数，即 Θ^{f_x} 和 Θ^{sp}。通过这种方式，脑信号适配器能够从人类脑信号记录中解码信息，以指导语言模型生成与感知续写在语义上相关的输出。

4.3.3　实验设置

4.3.3.1　实验数据集

本章在三个公开的 fMRI 数据集上测试 BrainLLM，分别是 Pereira’s 数据集[164]、Huth’s 数据集[192]和 Narratives 数据集[193]。所有数据集及其相关研究均获得伦理委员会的批准，并可用于基础研究。每位参与者均签署了知情同意书。Pereira’s 数据集收集了参与者在观看由维基百科风格句子组成的视觉刺激时的 BOLD 信号。与以往研究一致[194]，本文选择了参与实验 2 和实验 3 的参与者的脑数据。这涉及 5 名参与者，每位参与者在实验中观看了 627 个句子。本章使用了每个句子对应的发布的脑成像的 beta 系数（参见原始论文[164]）。Huth’s 数据集和 Narratives 数据集包含参与者在收听叙述性故事的听觉语言刺激时记录的原始 BOLD 响应。本章采用了这些数据集的官方发布的预处理版本（https://openneuro.org/datasets/ds003020/和https://openneuro.org/datasets/ds002345/）。Huth’s 数据集包括 8 名参与者的数据，每人收听了 27 个故事。因此，每位参与者贡献了 6 小时的神经数据，共计 9,244 个时间帧。Narratives 数据集最初包括 365 名参与者，本章选择了 28 名至少参与三次实验的个体。其中，8 名参与者参与了 4 个实验，20 名参与者参与了 3 个实验，每位参与者平均收集到 1,733 个时间帧的数据。

为了处理高维的 fMRI 数据，本章应用主成分分析[195]对所有数据集的全脑 BOLD 特征一致地降维至 $c = 1,000$。当对单一脑区进行分析时，原始信号将直接使用而不进行降维。基于此，本章根据每个时间帧的 BOLD 特征、呈现给参与者的对应刺激（感知续写），以及刺激之前的文本提示（如果有）构建了语言生成任务的数据样本。Pereira’s 数据集由参与者对于每个独立的句子的脑信号组成，不同句子呈现时没有重叠。本章将每个句子分为三个部分，每部分的词元数量大致相

等。通过将前三分之一作为文本提示，后三分之一作为感知续写，以及组合前两部分作为文本提示、最后一部分作为感知续写，本章基于每个句子构建了两个数据样本。对于 Huth's 数据集和 Narratives 数据集，语言刺激连续呈现给参与者。因此，本章基于 fMRI 扫描的重复时间的设置（Huth's 数据集中为 2 秒，Narratives 数据集中为 1.5 秒）视为一个时间帧来分割数据集。每个时间帧内的感知内容被选为感知续写。然后，本章使用一个长度为 1 到 3 个时间帧的滑动窗口选择感知内容出现之前的语言刺激作为文本提示。这一步为每个时间帧创建了 3 个数据样本，从而在构造尽可能多的样本的同时确保模型能够处理不同长度的文本提示。之后，数据样本按 3:1:1 的大致比例分成训练、验证和测试集。数据划分过程中确保了在训练、测试和验证集之间没有感知续写和脑记录的重叠。

4.3.3.2　语言生成的训练和推理设置

本章使用 Adam 优化器[125]训练 BrainLLM，学习率为 1×10^{-4}，批量大小为 8。由于预训练语言模型对显存的要求较高，本研究设置批量大小为 8。预热脑信号适配器的无监督训练的训练在 10 个 epoch 后停止。整个训练过程在 16 个具有 40 GB 内存的 A100 图形处理单元上进行，耗时约 14 小时完成。在测试集的推理中，本研究采用了束搜索方法。该方法维护了一个包含五个最可能序列的集合，并在每个生成步骤为每个序列生成续写。

4.3.4　实验结果

本章进行了三项评估来验证 BrainLLM 的性能：首先，我们将 BrainLLM 与一系列现有的开放词汇语言解码方法进行比较（见表 4.2）。其次，我们将 BrainLLM 的语言重构性能与对照模型 PerBrainLLM 进行比较。PerBrainLLM 通过随机置换脑信号作为不同预测任务的输入，使得文本刺激内容与脑信号之间没有对应关系。PerBrainLLM 相比无任何脑输入的标准语言模型（StdLLM）是一个更合理的对照。这是由于不同数据集有不同的语言习惯和解码分布，例如 Pereira's 数据集是维基百科风格的文本，Huth's 数据集是口语化的叙述性故事文本。因此即使文本刺激内容与脑信号之间没有对应关系，由于 PerBrainLLM 会倾向于解码语言风格相似的内容，其在语言相似性指标上相比 StdLLM 仍然具有显著的提升（见表 4.3）。最后，为了验证所提出的框架，我们还将 BrainLLM 与无任何脑输入的标准语言模型（StdLLM）以及其在不同架构选择下的变体进行比较。BrainLLM 的性能从三个方面进行评估：（1）胜率：BrainLLM 生成感知续写的概率是否高于对照模型 PerBrainLLM 或无任何脑输入的标准语言模型 StdLLM；（2）语言相似性指标：包括双语评估替补（Bilingual Evaluation Understudy，BLEU）、面向召回的评估替

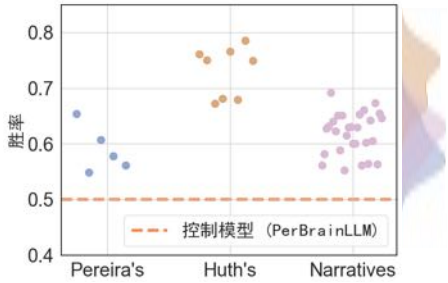
表 4.2　使用大语言模型进行脑解码的相关研究在不同设置下的性能

模型	验证[a]	数据集	性能（%）
Wang 等人[196]	TF	ZuCo	BLEU-1=40.1, ROUGE-1=30.1
UniCoRN[197]	TF	ZuCo	BLEU-1=35.9, ROUGE-1=39.1
	TF	Narratives	BLEU-1=62.9, ROUGE-1=59.5
DeWave[186]	TF	ZuCo	BLEU-1=41.4, ROUGE-1=30.7
NeuSpeech[198]	TF	Schoffelen	BLEU-1=78.1, ROUGE-1=83.7
	FTC	Schoffelen	BLEU-1=67.8, ROUGE-1=81.2
MAPGuide[199]	FTC	Huth’s	$BLEU_{zs}$=10.91, $METEOR_{zs}$=11.92
Tang 等人[53]	FTC	Huth’s	BLEU-1=24.09, WER=92.02, METEOR=16.67
PREDFT[200]	FTC	Huth’s	BLEU-1=25.98, ROUGE-1=19.61
BP-GPT[201]	FTC	Huth’s	BLEU-1=21.28, METEOR=20.31
BrainLLM（本方法）	FTC	Huth’s	BLEU-1=25.18, WER=92.02, METEOR=20.96
	TC	Pereira’s	BLEU-1=34.3, ROUGE-1=29.9, WER=75.8
	TC	Huth’s	BLEU-1=19.0, ROUGE-1=17.8, WER=89.2
	TC	Narratives	BLEU-1=13.8, ROUGE-1=13.0, WER=92.4

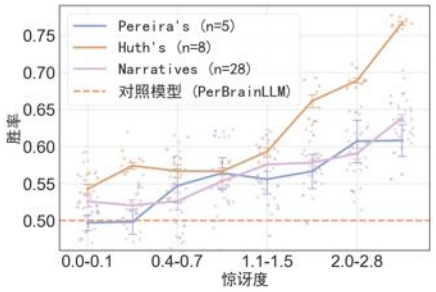
[a] FTC：全文构建，TF：教师强制，TC：文本续写。

补（Recall-Oriented Understudy for Gisting Evaluation，ROUGE）和词错误率（Word Error Rate，WER），用于计算感知续写与生成语言之间的相似性；（3）人类偏好：将 BrainLLM 和 PerBrainLLM 的输出展示给标注人员，判断哪个在语义上更接近感知续写。

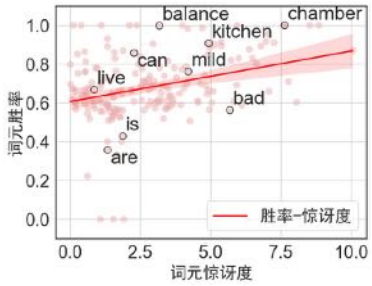
和大模型结合的大脑语义解码的相关研究根据评价方法可分为三类：（i）全文构建（Full-Text Completion，FTC）、（ii）教师强制（Teacher Forcing，TF）和（iii）文本续写（Text Completion，TC）。全文构建（FTC）意味着解码器需要仅根据脑信号构建完整文本，不基于任何一个提示词。教师强制（TF）意味着解码器可以通过使用实际输出作为下一个输入来预测下一个词元，而不是使用解码器自身预测的输出。Jo et al.[202]对 TF 的设置提出了质疑，因为提供实际输出作为输入会导致解码性能和真实环境差异较大。文本续写（TC）是本文提出的方法，解码器基于一部分语义上下文来构建文本续写。与 TF 不同，TC 和 FTC 不能利用实际输出作为输入，这使得结果的评价更为严格。TC 和 FTC 之间的一个区别在于，FTC 的性



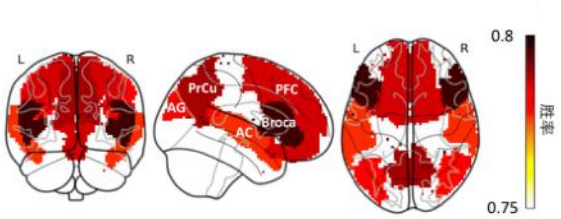
(a) 胜率显著高于 0.5



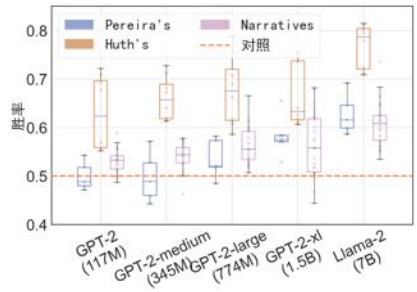
(b) 胜率和惊讶度正相关



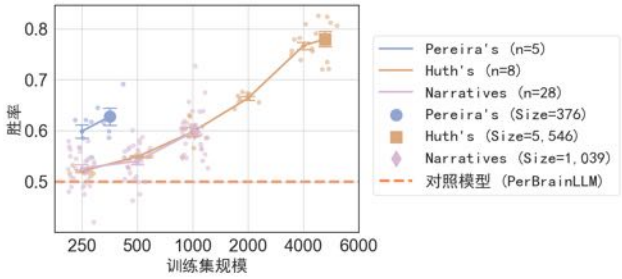
(c) 200 个随机选取的词元的胜率与惊讶度得分的散点图



(d) 在单个参与者中使用来自不同大脑皮层区域的脑信号时的胜率



(e) LLM 的参数规模与胜率正相关



(f) 胜率与训练数据的大小呈正相关

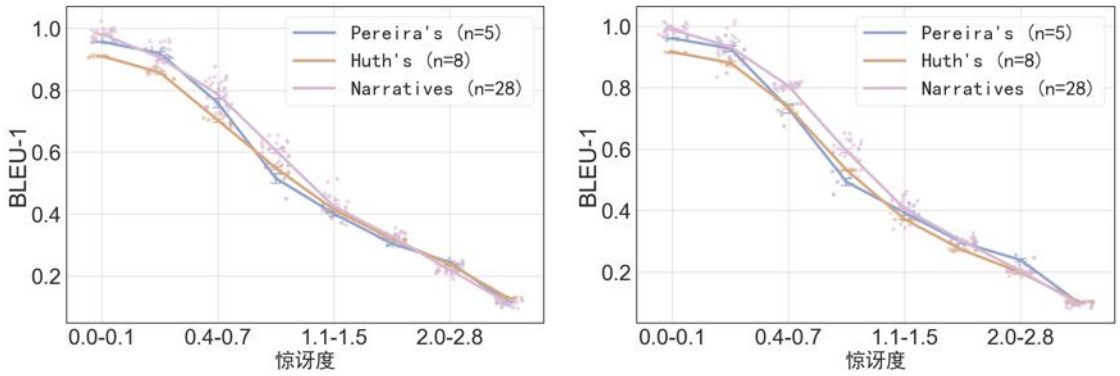
图 4.4　BrainLLM 对 PerBrainLLM 的胜率分析，胜率通过比较感知续写的生成概率进行计算。每个子图详细说明了不同方面的实验结果。

能和实际应用相比仍有显著差距。相比之下，TC 的性能离实际应用更接近，但要求解码器能够获取额外的用户相关上下文信息，以用作文本提示。从表 4.2中可以看到，在相同的 FTC 设置下，BrainLLM 相对于 Tang et al.[53]提出的的方法有显著提升，其基于语言模型预构建候选的下一个词元，然后基于脑信号从这些候选词元中进行选择。BrainLLM 在所有语言相似性指标上都优于他们的方法，BLEU-1 分数的改进达到 40.2%（更详细的比较参见表 4.4）。此外，BrainLLM 在性能上与同期的 BP-GPT 和 PREDFT 相当，尽管这两者也采用了类似的生成式方法，但它们在架构设计和训练方法上存在一些差异。在和对照模型 PerBrainLLM 的比较中，BrainLLM 的平均胜率在 Pereira’s 数据集、Huth’s 数据集和 Narratives 数据集上相比 PerBrainLLM 都具有显著更高的概率（见图 4.4(a)，错误发现率（False Discovery Rate，FDR）< 0.05（单侧非参数检验））。最高的平均胜率（73.1%）出现在 Huth’s

表 4.3　基于脑信号重建语言的任务在不同数据集下的语言相似度指标[a]

数据集	模型	BLEU-1(↑)	ROUGE-1(↑)	ROUGE-L(↑)	WER(↓)
Huth's	StdLLM	0.1500*	0.1360*	0.1310*	0.9200*
	PerBrainLLM	0.1668*	0.1536*	0.1474*	0.9109*
	BrainLLM	**0.1899**	**0.1780**	**0.1709**	**0.8916**
Pereira's	StdLLM	0.2415*	0.2133*	0.2096*	0.8349*
	PerBrainLLM	0.3269*	0.2815*	0.2751*	0.7783*
	BrainLLM	**0.3432**	**0.2987**	**0.2878**	**0.7576**
Narratives	StdLLM	0.0953*	0.0858*	0.0829*	0.9485*
	PerBrainLLM	0.1269*	0.1211*	0.1105*	0.9311*
	BrainLLM	**0.1375**	**0.1301**	**0.1209**	**0.9239**

[a] * 表示与 BrainLLM 之间的差异在 FDR < 0.05 的水平上显著（单侧非参数检验）。



(a) BrainLLM　　(b) PerBrainLLM

图 4.5　不同惊讶水平下续写内容的 BLEU-1 得分

数据集，该数据集中每个被试者的神经训练数据样本量最大（见图 4.4(f)）。这表明增加神经训练数据的规模可能会提高模型性能。在所有语言相似性指标上，本研究也观察到了类似的性能提升，如表 4.3所示。进一步，本章进行了人工评估实验，从 Amazon Mechanical Turk ①招募了 202 名标注员，他们被要求在 BrainLLM 和 PerBrainLLM 生成的输出之间进行偏好选择，或者在没有明确的偏好可以选择“难以区分”。在从 Huth's 数据集中随机选择的 3,000 对由 BrainLLM 和 PerBrainLLM 生成的文本对中，偏好分布为 48.4% 偏好 BrainLLM，39.2% 偏好 PerBrainLLM，12.4% 的文本对标注员而言难以区分。统计分析揭示了 BrainLLM 和 PerBrainLLM 之间的偏好存在显著差异（p=0.039，基于非参数检验）。

① www.mturk.com

不同惊讶水平下的语言生成性能　通过预测具有最高概率的下一个词元，大语言模型能够在给定文本提示的情况下生成结构良好且连贯的语言。该架构还通过估算预测误差信号，为建模文本续写中的惊讶度（即语言困惑度的对数，log(perplexity)）提供了统一的框架[203]。例如，“Nice to”后接“meet you”的可能性高于“take chances”，这意味着对于语言模型来说，“meet you”的惊讶水平低于“take chances”。通常，更高的惊讶水平表明语言模型认为生成感知续写的概率更低，因此也更具有挑战性。本章根据惊讶水平对测试数据进行划分，并分别在这些数据上评估 Brain-LLM 的性能。

如图 4.5(a)和图 4.5(b)所示，在 BLEU-1 的评估中，BrainLLM 和 PerBrainLLM 的性能均随着惊讶水平的提高而下降。对 BrainLLM，在 Pereira’s 数据集、Huth’s 数据集和 Narratives 数据集中，惊讶水平与 BLEU-1 分数之间的 Pearson 相关系数分别为-0.66、-0.52 和-0.56。这个观察表明，随着惊讶水平的增加，语言模型生成感知续写的难度增加。对 PerBrainLLM，在 Pereira’s 数据集、Huth’s 数据集和 Narratives 数据集中，惊讶水平与 BLEU-1 分数之间的 Pearson 相关系数分别为-0.67、-0.54 和-0.58。BrainLLM 的相关系数大于 PerBrainLLM，表明随着惊讶水平的增加，BrainLLM 的性能下降小于 PerBrainLLM。此外，本章分析了 BrainLLM 相较于 PerBrainLLM 在不同惊讶水平下的胜率，如图 4.4(b)所示。实验观察到，胜率随着惊讶水平的提高而增加。在 Pereira’s、Huth’s 和 Narratives 数据集中，惊讶水平与胜率之间存在显著的正相关性，皮尔逊相关系数分别为 $r = 0.09, 0.15, 0.08$（FDR < 0.05）。这表明，当大语言模型认为感知续写的内容更意外时，从脑信号中解码的信息可以显著提升生成过程。此外，具有较高惊讶水平和更高具体性的词元[204]与更高的胜率相关，其皮尔逊相关系数分别为 $r = 0.152$ 和 $r = 0.305$（见图 4.4(c)）。这表明 BrainLLM 在处理具有更明确含义的词元时表现更为有效。例如，具体名词如“chamber”和“kitchen”的胜率高于功能词如“is”和“are”。

文本提示的影响　通常，语言模型根据给定的文本提示生成语言。现有的自然语言处理研究[205]表明，随着文本提示长度的增加，生成内容的准确性会提高[205]。在大脑语义解码的场景下，本章进一步分析了文本提示长度对 BrainLLM 性能的影响，并观察到文本提示长度与胜率之间存在负相关性，在 Pereira’s、Huth’s 和 Narratives 数据集中，皮尔逊相关系数分别为 $r = -0.013, -0.059, -0.060$。这一现象可以部分解释为，较长的文本提示为大语言模型提供了更多的上下文信息，从而降低了感知续写的惊讶水平[173,206]，进而减少了脑输入信息的重要性。此外，Tikochinski et al.[207]提出，大语言模型可以处理大范围的上下文窗口，而大脑可能更倾向于关注最近感知到的内容。这种差异也可能影响脑信号解码的效果。

表 4.4　BrainLLM 与基于后验选择方法的语言重建模型（Tang et al.[53]）在 Huth's 数据集上的性能比较

设置	模型	BLEU-1(↑)	ROUGE-1(↑)	ROUGE-L(↑)	WER(↓)	胜率[a]
有文本提示	PerBrainLLM	0.1668*	0.1536*	0.1474*	0.9200*	0.5000*
	Tang et al.[53]	0.1675*	0.1537*	0.1483*	0.9197*	-
	BrainLLM	**0.1899**†	**0.1780**†	**0.1709**†	**0.8916**†	**0.7667**†
无文本提示	PerBrainLLM	0.0960*	0.0817*	0.0779*	0.9703*	0.5000*
	Tang et al.[53]	0.0967†,*	0.0818*	0.0788†,*	0.9700†,*	-
	BrainLLM	**0.1356**†	**0.1160**†	**0.1099**†	**0.9541**†	**0.8816**†

[a] 胜率表示相对于 PerBrainLLM 的胜率。†/∗ 表示在有无文本提示的设置下，使用 Wilcoxon 检验在 FDR < 0.5 水平上与 BrainLLM/PerBrainLLM 存在显著差异。由于基于后验选择的方法无法获得生成感知续写的概率，其胜率没有被汇报。由于 Tang et al.[53] 提出的方法是在无任何文本提示的设置下进行生成，而本章的方法是在有文本提示的设置下进行生成。因此，本章同时在两种设置下展示其性能比较。

此外，本章还研究了在没有任何文本提示的情况下从脑记录生成语言的性能。表 4.4展示了 BrainLLM 和 PerBrainLLM 在无文本提示情况下语言生成的性能。一方面，本章观察到 BrainLLM 在所有语言相似性指标上显著优于 PerBrainLLM。BrainLLM 相对于 PerBrainLLM 的胜率（Pereira's 数据集为 0.8885，Huth's 数据集为 0.8816，Narratives 数据集为 0.6728）甚至高于有文本提示时的胜率。在无文本提示时 BrainLLM 相对于 PerBrainLLM 的更高的胜率一定程度上和感知续写的高惊讶水平也是相关的。然而，无文本提示生成的语言相似性指标明显低于有文本提示生成的指标，导致其解码的内容质量离现实可用还有较大差距。这表明仅基于脑输入且没有任何文本提示生成语言仍然具有挑战性。

不同参数规模的语言模型的影响　本章的主要实验基于 Llama-2[169]（7B 参数）进行，为了进一步研究不同参数规模的语言模型的影响，本章测试了一系列其他具有不同参数规模的生成式语言模型，包括 GPT-2（117M 参数）、GPT-2-medium（345M 参数）、GPT-2-large（774M 参数）、以及 GPT-2-xl（1.5B 参数）。对于 PerBrainLLM 和 BrainLLM，语言相似性指标都随着语言模型参数数量的增加而显著提升（见表 4.5）。这一观察与现有认知一致：具有更多参数的大语言模型在语言生成方面表现更优[205,208]。有趣的是，尽管 PerBrainLLM 的性能随着参数数量的增加而提升，BrainLLM 相对于 PerBrainLLM 的胜率也随之增加（见图 4.4(e)）。这表明，随着参数数量的增加，大语言模型从脑信号输入中提取的信息也在发挥更大的作用。

表 4.5　Huth's 数据集中使用不同参数规模的语言模型的生成性能[a]

语言模型	模型	BLEU-1(↑)	ROUGE-1(↑)	ROUGE-L(↑)	WER(↓)
Llama-2 (7B)	StdLLM	0.1500*	0.1360*	0.1310*	0.9200*
	PerBrainLLM	0.1668	0.1536	0.1474	0.9109
	BrainLLM	**0.1899**	**0.1780**	**0.1709**	**0.8916**
GPT-2-xl (1.5B)	PerBrainLLM	0.1708*	0.1652*	0.1581*	0.9090*
	BrainLLM	0.1791	0.1729	0.1656	0.9022
GPT-2-large (774M)	PerBrainLLM	0.1657*	0.1584*	0.1516*	0.9132*
	BrainLLM	0.1762	0.1693	0.1616	0.9049
GPT-2-medium (345M)	PerBrainLLM	0.1640*	0.1549*	0.1489*	0.9140*
	BrainLLM	0.1667	0.1578	0.1514	0.9126
GPT-2 (117M)	PerBrainLLM	0.1088*	0.1059*	0.0997*	0.9516*
	BrainLLM	0.1096	0.1065	0.1011	0.9520

[a] * 表示在使用相同模型和相同数据集时，与 BrainLLM 有显著差异（Wilcoxon 检验中 FDR < 0.5）。

神经活动数据规模对训练的影响　本章在不同规模的神经活动数据上测试了 BrainLLM，并计算了其相对于 PerBrainLLM 的胜率。如图 4.4(f)所示，在 Huth's 数据集和 Narratives 数据集中，随着模型使用更多数据进行训练，语言生成性能稳步提升。现有研究[171,174]发现，扩大神经活动数据集的规模可以改善大脑中提取的语义表征与大语言模型的嵌入表征之间的映射。本章的实验结果进一步表明，在联合建模脑信号表示作为语言模型的输入时，扩大神经活动训练数据的规模也可以提高语言生成性能。

不同脑区输入的语言生成性能　本章探讨了以来源于不同脑区的脑记录作为输入生成语言的情况。图 4.4(d)展示了 BrainLLM 相对于 PerBrainLLM 的胜率，脑区包括布洛卡（Broca）区[209]、楔前叶（Precuneus，PrCu）[210]、前额叶皮层（Prefrontal Cortex，PFC）[211]、听觉皮层（Auditory Cortex，AC）[212]和角回（Angular Gyrus，AG）[213-214]，数据来源于 Huth's 数据集中随机选择的一名参与者（被试者 1）。本章观察到，BrainLLM 在所有语言处理相关脑区中的表现均显著优于 PerBrainLLM，其中在 Broca 区的胜率最高，为 0.8012。这一性能甚至超越了使用所有脑区信号时的结果，其原因可能是这些语言相关脑区的信噪比较高，并且在使用所有脑区时

使用的降维方法可能会导致信息损失。尽管如此，为避免在选择 ROI 时有偏，我们汇报的主要结果仍基于使用所有脑区信号构建的模型。

现有研究表明，在语言处理过程中，大脑中相当多的一部分皮层会被激活[215-216]。这表明与语言相关的不同脑区可能编码了重叠或相似的语言表示[217]，从而可以仅使用单一脑区生成语言。这一发现曾在基于分类而非生成的脑语言解码研究中被汇报[53,218]。

4.3.5　方法讨论

本章的研究表明，可以直接通过脑信号作为输入来生成开放词汇的语言内容，而不需要从预构建的语言候选中进行选择。与标准语言模型根据其训练数据仅生成最可能的语言续写不同，BrainLLM 的生成输出更符合人类参与者感知的语义文本内容。并且，基于提示微调的训练策略[185,219]，BrainLLM 大约只有 600 万个可训练参数，远小于 Llama-2 的 70 亿个参数。这个参数规模与现有的常用于脑解码的模型（如岭回归）在同一参数级别（例如，Tang 等[53]；Pereira's 等[164]），但实现了直接的语言生成，而不限制生成过程在预定义的候选语义对象的选择上。以下，本章对该方法的机理，与先前方法的比较，以及该方法的影响和未来扩展进行讨论。

4.3.5.1　如何将脑信号表征集成到计算语言生成模型中？

先前的研究表明，语言模型和人脑中的语义表示可以互相映射[171,184,220-223]。这些研究的关键发现包括探索训练语言模型时如何增强这种映射[224]，以及脑表示是否可以用于改进语言模型中的表示[171,173]。本章的方法与上述不同，因为脑信号和大语言模型中的语言表示之间的对齐不一定意味着二者可以在统一的计算框架中使用。BrainLLM 证明了使用从大脑解码的表示来丰富大语言模型输入的上下文信息的可行性。这种方法使得语言模型能够生成与人类参与者感知的语义匹配的连贯语言内容[219]。

BrainLLM 的成功可以归因于两个关键因素。首先，人脑编码的信息通常涵盖情境语义[164,173]。关于脑表示和语言模型表示之间映射的证据表明，情境语义可以被机器模型学习，从而实现有效的端到端的下一个词元生成训练。其次，语言模型综合能力的提升（尤其是参数量的提升）带来了“少样本学习”或“上下文学习”中的能力[225]。BrainLLM 利用这一能力，通过一个比语言模型预训练所需的数据规模小的神经数据集来训练脑信号的上下文表示。本章的实验还表明，随着模型参数规模的增加，BrainLLM 在性能提升方面比 PerBrainLLM 更大。

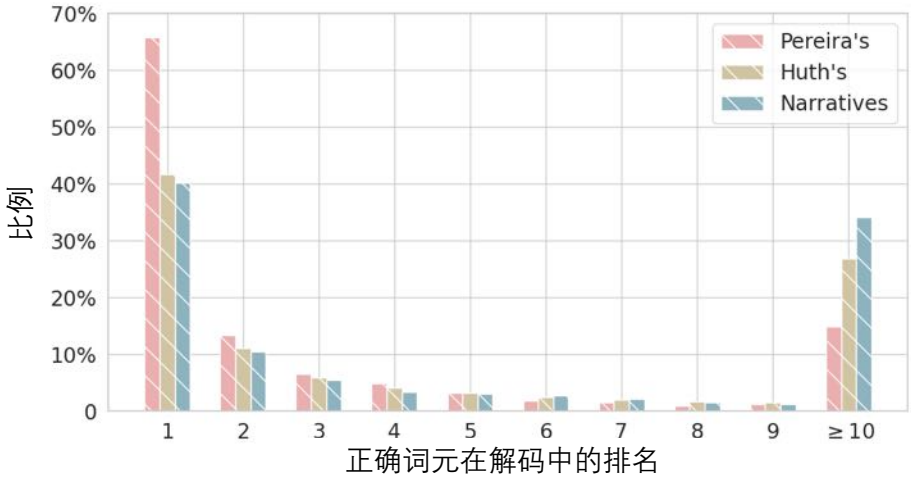


图 4.6　在三个数据集中使用 BrainLLM 进行语言生成过程中正确词元的排名

4.3.5.2　与先前工作的比较

大多数现有研究将语言重建任务视作一个分类任务，包括预定义一组语义候选（例如，词汇[162]，概念[164]，句子[226]），并使用映射函数来确定哪个候选与脑活动最匹配。预定义的步骤意味着这些方法无法构建超出预定义范围内的候选。一个例外是一项最近的研究[53]，该研究通过首先使用大语言模型预生成一些候选词元，然后从这些候选中选择与脑信号相匹配的词元来成功构建连续语义候选。

BrainLLM 和上述研究显著的不同之处在于直接使用了从大脑解码的表示作为生成式语言模型的输入。这种生成范式赋予其以下独特特性：首先，生成范式意味着语言重建可以通过逐个词元的构建实现，而不依赖于可能不正确的预先生成的候选。如果将每个词元的生成过程视作是一个分类任务，则 BrainLLM 在表现最佳的 Pereira’s 数据集上实现了高达 65.8% 的 top-1 准确率，在所有三个数据集上准确率均超过 40%（见图 4.6，图中较高的排名表示语言模型认为感知续写中的词元更有可能被生成，排名为 1 表示模型能够准确预测下一个词元。）。值得注意的是，这种准确率并不是在通常的有 2 至 50 个候选的分类任务中实现的，而是在有 32,000 个词元候选的生成设置中实现的。其次，BrainLLM 可以量化任何语义内容的生成概率，而不仅限于有限数量的语义候选。此功能可以更好地从语义空间的角度理解人类大脑中的语义分布，从而协助一些下游任务，例如在神经语言学分析中比较与不同语言特征相关的内容的生成概率，通过该语义分布来协助信息系统更好地理解人类用户等。最后，BrainLLM 在数据规模和参数规模方面的能力变化也表明，脑模态与生成式模型的结合具有良好的可扩展性。

近年来，生成式人工智能领域的许多研究激发并推动了脑解码研究的进展。生成式人工智能为从大脑解码信息提供了一条新途径，绕过了传统的分类设置。例如，除了本文探讨的语言重建之外，基于脑数据的视觉重建也从基于分类的模

型[227]发展到基于扩散的生成模型[228-230]。生成式人工智能的利用不仅限于从大脑解码信息，例如，它已被证明可以来解释人类视觉皮层的功能组织[231]。另一方面，一些研究探索了为什么脑信号有可能与这些生成模型联合建模。例如，Goldstein et al.[173], Lupyan et al.[232], Clark[233]的研究认为人脑有预测下一个词的倾向，这一现象得到了多项研究的支持。因此，我们相信生成式的方法是研究大脑中信息感知的一个有前景的方向。

4.3.5.3　影响与未来扩展

本章的研究展示了直接从脑信号中生成语言的可行性，并分析了其与先前方法的差异和优越性。由于生成范式的优势，BrainLLM 可以作为传统基于分类的方法的替代方案，尤其是在用户指令无法被限制在预定义候选集中的脑机接口应用中。例如，在搜索场景中，用户的信息需求是开放的而非受限于预定义的短语集合内的，因此 BrainLLM 可以帮助信息系统更好地在这种场景下解码用户信息需求。尽管 BrainLLM 表现优越，但开放词汇的解码仍然是极具挑战性的任务。本章观察到，在没有文本提示的情况下，BrainLLM 的输出性能离现实可用仍有差距（见表 4.4）。一个有前景的未来方向是将 BrainLLM 与外部模块集成，以推断一些文本提示来协助语言生成过程，例如结合其他类型的脑机接口和能够检测任务相关上下文的方法。例如，基于运动表示的脑机接口[179,234-235]已在现实场景中展示了较高的可用性，但这些接口需要用户进行充分的训练，并在输入过程中付出大量的努力[52,234]。相比之下，BrainLLM 在用户感知过程中就能够有效解码来自视觉和听觉刺激的语义内容。因此，结合两种类型的脑机接口可能会产生更有效的应用：基于运动的脑机接口生成初始文本提示，并通过 BrainLLM 的语言续写对用户大脑中的语义内容进行高质量、低成本的解码。

此外，BrainLLM 量化了在给定文本提示时，基于用户的脑信号生成感知续写的概率。因此，它可以用于研究人脑中编码的语义信息，而无需预定义有限的语言刺激集。例如，本章探讨了脑信号在不同惊讶水平、上下文长度和不同脑区带来的性能提升。这种方法还可以扩展现有的关于研究脑语言表征和感知的范式。例如，在神经语言学研究[137]中，研究人员通常通过操控和预定义具有各种语言特征的语言刺激来研究其对脑反应的影响。BrainLLM 允许我们在更自然的设置中收集脑数据，并通过比较具有不同语言特征的语义内容的生成概率来进行分析。例如可以用于探索不同人群对各种语言内容是否有不同的期望，以及哪些脑区与特定的语言内容更密切相关。此外，现有研究表明，人脑中的语义信息是上下文感知的[218]，例如，大脑对“快”的反应在理解“快车”和“快餐”这两个词汇时是不同的。由于本章的方法也是基于上下文（文本提示）的生成，因此可以用于探索上

下文信息对脑反应的影响。例如通过比较语言重建的性能来探索不同脑区与上下文语义特征之间的联系。

最后，有一些研究表明，计算语言建模可以从人类对语言的反应中获得一些相关见解，尤其是脑反应[184]。本章的实验表明，个性化的脑信号可能会优化语言生成过程，特别是在真实输出的生成概率对于没有引入对应脑信号输入的语言模型较低的情况下。这表明了可以基于脑反应的个体差异，训练更个性化的、更有效的语言模型。例如，BrainLLM 估计的生成概率可以促进语言模型的训练，以产生更符合人类用户期望的内容。基于人类反馈的语言模型训练已经成为语言模型后训练中重要的一环，该后训练的数据在传统上主要基于用户的行为信号反馈[236]。然而，行为信号仅提供了一维的偏好性反馈，BrainLLM 则能够在整个词汇分布中提供多维反馈，这能够为模型训练提供更丰富的信息。

4.4　大脑语义解码在查询增强场景的应用

基于第 4.3节介绍的大脑语义解码的定义和方法，本节进一步研究了从大脑中解码出来的语义信息如何被用于信息检索场景下的查询增强（也称为查询扩展）。本章将该方法称为基于脑信号的查询增强（Query Augmentation with Brain Signals，Brain-Aug），Brain-Aug 沿袭了 BrainLLM 的架构，但在推理策略和训练设置上有所不同。本节将依次介绍基于脑机接口的查询增强任务的任务设置，Brain-Aug 所采用的面向排序的推理策略，以及查询增强任务的实验结果和分析。

4.4.1　任务设置

任务流程　Brain-Aug 直接利用大脑信号增强用户提交的查询表示，其流程如图 4.7所示。Brain-Aug 的输入包括原始查询和与用户在查询语境下的认知反应相关的脑信号。本节采用和第 4.3节一致的方法训练一个脑信号适配器，将脑信号的表示与语言模型中文本嵌入的表示空间对齐，从而联合建模脑反应和原始查询创建一个统一的提示表示。基于该脑信号适配器和预训练大语言模型，Brain-Aug 可以生成原始查询的续写。与第 4.3节不同的是，在续写的过程中，我们使用了面向排序的推理方法来提升续写内容在不同文档间的可区分性，以提高排序性能（见第 4.4.2节）。以图 4.7为例，原始查询“Raspberry”（该例子来自于实验中的 Pereira’s 数据集）被扩展为“Raspberry is eaten fresh or cooked.”在这种情况下，和“蓝莓的食用”子主题相关的文档比那些关于“蓝莓的营养”或“蓝莓派”的文档排名更高。最后，本节根据扩展后的查询在文档排序中的性能来评估 Brain-Aug。

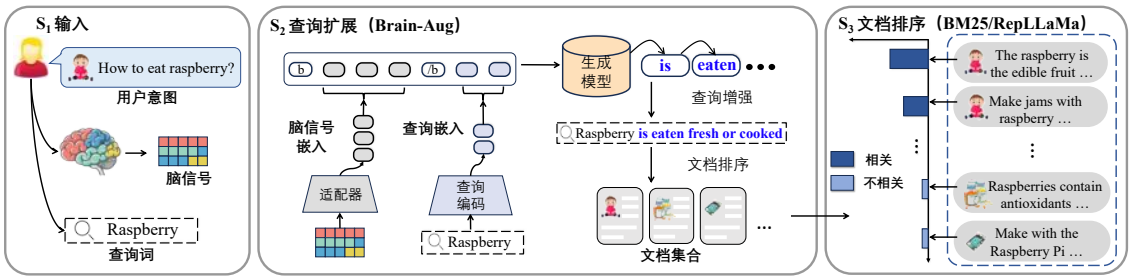


图 4.7　使用大脑信号进行查询扩展的过程（Brain-Aug）

数据集与任务构建　和第 4.3节一致，本节采用了三个公开可用的 fMRI 数据集，分别是 Pereira's 数据集[164]、Huth's 数据集[192]以及 Narratives 数据集[193]。不同的是，本节将这些数据集中的文本刺激处理成由文档语料库和查询集合组成的数据集，以构建检索任务。本节参考 Izacard et al.[237]和 Lee et al.[238]的方法，从现有的 fMRI 数据集中提取查询和文档。具体来说，本节在文档中选择一个文本片段作为伪查询，并将对应的文档视为该查询的相关文档。形式化地，对于文档 $D=\{w_1,\cdots,w_m\}$，我们提取一个片段 $Q=\{w_l,w_{l+1},\cdots,w_r\}$ 以构成一个相关的查询-文档对 $\{Q,D\backslash Q\}$，其中 $D\backslash Q=\{w_1,\cdots,w_{l-1},w_{r+1},\cdots,w_m\}$。

在 Pereira's 数据集中，每个文档由 3–4 个句子组成，这些句子逐一作为视觉刺激呈现给用户。由于单个句子的长度过长，不适合作为查询，我们将每个句子的前三分之一和三分之二截取下来，分别构造两个查询，从而为每个文档生成 6–8 个相关的查询-文档对。在 Huth's 数据集和 Narratives 数据集中，连续的内容以听觉刺激的形式呈现给用户。我们采用固定的 20 秒时间间隔（对应 10 次 fMRI 扫描）将刺激分割为文档，然后基于 1-3 个时间帧的时间间隔从文档中分割出不同长度的查询。由于不同参与者的脑数据存在差异，本节为每位参与者分别训练模型，并在参与者的每组数据上使用五折交叉验证评估 Brain-Aug。数据样本根据其所属的文档随机分为五折。在每次交叉验证中，将其中一折数据作为测试集，其余四折数据作为训练集和验证集，训练集、验证集和测试集的比例约为 3:1:1。

真实标签构建　为了利用第 4.3节中提出的方法训练脑信号适配器，我们需要构建续写内容的真实标签。在实际场景中，用户往往不擅长撰写清晰明确的查询，因此真实标签往往难以获取。为解决这一问题，一个常用的做法是认为用户意图的准确表示可以通过相关文档的内容反映出来。基于该假设，本节将这些相关文档中涉及查询词的语义片段设定为真实标签，并训练模型根据用户的脑信号和初始查询重建相关文档中的相关语义片段。具体而言，查询内容所对应的后续内容 $M=\{w_{r+1},\cdots,w_m\}\in D\backslash Q$ 被作为续写的真实标签，并采用与第 4.3节一致的方法进行脑信号适配器的训练。

训练与评估设置　接下来，本节将原始查询与其生成的续写结合作为增强的查询，并评估其在文档排序任务上的性能。本节选择两个排序模型来充分验证查询在文档排序任务上的性能，即稀疏排序模型 **BM25**[22]和密集排序模型 **RepLLaMA**[239]。本节采用文档排序指标来进行结果评价，包括不同截断位置的 NDCG[240]（NDCG@10，NDCG@20）、Recall@20 以及 MAP[116]。

4.4.2　面向排序的推理策略

在推理阶段，生成的续写内容应该在充分反映用户语义的同时，也应能够对不同的文档有比较好的区分效果。例如如果我们将 “Raspberry” 续写为 “Raspberry is popular”，则该续写内容可能也是符合用户语义的，但对于不同子主题的文档的区分效果较差，无法有效提升文档排序性能。因此，本研究在生成查询续写 $\hat{M} = \{\hat{m}_1, \cdots, \hat{m}_k\}$ 时，结合了词汇表中每个词元的逆文档频率（Inverse Document Frequency，IDF）信息[241]。设 IDF($\hat{m}$) 为词元 $\hat{m}$ 的 IDF，在推理阶段每个词元 $\hat{m}_i \in \hat{M}$ 的生成概率可以被计算为：

$$P_{\text{inf}}(\hat{m}_i) = \frac{P_{\text{LM}}(\hat{m}_i) + \alpha\,\text{IDF}(\hat{m}_i)}{\sum_{m \in \text{Vocab}}(P_{\text{LM}}(m) + \alpha\,\text{IDF}(m))} \tag{4.4}$$

其中 $P_{\text{LM}}(m) = P_{\text{LM}}(m \mid \{\hat{m}_1, \cdots, \hat{m}_{i-1}\}, S; \Theta)$ 表示在给定先前生成的词元 $\{\hat{m}_1, \cdots, \hat{m}_{i-1}\}$ 情况下，生成下一个词元 m 的概率，α 是一个超参数，Vocab 指代语言模型的词汇表。这种方法确保查询的续写不仅在语境上相关，而且生成内容可以在检索过程中有效地区分文档。

4.4.3　基线与对照

为了评估 Brain-Aug 是否有助于文档排序，我们将其文档排序性能与多个基线和对照进行比较。

对于基线，本节选择：（1）**原始查询**。（2）利用伪相关信号扩展的查询（记为 **Unsup-Aug**）。对于 BM25 排序模型，本节采用 RM3[129]作为 Unsup-Aug，通过从初始检索中排名靠前的文档中选择相关的词语内容来扩展查询。对于 RepLLaMA 排序模型，本节采用 Rocchio[181]作为 Unsup-Aug，通过调整查询向量的表示使其更接近排名靠前的文档的语义向量表示。（3）本节还报告了先使用 Brain-Aug 然后使用 Unsup-Aug 的额外结果，记为 **Unsup+Brain-Aug**。

对于对照模型，本节选择 Brain-Aug 的变体或消融版本。第一个对照是没有任何脑输入的 Brain-Aug（记为 **w/o Brain**），因此查询续写仅依赖于原始查询和语言模型自身的生成能力。第二个对照是使用随机采样脑输入的 Brain-Aug（记为 **RS Brain**）。RS Brain 通过从同一数据集中随机选择的脑输入进行采样。最后一个对

表 4.6　基于 BM25 的文档排序性能在所有被试者上的平均表现[a]

数据集	查询	N@10	N@20	R@20	MAP
Pereira's	原始查询	0.643*,†	0.664*,†	0.888*,†	0.594*,†
	Unsup-Aug	0.646*,†	0.655*,†	0.924*,†	0.590*,†
	Brain-Aug★	0.671	**0.691**	**0.941**	**0.618**
	Unsup+Brain-Aug★	**0.673**	0.686	0.936	0.615
Huth's	原始查询	0.297*,†	0.326*,†	0.536*,†	0.264*,†
	Unsup-Aug	0.291*,†	0.320*,†	0.575†	0.259*,†
	Brain-Aug★	0.306	0.340	0.569†	**0.273**
	Unsup+Brain-Aug★	**0.309**	**0.342**	**0.580**	0.269
Narratives	原始查询	0.419*,†	0.434*,†	0.629*,†	0.355*,†
	Unsup-Aug	0.440	0.452†	0.670†	0.367*,†
	Brain-Aug★	0.441	0.458	0.669	**0.382**
	Unsup+Brain-Aug★	**0.445**	**0.462**	**0.678**	**0.382**

[a] 我们的方法（*Brain-Aug* 和 *Unsup+Brain-Aug*）标记为 ★。最佳结果以**加粗**形式表示，次优结果用 下划线 标记。∗/† 分别表示 Brain-Aug / Brain+Unsup 显著优于该基线（$p < 0.05$，配对 T 检验）。

照是不使用面向排序的推理策略的 Brain-Aug，在其中每个词元的生成概率在计算时未考虑 IDF 权重（记为 **w/o IDF**）。

4.4.4　文档排序实验结果与分析

4.4.4.1　整体性能

表 4.6和表 4.7分别展示了基于 BM25 和 RepLLaMa 为排序模型时，原始查询、利用无监督信号扩展的查询（Unsup-Aug）和利用脑信号扩展的查询（Brain-Aug）的文档排序性能。本节观察到以下几点：

（1）无论是使用 BM25 还是 RepLLaMa 作为排序模型，Brain-Aug 的性能都显著优于原始查询和 Unsup-Aug。例如，根据 NDCG@20 的结果显示，Brain-Aug 在 Pereira's 数据集、Huth's 数据集和 Narratives 数据集上的表现分别将原始查询的得分提升了 0.027、0.014 和 0.024。唯一的例外是在 Pereira's 数据集上使用 RepLLaMa 作为排序模型和 MAP 作为评价指标时。这种例外可能是由于 RepLLaMa 在 Pereira's 数据集上基于原始查询就有较高的性能，我们将在观察（3）中讨论这一点。

表 4.7　基于 RepLLaMa 的文档排序性能在所有被试者上的平均表现[a]

数据集	查询	N@10	N@20	R@20	MAP
Pereira's	原始查询	0.878	0.881*,†	0.964*,†	0.858
	Unsup-Aug	0.872*,†	0.877*,†	0.951*,†	0.855
	Brain-Aug★	**0.883**	**0.887**	**0.980**	**0.859**
	Unsup+Brain-Aug★	0.878	0.882	0.975	0.853
Huth's	原始查询	0.299	0.328*,†	0.520*,†	0.275*,†
	Unsup-Aug	0.302*,†	0.333*,†	0.537*,†	0.276*,†
	Brain-Aug★	**0.310**	**0.342**	0.550	**0.281**
	Unsup+Brain-Aug★	0.308	0.340	**0.552**	0.279
Narratives	原始查询	0.413*,†	0.426*,†	0.611*,†	0.351*,†
	Unsup-Aug	0.416*,†	0.431*,†	0.629*,†	0.356*,†
	Brain-Aug★	0.430	**0.446**	0.641	**0.382**
	Unsup+Brain-Aug★	**0.432**	**0.446**	**0.642**	0.380

[a] 我们的方法（*Brain-Aug* 和 *Unsup+Brain-Aug*）标记为★。最佳结果以**加粗**形式表示，次优结果用下划线标记。*/† 分别表示 Brain-Aug/Brain+Unsup 显著优于该基线（$p < 0.05$，配对 T 检验）。

（2）在不同的数据集和指标下，Unsup-Aug 并未始终优于原始查询。使用 BM25 作为排序模型时，Unsup-Aug 和原始查询在 Recall@20 指标上的显著差异表明，Unsup-Aug 通过扩充查询词的语义表示空间，从而提高了召回率。当 Brain-Aug 与 Unsup-Aug 结合（Unsup+Brain-Aug）时，实验观察到其性能要优于 Unsup-Aug。这突显了脑信号在查询扩展中的有效性以及将其与其他信号结合的潜力。

（3）在 Huth's 数据集和 Narratives 数据集上，RepLLaMa 和 BM25 的性能差异不大。这表明在零训练样本、跨领域的设置中（数据集与 RepLLaMa 的训练数据不同），密集检索模型如 RepLLaMa 不一定优于传统稀疏检索模型如 BM25。这种现象也在 BEIR 数据集中观察到[242]。然而，在 Pereira's 数据集上，RepLLaMa 相比 BM25 有显著的性能提升。RepLLaMa 在 Pereira's 数据集上的出色表现可能是因为其文本内容（基于维基百科）很可能在 RepLLaMa 的预训练中被使用。

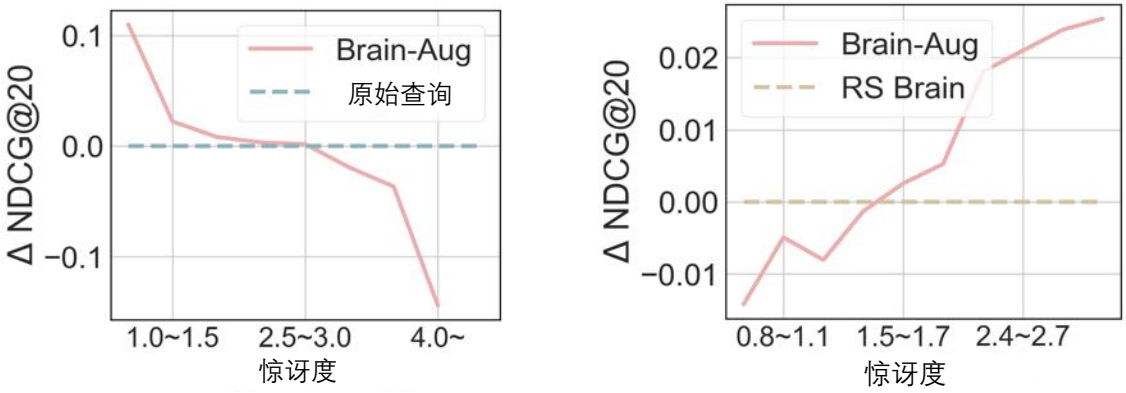
4.4.4.2　Brain-Aug 的消融实验

接下来，本章研究脑信号和面向排序的推理策略对 Brain-Aug 的贡献。实验结果如表 4.8所示。首先，本章观察到移除脑信号输入（w/o Brain）或随机采样脑信号

表 4.8　Brain-Aug 与其变体在 BM25 排序模型下的文档排序性能比较[a]

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

[a] 最佳结果以**加粗**显示。* 表示 *Brain-Aug* 与基线在配对 T 检验中有显著差异（$p < 0.05$）。



(a) Brain-Aug vs. 原始查询　　(b) Brain-Aug vs. Brain

图 4.8　在 Pereira's 数据集中，文档排序性能与真实查询续写惊讶度之间的关系。"RS Brain"表示将 Brain-Aug 的脑输入随机化的消融版本。*Δ* NDCG@20 表示 Brain-Aug 相较于 RS Brain 或原始查询的性能提升。

输入（RS Brain）的变体会导致性能下降。这表明从脑信号中解码的当前场景下的语义信息确实能够增强查询。其次，尽管 RS Brain 在生成真实内容的概率上优于 w/o Brain 方法（可参考第 4.3节中表 4.3里关于 PerBrainLLM 和 StdLLM 的对比），但在 Huth's 数据集和 Narratives 数据集上，它并没有实现更好的文档排序性能。这可能是因为 RS Brain 生成的内容与整个数据集的词元分布更匹配，从而整体地提高了生成真实内容的概率。但由于 RS Brain 缺乏与查询语境相关的语义，因此生成内容未能被有效地用于区分数据集中的不同文档。第三，本章还观察到，与没有使用面向排序的推理策略的消融版本（w/o IDF）相比，Brain-Aug 的性能显著提升。这表明生成可用于区分不同文档相关性的内容的重要性。

表 4.9　Pereira's 数据集中增强查询和原始查询基于 BM25 进行文档排序的示例[a]

方法	查询内容	排名最高的文档	相关性
原始查询	The shaking can	d_{21}: The wind from the hurricane shook the house, shattering a window ... Later that night, with the wind shaking the house, ...	0
Unsup-Aug	The shaking can from house wind	d_{21}: The wind from the hurricane shook the house, shattering a window ... Later that night, with the wind shaking the house ...	0
RS Brain	The shaking can last anywhere from a few seconds to several minutes	d_{21}: The wind from the hurricane shook the house, shattering a window in the kitchen. ... Later that night, with the wind shaking the house, we fell asleep huddled on the sofa.	0
Brain-Aug	The shaking can be caused by an earthquake	d_{13}: Earthquakes shake the ground and can knock down buildings and other structures. also trigger landslides and volcanic activity. Most earthquakes are caused by ...	1

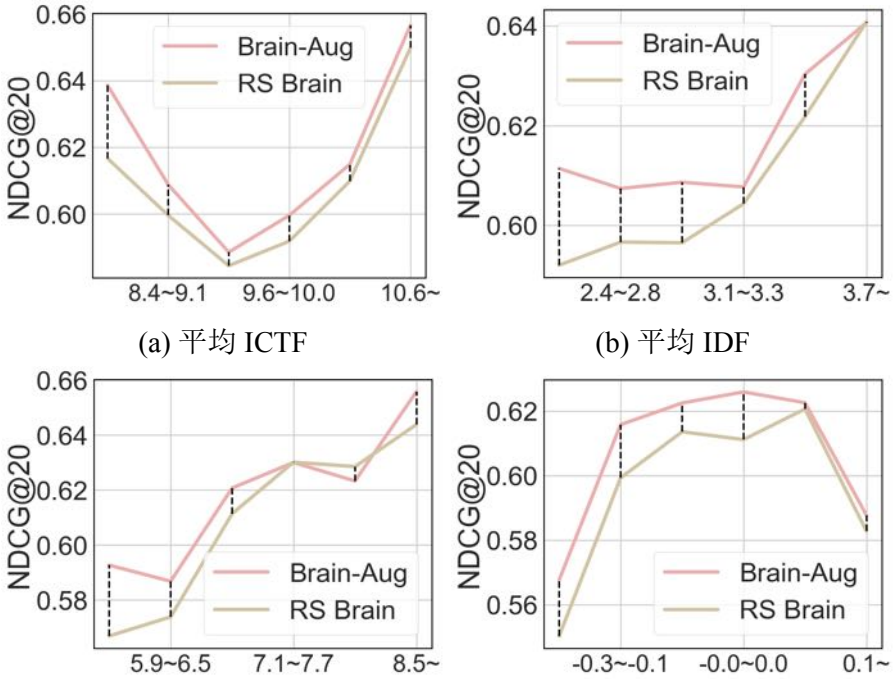
[a] 最佳结果以**加粗**显示。* 表示 *Brain-Aug* 与基线在配对 T 检验中有显著差异（$p < 0.05$）。

4.4.4.3　文档排序性能与查询生成困惑度之间的关系

图 4.8展示了 Brain-Aug 和 RS Brain 的文档排序性能与使用 RS Brain 测量的查询续写困惑度之间的关系。较低的困惑度表明生成更准确查询续写的可能性更高。如图 4.8(a)所示，较低的困惑度导致文档排序性能的提升，这是因为这种情况下续写的内容可能更为准确。而图 4.8(b)显示出了不同的趋势：当困惑度较高时，Brain-Aug 相较于其消融版本 RS Brain 的性能提升更大。这意味着在生成准确查询续写比较困难时，从脑信号中解码的语义更加有益。这一观察结果与第 4.3节中的发现一致，即在生成具有较高不确定性的续写时，加入脑信号会带来更显著的性能提升。

4.4.4.4　案例研究

表 4.9提供了一个基于 BrainLLM 进行查询增强的案例，原始查询是 "The shaking can"，取自 Pereira's 数据集中的文档 d_{13}。Brain-Aug 使用脑信号将查询扩展为 "be caused by an earthquake." 因此，与地震主题相关的文档 d_{13} 在搜索结果中被正确地排在头部。相比之下，当使用原始查询，利用无监督信号扩展查询，或利用随机采样的脑信号扩展查询时，与风暴主题相关的文档 d_{21} 被错误地排在头部。此案例研究展示了在查询扩展过程中结合脑信号中解码的语义信息带来的显著提升。



(a) 平均 ICTF　　(b) 平均 IDF

(c) 特异性　　(d) 清晰性

图 4.9　Pereira’s 数据集的文档排序性能与不同查询特征的关系

4.4.4.5　Brain-Aug 在不同特征查询上的性能分析

接下来，本章通过根据查询特征对查询进行分组，研究 Brain-Aug 在不同查询上所取得的性能提升。本章选择了四个查询特征：三个基于查询词元的预检索特征，即逆语料库词频（Inverse Corpus Term Frequency，*ICTF*）、*IDF* 和特异性分数[243]，以及一个基于检索文档信息的后检索特征，即清晰性分数[161,244]。一般而言，较大的 *ICTF*、*IDF*、特异性和清晰性分数对应于更明确的查询，通常会导致更好的检索质量。图 4.9展示了 Pereira’s 数据集的文档排序性能与不同查询特征的关系。基于图 4.9，本章有两个关键发现：

（1）当平均 IDF、特异性分数和清晰性分数增加时，Brain-Aug 和 RS Brain 的检索性能均有所提升。这表明更具体、明确的查询通常具有更好的检索性能。

（2）当这些特征的数值降低时，Brain-Aug 相对于 RS Brain 的性能提升更为显著。具体而言，Brain-Aug 相较于 RS Brain 在 NDCG@20 的提升与平均 ICTF、平均 IDF、特异性和清晰性分数之间皮尔森相关系数分别为 −0.14、−0.19、−0.17 和 −0.32，均为显著的负相关。这表明基于脑信号解码的语义信息在语义更模糊的查询中带来的性能提升更大。

4.5　本章小结

本章探索了如何从大脑信号中解码用户信息需求，并将其用于信息获取系统。现有的脑语言解码技术研究都将解码任务视为分类任务，限制了表达复杂多样的用户信息需求的能力。因此，本章提出了 BrainLLM 方法，将从大脑信号中解码的语义表征直接用于语言生成过程，在实现开放语境下的用户信息需求解码的同时，显著提高了在三个 fMRI 数据集上的解码性能。本章对 BrainLLM 进行了进一步的分析，发现当解码内容具有较高的惊讶度时，BrainLLM 的性能提升更为显著。此外，本章还发现了 BrainLLM 的能力随着语言模型的参数量增加而增加，这暗示了更强的语言模型在脑语言解码任务上更大的潜力。进一步，为应对信息系统中用户需求解码与系统反馈整合的挑战，本章将 BrainLLM 和查询扩展技术结合，设计了 Brain-Aug。Brain-Aug 在 BrainLLM 的基础上，引入了面向排序的推理策略，实现以大脑信号作为额外输入增强查询的表示能力，并进一步提高了搜索排序的性能。实验结果显示，Brain-Aug 能够显著提高查询的表征能力并优化搜索排序，其效果不但优于传统的查询扩展方法，也能与这些已有的传统方法结合。此外，本章发现 Brain-Aug 在处理模糊查询时表现尤为优异，这揭示了其在信息检索中的应用场景。本章的研究为基于大脑信号进行信息需求解码的系统提供了方法支撑和实验验证，为后续构建包括用户反馈建模等不同功能在内的基于脑机接口的交互式信息检索系统提供了基础。

本章相关成果发表于 SCI 一区期刊 Nature Communications Biology[17] 和 CCF A 类会议 Multimedia 2024 长文[18]。

第 5 章　基于脑机接口的用户反馈建模

5.1　本章引言

上一章详细阐述了借助脑机接口技术对用户的信息需求进行解码的方法设计，并通过基于查询增强的下游应用，充分验证了该技术在信息检索系统中所具备的有效性。本章则从用户反馈环路这一新的视角切入，探索如何通过脑机接口采集用户在信息交互过程中的生物信号反馈，构建更高效的智能交互系统。信息检索系统中的用户反馈环路构建是优化检索性能和提升用户体验的重要方法。例如，信息检索系统可以基于用户认为相关的文档更好地建模用户的兴趣和意图，从而重新排序信息内容。这种技术被称为相关性反馈技术[245-246]，并广泛地用于交互式信息系统的场景中[178,246-247]。

传统上，用户反馈信号可以分为两类：显式用户信号（如对搜索结果的主动标注，点赞行为等）和隐式用户信号（如点击、驻留时间等）。显式用户信号指的是用户需要对给定的文档集合主动提供明确的相关性判断。这给用户增加了额外的认知负担，因此在实际的信息系统中难以广泛应用。另一方面，以点击为代表的隐式信号是信息检索系统中长期以来进行搜索评估[248]和相关性建模[249-250]的重要用户反馈。然而，隐式信号是从用户行为中推断出的文档相关性的间接预测，因而常常存在有偏差和不准确的情况[98]。虽然现有研究尝试通过融合眼动追踪、鼠标移动等更多隐式信号来减轻单一信号的偏差，但实际经验表明基于这些间接信号推断的用户的真实相关性理解仍存在较大的误差[251]。例如，研究者发现眼动追踪信号对相关性解码的可信度有限，这是因为注视某个文档并不一定能保证其内容真正相关[252]。

脑机接口为突破这些局限提供了新的技术路径。例如，一些现有研究[69]发现相关和不相关的搜索结果引发的脑响应之间存在潜在差异。这些差异表明脑信号可以作为测量相关性的一种替代信号。与现有信号相比，脑信号直接反映用户的心理活动，因此不易受用户行为的固有偏差的影响。为了探索脑机接口在用户反馈建模任务上的效果，本章开展了两个用户研究和对应的信息环路构建的探索。这两个用户研究的整体范式如图 5.1所示，用户会首先阅读一个检索任务，然后可以通过互联网搜索来获取这些任务的答案。在这个过程中，本研究会采集用户的脑电信号。最后，本研究会采集用户对于信息内容的显式反馈和整体的信息检索体验。本章基于这两个用户研究，探索了用户脑电是否能够被用于相关性估计，以及基于用户反馈的相关性估计如何能被用于下游的相关性反馈任务，提升搜索结

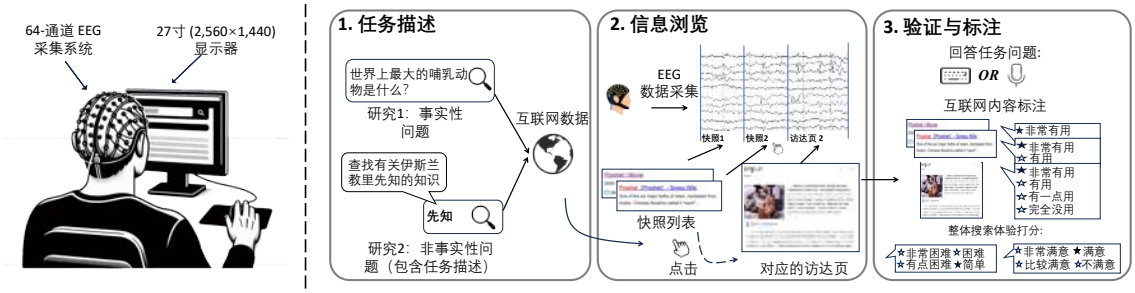


图 5.1　本章开展的两个用户研究的实验流程

果重排序性能。

第一个用户研究聚焦于事实性问题检索的场景，并探索了一个比较极端的情况，即“零点击”搜索[①]。“零点击”搜索指的是用户在搜索会话中未和信息内容进行点击交互。在传统上，点击会被视为相关性的正面信号，而无点击（尤其是查看结果后却没有点击）则被视为一种负面信号。然而，随着搜索引擎技术的进步，用户在没有发生点击的情况下就满意地完成搜索任务的情况变得越来越常见，尤其是对于一些事实性问题的检索。当前，搜索引擎返回的信息内容远不止传统的“十个蓝色链接”，而是会通过更丰富的内容摘要、内容整合等形式，让用户在付出更少的点击的情况下满足用户的信息需求。图 5.2展示了三个真实世界搜索结果的例子，其中有两个结果的点击必要性为 0，但其中第一个结果却能给用户提供有用的信息。报道显示，2020 年谷歌的“零点击”搜索会话在所有搜索会话中的占比已上升至近 65%[②]。因此，理解用户在没有发生点击行为的场景下，是否满意地完成搜索任务并获取有用信息成为一项重要挑战并受到广泛关注。为了应对该挑战，本章采集了一些“零点击”搜索场景下的用户脑电信号，分析了无点击结果的相关性和脑电信号的关联。一方面，本章揭示了与这些结果的相关性判断有关联性的脑区，包括侧颞叶、额叶及枕叶等。另一方面，本研究探索了如何用脑电信号更好地预测相关性，并设计了重排序策略，基于预测的搜索结果相关性来优化搜索结果排序（相当于第二个用户研究中的回顾式相关性反馈）。实验结果证明，在无点击场景下，脑电信号可作为用户反馈信号的替代。

第二个用户研究探索了非事实性问题检索的场景。与具有明确答案边界的事实性问题不同，非事实性问题的答案通常并不唯一，需要用户通过与多个搜索结果的交互来综合信息并形成最终结论。由于此类问题本身具有答案的开放性和检索目标的动态演化特性，其解决过程相比事实性问题表现出显著更高的复杂度。在该场景下，传统的点击信号和伪相关性信号等常常不足以为系统建模用户在完成检索任务过程中的反馈提供足够的支持。本章设计了一个将脑信号与伪相关性信

① https://sparktoro.com/blog/less-than-half-of-google-searches-now-result-in-a-click/
② https://www.perficient.com/insights/research-hub/zero-click-study

图 5.2　信息检索任务中文档的点击必要性和相关性/有用性并不一致

号和点击信号结合的相关性反馈框架，该框架可用于两种不同的相关性反馈设置，即交互式相关性反馈（Interactive Relevance Feedback，IRF）和回顾式相关性反馈（Retrospective Relevance Feedback，RRF）。交互式相关性反馈[253]在搜索结果逐个呈现给用户的同时动态地进行文档排序，基于用户反馈信号的增量信息不断调整检索策略[181]。而回顾式相关性反馈则被应用于在搜索过程结束后，对该查询相关的文档进行重新排序，这与第一个用户研究中涉及的搜索结果重排序任务类似。回顾式相关性反馈无法直接协助当前的搜索进程，但可以帮助具有潜在的类似搜索意图的搜索过程[254]。相比第一个用户研究，本章增加了交互式相关性反馈的研究，这是因为非事实性问题检索的场景下，用户在搜索过程中可能需要多次交互来完成搜索任务，因此需要一种能够动态调整检索策略的反馈机制。本章首先探讨了脑信号能否在复杂的非事实性问题场景下用于有效的相关性估计。接下来，在这两个相关性反馈设置下，本章还探索了所提出的框架是否能改善相关性反馈的性能，以及脑信号相比其他传统信号能发挥多大的作用。本章发现脑信号在点击信号存在偏差的搜索场景尤其有效。基于这些发现，本章进一步提出了一种在该相关性反馈框架下融合不同信号的算法，通过根据特定的搜索场景自适应地调整脑信号与其他信号的融合权重来提升相关性反馈的表现。实验验证了该算法的有效性，为基于脑机接口的信息系统环路中如何融合不同的用户信号提供了建设性的意见。

本章的结构如下：第 5.2节介绍相关工作，包括相关性反馈技术、零点击搜索

和搜索结果的有用性等。第 5.3节介绍了“零点击”搜索场景下的用户实验和建模方法。第 5.4节介绍了非事实性问题检索场景下的用户实验和建模方法。第 5.5节对本章进行小结。

5.2　相关工作

5.2.1　相关性反馈技术

相关性反馈技术主要有两大类：基于向量空间模型的方法（例如 Rocchio[33]）和基于统计语言模型的方法（例如 RM3[129]）。基于向量空间模型的方法通过将查询向量调整得更接近相关文档[33]，而基于统计语言模型的方法通过选择相关词语来扩展查询[129]。最近，随着神经检索方法的发展[255-256]，研究者们尝试了探索基于 BERT 的神经网络编码器在相关性反馈中的有效性[257-260]。例如，Wang et al.[258], Zheng et al.[261]采用基于向量空间模型的设计，将 BERT 编码的查询嵌入与文档嵌入融合。Zheng et al.[261]提出了 BERT-QE，将潜在的相关文档视为用户提交的额外查询，并使用基于 BERT 的排序器进行文档重排。此外，相关性反馈还存在不同的范式。例如，Aalbersberg[253]提出了交互式相关性反馈，该方法以增量的形式在每轮用户交互后对下一批文档重新排序。除了交互式相关性反馈，本章还研究了对历史文档的相关性估计，即回顾性相关性反馈。回顾性相关性反馈通过重新排序搜索结果或重新训练用于估计的相关性信号的排序模型等形式[254]，可以协助未来具有类似意图的潜在搜索过程。在典型的互联网检索之外的信息获取场景中，研究者也设计了一些场景驱动的相关性反馈技术，例如用于会话式搜索[181]和产品搜索[262]的方法。

在这些相关性反馈技术中，如何获取准确的相关性估计是其有效性的关键。标准的相关性反馈技术需要对固定批量的文档进行相关性评估。由于标准的相关性反馈过程给用户增加了额外的认知负担，先前的研究探索了多种替代信号来估计搜索结果的相关性。这些信号大致可以分为两类：伪相关性信号和隐式信号。伪相关性信号的基本思想是简单地将排名靠前的文档视为相关文档，并从这些文档中选择一些文本内容进行查询重写。该方法通常用于缓解查询术语不匹配的问题[129,257]。伪相关性信号不基于任何用户信号，使其易于部署并被广泛使用。然而，伪相关性信号的质量高度依赖于初始检索的有效性[183]，并且在提交的查询模糊或不明确时尤其不稳定。Li et al.[183]观察到，如果初始的检索性能较弱，伪相关性信号的表现往往会受到限制。除了伪相关性信号，隐式信号也被广泛研究，即通过用户行为（例如点击[246]、停留时间[263]和眼动追踪[251]）来推断用户的相关性偏好。在所有隐式信号中，点击信号最为广泛使用，并被认为是相关性的指示标

志[246]。然而，点击信号是相关性的一种间接表现[98]，并且在某些搜索场景中可能存在偏差[264-266]，即被用户点击的文档很多时候并不相关（如“点击诱饵”问题，具有误导性标题的文档能够吸引点击[5]），没有被点击的文档也不代表不相关。

传统上，相关性反馈仅基于一种信号类型（例如伪相关性信号[257-258,260]或隐式信号[267-268]），由于这些相关性反馈信号的准确性和固有偏差，其性能受到了限制[183]。为了解决此偏差问题，有一些研究设计了多种鲁棒结构[269-271]来处理有偏的信号。例如，Lv et al.[270]提出学习一个自适应系数，从而在相关性反馈信号不可靠时避免过度依赖相关性反馈信息。此外，还存在一系列关于点击模型的研究[26,272]，旨在从偏差点击信号中估计相关性。尽管如此，由于用户的交互行为极其多样且难以捕捉，这些方法的性能仍然有限[273]。因此，如何获取更准确的相关性反馈信号是相关性反馈技术进步的关键。

5.2.2　零点击搜索

“零点击”指的是在没有点击任何搜索结果的情况下结束搜索会话。但“零点击”并不意味着用户没有找到满足其信息需求的内容。近年来，商业搜索引擎一直尝试通过提取高质量的摘要或构建增强的搜索结果来改善用户体验，以便用户可以用尽可能少的努力（如点击）获取信息需求。因此，“零点击”在当下的含义更多的是指用户通过 SERP 的浏览直接满足了信息需求，因此不需要点击。例如，抖音、腾讯等商业搜索引擎都推出了“AI 搜”的产品，利用大语言模型的能力来整合信息内容，直接给用户提供高质量的答案，而不必和每个搜索结果交互。

随着信息检索技术的不断发展，“零点击”的现象变得越来越普遍。例如，Li et al.[265]研究了桌面和移动搜索日志中没有点击就结束搜索会话的现象，发现有大量的会话属于良性的“零点击”（即用户没有点击就顺利完成了搜索任务），尤其是在移动搜索中。此外，一些研究人员借助页面内容和用户与搜索引擎结果页的交互，在桌面[274]和移动设备[266,275]上检测“零点击”场景下的用户满意度。在搜索评估方面，Khabsa et al.[41]提出了一种在线指标，避免仅使用点击作为正面信号，纳入对“零点击”现象的考虑。更进一步，研究者们探索了一种理想的信息获取场景，称为“零查询”。系统无需等待用户输入查询和点击搜索结果，而是自主决定何时提供何种信息[276]。

此外，有另一类研究通过考虑结果级别的点击必要性来解决这一问题，这通常需要外部人员的标注。例如，Luo et al.[277]提出了点击必要性的概念以及一种移动搜索下的新评估指标。Zhang et al.[278]提出了一种联合相关性估计模型，使用点击必要性作为特征，并在性能上优于最先进的搜索排序方案。除了收集点击必要性的外部评估者标注外，一些研究将点击必要性视为可训练的参数以构建点击模

型，尤其在移动搜索中取得了更好的性能[279]。然而，点击必要性和无点击结果的相关性并不直接相关，一个点击必要性低的搜索结果不论有用还是无用都很可能没有被点击。本章的贡献在于补充了现有关于“零点击”搜索的研究，利用无点击行为中的脑活动来对用户的认知过程进行更有效的建模，并探索其在估计结果相关性方面的有效性。

5.2.3　搜索结果的有用性

有用性是互联网搜索中的关键概念。与第三方评估者标注的相关性不同，有用性代表了用户对搜索结果是否能够满足其信息需求的看法[280]，是一种主观的相关性。例如，Mao et al.[281]发现被第三方标注者标注为高相关性的信息内容并不一定对用户有用。他们揭示有用性与用户满意度的一致性要高于由外部评估者标注的相关性。基于这些发现，他们进一步提出了桌面搜索场景[282]和移动搜索场景[283]中的有用性预测模型。本章的用户实验采集了用户自己标注的有用性来作为相关性，因此不区分有用性和相关性两个概念。

5.3　事实性搜索任务中的“零点击”场景探究

5.3.1　数据采集

5.3.1.1　实验任务构建

本章首先从搜索结果相关性（Search Result Relevance，SRR）数据集[278]中采样了 150 个查询用于用户研究。使用该数据集的原因有两个：（1）它包含大量的真实查询日志、搜索结果快照和访达页内容，每个查询有十个对应的结果；（2）它根据搜索结果的展示风格提供了结果类型的人类标注。为了确保查询任务有更大的可能导致“零点击”场景的发生，并且查询词易于被参与者理解，本研究的采样基于以下两个标准：（1）该查询在数据集的日志中没有点击交互；（2）查询具有清晰明确的描述。

之后，本研究招募了 15 名标注者对所采样的查询对应的搜索结果进行标注。对于每个结果，其点击必要性（二元）和相关性（基于四级 Likert 量表）由至少三名不同的标注者判断，最终标注结果取标注者评分的中位数。最终，本研究从 150 个查询中筛选出 90 个查询，这些查询至少包含五个被标注为点击必要性较低（$\leqslant$ 0.5）的搜索结果。

5.3.1.2　参与者

本研究招募了 18 名年龄在 19 至 26 岁之间的大学生（均值 =21.56，标准差 =1.82）。参与者数量和现有的基于 EEG 的用户研究相当（例如，Duan et al.[118]的研究中的 15 名参与者和 Allegretti et al.[71]的研究中的 20 名参与者）。其中包括十名男性和八名女性，主要专业为计算机科学、物理学、艺术和工程等。所有参与者都汇报熟悉搜索引擎的基本使用，平均每天或每两天就会使用一次搜索引擎。整个实验任务大约需要两小时完成：其中 50 分钟用于准备和休息，60 分钟用于主任务，10 分钟用于实验问卷填写。每位参与者在认真完成所有任务后将获得 240 元人民币的报酬。

5.3.1.3　实验流程

本次用户研究获得了清华大学心理学系伦理委员会的批准（2021 伦审第 18 号）。在实验开始时，参与者将填写一个包含人口统计信息的问卷，并签署关于安全和隐私保护的知情同意书（参见图 6.1）。然后，他们将阅读用户实验的说明，了解用户研究期间完成搜索任务的流程。在进入主实验阶段之前，参与者将完成一个包含两个搜索任务的训练阶段，以确保他们熟悉搜索任务的流程。参与者被要求在 60 分钟内尽可能多地完成搜索任务，并且可以在任务之间休息，休息时间不计入 60 分钟的时间限制。

主实验中的每个搜索任务包含相同的步骤：

1. 参与者查看从数据集中随机选择的一个查询及其相关描述，查看结束后可以按下按钮进入搜索阶段。
2. 屏幕中心呈现 1.5 秒的十字符号以提示注视点的位置。接下来，我们会从该查询下对应的搜索结果列表中随机选择一个搜索结果，该搜索结果的快照（示例参见图 5.2）会出现在屏幕上，持续 2.5 秒。之后，用户可以从“跳过”、“点击”和“结束搜索”这三个选项中选择。这一过程遵循之前的研究[66]，通过在刺激出现后 2.5 秒内禁止用户操作，避免了与移动光标和点击按钮相关的脑活动干扰。该 2.5 秒内的脑信号被用于本章的分析和实验。
3. 当参与者选择“点击”一个搜索结果时，将显示相应结果的访达页。在参与者浏览完该访达页后，参与者可以选择结束搜索或继续浏览下一个结果。
4. 如果参与者结束搜索，他们将看到一个验证与评价页面。在此页面上，他们需要通过语音输入简要回答搜索任务，并报告他们感知到的任务难度（五级 Likert 量表）和对每个结果标注相关性（四级 Likert 量表）。

每位参与者完成的搜索任务顺序以及搜索结果的展示顺序均经过随机化处理。

表 5.1　参与者标注的搜索结果相关性在点击/无点击场景下的数量

响应	相关性			
	1	2	3	4
点击	14.0 (±14)	7.4 (±7)	10.2 (±11)	12.8 (±13)
无点击	101.2 (±53)	30.8 (±21)	31.5 (±20)	52.5 (±16)

在正式实验开始之前，本章还进行了一项涉及四个额外用户的试点研究，以调整实验设置，包括注视十字和结果的显示时间、训练任务的数量等。

5.3.1.4　设备与 EEG 预处理

实验使用 27 英寸显示器（分辨率 2,560×1,440）的台式电脑与 Chrome 浏览器。采用 Scan NuAmps Express 系统（Compumedics Ltd.）与 64 通道 Quik-Cap 电极帽（按国际 10-20 系统布置）采集 EEG 信号。EEG 设备的通道/电极分布可参见图 3.8(b)。电极阻抗控制在 $25k\Omega$ 以下，采样率 1,000Hz。

预处理流程和第三章类似，采用了标准的方法[19,69]。本研究截取了刺激呈现后 2,500ms 的 EEG 数据，依次进行平均乳突参考、500ms 前刺激基线校正、0.5-50Hz 带通滤波和降采样至 500Hz。为满足实时相关性反馈需求，本研究并未采用独立成分分析等更耗时的伪迹去除方法。

5.3.1.5　数据的统计信息

收集的数据集由 90 个搜索任务中的 1,252 次交互组成，参与者平均每个任务下会浏览 3.61（标准差 =2.24）个搜索结果。每名参与者平均完成 69.56（标准差 =12.23）个任务，并检查 250.78（标准差 =56.53）个搜索结果。表 5.1展示了参与者标注的搜索结果相关性的数量分布（点击和无点击）。本研究观察到大约 85.9% 的搜索结果未被点击，其中 46.8% 是“完全没用”（相关性 =1），其次是“非常有用”（相关性 =4），而“相当有用”（相关性 =3）和“有点有用”（相关性 =2）较少。

5.3.2　数据分析

5.3.2.1　脑电信号预处理

脑电数据通常包含与电源线、眨眼、身体运动等相关的噪声，需按标准流程预处理以便进一步分析。本研究采用的标准预处理流程与第 3.4节类似，包括：重参考至平均乳突、基线校正、低通 50Hz 和高通 0.5Hz 滤波、伪迹去除、降采样至 500Hz。对于伪迹去除，本研究采用参数化噪声协方差模型[142]以消除眼动、心电

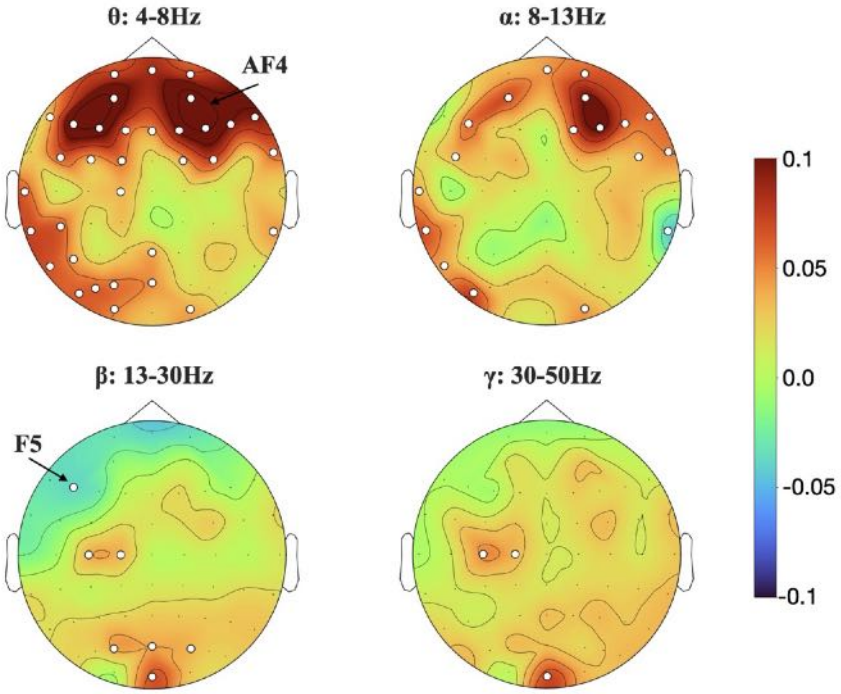


图 5.3　参与者的相关性标注与脑电频谱功率的皮尔森相关系数。高亮的脑电通道表示相关性水平在 $p < 0.05$ 时显著。

和肌电相关伪迹。随后，本研究根据每个搜索结果的呈现时间点提取感兴趣的脑电片段（实验设置中为 2,500ms），并利用刺激前 0-1,500ms 时段的脑电再次进行基线校正。

5.3.2.2　脑电频谱与内容相关性的关联分析

基于去噪后的脑电数据，本研究分析了用户感知的相关性与其不同频段脑电频谱功率的关系。该方法被广泛应用于探究不同类型刺激下的大脑活动差异[284]。对于每个通道，本研究根据 Welch 方法提取 4-50Hz 范围内的频谱功率，并分别在 θ（4-8Hz）、α（8-13Hz）、β（13-30Hz）和 γ（30-50Hz）频段求平均，获取脑电节律功率。接下来，本研究计算所有参与者脑电频率功率变化与无点击结果的相关性的平均皮尔逊相关系数，结果如图 5.3所示。

从图 5.3中可以看出，θ 和 α 频段比 β 和 γ 频段有更多相关性显著的 EEG 通道，其中最显著的电极出现在 AF4。θ 和 α 频段与认知表现相关[153]。这些通道的相关系数在这些频段都大于零，表明当参与者检验高相关性的内容时脑电的频谱功率显著增强。另一发现是 θ 和 α 频段，有显著相关性的脑电通道主要分布于额叶和左颞叶区域。基于 fMRI 的相关性感知研究[145]表明，处理相关与非相关文档时这些脑区（额叶和左颞叶）活动存在差异。本章的研究进一步地验证了这个差异。

对于 β 和 γ 频段，相关性与脑电频谱功率出现显著相关性的通道主要出现在枕叶和左中央区，其中最显著的电极为 F5。先前研究[118,285]表明高频脑电信号（如 β 和 γ 频段）和高级认知功能更相关，特别是和用户情感相关。因此可以推测信息需求是否被实现及用户是否满意可能诱发类似积极情感的高级认知模式。近期的一些相关研究揭示，内侧前额叶皮层、左额下回和枕中回等脑区在信息需求被实现时的执行功能存在显著差异[65]。本章在这些频段发现的显著活跃的脑区（左额叶和枕叶）与他们的研究部分吻合，说明这些区域可能与用户对有助于信息需求满足的有用结果的响应相关。

5.3.3　基于脑信号的相关性估计实验

5.3.3.1　实验设置

为了探索脑信号在相关性估计中的有效性，本章比较了不同的输入，包括脑信号、内容信息、上下文信息及这些特征的组合。

内容/上下文信息　在相关性判断的研究中，Mao et al.[282]研究了影响桌面搜索场景中相关性判断的因素，并将其研究扩展到了移动设备上[283]。他们使用两种特性来进行相关性估计任务：（1）内容信息（即当前搜索结果的特征）和（2）上下文信息（即交互历史的特征）。本实验中使用的模型特征主要继承自他们的研究，并排除了与用户行为相关的那些特征，这是因为无点击结果没有停留时间、滚动等行为数据。此外，本章补充了结果类型的内容特征，这在他们的研究中未被使用，但被 Williams et al.[275]证明是与点击必要性相关的因素。结果类型根据数据集中给出的展示风格分为 19 类，例如“自然结果”和“垂直问答”。因此，在本章的实验中，内容特征包括查询-结果对的相似度排名（基于 BERT 编码计算）、查询-结果对的 *BM25* 排名和结果类型。上下文特征包括与先前搜索结果的平均/最大相似度排名、与先前搜索结果的平均/最大/累积相关性评分和先前搜索结果的数量。

脑信号　对于频域特征，和第 3.4节一致，本章提取了 DE 作为脑信号的特征，包括 62 个不同通道（不包括重参考通道 M1 和 M2）上的五个频段（δ: 0.5-4Hz, θ: 4-8Hz, α: 8-13Hz, β: 14-30Hz, γ: 30-50Hz）。

对于时域特征，Jia et al.[286]认为其与常用频谱特征具有较好的互补性。在信息检索领域，时域特征被证明与相关性判断和信息搜索中的决策相关[287]。因此，本章在实验中通过将每个通道的原始脑电数据降采样至 50Hz 来提取时域特征。

模型　本章在实验中分别采用了两种模型：一种是拓扑不变的 DT 模型，另一种是拓扑感知的时频空网络（Spatial-Spectral-Temporal based Network，SST）[286]。DT 在机器学习任务中被广泛使用，因为它可以自动选择和组合脑电特征，并考虑它们与预测值的相关性。在本章的实验中，仅将频谱特征输入 DT 分类器，以避免高维问题。SST 应用注意力机制自适应捕捉频谱和时间信息中的判别模式，在一些脑电预测任务中被报告能达到最先进的性能。它可以自动捕捉相邻或对称通道的脑电特征，这其中可能包含重要的信息[114]。

至于内容和上下文信息的建模，本章在数据集中比较了 DT、MLP 和 SVM 的性能。其中，DT 表现最好，Mao et al.[282]也在相关性估计实验中应用了它。因此，本章使用 DT 进行上下文信息的建模。随后，本章还引入了考虑不同信号源的融合模型。本章使用权衡参数 λ（从.05 到.95 的 19 个值）进行网格搜索，以融合基于内容/上下文信息和脑信号的模型的估计得分。此外，本章也报告了使用不同权衡参数 λ 设置下融合模型的性能。

定义　为避免歧义，本章使用 F^f 表示使用特征 f（内容、上下文或脑信号）的模型 F（DT 或 SST）。cn、cx 和 bs 分别表示内容特征、上下文特征和脑信号特征。+ 表示使用权衡参数 λ 来融合不同模型的得分。例如，$\mathrm{DT}^{cn,cx} + \mathrm{SST}^{bs}$ 表示使用内容和上下文特征的 DT 与使用脑信号的 SST 的融合模型。

训练策略与评价　本章仅考虑相关性得分为 1 和 4 的样本，将该任务简化为二分类问题。其原因是：（1）评分 1（“完全无用”）和 4（“非常有用”）是边界情况下的相关性判断，因此它们的噪声比评分 2 和 3 少。（2）标注得分为 1 和 4 的样本占据了大部分（71.2%）的搜索结果。

为了验证不同应用场景下的性能，本章在实验中使用了两种数据划分的策略：*用户独立*和*任务独立*。在用户独立策略中，对每个参与者的数据进行验证时，使用其他参与者的数据来训练监督模型。任务独立策略将任务划分为十折，并在验证每一折时使用其余折进行训练。

在评价指标方面，本章遵循 Mao et al.[282]的工作，使用 AUC 作为任务的评价指标，以应对不平衡类的问题，并报告不同折间的 AUC 的标准差。

本节首先详细说明使用不同信息源进行相关性估计的整体性能，以展示脑信号在解决相关性估计方面的有效性。然后，本节分析了不同信息源和实验设置（例如，任务难度和时间间隔长度）对于使用脑信号进行相关性估计的影响。

表 5.2　使用不同信息源输入进行相关性估计的性能表现[a]

模型[b]	用户独立		任务独立	
	AUC	STD	AUC	STD
DT^{cn}	0.619**	0.040	0.593**	0.080
DT^{cx}	0.664**	0.047	0.585**	0.049
DT^{bs}	0.585**	0.047	0.642	0.033
SST^{bs}	0.654**	0.043	0.655	0.037
$DT^{cn,cx}$	0.672**	0.049	0.614*	0.067
$DT^{cn,cx} + DT^{bs}$	0.687**	0.049	0.683	0.049
$DT^{cn,cx} + SST^{bs}$	**0.718**	0.040	**0.687**	0.050

[a] *cn*、*cx* 和 *bs* 分别表示内容、上下文和脑信号。+ 表示融合两种信号。*/** 表示与 $DT^{cn,cx} + SST^{bs}$ 相比，性能差异基于 p<0.05/0.01 显著。

[b] F^f 表示使用特征 f 的模型 F。

5.3.3.2　整体实验结果

表 5.2展示了基于各种信息源（即内容、上下文、脑信号及其组合）的不同模型的相关性估计的整体性能。从表 5.2中，本节有以下几点发现：

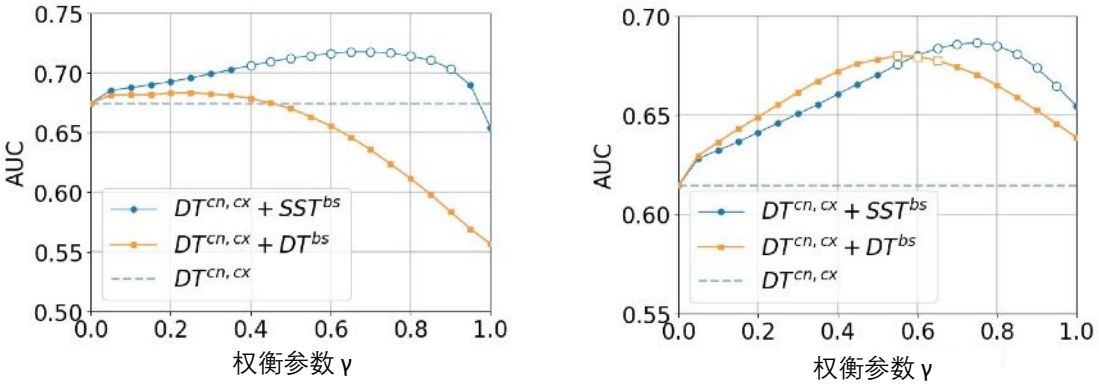
（1）对于两种训练策略，使用所有特征的模型比未考虑脑信号的模型性能显著更好。其中，$DT^{cn,cx} + SST^{bs}$ 表现出了最佳的性能，该方法使用 SST 进行脑信号建模，并将其与 $DT^{cn,cx}$ 的得分进行加权求和。这一观察表明，脑信号补充了包括内容和上下文特征在内的传统信息，并且有助于相关性估计。

（2）在基于脑电特征的模型当中，SST 比 DT 表现更好，尤其是在用户独立的策略中。这表明在脑信号建模中考虑拓扑结构并利用注意力机制是有效的。

（3）在基于脑电特征的模型中，用户独立策略的性能通常比任务独立策略差，尤其是在使用 DT 作为分类模型时，这种差异更加明显。此外，用户独立策略中的标准差也大于任务独立策略的标准差。原因可能有两个方面：一方面，先前的研究有揭示过一种“脑电不敏感”现象[288]，表明大约 15 – 30% 的用户可能会在脑机接口系统中表现不佳。另一方面，不同用户的脑信号存在个性化的差异[289]。进一步，本节的实验还发现，跨被试者的个性化脑电差异对于深度神经网络 SST 的影响比传统分类器 DT 更小。

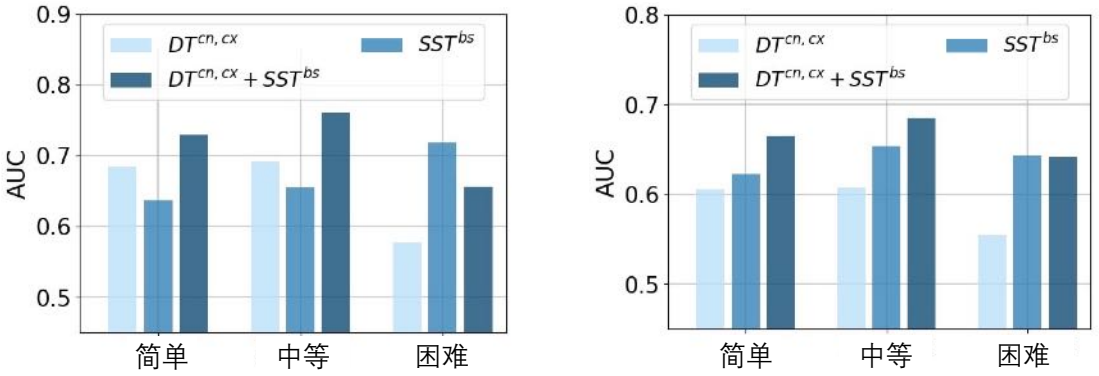
（4）对于不使用脑信号的模型，它们在任务独立训练策略中的表现比用户独立策略要差。其原因是某些内容和上下文特征（例如，BM25 排名）与任务相关。

因此，它不能很好地在任务独立的实验设置中发挥作用。然而，使用脑信号的模型在任务独立策略中的表现并不逊于用户独立策略。其原因可能在于这些模型能够直接捕捉用户反馈，从而相比仅应用传统特征的模型，受到任务影响较小。



(a) 用户独立　　(b) 任务独立

图 5.4　使用不同权衡参数 γ 的相关性估计性能



(a) 用户独立　　(b) 任务独立

图 5.5　在不同任务难度下的相关性估计性能

5.3.3.3　信息源分析

本节通过使用权衡参数 γ 结合脑信号和内容/上下文特征的信息源估计的分数，设计了融合模型，并进行以下分析：（1）哪些 γ 的取值下融合模型的性能显著优于 $DT^{cn,cx}$ 以及（2）融合模型是否对 γ 敏感。在图 5.4中，本节展示了使用所有特征的模型（$DT^{cn,cx} + DT^{bs}$ 和 $DT^{cn,cx} + SST^{bs}$）在不同权衡参数 γ 下的性能。如果 $\gamma = 0$，模型会退化为仅使用内容和上下文特征的 $DT^{cn,cx}$。而如果 $\gamma = 1$，模型会与仅使用脑信号的 DT^{bs} 或 SST^{bs} 一致。空心点表示融合模型在配对检验中显著优于 $DT^{cn,cx}$（$p<0.05$）。由于 SST 在脑信号建模中表现更好，本节主要讨论融合模型 $DT^{cn,cx} + SST^{bs}$，并有以下两个发现。

一方面，随着 γ 的增加，$DT^{cn,cx} + SST^{bs}$ 先单调增加到最佳性能，然后在达到

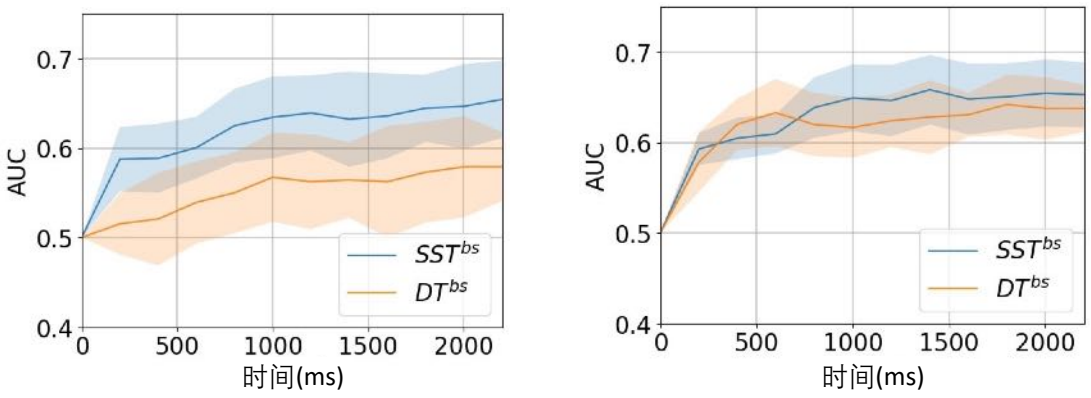
γ 的最佳取值后逐渐下降。这一发现表明，将传统信息和脑信号结合在一起比仅使用一种信号效果更好。

另一方面，在用户独立策略中，当 $0.4 \leqslant \gamma \leqslant 0.85$ 时，$DT^{cn,cx} + SST^{bs}$ 显著优于 $DT^{cn,cx}$；在任务独立策略中，当 $0.55 \leqslant \gamma \leqslant 0.9$ 时，效果显著。然而，在 $0.15 \leqslant \gamma \leqslant 0.85$ 范围内改变 γ 时，两种训练策略并没有显著的差异。这意味着融合模型可能对该参数不敏感。

5.3.3.4　任务难度分析

任务难度是指用户对完成搜索任务所需努力的评估[290]，是搜索引擎优化的重要因素。本节将用户在实验中标注的任务难度分为三组：简单（非常简单和简单）、中等（既不简单也不难）和困难（困难和非常困难）。在此基础上，本节计算了相关性估计的性能，如图 5.5所示。

在两种训练策略中，$DT^{cn,cx}$ 在困难任务中的表现比在简单任务中差。特别是在用户独立的训练策略中，重复测量方差分析显示不同任务难度水平之间存在显著差异（$F[20,2]=4.24$，$p<0.05$）。事后 Bonferroni 检验进一步揭示了基于传统特征的模型 $DT^{cn,cx}$ 在困难任务中的表现比简单和中等任务差（$p<0.05$）。相比之下，基于脑信号的模型（SST^{bs}）的性能不会随着任务难度的增加而下降。特别是在用户独立策略中，尽管不显著，但其在困难任务中的表现甚至优于简单和中等搜索任务。这意味着相关性判断的脑活动可能在不同任务难度上都具有相似的模式，所以并不会受到任务难度的太大影响。



(a) 用户独立　　(b) 任务独立

图 5.6　使用刺激出现后不同时间间隔的脑信号进行相关性估计的性能

5.3.3.5　时间间隔分析

由于脑活动与用户接受信息内容的时间相关，本节进一步探索从不同长度的时间间隔中提取的脑信号对相关性估计模型性能的影响。图 5.6展示了两种基于脑

电的模型的实验结果，模型所使用的脑信号特征来自于时间区间 $[0, T]$，其中 T 从 0ms 到 2,500ms。当 $T = 0$ 时，相当于模型没有获得任何输入。随着时间间隔长度的增长，SST^{bs} 和 DT^{bs} 的性能的提升逐渐减缓。在两种训练策略中，800ms 之后的模型性能没有显著差异。这一发现与现有工作一致，表明我们的大脑需要大约 800ms 来判断视觉呈现的刺激的相关性[71]。

5.3.4　搜索结果重排序实验

本节设计了两种搜索结果重排序方法，即个性化模型（Personalized Model，PM）和通用意图模型（Generalized Intent Model，GIM）。PM和GIM基于第 5.3.3节中预测的相关性，实现搜索结果的重排序，以提升用户的信息获取效率。

5.3.4.1　问题定义

本节将第 t 个搜索任务中第 p 个用户交互的第 i 个搜索结果记为 $d_p^{t,i}$，并将用户对它的相关性判断记为 $u_p^{t,i}$。需要注意的是，不同用户在同一个搜索任务中交互的搜索结果是不同的，即 $\langle d_{p_j}^{t,j_1}, d_{p_j}^{t,j_2}, ...\rangle \neq \langle d_{p_k}^{t,k_1}, d_{p_k}^{t,k_2}, ...\rangle, j \neq k$。这是因为搜索结果列表是随机排列的，且用户可以随时中断搜索。搜索结果重排序的目标是根据预测的相关性得分来获取更好的搜索结果排序。

5.3.4.2　方法

基线　本节采用两个基线模型进行比较：概率检索模型 BM25[22]和预训练模型 BERT[291]。这些模型仅考虑查询和搜索结果的内容信息，而没有考虑搜索过程中的用户信号。

个性化模型　个性化模型基于每个用户的预测相关性分数对搜索结果进行重排序。重排序列表 σ_p^t 可表述为：

$$\sigma_p^t = \S(\langle d_p^{t,i_1}, d_p^{t,i_2}, ...\rangle, \langle \hat{r}_p^{t,i_1}, \hat{r}_p^{t,i_2}, ...\rangle) \tag{5.1}$$

其中 $\hat{r}_p^{t,i}$ 是由相关性估计模型预测的相关性分数，§ 表示根据预测的相关性分数 $\langle \hat{r}_p^{t,i_1}, \hat{r}_p^{t,i_2}, ...\rangle$ 对 $\langle d_p^{t,i_1}, d_p^{t,i_2}, ...\rangle$ 进行排序的函数。

通用意图模型　通用意图模型的目标是利用用户的群体智慧生成该查询下更准确且通用的意图表示，如算法5.1所示。对于给定用户 p_i 和任务 t，我们首先聚合搜索结果向量 $\langle \vec{d}_{p_j}^{t,j_1}, \vec{d}_{p_j}^{t,j_2}, ...\rangle, j \neq i$，即根据它们的预测相关性分数 $\langle \hat{r}_{p_j}^{t,j_1}, \hat{r}_{p_j}^{t,j_2}, ...\rangle, j \neq i$

算法 5.1 跨用户的通用意图模型算法

输入: 目标用户 i 的候选文档 $\langle d_{p_i}^{t,i_1}, d_{p_i}^{t,i_2}, ...\rangle$;
　　其他用户 j 交互过的文档 $\langle d_{p_j}^{t,j_1}, d_{p_j}^{t,j_2}, ...\rangle$;
　　基于脑信号估计的相关性分数 $\langle \hat{r}_{p_j}^{t,j_1}, \hat{r}_{p_j}^{t,j_2}, ...\rangle, j \neq i$。
数据:
　　预测相关性分数列表 $\langle s_{p_i}^{t,i_1}, s_{p_i}^{t,i_2}, ...\rangle$; 意图向量 $\vec{I}$;
初始化:
　　$s_{p_i}^{t,i} = 0; \vec{I} = \vec{0}; \vec{d_p^{t,i}} = BERT(d_p^{t,i})$;
　　$\overline{\hat{r}_{p_i}^t} = \text{average}\{\hat{r}_{p_j}^{t,j_k} \in \langle \hat{r}_{p_j}^{t,j_1}, \hat{r}_{p_j}^{t,j_2}, ...\rangle, j \neq i\}$;
for $j \neq i$ **do**
　for $d_{p_j}^{t,j_k} \in \langle d_{p_j}^{t,j_1}, d_{p_j}^{t,j_2}, ...\rangle$ **do**
　　$\vec{I} = \vec{I} + \vec{d_{p_j}^{t,j_k}} \cdot (\hat{r}_{p_j}^{t,j_k} - \overline{\hat{r}_{p_i}^t})$;
　end for
end for
for $s_{p_i}^{t,i_k} \in \langle s_{p_i}^{t,i_1}, s_{p_i}^{t,i_2}, ...\rangle$ **do**
　$s_{p_i}^{t,i_k} = \text{cosine\_similarity}(\vec{I}, \vec{d_{p_i}^{t,i_k}})$;
end for
$\sigma_{p_i}^t = \S(\langle d_{p_i}^{t,i_1}, d_{p_i}^{t,i_2}, ...\rangle, \langle s_{p_i}^{t,i_1}, s_{p_i}^{t,i_2}, ...\rangle)$;
返回 $\sigma_{p_i}^t$;

生成意图向量 $\boldsymbol{I}$，其中 $\vec{d_{p_j}^{t,j_k}}$ 是使用预训练的“bert-chinese-base”编码器[24]对 $d_{p_j}^{t,j_k}$ 的表示。然后，我们通过计算意图向量 $\boldsymbol{I}$ 与搜索结果向量 $\vec{d_p^{t,i}}$ 的余弦相似度，为用户 p 和任务 t 生成搜索结果分数 $s_p^{t,i}$。最后，我们根据它们的分数 $\langle s_p^{t,i_1}, s_p^{t,i_2}, ...\rangle$ 进行排序来计算重排序列表 $\sigma_{p_i}^t$。

5.3.4.3　实验设置

搜索结果重排序实验基于 $\text{DT}^{cn,cx}$、SST^{bs} 和 $\text{DT}^{cn,cx} + \text{SST}^{bs}$ 估计的相关性分数进行，这是因为这些方法在使用相同输入信息源的情况下优于其他模型。本节发现，在任务独立和用户独立的数据划分中，搜索结果重排序的实验结果表现出高度的一致性。因此，本节仅汇报任务独立策略中的实验结果。为避免歧义，本节使用 $\boldsymbol{F}^f$ 表示使用特征 f（取值包括 cn，cx 和 bs）的模型 $\boldsymbol{F}$（取值包括 BM25，BERT，PM和GIM）。例如，BM25 表示利用内容特征的 BM25 模型，$\text{GIM}^{cn,cx,bs}$ 表示基于 $\text{DT}^{cn,cx} + \text{SST}^{bs}$ 相关性分数的通用意图模型。

为了比较不同模型和特征的性能，本节使用了两个常见的评估指标：NDCG[240]和平均倒数排名（Mean Reciprocal Rank，MRR）[116]。由于搜索任务中文档的平均数量为 3.41，本节使用相对较小的三个截止位置 $\{1, 3, 5\}$ 计算 NDCG，即 NDCG@$\{1, 3, 5\}$。

表 5.3　搜索结果重排序任务在不同信息源和不同模型下的性能[a]

Model	NDCG@1	NDCG@3	NDCG@5	MRR
$BM25^{cn}$	0.407*	0.672*	0.725*	0.621*
$BERT^{cn}$	0.399*	0.691*	0.737*	0.655*
$PM^{cn,cx}$	0.446*	0.714*	0.751*	0.677*
PM^{bs}	0.457*	0.725*	0.764*	0.691*
$PM^{cn,cx,bs}$	0.522*	0.752*	0.787*	0.726*
$GIM^{cn,cx}$	0.490*	0.739*	0.775*	0.709*
GIM^{bs}	0.571	0.776	0.811	0.754
$GIM^{cn,cx,bs}$	**0.591**	**0.787**	**0.814**	**0.764**

[a] F^f 表示使用特征 f 的模型 F。cn、cx 和 bs 分别表示内容、上下文和脑信号。* 表示与 $GIM^{cn,cx,bs}$ 的性能差异在 $p<0.01$ 的水平上显著。

5.3.4.4　实验结果

表 5.3展示了使用不同特征的重排序方法的排名性能，其主要观察如下：

（1）$BERT^{cn}$ 在大多数评价指标（即 NDCG@3、NDCG@5 和 MRR）上的表现优于 $BM25^{cn}$。然而，在所有评估指标中，它们的表现显著低于利用了用户信号的个性化模型和通用意图模型。这是因为搜索结果的相关性不能仅通过语义相似度分数来简单判断[292]。

（2）对于个性化模型和通用意图模型，使用脑信号的模型（即 $PM^{cn,cx,bs}$ 和 $GIM^{cn,cx,bs}$）表现显著优于仅使用内容和上下文信息的模型（即 $PM^{cn,cx}$ 和 $GIM^{cn,cx}$）。这一结果表明，利用脑信号可以提高重排序性能。

（3）$GIM^{cn,cx,bs}$ 的表现显著优于 $PM^{cn,cx,bs}$，这表明通用意图模型要更为有效。此外，通用意图模型通过在相应的搜索任务中建模搜索意图来对结果列表进行重排序，因此可以适用于新用户和新搜索结果等场景。相比之下，个性化模型基于预测相关性分数对搜索结果进行重排序，因此不适用于新用户和新搜索结果等场景。个性化模型的表现不如通用意图模型，其原因很可能在于个性化模型在仅建模单个用户的搜索行为和脑活动时存在不稳定性。然而，值得一提的是，个性化模型在考虑用户个性化信息方面具有优势，因此在一些需要高度个性化的潜在场景中可能会更有效。

5.4　非事实性搜索任务探究

5.4.1　数据采集

本章开展了一个模拟真实的非事实性搜索过程的用户实验。参与者需通过交互网页完成搜索任务，期间其脑电信号被同步采集。本实验遵循清华大学心理学系伦理委员会批准的伦理规范（2021 伦审第 18 号），并采取了多重措施保护参与者隐私，包括数据匿名化处理、实验前获取知情同意书（参见图 6.1），并允许参与者随时终止实验。

5.4.1.1　参与者

本研究招募了 21 名年龄介于 19 至 27 岁的参与者（8 名女性，13 名男性，年龄均值 =23.85，标准差 =2.28）。所有参与者均为汉语母语者，并汇报称熟悉搜索引擎的使用。实验流程包含 1 小时神经信号采集、累计 30 分钟的休息、30 分钟的准备时间和 30 分钟问卷填写与设备佩戴指导。参与者按每小时 100 元的标准获得报酬，总酬劳约 200-300 元。

表 5.4　用户研究中使用的查询示例

查询	来源	任务描述
先知	iMine 0001	（i）查找关于伊斯兰教“先知”的信息；（ii）探索“先知”在一般领域中的概念；（iii）搜索一部名为“先知”的电影的高质量音源。
波斯猫	iMine 0002	（i）了解波斯猫的概念和特征；（ii）下载波斯猫的图片；（iii）了解波斯猫的市场价格；（iV）阅读以波斯猫为主题的书籍。
乘法表	iMine 0042	（i）下载乘法表；（ii）学习记忆乘法表的技巧。
钢笔	TREC 20582	（i）了解钢笔及其起源；（ii）查看钢笔的品牌和价格。
茶	TREC 20906	（i）了解茶的好处，如营养成分和功效；（ii）为课程演示了解茶的种类；（iii）探索茶的准备方法；（iv）探索不同商业品牌的茶。
牛奶实验	TREC 21092	（i）了解石蕊与牛奶之间化学反应的相关信息；（ii）了解与牛奶或酸奶的感官实验相关的信息；（iii）了解与牛奶有关的一般科学实验。

5.4.1.2　实验刺激材料准备

本研究构建了包含 100 个查询及其对应文档的数据集（每个查询平均对应 39.5 篇文档）。数据集已开源于https://github.com/THUIR/Brain-Relevance-Feedback。

查询与文档集构建　所有的查询选自数据集 NTCIR-11 IMine[293]和 TREC-2009[294]。数据集选择的依据如下：（1）数据集间相互独立；（2）多数查询简短且主题宽泛，便于设计具有不同意图的非事实性检索任务。本研究从中共选取 100 个候选查询，其中 50 个来自 NTCIR-11 IMine，50 个来自 TREC-2009。NTCIR-11 IMine 的中文查询（共 50 个）全部被选择。TREC-2009 的查询经筛选后翻译为中文，并排除中国用户可能不熟悉的特定实体（如“堪萨斯城”、“沃尼奇”等）。表 5.4展示了一些样例的查询。随后，本研究使用中文搜索引擎搜狗①为每个查询检索相关文档。从每条文档中提取文档快照（搜索引擎结果页中的文档快照）并爬取访达页（点击文档后跳转的网页）。所有网页经人工校验，剔除无效或爬取失败的文档，最终每个查询平均有 39.5 篇文档相关，每篇文档包含一个内容快照与访达页内容。

任务描述构建　本研究为每个查询撰写 2-5 个子任务描述（见表 5.4）。例如，查询“先知”的任务描述包括：（1）查找关于伊斯兰教“先知”的信息；（2）探索“先知”在一般领域中的概念；（3）搜索一部名为“先知”的电影的高质量音源。任务描述生成参考 Liu et al.[293]提出的子主题挖掘方法：由两位信息检索专业的博士生独立地将文档聚类并生成子任务描述，经讨论后整合和合并各自提出的子任务描述，最终每个查询平均会生成 2.6 个任务描述（标准差 =1.0）。

搜索结果快照标注　为评估相关性反馈性能，本研究另外招募了 9 名有搜索引擎使用经验的中文用户对快照进行相关性标注（相关/不相关）。每篇快照由 3 人标注，根据多数的标注来确定最终标签。标注者间 Fleiss’s κ 系数为 0.76，达到较高的一致性。由于成本限制，访达页相关性由参与者在完成搜索的同时自行标注。

5.4.1.3　实验流程

整体流程　实验开始前，被试者会填写人口统计学信息（年龄、专业等）及搜索习惯问卷。随后进行任务说明培训，培训包含两个和主任务流程一致的练习任务。正式任务限时 1.5 小时，参与者需完成尽可能多的搜索任务，在每完成 5 个任务后

① www.sogou.com.

可以短暂地休息。参与者平均每小时完成 36.8 个任务（标准差 =10.3）。实验结束后，参与者会填写一个搜索体验问卷。

主任务流程　主任务流程包括 4 个步骤，系统界面详见图 5.7。

S_1：用户查看查询词及关于该查询词随机分配的子任务描述（参见图 5.7(b)）。

S_2：一个注视点在屏幕中央呈现 0.5 秒，随后呈现一个文档快照（含标题、摘要和封面图片）。文档快照呈现的前 2 秒不可点击（避免点击动作产生脑电干扰），随后该快照的边框高亮并变为可点击状态。快照页面的内容参见图 5.7(c)与图 5.7(d)。参与者可点击进入访达页（见下一步），继续浏览（展示下一条快照）或结束搜索。为贴近真实搜索场景，实验中允许用户在该步骤回看或点击先前浏览过的文档，但仅采集首次浏览时的脑电数据被用于后续实验分析。

S_3：访达页内容被展示（访达页较文档快照通常包含更详实的内容，见图 5.7(e)）。用户在进行至少 2 秒的浏览后可继续搜索（返回前一步）或结束搜索。

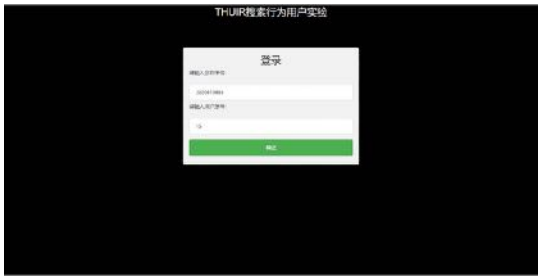
S_4：参与者对浏览过的文档进行相关性/有用性评分（四级 Likert 量表：1 对应“完全不相关”，4 对应“完全相关”）。参与者被要求独立地标注文档快照与访达页相关性/有用性，这是因为二者可能不一致（如文档快照看起来很相关，但访达页实际上不符合用户需求[5]）。

本实验所用设备和数据预处理与第 5.3节一致，唯一的差别是由于实验范式不同，选取的 EEG 数据片段为刺激后的 2,000ms。

5.4.1.4　用户研究的系统界面

图 5.7展示了用户研究的界面（以选自 TREC 20582 的查询“茶”为例）。在所有页面中，背景颜色设为黑色，这是用于减少无关因素的干扰的常见做法[295]。

图 5.7(a)中展示了系统的登录页面，参与者输入他们的学生 ID（若参与者非学生时使用随机生成的 ID）以及随机分配的用户 ID。登录后，搜索任务会依次呈现给参与者。图 5.7(b)-5.7(f)提供了搜索任务中不同步骤的截图。其中，图 5.7(b) 是搜索任务描述页面，参与者被要求仔细阅读顶部的查询词（例如“茶”），然后是任务描述（例如“为课程演示探索茶的种类，寻找来自两个或更多相关网站的信息”）。随后，参与者可以点击“开始搜索”进入搜索任务。图 5.7(c)和 5.7(d)展示了搜索任务的文档快照页面。最初的两秒钟内，文档的摘要是不可点击的（如图 5.7(c)所示），以确保排除与点击动作相关的脑活动[19,66]。在此期间之后，参与者有四个可选操作，如图 5.7(d)所示：(1) 点击文档进入相应的访达页，(2) 点击“下一个文档”查看下一个文档，(3) 点击“结束搜索”进入标注页面，(4) 浏览屏幕右侧显示



(a) 登录页面。

(b) S_1: 搜索任务描述页面。

(c) S_2: 快照页面（前 2 秒）。

(d) S_2: 快照页面（2 秒后）。

(e) S_3: 访达页面。

(f) S_4: 标注页面。

图 5.7　用户研究的系统界面

的曾经交互的文档并可以再次交互（例如点击）。图 5.7(e)展示了访达页，参与者可以滚动查看访达页上的所有内容。检查完毕后，参与者可以为访达页面标注相关性分数，并通过点击“返回”继续进行其他文档快照的浏览。结束搜索后，参与者会被呈现图 5.7(f)所示的标注页面。在此页面上，参与者需要简要口述搜索任务的答案，然后为每个文档快照标注相关性/有用性评分。

5.4.1.5　问卷分析

实验后问卷显示：57.1% 参与者认为实验流程与日常搜索差异较小，23.8% 认为相当，14.3% 认为差异较大，4.8% 认为差异极小，这表明实验的设计基本比较符合现实。用户反馈的主要差异在于不支持查询重构等行为，这是因为本章主要聚焦在单轮检索场景。本研究还采集了用户的搜索体验维度（舒适/压力/清晰/熟悉/有趣/困难）。在该用户研究中，平均而言用户认为实验任务舒适、清晰、熟悉且有趣，压力与困难程度适中。

5.4.1.6　数据统计

数据集包含 21 名参与者的 979 个搜索任务，人均完成 46.6 个任务（标准差=16.6）。在每个任务下，用户平均浏览 10.9 篇文档（总计 10,670 篇），点击 1.9 篇（总计 1,820 次）。参与者的标注显示：文档快照的平均相关性为 1.72，访达页的平均相关性为 2.67（仅点击文档有访达页评分）。参与者标注与第三方标注的 Kendall’s τ 相关系数为 0.66（p<$1e^{-3}$），显示较高一致性。数据集总共包含参与者浏览文档快照时的 10,670 段 EEG 信号及查看访达页时的 1,820 段 EEG 信号。

5.4.2　相关性反馈框架设计

5.4.2.1　问题定义

假设用户提交的查询为 q，与查询 q 关联的文档列表为 $\mathcal{D}=\{d_1,d_2,...d_n\}$。进一步，假设用户在结束搜索前浏览了查询 q 下的 h_{max} 个文档。在过程中的某个状态，$h\in\{1,2,...,h_{max}\}$ 个历史文档 $\mathcal{D}_h=\{d_1,d_2,...,d_h\}$ 已经被用户浏览，还有 $n-h$ 个候选文档 $\mathcal{D}_u=\{d_{h+1},d_{h+2},...,d_n\}$ 尚未被浏览。

本研究的相关性反馈框架旨在通过从历史文档 $\mathcal{D}_h$ 的交互中获取反馈信号，用于两个相关性反馈任务，即交互式相关性反馈（Interactive Relevance Feedback，IRF）和回顾式相关性反馈（Retrospective Relevance Feedback，RRF）。IRF 通过重新估计候选文档的相关性 $\boldsymbol{R}^{it}=\{r_{h+1}^{it},...,r_n^{it}\}$ 来重排序。它可以用于实时的搜索场景：随着检查文档数量 h 的增加，IRF 迭代地重新排序 $\mathcal{D}_u$ 以优化当前搜索过程。另一方面，RRF 通过重新估计历史文档的相关性 $\boldsymbol{R}^{re}=\{r_1^{re},...,r_h^{re}\}$ 来重排序历史文档 $\mathcal{D}_h$。它对当前搜索过程并没有直接帮助，但可以用于在具有类似意图的未来搜索过程中提供更好的排序列表。

为了评估 IRF 和 RRF，本研究使用基于排序的指标 Π（例如，NDCG[240]）。假设 $\mathcal{D}_u$ 和 $\mathcal{D}_h$ 的真实相关性分别为 $\boldsymbol{R}^{gu}=\{r_{h+1}^{gu},...,r_n^{gu}\}$ 和 $\boldsymbol{R}^{gh}=\{r_1^{gh},...,r_h^{gh}\}$，其中 r_i^{gu}（$i>h$）或 r_i^{gh}（$i\leqslant h$）是第 i 个文档的真实相关性。IRF 和 RRF 的性能可以分别表示为 $\Pi(\boldsymbol{R}^{gu},\boldsymbol{R}^{it})$ 和 $\Pi(\boldsymbol{R}^{gh},\boldsymbol{R}^{re})$。

5.4.2.2　相关性反馈框架

所提出的相关性反馈框架包括以下步骤：（1）伪相关性信号、点击信号和脑信号分别独立地转换为相关性得分 $\boldsymbol{R}^{p}$、$\boldsymbol{R}^{c}$ 和 $\boldsymbol{R}^{bs}$。脑相关性得分 $\boldsymbol{R}^{bs}$ 从文档快照的脑相关性得分 $\boldsymbol{R}^{\text{snippet}}$ 和文档访达页的脑相关性得分 $\boldsymbol{R}^{\text{land}}$ 中选择。（2）对于第 i 个历史文档 $d_i\in\mathcal{D}_h$，反馈信号 $r_i^p\in\boldsymbol{R}^p$，$r_i^c\in\boldsymbol{R}^c$ 和 $r_i^{bs}\in\boldsymbol{R}^{bs}$ 分别被融合成 IRF 和 RRF 的相关性得分 $r_i^{it,h}\in\boldsymbol{R}^{it,h}$ 和 $r_i^{re}\in\boldsymbol{R}^{re}$。为简单起见，本研究用 F_Θ 表示使用

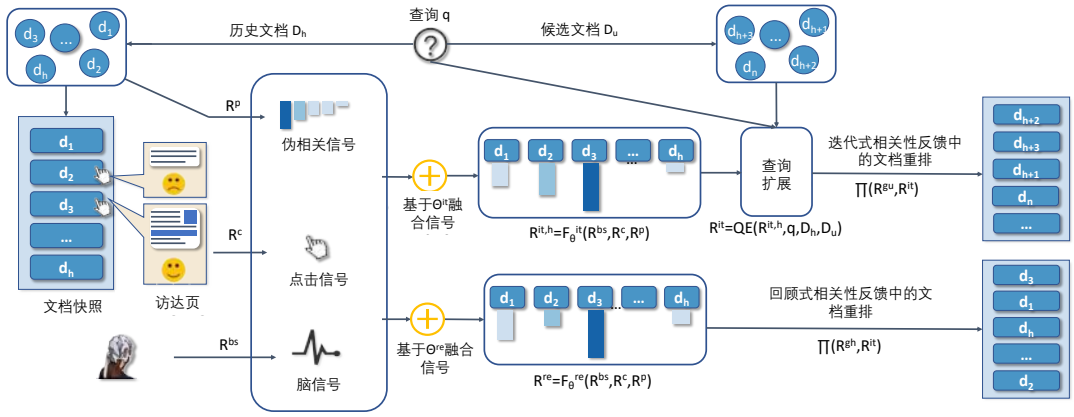


图 5.8　相关性反馈框架的结构。在搜索过程中，除了系统提供的伪相关性信号外，本研究还收集了用户的脑信号和点击信号。这些信号随后被融合成两个相关性得分，分别用于 IRF 和 RRF 任务。然后，RRF 中的融合相关性得分直接用于重新估计历史文档的相关性。而在 IRF 中，候选文档可以通过一个查询扩展模块重新排序，该模块利用 IRF 的融合相关性得分来提取历史文档的信息。

参数 Θ 进行的融合，例如 $R^{it} = F_{\Theta^{it}}(R^{bs}, R^{c}, R^{p})$。（3）在 RRF 中，历史文档 $\mathcal{D}_h$ 是根据融合的相关性得分 R^{re} 来直接重排序的。（4）IRF 通过一个查询扩展模块 QE 为候选文档生成相关性得分，该模块可以被公式化为 $R^{it} = \mathrm{QE}(R^{it,h}, q, \mathcal{D}_h, \mathcal{D}_u)$，其中历史文档的融合相关性得分 $R^{it,h}$ 被用于平衡每个历史文档在查询扩展模块 QE 中的重要性。

信号准备　为了对每个文档生成高质量的相关性估计，本研究独立地从伪相关性信号、点击信号和脑信号中获取相关性得分，并将其融合，详细如下：

- **伪相关性得分。**通过测量 q 与每个文档 $d_i \in \mathcal{D}_h$ 之间的语义相似性，本研究生成伪相关性得分 $R^{p} = \{r_1^{p}, r_2^{p}, ..., r_h^{p}\}$，其中 $r_i^{p} \in [0, 1]$ 是 d_i 的排序分数。对于查询 q 和文档 d，本相关性反馈框架中基于 BERT[255,291] 来测量它们的语义相似性，表示为 $\mathrm{BERT}(q, d)$。BERT 通过在 T^{2}Ranking 数据集 ① 上微调 BERT-Chinese ② 来初始化，使用数据集原始论文中的相同流程和可用代码[296]。该数据集包含了从搜狗搜索日志中提取的查询和文档的人类标注，与我们的用户研究构建的数据集类似。此外，我们使用快照内容代表文档 d，以估计语义相似性，因为它反映了用户在搜索引擎结果页上浏览到的内容，同时有效地涵盖了文档的主题并过滤掉多余的细节。
- **基于点击的相关性得分。**基于点击的相关性得分 $R^{c} = \{r_1^{c}, r_2^{c}, \cdots, r_h^{c}\}$ 是从用户的点击行为中生成的，其中 $r_i^{c} = 0$（或 1）表示用户跳过（或点击）第 i 个文档。

① https://github.com/THUIR/T2Ranking
② https://github.com/ymcui/Chinese-BERT-wwm

- **基于脑的相关性得分。** 使用 EEG 设备，收集用户对前 h 个文档的快照内容和访达页内容的脑响应，表示为 $\boldsymbol{X}^{\text{snippet}}$ 和 $\boldsymbol{X}^{\text{land}}$。其中一个脑信号样本 $x \in \{x_i^{\text{snippet}}, x_i^{\text{land}}\}$，表示对第 i 个文档的快照或访达页的脑响应。x 是空间 $\mathbb{R}^t$ 中的一个向量，其中 t 表示 EEG 特征的长度。基于脑解码模型 $\boldsymbol{G}$（详见第 5.4.3节），$\boldsymbol{X}^{\text{snippet}}$ 和 $\boldsymbol{X}^{\text{land}}$ 被转化为脑相关性得分 $\boldsymbol{R}^{\text{snippet}}$ 和 $\boldsymbol{R}^{\text{land}}$（分别对应快照和访达页）。注意，由于用户对第 i 个文档的访达页内容的脑响应（由 x_i^{land} 表示）在用户没有点击文档并进入访达页时并不存在，因此 r_i^{land} 并不是对所有样本都可用的。基于 $\boldsymbol{R}^{\text{snippet}}$ 和 $\boldsymbol{R}^{\text{land}}$，我们生成 $\boldsymbol{R}^{bs} = \{r_1^{bs}, \cdots, r_h^{bs}\}$，表示文档 $\mathcal{D}_h$ 的脑相关性得分。本研究分别为 IRF 和 RRF 使用不同的原则来计算 $\boldsymbol{R}^{bs}$。理由和细节如下：

  （1） IRF 在查询扩展过程中从 $\mathcal{D}_h$ 中提取信息。这是由于快照通常比访达页更简洁，且包含的噪声更少。由于快照内容而不是访达页内容被用来代表文档。因此，我们简单地设置 $r_i^{bs} = r_i^{\text{snippet}}$。

  （2） RRF 的目标是重排序历史文档 $\mathcal{D}_h$。如果一个文档 d_i 有一个吸引人的快照但访达页质量低下，通常会导致不良的用户体验[5]，因此不应被作为排名靠前的文档。因此，如果 r_i^{land} 可用，则 r_i^{bs} 被赋予 r_i^{land} 的值；否则，我们将 r_i^{bs} 设置为 r_i^{snippet}。

相关性得分融合　对于第 i 个文档（$i \leqslant h$），其融合相关性得分 $r_i^{it,h}$ 和 r_i^{re} 是基于伪相关性信号（r_i^p）、点击信号（r_i^c）和脑信号（r_i^{bs}）的相关性得分的加权和，具体可以表示为：

$$r_i^{it,h} = \theta^{it,bs} \cdot r_i^{bs} + \theta^{it,c} \cdot r_i^c + \theta^{it,p} \cdot r_i^p \tag{5.2}$$

$$r_i^{re} = \theta^{re,bs} \cdot r_i^{bs} + \theta^{re,c} \cdot r_i^c + \theta^{re,p} \cdot r_i^t \tag{5.3}$$

其中 $\Theta^{it} = \{\theta^{it,bs}, \theta^{it,c}, \theta^{it,p}\}$ 和 $\Theta^{re} = \{\theta^{re,bs}, \theta^{re,c}, \theta^{re,p}\}$ 是融合参数。本节为 IRF 和 RRF 使用不同的融合参数，原因如下：

（1） IRF 中生成 $\boldsymbol{R}^{bs}$ 和 RRF 中生成 $\boldsymbol{R}^{bs}$ 的依据不同，因此融合权重应有所不同。

（2） $\boldsymbol{R}^{it,h}$ 被用于从 $\mathcal{D}_h$ 的快照中提取文本信息来进行查询增强，而 $\boldsymbol{R}^{re}$ 被用于重新估计 $\mathcal{D}_h$ 的相关性。因此，它们之间存在固有的差异，例如 $\boldsymbol{R}^{it,h}$ 强调文档的快照内容是否相关，而 $\boldsymbol{R}^{re}$ 也应反映访达页的质量信息。

其中，Θ^{it} 和 Θ^{re} 可以是固定的，也可以根据不同的搜索场景进行调整，这一点在

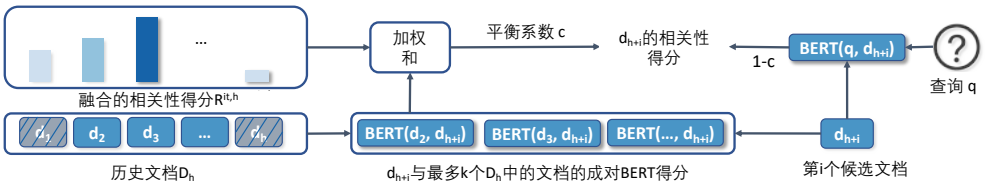


图 5.9　查询扩展模块的示意图

后文中会进行进一步分析。

查询扩展　查询扩展模块 QE 的结构如图 5.9所示。查询扩展模块基于融合相关性得分 $R^{it,h}$ 和历史文档 $\mathcal{D}_h$ 来重排序候选文档 $\mathcal{D}_u$。和第 5.3节一致，本节在相关性反馈框架中采用了 BERT 用于文档语义编码，并对相关性反馈算法进行了以下扩展：（1）将该模块扩展到 IRF 的交互设置中。（2）引入了从各种信号融合而来的相关性得分。查询扩展模块的详细步骤如下：

首先，我们从历史文档 $\mathcal{D}_h$ 中选择最多 k 个（k 在本研究中固定设为 10）融合相关性得分最高的文档，记为 $\mathcal{D}_s = \{d_{s_1}, \cdots, d_{s_k}\}$，对应的融合相关性得分为 $r_i^{it,h}$（$i \in \{s_1, s_2, \cdots, s_k\}$）。其次，对于每个候选文档 d_{h+i}，我们计算它与 $\mathcal{D}_s$ 中每个文档 d_{s_j} 的 BERT 分数 $\text{BERT}(d_{s_j}, d_{h+i})$。第三，计算出的 BERT 分数根据融合相关性得分 $R^{it,h}$ 进行加权求和，以获取第 $h+i$ 个文档的相关性得分 r_{h+i}^f：

$$r_{h+i}^f = \sum_{j=1}^{k} \frac{e^{r_{s_k}^{it,h}}}{\sum_{l=1}^{k} e^{r_{it,h}^{s_l}}} \cdot \text{BERT}(d_{s_j}, d_{h+i}) \tag{5.4}$$

第四，候选文档 d_{h+i} 的相关性得分通过其与初始查询 q 的 BERT 分数和相关性得分 r_{h+i}^f 进行加权求和：

$$r_{h+i}^{it} = r_{h+i}^f \cdot c + \text{BERT}(q, d_{h+i}) \cdot (1-c) \tag{5.5}$$

其中 c 是一个系数（设为 0.1），用于平衡反馈信息和初始查询的影响，这在现有相关性反馈方法中被广泛应用[33,297]。最后，候选文档 $\mathcal{D}_u$ 通过 $R^{it} = \{r_{h+1}^{it}, \cdots, r_n^{it}\}$ 重排序。

相关性反馈的训练与评价　算法 5.2展示了相关性反馈实验的训练和评估流程。与现有文献[19,62]中常用的不考虑数据样本自然顺序的数据集划分不同，本节采用按时间点划分的策略来训练和评估相关性反馈模型。当新的搜索任务呈现给参与者 u 时，本方法将首先训练和准备一个脑解码模型（如算法 5.2第 3-4 行所示），该模型可以基于来自相同参与者先前搜索任务的脑数据（个性化模型 G_p）或基于其他参

算法 5.2 整体相关性反馈流程与实验设置

1: **输入:** 用户 u，一系列搜索任务，每个任务由查询集 $\mathcal{Q}$ 和每个查询 $q \in \mathcal{Q}$ 对应的文档集 $\mathcal{D}$ 构成，从其他参与者的脑记录中训练的通用脑解码模型 G_g，文本排序模型 BERT。
2: **初始化:** 以随机参数初始化的个性化模型 G_p，超参数 $\Theta^{it} = \{\theta^{it}_{bs}, \theta^{it}_{c}, \theta^{it}_{p}\}$ 和 $\Theta^{re} = \{\theta^{re}_{bs}, \theta^{re}_{c}, \theta^{re}_{p}\}$，所有收集的脑数据样本 $X_{\text{global}} = []$，所有收集的用户标注 $\mathcal{R}_{\text{global}} = []$。
3: **输出：** IRF 和 RRF 的平均文档重排序性能 S_{IRF} 和 S_{RRF}。
4: 初始化脑解码模型 G 为 G_g，并将 S_{IRF} 和 S_{RRF} 初始化为空列表 []。
5: **for** $q \in \mathcal{Q}$ **do**
6: 　使用 X_{global} 和 R_{global} 训练 G_p。
7: 　如果收集的脑数据样本 X 的数量达到 100，将 G 转换为 G_p。
　　// 通过在搜索过程中重新排序候选文档来评估 IRF 性能。
　　// h_{max} 是用户 u 总共交互过的文档数量。
8: 　**for** $h \in \{1, ..., h_{max}\}$ **do**
9: 　　收集用户 u 关于文档 $d \in \{d_1, ..., d_h\}$ 的脑响应 X。
10: 　**end for**
11: 　使用脑解码模型 G 生成 R^{bs}。
12: 　根据用户 u 的点击行为生成 R^c，根据基于文本的排序分数 BERT(q, D) 计算 R^p。
13: 　$R^{it} = \text{QE}^{F_{\Theta^{it}}(R^{bs}, R^c, R^p)}$。　// 为 IRF 生成组合相关性分数。
14: 　利用第三方相关性标注为 $d \in \{d_{h+1}, ..., d_n\}$ 作为 R^{gu}。
15: 　S_{IRF}.append($\Pi(R^{gu}, R^{it})$)。　// 计算 IRF 的基于排序的指标。
16: 　// 在搜索结束后通过重新排序历史文档来评估 RRF 性能。
17: 　收集用户 u 对应于文档 $d \in \{d_1, ..., d_{h_{max}}\}$ 的脑响应 X。
18: 　为文档 $d \in \{d_1, ..., d_{h_{max}}\}$ 生成 R^{bs}，R^c 和 R^p。
19: 　$R^{re} = F_{\Theta^{re}}(R^{bs}, R^c, R^p)$。　// 为 RRF 生成组合相关性分数。
20: 　收集用户标注为 $d \in \{d_1, ..., d_{h_{max}}\}$ 作为 R^{gh}。
21: 　S_{RRF}.append($\Pi(R^{gh}, R^{re})$)。　// 计算 RRF 的基于排序的指标。
22: 　// 扩展 X_{global} 和 R_{global} 用于脑解码模型的训练。
23: 　X_{global}.append(X)，R_{global}.append(R^{gh})。
24: **end for**
25: $S_{\text{IRF}} = \text{Average}(S_{\text{IRF}}), S_{\text{RRF}} = \text{Average}(S_{\text{RRF}})$
26: **返回** $S_{\text{IRF}}, S_{\text{RRF}}$;

与者的脑数据（通用模型 G_g）进行训练。因为脑信号在不同个体间存在差异，为每个参与者训练个性化模型 G_p 是必要的，这一点已在之前的研究中得到证实[84,298]。然而，每个参与者采集的数据量在搜索过程开始时可能不足以训练 G_p。因此，本研究采用使用其他参与者数据训练的通用模型 G_g 作为替代，直到其收集的数据量达到最低要求的大小（设为 100 个数据样本）。

随后，分别如算法 5.2第 5-13 行和第 14-18 行所述，本研究开展 IRF 和 RRF 的实验。对于 IRF，从历史文档 $\mathcal{D}_h$ 中收集有价值的相关性反馈信号，利用相关性反馈方法（详见第 5.4.2节）生成候选文档 $\mathcal{D}_u$ 的相关性得分。鉴于候选文档可能没有用户标注，IRF 性能使用第三方标注 R^{gu} 进行评估，随着历史文档数量 h 的增加逐步进行。因此，对于查询 q，本研究会计算 h_{max} 个基于排序的指标 $\Pi(R^{gu}, R^{it})$。

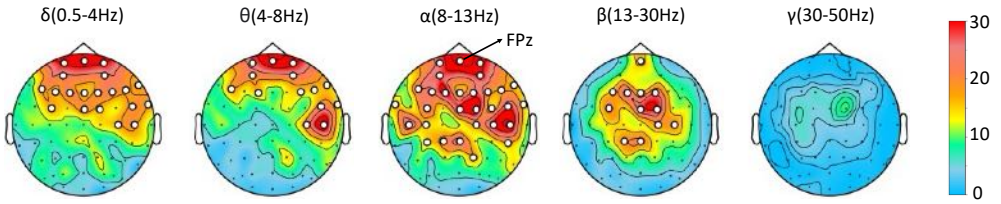


图 5.10　相关和不相关网页的脑反应差异显著性地形图（F 值）。突出显示的脑电通道表示差异在 $p < 1 \times 10^{-3}$ 水平上显著。

与 IRF 不同，RRF 的性能评估在当前搜索查询结束时进行。通过 $\Pi(R^{gh}, R^{re})$ 计算 $\mathcal{D}_{h_{max}}$ 的重排序性能，其中 R^{re} 是所提出方法计算的相关性得分，R^{gh} 表示用户的标注。

5.4.3　基于脑信号的相关性估计实验

本节介绍将用户浏览文档快照和访达页内容的脑反应解码为基于脑的相关性估计的实验。本节首先说明脑信号的特征提取、解码模型选择和评估方法；然后通过对提取特征的统计分析展示了脑信号在相关和不相关文档之间的差异；最后汇报解码实验的效果以评估基于脑信号的相关性估计性能。

特征提取： 和第 5.3节一致，本节采用 62 个通道和五个频段上的 DE 作为特征。因此，脑电数据样本被预处理为大小为 62×5 的输入向量。

解码模型选择： 与第 5.3节不同，本节采用高斯核的支持向量机（SVM）作为基础脑解码模型。一方面，SVM 在脑解码算法中也是相当有效的[118,299]。另一方面，与第 5.3节中使用的神经网络方法相比，该方法所需的计算量较少[84]，能够实现在现实系统中的在线训练和推理的需求，从而满足本节在第 5.3节基础上新增的交互式相关性反馈任务的要求。

评估： 本节评估脑解码模型 G（G_p 或 G_g）在相关和不相关文档的分类能力方面的表现。由于相关性标注 1 占 68.2% 的数据样本，相关性标注 1 被视为负样本（不相关），标注 2-4 为正样本（相关）。因此，分类问题被转化为二元分类问题，本节通过 AUC 来衡量其性能。这里，本节使用了同一个分类模型 G（同时作为 G_p 和 G_g）来对快照和访达页的脑反应进行分类，因为其性能要优于对快照和访达页的脑反应训练独立的模型。

5.4.3.1　特征分析

本节计算每个脑电通道在两种条件（相关和不相关文档）下的 DE 特征的平均值。图 5.10显示了 ANOVA 检验下 62 个脑电通道的 F 值，突出显示的通道表示差异在 $p < 1 \times 10^{-3}$ 水平上显著。本节观察到在 δ、θ、α 和 β 频段中都存在显著的通道。其中，最显著的差异出现在 α 频段的 FPz 通道（$F[1, 20] = 39.03, p = 4.2 \times 10^{-6}$, $M_{\text{diff}} = -0.40 \ln(Hz)$）。这一发现表明，大脑对相关和不相关信息内容的反应是有显著差异的。

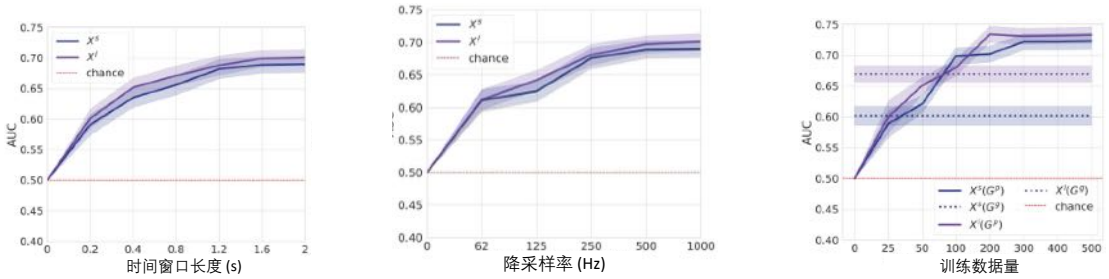
此外，本节观察到在 δ、θ 和 α 频段中，最显著的差异出现在额叶，这与上一章中的发现一致。另一方面，本节也观察到在中央脑区，尤其是在 β 和 α 频段中有显著的神经差异。Pinkosova et al.[69], Allegretti et al.[71], Yang et al.[300]在中央脑区也有类似的观察。尽管这些研究之间的设置不同（基于视觉[71]，文本[69]）以及本研究的多模态网页内容作为实验刺激），但中央脑区的共同发现表明该区域的脑功能与相关性判断之间可能存在联系。对此潜在联系的一个解释是相关性判断期间的记忆处理[69]，例如通过回忆相关知识来识别信息内容是否相关。

除了这些共性观察外，本节还有一些与先前研究不同的发现。例如，第 5.3节中的事实性搜索任务里最显著的差异在 β 频段观察到，而在本节中的非事实性搜索任务下，α 频段中的差异比在 β 频段中的更大。这种差异的一个可能原因是搜索任务的类型不同：在事实性问题中参与者的注意力水平（反映在 β 频段[301]）在找到直接答案（一个或两个词）时会有很大变化。而在非事实性的任务中，参与者需要通过更深入的理解来判断文档的相关性。因此，另一些神经现象，如效价和记忆等，可能在非事实性任务的搜索中起主要作用，这也在现有研究[302]中有所揭示。

5.4.3.2　脑解码性能

对访达页内容相关性估计的平均 AUC 性能为 0.701（标准差 =0.059），略好于快照内容（X^{snippet}）的 0.690 （标准差 =0.060）。此外，本节观察到个性化模型 G_p（AUC 在 X^{land} 和 X^{snippet} 上分别为 0.691 和 0.681）的总体表现显著优于通用模型 G_g（AUC=0.670（X^{land}）和 AUC=0.603（X^{snippet}））。这验证了 G_p 优于 G_g 的假设，即为每个参与者训练个性化的模型有助于分类性能。然而，当收集到的个体数据量不足（前 100 个样本）时，G_p 的性能（AUC=0.584）不如通用模型 G_g（AUC=0.627）。因此，本节采用了基于时序分割的策略，对任意一个被试者，先用通用模型 G_g 进行搜索结果相关性估计，再转换为个性化模型 G_p。这个策略更接近现实场景，模拟了新用户开始使用本系统时的冷启动情况。

注意到，由于脑电数据通常包含噪声，通过脑电信号解码相关性的性能并不



(a) 时间长度　　(b) 降采样率　　(c) 训练数据样本数量

图 5.11　在不同时间长度、降采样率和用于训练的个性化数据样本数量下相关性估计的性能。阴影区域表示标准误差。

完美[19,155]。然而，基于脑电的相关性预测相对于显式标注存在一定优势，即脑电信号的采集是实时的，且不会干扰用户的搜索过程。

5.4.3.3　敏感性分析

第 5.3.1.4节详细说明了数据预处理的方法，这与现有的神经科学研究[19,69]的常见设置是一致的。为了探索和理解这些设置如何影响相关性估计的性能，本节对数据预处理的方法进行了敏感性分析，特别是时间窗口的长度和降采样率，如图 5.6和图 5.11(b)所示。从图 5.11(a)可以看出，随着时间窗口持续增加，相关性估计性能有所提升，并在 1,600ms 到 2,000ms 之间趋于稳定。现有文献[19,71]报告说，用户对于相关和不相关文档的脑信号差异在 800ms 开始出现。实验发现如果能采集更长的时间窗口，该相关性估计的性能仍有改进空间。另一方面，本实验发现降采样的频率越高，相关性估计性能也越高，如图 5.11(b)所示。为了减少计算复杂度，特别是在涉及交互式相关性反馈的实时场景中，本实验中采用了一致的 500Hz 的采样率。这确保了与不进行降采样时相当的性能，并且在后续的模型构建中所需的计算资源更少。

另外，本节还进行了关于训练使用的个体数据样本数量的敏感性分析。此分析使用了不同规模的脑数据样本 $\boldsymbol{X}$ 训练脑解码模型 G^p，并使用额外的 100 个数据样本评估其性能。图 5.11(c)显示了以 AUC 为单位的分类性能。如图 5.11(c)所示，使用通用模型（G^g）也具有显著高于 0.5 的 AUC，这意味着我们可以在不使用用户的个人数据进行模型训练的情况下也实现一定程度上的相关性估计。此外，实验中训练的个性化模型 G^p 即使在只有 100 个数据样本时也能达到比较好的表现，这意味着在实际中，我们只需收集有限的用户数据来训练模型，就可以初始化这样的相关性反馈系统。

5.4.3.4　混合效应分析

在数据收集过程中，我们随机化了任务顺序和搜索文档的顺序，以尽量减少潜在混杂因素的风险。尽管采取了这些预防措施，但完全消除所有混杂变量的影响是不可能的。在此，我们讨论了可能影响我们分析观察的统计稳健性和有效性的各种混杂因素。根据现有文献[19]，我们考虑的混杂因素包括：个体差异（I）、任务排序（O^t）、文档排序（O^d）、文档中的单词数（W）和显示文档的图像大小（S）。我们采用线性混合模型来建模通过 EEG 频谱功率测量的脑活动（X，选取最显著的 FPz 通道上 α 波段功率）与文档的真实相关性（R^{gh}）之间的依赖关系，该模型形式化为：

$$X = (\beta_u + i_u)R^{gh} + \beta_w W + \beta_t O^t + \beta_d O^d + \beta_s S + I + \beta_0 + e, \tag{5.6}$$

其中 e 表示全局残差，β_0 表示全局截距，β_u、β_w、β_t、β_d、β_s 表示对应上述混杂因素效应的系数。I 是个体差异效应，i_u 是和参与者 u 关联的系数。脑活动 X 通过如第 5.4.3节中所述的频谱功率估计，文档相关性 R^{gh} 的取值范围从 1 到 4。

表 5.5　混合线性模型下脑电特征与不同变量的关联性分析

效应	相关系数	标准差	z 统计量	p>z
词数	0.000	0.000	1.247	0.106
文档位置	-0.007	0.004	-2.320	0.010
任务次序	-0.001	0.000	-1.542	0.062
快照面积	0.001	0.000	0.099	0.921
文档相关性	0.041	0.005	8.921	0.000

表 5.5展示了固定效应（即文档相关性）和随机效应（即词数、文档位置、任务次序和快照面积）在混合线性模型中的相关系数和显著性。从表 5.5中，我们观察到文档相关性对脑反应产生了显著影响。与混杂变量相比，这一效应尤为突出。在混杂因素中，只有文档排序对结果显示出显著影响，这可能和搜索引擎中常见的位置偏差效应[27,254]有关，即用户通常认为列表顶部的文档更相关。相反，其他混杂因素显示出最小且无统计显著性的影响。这表明，尽管有多种因素影响脑反应，但文档的相关性发挥了主要作用。

5.4.4　相关性反馈实验

5.4.4.1　实验设置

评估策略　在交互式相关性反馈任务中，本节以交互的方式评估候选文档 $\mathcal{D}_u = \{d_{h+1}, d_{h+2}, \cdots, d_n\}$ 的重排序性能，即随着 h 从 1 增加到 h_{max}，计算平均的 $\Pi(R^{gu}, R^{it})$。由于我们不要求参与者为候选文档 $\mathcal{D}_u$ 标注相关性，我们使用第三方标注作为真实相关性 R^{gu}。在回顾式相关性反馈中，我们计算历史文档 $\mathcal{D}_h = \{d_1, d_2, \cdots, d_h\}$ 的评价指标 $\Pi(R^{gh}, R^{re})$，其中 $h = h_{max}$，并基于用户的标注作为 R^{gh}。由于访达页包含的文档内容比快照内容多，如果文档被点击，则文档的真实相关性基于访达页的标注，否则基于文档快照的标注作为替代。对于基于排序的评估指标 Π，我们采用平均准确率（MAP）和在不同的截止点上（1、3、5 和 10）的归一化折扣累积增益（NDCG）[303]。最后，本节基于双边 T 检验来测量由不同方法和信号实现的重排序性能的差异的显著性。

基线　对于交互式相关性反馈，基线包括三种不使用任何用户信号的重排序策略：BM25[22]，BERT (R^p)[255]（根据 BERT 计算的相关性得分进行重排序，相当于基于 R^p 进行重排序），以及 Sogou（使用搜狗搜索引擎中的原始排序）。此外，本节报告了所提出的基于 BERT 的查询扩展方法 $\text{QE}^{F_{\Theta^{it}}(R^{bs}, R^c, R^p)}$ 及其消融版本（QE^{R^p}，$\text{QE}^{F_{\Theta^{it}}(R^c, R^p)}$ 和 $\text{QE}^{F_{\Theta^{it}}(R^{bs}, R^p)}$）的性能。除了提出的相关性反馈方法外，本节还报告了传统相关性反馈方法 RM3[129] 的性能。RM3 的实现基于 Lavrenko et al.[129] 使用的参数，并从 $\{3, 5, 10\}$ 中选择重写的词的数量。对于回顾式相关性反馈，基线包括三种不使用用户信号的重排序策略：BM25，BERT 和 Sogou。除了结合所有信号的相关性反馈框架（$F_{\Theta^{re}}(R^{bs}, R^c, R^p)$），本节还报告了其消融版本 $F_{\theta^{re}}(R^c, R^p)$（无脑信号）和 $F_{\Theta^{re}}(R^{bs}, R^p)$（无点击信号）的性能。

融合参数的选择　每个融合参数 $\theta^{*,\dagger}$（$* \in \{it, re\}$，$\dagger \in \{bs, c, p\}$）从 $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ 中选择。本节的实验首先探索了固定的 Θ^{it} 和 Θ^{re}。为了初始化该固定的参数，本节随机抽取一个包含 200 个搜索任务的子集，并测试基于 NDCG@10 的文档重排序性能。然后，在这个子集上选择所有组合中最优的参数组合，即在交互式相关性反馈中的 $\Theta^{it,bs:c:p} = 3:1:1$，在回顾式相关性反馈中的 $\Theta^{re,bs:c:p} = 5:2:0$。实验观察到，与其他参数相比，所选参数在整个数据集中也能获得最佳性能，这表明所选参数是比较鲁棒的。

由于在选择回顾式相关性反馈参数时 $\theta^{re,p}$ 为 0，本研究进一步在 0-0.2 范围内对其进行搜索。实验观察到 $F_{\Theta^{re}}(R^{bs}, R^c, R^p)$ 的性能在 0-0.14 范围内没有显著差异

（在 0.06 时达到峰值）。尽管反馈参数 $\theta^{re,p}$ 的最佳取值非常小，但不代表伪相关性信号就完全没有作用，实际上，BERT(R^p) 的性能依然是显著高于随机的排序。这表明伪相关性信号对回顾式相关性反馈是有效的，但与用户信号（R^c 和 R^{bs}）相比，其收益就非常小了。这是由于本实验中的文档是从向搜狗搜索引擎提交查询时的顶部文档中选择的。因此，它们中的大多数已经与查询项在语义上相关。在这种情况下，用户的真实反馈比语义相关性上的细微差异更重要。

5.4.4.2　总体结果

交互式相关性反馈性能　表 5.6展示了交互式相关性反馈中的文档重排序性能。表 5.6中揭示了以下发现：（1）使用用户信号的方法（即 $\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^p)}$，$\mathrm{QE}^{F_{\Theta^{it}}(R^c,R^p)}$ 和 $\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$）优于不考虑用户信号的方法（即 BM25，BERT(R^p)，Sogou 和 QE^{R^p}）。这表明从用户交互中提取的信号有助于改善文档重排序性能。（2）结合脑信号预测的相关性得分 R^{bs} 的查询扩展方法获得了显著的性能提升。$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^p)}$（$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$）与 QE^{R^p}（$\mathrm{QE}^{F_{\Theta^{it}}(R^c,R^p)}$）的性能（基于 NDCG@10）之间的差异为 5.3%（1.5%），在配对 T 检验中在 $p = 7.8 \times 10^{-16}$（1.5×10^{-6}）的水平上显著。这表明脑信号可以为现有信号（伪相关信号或伪相关信号和点击信号的组合）提供额外的信息。（3）提出的基于 BERT 的查询扩展方法比传统方法 RM3 更有效。这可能表明 BERT 在捕捉文档的语义表示方面优于统计语言模型。

表 5.6　在交互式相关性反馈中的文档重排序性能[a]

Method[b]	**NDCG@1**	**NDCG@3**	**NDCG@5**	**NDCG@10**	**MAP**
BM25	0.2031*	0.2265*	0.2449*	0.3031*	0.3048*
Sogou	0.2085*	0.2284*	0.2481*	0.3065*	0.3214*
BERT（R^p）	0.2060*	0.2326*	0.2579*	0.3221*	0.3033*
$QE^{F_{\Theta^{it}}(R^p)}$	0.2306*	0.2361*	0.2582*	0.3205*	0.3279*
$QE^{F_{\Theta^{it}}(R^{bs},R^p)}$	0.2477*	0.2569*	0.2774*	0.3374*	0.3235*
$QE^{F_{\Theta^{it}}(R^c,R^p)}$	0.2842*	0.2952*	0.3124*	0.3690*	0.3708*
$RM3^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$	0.2332*	0.2523*	0.2717*	0.3289*	0.3337*
$QE^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$	**0.2948**	**0.3024**	**0.3191**	**0.3747**	**0.3744**

[a] $*$ 表示与 $QE^{F_{\Theta^{it}}(R^{bs},R^c,R^p)}$ 相比性能存在显著差异，显著水平为 $p < 1 \times 10^{-3}$。

[b] $R^{\dagger}$ 表示基于信号 † 的相关性估计。*bs*、*c* 和 *p* 分别表示脑信号、点击信号和伪相关性信号。

回顾式相关性反馈性能　表 5.7展示了回顾式相关性反馈中的文档重排序性能。本节观察到和在交互式相关性反馈中类似的发现。首先，使用用户信号的模型（即 $F_{\Theta^{re}}(R^{bs}, R^{p})$，$F_{\Theta^{re}}(R^{c}, R^{p})$ 和 $F_{\theta^{re}}(R^{bs}, R^{c}, R^{p})$）通常比不使用用户信号的模型（即 BM25，BERT 和 Sogou）表现更好。其次，脑信号可以提升相关性反馈性能，例如，$F_{\theta^{re}}(R^{bs}, R^{c}, R^{p})$ 在 NDCG@10 上相比 $F_{\theta^{re}}(R^{c}, R^{p})$ 提高了 7.4%。

除了相似之处，本节还注意到交互式相关性反馈和回顾式相关性反馈之间的差异。例如，本节观察到通过添加用户信号（即脑信号或点击信号或它们的组合）实现的性能提升在回顾式相关性反馈中比在交互式相关性反馈中更大。例如在回顾式相关性反馈中，$F_{\theta^{re}}(R^{bs}, R^{c}, R^{p})$ 与 $F^{re}_{\Theta}(R^{c}, R^{p})$（$F_{\theta^{re}}(R^{bs}, R^{c}, R^{p})$ 与 R^{p}）之间的性能提升在 NDCG@10 上为 7.4%（37.3%）。然而，在交互式相关性反馈中，$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 与 $\mathrm{QE}^{F_{\Theta^{it}}(R^{c},R^{p})}$（$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 与 $\mathrm{QE}^{R^{p}}$）之间的性能提升较小，基于 NDCG@10 为 1.5%（16.9%）。一个可能的原因是排序候选文档比排序历史文档更困难。有了准确的相关性反馈信号，本方法可以对历史文档的相关性做更好的估计。但实际并不能保证候选的相关文档与用户观察到的历史相关文档一定相似，从而对候选文档也做出比较好的估计。

案例研究　表 5.8展示了参与者（序号 1）和查询 “先知” 的交互式相关性反馈和回顾式相关性反馈结果示例，任务描述为 “探索一般领域的先知概念”。在这种情况下，参与者浏览了从 d1 到 d6 的历史文档。对于交互式相关性反馈任务，目标是对未见的文档（D^{u}）d7 到 d12 进行重排序，而回顾式相关性反馈任务则专注于对历史文档（D^{h}）从 d1 到 d6 进行重排序。表 5.8(a)展示了这些文档的标题。根据任务描述，文档 d4、d6 和 d11 具有高度相关性（相关性标注为 4）。另一方面，文档 d2

表 5.7　在回顾式相关性反馈中的文档重排序性能[a]

Method[b]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
BM25	0.3113*	0.3831*	0.4703*	0.5587*	0.5477*
BERT（R^{p}）	0.3218*	0.3846*	0.4626*	0.5605*	0.5534*
Sogou	0.3313*	0.4046*	0.4940*	0.5848*	0.5633*
$F_{\Theta^{re}}(R^{bs}, R^{p})$	0.5074*	0.5494*	0.6254*	0.6973*	0.6744*
$F_{\Theta^{re}}(R^{c}, R^{p})$	0.5426*	0.5936*	0.6578*	0.7161*	0.6694*
$F_{\Theta^{re}}(R^{bs}, R^{c}, R^{p})$	**0.6350**	**0.6617**	**0.7171**	**0.7693**	**0.8009**

[a] $*$ 表示与 $F_{\Theta^{re}}(R^{bs}, R^{c}, R^{p})$ 相比性能存在显著差异，显著水平为 $p < 1 \times 10^{-3}$。

[b] $R^{\dagger}$ 表示基于信号 † 的相关性估计。bs、c 和 p 分别表示脑信号、点击信号和伪相关性信号。

表 5.8　一个交互式相关性反馈和回顾式相关性反馈的示例

(a) 历史文档（D^h）和候选文档（D^u）的标题

查询：先知	
历史文档 (D^h)	候选文档 (D^u)
d1: 先知 - 法国，暴力，犯罪	d7: 诗歌“先知”的分析和翻译
d2: 先知默罕默德的奇迹	d8: 圣经里先知的含义是什么？-搜狗问问
d3: 纪伯伦的诗《先知》	d9:《先知》电影：免费在线资源
d4: 先知 - 搜狗百科	d10: 先知: 纪伯伦
d5: 第五人格 - 先知角色的介绍	d11: 不同文化下的先知
d6: 先知的一般概念	d12: 先知 - 中国词典

(b) 历史文档 D^h 的估计基础相关性分数，R^c、R^p 和 R^{bs} 分别代表基于点击、伪相关性和基于脑信号的相关性分数

D^h	R^c	R^p	R^{bs}
d1	0	0.6	0.3
d2	0	0.3	0.6
d3	0	0.4	0.3
d4	1	0.4	0.7
d5	0	0.3	0.2
d6	1	0.5	0.6

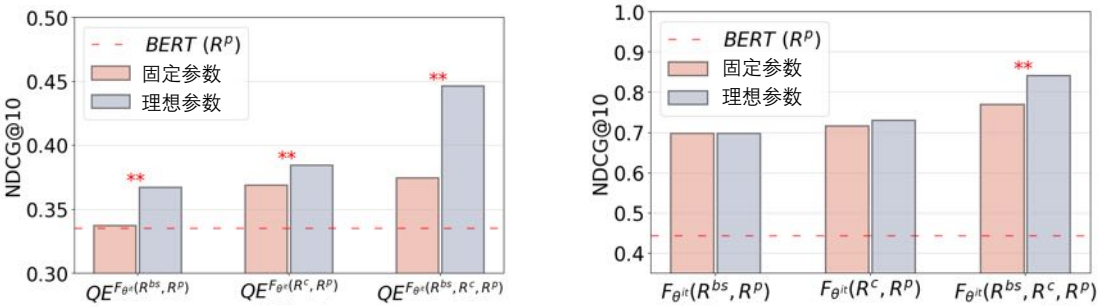
(c) 在有无脑信号的相关性反馈模型中（w R^{bs} 和 w/o R^{bs}），交互式相关性反馈和回顾式相关性反馈任务的重排序文档列表

IRF	w R^{bs}	d11,d12,d8,d7,d9,d10
	w/o R^{bs}	d12,d11,d8,d7, d9,d10
RRF	w R^{bs}	d4,d6,d2,d1,d3,d5
	w/o R^{bs}	d6,d4,d1,d3,d2,d5

注：该示例来自于参与者 ID 1 和查询 “先知”，任务描述为 “探索一般领域的先知概念”。
相关性标注为 4 和 2-3 的文档分别用紫色和浅紫色突出显示。标注为 1 的文档用黑色表示。

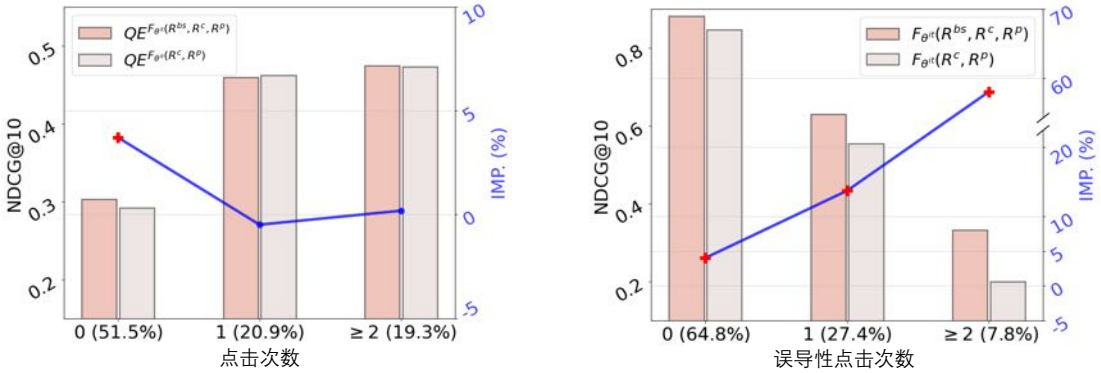
和 d8 部分相关，因为它们分别与伊斯兰教和基督教的先知有关。两者都是 “先知” 在更广泛概念下的文档。文档 d12 也部分相关，因为 “先知” 的字典解释可能提供一些有用的信息。

表 5.8(b)展示了对历史文档 D^h 估计的相关性分数。从表 5.8(b)可以观察到，R^p 与真实相关性不太一致，因为它仅基于语义上和查询词 “先知” 的匹配程度。另一方面，R^c 和 R^{bs} 与真实相关性更加一致。其中，R^{bs} 可以为一些特殊情况提供额外的信息，例如在两个文档均未点击或已点击时，判断两个文档相对的相关性。如表 5.8(c)所示，在回顾式相关性反馈任务中，d1，d2，d3 和 d5 是无点击交互的文档。但利用了 R^{bs} 提供的信息实现的相关性反馈技术能够准确地将 d2 排在它们中的顶部。此外，用户的脑信号反馈提供的信息能表明用户对已点击文档 d4 和 d6 不同程度上的满意度。因此，在交互式相关性反馈任务中，包含 R^{bs} 的相关性反馈将 d11 排在 d12 之前，因为 d11 在语义上与最满意的历史文档（即 d6）更接近。



(a) 交互式相关性反馈　　(b) 回顾式相关性反馈

图 5.12　固定和理想融合参数 Θ 下的相关性反馈性能。** 表示在 $p < 0.01$ 水平下，使用固定和理想融合参数的相关性反馈方法之间的显著差异，采用配对 T 检验。



(a) 交互式相关性反馈　　(b) 回顾式相关性反馈

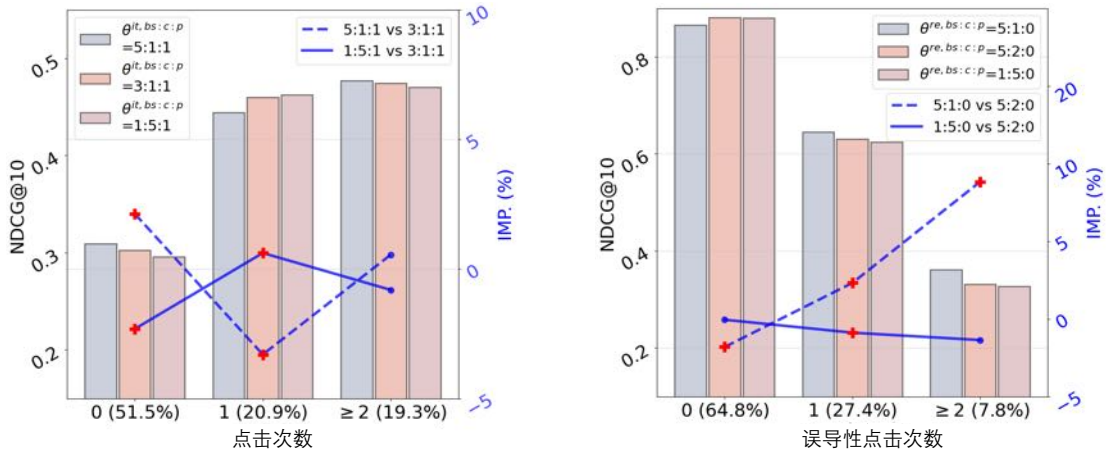
图 5.13　不同搜索场景下有无脑信号的相关性反馈性能。+表示在配对 T 检验中，模型的提升（IMP.）显著（$p < 0.05$）。

5.4.5　信号融合的权重分析

上述实验使用了固定的权重 Θ^{it} 和 Θ^{re}。然而，在不同搜索场景下，每个数据样本对于相关性反馈信号的依赖程度可能是不一样的。因此，本节进行了一项理想实验，自适应地选择理想的融合参数 Θ^{it} 和 Θ^{re} 以实现最佳的文档重排序性能。对于每个数据样本，在算法 5.2中第 10 行的 Θ^{it} 和第 16 行的 Θ^{re}）中，本节从集合 $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ 中搜索 $\theta^{*,\dagger}$ 的值，其中 $* \in \{it, re\}$，$\dagger \in \{bs, c, p\}$。随后，本实验根据每个不同数据样本的 NDCG@10 确定参数 Θ^{it} 和 Θ^{re} 的最佳组合。注意，理想实验的性能实际上泄露了每个样本对于不同相关性反馈信号的依赖程度，在现实中是无法实现的。该实验旨在比较实际实验结果与最理想结果之间的差异，从而探索相关性反馈框架中自适应融合权重的潜力。

图 5.12展示了固定和理想融合参数 Θ 下的相关性反馈性能。从图 5.12中观察到，在交互式相关性反馈和回顾式相关性反馈任务中的所有方法中，基于理想融合参数的相关性反馈优于具有固定融合参数的相关性反馈。此外，基于固定/理想融合参数的相关性反馈在交互式相关性反馈中的性能差异大于在回顾式相关性反

馈中的差异。特别是，在交互式相关性反馈中，使用 $\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 的方法，性能提升达到 19.1%（NDCG@10 从 0.3747 提升到 0.4463），在 $p < 1 \times 10^{-2}$ 的水平上显著。这表明合理地融合脑信号、点击信号和伪相关信号对交互式相关性反馈有着比回顾式相关性反馈更大的潜力。



(a) 交互式相关性反馈　　(b) 回顾式相关性反馈

图 5.14　在不同搜索场景和不同融合参数 Θ 下的相关性反馈性能。+表示改进（IMP.）在 $p < 0.05$ 的水平上显著。

5.4.6　搜索场景分析

本节进一步探讨了点击信号缺失或存在偏差的搜索场景中的相关性反馈性能。实验观察到，在点击缺失时，脑信号通过估计无点击结果的相关性发挥了特别重要的作用（在交互式相关性反馈中分析，因为在任何点击发生之前启动相关性反馈在交互过程中是重要的）。此外，脑信号在发生误导性点击的情况下（在回顾式相关性反馈中分析，以便我们可以为未来的潜在搜索过程重排序可能导致误导性点击的结果）对于相关性反馈也有比较大的帮助。

无点击场景　在第 5.3节中提到，不是所有无点击的结果都是不相关的。例如，一些无点击的结果可能包含有用的信息，甚至吸引用户回访的行为[304]。本节的用户研究也观察到了无点击文档被标注了不同的相关性分数（1（77.6%）、2（7.4%）、3（4.6%）和 4（10.2%））。基于以上观察，本节旨在探索脑信号是否可以估计无点击结果的相关性，并在用户的点击行为发生之前就启动交互式相关性反馈来提升文档排序性能。

图 5.13(a)展示了在不同点击次数的搜索场景中的交互式相关性反馈性能。从图 5.13(a)中可以观察到，$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 与其消融版本 $\mathrm{QE}^{F_{\Theta^{it}}(R^{c},R^{p})}$ 之间的性能提升在无点击搜索场景中最大（在 NDCG@10 方面提高了 3.7%）。相反，它们的性能

差异在具有不小于一次点击的搜索场景中并不显著。由于交互式相关性反馈是一个交互的过程，随着历史文档数量的增加，点击次数也会增加（Pearson’s $r = 0.45$（$p < 1 \times 10^{-3}$））。因此，本节也观察到在搜索过程的开始阶段，$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 和 $\mathrm{QE}^{F_{\Theta^{it}}(R^{c},R^{p})}$ 之间的性能差异更大（在 NDCG@10 方面为 3.3% 的提升，搜索会话长度 $h \leqslant 4$），而在随后的搜索过程中（搜索会话长度 $h > 4$）则仅有 0.6% 的提升。这表明脑信号在搜索初期发挥着重要作用，而搜索过程的开始阶段正是提升用户搜索体验的关键时刻。

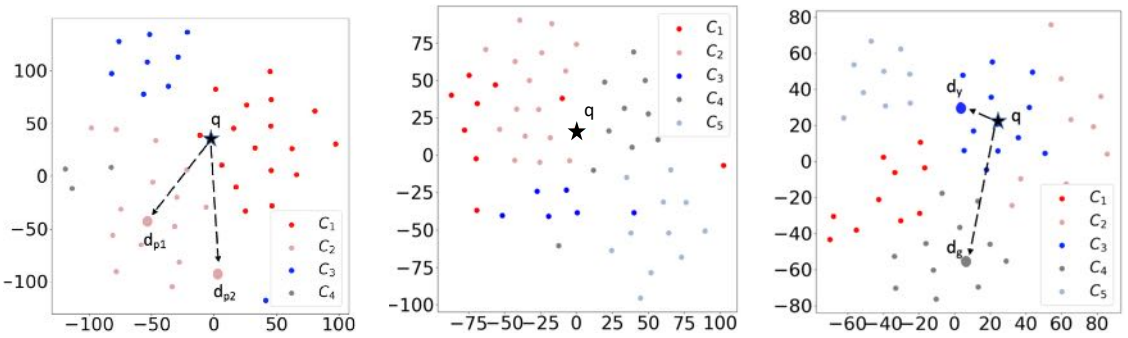
此外，本节分析了交互式相关性反馈性能如何随融合参数 Θ^{it} 的选择而变化。我们从 {3:1:1, 5:1:1, 1:5:1} 中选择 $\Theta^{it,bs:c:p}$，其中 3:1:1 是所有数据样本中的平均最佳参数，而 5:1:1 和 1:1:5 是分别强调脑信号和点击信号重要性的融合参数。如图 5.14(a)所示，在无点击搜索场景中，结合更高权重的脑信号（$\Theta^{it,bs:c:p} = 5{:}1{:}1$）显著优于其他融合参数 $\Theta^{it,bs:c:p} = 3{:}1{:}1$ 和 $\Theta^{it,bs:c:p} = 1{:}5{:}1$），在配对 T 检验中分别基于 $p < 2.6 \times 10^{-10}$ 和 $p < 6.3 \times 10^{-17}$ 显著。此外，如果我们简单地在无点击场景中将参数 Θ^{it} 设置为 5:1:1，而在其他场景中设置为 3:1:1，则交互式相关性反馈性能在 NDCG@10 方面可以达到 0.3781，显著优于使用固定融合参数所获得的性能（$p = 2.7 \times 10^{-10}$）。这揭示了根据搜索上下文自适应地融合不同相关性反馈信号的潜在收益。

误导性点击场景　误导性点击表示用户点击的文档不相关，可能导致不佳的搜索体验。这通常发生在文档的快照很吸引人，但其访达页内容不令人满意的时候[5]。在本实验中，如果将点击的文档访达页的相关性标注为 1（“完全不相关”）或 2（“不相关”）定义为“误导性点击”，那么有 21.8% 的点击属于“误导性点击”。

如图 5.13(b)所示，实验观察到随着误导性点击数量的增加，$F_{\Theta^{re}}(R^{bs}, R^{c}, R^{p})$ 和 $F_{\Theta^{re}}(R^{c}, R^{p})$ 之间的性能差异更大。这表明脑信号在误导性点击频繁发生的场景中可以带来更多的增益。此外，我们探索了融合参数 Θ^{re} 的不同取值，包括平均表现最佳的取值（$\Theta^{re,bs:c:p} = 5{:}2{:}0$）、强调脑信号的取值（$\Theta^{re,bs:c:p} = 5{:}1{:}0$）和强调点击信号的取值（$\Theta^{re,bs:c:p} = 1{:}5{:}0$）。从图 5.14(b)中，实验观察到当存在误导性点击时，融合参数 $\Theta^{re,bs:c:p} = 5{:}1{:}0$ 明显优于使用平均最佳参数 $\Theta^{re,bs:c:p} = 5{:}2{:}0$。这强调了在可能存在负面或不当点击的情况下优先考虑脑信号的必要性。如果我们简单地在至少发生一次误导性点击的搜索场景中采用参数 $\Theta^{re,bs:c:p} = 5{:}1{:}0$，而在其他场景中使用 5:2:0，则回顾式相关性反馈性能在 NDCG@10 方面可以达到 0.7756（显著优于使用固定的参数，配对 T 检验的 $p = 1.0 \times 10^{-3}$）。尽管误导性点击的数量在实际中是无法知道的，这一观察仍揭示了更好地将脑信号结合到相关性反馈任务中的可能性。

5.4.7　相关性反馈信号的自适应融合探究

在第 5.4.5节中展示了不同相关性反馈信号在各种搜索场景中的重要性有所不同。这表明动态地进行相关性反馈信号融合存在一定的潜力，尤其是在交互式相关性反馈任务中。基于这些观察，本节提出并实验了一种在不同搜索场景下自适应地计算融合权重的相关性反馈信号融合方法。



(a) 查询 “先知”.　(b) 查询 “茶”.　(c) 查询 “婚戒”.

图 5.15　查询和其对应的不同意图下的文档（用不同颜色表示）的 BERT 嵌入的 T-SNE 图

5.4.7.1　建模动机

如第 5.4.5节中所讨论的，在交互式相关性反馈中设置自适应的融合参数比在回顾式相关性反馈中更为重要。因此，本节着重于在交互式相关性反馈中应用自适应相关性反馈信号融合方法。在非事实性搜索任务中，提交给搜索引擎的查询通常是比较宽泛的，可能与不同的子主题相关。例如，如图 5.15(a)所示，与查询“先知”对应的文档被分类为四个簇 C_1、C_2、C_3 和 C_4，每个簇代表一个不同的子主题。假设用户的搜索意图与簇 C_2 中的文档相关，并对文档 d_{p1} 显示出较高的相关性反馈评分。在这种情况下，交互式相关性反馈将使原始查询表示更接近 d_{p1}。因此，簇 C_2 中的其他文档（例如 d_{p2}）将获得比其他簇中的文档更高的相关性估计。

本节将可能出现的一个搜索场景表示为 $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$，其中用户在查询 q 下已查看了 h 个文档，用户的点击次数为 n_h（$n_h \in \{1, 2, ..., h\}$）。融合参数 $\Theta^{it,S_c} = \{\theta^{it,bs,S_c}, \theta^{it,c,S_c}, \theta^{it,p,S_c}\}$ 应为不同的搜索场景 S_c 自适应地选择，具体原因如下：首先，为了有效利用 $\mathcal{D}_h$ 中的文档作为反馈信号，应考虑这些文档在其各自簇中的代表性。例如，如图 5.15(a)所示，文档 d_{p1} 和文档 d_{p2} 都属于相同的文档簇 C_2。然而，d_{p1} 比 d_{p2} 更具代表性，因为它更接近于 C_2 中文档的平均表示。因此，d_{p1} 的反馈信息比 d_{p2} 更有价值，因为它在 C_2 中的文档中更具代表性。为了提升相关性反馈性能，当 d_{p1}（d_{p2}）属于反馈文档 $\mathcal{D}_h$ 时，Θ^{it,S_c} 应被赋予更高（更低）的权重。其次，文档的点击概率不仅受其相关性影响，还受其他因素影响，例如一些

算法 5.3 自适应相关性反馈信号组合

输入: 一个搜索场景 $Sc = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$，其中 n_h 从 $\{1, 2, \cdots, h\}$ 中选择；文档聚类 $\mathcal{D} = \{C_1, C_2, \cdots, C_{qm}\}$。模拟次数 N。
初始化: 所有候选融合参数 $\Theta^{it} = \{\theta^{it,bs}, \theta^{it,c}, \theta^{it,p}\}$，其中 $\theta^{it,bs}$、$\theta^{it,c}$、$\theta^{it,p}$ 从 $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ 中选择；搜索场景 Sc 的最佳融合参数 $\Theta^{it,Sc}$，初始化为 $\{0, 0, 0\}$；最佳模拟性能 $\hat{\Pi} = 0$。
输出： 自适应相关性反馈融合参数 $\Theta^{it,Sc}$。
for 每个 Θ^{it} **do**
    性能总和 $\Pi = 0$;
    for 每个 $C_j \in \{C_1, C_2, \cdots, C_{qm}\}$ **do**
        在假设 C_j 中的文档为相关时，模拟 N 个可能的基于点击和基于脑信号的相关性得分 $\{R^{c,Sc}, R^{bs,Sc}\}$。
        计算使用模拟的相关性得分 $\{R^{c,Sc}, R^{bs,Sc}\}$ 和候选融合参数 Θ^{it} 的平均 RF 性能 Π_{C_j}。
        $\Pi = \Pi + \Pi_{C_j}$;
    end for
    if $\Pi > \hat{\Pi}$ **then**
        $\Theta^{it,Sc} = \Theta^{it}$;
        $\hat{\Pi} = \Pi$;
    end if
end for
返回: $\Theta^{it,Sc}$;

不相关的文档可能会因“点击诱饵”问题而吸引用户点击。因此，点击信号是否可靠和我们应如何平衡点击信号和脑信号的融合权重有关。最后，$\mathcal{D}_h$ 中不同簇下的文档与原始查询 q 的距离也是需要考虑的因素。例如，如图 5.15(c)所示，d_y 的语义表征比 d_g 的语义表征更接近于原始查询 q。因此，应根据反馈文档 $\mathcal{D}_h$ 中的文档与原始查询 q 的在语义向量表征上的差异来设置 θ^{it,p,S_c} 的权重。

5.4.7.2　总体流程

如算法 5.3所示，自适应相关性反馈信号融合方法为可能的搜索场景 $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$ 生成自适应融合参数 Θ^{it,S_c}。该方法包括以下过程：首先，将查询 q 对应的文档 $\mathcal{D}$ 聚类为 q_m 个簇 $\{C_1, C_2, ..., C_{q_m}\}$。其次，对于每个文档簇，，假设此簇内的文档与用户的搜索意图相关，并在此假设下模拟用户可能的基于点击和基于脑信号的相关性得分 $\{R^{c,S_c}, R^{bs,S_c}\}$。最后，考虑在所有的 q_m 个文档簇 C_j（$j \in \{1, \cdots, q_m\}$）下融合用户信号 $\{R^{c,S_c}, R^{bs,S_c}\}$ 时，平均相关性反馈性能最好的融合参数 Θ^{it,S_c}。通过上述过程，该方法可以计算在搜索过程中能结合实际用户信号的自适应相关性反馈融合参数 Θ^{it,S_c}。

5.4.7.3　准备工作

本节按照 Liu et al.[293]的方法将查询 q 对应的文档 D 聚类为 q_m 个簇 $\{C_1, C_2, ..., C_{q_m}\}$。Liu et al.[293]假设一个查询可能包含多个子主题，并采用人工标注的方法将每个文档分类到其中一个子主题中。

5.4.7.4　基于点击和基于脑信号的相关性得分融合

对于一个搜索场景 $S_c = \{q, D_h, D_u, n_h\}$，本方法将融合基于点击和基于脑信号的相关性得分 R^{c,S_c} 和 R^{bs,S_c}。首先，假设查询 q 与 q_m 个搜索意图相关，其中第 j 个搜索意图与第 j 个簇 C_j 有关（$j \in \{1, \cdots, q_m\}$）。然后，对每个簇 C_j 进行迭代，并在每次迭代中模拟 N（设为 20）组用户行为。该模拟将为每个场景 Sc 生成基于点击的相关性得分 $R^{c,Sc}$ 和基于脑信号的相关性得分 $R^{bs,Sc}$。注意，本方法简单地假设每个查询对应的所有搜索意图具有均匀的可能性。因此，对于每个文档簇 C_j，模拟次数被设为相同的数值 N。

当对第 j 个簇 C_j 进行迭代时，第 i 个文档的基于点击的相关性得分，记作 $r_i^{c,Sc}$，在其总和为 n_h 的约束下，按照伯努利分布进行模拟：

$$\forall i, r_i^c, Sc \sim \begin{cases} \text{Bernoulli}(p_{c,\text{rel}}), & \text{当} \quad d_i \in C_j \\ \text{Bernoulli}(p_{c,\text{irel}}), & \text{当} \quad d_i \notin C_j \end{cases} \quad | \sum_i^h r_i^{c,sc} = n_h \tag{5.7}$$

其中，$p_{c,\text{rel}}$（$p_{c,\text{irel}}$）是通过用户研究中的相关（不相关）文档中脑信号相关性得分的分布用区间估计方法推断出的参数，其中 $p_{c,\text{rel}}$（$p_{c,\text{irel}}$）表示相关（不相关）文档被点击的可能性。该模拟过程确保每个 $r_i^{c,Sc}$ 遵循伯努利分布，并且模拟的总点击次数为 n_h。

另一方面，第 i 个文档 d_i 的基于脑信号的相关性得分 r_i^{bs} 是按照正态分布模拟的：

$$r_i^{bs} \sim \begin{cases} \text{Normal}(\mu_{bs,\text{rel}}, \sigma_{bs,\text{rel}}), & \text{if} \quad d_i \in C_j \\ \text{Normal}(\mu_{bs,\text{irel}}, \sigma_{bs,\text{irel}}), & \text{if} \quad d_i \notin C_j \end{cases} \tag{5.8}$$

其中，$\mu_{bs,\text{rel}}$ 和 $\sigma_{bs,\text{rel}}$（$\mu_{bs,\text{rel}}$ 和 $\sigma_{bs,\text{irel}}$）是通过用户研究中的相关（不相关）文档中脑信号相关性得分的分布以区间估计推断出的参数。

5.4.7.5　最优融合参数搜索

模拟过程结束后，对于搜索场景 $S_c = \{q, \mathcal{D}_h, \mathcal{D}_u, n_h\}$，我们为每个文档簇 C_j（$j \in \{1, \cdots, q_m\}$）模拟 N 个对应的 R^c 和 R^{bs}，然后分别搜索最优融合参数 $\Theta^{it,Sc}$ 和 $\Theta^{re,Sc}$。对于所有参数，即 $\theta^{it,bs,Sc}$、$\theta^{it,c,Sc}$ 和 $\theta^{it,p,Sc}$，都从 $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ 中选择。接下来，计算排名评估指标 $\Pi(R^{gu}, R^{it})$，其中 R^{gu} 通过假设在簇 C_j（$j \in \{1, \cdots, q_m\}$）内的文档为相关来进行评估：

$$r_i^{gu} = \begin{cases} 1, & \text{当} \quad d_i \in C_j \\ 0, & \text{当} \quad d_i \notin C_j \end{cases} \tag{5.9}$$

然后，选择在所有文档簇 C_j（$j \in \{1, \cdots, q_m\}$）中，基于所有 $N \cdot q_m$ 个模拟的用户信号 R^c 和 R^{bs} 的平均相关性反馈性能最佳的融合参数（在我们的实验中，Π 设置为 NDCG@10）。由于搜索引擎的用户在提交查询时对哪个子主题感兴趣是无法预先知道的，因此，该方法提出的平均最佳性能的最优融合参数是通过对假设所有子主题均等地和该查询词相关而得到的。

表 5.9　固定和自适应的融合参数在交互式相关性反馈中的文档重排序性能[a]

Method[b]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
$QE^{F_{\Theta^{it}}(R^c, R^p)}$	0.2842*	0.2952*	0.3124*	0.3690*	0.3708*
$QE^{F_{\Theta^{it}}(R^{bs}, R^c, R^p)}$	0.3056*	0.3124*	0.3285*	0.3845*	0.3826*
$QE^{F_{\Theta^{it,Sc}}(R^c, R^p)}$	0.2948*	0.3024*	0.3191*	0.3747*	0.3744*
$QE^{F_{\Theta^{it,Sc}}(R^{bs}, R^c, R^p)}$	**0.3126**	**0.3258**	**0.3505**	**0.4183**	**0.4061**

[a] $*$ 表示性能与 $QE^{F_{\Theta^{it,Sc}}(R^{bs}, R^c, R^p)}$ 存在显著差异，显著水平为 $p < 1 \times 10^{-3}$。

[b] $R^{\dagger}$ 表示基于信号 † 的相关性估计。bs、c 和 p 分别表示脑信号、点击信号和伪相关性信号。Θ^{it} 和 $\Theta^{it,Sc}$ 分别表示固定和自适应融合参数。

表 5.10　固定和自适应的融合参数在回顾式相关性反馈中的文档重排序性能[a]

Method[b]	NDCG@1	NDCG@3	NDCG@5	NDCG@10	MAP
$QE^{F_{\Theta^{it}}(R^{bs}, R^c, R^p)}$	**0.6350**	0.6617	**0.7171**	0.7693	**0.8009**
$QE^{F_{\Theta^{it,Sc}}(R^{bs}, R^c, R^p)}$	**0.6350**	**0.6622**	0.7170	**0.7694**	0.8004

[a] 使用自适应融合参数 $\Theta^{re,Sc}$ 代替固定融合参数 Θ^{re} 时未产生任何显著差异。

[b] $R^{\dagger}$ 表示基于信号 † 的相关性估计。bs、c 和 p 分别表示脑信号、点击信号和伪相关性信号。Θ^{it} 和 $\Theta^{it,Sc}$ 分别表示固定和自适应融合参数。

5.4.7.6　实验结果

表 5.9展示了使用自适应相关性反馈信号融合方法和固定参数融合方法的交互式相关性反馈实验结果。从表 5.9中，可以观察到使用自适应的信号融合方法显著提高了交互式相关性反馈的性能。这个发现验证了第 5.4.5节中的分析。具体来说，使用固定融合参数时，脑信号带来的额外提升在 NDCG@10 方面为 1.5%（$\mathrm{QE}^{F_{\Theta^{it}}(R^{bs},R^{c},R^{p})}$ 与 $\mathrm{QE}^{F_{\Theta^{it}}(R^{c},R^{p})}$ 相比）。另一方面，$\mathrm{QE}^{F_{\Theta}^{it,Sc}(R^{bs},R^{c},R^{p})}$ 和 $\mathrm{QE}^{F_{\Theta}^{it,Sc}(R^{c},R^{p})}$ 的性能差异在 NDCG@10 方面为 8.8%。这表明自适应相关性反馈信号融合方法能够进一步利用脑信号带来的信息增益。

对于回顾式相关性反馈，采用相同的自适应相关性反馈信号融合方法没有导致显著的性能差异，如表 5.10所示。例如，$F_{\Theta}^{it,Sc}(R^{bs}, R^{c}, R^{p})$ 和 $F_{\Theta}^{it}(R^{bs}, R^{c}, R^{p})$ 在 NDCG@10 方面的性能差异并不显著，其中 $\Theta^{it,Sc}$ 是回顾式相关性反馈的自适应融合参数。这可能是由于自适应相关性反馈信号融合方法在回顾式相关性反馈中的潜力有限（见第 5.4.5节）。

5.5　本章小结

本章分别在“零点击”的事实性问题的场景和复杂的非事实性问题的场景下探索了脑信号在信息检索系统中作为一种反馈信号的潜力。在“零点击”的事实性问题的场景下，我们验证了脑信号可以解决传统信号稀缺的问题，实现该场景下更准确的用户状态理解和更合理的搜索结果相关性估计。实验发现脑信号能使搜索结果重排序在 NDCG@1 上提升 40% 以上，这一显著提升在不改变传统搜索范式的情况下仅凭借搜索算法能力的提高是难以实现的。而在复杂的非事实性问题的场景下，用户的信息需求很难被查询词所精准表达，因此反馈信号对于信息系统理解用户和提供优质内容是非常重要的。在该场景下，本研究提出了一种创新性的相关性反馈框架，将伪相关性信号、点击信号和脑信号结合用于文档重排序。本章在交互式相关性反馈和回顾式相关性反馈两种不同设置下验证了脑信号的有效性，展示了其在提升信息检索性能方面的潜力。实验结果表明，对于交互式相关性反馈任务，脑信号特别是在搜索过程的初始阶段或缺少点击反馈的情况下具有显著的帮助作用。此外，本研究还揭示了脑信号能够纠正“误导性点击”带来的有偏用户反馈，并提高回顾式相关性反馈的性能。通过进一步分析不同搜索场景中信号权重的融合对相关性反馈性能的影响，发现脑信号和点击信号的重要性因场景而异。因此，本章进一步提出了一种自适应的信号融合策略，以应对不同的搜索场景，从而实现进一步的搜索结果重排序性能提升。本章内容在前两章的基础上，为构建能感知用户反馈的交互式信息检索系统提供了方法支撑和实验验

证，打破现有交互范式的束缚，以脑机接口技术为核心突破口，旨在构建更智能、更高效、更贴合用户需求的信息系统。

本章相关成果发表于 CCF A 类会议 SIGIR 2022 长文[19]和 CCF A 类期刊 TOIS[20]。

第 6 章　总结与展望

89 年前，图灵提出了计算机的概念。计算机的发明、互联网的普及与信息检索技术的飞跃，极大地加速了信息传播，提升了信息获取的便捷性。互联网和信息检索系统似乎已经成为人类大脑的外接设备[2]，协助人类在完成日常任务时获取信息。而在计算机提出后的第 37 年，脑机接口的概念被引入，用于描述一种将脑电信号转换为计算机信号的控制系统。脑机接口的愿景在于革新传统的人机交互范式，构建超越文本、图像等传统媒介的、更为直接高效的信息交流通路。然而，现有的脑机接口距离实现“心灵感应”的幻想还比较遥远，其研究主要还集中在硬件和解码技术本身，而没有和信息获取相关的技术相结合。鉴于此，本文的核心议题在于探索现有脑机接口技术能为信息检索系统注入何种新活力，并为高效的信息获取与利用开辟何种新路径。尽管由于脑机接口设备在便携性、隐私性和成本方面的限制，本研究仍处于起步阶段，基于脑机接口的信息检索系统在短期内仍难以普及。但值得注意的是，近年来脑机接口技术在科研及临床领域的应用广度与深度持续拓展，其成本与便携性亦在稳步改善。因此，前瞻性地探索脑机接口与信息系统融合的理论与实践路径，具有重要的战略意义与学术价值。本章将首先总结本文的主要研究内容和贡献，然后讨论脑机接口与信息检索系统结合的潜在应用场景和提出未来研究方向。

6.1　研究工作总结

本文针对基于脑机接口的信息检索技术中的关键挑战，提出了一系列创新的研究方法，旨在利用脑机接口提升信息检索过程中的用户认知理解、信息需求解码以及反馈建模的效果。本文首先挖掘了脑信号在用户认知建模中的潜力，利用其相较于传统信号的优势，以进行更准确的满意度建模和更细粒度的阅读理解过程建模。接着，本文针对传统信息检索系统中用户信息需求表达不准确的问题，研究了基于脑信号的开放环境语言解码，并将其应用于查询增强任务，以提升检索系统返回优质信息的能力。最后，本文将脑机接口应用于用户反馈建模，并将脑信号与传统信号融合，构建了更高效的基于用户反馈的信息检索环路优化方案。通过系统地探索和分析脑机接口对现有信息检索的增强，本文的研究在以下几个方面取得了重要进展：

1. **用户认知过程理解：**此部分工作致力于突破传统用户研究技术（如眼动追踪、行为观察及显式标注）的局限，实现更深层次、基于神经活动的用户认知过

程建模。基于脑信号与人类大脑认知活动直接相关的特点，本文首次将脑信号与用户满意度建模相结合，提出了脑拓扑结构自适应网络，有效应对了传统信息系统在准确评估用户满意度方面的挑战。脑拓扑结构自适应网络通过自适应地学习和脑拓扑结构相关的信号融合方法来聚合不同 EEG 电极的信息，在用户满意度建模任务上，展现出超越传统拓扑无关及拓扑固定网络的性能优势。进一步，本文在搜索和推荐场景中验证了基于该网络推断的用户满意度的应用潜力。另一方面，基于脑机接口可以提供更细粒度的用户信号的特点，本文开展了一个用户研究，深入探讨了用户在文本阅读理解过程中的细粒度认知机制。本研究揭示了在文本阅读理解过程中，认知负荷、逻辑推理及语义匹配等心智活动在词汇级别上的神经生理对应物。基于这些发现，本文提出了若干信息检索系统设计的见解，并构建了基于 EEG 的用户阅读状态检测模型，实现了词、句粒度的用户认知状态实时感知。

2. **用户信息需求解码：**此部分工作聚焦于用户信息需求的神经解码，旨在辅助信息检索系统捕获超越用户文本输入的深层意图，以实现更精准高效的检索。本文首先引入了一种自回归的脑语言生成方法，实现从功能性磁共振成像数据到语言的开放式解码，其效果甚至不逊色于一些只针对有限解码目标的分类式方法。接下来，本文将该方法应用于查询增强的场景中，基于用户的脑信号对查询词进行续写，以实现更准确的用户信息需求表达。该方法显著提升了下游文档排序任务的准确率，尤其在处理模糊或歧义性用户查询时，性能增益更为显著。
3. **用户反馈建模：**该工作旨在利用脑电信号作为反馈信号来源，解决传统信号缺失或有偏的问题。本文分别在事实性搜索和非事实性搜索两个场景下进行了探究。在事实性搜索的场景中，本文聚焦于缺乏传统点击反馈的场景，并分析了脑机接口信号在该场景下建模用户反馈的有效性，设计了基于该信号进行个性化和通用意图建模的文档重排序方法。在非事实性搜索的场景中，用户通常需要进行多轮的搜索交互才能获得满意的结果。本文设计了一个融合传统反馈信号和脑信号的相关性反馈方法，分别用于交互式和回顾式相关性反馈任务中，以实现更高效的搜索。用户反馈建模方面的研究探索了脑信号用于更高效的信息获取环路构建的可行性，为未来基于脑机接口的智能信息系统的开发提供了有力的方法支持。

综上所述，本研究探索了脑机接口应用于信息检索系统中的潜力，以提升人与信息交互的效率。本文的研究表明，将脑信号应用于信息检索的各个环节，可以更精准地理解用户的认知状态和信息需求，从而显著提升信息检索系统的用户满

意度和性能。这些成果不仅为脑机接口在更广泛信息技术领域的应用开辟了新思路，也为发展下一代智能化、个性化人机交互范式奠定了重要的理论与技术基石。

6.2　未来工作展望

尽管本研究在脑机接口与信息检索的交叉领域取得了一系列进展，但此方向的探索仍处于初级阶段，诸多理论与技术瓶颈尚待突破。例如，本文的研究仍受限于传统的信息载体和交互方式，主要集中在最经典的搜索和推荐系统，对数据隐私的问题还需要更深入的研究等。未来的研究可以考虑以下几个方面：

信息的载体　本文的研究方法主要基于现有的信息检索系统，这些系统以文本和富文本作为主要的信息载体。然而，信息的载体形式多种多样，除了文本之外，还有图像、音频、视频等多媒体形式。更进一步，从认知哲学的视角审视，人类思维本身具有高度抽象性，其在大脑中的表征可能并不依赖于特定外部媒介形态。脑机接口作为一种利用电信号进行信息传输的技术，也应被视为一种重要的信息载体。因此，未来的研究还应考虑如何拓展信息载体的多样性。例如，脑机接口既可以作为人脑信号的采集设备，也可以实现向人脑的输出，如经颅电刺激技术。此外，在元宇宙中，信息的载体被呈现为一种沉浸式的环境，用户可以进入一个完全数字化的世界，增强用户的参与感和体验深度。如何在脑机接口、元宇宙等新技术的赋能下，构建信息载体的最终形态和人与信息的全新交互方式，是未来研究的重要方向。

广义的信息检索场景　本文的研究主要集中在经典的搜索和推荐系统，然而，信息检索的概念可以是更广泛的。例如，驾驶行为分析[305]，军事武器控制[306]，元宇宙[307]和智能助手[308]等都是广义的信息检索或信息交互场景下重要的研究课题。在这些多样的场景中，探索脑机接口的潜力，构建一种跨场景跨任务的元交互范式，是未来研究的重要方向。

数据隐私与伦理问题　脑机接口直接访问和解码脑信号的特点可能会导致对个体思想的监控，引发了关于隐私、操控和自由意志的担忧[309]。尽管这项技术目前处于非常早期的阶段，这样的应用似乎还很遥远，但提前思考脑机接口与信息检索系统结合带来的数据隐私问题是非常重要的。一方面，端侧计算、联邦学习和差分隐私等技术的发展，可以一定程度上缓解这些数据隐私问题。另一方面，构建负责任的信息交互系统必须从系统设计的源头贯彻“隐私保护设计”的原则，确

保仅采集和处理与特定任务直接相关且最小化的非敏感用户信息，这需要脑机接口技术与信息系统在设计理念与实现路径上的深度协同与创新。

参考文献

[1] Reber P. What is the memory capacity of the human brain[J]. Scientific American, 2010, 4: 2010.

[2] Cerf V G. Cognitive implants[J]. Communications of the ACM, 2014, 57(2): 7-7.

[3] Sparrow B, Liu J, Wegner D M. Google effects on memory: Cognitive consequences of having information at our fingertips[J]. Science, 2011, 333(6043): 776-778.

[4] Kacprzak E, Koesten L M, Ibáñez L D, et al. A query log analysis of dataset search[C]// Proceedings of the 17th International Conference on Web Engineering, ICWE 2017. Springer, 2017: 429-436.

[5] Lu H, Zhang M, Ma S. Between clicks and satisfaction: Study on multi-phase user preferences and satisfaction for online news reading[C]//The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 435-444.

[6] Huang L, van Luijtelaar G. Brain computer interface for epilepsy treatment[J]. Brain-Computer Interface Systems-Recent Progress and Future Prospects, 2013.

[7] Kawala-Sterniuk A, Browarska N, Al-Bakri A, et al. Summary of over fifty years with brain-computer interfaces—a review[J]. Brain Sciences, 2021, 11(1): 43.

[8] Nicolas-Alonso L F, Gomez-Gil J. Brain computer interfaces, a review[J]. Sensors, 2012, 12 (2): 1211-1279.

[9] Wang W, Feng F, He X, et al. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021: 1288-1297.

[10] McGuire N, Moshfeghi Y. Prediction of the realisation of an information need: an eeg study[C]// Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2584-2588.

[11] Chen X, Wang Y, Nakanishi M, et al. High-speed spelling with a noninvasive brain-computer interface[J]. Proceedings of the National Academy of Sciences, 2015, 112(44): E6058-E6067.

[12] Wang X W, Nie D, Lu B L. Emotional state classification from eeg data using machine learning approach[J]. Neurocomputing, 2014, 129: 94-106.

[13] Liu Y, Mao J, Xie X, et al. Challenges in designing a brain-machine search interface[C]//ACM SIGIR Forum: Vol. 54. ACM New York, NY, USA, 2021: 1-13.

[14] Ye Z, Xie X, Liu Y, et al. Towards a better understanding of human reading comprehension with brain signals[C]//Proceedings of the World Wide Web Conference 2022. 2022: 380-391.

[15] Ye Z, Xie X, Liu Y, et al. Brain topography adaptive network for satisfaction modeling in interactive information access system[C]//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 90-100.

[16] Ye Z, Ai Q, Liu Y. Brain-computer interface meets information retrieval: Perspective on next-generation information system[C]//Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding. 2024: 61-65.

[17] Ye Z, Ai Q, Liu Y, et al. Generative language reconstruction from brain recordings[J/OL]. Communications Biology, 2025, 8(1): 346. DOI: 10.1038/s42003-025-07731-7.

[18] Ye Z, Zhan J, Ai Q, et al. Query augmentation with brain signals[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 7561-7570.

[19] Ye Z, Xie X, Liu Y, et al. Why don't you click: Understanding non-click results in web search with brain signals[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 633-645.

[20] Ye Z, Xie X, Ai Q, et al. Relevance feedback with brain signals[J]. ACM Transactions on Information Systems, 2024, 42(4): 1-37.

[21] 潘正源, 李樵, 李月琳, 等. 智能信息检索研究范式的演进, 反思与前瞻[J]. 图书馆论坛, 2024, 44(1): 137-150.

[22] Robertson S, Zaragoza H, et al. The probabilistic relevance framework: Bm25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.

[23] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval [J]. Journal of Documentation, 1972, 28(1): 11-21.

[24] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[25] Zhu Y, Yuan H, Wang S, et al. Large language models for information retrieval: A survey[A]. 2023. arXiv: 2308.07107.

[26] Chuklin A, Markov I, De Rijke M. Click models for web search[M]. Springer Nature, 2022.

[27] Azzopardi L. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval[C]//Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. 2021: 27-37.

[28] Agarwal A, Wang X, Li C, et al. Addressing trust bias for unbiased learning-to-rank[C]// Proceedings of the World Wide Web Conference. 2019: 4-14.

[29] Chen N, Liu J, Fang H, et al. Decoy effect in search interaction: Understanding user behavior and measuring system vulnerability[J]. ACM Transactions on Information Systems, 2025, 43 (2): 1-58.

[30] Xie X, Liu Y, Wang X, et al. Investigating examination behavior of image search users[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 275-284.

[31] Li X, Mao J, Wang C, et al. Teach machine how to read: Reading behavior inspired relevance estimation[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 795-804.

[32] Liu J, Jung Y J. Interest development, knowledge learning, and interactive ir: Toward a state-based approach to search as learning[C]//Proceedings of the 2021 Conference on Human Information Interaction and Retrieval. 2021: 239-248.

[33] Rocchio Jr J J. Relevance feedback in information retrieval[J]. The SMART Retrieval System: Experiments in Automatic Document Processing, 1971.

[34] Chen J, Mao J, Liu Y, et al. A hybrid framework for session context modeling[J]. ACM Transactions on Information Systems, 2021, 39(3): 1-35.

[35] Bi K, Ai Q, Croft W B. Asking clarifying questions based on negative feedback in conversational search[C]//Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 2021: 157-166.

[36] Liu B, Wu Y, Zhang F, et al. Query generation and buffer mechanism: Towards a better conversational agent for legal case retrieval[J]. Information Processing & Management, 2022, 59 (5): 103051.

[37] Xie X, Mao J, Liu Y, et al. Improving web image search with contextual information[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1683-1692.

[38] Datta D, Varma S, Singh S K, et al. Multimodal retrieval using mutual information based textual query reformulation[J]. Expert Systems with Applications, 2017, 68: 81-92.

[39] Knees P, Schedl M. Music retrieval and recommendation: A tutorial overview[C/OL]//SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery, 2015: 1133–1136. DOI: 10.1145/2766462.2767880.

[40] Lu H, Zhang M, Ma W, et al. Quality effects on user preferences and behaviors in mobile news streaming[C]//Proceedings of the 2019 World Wide Web Conference. 2019: 1187-1197.

[41] Khabsa M, Crook A, Awadallah A H, et al. Learning to account for good abandonment in search success metrics[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016: 1893-1896.

[42] Price C J. The anatomy of language: A review of 100 fmri studies published in 2009[J]. Annals of the New York Academy of Sciences, 2010, 1191(1): 62-88.

[43] Beres A M. Time is of the essence: A review of electroencephalography (eeg) and event-related brain potentials (erps) in language research[J]. Applied Psychophysiology and Biofeedback, 2017, 42: 247-255.

[44] Sergent J. Brain-imaging studies of cognitive functions[J]. Trends in Neurosciences, 1994, 17 (6): 221-227.

[45] Rissman J, Wagner A D. Distributed representations in memory: Insights from functional brain imaging[J]. Annual Review of Psychology, 2012, 63(1): 101-128.

[46] Li Y, Pazdera J K, Kahana M J. Eeg decoders track memory dynamics[J]. Nature Communications, 2024, 15(1): 2981.

[47] Cabeza R, Kapur S, Craik F I, et al. Functional neuroanatomy of recall and recognition: A pet study of episodic memory[J]. Journal of Cognitive Neuroscience, 1997, 9(2): 254-265.

[48] Parvizi J, Kastner S. Promises and limitations of human intracranial electroencephalography [J]. Nature Neuroscience, 2018, 21(4): 474-483.

[49] Flesher S N, Downey J E, Weiss J M, et al. A brain-computer interface that evokes tactile sensations improves robotic arm control[J]. Science, 2021, 372(6544): 831-836.

[50] Naddaf M. Science in 2025: the events to watch for in the coming year[J]. Nature, 2025, 637 (8044): 9-11.

[51] Guan Z, Zhang X, Huang W, et al. A method for detecting depression in adolescence based on an affective brain-computer interface and resting-state electroencephalogram signals[J]. Neuroscience Bulletin, 2024: 1-15.

[52] Anumanchipalli G K, Chartier J, Chang E F. Speech synthesis from neural decoding of spoken sentences[J]. Nature, 2019, 568(7753): 493-498.

[53] Tang J, LeBel A, Jain S, et al. Semantic reconstruction of continuous language from non-invasive brain recordings[J]. Nature Neuroscience, 2023: 1-9.

[54] Takagi Y, Nishimoto S. High-resolution image reconstruction with latent diffusion models from human brain activity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14453-14463.

[55] Zheng W L, Dong B N, Lu B L. Multimodal emotion recognition using eeg and eye tracking data[C]//2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, 2014: 5040-5043.

[56] Zhang Y, Qin F, Liu B, et al. Wearable neurophysiological recordings in middle-school classroom correlate with students' academic performance[J]. Frontiers in Human Neuroscience, 2018, 12: 457.

[57] Marler T, Bartels E, Binnendijk A. Brain-computer interfaces: U.s. military applications and implications, an initial assessment[EB/OL]. United States: RAND Corporation, 2020. https://coilink.org/20.500.12592/hkmbf5.

[58] Zander T O, Krol L R, Birbaumer N P, et al. Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity[J]. Proceedings of the National Academy of Sciences, 2016, 113(52): 14898-14903.

[59] Mugler E M, Ruf C A, Halder S, et al. Design and implementation of a p300-based brain-computer interface for controlling an internet browser[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2010, 18(6): 599-609.

[60] Chen X, Ye Z, Xie X, et al. Web search via an efficient and effective brain-machine interface [C]//Proceedings of the 15th ACM International Conference on Web Search and Data Mining. 2022: 1569-1572.

[61] Rao R P, Stocco A, Bryan M, et al. A direct brain-to-brain interface in humans[J]. PloS one, 2014, 9(11): e111332.

[62] Eugster M J, Ruotsalo T, Spapé M M, et al. Predicting term-relevance from brain signals[C]// Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 425-434.

[63] Jacucci G, Barral O, Daee P, et al. Integrating neurophysiologic relevance feedback in intent modeling for information retrieval[J]. Journal of the Association for Information Science and Technology, 2019, 70(9): 917-930.

[64] Miranda-Correa J A, Abadi M K, Sebe N, et al. Amigos: A dataset for affect, personality and mood research on individuals and groups[J]. IEEE Transactions on Affective Computing, 2018, 12(2): 479-493.

[65] Moshfeghi Y, Triantafillou P, Pollick F. Towards predicting a realisation of an information need based on brain signals[C]//Proceedings of the World Wide Web Conference. 2019: 1300-1309.

[66] Moshfeghi Y, Triantafillou P, Pollick F E. Understanding information need: An fMRI study[C]// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 2016: 335-344.

[67] Luck S J. An introduction to the event-related potential technique[M]. MIT Press, 2014.

[68] Kangassalo L, Spapé M, Jacucci G, et al. Why do users issue good queries? neural correlates of term specificity[C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 375-384.

[69] Pinkosova Z, McGeown W J, Moshfeghi Y. The cortical activity of graded relevance[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 299-308.

[70] Ji K, Hettiachchi D, Salim F D, et al. Characterizing information seeking processes with multiple physiological signals[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 1006-1017.

[71] Allegretti M, Moshfeghi Y, Hadjigeorgieva M, et al. When relevance judgement is happening? An EEG-based study[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 719-722.

[72] Eugster M J, Ruotsalo T, Spapé M M, et al. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals[J]. Scientific Reports, 2016, 6 (1): 38580.

[73] Golenia J E, Wenzel M A, Bogojeski M, et al. Implicit relevance feedback from electroencephalography and eye tracking in image search[J]. Journal of Neural Engineering, 2018, 15 (2): 026002.

[74] Gwizdka J. Inferring web page relevance using pupillometry and single channel eeg[M]// Information Systems and Neuroscience. Springer, 2018: 175-183.

[75] Michalkova D, Rodriguez M P, Moshfeghi Y. Understanding feeling-of-knowing in information search: An eeg study[J]. ACM Transactions on Information Systems, 2024, 42(3): 1-30.

[76] Baumgartner T, Valko L, Esslen M, et al. Neural correlate of spatial presence in an arousing and noninteractive virtual reality: An eeg and psychophysiology study[J]. CyberPsychology & Behavior, 2006, 9(1): 30-45.

[77] Martindale C, Moore K. Priming, prototypicality, and preference.[J]. Journal of Experimental Psychology: Human Perception and Performance, 1988, 14(4): 661.

[78] Silvia P J. Cognitive appraisals and interest in visual art: Exploring an appraisal theory of aesthetic emotions[J]. Empirical Studies of the Arts, 2005, 23(2): 119-133.

[79] Berlyne D E. Novelty, complexity, and hedonic value[J]. Perception & Psychophysics, 1970, 8(5): 279-286.

[80] Nadal M, Munar E, Capo M A, et al. Towards a framework for the study of the neural correlates of aesthetic preference[J]. Spatial Vision, 2008, 21(3): 379.

[81] Lawhern V J, Solon A J, Waytowich N R, et al. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces[J]. Journal of Neural Engineering, 2018, 15 (5): 056013.

[82] Kostas D, Aroca-Ouellette S, Rudzicz F. Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data[J]. Frontiers in Human Neuroscience, 2021, 15: 653659.

[83] Song T, Zheng W, Song P, et al. Eeg emotion recognition using dynamical graph convolutional neural networks[J]. IEEE Transactions on Affective Computing, 2018, 11(3): 532-541.

[84] Zhong P, Wang D, Miao C. Eeg-based emotion recognition using regularized graph neural networks[J/OL]. IEEE Transactions on Affective Computing, 2022, 13(3): 1290-1301. DOI: 10.1109/TAFFC.2020.2994159.

[85] Jia Z, Lin Y, Wang J, et al. Hetemotionnet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 1047-1056.

[86] Li R, Wang Y, Lu B L. A multi-domain adaptive graph convolutional network for eeg-based emotion recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 5565-5573.

[87] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30.

[88] Luck S J, Woodman G F, Vogel E K. Event-related potential studies of attention[J]. Trends in Cognitive Sciences, 2000, 4(11): 432-440.

[89] Raney G E. Monitoring changes in cognitive load during reading: An event-related brain potential and reaction time analysis.[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1993, 19(1): 51.

[90] Kelly D, et al. Methods for evaluating interactive information retrieval systems with users[J]. Foundations and Trends® in Information Retrieval, 2009, 3(1–2): 1-224.

[91] Ali R, Beg M S. An overview of web search evaluation methods[J]. Computers & Electrical Engineering, 2011, 37(6): 835-848.

[92] Liu M, Liu Y, Mao J, et al. ” satisfaction with failure” or ” unsatisfied success” investigating the relationship between search success and user satisfaction[C]//Proceedings of the 2018 World Wide Web Conference. 2018: 1533-1542.

[93] Fox S, Karnawat K, Mydland M, et al. Evaluating implicit measures to improve web search[J]. ACM Transactions on Information Systems, 2005, 23(2): 147-168.

[94] Hassan A, White R W, Dumais S T, et al. Struggling or exploring? disambiguating long search sessions[C]//Proceedings of the 7th ACM international conference on Web search and data mining. 2014: 53-62.

[95] Chen Y, Liu Y, Zhang M, et al. User satisfaction prediction with mouse movement information in heterogeneous search environment[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(11): 2470-2483.

[96] Wu Y, Liu Y, Tsai Y H R, et al. Investigating the role of eye movements and physiological signals in search satisfaction prediction using geometric analysis[J]. Journal of the Association for Information Science and Technology, 2019, 70(9): 981-999.

[97] Amatriain X, Pujol J M, Oliver N. I like it... i like it not: Evaluating user ratings noise in recommender systems[C]//International Conference on User Modeling, Adaptation, and Personalization. Springer, 2009: 247-258.

[98] Liu Y, Wang C, Zhou K, et al. From skimming to reading: A two-stage examination model for web search[C]//Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. 2014: 849-858.

[99] Crowder R G, Wagner R K. The psychology of reading: An introduction[M]. Oxford University Press, 1992.

[100] Liu Y, Chen Y, Tang J, et al. Different users, different opinions: Predicting search satisfaction with mouse movement information[C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 493-502.

[101] Li X, Liu Y, Mao J, et al. Understanding reading attention distribution during relevance judgement[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 733-742.

[102] Gwizdka J. Characterizing relevance with eye-tracking measures[C]//Proceedings of the 5th Information Interaction in Context Symposium. 2014: 58-67.

[103] Cole M J, Gwizdka J, Liu C, et al. Inferring user knowledge level from eye movement patterns [J]. Information Processing & Management, 2013, 49(5): 1075-1091.

[104] Hsiao J H w. Visual field differences in visual word recognition can emerge purely from perceptual learning: Evidence from modeling chinese character pronunciation[J]. Brain and Language, 2011, 119(2): 89-98.

[105] Newman S D, Ikuta T, Burns Jr T. The effect of semantic relatedness on syntactic analysis: An fmri study[J]. Brain and Language, 2010, 113(2): 51-58.

[106] Jia Z, Lin Y, Wang J, et al. Graphsleepnet: Adaptive spatial-temporal graph convolutional networks for sleep stage classification[C]//Proceedings of the 29th International Joint Conference on Artificial Intelligence. 2020: 1324-1330.

[107] Supratak A, Guo Y. Tinysleepnet: An efficient deep learning model for sleep stage scoring based on raw single-channel eeg[C]//2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2020: 641-644.

[108] Guan S, Zhao K, Yang S. Motor imagery EEG classification based on decision tree framework and riemannian geometry[J/OL]. Comput. Intell. Neurosci., 2019, 2019: 5627156:1-5627156:13. DOI: 10.1155/2019/5627156.

[109] Li X, Song D, Zhang P, et al. Emotion recognition from multi-channel eeg data through convolutional recurrent neural network[C]//2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2016: 352-359.

[110] Zhang G, Yu M, Liu Y, et al. Sparsedgcnn: Recognizing emotion from multichannel EEG signals[J/OL]. IEEE Transactions on Affective Computing, 2023, 14(1): 537-548. DOI: 10.1109/TAFFC.2021.3051332.

[111] Ho T C, Connolly C G, Blom E H, et al. Emotion-dependent functional connectivity of the default mode network in adolescent depression[J]. Biological Psychiatry, 2015, 78(9): 635-646.

[112] Yin Z, Li J, Zhang Y, et al. Functional brain network analysis of schizophrenic patients with positive and negative syndrome based on mutual information of eeg time series[J]. Biomedical Signal Processing and Control, 2017, 31: 331-338.

[113] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information[J]. Physical Review E, 2004, 69(6): 066138.

[114] Schmidt L A, Trainor L J. Frontal brain electrical activity (eeg) distinguishes valence and intensity of musical emotions[J]. Cognition & Emotion, 2001, 15(4): 487-500.

[115] Cheng H T, Koc L, Harmsen J, et al. Wide & deep learning for recommender systems[C]// Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. 2016: 7-10.

[116] Järvelin K, Kekäläinen J. Ir evaluation methods for retrieving highly relevant documents[C]// ACM SIGIR Forum: Vol. 51. ACM New York, NY, USA, 2017: 243-250.

[117] Zhu M. Recall, precision and average precision[J]. Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 2004, 2(30): 6.

[118] Duan R N, Zhu J Y, Lu B L. Differential entropy feature for eeg-based emotion classification [C]//2013 6th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE, 2013: 81-84.

[119] Zerveas G, Jayaraman S, Patel D, et al. A transformer-based framework for multivariate time series representation learning[C]//Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. 2021: 2114-2124.

[120] Pyakillya B, Kazachenko N, Mikhailovsky N. Deep learning for ecg classification[C]//Journal of Physics: Conference Series: Vol. 913. IOP Publishing, 2017: 012004.

[121] Suykens J A, Vandewalle J. Least squares support vector machine classifiers[J]. Neural Processing Letters, 1999, 9(3): 293-300.

[122] Safavian S R, Landgrebe D. A survey of decision tree classifier methodology[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 21(3): 660-674.

[123] Davis III K M, Spapé M, Ruotsalo T. Collaborative filtering with preferences inferred from brain signals[C]//Proceedings of the World Wide Web Conference 2021. 2021: 602-611.

[124] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in python[J]. the Journal of Machine Learning Research, 2011, 12: 2825-2830.

[125] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//Bengio Y, LeCun Y. 3rd International Conference on Learning Representations. 2015.

[126] Yao D, Qin Y, Hu S, et al. Which reference should we use for eeg and erp practice?[J]. Brain Topography, 2019, 32(4): 530-549.

[127] Gjoreski M, Mitrevski B, Luštrek M, et al. An inter-domain study for arousal recognition from physiological signals[J]. Informatica, 2018, 42(1).

[128] Hu J, Wang C, Jia Q, et al. Scalingnet: Extracting features from raw eeg data for emotion recognition[J]. Neurocomputing, 2021, 463: 177-184.

[129] Lavrenko V, Croft W B. Relevance-based language models[C]//ACM SIGIR Forum: Vol. 51. ACM New York, NY, USA, 2017: 260-267.

[130] Richardson M, Dominowska E, Ragno R. Predicting clicks: Estimating the click-through rate for new ads[C]//Proceedings of the 16th International Conference on World Wide Web. 2007: 521-530.

[131] Rendle S. Factorization machines[C]//2010 IEEE International Conference on Data Mining. IEEE, 2010: 995-1000.

[132] Zhao W X, Mu S, Hou Y, et al. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 4653-4664.

[133] Powers D M W. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation[A]. 2020. arXiv: 2308.07107.

[134] Aldayel M, Ykhlef M, Al-Nafjan A. Deep learning for eeg-based preference classification in neuromarketing[J]. Applied Sciences, 2020, 10(4): 1525.

[135] Wang W, Feng F, He X, et al. Denoising implicit feedback for recommendation[C]//Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 373-381.

[136] Li P, Li W, He Z, et al. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering[A/OL]. 2016. arXiv: 1607.06275. https://arxiv.org/abs/1607.06275.

[137] Kutas M, Hillyard S A. Reading senseless sentences: Brain potentials reflect semantic incongruity[J]. Science, 1980, 207(4427): 203-205.

[138] Dimigen O, Sommer W, Hohlfeld A, et al. Coregistration of eye movements and eeg in natural reading: Analyses and review.[J]. Journal of Experimental Psychology: General, 2011, 140(4): 552.

[139] Jiang X, Zhou X. An alternative structure rescues failed semantics? strong global expectancy reduces local-mismatch n400 in chinese flexible structures[J]. Neuropsychologia, 2020, 140: 107380.

[140] Craswell N, Mitra B, Yilmaz E, et al. Overview of the trec 2019 deep learning track[A/OL]. 2020. arXiv: 2003.07820. https://arxiv.org/abs/2003.07820.

[141] Blackwood D, Muir W. Cognitive brain potentials and their application[J]. The British Journal of Psychiatry, 1990, 157(S9): 96-101.

[142] Huizenga H M, De Munck J C, Waldorp L J, et al. Spatiotemporal eeg/meg source analysis based on a parametric noise covariance model[J]. IEEE Transactions on Biomedical Engineering, 2002, 49(6): 533-539.

[143] Lehmann D, Skrandies W. Reference-free identification of components of checkerboard-evoked multichannel potential fields[J]. Electroencephalography and Clinical Neurophysiology, 1980, 48(6): 609-621.

[144] Vogel E K, Luck S J. The visual n1 component as an index of a discrimination process[J]. Psychophysiology, 2000, 37(2): 190-203.

[145] Moshfeghi Y, Pinto L R, Pollick F E, et al. Understanding relevance: An fmri study[C]// European Conference on Information Retrieval. Springer, 2013: 14-25.

[146] Kutas M, Hillyard S A. Brain potentials during reading reflect word expectancy and semantic association[J]. Nature, 1984, 307(5947): 161-163.

[147] Hoeks J C, Stowe L A, Doedens G. Seeing words in context: the interaction of lexical and sentence level information during reading[J]. Cognitive Brain Research, 2004, 19(1): 59-73.

[148] Van Herten M, Kolk H H, Chwilla D J. An erp study of p600 effects elicited by semantic anomalies[J]. Cognitive Brain Research, 2005, 22(2): 241-255.

[149] Burkhardt P. Inferential bridging relations reveal distinct neural mechanisms: Evidence from event-related brain potentials[J]. Brain and Language, 2006, 98(2): 159-168.

[150] Jiang C G, Wang J, Liu X H, et al. The neural correlates of effortful cognitive processing deficits in schizophrenia: An erp study[J]. Frontiers in Human Neuroscience, 2021, 15.

[151] Liu T, Goldberg L, Gao S, et al. An online brain-computer interface using non-flashing visual evoked potentials[J]. Journal of Neural Engineering, 2010, 7(3): 036003.

[152] Harmony T, Fernández T, Silva J, et al. Eeg delta activity: an indicator of attention to internal processing during performance of mental tasks[J]. International journal of psychophysiology, 1996, 24(1-2): 161-171.

[153] Klimesch W. Eeg alpha and theta oscillations reflect cognitive and memory performance: A review and analysis[J]. Brain Research Reviews, 1999, 29(2-3): 169-195.

[154] Penolazzi B, Angrilli A, Job R. Gamma eeg activity induced by semantic violation during sentence reading[J]. Neuroscience Letters, 2009, 465(1): 74-78.

[155] Gwizdka J, Hosseini R, Cole M, et al. Temporal dynamics of eye-tracking and eeg during reading and relevance decisions[J]. Journal of the Association for Information Science and Technology, 2017, 68(10): 2299-2312.

[156] Zhang Y, Zhao Q, Jin J, et al. A novel bci based on erp components sensitive to configural processing of human faces[J]. Journal of Neural Engineering, 2012, 9(2): 026018.

[157] Lundberg S M, Lee S. A unified approach to interpreting model predictions[C]//Guyon I, von Luxburg U, Bengio S, et al. Advances in Neural Information Processing Systems. 2017: 4765-4774.

[158] Sur S, Sinha V. Event-related potential: An overview[J]. Industrial Psychiatry Journal, 2009, 18(1): 70.

[159] Zheng Y, Mao J, Liu Y, et al. Human behavior inspired machine reading comprehension[C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 425-434.

[160] Yano Y, Tagami Y, Tajima A. Quantifying query ambiguity with topic distributions[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016: 1877-1880.

[161] Cronen-Townsend S, Zhou Y, Croft W B. Predicting query performance[C]//Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2002: 299-306.

[162] Mitchell T M, Shinkareva S V, Carlson A, et al. Predicting human brain activity associated with the meanings of nouns[J]. Science, 2008, 320(5880): 1191-1195.

[163] Pei X, Barbour D L, Leuthardt E C, et al. Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans[J]. Journal of Neural Engineering, 2011, 8(4): 046028.

[164] Pereira F, Lou B, Pritchett B, et al. Toward a universal decoder of linguistic meaning from brain activation[J]. Nature Communications, 2018, 9(1): 963.

[165] Kivisaari S L, van Vliet M, Hultén A, et al. Reconstructing meaning from bits of information [J]. Nature Communications, 2019, 10(1): 927.

[166] Moses D A, Metzger S L, Liu J R, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria[J]. New England Journal of Medicine, 2021, 385(3): 217-227.

[167] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.

[168] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[169] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models [A/OL]. 2023. arXiv: 2302.13971. https://arxiv.org/abs/2302.13971.

[170] Affolter N, Egressy B, Pascual D, et al. Brain2word: Decoding brain activity for language generation[A/OL]. 2020. arXiv: 2009.04765. https://arxiv.org/abs/2009.04765.

[171] Toneva M, Wehbe L. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)[J]. Advances in Neural Information Processing Systems, 2019, 32.

[172] Antonello R, Huth A. Predictive coding or just feature discovery? an alternative account of why language models fit brain data[J]. Neurobiology of Language, 2024, 5(1): 64-79.

[173] Goldstein A, Zada Z, Buchnik E, et al. Shared computational principles for language processing in humans and deep language models[J]. Nature Neuroscience, 2022, 25(3): 369-380.

[174] Antonello R, Vaidya A, Huth A. Scaling laws for language encoding models in fmri[J]. Advances in Neural Information Processing Systems, 2023, 36: 21895-21907.

[175] Liu X, Zhou M, Shi G, et al. Coupling artificial neurons in bert and biological neurons in the human brain[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 37. 2023: 8888-8896.

[176] Mei Q, Zhou D, Church K. Query suggestion using hitting time[C]//Proceedings of the 17th ACM Conference on Information and Knowledge Management. 2008: 469-478.

[177] Ahmad W U, Chang K W, Wang H. Context attentive document ranking and query suggestion [C]//Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2019: 385-394.

[178] Pereira M, Etemad E, Paulovich F. Iterative learning to rank from explicit relevance feedback [C]//Proceedings of the 35th Annual ACM Symposium on Applied Computing. 2020: 698-705.

[179] Willett F R, Avansino D T, Hochberg L R, et al. High-performance brain-to-text communication via handwriting[J]. Nature, 2021, 593(7858): 249-254.

[180] Défossez A, Caucheteux C, Rapin J, et al. Decoding speech perception from non-invasive brain recordings[J]. Nature Machine Intelligence, 2023: 1-11.

[181] Bi K, Ai Q, Croft W B. Iterative relevance feedback for answer passage retrieval with passage-level semantic match[C]//Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41. Springer, 2019: 558-572.

[182] Li H, Scells H, Zuccon G. Systematic review automation tools for end-to-end query formulation [C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 2141-2144.

[183] Li H, Mourad A, Koopman B, et al. How does feedback signal quality impact effectiveness of pseudo relevance feedback for passage retrieval?[C]//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022: 2154-2158.

[184] Toneva M. Bridging language in machines with language in the brain[D]. Carnegie Mellon University, 2021.

[185] Liu X, Zheng Y, Du Z, et al. GPT understands, too[J/OL]. AI Open, 2024, 5: 208-215. DOI: 10.1016/J.AIOPEN.2023.08.012.

[186] Duan Y, Chau C, Wang Z, et al. Dewave: Discrete encoding of eeg waves for eeg to text translation[J]. Advances in Neural Information Processing Systems, 2024, 36.

[187] Fathullah Y, Wu C, Lakomkin E, et al. Audiochatllama: Towards general-purpose speech abilities for llms[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2024: 5522-5532.

[188] Chu Y, Xu J, Yang Q, et al. Qwen2-audio technical report[A/OL]. 2024. arXiv: 2407.10759. https://arxiv.org/abs/2407.10759.

[189] Huang S, Dong L, Wang W, et al. Language is not all you need: Aligning perception with language models[J]. Advances in Neural Information Processing Systems, 2024, 36.

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

[191] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in Neural Information Processing Systems, 2023, 36: 34892-34916.

[192] LeBel A, Wagner L, Jain S, et al. A natural language fmri dataset for voxelwise encoding models [J]. Scientific Data, 2023, 10(1): 555.

[193] Nastase S A, Liu Y F, Hillman H, et al. The “narratives” fmri dataset for evaluating models of naturalistic language comprehension[J]. Scientific Data, 2021, 8(1): 250.

[194] Luo Y, Xu M, Xiong D. Cogtaskonomy: Cognitively inspired task taxonomy is beneficial to transfer learning in nlp[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 904-920.

[195] Abdi H, Williams L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459.

[196] Wang Z, Ji H. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 36. 2022: 5350-5358.

[197] Xi N, Zhao S, Wang H, et al. Unicorn: Unified cognitive signal reconstruction bridging cognitive signals and human language[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. 2023: 13277-13291.

[198] Yang Y, Duan Y, Zhang Q, et al. Neuspeech: Decode neural signal as speech[A/OL]. 2024. arXiv: 2403.01748. https://arxiv.org/abs/2403.01748.

[199] Zhao X, Sun J, Wang S, et al. Mapguide: A simple yet effective method to reconstruct continuous language from brain activities[A/OL]. 2024. arXiv: 2403.17516. https://arxiv.org/abs/2403.17516.

[200] Yin C, Ye Z, Li P. Language reconstruction with brain predictive coding from fmri data[A/OL]. 2024. arXiv: 2405.11597. https://arxiv.org/abs/2405.11597.

[201] Chen X, Du C, Liu C, et al. Open-vocabulary auditory neural decoding using fmri-prompted llm[A/OL]. 2024. arXiv: 2405.07840. https://arxiv.org/abs/2405.07840.

[202] Jo H, Yang Y, Han J, et al. Are eeg-to-text models working?[A/OL]. 2024. arXiv: 2405.06459. https://arxiv.org/abs/2405.06459.

[203] Meister C, Cotterell R. Language model evaluation beyond perplexity[A/OL]. 2021. arXiv: 2106.00085. https://arxiv.org/abs/2106.00085.

[204] Brysbaert M, Warriner A B, Kuperman V. Concreteness ratings for 40 thousand generally known english word lemmas[J]. Behavior Research Methods, 2014, 46: 904-911.

[205] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[A/OL]. 2020. arXiv: 2001.08361. https://arxiv.org/abs/2001.08361.

[206] Ganguli D, Hernandez D, Lovitt L, et al. Predictability and surprise in large generative models [C]//Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022: 1747-1764.

[207] Tikochinski R, Goldstein A, Meiri Y, et al. Incremental accumulation of linguistic context in artificial and biological neural networks[J]. Nature Communications, 2025, 16(1): 803.

[208] Zhao W X, Zhou K, Li J, et al. A survey of large language models[A/OL]. 2024. arXiv: 2303.18223. https://arxiv.org/abs/2303.18223.

[209] Musso M, Moro A, Glauche V, et al. Broca's area and the language instinct[J]. Nature Neuroscience, 2003, 6(7): 774-781.

[210] Chee M W, Soon C S, Lee H L, et al. Left insula activation: A marker for language attainment in bilinguals[J]. Proceedings of the National Academy of Sciences, 2004, 101(42): 15265-15270.

[211] Gabrieli J D, Poldrack R A, Desmond J E. The role of left prefrontal cortex in language and memory[J]. Proceedings of the National Academy of Sciences, 1998, 95(3): 906-913.

[212] Salmelin R, Schnitzler A, Parkkonen L, et al. Native language, gender, and functional organization of the auditory cortex[J]. Proceedings of the National Academy of Sciences, 1999, 96 (18): 10460-10465.

[213] Van Ettinger-Veenstra H, McAllister A, Lundberg P, et al. Higher language ability is related to angular gyrus activation increase during semantic processing, independent of sentence incongruency[J]. Frontiers in Human Neuroscience, 2016, 10: 110.

[214] Price A R, Bonner M F, Peelle J E, et al. Converging evidence for the neuroanatomic basis of combinatorial semantics in the angular gyrus[J]. Journal of Neuroscience, 2015, 35(7): 3276-3284.

[215] Lerner Y, Honey C J, Silbert L J, et al. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story[J]. Journal of Neuroscience, 2011, 31(8): 2906-2915.

[216] Binder J R, Desai R H. The neurobiology of semantic memory[J]. Trends in Cognitive Sciences, 2011, 15(11): 527-536.

[217] Keller T A, Carpenter P A, Just M A. The neural bases of sentence comprehension: A fmri examination of syntactic and lexical processing[J]. Cerebral Cortex, 2001, 11(3): 223-237.

[218] Caucheteux C, King J R. Brains and algorithms partially converge in natural language processing[J]. Communications Biology, 2022, 5(1): 134.

[219] Liu X, Ji K, Fu Y, et al. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 61-68.

[220] Schrimpf M, Blank I A, Tuckute G, et al. The neural architecture of language: Integrative modeling converges on predictive processing[J]. Proceedings of the National Academy of Sciences, 2021, 118(45): e2105646118.

[221] Hale J T, Campanelli L, Li J, et al. Neurocomputational models of language processing[J]. Annual Review of Linguistics, 2022, 8: 427-446.

[222] Anderson A J, Kiela D, Binder J R, et al. Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning[J]. Journal of Neuroscience, 2021, 41(18): 4100-4119.

[223] Sun J, Wang S, Zhang J, et al. Neural encoding and decoding with distributed sentence representations[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(2): 589-603.

[224] Aw K L, Toneva M. Training language models to summarize narratives improves brain alignment[C]//11th International Conference on Learning Representations. 2023.

[225] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning [A/OL]. 2021. arXiv: 2104.08691. https://arxiv.org/abs/2104.08691.

[226] Sun J, Wang S, Zhang J, et al. Towards sentence-level brain decoding with distributed representations[C]//Proceedings of the AAAI Conference on Artificial Intelligence: Vol. 33. 2019: 7047-7054.

[227] Nishimoto S, Vu A T, Naselaris T, et al. Reconstructing visual experiences from brain activity evoked by natural movies[J]. Current Biology, 2011, 21(19): 1641-1646.

[228] Scotti P S, Tripathy M, Villanueva C K T, et al. Mindeye2: Shared-subject models enable fmri-to-image with 1 hour of data[A/OL]. 2024. arXiv: 2403.11207. https://arxiv.org/abs/2403.11207.

[229] Scotti P, Banerjee A, Goode J, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors[J]. Advances in Neural Information Processing Systems, 2024, 36.

[230] Ozcelik F, VanRullen R. Natural scene reconstruction from fmri signals using generative latent diffusion[J]. Scientific Reports, 2023, 13(1): 15666.

[231] Luo A, Henderson M, Wehbe L, et al. Brain diffusion for visual exploration: Cortical discovery using large scale generative models[J]. Advances in Neural Information Processing Systems, 2024, 36.

[232] Lupyan G, Clark A. Words and the world: Predictive coding and the language-perception-cognition interface[J]. Current Directions in Psychological Science, 2015, 24(4): 279-284.

[233] Clark A. Whatever next? predictive brains, situated agents, and the future of cognitive science [J]. Behavioral and Brain Sciences, 2013, 36(3): 181-204.

[234] Zhu D, Bieger J, Garcia Molina G, et al. A survey of stimulation methods used in ssvep-based bcis[J]. Computational Intelligence and Neuroscience, 2010, 2010.

[235] Metzger S L, Littlejohn K T, Silva A B, et al. A high-performance neuroprosthesis for speech decoding and avatar control[J]. Nature, 2023, 620(7976): 1037-1046.

[236] Bai Y, Jones A, Ndousse K, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback[A/OL]. 2022. arXiv: 2204.05862. https://arxiv.org/abs/2204.05862.

[237] Izacard G, Caron M, Hosseini L, et al. Unsupervised dense information retrieval with contrastive learning[A/OL]. 2022. arXiv: 2112.09118. https://arxiv.org/abs/2112.09118.

[238] Lee K, Chang M W, Toutanova K. Latent retrieval for weakly supervised open domain question answering[A/OL]. 2019. arXiv: 1906.00300. https://arxiv.org/abs/1906.00300.

[239] Ma X, Wang L, Yang N, et al. Fine-tuning llama for multi-stage text retrieval[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2421-2425.

[240] Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques[J]. ACM Transactions on Information Systems, 2002, 20(4): 422-446.

[241] Robertson S. Understanding inverse document frequency: On theoretical arguments for IDF [J]. Journal of Documentation, 2004, 60(5): 503-520.

[242] Thakur N, Reimers N, Rücklé A, et al. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models[A/OL]. 2021. arXiv: 2104.08663. https://arxiv.org/abs/2104.08663.

[243] Shtok A, Kurland O, Carmel D, et al. Predicting query performance by query-drift estimation [J]. ACM Transactions on Information Systems, 2012, 30(2): 1-35.

[244] Meng C, Arabzadeh N, Aliannejadi M, et al. Query performance prediction: From ad-hoc to conversational search[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 2583-2593.

[245] Harman D. Relevance feedback revisited[C]//Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1992: 1-10.

[246] White R W, Ruthven I, Jose J M. The use of implicit evidence for relevance feedback in web retrieval[C]//Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25–27, 2002 Proceedings 24. Springer, 2002: 93-109.

[247] Montazeralghaem A, Zamani H, Allan J. A reinforcement learning framework for relevance feedback[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 59-68.

[248] Joachims T, et al. Evaluating retrieval performance using clickthrough data.[M]. Citeseer, 2003.

[249] Radlinski F, Joachims T. Query chains: Learning to rank from implicit feedback[C]// Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 2005: 239-248.

[250] Joachims T, Granka L, Pan B, et al. Accurately interpreting clickthrough data as implicit feedback[C]//ACM SIGIR Forum: Vol. 51. Acm New York, NY, USA, 2017: 4-11.

[251] Akuma S. Eye gaze relevance feedback indicators for information retrieval[J]. International Journal of Intelligent Systems and Applications (IJISA), 2022, 14(1): 57-65.

[252] Mao J, Liu Y, Zhang M, et al. Estimating credibility of user clicks with mouse movement and eye-tracking information[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 2014: 263-274.

[253] Aalbersberg I J. Incremental relevance feedback[C]//Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval. 1992: 11-22.

[254] Joachims T, Swaminathan A, Schnabel T. Unbiased learning-to-rank with biased feedback[C]// Proceedings of the Tenth ACM International Conference on Web Search and Data Mining. 2017: 781-789.

[255] Nogueira R, Cho K. Passage re-ranking with bert[A/OL]. 2020. arXiv: 1901.04085. https: //arxiv.org/abs/1901.04085.

[256] Yang W, Xie Y, Lin A, et al. End-to-end open-domain question answering with bertserini [C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2019. http: //dx.doi.org/10.18653/v1/N19-4013.

[257] Lin S C, Yang J H, Nogueira R, et al. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting[J]. ACM Transactions on Information Systems, 2021, 39(4): 1-29.

[258] Wang X, Macdonald C, Tonellotto N, et al. Pseudo-relevance feedback for multiple representation dense retrieval[C]//Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval. 2021: 297-306.

[259] Wang X, Macdonald C, Tonellotto N, et al. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval[J]. ACM Transactions on the Web, 2023, 17(1): 1-39.

[260] Yu H, Xiong C, Callan J. Improving query representations for dense retrieval with pseudo relevance feedback[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 3592-3596.

[261] Zheng Z, Hui K, He B, et al. Bert-qe: Contextualized query expansion for document re-ranking [A/OL]. 2020. arXiv: 2009.07258. https://arxiv.org/abs/2009.07258.

[262] Bi K, Ai Q, Zhang Y, et al. Conversational product search based on negative feedback[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 359-368.

[263] Morita M, Shinoda Y. Information filtering based on user behavior analysis and best match text retrieval[C]//Proceedings of the Seventeenth International ACM SIGIR Conference on Research and Development in Information Retrieval. Springer, 1994: 272-281.

[264] Dolma Y, Kalani R, Agrawal A, et al. Improving bounce rate prediction for rare queries by leveraging landing page signals[C]//Companion Proceedings of the Web Conference 2021. 2021: 1-6.

[265] Li J, Huffman S, Tokuda A. Good abandonment in mobile and pc internet search[C]// Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009: 43-50.

[266] Wu D, Dong J, Shi L, et al. Credibility assessment of good abandonment results in mobile search[J]. Information Processing & Management, 2020, 57(6): 102350.

[267] Claypool M, Le P, Wased M, et al. Implicit interest indicators[C]//Proceedings of the 6th International Conference on Intelligent User Interfaces. 2001: 33-40.

[268] White R W, Jose J M, Ruthven I. An implicit feedback approach for interactive information retrieval[J]. Information Processing & Management, 2006, 42(1): 166-190.

[269] Can E F, Croft W B, Manmatha R. Incorporating query-specific feedback into learning-to-rank models[C]//Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. 2014: 1035-1038.

[270] Lv Y, Zhai C. Adaptive relevance feedback in information retrieval[C]//Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009: 255-264.

[271] Yin H, Cui B, Chen L, et al. A temporal context-aware model for user behavior modeling in social media systems[C]//Proceedings of the 2014 ACM SIGMOD international conference on Management of data. 2014: 1543-1554.

[272] Zhang R, Xie X, Mao J, et al. Constructing a comparison-based click model for web search [C]//Proceedings of the Web Conference 2021. 2021: 270-283.

[273] Zhang F, Liu Y, Mao J, et al. User behavior modeling for web search evaluation[J]. AI Open, 2020, 1: 40-56.

[274] Diriye A, White R, Buscher G, et al. Leaving so soon? understanding and predicting web search abandonment rationales[C]//Proceedings of the 21st ACM International Conference on Information and Knowledge Management. 2012: 1025-1034.

[275] Williams K, Kiseleva J, Crook A C, et al. Detecting good abandonment in mobile search[C]// Proceedings of the 25th International Conference on World Wide Web. 2016: 495-505.

[276] Sakai T. Towards zero-click mobile ir evaluation: Knowing what and knowing when[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2012: 1157-1158.

[277] Luo C, Liu Y, Sakai T, et al. Evaluating mobile search with height-biased gain[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 435-444.

[278] Zhang J, Liu Y, Ma S, et al. Relevance estimation with multiple information sources on search engine result pages[C]//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 627-636.

[279] Mao J, Luo C, Zhang M, et al. Constructing click models for mobile search[C]//Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. 2018: 775-784.

[280] Voorhees E M. The philosophy of information retrieval evaluation[C]//Workshop of the Cross-Language Evaluation Forum for European Languages. Springer, 2001: 355-370.

[281] Mao J, Liu Y, Zhou K, et al. When does relevance mean usefulness and user satisfaction in web search?[C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2016: 463-472.

[282] Mao J, Liu Y, Luan H, et al. Understanding and predicting usefulness judgment in web search [C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 1169-1172.

[283] Mao J, Liu Y, Kando N, et al. Investigating result usefulness in mobile search[C]//European Conference on Information Retrieval. Springer, 2018: 223-236.

[284] Koelstra S, Muhl C, Soleymani M, et al. Deap: A database for emotion analysis; using physiological signals[J]. IEEE Transactions on Affective Computing, 2011, 3(1): 18-31.

[285] Yang K, Tong L, Shu J, et al. High gamma band eeg closely related to emotion: Evidence from functional network[J]. Frontiers in Human Neuroscience, 2020, 14: 89.

[286] Jia Z, Lin Y, Cai X, et al. Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition[C]//Proceedings of the 28th ACM International Conference on Multimedia. 2020: 2909-2917.

[287] Frey A, Ionescu G, Lemaire B, et al. Decision-making in information seeking on texts: An eye-fixation-related potentials investigation[J]. Frontiers in Systems Neuroscience, 2013, 7: 39.

[288] Vidaurre C, Blankertz B. Towards a cure for bci illiteracy[J]. Brain Topography, 2010, 23(2): 194-198.

[289] Zheng W L, Lu B L. Personalizing eeg-based affective models with transfer learning[C]// Proceedings of the 25th International Joint Conference on Artificial Intelligence. 2016: 2732-2738.

[290] Arguello J. Predicting search task difficulty[C]//European Conference on Information Retrieval. Springer, 2014: 88-99.

[291] Yates A, Nogueira R, Lin J. Pretrained transformers for text ranking: Bert and beyond[C]// Proceedings of the 14th ACM International Conference on Web Search and Data Mining. 2021: 1154-1156.

[292] Borlund P. The concept of relevance in ir[J]. Journal of the American Society for information Science and Technology, 2003, 54(10): 913-925.

[293] Liu Y, Song R, Zhang M, et al. Overview of the ntcir-11 imine task[C]//Proceedings of NTCIR-11. 2014.

[294] Clarke C L, Craswell N, Soboroff I. Overview of the trec 2009 web track.[C]//Trec: Vol. 9. 2009: 20-29.

[295] Mudrik L, Lamy D, Deouell L Y. Erp evidence for context congruity effects during simultaneous object–scene processing[J]. Neuropsychologia, 2010, 48(2): 507-517.

[296] Xie X, Dong Q, Wang B, et al. T2ranking: A large-scale chinese benchmark for passage ranking [C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 2681-2690.

[297] Bi K, Ai Q, Croft W B. Revisiting iterative relevance feedback for document and passage retrieval[A/OL]. 2019. arXiv: 1812.05731. https://arxiv.org/abs/1812.05731.

[298] Lan Z, Sourina O, Wang L, et al. Domain adaptation techniques for eeg-based emotion recognition: A comparative study on two public datasets[J]. IEEE Transactions on Cognitive and Developmental Systems, 2018, 11(1): 85-94.

[299] Bhardwaj A, Gupta A, Jain P, et al. Classification of human emotions from eeg signals using svm and lda classifiers[C]//2015 2nd International Conference on Signal Processing and Integrated Networks. IEEE, 2015: 180-185.

[300] Yang H, Laforge G, Stojanoski B, et al. Late positive complex in event-related potentials tracks memory signals when they are decision relevant[J]. Scientific Reports, 2019, 9(1): 9469.

[301] Pfurtscheller G, Da Silva F L. Event-related eeg/meg synchronization and desynchronization: Basic principles[J]. Clinical Neurophysiology, 1999, 110(11): 1842-1857.

[302] Moshfeghi Y, Jose J M. On cognition, emotion, and interaction aspects of search tasks with different search intentions[C]//Proceedings of the 22nd International Conference on World Wide Web. 2013: 931-942.

[303] Clough P, Sanderson M. Evaluating the performance of information retrieval systems using test collections[J]. Information Research, 2013, 18(2): 18-2.

[304] Xu D, Liu Y, Zhang M, et al. Incorporating revisiting behaviors into click models[C]// Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. 2012: 303-312.

[305] Lin C T, Wu R C, Jung T P, et al. Estimating driving performance based on eeg spectrum analysis[J]. EURASIP Journal on Advances in Signal Processing, 2005, 2005: 1-10.

[306] Czech A. Brain-computer interface use to control military weapons and tools[C]//Control, Computer Engineering and Neuroscience: Proceedings of IC Brain Computer Interface 2021. Springer, 2021: 196-204.

[307] Lotte F, Faller J, Guger C, et al. Combining bci with virtual reality: Towards new applications and improved bci[J]. Towards Practical Brain-Computer Interfaces: Bridging the Gap from Research to Real-World Applications, 2013: 197-220.

[308] Velasco-Álvarez F, Fernández-Rodríguez Á, Vizcaíno-Martín F J, et al. Brain–computer interface (bci) control of a virtual assistant in a smartphone to manage messaging applications[J]. Sensors, 2021, 21(11): 3716.

[309] Rainey S, Martin S, Christen A, et al. Brain recording, mind-reading, and neurotechnology: Ethical issues from consumer devices to brain-based speech decoding[J]. Science and Engineering Ethics, 2020, 26: 2295-2311.

补充内容

心理认知实验研究
受试者同意书

介绍

感谢您与我们共同完成此项科学实验，本实验以探讨大脑相关性认知为目的，以头皮脑电（EEG）及其相关的事件相关电位（任务所触发的脑电）技术为手段。脑电是一种使用电生理指标记录大脑活动的方法，它以完全无创的方式，记录大脑活动时的电波变化，是脑神经细胞的电生理活动在大脑皮层或头皮表面的总体反映。该技术不会引起任何生理心理的负面影响，您可以安心地参加实验。

这份知情同意书介绍了本研究的基本信息，为了确保您了解本研究，请认真阅读，并在最后签名，如有任何疑问，请询问主试。

研究目的

本研究试图考察心理行为特征。

研究过程

本研究您将在鼠标、键盘上完成按键反应。

风险，压力或不适

本研究不存在任何创伤性操作。

研究受益

人类的心理特征与生物学基础息息相关，您提供的数据将推进这一领域的研究，帮助我们更好的理解大脑对文本相关性判断的加工过程。研究设计不会使您直接受益。

自愿参与/退出

所有实验，受试者自愿参加，并且随时可以提出退出的要求，不会受到任何惩罚。

信息保密

如果您决定参加本研究，您参加本研究及在研究过程中的个人资料，将严格保密，所有的研究组成员，都要求对您的身份信息进行保密，可以识别您身份的信息，不会透露给研究组以外的成员。日后研究成果发表内容均为研究总体的平均结果，不会涉及您的个人信息。

联系方式

在研究中，如果您有任何与本研究有关的疑问，或者您认为在参加研究中受到了不公正对待，都可以随时向研究人员提出；如果在研究过程中有任何重要的新信息，可能会影响到您的继续参加的意愿时，研究人员会及时与您协商。

研究人员 邮箱：********** 电话：**********

当事人声明

本人在参加该实验前，已了解研究的内容且研究人员也已经解答所有提出的问题，并告知了潜在风险。

我_____，出生日期_____，自愿参加此项研究，我允许研究人员使用我在知情同意书中提到的信息。

签名_____ 日期 _____

研究人员姓名_____ 签名_____ 日期_____

图 6.1　脑电实验的知情同意书（隐去主试者和被试者信息）

致 谢

在这段漫长而充实的博士旅程中，我得到了许多人的支持和帮助。首先，感谢我的导师刘奕群教授。他在学术上给予我悉心指导，在课题的每一个阶段都给予了无私的支持，为我的研究扫清了许多障碍。他的远见和支持是我能够在这个交叉领域取得进展的重要保障，他的言传身教将使我终生受益。此外，感谢课题组的艾清遥副教授、张敏教授、马少平教授、金奕江老师和吴玥悦老师。他们的指导让我在学术道路上获得了更多的知识和见解，也让我在做人处事上受益匪浅。

我还要感谢课题组的毛佳昕老师和谢晓晖老师在我科研初期给予的耐心指导，帮我打下了坚实的基础。感谢课题组的孙筝老师和冯蕊老师在日常事务方面提供的帮助。感谢研究小组的陈雪松、吴之璟、朱书琦、张韶润、杜邦得、张开元、李佳祺、汪何希、Monika、王志红的合作，你们在实验设计、数据分析等方面的帮助和讨论是我能够取得科研进展的重要保障。感谢课题组的周雨佳、邵韵秋、李祥圣、陈佳、张帆、张瑞喆、储著敏、王晨阳、詹靖涛、刘布楼、马奕潇、方言、王亦凡、李海涛、苏炜航、涂奕腾、董骞、杨圣豪、陈海天、胡依然、王贝宁、李涵宇、汪佳茵、张潇宇、郭世圆、何祉瑜、庾源清、王哲凡、王畅越、陈俊杰、李心怡、郭志强等同学，你们对我的支持让我感受到了课题组大家庭的温暖。我还要特别感谢李佳玉同学对我的无私帮助和鼓励，与我一同迈向美好的未来。作为清华大学信息检索课题组的一份子，我感到非常荣幸。

在哥本哈根大学和阿姆斯特丹大学进行的合作研究期间，承蒙 Christina 教授、Tuukka 副教授和 Maarten 院士的热心指导与帮助，不胜感激。同样，我也非常高兴在访问期间认识了李秋池助理教授，以及张逸杰、Theresia、Pietro、黄瑾、李明、赵愉悦等同学，你们的友谊和合作让我在异国他乡感受到了家的温暖。

此外，感谢课题组外的高晓榕教授和王玉玲老师在跨学科领域交流中给予的指导。感谢陪伴我度过这段旅程的陈伟浩、蔚兆洋、李广普、于天宇、徐一支、郭铭浩、苏昊溪、李杨毅等朋友，一路上有你们，我才能坚持下来。

最后，我衷心感谢清华大学为我提供的优越平台，以及党和政府创造的良好环境，让我得以在学术道路上自由探索。同时，我也想感谢自己。在这段旅程中，我勇敢地追寻梦想，提升了自己独立分析和解决问题的能力、抗压抗挫的能力，以及快速适应变化的能力。我还要衷心地感谢我的父母和家人给予我的强大的后盾，你们的支持和理解是我坚持不懈的最大动力。感谢所有在这段旅程中给予我帮助和支持的人，谢谢你们！

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：叶子逸 日 期：2025.5.15

个人简历、在学期间完成的相关学术成果

个人简历

1998 年 03 月 05 日出生于福建省厦门市。

2016 年 9 月考入清华大学新雅书院（计算机科学与技术方向），2020 年 7 月本科毕业并获得工学学士学位。

2020 年 9 月免试进入清华大学计算机科学与技术系攻读博士至今。

在学期间完成的相关学术成果

学术论文：

[1] ***Ziyi Ye***, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Min Zhang, Christina Lioma, and Tuukka Ruotsalo. *Generative Language Reconstruction from Brain Recordings*. Nature Communications Biology, 8, 346 (2025). (**SCI 一区**) *[链接]*.

[2] ***Ziyi Ye***, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. *Brain Topography Adaptive Satisfaction Modeling for Interactive Information Access*. Proceedings of the 30th ACM International Conference on Multimedia, 90-100. (**CCF-A**, Full paper) *[链接]*.

[3] ***Ziyi Ye***, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuancheng Li, Jiaji Li, Xuesong Chen, Min Zhang, and Shaoping Ma. *Why Don't You Click: Understanding Non-click Results in Web Search with Brain Signals*. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 633-645. (**CCF-A**, Full paper) *[链接]*.

[4] ***Ziyi Ye***, Jingtao Zhan, Qingyao Ai, Yiqun Liu, Maarten de Rijke, Christina Lioma, and Tuukka Ruotsalo. *Query Augmentation with Brain Signals*. Proceedings of the 32nd ACM International Conference on Multimedia, 7561-7570. (**CCF-A**, Full paper) *[链接]*.

[5] ***Ziyi Ye***, Xiaohui Xie, Yiqun Liu, Zhihong Wang, Xuesong Chen, Min Zhang, and Shaoping Ma. *Towards a Better Understanding of Human Reading Comprehension with Brain Signals*. Proceedings of the ACM Web Conference 2022, 380-391. (**CCF-A**, Full paper) *[链接]*.

[6] ***Ziyi Ye***, Xiaohui Xie, Qingyao Ai, Yiqun Liu, Zhihong Wang, Min Zhang, and Shaoping Ma. *Relevance Feedback with Brain Signals*. ACM Transactions on Information Systems 42 (4), 1-37 (2024). (**CCF-A**, Journal paper) *[链接]*.

[7] ***Ziyi Ye***, Qingyao Ai, and Yiqun Liu. *Brain-Computer Interface Meets Informa-*

tion Retrieval: Perspective on Next-generation Information System. Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding, 61-65. (**CCF-A**, Workshop paper) *[链接]*.

[8] ***Ziyi Ye***, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. *Learning LLM-as-a-Judge for Preference Alignment*. The Thirteenth International Conference on Learning Representations. (**TH-CPL-A**, Full paper) *[链接]*.

[9] ***Ziyi Ye***, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. *Investigating COVID-19-Related Query Logs of Chinese Search Engine Users*. Proceedings of the Association for Information Science and Technology, 57(1), e424 (2020). (Poster paper). *[链接]*.

[10] Xuesong Chen, ***Ziyi Ye***, Xiaohui Xie, Yiqun Liu, Xiaorong Gao, Weihang Su, Shuqi Zhu, Yike Sun, Min Zhang, and Shaoping Ma. *Web search via an efficient and effective brain-machine interface*. Proceedings of the fifteenth ACM international conference on web search and data mining, 1569-1572. (**TH-CPL-A**, Demo paper). *[链接]*.

[11] Jia Chen, Jiaxin Mao, Yiqun Liu, ***Ziyi Ye***, Weizhi Ma, Chao Wang, Min Zhang, and Shaoping Ma. *A Hybrid Framework for Session Context Modeling*. ACM Transactions on Information Systems 39 (3), 1-35. (**CCF-A**, Journal paper). *[链接]*.

[12] Shaorun Zhang, Zhiyu He, ***Ziyi Ye***, Peijie Sun, Qingyao Ai, Min Zhang, and Yiqun Liu. *EEG-SVRec: An EEG Dataset with User Multidimensional Affective Engagement Labels in Short Video Recommendation*. Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 698-708. (**CCF-A**, Resource paper). *[链接]*.

专利：

[13] 刘奕群，陈雪松，谢晓晖，**叶子逸**，朱书琦，张敏，马少平。网页搜索方法及装置、电子设备和存储介质: 中国，CN114020158B，2023-07-25。

[14] Yiqun Liu, Xuesong Chen, Xiaohui Xie, **Ziyi Ye**, Shuqi Zhu, Weihang Su, Xiaorong Gao, Yike Sun, Min Zhang, Shaoping Ma. Method, apparatus, electronic device, and storage medium for web search: 美国, US20230169130A1.

指导教师评语

信息检索对于信息化社会充分利用信息、提升信息获取效率具有重要意义，信息检索作为一种人类获取信息的方式，如何与人类用户进行交互的问题始终是其中核心的研究方向之一。本论文创新性的研究了脑机接口技术在信息检索交互过程中可能发挥作用的路径，并对这种路径与传统信息检索交互方式的差别进行了比较。

论文主要针对如何利用脑机接口更好的理解人类用户的信息需求，以及获得人类用户对于搜索结果的满意度及相关性反馈开展了研究，验证了直接利用脑电信号进行信息检索交互的可能性，取得了一系列开创性的研究成果，在信息与生物医学交叉学科领域进行了探索，并构建了初步的验证系统。

在完成论文工作的过程中，作者体现了较强的理论研究和实践创新能力。论文取得了较突出的创新成果，结构组织较为严谨，内容较为详实全面，是一篇优秀的博士论文。

答辩委员会决议书

该论文研究了基于脑机接口的信息检索技术，选题具有重要的理论意义和应用价值。

论文取得的主要创新成果如下：

（1）针对脑机信号的时空特性，分别提出了一种基于脑拓扑结构的用户满意度检测模型和一种细粒度用户阅读过程理解模型，对信息获取过程中的脑认知机制进行了系统地分析和研究。

（2）提出了一种生成式的脑信号解码方法，实现了基于脑机接口的开放词表语言内容生成，显著改进了信息检索系统查询扩展的性能。

（3）针对事实性搜索和非事实性搜索的场景，提出了基于脑机接口的用户意图建模方法，在信息检索相关性反馈任务上性能取得显著提高。

论文工作表明，该同学已经掌握了本学科坚实全面的基础理论和系统深入的专门知识，独立从事学术研究的能力强。论文结构完整、层次清晰、写作规范，达到了工学博士学位论文要求，是一篇优秀的博士学位论文。答辩过程中，阐述清楚，回答问题正确。

经答辩委员会无记名表决，一致同意通过论文答辩，并建议授予叶子逸工学博士学位。