Evaluating the Usefulness of Multiple Functionalities for Filtering Topics within Neuroscience Literature Exploration

Yuxuan Zhuang (5273773)
Utrecht University
Utrecht, The Netherlands

Abstract

Neuroscientists need to analyze a vast amount of literature to find potentially fruitful experiments. Topic-based literature exploration is a useful means to analyze many publications simultaneously because it provides an overview of the relations between topics: the co-occurrence of brain regions and brain diseases topics within the same sentences of a publication often implies a relation between them. In order to incorporate DatAR into the daily workflow of neuroscience literature exploration, multiple functionalities should be provided to complete tasks with full scope. Neuroscientists have indicated that filtering topics is useful, involving at least two rounds and multitasking, including identifying, comparing, and verifying identified relations. Therefore, our goal is to investigate the extent to which multiple functionalities are useful for filtering topics.

We follow a user-centered design approach. We (i) identify the user tasks of filtering topics is indeed useful with neuroscientists; (ii) identify the representative tasks for filtering topics; (iii) design the evaluation approach. The evaluation is first conducted using the initial version of DatAR (V1), followed by the iteration of multiple functionalities based on the user requirements collected from the first round of evaluation. The developed compare version of DatAR (V2) introduces five additional aspects, which are intended to improve the accuracy and sufficiency of the identified relations provided by multiple functionalities. After conducting the second round of evaluation based on V2 with six neuroscientists, the results indicate valuable insights addressing our original goal of investigating the usefulness of multiple functionalities for filtering topics. Specifically, the findings show that (i) the identified relations provided by multiple functionalities are useful for filtering topics, directly supporting our goal; (ii) the usefulness of multiple functionalities could be enhanced by providing more sufficient, latest, and closely related topics, further aligning with the goal of maximizing filtering effectiveness; and (iii) enabling the recording and comparison of differences in identified relations across multiple rounds of topic filtering, improving the visualization of identified relations in complex multitasking scenarios, and offering guidance with sufficient detailed information for filtered topics could support the integration of multiple functionalities into the daily workflow of neuroscience research. These results collectively validate and extend our goal by pinpointing specific improvements that could facilitate effective topic filtering.

1 Introduction

Neuroscientists need to identify potentially fruitful experiments by exploring a vast amount of literature. They first need to explore the relations between, such as brain diseases and brain regions. Topic-based literature exploration [2] is a useful means to help neuroscientists obtain an overview of relations between brain related topics without the need of reading individual publications one by one. For example, neuroscientists were able to use the direct relation between topics, which indicate that these two brain topics exist in the same sentence source from literature, to explore which brain regions affect which brain disease and vice versa, such as the brain region *cerebellar vermis* and the brain disease *bipolar disorder* (see Fig. 1) [20].

Previous work has developed different functionalities in DatAR prototype, an immersive analytics(IA) system [9] to assist neuroscientists in finding relations between brain diseases and brain regions [21, 27]. Each functionalities support a specific aspect of literature exploration task. In this study, we identify eight functionalities for multiple functionalities integration from previous DatAR's works [19, 20, 26] to complete tasks with full scope and objectives (see Section 2.2). Neuroscientists can use the multiple functionalities to explore the relations between brain regions and brain diseases topics within neuroscience literature exploration. The multiple functionalities integration and interaction are intend to discover areas that require additional consideration or improvements within neuroscience literature exploration, making it more meaningful to find potentially fruitful experiments.

The motivation for our study originates from the expressed desire of junior neuroscientists in incorporating DatAR into daily neuroscience literature exploration workflow, based on focus group discussions held in December 2023. Neuroscientists described the user task of filtering topics within neuroscience literature exploration, beginning with selecting a topic of interest, followed by finding more related topics based on identified relations, and concluding with the final validation of the selected topics (see Section 3.1). After at least two rounds of topic filtering process, neuroscientists could ensure that their costly experiments are likely to contribute to the literature. We use multiple functionalities to support the topics filtering tasks from identifying the relations, comparing the relations between brain related topics and verifying the displayed relations (see Section 2.1). Evaluating the usefulness of multiple functionalities could provide us with further insights on how to integrate DatAR into neuroscience workflows in the future. We define 'Usefulness' as the extent to which neuroscientists think that using multiple functionalities could aid neuroscience literature exploration [23]. The research question of our study is: ***RQ: To what extent are the multiple functionalities for filtering topics useful within neuroscience literature exploration?***

We follow a user-centered design approach, collaborating closely with neuroscientists at every step of the process. First we need to confirm that the user task of filtering topics is indeed useful within neuroscience literature exploration (see Section 3.1.1); then identify evaluation tasks in filtering topics (see Section 3.1.2). Five evaluation tasks were identified from previous DatAR's studies [17, 19, 27] as each tasks representing one or several aspects of the topics filtering process. We then design an evaluation method to assess the usefulness of multiple functionalities. Representative tasks are needed for the evaluation. By conducting the semi-structured interviews with two senior neuroscientists, they provided one or two representative tasks for each evaluation task based on their domain knowledge (see Section 3.1.3). In order to gain insight into the process of filtering topics, two rounds of evaluation are adopted (see Section 3.2). The first round of evaluation is conducted based on V1 (see Section 3.2.1) to collect the usefulness feedback from neuroscientists through both qualitative interview questions and quantitative questionnaires. Based on the feedback collected from the initial evaluation, we identify five user requirements (URs) including adding additional infomation that intend to enhance the accuracy, sufficiency of the results provided by multiple functionalities. These URs are applied to multiple functionalities and incorporated into the compare version of DatAR (V2). V2 is then utilized in the second round of evaluation. The results of two rounds of evaluation are serve as a whole to provide insight into the usefulness of multiple functionalities within the user tasks of filtering topics (see Section 3.2.2).

Section 4 describes the data source and development environment for two versions of DatAR. Section 6 presents the results collected from two rounds of evaluations (see Section 5). Section 7 summarizes the findings from our analysis, outlining the discussion points obtained from both rounds of evaluation. Section 8 summarizes the findings of our research and highlights the potential for further research.

2 Related Work

We first explore the usefulness of filtering topics within neuroscience literature exploration (see Section 2.1). Then describe the existing DatAR prototype and its functionalities in Section 2.2. The evaluation in this study will be limited to eight functionalities integration. Section 2.3 then discusses the TAM model and related metrics. Six identified metrics will be used to evaluate the usefulness of multiple functionalities in the evaluation procedure.

2.1 Filtering Topics within Literature Exploration

Literature exploration has been divided into 5-6 different stages, from selecting topics, filtering topics to verifying topics [8]. The stage of filtering topics has been further investigated, after selecting topics, the main user task in the second stage includes conducting a comprehensive search of topics, followed by filtering and excluding topics based on specific criteria, or finding and verifying the relationships between topics [7].

The process of filtering topics involves the substages of preliminary selection and topic identification, in-depth analysis and topic refinement, and final validation of the selected topics [5]. Researchers could already have 3-5 initial topics of interest derived from their previous studies, so they typically starting from finding more related research topics by identifying existing relations between the initial topics. Researchers then use traditional literature exploration tools to further verify the identified relations, including going through the literature source to explore the experimental methods and results provided by the displayed relations [15]. When filtering topics, researchers are constantly adding new topics from the literature and filtering again based on the added topics to identify useful and up-to-date research topics. The process of filtering topics is usually repeated for at least two rounds to ensure that the filtered topics are useful for finding potentially fruitful experiments [5, 15]. The identified process of filtering topics and its necessity for researchers within literature exploration form the basis for evaluating the usefulness of multiple functionalities for filtering topics in our study

2.2 Existing DatAR Prototype

Neoroscientists are able to explore brain-related topics through a number of functionalities and visualisations implemented as widgets [21]. We identified eight functionalities from previous DatAR's works for multiple functionalities integration [19, 20, 26] (as shown in Figures 1 and 2). Each functionality support a specific aspect of neuroscience literature exploration task. The detailed descriptions and literature source of each functionality are as follows:

- **Brain Regions Visualisation** (Figure 1, Label 1): This widget displays 274 brain regions in the form of 3D AR spheres, with the position of each sphere determined by the 3D coordinates of the brain region [22]. Researchers can see highlighted 3D brain region spheres, which represent the identified direct or indirect relations associated with a specific brain disease topic.

- **Direct Relation Explorer** (Figure 1 and 2, Label 2): This widget performs a search for direct relations between either one brain region and various brain diseases or one brain disease and several brain regions within the PubMed database [22]. Researchers can place brain-related topics and specific intermediate topics they want to explore, and then connect this widget with Brain Regions Visualisation or Brain Disease Topic Model. The visualization of these direct relations is depicted with pink spheres in a 3D AR environment.
- **Max-Min Cooccurrences Filter Widget** (Figure 1, Label 3): This widget could filter out a select number of co-occurrences. Researchers can connect this widget with Direct Relation Explorer and represent these infrequently mentioned relations as yellow spheres in 3D AR environment [22].
- **Sentences Extractor Widget** (Figure 1, Label 4): This widget can query and display the sentences source in the title or abstract of the identified relations [22]. Researchers can go back to the publication to read what was said about these two topics to further verify the displayed relations.
- **Indirect Relations Querier** (Figure 2, Label 5): This widget performs a search for unknown indirect relations using gene, protein or mental process as intermediate topics. Indirect relations suggest that while Topic A may not directly co-occur with Topic B in literature, a relation between A and B could be existing based on the relations of one or more brain topics involving both Topic A and Topic B [26]. Researchers can place brain-related topics and specific intermediate topics they want to explore, and then connect this widget with Brain Regions Visualisation or Brain Disease Topic Model. The visualization of these direct relations is depicted with green spheres in a 3D AR environment.
- **Brain Disease Topic Model** (Figure 2, Label 6): This widget can display a topic model of brain diseases, where diseases with similar semantic properties are displayed adjacently in 3D AR [22]. The distance between any two diseases indicates their similarity based on all co-occurrences between all topics. Researchers can see highlighted 3D brain disease spheres, which represent the identified direct or indirect relations associated with a specific brain region topic.
- **Resource Sphere Inspector Widget** (Figure 2, Label 7): This widget can display all brain related topics from data source [22]. Researchers can extract all resource spheres that represent the brain regions or brain diseases topics.
- **Comparison Widget** (Figure 2, Label 8): This widget can compare the identified relations between two brain related topics [19]. Researchers can identify similar and different brain regions affected by different brain diseases, or similar and different brain diseases that affect in different brain regions.

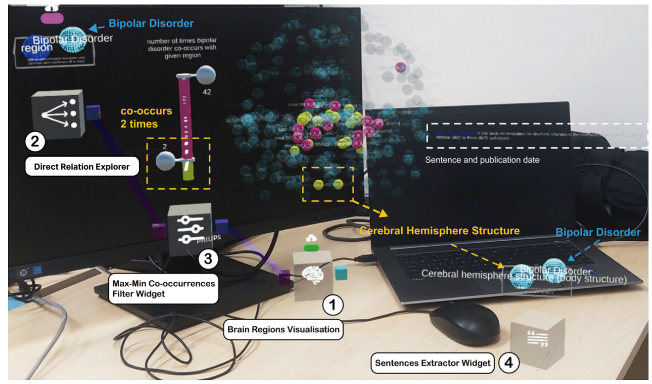

Figure 1. Widgets in the DatAR prototype 1) Brain Regions Visualisation 2) Direct Relation Explorer 3) Max-Min Cooccurrences Filter Widget 4) Sentences Extractor Widget

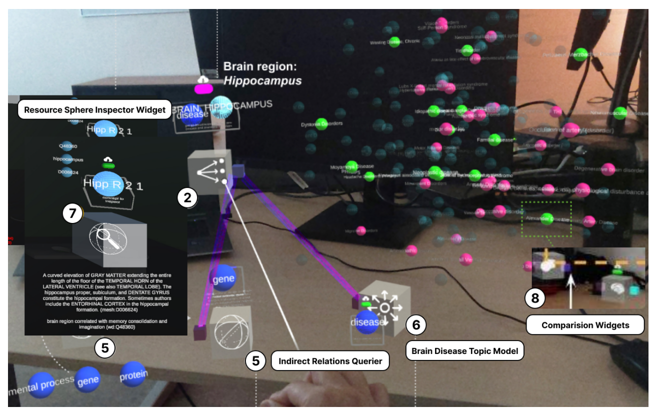

Figure 2. Widgets in the DatAR prototype 5) Indirect Relations Querier 6) Brain Disease Topic Model 7) Resource Sphere Inspector Widget 8) Comparison Widget

2.3 Evaluating the Usefulness for Filtering Topics: Model and Metrics

Usefulness evaluation is used to explore the interactions between technology and users in the field of Human-Computer Interaction (HCI) empirical research. Usefulness measures the actual value of the technology to users through metrics such as ease of use, functionality, and usability. The evaluation of usefulness is a core aspect of assessing the impact of information systems [24].

While previous DatAR works [19, 26] provided some insights on the usefulness of functionalities based on one or two user tasks, they have limited the scope of understanding multiple functionalities perform across a wider range of neuroscience literature exploration tasks. In contrast, recent work integrated comparison and co-occurrence exploration functionalities within the DatAR prototype, allowing for a broader assessment. Neuroscientists and visualization experts were invited to assess the visualization, navigation, and performance of multiple functionalities based on user tasks to find and compare known direct relations. The findings suggested that multiple functionalities were effective in helping neuroscientists find potentially fruitful experiments through identifying and comparing the relations between topics, but further research is necessary to refine and validate these results across a more diverse set of use cases.

Our research focuses on evaluating the usefulness of multiple functionalities for filtering topics. Filtering topics requires multiple rounds of utilizing multiple functionalities based on the initial version of DatAR (V1) to conduct multiple user tasks involving the exploration, comparison, and validation of topics. Unlike the single evaluation of one user task in previous works, we need to further consider the evaluation models and metrics:

Usefulness evaluation metrics from TAM: Perceived Usefulness (PU) The Technology Acceptance Model (TAM) was developed by Davis (1989) [3]. It is a framework for evaluating the usefulness of information systems. The model suggests that users' acceptance and use of technology are influenced by Perceived Ease of Use (PEoU) and Perceived Usefulness (PU). PU measures the extent to which users believe that using a particular system/technology will enhance their job performance. The effectiveness of using PU to measure the extent to which users believe that using a specific technology will improve their work performance has already been validated in previous studies on technologies specific to the marketing and education fields [6, 10, 12, 16]. In our study, Perceived Usefulness (PU) can be used to evaluate how beneficial neuroscientists perceive the representative tasks conducted by using multiple functionalities to find fruitful experiments.

Usefulness evaluation metrics from Extended TAM: Result Demonstrability (RD), Job Relevance (JR), and Output Quality (OQ) The extended versions of TAM, namely TAM2 (Wu, 2011) [25] and TAM3 (Faqih, 2015) [4], incorporate metrics such as Result Demonstrability (RD), Job Relevance (JR), and Output Quality (OQ). These metrics have been applied to Augmented Reality (AR) applications [6, 10, 12, 16]. The validity of these metrics was confirmed in evaluations of AR applications related to education by Pribeanu (2012) [13] and Balog (2009) [1]. In our study, Result Demonstrability (RD) can be used to evaluate the clarity of the results provided by multiple functionalities for both the users and other neuroscientists. Job Relevance (JR) can be used to assess the relevance of the result provided by multiple functionalities to neuroscientists' future research. Output Quality (OQ) can be used to evaluates the quality of results produced by multiple functionalities, such as clarity, understandability, and accuracy of the identified relations.

Usefulness evaluation metrics from other sources: Perceived Informativeness (PI), Perceived Enjoyment (PE) Perceived Informativeness (PI) could provide insight in long-term acceptance and perceived usefulness of a technology [14]. This metric could reflect the extent to which DatAR integration into the daily neuroscience research workflow is beneficial in our study. Perceived Enjoyment (PE) could also influence acceptance and perceived usefulness of technology, which has been used to provide insights into the usefulness of new technologies in the retail industry [14]. In our study, Perceived Informativeness (PI) can be used to assesses how informative multiple functionalities are perceived to be, specifically in providing sufficient and relevant information to find fruitful experiments. Perceived Enjoyment (PE) can be used to measures how enjoyable the use of multiple functionalities is perceived to be.

3 Method

We follow a user-centered design approach of problem discovery to design evaluation approach for an identified user task. Neuroscientists were consulted at every stage of our work. We identified a useful user task of filtering topics (see Section 3.1.1). We then identify five evaluation tasks involved in the topic filtering process (see Section 3.1.2). These five evaluation tasks represent different sub-aspects of topic filtering process. We then identify one or two representative tasks corresponding to each of the evaluation tasks (see Section 3.1.3). These representative tasks are conducted by neuroscientists during the evaluation of multiple functionalities.

Building on the user-centered design approach, we seek to obtain feedback from neuroscientists through two rounds of evaluation (see Fig. 3), which aim to match the process of multiple rounds of topic filtering process in real-world research workflow. We use different version of DatAR prototype in two rounds of evaluation.

After conducting the first round of evaluation based on initial version of DatAR(V1) (see Section 3.2.1) to gather feedback on improvements from neuroscientists, we developed a compare version of DatAR (V2) (see Section 3.2.2) based on the user requirements collected during the first round. These additional aspects including adding publication date, providing specific intermediate topics, etc. applied in V2 were intend to enhance the accuracy and sufficiency of identified relations provided by multiple functionalities to better support the user task of filtering topics. Comparing the results from the two versions of DatAR will offer us a deeper understanding of the usefulness of multiple functionalities.

We design a questionnaire to evaluate the usefulness of multiple functionalities. Responses to the 55 questions are rated on a 7-point Likert scale. The evaluation is centered on six metrics from the TAM model [3]. We also conduct semi-structured interviews to collect participants' qualitative insights into the usefulness of multiple functionalities in the two versions of multiple functionalities (see Section 3.2).

3.1 Identifying User Task and Representative Evaluation Tasks

We first explain why the user task of filtering topics is useful (see Section 3.1.1). We then identify five evaluation tasks which represent different sub-aspects of filtering topics, ranging from identifying relations to verifying the identified relations, in collaboration with two neuroscientists (see Section 3.1.2). We determine one or two representative tasks for each evaluation task for filtering topics (see Section 3.1.3).

3.1.1 Assessing the User Task of Filtering Topics. In order to assess whether the user task of filtering topics is useful within neuroscience literature exploration, we interviewed three junior neuroscientists (P1, P2, P3) and two senior neuroscientists (P4, P5). We first explore the general process of neuroscience literature exploration, we conducted a focus group discussion around two questions: *"When exploring neuroscience literature, what do you typically do, and what steps are involved in this process?"*

Participants (P1, P2, P3) indicated that they normally start with 3 to 5 existing research topics of interest. They (P1, P2, P3) are more likely to focus on the process of finding more relevant topics and filtering these identifed topics based on their domain knowledge and literature source. Two participants (P2, P4) detailed their topic filtering process, where approximately 50 relevant topics are initially selected from a vast amount of literature. They then identified, compared, and verified the relations between topics to determine fewer than 10 topics that are useful for finding potentially fruitful experiments. The filtering topics process described by two participants (P2, P4) indicated that the user task of filtering topics could be useful within neuroscience literature exploration.

We then further discussed the specific process of filtering topics with the same participants: *"What do you typically do when filtering topics?"* and *"Does the process of filtering topics need to be repeated, and how long does it usually take?"* These questions aimed to confirm whether filtering topics is useful within neuroscience literature exploration.

Four of five participants (P2-P5) indicated that the user tasks of filtering topics ensure that their costly experiments are likely to contribute to the literature. They (P2,P4, P5) suggested that multiple rounds of filtering topics were required to find useful topics and relations for further research. For example, one participant (P4) focusing on *Alzheimer's Disease* mentioned that medications such as *Rivastigmine* were commonly used to target directly associated brain regions, such as the *temporal lobe*, thereby slowing the progression of language degeneration in patients. He/She (P4) believed that filtering more related brain region topics related to *Alzheimer's Disease* could greatly aid subsequent experiments and clinical treatments. *Alzheimer's Disease* has also been verified to closely connected to *Vascular Diseases* with increasing age in the previous research (P5). Participants (P4, P5) hoped to filter out brain region topics that are either commonly affected by or differ between these two diseases, making them a starting point for further experiments.

3.1.2 Identifying Evaluation Tasks in Filtering Topics. The user task of filtering topics involves different aspects, from identifying and comparing relations to verifying identified relations (see Section 3.1.1). To gain insights into the usefulness of multiple functionalities for filtering topics, we need to identify several evaluation tasks that can represent different aspects of topic filtering process. Each evaluation task can address one or more aspects of the filtering process. For example, the evaluation tasks of identifying direct relations could represent the process of finding around 50 related topics in filtering topics. The combination of these evaluation tasks should be considered representing a complete topic filtering process.

In order to identify the evaluation tasks in filtering topics, we fist reviewed related literature from previous DatAR's works [18, 20–22, 27]. Seven evaluation tasks that could be useful for filtering topics have been identified and listed in Table 1.

After identifying seven potentially evaluation tasks in Table 1, we investigated whether these evaluation tasks represent one or several aspects in filtering topics. We conducted semi-structured interviews with two senior neuroscientists (P4, P5). They were asked to determine whether these evaluation tasks were necessary to represent a certain aspect of filtering topics, based on their domain knowledge and previous experience. The questions in the semi-structured interviews included: *"To what extent do you think this evaluation task is necessary in the topic filtering process?"* and *"If this evaluation task is necessary, which aspects of the topic filtering process does it represent?"* Five out of seven evaluation tasks have been identified from Table 1. The background information and detailed description of each identified evaluation task are as follows:

Finding known direct relations. (Table 1 No.5 and Table 2 No.1) This evaluation task was originally derived from neuroscientist Cunqing Huangfu. Direct relations indicate that Brain Topics A and B co-occur in the same sentence sources. All neuroscientists (P4, P5) confirmed the necessity of finding known direct relations for filtering topics. They indicated that identified direct relations could enable them to find more related topics from the initial topics of interest.

Finding unknown indirect relations. (Table 1 No.6 and Table 2 No.2) Indirect relations mean that for some

Table 1. Potentially useful evaluation tasks for filtering topics from previous works.

No.	Evaluation Task	Representative Task from Previous Works	Source
1	Delving deeper into information about brain topics	Selecting an interesting brain topic, examining its detailed descriptions and encapsulated themes.	[20]
2	Locating relevant diseases in the topic model when exploring brain topics	Searching for diseases related to the amygdala, suggesting that semantically similar brain diseases may affect similar brain regions.	[18, 20]
3	Identifying brain regions often mentioned with specific brain diseases	Finding the most frequently mentioned brain regions in discussions of depression/bipolar disorder.	[20–22, 27]
4	Comparing two diseases in terms of brain-related topics	Comparing the most frequently mentioned brain regions in discussions of depression and anxiety, determining the amygdala as an interesting brain region for further research.	[18, 21]
5	Finding known direct relations	Searching for brain diseases that co-occur with the amygdala more than 400 times in the literature. Finding brain regions that co-occur less than 2 times with bipolar disorder in the literature.	[18, 20, 22, 27]
6	Finding unknown indirect relations	Selecting genes as an intermediary topic to find indirect relations between bipolar disorder and the cerebellar vermis. Selecting genes as an intermediary topic to find indirect relations between the hippocampus and Alexander disease. Selecting psychological processes as an intermediary topic to find indirect relations between bipolar disorder and the cerebellar vermis.	[18, 20, 27]
7	Verifying the displayed relations between brain regions and brain diseases	Querying literature sources for co-occurrence of specific brain regions and brain diseases (e.g., sentences showing the relation between a related disease and the amygdala), accessing sentences from original documents, and assessing the contribution to the relation positively or negatively.	[18, 20, 22, 27]

brain Topic A, there is no direct co-occurrence with brain Topic B in the literature. However, by identifying one or more brain topics that co-occur with both A and B, a relation between A and B can be inferred. Boyu's work [27] verified that indirect relations could be useful for filtering topics, as rare mentioned indirect topics may reveal a new research domain. One neuroscientist (P5) agreed on this evaluation task could aid for filtering topics. He/She (P5) believed that revealing unknown relations between existing topics could assist in finding more potentially useful topics.

Comparing the direct/indirect relations of two diseases in terms of brain-related topics. (Table 1 No.4 and Table 2 No.3) Comparison functionalities could be used to compare the known direct relations associated with two brain diseases. Identifying brain regions affected by both brain diseases is crucial for understanding the relations between various brain diseases. All neuroscientists (P4, P5) indicated that extending the evaluation task to compare the unknown indirect relations of two brain diseases is meaningful for filtering topics. The comparison between two brain related topics could assist in adding new topics or narrowing down the existing topics for topic filtering process.

Comparing the direct/indirect relations of two brain regions in terms of disease-related topics. (Table 1 No.4 and Table 2 No.4) Brain Disease Topic model provided neuroscientists with a brief summary indicating which brain diseases are semantically similar. However, it did not explain why these similarities existed. Ghazaleh's work supported comparison of two brain regions to provide region-disease-related information. P5 agreed on the necessity of this evaluation task for filtering topics and also hoped to extend the evaluation task to the indirect relations of two brain regions.

Verifying the displayed relations between brain regions and brain diseases. (Table 1 No.7 and Table 2 No.5) Neuroscientists (P4, P5) wanted to verify whether the displayed relations from previous tasks are useful for their future research. Troost's work [22] supported researchers in using sentences extractor to find sentence sources for co-occurring relations. Both neuroscientists (P4, P5) considered this evaluation task crucial for filtering topics. The detailed descriptions of the identified relations could help them determine whether the filtered topics provide relevant and effective information for future research.

3.1.3 Determining Representative Tasks for Filtering Topics.

After identifying five evaluation tasks in filtering topics in Section 3.1.2, to allow participants to evaluate the usefulness of multiple functionalities for filtering topics, we need to determine representative tasks that could be conducted during the evaluation procedure (see Section 3.2). We discussed with two of senior neuroscientists (P4, P5) to find one or two representative tasks that would be suitable for each evaluation task.

The representative tasks for first evaluation task aimed to find direct relations between *temporal lobe* and *Alzheimer's Disease*. Neuroscientist (P4) believed that finding brain diseases that co-occur more than 80 times with *temporal lobe* in the literature can help further explore potential preventive and therapeutic strategies (see Section 3.1.1), as shown in Fig 5. Additionally, the neurologist (P5) pointed out that *Vascular Diseases* could serve as another example of the representative task, provided for exploration by neuroscientists in different fields. These two representatibe tasks were suitable for participants to conduct in the evaluation procedure.

Neuroscientists (P4, P5) could also find that *Alzheimer's Disease* not only has direct relations with the *temporal lobe* but also possesses unknown indirect relations with *parietal lobe*. They hoped to understand which other brain regions have indirect relations with *Alzheimer's Disease*. Therefore, using genes such as MAPT as intermediary topics to explore the relations between *Alzheimer's Disease* or *Vascular Diseases* and brain regions could serve as the representative task for the evaluation task of finding unknown indirect relations (see Fig 4).

Neuroscientists (P4, P5) were aware that the *parietal lobe* and *temporal lobe* jointly contribute to both *Alzheimer's Disease* and *Vascular Diseases*. Neuroscientist (P4) wished to explore how they might jointly contribute to the same pathological conditions. More related topics could pave new paths for drug development targeting these two diseases. Thus, comparing brain regions commonly affected by these two diseases, as well as other diseases affected by these two brain regions, could serve as two representative tasks for the evaluation tasks of comparing direct/indirect relations (see Fig 6 and 8).

Finally, to verify that the identified relations are truly useful, neuroscientists (P4, P5) need to evaluate sources of sentences. A suitable representative task was verifying the relations between *Alzheimer's Disease* and *corpus striatum structure*.

3.2 Designing Evaluation Process for filtering topics

We conduct two rounds of evaluation [11] to match the process of filtering topics and achieve an in-depth understanding of the usefulness of multiple functionalities (RQ). In the first round of evaluation, a group of neuroscientists are invited to conduct the representative tasks using multiple functionalities based on V1 (see Section 3.2.1). After iteration of the multiple functionalities based on the user requirements from first round (see Section 6.1), the same group of neuroscientists are asked to re-evaluate the usefulness of multiple functionalities based on V2 (see Section 3.2.2). Figure 3 shows the entire evaluation process of multiple functionalities. The detail descriptions can be found in subsections.

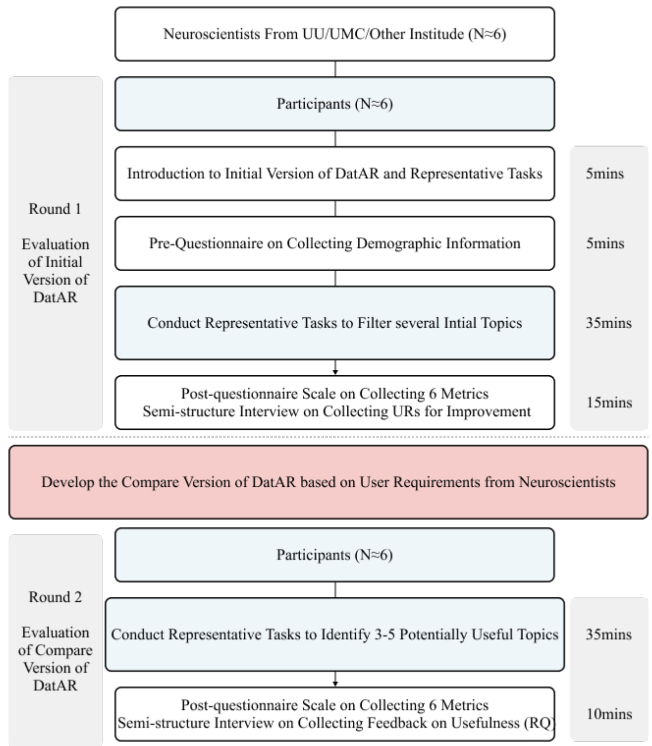

Figure 3. Evaluation and iteration process of multiple functionalities for filtering topics

3.2.1 Evaluation of Multiple Functionalities based on Initial Version of DatAR (V1).

The first round of evaluation aims to collect potential improvements on multiple functionalities and gather initial insight into usefulness of multiple functionalities for filtering topics. We first identify the scope of multiple functionalities integration in our study. We use eight functionalities integration (see Section 2.2) from previous studies to develop the initial version of DatAR (V1). These eight functionalities enable neuroscientists to conduct the representative tasks (see Table 2) for filtering topics in the first round of evaluation.

We use both quantitative and qualitative methods to gain a comprehensive understanding of the usefulness of multiple functionalities (RQ). We employ a survey with a 7-point Likert scale to provide measurable and comparable data allowing us to statistically assess participants' responses. The questions in the questionnaire are set up corresponding to the representative tasks based on six metrics: **Output Quality**

Table 2. Useful representative tasks for filtering topic based on the interview with two neuroscientists. These representative tasks are conducted by participants in the multiple rounds of evaluation procedures.

No.	Evaluation Task	Representative Task Provided by Two Neuroscientists	Source
1	Finding known direct relations.	Finding brain diseases that co-occur more than 80 times with temporal lobe in the literature. Finding brain regions that co-occur more than 30 times with Vascular Diseases in the literature.	[18, 20–22, 27], P4, P5
2	Finding unknown indirect relations.	Selecting gene as an intermediary topic to find indirect relations between Alzheimer's Disease and the brain regions. Selecting gene as an intermediary topic to find indirect relations between Vascular Diseases and the brain regions.	[18, 20, 27], P5
3	Comparing the direct/indirect relations of two diseases in terms of brain-related topics.	Comparing the most frequently mentioned brain regions in discussions of Vascular Diseases and Alzheimer's Disease, determining interesting brain regions for further research. Then selecting gene as an intermediary topic to find more brain regions that have indirect relation with brain diseases.	[18, 21], P4, P5
4	Comparing the direct/indirect relations of two brain regions in terms of disease-related topics.	Comparing the most frequently mentioned brain diseases in discussions of temporal lobe and parietal lobe, determining an interesting brain disease for further research. Then selecting gene as an intermediary topic to find more brain diseases that have indirect relation with brain regions.	P4, P5
5	Verifying the displayed relations between brain regions and brain diseases.	Querying literature sources for co-occurrence of corpus striatum structure and Alzheimer's Disease, accessing sentences from original documents, and assessing the contribution to the relation positively or negatively. Querying literature sources for co-occurrence of nucleus accumbens and Alzheimer's Disease, accessing sentences from original documents, and assessing the contribution to the relation positively or negatively.	[18, 20, 22, 27], P5

(OQ), **Perceived Informativeness (PI)**, **Result Demonstrability (RD)**, **Perceived Usefulness (PU)**, **Perceived Enjoyment (PE)**, **Job Relevance (JR)** (see Section 2.3).

For the qualitative method, we conduct a semi-structured interview to collect the subjective feedback on **Usefulness** at the end of the first evaluation. Usefulness feedback are intended to answer to what extent neuroscientists think that using multiple functionalities could aid filtering topics [23]. The feedback provided by neuroscientists are recorded and use for the iteration of multiple functionalities in the next phase. We also collect the subjective feedback on**Functionality**, and **Visualization** of multiple functionalities to provide additional insights on the multiple functionalities from usabilty, visualization aspects and so on.

3.2.2 Evaluation of Multiple Functionalities based on Compare Version of DatAR (V2).

Compare version of DatAR (V2) is developed based on user requirements (URs) from the first round of evaluation (see Section 6.2). No new widgets are created to keep the consistency of the result. The additional aspects of functionalities for each task are as follows:

When participants conduct the task of finding known direct relations (Task 1, Fig 5), the iterated functionalities additionally support filtering topics by publication date (UR4). Participants can also find unknown indirect relations between brain diseases and brain regions (Task 2, Fig 4) using multiple specific intermediate topics, such as two genes (UR5). Publication date (UR4) and specific intermediate topics (UR5) are also available when comparing the direct or indirect relations between two brain topics (Task 3 and 4, Fig 6, 8). Finally, for verifying the identified relations (Task 5, Fig 7), the multiple functionalities additionally display the title (UR3), authors (UR2), and link to full literature (UR1) of the sentence source in compare version of DatAR (V2).

We invite the same group of neuroscientists in the second round of evaluation to gain insight on whether the additional aspects applied on compare version of DatAR (V2) could enhance the usefulness of multiple functionalities. They conduct the same representative tasks identified in Section 3.2 to further explore and validate the identified topics. After conducting the representative tasks, we used the same questionnaire in Section 3.2.1 to collect quantitative feedback related to six metrics. We then conduct semi-structured interviews with the neuroscientists to collect their subjective feedback on **Usefulness** aspect of multiple functionalities. One question is whether neuroscientists think the topics filtered after two rounds are useful for finding potentially fruitful experiments. Given the number of topic filtering rounds typically performed by neuroscientists and the time constraints of this study, we decided not to proceed with further evaluation of the usefulness of multiple functionalities.

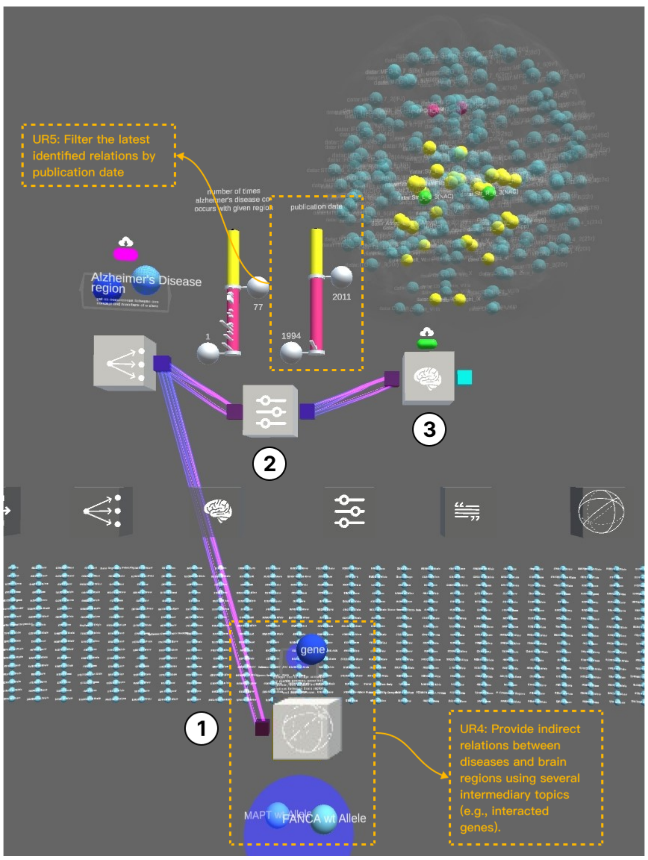

Figure 4. First-person view shows a user conducting the representative task of "Finding unknown indirect relations" using the Indirect Relation Querier (1), Max-Min Cooccurrences Filter Widget (2) and Brain Regions Visualisation (3). The functionalities for filtering publication date and using several specific genes as intermediate topics within the orange outline is an iteration based on UR4, applied to Indirect Relation Querier; and UR5, applied to Max-Min Cooccurrences Filter Widget.

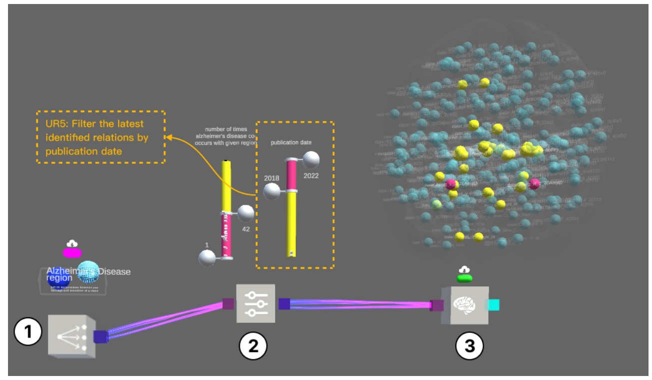

Figure 5. First-person view shows a user conducting the representative task of "Finding known direct relations" using the Direct Relation Explorer (1), Max-Min Cooccurrences Filter Widget (2) and Brain Regions Visualisation (3). The functionality for filtering publication date within the orange outline is an iteration based on UR5, applied to Max-Min Cooccurrences Filter Widget.

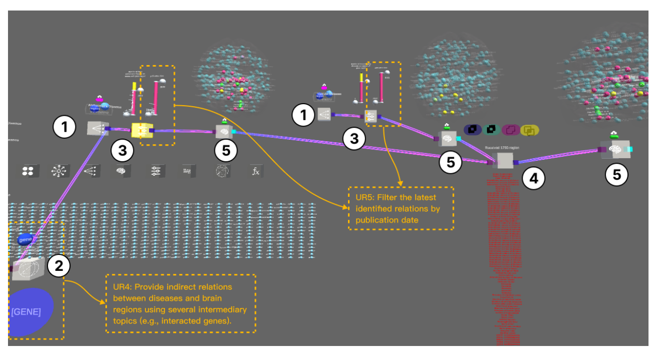

Figure 6. First-person view shows a user conducting the representative task of "Comparing the direct/indirect relations of two disaeses in terms of brain-related topics" using the Direct Relation Explorer (1), Indirect Relation Querier (2), Max-Min Cooccurrences Filter Widget (3), Comparison Widget and Brain Regions Visualisation (4). The functionalities for filtering time and using several specific genes as intermediate topics within the orange outline is an iteration based on UR4, applied to Indirect Relation Querier; and UR5, applied to Max-Min Cooccurrences Filter Widget.

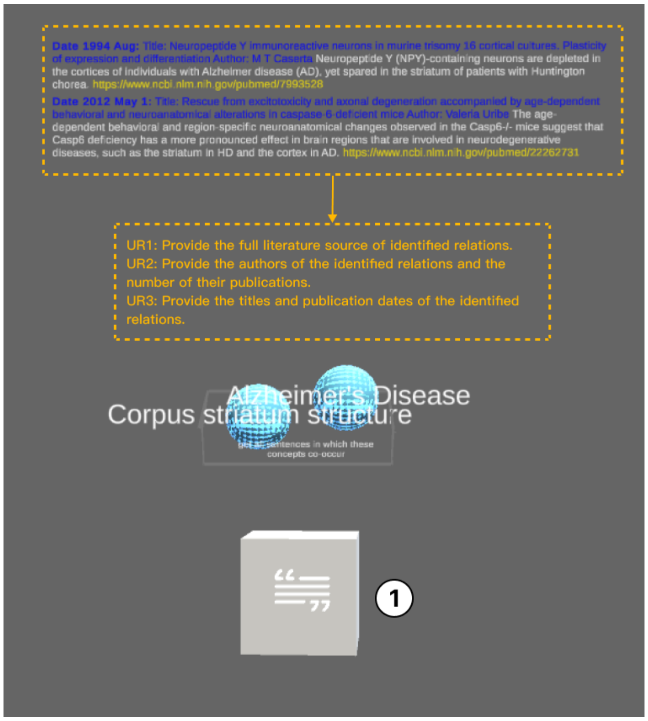

Figure 7. First-person view shows a user conducting the representative task of "Verifying the displayed relations between brain regions and brain diseases" using the Resource Sphere Inspector Widget and Sentences Extractor Widget (1). The functionalities for showing author, title and link to full literature within the orange outline is an iteration based on UR1, UR2 and UR3 applied to Sentences Extractor Widget.

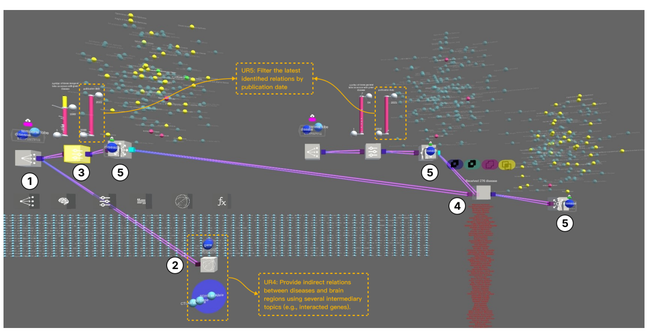

Figure 8. First-person view shows a user conducting the representative task of "Comparing the direct/indirect relations of two brain regions in terms of disease-related topics" using the Direct Relation Explorer (1), Indirect Relation Querier (2), Max-Min Cooccurrences Filter Widget (3), Comparison Widget and Brain Disease Topic Model (4). The functionalities for filtering time and using several specific genes as intermediate topics within the orange outline is an iteration based on UR4, applying on Indirect Relation Querier; and UR5, applying on Max-Min Cooccurrences Filter Widget.

4 Implementation

This section describe the data source (see Section 4.1) and development environment (see Section 4.2) of implementation on the V1 and V2.

4.1 Data Source

We use the Neuroscience Knowledge Graphs of Brain Science (KGBS) available from Triply. This database contains co-occurrence information of topics mentioned in the abstracts and titles of the literature. The literature is collected from neuroscience-related publications available on PubMed. The two different version of DatAR prototype both requires the following data:

- Brain Topic A
- Related Brain Topic B
- Brain category of A (brain disease or brain region)
- Brain category of B (brain disease or brain region)
- Gene/Portein/Mental Process Topic C
- Number of co-occurrences between A and B
- Sources of co-occurrence between A and B (Including title, publication date, authors, PubMed link and sentence source)
- Number of co-occurrences between A and B with C as intermediate topics
- Sources of co-occurrence between A and B with C as intermediate topics (Including title, publication date, authors, PubMed link and sentence source)

4.2 Device

We develop and build the DatAR prototype using Unity3D (version 2020.3.15f2) and the MRTK package (version 2.7.0). The prototype runs on the HoloLens 2 head-mounted display. The C# code for the V1 and V2 is stored in a separate branch of the DatAR tool at Utrecht University.

5 Evaluation

To address our research questions, we conduct two rounds of evaluation. We ask participants to reflect on the use of multiple functionalities for filtering topics, and collect their feedback through both quantitative and qualitative methods.

5.1 Evaluation of Multiple Functionalities based on V1

This section describes the first round of evaluation based on the V1 (see Section 3.2.1). Section 5.1.1 describes the profiles of neuroscientists involved in this study. Section 5.1.2 outlines the comprehensive session from introduction to task involvement. To quantitatively assess the usefulness of multiple functionalities across six metrics, we describe the survey questions in Section 5.1.3. Finally, the semi-structured interview questions explained in Section 5.1.4 are used to collect the subjective feedback for the iteration of multiple functionalities.

Table 3. Backgrounds of participants in evaluation of multiple functionalities. All participants take part in two rounds of evaluation for filtering topics (Section 4 and 6).

No.	Research Topics	Academic Qualifications	Experience in AR	Experience in Systematic Literature Exploration	Duration for Filtering Topics	Tools Usage
P1	Bioinformatics	PhD student (1st year)	Little knowledge	Familiar	Several months	PubMed, Google scholar
P2	Healthcare	PhD student (1st year)	Little knowledge	Familiar	One month	Sciencedirect, Webofscience, Covidence
P3	Childhood brain cancer	PhD student (3rd year)	Little knowledge	Familiar	Several weeks	Pubmed, Vosview
P4	Regenerative Medicine	Junior Research Assistant	Little knowledge	Professional	Several weeks	PubMed
P5	Bioinformatics	Junior Research Assistant	Little knowledge	Little knowledge	One month	Google scholar, Connected papers
P6	Childhood brain cancer (stem cell)	PhD student (1st year)	Not known	Familiar	One month	PubMed

5.1.1 Participants. We recruit six participants (P1-P6, Table 3) from research institutions including UU, UMC, and PMC. The participants' academic qualification, research topics, and experiences with AR and systematic literature exploration are provided in Table 3. All participants have no experience in AR/VR, with one participant (P6) being unknown of AR technology, which might lead to different feedback on the multiple functionalities. Participants have extensive experience in systematic literature reviews except for one participant (P5). The primary tool used for neuroscience literature exploration is traditional search engines, with PubMed being commonly used (by P1, P3, P4, P6). Two participants (P2, P5) mention the topic-based literature exploration tool *Connected Papers*, stating that although this tool only showing relations between topics through 2D visualization, it is still very helpful for neuroscience literature exploration.

The duration required for filtering topics vary from weeks to months, depending on the direction and complexity of future research. All participants indicate that the process of filtering topics needs to be repeated multiple times to find potentially fruitful experiments, which enhances the necessity and credibility of the multiple rounds evaluation results of this study.

5.1.2 Procedure. Each evaluation sessions take around 1 hour. The evaluations are conducted in June 2024. The procedure is as follows:

- **Step 1:** We collect demographic information from six participants through a questionnaire and request them to sign an informed consent form. The demographic information includes participants' research topics, academic qualifications, etc. (see Table 3). This is to prove that the participants' domain knowledge is sufficient to understand the representative tasks conducted during the evaluation procedure and to provide valuable insights. As we are not concerned with how age or gender might affect our results, we do not collect these information.
- **Step 2:** Participants are also asked to report the literature exploration tools they have previously used, the normally duration needed to identify a research topic through literature exploration, and their experience with AR devices.
- **Step 3:** We describe the aim of the evaluation session and provide the introduction to the V1 used for exploring and analyzing topics related to the brain.
- **Step 4:** Participants are required to wear the AR device (Hololens 2) and conduct one of the representative tasks from among five user tasks (see Table 2). Participants can choose to conduct the representative task that is related to their field of study within each user task.
- **Step 5:** After conducting the representative tasks, participants are asked to fill out a survey (see Section 5.1.3) to assess the **Usefulness** of multiple functionalities.
- **Step 6:** After completing the survey, we conduct a semi-structured interview lasting about 15 minutes including five questions (see Section 5.1.4). Participants provide suggestions for improvements in **Usefulness**,

Functionality, and **Visualization** aspects of multiple functionalities based on the V1.

5.1.3 Survey for Usefulness of Multiple Functionalities.

Participants evaluate the usefulness of multiple functionalities for filtering topics based on the V1. The survey includes a total of 55 questions. These questions are organized according to 5 representative tasks (see Table 2). Each representative task contains 11 questions corresponding to 6 metrics (see Section 3.2.1). Participants rated each question using a 7-point Likert scale.

For example, for representative task 1: *"Finding brain diseases that co-occur more than 80 times with the temporal lobe in the literature"*, the question used to assess perceived usefulness (PU) is: *"To what extent do the identified direct relations by conducting this task contribute to your research?"* Participants rate their agreement with this statement from 1 (strongly disagree) to 7 (strongly agree). In order to avoid redundancy, the questions included in the survey are listed in Figure 9 and 10.

5.1.4 Semi-structured Interview for Iteration of Multiple Functionalities.

After participants fill in the questionnaire, we conduct a semi-structured interview with them to collect the subjective feedback on three aspects of multiple functionalities: **Usefulness**, **Functionality**, and **Visualization**.

Usefulness is defined as the extent to which neuroscientists think that using multiple functionalities could aid neuroscience literature exploration for filtering topics, which serve as the main aspect of our study. We initially guide participants to find useful comments from first impressions and open-ended questions: *"Can you describe your impressions of using the DatAR prototype? Which aspects stood out to you as potentially beneficial for your research or you find it interesting?"* We then ask participants to provide suggestions for multiple functionalities for the next round of topic filtering. The iteration aims to improve the usefulness of multiple functionalities: *"Based on your experience, what additional information would be useful for you in your further research?"* and *"Are there any other tasks or functions that the multiple functionalities currently do not support but you feel are useful for your further research? Please describe."*

Functionality and **Visualization** are defined as the extent to which neuroscientists find the use of multiple functionalities pleasant and think it requires further improvement, which serve as the additional aspect of our study. These two aspects provide further insights into multiple functionalities from different aspects. Since these aspects are not the primary focus of our study, the questions are relatively brief: "*What extensions of the implemented functionalities and visual representations would you like to have?"*

5.2 Evaluation of Multiple Functionalities based on V2

We conduct the second round of evaluation to evaluate the usefulness of multiple functionalities after applying URs. We invite the same group of neuroscientists to participate in this study. We measured the participants' feedback on the usefulness of the multiple functionalities through a quantitative survey and semi-structured interviews.

5.2.1 Participants.

All six neuroscientists come from the first round of evaluation. The purpose of inviting the same participants is to have them provide an understanding of the usefulness of multiple functionalities for multiple rounds of topic filtering (RQ), as well as their observations on multiple functionalities before and after the iteration based on their experience and reflection.

5.2.2 Procedure.

Each evaluation sessions take around 1 hour. The evaluations are conducted in August 2024. The procedure of the second round of evaluation is similar to that of the first round and is shown as follows:

- **Step 1:** We describe the aim of this (second round of) evaluation session and introduce the V2 including the iterated multiple functionalities and identified topics from the first round of evaluation.
- **Step 2:** Participants are required to wear the AR device (Hololens 2) and conduct the same representative tasks as in the first round of evaluation (see Table 2).
- **Step 3:** After conducting the representative tasks, participants are asked to fill out the same survey as in the first round of evaluation (see Section 5.1.3) to assess the **Usefulness** of multiple functionalities based on the V2.
- **Step 4:** After completing the survey, we conduct a semi-structured interview lasting about 20 minutes, including six questions (see Section 6.2). Participants provide feedback on **Usefulness** of multiple functionalities from different aspects based on the V2.

5.2.3 Semi-structured Interview for Usefulness of Multiple Functionalities.

At the end of the second round of evaluation, we conduct semi-structured interviews with participants to collect feedback on the **Usefulness** of multiple functionalities for multiple rounds of filtering topics. We focus on feedback regarding the usefulness of multiple functionalities in the following three aspects.

Usefulness of filtering topics compare to V1. We guide participants to provide insights from first impressions based on V2 prototype compare to using multiple functionalities for filtering topics for the first round of evaluation: *"Can you describe your impressions of using the V2 compared to the V1?"*

Usefulness of filtering and verifying topics based on V2. The second aspect of the interview questions focus on the detailed aspects of the V2. We first review which aspects of DatAR have been iterated. Then we guide participants to give feedback on how useful the multiple functionalities are for filtering and verifying topics: *"Do you think the additional aspects help you filter and verify topics?"* and *"Which aspects stand out to you as potentially beneficial for you to filter and verify topics?"*

Usefulness of multi-rounds of filtering topics for further research. The third aspect of the interview questions focus on taking the multiple rounds of filtering topics as a whole. First guide the participants to answer whether or not the multiple functionalities assist them in finding potentially fruitful experiments after 2 rounds of filtering topics: *"Do you think filtered topics are useful for you to find fruitful experiments after 2 rounds of filtering topics?"* Then ask whether they need more rounds or additional information to filter topics: *"Do you think filtering topics again would be useful for you to find fruitful experiments?"* In the end, we guide them to provide other potentially useful requirements to improve the DatAR prototype within neuroscience literature exploration: *"Which functionalities do you think need to be improved in the future?"* and *"Do you need other functionalities to filter topics? Please briefly describe the scenarios."*

6 Result

Section 6.1 describe the qualitative and quantitative results of the evaluation of the V1. We use Likert scale items to measure the participants' responses. Section 6.2 describe the results of the evaluation of the V2, and by comparing the results with those from the first round of evaluation, we comprehensively explain the usefulness of the multiple functionalities for filtering topics (RQ).

6.1 Result of V1

This section addresses the usefulness of multiple functionalities of the V1 for filtering topics. The results include responses from 6 neuroscientists to 55 questions (each representative task includes 11 questions, each question corresponding to one of the 6 metrics, see Figure 9 and 10). Each question is rated on a scale from 1 (low score) to 7 (high score). Neuroscientists are then asked to answer semi-structured interview questions for iteration of the V1.

6.1.1 Interpretation.

To statistically address our research questions, we calculate the minimum, maximum, mean, standard deviation, variance, number of responses, and total score of the quantitative survey. These statistics help illustrate the central tendencies and consistency of responses, offering insight into the reliability and spread of the data. This detailed overview supports a more nuanced interpretation of survey results.

Summary of result. All neuroscientists somewhat agree (ms>5) on the usefulness of multiple functionalities, indicating that these functionalities play a role in assisting them for filtering topics. However, three neuroscientists (P1, P4, P6) have given scores of 3 or lower on several metrics, including perceived enjoyment (PE), job relevance (JR) and result demonstrability (RD), indicating that they hold differing opinions in certain aspects.

For "Finding known direct relations" (Task 1), the high average scores on all questions indicate that the majority of participants (P1, P2, P3, P5) agree that multiple functionalities are useful to find direct relations. The results are apparent (RD), sufficient (PI) and easy to understand (OQ). Only one participant (P5) finds it difficult to verify whether the identified relations are useful for their research. He/She states that the result provided by traditional topic-based literature exploration tools could be more accurate and efficient for filtering topics compared to multiple functionalities in AR environment.

For "Finding unknown indirect relations" (Task 2), half of the participants (P1, P3, P4) are doubtful about the relevance of the identified unknown indirect relations to their research (JR). Two participants (P1, P4) feel somewhat unpleasant (PE) due to the complexity and time-consuming nature when using multiple functionalities to find indirect relations. Additionally, two participants (P4, P6) find it difficult to judge the contribution of the identified indirect relations to their research (PU) and suggest that more detailed information about the identified indirect relations, such as specific genes or disease subtypes, would be helpful.

When comparing direct/indirect relations (Tasks 3 and 4), half of the participants (P1, P4, P5) show subjective differences in their level of perceived enjoyment (PE) while conducting representative tasks. They (P1, P2) indicate that overlapping information on visualization among brain topics further decreases their perceive usefulness when conducting representative tasks. Additionally, one participant (P6) questions whether the union results provided by comparison functionalities are apparent (RD), as the additional identified relations could not assist them in filtering and narrowing down the scope of topics.

Finally, for "Verifying the displayed relations between brain regions and brain diseases" (Task 5), although participants highly agree that the sentence source could help verify identified topics, three of them (P3, P4, P6) hold neutral or negative attitudes towards the sufficiency of the provided information (PI) and the relevance of the identified topics (JR) to their research. They (P2, P5) suggest providing more detailed information related to the identified topics. The detailed information of quantitative surveys in terms of statistics and metrics is as follows:

Distribution of average score. Neuroscientists gave an average score of 5.42 for "Finding known direct relations" (Task 1), with four participants (P1, P2, P3, P5) scoring 5 or above on all questions, indicating the general agreement on the usefulness of using the multiple functionalities for finding direct relations between interested brain topics. Similarly, "Verifying the displayed relations between brain regions and brain diseases" (Task 5) received an average score of 5.79, suggesting that the sentence source could be helpful for filtering topics, except for two participants (P4, P6) who scored below 4 on three questions. The distribution of average scores for "Comparing direct/indirect relations" (Tasks 3 and 4) is relatively high with two participants (P2, P3) giving higher overall scores.

Degree of variation. There is a higher degree of variation in scores for "Comparing the direct/indirect relations of two brain regions in terms of disease-related topics" (Task 4) (sd4=1.15) and "Verifying the displayed relations between brain regions and brain diseases" (Task 5) (sd5=1.09), with 3 participants(P1, P4, P6) giving lower scores. This high variation score indicates that there is a divergence of opinions among neuroscientists regarding the usefulness of multiple functionalities when comparing or verifying the filtered topics.

Metrics score. For job relevance (JR), neuroscientists give lower average scores for "Finding unknown indirect relations" (Task 2) (ms2=4.92) compared to other tasks, with 3 participants(P1, P3, P4) expressing less positive views on the relevance of the indirect relations found to their research.

For perceived usefulness (PU), neuroscientists give the highest average scores for "Verifying the displayed relations between brain regions and brain diseases" (Task 5) (ms5=6.33), indicating that participants believe the provided sentence sources can assist them in filtering the identified relations. However, for "Finding unknown indirect relations" (Task 2), the average score is 4.8, with 2 participants (P1, P3) expressing negative views on the perceived usefulness of the identified indirect relations, suggesting that participants believe understanding these indirect relationships requires some prior knowledge.

For perceived enjoyment (PE), 3 participants (P1, P4, P6) express below neutral attitudes for the tasks, aside from "Finding known direct relations" (Task 1) and "Finding unknown indirect relations" (Task 2), suggesting that as task complexity increases, participants increasingly feel unpleasant when using multiple functionalities for filtering topics.

For output quality (OQ) and perceived informativeness (PI), the average scores for all tasks are greater than 5, indicating that participants believe the identified relations and topics are sufficient, accurate, and easy to understand for their research.

For result demonstrability (RD), two participants (P1, P4) find it difficult to understand the identified indirect relations (Task 2) and the union/difference sets between brain regions/brain diseases (Task 3 and 4). Understanding these results requires prior knowledge and time to grasp their meaning.

6.1.2 Proposed Enhancement.

Despite the quantitative feedback on the V1, participants express the desire for more detailed information for filtering topics during semi-structured interviews in the follow aspect:

Usefulness: Five neuroscientists indicate that to further filter the 20 topics down to 5-6 potentially fruitful experimental topics, detailed information is necessary (P1, P2, P3, P5, P6). Half of the participants emphasize the usefulness of publication date (P1, P2, P5) and original text links (P1, P2, P3, P5, P6) for further topic filtering. Two participants express the need to filter relations by the most recent 10 years to ensure that the identified relations are up-to-date and useful for finding potentially fruitful experiments (P2, P5). Moreover, by using specific disease subtypes, gene chains (a series of interacting genes), or a set of related structural proteins as intermediate topics, neuroscientists could obtain more detailed information on indirect relationships relevant to their research fields (P5, P6). While author (P1), article citation count (P5), and journal impact factor (P3) are also considered useful for filtering topics, the latter two are dismissed in subsequent focus group discussions due to potential bias (P1, P6).

Functionalities and Visualization: Half of the participants (P1, P5, P6) agree that the color scheme representing direct and indirect relations in the DatAR prototype is appropriate. One participant mentions that the color coding needs optimization when comparing the relationships between two topics (P3). Two participants believe that using immersive analysis techniques in an AR environment could help researchers quickly map relations to real brain regions, given the complexity of the brain's 3D structure (P3, P6). Although 3 participants (P3, P5, P6) find the usefulness of multiple functionalities to be high, adding functions such as batch literature source exporter(P1,P3) could enhance the ease of use when filtering topics. Additionally, because filtering topics requires conducting multiple tasks at the same time, two participants suggested adding multi-task management windows or other similar functionalities to improve efficiency and prevent information overlap in the workspace.

These feedback are collected and transferred into 13 potential user requirements (URs), as shown in Table 4. We then conduct a focus group interview with two neuroscientists (P1, P6) to discuss the collected URs based on neuroscientists' domain knowledge. 5 URs are identified for implementation on V2 based on conclusions from the interview and limitations of the dataset used by DatAR. These URs are marked with a check in the last column of the table and the detail descriptions are as follow. The identified 5 URs are used for iterated the multiple functionalities (see Section 3.2.2).

UR1: Provide the full literature source of identified relations. Participants (P1, P6) should be able to access the full literature of the identified relations. This information can be provided in the form of links. The provided full literature could help neuroscientists further verify the usefulness of identified relations, for example, by reading the full literature to filter out confounding effects (P6). Since displaying full

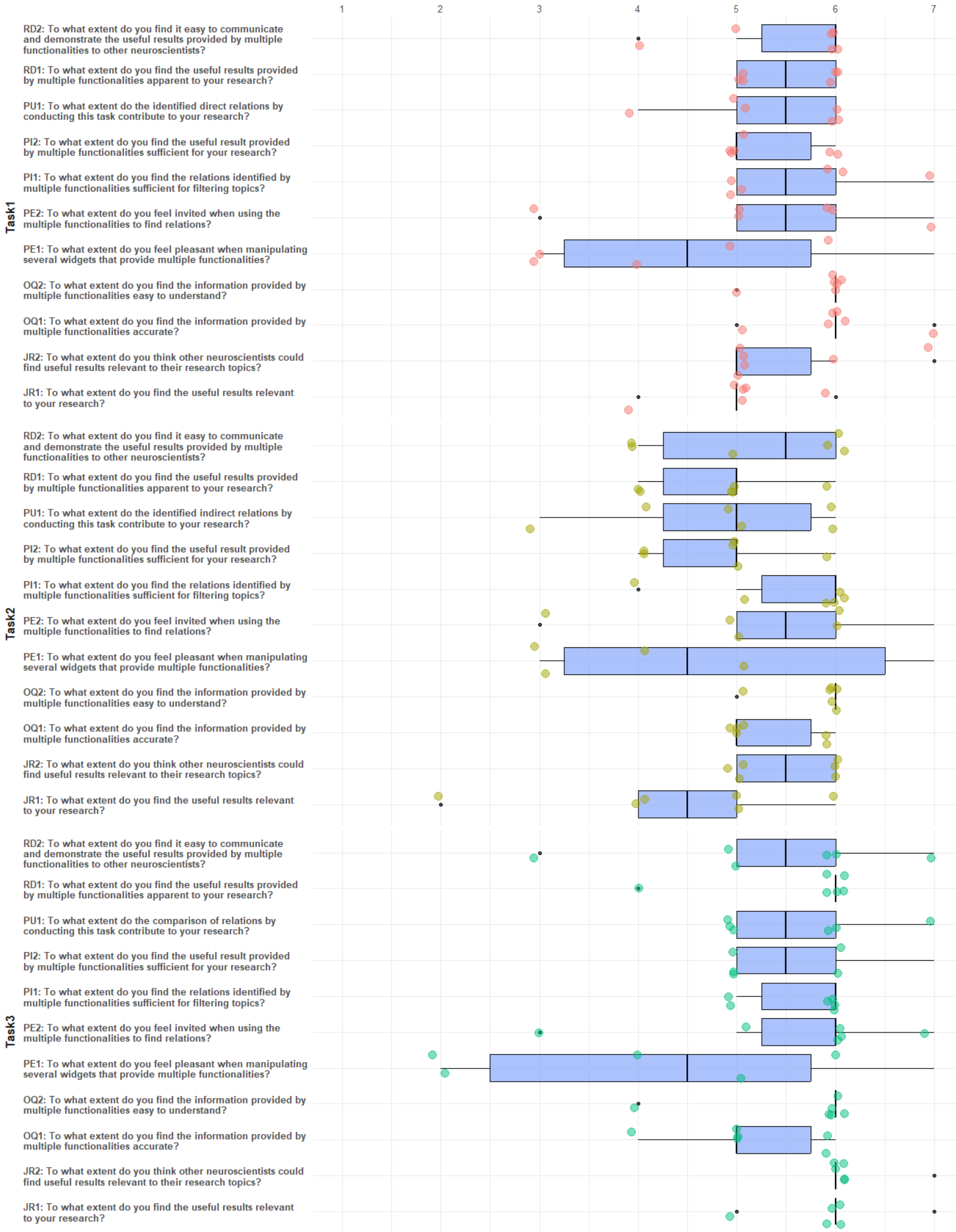

Figure 9. Quantitative result (Task1 to Task3) of the evaluation on V1. Six neuroscientists responded to questions based on the conducted representative tasks using a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Each question corresponds to one of the six metrics aspects. The box plot displays the average scores and standard deviations of the ratings. The small circles are the ratings given by participants.

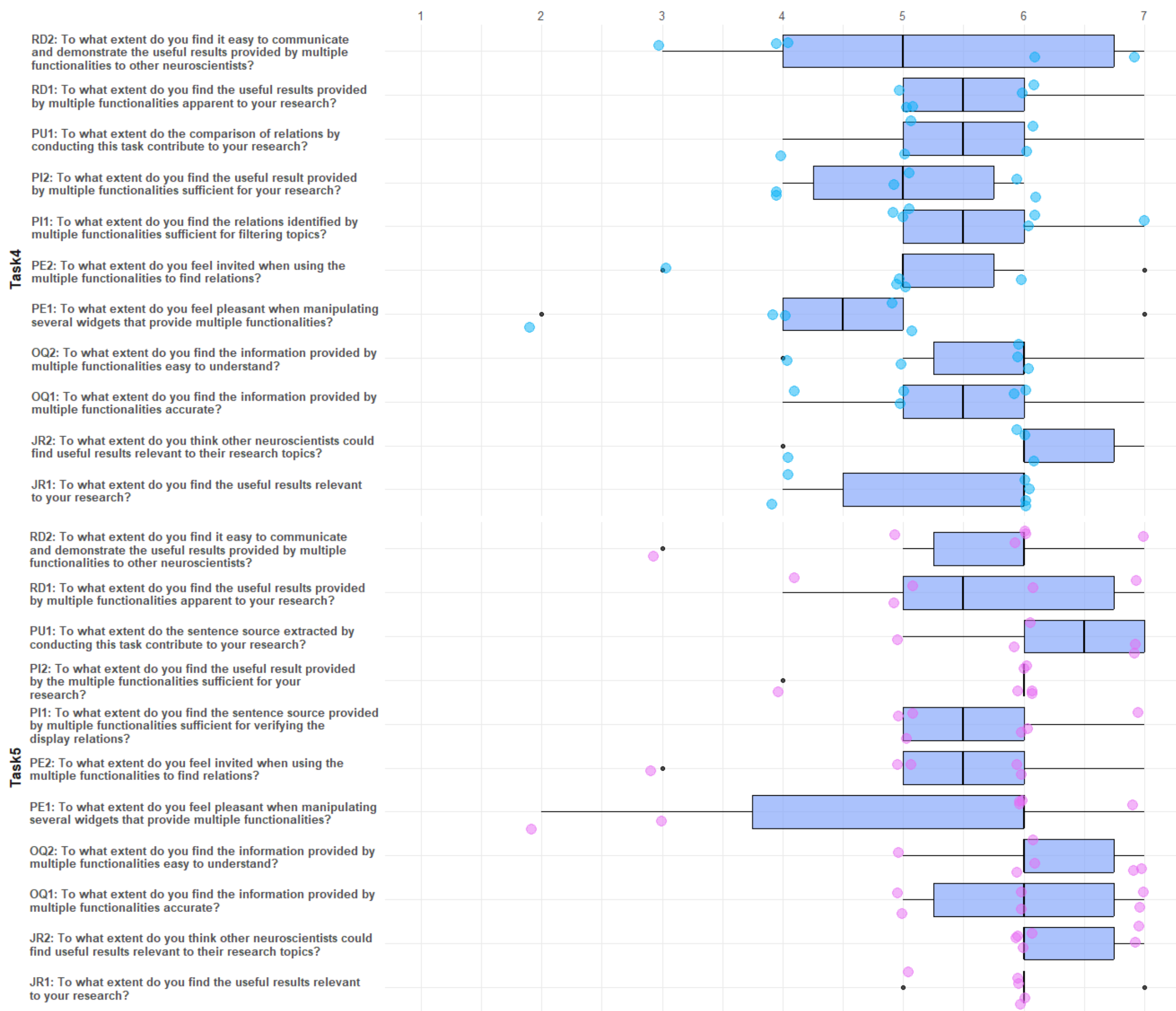

Figure 10. Quantitative result (Task4 and Task5) of the evaluation on V1. Six neuroscientists responded to questions based on the conducted representative tasks using a 7-point Likert scale (1=strongly disagree, 7=strongly agree). Each question corresponds to one of the six metrics aspects. The box plot displays the average scores and standard deviations of the ratings. The small circles are the ratings given by participants.

literature for each identified relations could clutter the user interface, it is a better choice to show the links only when verifying topics (P1, P6).

UR2: Provide the authors of the identified relations. One participant (P1) express the desire to identify the authors mentioning in these relations when verifying topics. Author information can help neuroscientists quickly understand the representative scientists or authority in specific areas of neuroscience. Another participant (P6) hold a neutral view, stating that author information might be useful for verifying topics, but it needs to be taken into consideration with other information.

UR3: Provide the titles and publication dates of the identified relations. The titles of literature are equally important for verifying topics. Participants first need to see the titles of literature to understand the research topics of previous work. They then further read the sentence sources to validate the identified relations based on their domain knowledge.

UR4: Provide indirect relations between diseases and brain regions using several intermediate topics. One participant (P6) express the desire to display indirect relations using a series of genes (pathways) as intermediate topics during the first evaluation. For example, in gene knockouts techniques, there is often a series of genes with the same pathway. Identifying the indirect relations among a series of genes is useful because single genes may have few or no relations, and multiple genes often collectively influence a disease or brain region.

Table 4. User requirements collected from the semi-structured interview in the first round of evaluation through UCD process. Five identified user requirements with tick mark in the last column are applied to the V2.

ID	User Requirements	Categorization	Source	Improvement
UR1	Provide the full literature source of identified relations.	Usefulness	P1,P3,P4,P5	✓
UR2	Provide the authors of the identified relations.	Usefulness	P1,P3	✓
UR3	Provide the titles and publication dates of the identified relations.	Usefulness	P1,P2,P5	✓
UR4	Provide indirect relations between diseases and brain regions using several intermediate topics (e.g., interacted genes).	Usefulness	P6	✓
UR5	Filter the latest identified relations by publication date.	Usefulness	P2,P5	✓
UR6	Provide the experimental results and images of the identified relations from literature source.	Usefulness	P3,P5	
UR7	Provide the relations between specific disease subtypes and brain regions	Usefulness	P5	
UR8	Filter valuable relations by the number of citations of the articles, enlarging the volume of spheres for highly cited relations' topics.	Usefulness	P5	
UR9	Filter valuable relations by journal impact factor.	Usefulness	P3	
UR10	Allow zooming of topic model and brain regions visualization.	Functionality	P1	
UR11	Allow batch export of full literature source of identified relations.	Functionality	P1	
UR12	Avoid information overlap when displaying multiple tasks.	Visualization	P1,P2	
UR13	Use different colors to distinguish union relations between two topics.	Visualization	P3	

During the interview, participants (P1, P6) further express the need to find indirect relations using several intermediate topics, such as identifying relations between Alzheimer's and brain regions using multiple proteins as intermediate topics. Then, participants can narrow down the range of intermediate topics to further investigate if other brain-related topics have stronger connections to Alzheimer's.

UR5: Filter the latest identified relations by publication date. Previously verified relations between brain regions and brain diseases might be overturned by recent studies (P1). Participants (P1, P6) should be able to quickly find literature from specific time periods when filtering topics, ensuring that the findings are truly useful for future research (for example, showing only co-occurrences from the past ten years). Moreover, displaying the temporal distribution of relations can visually represent trends in research popularity, allowing participant (P6) to further investigate, for instance, the surge of literature in specific years.

6.2 Result of V2

The same group of participants provide feedback on the usefulness of multiple functionalities for filtering topics based on the V2. The questions in the questionnaire are the same as those in the first round of evaluation. The quantitative result can be seen in Figure 11 and 12. Neuroscientists then answer semi-structured interview questions (Section 5.2.3). These questions focus not only on the V2 but also take the multiple rounds of filtering topics as a whole into consideration.

6.2.1 Interpretation. Summary of results.

All neuroscientists somewhat agree (mean score > 5) on the usefulness of multiple functionalities, which is similar to the evaluation of the V1. Most participants (P1, P3, P5, P6) express a more positive attitude toward the multiple functionalities of the V2 when finding unknown indirect relations (Task 2) and comparing the direct/indirect relations of two diseases in terms of brain-related topics (Task 3), implying that the additional aspects, including publication dates and specific multiple intermediate topics, are useful improvements for filtering topics. However, four participants (P3, P4, P5, P6) give lower scores in job relevance (JR), output quality (OQ), and perceived informativeness (PI), suggesting that neuroscientists have stricter criteria for the credibility, sufficiency, and relevance of results to their own research fields when further filtering topics to find potentially fruitful experiments. In addition, the perceived enjoyment (PE) scores significantly decline when finding and comparing relations between topics, revealing that usability issues impact how neuroscientists perceive the usefulness of multiple functionalities.

Distribution of average score. The average score in the second round of evaluation shows a different trend compared to the first round. Scores for "Finding unknown indirect relations" (Task 2) (ms2=5.18) and "Comparing the direct/indirect relations of two brain regions in terms of disease-related topics" (Task 4) (ms4=5.41) increase, indicating that additional aspects such as publication dates and multiple specific intermediate topics, improve the accuracy of the identified relations and contribute to narrow down the scope of filtered topics. The average score for "Verifying the displayed relations between brain regions and brain diseases" (Task 5) (ms5=5.35) shows the largest decrease (0.44). Participants (P5, P6) mention that the additional information including title and author helps them verify the identified relations more efficiently, but because the lack of field-specific information, it does not contribute to their research. However, the increase in the lowest score in these tasks suggests that participants still agree that verifying relations through sentence sources is somewhat useful for filtering topics.

Degree of variation. The score distribution for "Comparing the direct/indirect relations of two brain regions in terms of disease-related topics" (Task 4) (sd4=0.96) and "Verifying the displayed relations between brain regions and brain diseases" (Task 5) (sd5=0.92) is more centralized, while the score distribution for the other three tasks remains largely unchanged. The reduction in variation indicates that neuroscientists' opinions on the usefulness of multiple functionalities for further comparison and verification of identified relations are converging.

Metrics score. For job relevance (JR) and output quality (OQ), the average scores decline for all tasks except "Finding unknown indirect relations" (Task 2). Half of the participants (P3, P5, P6) mention that the identified relations irrelevant to their research fields. Insufficient data and inconsistency in topic terminology also reduce the credibility of the results, which is confirmed in the subsequent semi-structured interviews.

For perceived usefulness (PU), neuroscientists show a significant increase in the average score for finding unknown indirect relations (Task 2) (ms2=5.67). Five participants (P1, P2, P3, P4, P5, P6) find the multiple specific intermediate topics useful for filtering topics for their research.

For perceived enjoyment (PE), the average scores drop significantly for all tasks except "Verifying the displayed relations between brain regions and brain diseases" (Task 5), indicating that participants feel complex and inefficient when using the multiple functionalities in the V2. The result emphasize the usability and user experience issues during multiple rounds of filtering topics.

For perceived informativeness (PI), the average scores decline for "Comparing the direct/indirect relations of two diseases in terms of brain-related topics" (Task 3) (ms3=5.17) and "Verifying the displayed relations between brain regions and brain diseases" (Task 5) (ms5=5.08), indicating that although the additional aspects including title, author and link to full literature offer more detailed information about the identified relation, participants feel that the results are insufficient for their research. This feedback highlights the importance of offering sufficient latest topics while ensuring their relevance to the neuroscientists' specific research areas.

For result demonstrability (RD), the average scores for some tasks slightly decline, indicating that understanding the further filtered topics requires slightly higher time investment and prior knowledge.

6.2.2 Proposed Qualitative Feedback.

During the semi-structured interviews, participants provide feedback on three aspects of the usefulness of multiple functionalities:

Usefulness of filtering topics compared to the 1st round. Participants are asked to describe their impressions of using the version 1 of the DatAR prototype compared to their initial experience with version 2. This question, also used in the 1st round of evaluation, aims to gather insights into participants' first impressions and the perceived improvements in the new prototype's functionalities. All six participants agree that the V2 provide more comprehensive and useful information to help filter topics, especially in finding specific and research-relevant topics. However, three participants (P1, P2, P4) express a neutral or lower attitude toward the usability of multiple functionalities in the V2. One participant (P2) finds the multiple functionalities to be complex in operation, significantly increasing the time required to complete the representative tasks. Another participant (P4) also expresses concerns about the ease of use of multiple functionalities, as they need more steps to achieve the desired results.

Usefulness of filtering and verifying topics in the 2nd round. Participants are asked to provide feedback on specific aspects of the version 2 of the DatAR prototype, which includes new functionalities such as publication date, author information, links to full literature, and the selection of multiple intermediate topics. The aim was to understand how these additions support filtering and verifying topics. Four participants (P1, P2, P4, P5) find that the additional information provided by the V2 helps them further filter and verify topics. Among them, two participants (P1, P2) express a positive attitude, stating that the publication date filter helps them further narrow down the range of latest identified topics to be verified to fewer than 10. Two participants (P4, P5) partially agree with this statement, with only one participant (P4) mentions that some of the additional information, such as provided specific intermediate topic genes

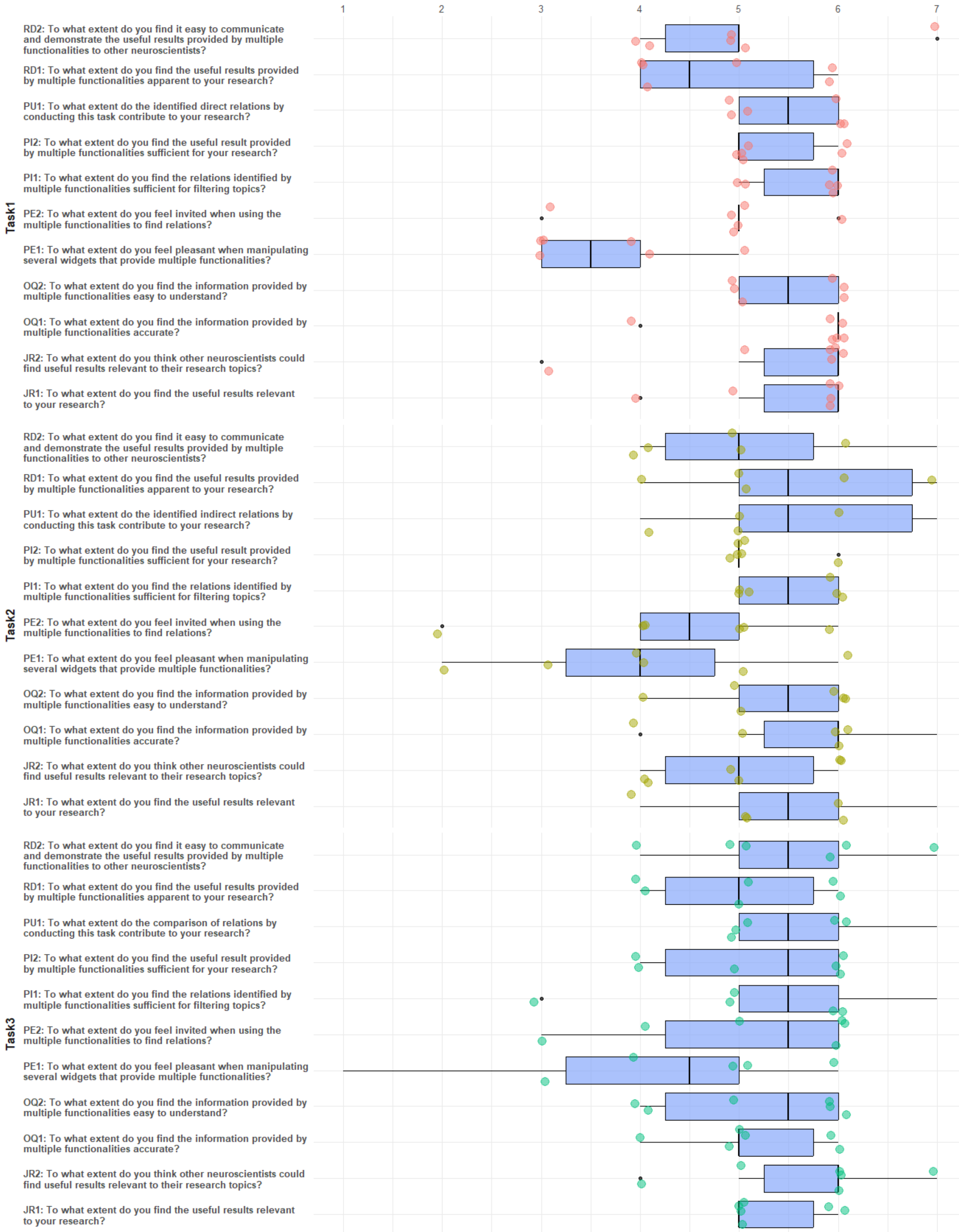

Figure 11. Quantitative result (Task1 to Task3) of the evaluation on V2. The same group of neuroscientists responded to the same questionnaire from the first round of evaluation. The box plot displays the average scores and standard deviations of the ratings. The small circles are the ratings given by participants.

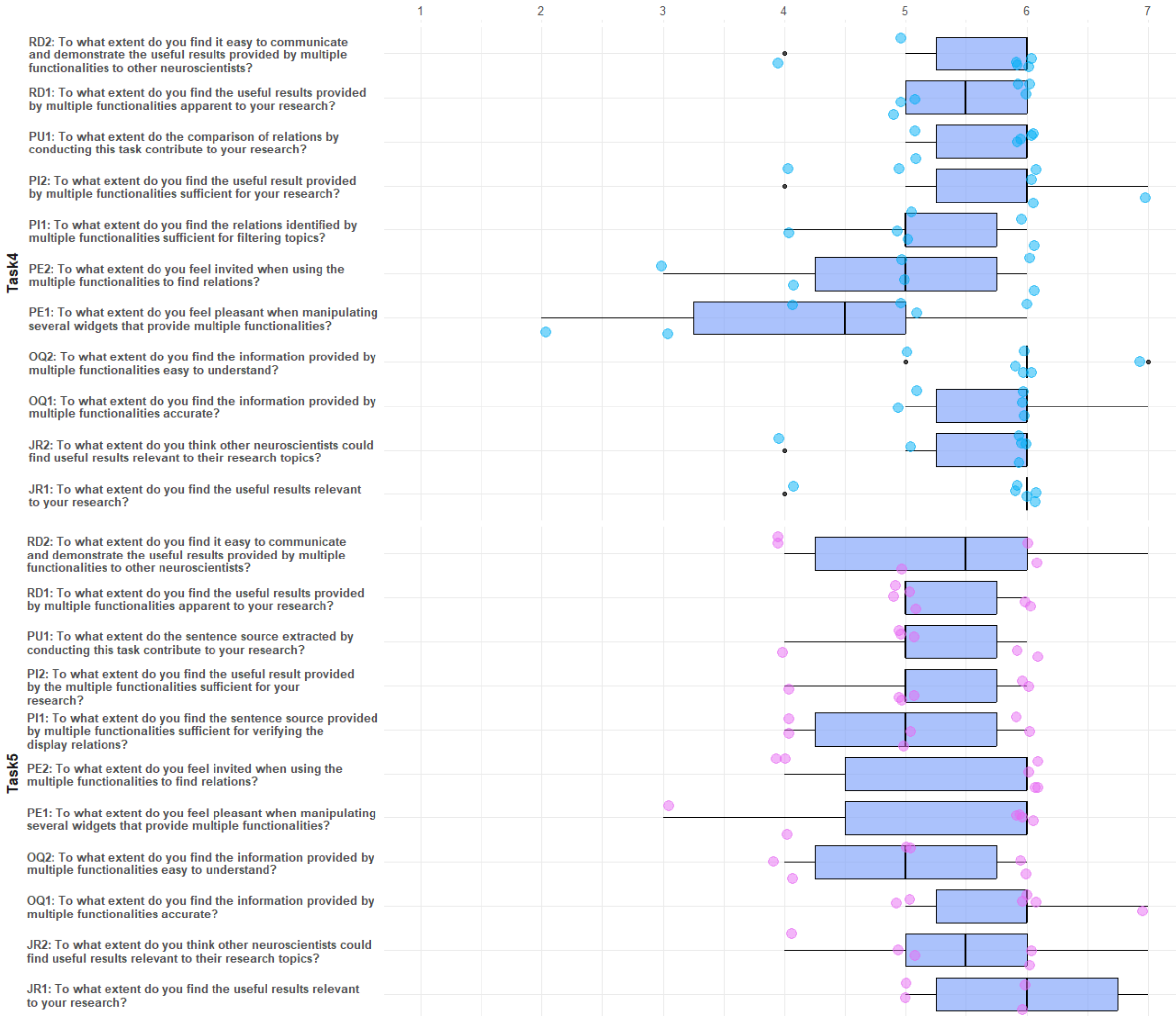

Figure 12. Quantitative result (Task4 and Task5) of the evaluation on V2. The same group of neuroscientists responded to the same questionnaire from the first round of evaluation. The box plot displays the average scores and standard deviations of the ratings. The small circles are the ratings given by participants.

irrelevant to their research, may not be useful for filtering topics.

Regarding the additional information including publication date, authors, links to full literature, and the selection of multiple intermediate topics, four participants (P1, P2, P4, P6) consider publication date to be the most potentially beneficial, followed by links to full literature (P1, P3, P5). Two participants (P5, P6) think that multiple specific intermediate topics are helpful for filtering topics, but information such as authors, titles, and literature links cannot directly help in filtering topics. They still need to read the full texts on PubMed to further verify the identified relations.

Usefulness of multi-rounds of filtering topics for further research. Participants are asked whether multiple rounds of filtering topics, as implemented in the DatAR prototype, assist them in identifying potentially fruitful experiments. One participant (P2) believes that multiple rounds of filtering topics can narrow down the topics to fewer than 10, which is useful for finding potentially fruitful experiments. The other five participants (P1, P3, P4, P5, P6) partially agree with this view, noting that neuroscientists need other supplementary information to determine whether the identified topics are truly useful for their research such as the context of the sentence source.

Regarding whether using multiple functionalities again to filter topics is useful for finding potentially fruitful experiments, four participants (P2, P3, P4, P5) express a neutral attitude. After obtaining around ten verified topics in the second round of evaluation, participants need to read the specific literature to find useful information for future research, such as pathogenic mechanisms of genes or experimental methods. If the identified relations are not sufficient for future research, participants may use multiple functionalities again to filter topics to find more potentially useful relations or interesting topics. Three participants (P1, P3, P6) describe potential scenarios for using multiple functionalities to filter topics again: neuroscientists may use the multiple functionalities again based on the identified topics to filter topics of interest, such as other brain disease topics related to the same brain region using the same specific gene as an intermediate topic.

6.2.3 Proposed Enhancement. Integrating with traditional literature exploration tools. Four participants (P1, P2, P4, P5) suggest that being able to read the full literature directly on a 2D screen for the identified topics is beneficial, as it allows them to further verify whether the identified relations are truly useful for future research. Another participant (P3) states that it depends on whether the provided sentence sources offer enough valid information. If the information is sufficient, there may be no need to browse the full literature again.

Providing information within a specific field of neuroscience Five participants (P1, P3, P4, P5, P6) believe that the current DatAR prototype provides general topics in neuroscience, which is suitable for junior researchers who have little or no understanding of neuroscience. For senior researchers in specific fields of neuroscience, it would be more useful to identify potentially fruitful experiments if the DatAR prototype could display topic information related to their own research aspect.

Improving the usability of multiple functionalities. Three participants (P1, P2, P4) indicate that the ease of use of multiple functionalities significantly affects their perceived usefulness when filtering topics. Providing a more convenient interaction method could enhance the user experience and efficiency. For example, using more efficient interaction methods rather than sliding to precisely filter by publication date and co-occurrence count (P2, P3).

Providing detailed information on sentences sources. Two participants (P2, P6) state that information such as the citation count of literature is useful for making an initial judgment on the usefulness of the identified relations, as this information can reflect the authority of the source. Another participant (P3) mentions that the name of the journal is also an important reference for filtering topics.

Data completeness and consistency. Two participants (P5, P6) mention that due to the ongoing development in the field of neuroscience, the naming of brain regions and brain disease topics is constantly evolving. Ensuring consistency in data entry could improve the reliability of the identified relations.

Displaying changes in identified relations. One participant (P4) suggests the need to record and compare the changes in identified relations between multiple rounds of filtering filtering, as these changes may reflect shifts in research trends, which could be useful for future research.

7 Discussion

We discuss the feedback and observations collected from two rounds of evaluation on the initial and V2. Our ideas emphasize future development strategies for more deeply integrating DatAR into neuroscience literature exploration for daily research.

7.1 Usability Concerns in Multiple Rounds and Multitasking Topic Filtering

Enhancing functionality for filtering topics. Participants report that conducting in parallel with multiple representative tasks and filtering topics multiple times is complex and time-consuming. To improve the usability of multiple functionalities, aspects include allowing the recording and comparison of differences in identified relations during multiple rounds of filtering topics, providing functional templates for commonly used representative tasks that participants can invoke as needed, and allowing the use of voice commands to connect multiple functionalities as an alternative to the existing drag-and-drop gesture, thereby simplifying the interaction process.

Enhancing visualization for filtering topics. Participants emphasize the need to improve visualization of identified relations based on complex multitasking scenarios when filtering topics. The potential improvements include highlighting only the topics or relations identified in the tasks of interest while diminishing the visualization of other tasks, allowing the selection to display only parts of the brain region visualization or disease topic-model relevant to the participants' research, such as displaying the subdivision of the brain. Additionally, employing different colors for visualization can aid in intuitively comparing the differences in identified relations across various topics during multitasking for filtering topics.

Refining guidance and basic functionalities. Participants suggest that the DatAR prototype needs to consider necessary basic functionalities in future development to enhance usability. The potential improvements include providing initial guidance to help participants familiarize themselves with the functionalities and adding a result export function that allows participants to export identified relations and filtered topics, thus facilitating better verification for subsequent literature exploration for filtering topics.

7.2 Integrating DatAR into Daily Literature Exploration Process

Integrating multiple functionalities into the traditional literature exploration process. Qualitative results mention possible scenarios for reusing multiple functionalities to filter topics. Neuroscientists alternately use multiple functionalities to identify potentially useful relations and traditional literature analysis tools to verify these identified relations through detailed information until they find potentially fruitful experiments. This feedback emphasizes the necessity of incorporating multiple functionalities into the traditional literature exploration process. A possible direction is to allow the synchronization of identified relation source data to literature-based discovery platforms for further analysis.

Integrating sufficient detailed data into the DatAR prototype. Another possible improvement is to expand the scope of the database to display more detailed information within immersive analysis environment, such as providing context for sentence sources of identified relations and showing the citation of the identified terms or topics to enhance the credibility. The visualization of textual information needs further consideration to avoid overlap and other issues that affect readability, especially when handling large amounts of data.

7.3 Delving into Real-World Neuroscience Research

Providing topics on specific neuroscience areas. Participants report that current topics cover general areas of neuroscience, which could be verified to be not very useful for future research. They suggest that there should be an integration of the latest and subfield-specific topics. Moreover, ensuring the consistency of the terms of integrated topics could enhance the credibility of the identified relations. These improvements allow neuroscientists to explore specific topics related to their research and match real neuroscience literature exploration scenarios.

Applying to real-world neuroscience research. Participants express the desire to integrate multiple functionalities into real research in the evaluation. After integrating the latest or subfield-specific topics, it would be valuable to apply multiple functionalities to the ongoing process of a specific real-world neuroscience research. This could involve using a shadow-following approach to observe a neuroscientist within literature exploration, examining the future development and application of DatAR, and exploring how multiple functionalities could be integrated with traditional literature exploration tools to assist this neuroscientist in finding potentially fruitful experiments.

8 Conclusion and Future Work

We integrate and present multiple functionalities from previous work in this study. These functionalities aim to find, compare, and verify the complex relations between brain-related topics. We follow the evaluation and iteration approach to assist neuroscientists in conducting two rounds and multitasking topic filtering. Our goal is to evaluate the usefulness of multiple functionalities to gain insights on how to integrate DatAR into neuroscience workflows in the future.

Our research shows the potential of multiple functionalities to support neuroscience literature exploration for filtering topics. Key aspects that positively influence the usefulness of multiple functionalities include the accuracy of identified relations, the sufficiency of the source information for displayed relations, and the relevance of identified relations to the researchers' research. Furthermore, while additional aspects of multiple functionalities, such as publication date and specific intermediate topics, assist in filtering topics, it is necessary to consider the negative impact of usability issues. These issues include the lack of efficiency enhancing functionalities for multitasking in topics filtering scenarios, interaction complexity, and the effect of visualization on usefulness.

We will continue collaborating with neuroscientists, incorporating participant feedback to improve the usefulness of multiple functionalities. These aspects include enhancing functionalities for multiple rounds and multitasking topic filtering, improving visual feedback on identified relations, and providing guidance along with sufficient detailed information for the filtered topics. Another future direction is to place multiple functionalities within a longitudinal, real-world neuroscience research setting to understand how these functionalities can help neuroscientists find potentially fruitful experiments. The focus is on integrating DatAR into traditional literature exploration process and allowing neuroscientists to delve deeply into the context of specific topics related to their research.

In conclusion, we view this work as contributing valuable insights into the usefulness of multiple functionalities, aiding neuroscientists in their multiple rounds and multitasking topic filtering. Deepening our understanding of how these functionalities are applied in real-world neuroscience research will further advance the integration of DatAR into neuroscience daily research in the future.

References

[1] A Balog and C Pribeanu. 2009. Developing a measurement scale for the evaluation of AR-based educational systems. *Studies in Informatics and Control* 18, 2 (2009), 137–148.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[3] Fred D Davis et al. 1989. Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption* (1989), 205–219.

[4] Khaled MS Faqih and Mohammed-Issa Riad Mousa Jaradat. 2015. Assessing the moderating effect of gender differences and individualism-collectivism at individual-level on the adoption of mobile commerce technology: TAM3 perspective. *Journal of Retailing and Consumer Services* 22 (2015), 37–52.

[5] Chris Hart. 2018. Doing a literature review: Releasing the research imagination. (2018).

[6] Tseng-Lung Huang and Shuling Liao. 2015. A model of acceptance of augmented-reality interactive technology: the moderating role of cognitive innovativeness. *Electronic Commerce Research* 15 (2015), 269–295.
[7] KS Khan, A Bueno-Cavanillas, and J Zamora. 2022. Systematic reviews in five steps: II. Identifying relevant literature. *Semergen* 48, 6 (2022), 431–436.
[8] Lawrence A Machi and Brenda T McEvoy. 2009. The literature review: Six steps to success. (2009).
[9] Kim Marriott, Falk Schreiber, Tim Dwyer, Karsten Klein, Nathalie Henry Riche, Takayuki Itoh, Wolfgang Stuerzlinger, and Bruce H Thomas. 2018. *Immersive analytics*. Vol. 11190. Springer.
[10] Graeme McLean and Alan Wilson. 2019. Shopping in the digital world: Examining customer engagement through augmented reality mobile applications. *Computers in Human Behavior* 101 (2019), 210–224.
[11] Jakob Nielsen and Thomas K Landauer. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 206–213.
[12] Eleonora Pantano, Alexandra Rese, and Daniel Baier. 2017. Enhancing the online decision-making process by using augmented reality: A two country comparison of youth markets. *Journal of Retailing and Consumer Services* 38 (2017), 81–95.
[13] Costin Pribeanu et al. 2012. Using formative measurement models to evaluate the educational and motivational value of an AR-based application. *Problems of Education in the 21st Century* 50, 1 (2012), 70–79.
[14] Alexandra Rese, Daniel Baier, Andreas Geyer-Schulz, and Stefanie Schreiber. 2017. How augmented reality apps are accepted by consumers: A comparative analysis using scales and opinions. *Technological Forecasting and Social Change* 124 (2017), 306–319.
[15] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *Journal of Biomedical Informatics* 121 (2021), 103865.
[16] Philipp Spreer and Katrin Kallweit. 2014. Augmented reality in retail: assessing the acceptance and potential for multimedia product presentation at the PoS. *Transactions on Marketing Research* 1, 1 (2014), 20–35.
[17] Kuttimani Tamilmani, Nripendra P Rana, Samuel Fosso Wamba, and Rohita Dwivedi. 2021. The extended Unified Theory of Acceptance and Use of Technology (UTAUT2): A systematic literature review and theory evaluation. *International Journal of Information Management* 57 (2021), 102269.
[18] Ghazaleh Tanhaei. 2022. Exploring Neuroscience Literature and Understanding Relations Between Brain-Related Topics-Using Augmented Reality. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*. 387–390.
[19] Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2023. DatAR: Comparing Relationships between Brain Regions and Diseases: An Immersive AR Tool for Neuroscientists. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*. 373–375.
[20] Ghazaleh Tanhaei, Ivar Troost, Lynda Hardman, and Wolfgang Hürst. 2022. Designing a topic-based literature exploration tool in AR-an exploratory study for neuroscience. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 471–476.
[21] Ivar Troost, Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2020. DatAR: An immersive literature exploration environment for neuroscientists. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 55–56.
[22] Ivar Troost, Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2021. Exploring relations in neuroscientific literature using augmented reality: A design study. In *Designing Interactive Systems Conference 2021*. 266–274.
[23] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.
[24] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.
[25] Mei-Ying Wu, Han-Ping Chou, Yung-Chien Weng, and Yen-Han Huang. 2011. TAM-2 based study of website user behavior-using web 2.0 websites as an example. *WSEAS Transactions on Business and Economics* 4, 8 (2011), 133–151.
[26] Boyu Xu. 2024. Supporting Neuroscience Literature Exploration by Utilising Indirect Relations between Topics in Augmented Reality. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 457–460.
[27] Boyu Xu, Ghazaleh Tanhaei, Lynda Hardman, and Wolfgang Hürst. 2024. DatAR: Supporting Neuroscience Literature Exploration by Finding Relations Between Topics in Augmented Reality. In *International Conference on Multimedia Modeling*. Springer, 295–300.