Designing and Testing an AI System to Understand Student Muddy Card Responses

THOMAS JAMES ELTON

[redacted]

Supervisor: Dr. Jonathan Kay Kummerfeld

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Advanced Studies (Honours)

School of Computer Science
The University of Sydney
Australia

30 May 2025

THE UNIVERSITY OF
SYDNEY

Student Plagiarism: Compliance Statement

I Thomas James Elton certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This work is substantially my own, and to the extent that any part of this work is not my own I have indicated that it is not my own by acknowledging the source of that part or those parts of the work.

I also acknowledge the use of generative artificial intelligence (GAI) tools in this thesis. In particular, GAI was used to assist with coding and LaTeXquestions. However, ideas were still my own. GAI was also used for understanding particular concepts in literature, and occasionally used for finding research papers. I also acknowledge the use of Grammarly for writing assistance.

Name: Thomas James Elton

Signature: [redacted] **Date**: 29/05/2025

Abstract

Muddy cards are an active learning technique where students record the most confusing point from a lecture. They improve students' memory retention and self-efficacy, and aid in developing their understanding of course material. Despite their many benefits, reading and analysing muddy card responses requires a substantial time commitment from teachers. In this thesis, we developed a system that streamlines the process of collecting and analysing muddy cards. This system uses embedding algorithms and agglomerative clustering to enable teachers to quickly identify the most commonly occurring confusing points in their lecture. We have also implemented a 'student-assisted' approach, where students identify peer responses that are semantically similar to what they entered to improve the quality of clustering.

To explore the performance of the embedding models central to the muddy card system, we manually clustered 2,327 muddy cards (split into eight samples). We developed a new clustering metric called the 'student questions answered satisfaction' (SQAS) score to better measure clustering quality for muddy card applications. Using the SQAS score, we found that no embedding model markedly outperformed the others. This is in contrast to embedding benchmarks, which do reveal differences between embedding models. When incorporating the 'student-assisted' data in clustering, the 'multi-evidence' approach consistently improves SQAS score performance.

To investigate the effectiveness of our muddy card system, we conducted a user study with 20 units at the University of Sydney. On a two-week cycle, teachers would use our clustering interface, and as a control, a baseline version with simple options for alphabetically sorting responses. Students and teachers completed an end-of-study survey, and 13 teachers agreed to be interviewed regarding their experience with the system. During the study, there was a low student muddy card response rate. Through interviews, teachers explained that students do not generally complete optional activities, especially if they do not see the benefit in participating. We found that lecturers tended to prefer the clustering interface over the baseline interface.

To build upon this study's findings, we propose a follow-up user study where the muddy card system is modified to be more conducive to live lecture analysis. We hypothesise that this will increase the low response rate. The follow-up study will also include international universities to investigate the effect of societal differences on the uptake of muddy cards.

Acknowledgements

Firstly, I would like to thank the many different lecturers and students who tested our muddy card system. This research would not have been possible without their involvement!

On a more personal note, I would like to thank all the people who have helped me develop into the researcher I am today. It has been a blessing to have so many people in my life who have supported and encouraged me to pursue science. I am grateful to all my high-school science teachers who encouraged me to be curious, especially through student research projects. I would like to specifically thank my first high school science teacher, Mr. Tom Riley. As I reflect on my path to today, Mr. Riley played a pivotal role in developing my love for science, and I hope to encourage students like he did for me one day!

I would also like to thank those who have devoted substantial time to supervising me on large research projects. I would like to thank Dr. Roger Kennett, Dr. Stephen George-Williams, Dr. Reyne Pullen, and Dr. Jonathan Kummerfeld. I appreciate their supervision and mentorship, as well as the extensive time they have invested in developing me into the researcher I am today. I always felt 100% supported under their supervision. I am very thankful for Dr. Kummerfeld's supervision and mentorship over this thesis. As a supervisor, he went above and beyond, and I am grateful for his support!

My research inspirations stem from a passion for teaching. I am grateful to those who have helped me develop as an educator over the last few years at the different institutions I've had the pleasure of working at. In particular, I am very grateful to Prof. Diana Warren for encouraging and supporting me as I pursue tertiary data science teaching.

I am grateful to the family and friends who have supported me throughout life, but especially during Years 11-12 and university. There were many ups and downs, but I knew people were always there for me. I am especially grateful for my Mum (Colleen), Dad (David), and brother (Darcy), who have always been there for me. I am also grateful for all the friends I have made over my studies, which even crossed international boundaries after my year abroad. A shoutout to James Wurzer, who, during one of our calls, helped me load test the muddy card system by opening as many tabs as we could on our computers. Once again, my friends have always been there for me.

I am grateful to Jesus, my firm foundation. *John 3:16*, *Psalm 23:4*, *Joshua 1:9*.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

Muddy cards are an active learning technique where, following a lecture, students record the most confusing (i.e. muddiest) point from the lecture. Examples of muddy card responses include 'Why is sunlight needed in photosynthesis?' or 'How is merge sort considered a divide+conquer algorithm?' These muddy card responses are collected and subsequently analysed by the instructor. The instructor will read each muddy card response and arrange them into groups of similarity, where all the muddy cards in a group describe the same underlying confusion from the lecture (Mosteller, 1989). With the most confusing points identified, the instructor will often address the most commonly raised points of confusion through methods such as clarifying in the following lecture (King, 2011) or on the course website (Hall et al., 2002), or making changes for future offerings of the course (Edström et al., 2007).

However, while muddy cards only take a few minutes to administer, they require a substantial amount of time to analyse (Adams, 2004; Kessler and Nadjm-Tehrani, 2002). For a class of approximately 50 students, it took an instructor 30 to 45 minutes to summarise the responses (Mosteller, 1989). This makes muddy cards infeasible when considering the large class enrolments at many universities. For example, many classes at the University of Sydney contain hundreds of students, and some have thousands. It would take too long for instructors to read all the students' responses for classes of these sizes.

In an attempt to address this temporal challenge, this thesis explores our technological muddy card system that streamlines the process of student data collection and instructor analysis. Using our system, after a lecture, students will visit the *student interface*, which is a website where they can record what was least clear to them in the lecture. After this, the student will be presented with peer responses that the system deems similar in meaning to the muddiest point entered by the student. Here, the students aid the instructor in the analysis process by identifying similar responses, which we refer to as the *student-assisted approach*.

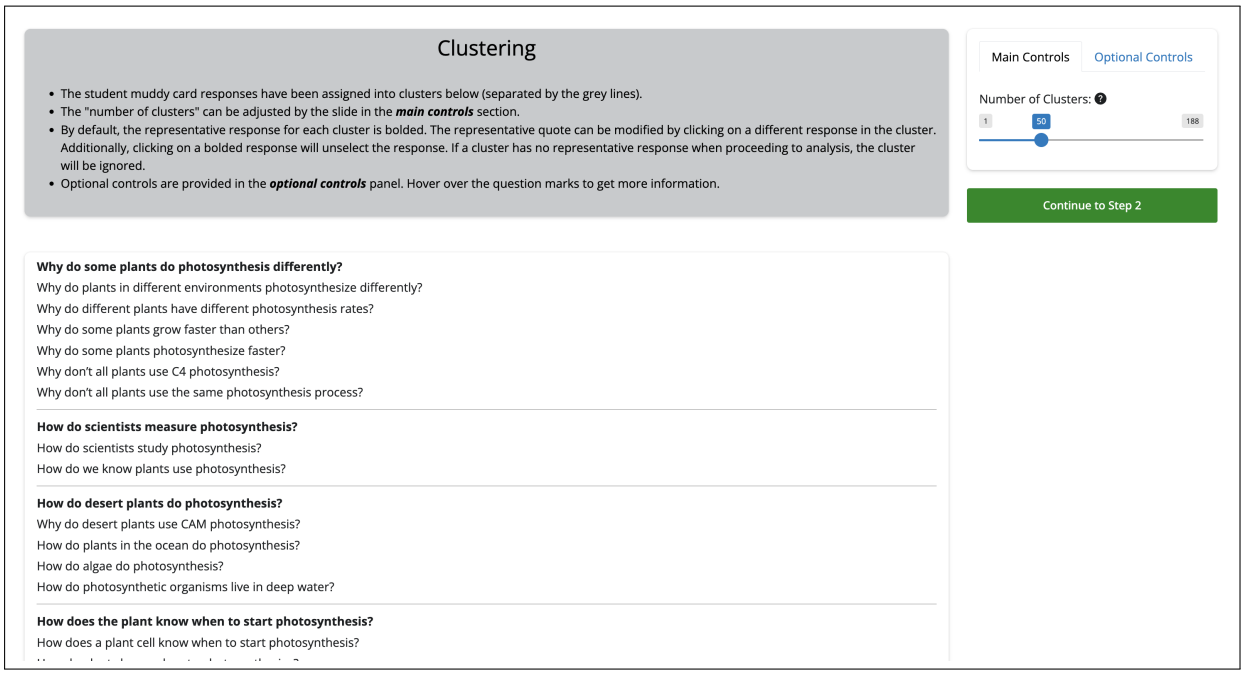


FIGURE 1.1: In the teacher interface, student responses are arranged down the screen, grouped by semantic similarity. Each cluster's representative quote is in bold.

Once the students record their muddy card responses, the instructor will log into the *teacher interface*, which uses artificial intelligence to cluster/group the muddy card responses into groups of semantic similarity (groups of similar meaning). To ensure the instructor remains an active participant in the muddy card analysis, they can use a slider to fine-tune the granularity of the clusters, which adjusts the number of clusters into which the data is split. This slider is also used, as the system cannot perfectly define the clusters, so the instructor is helping to find the correct set of clusters. The system selects one student response from each cluster to be that cluster's representative quote, with the representative quote in bold. This is presented in Figure 1.1.

To validate the muddy card system, one needs to consider both the performance of the internal algorithms and the teacher and student perspectives on the system's effectiveness. Because of this, this thesis will be broadly separated into two research questions. The first research question will explore the internal algorithms, with much attention placed on the performance of the internal natural language processing (NLP) techniques. The second research question will explore the student and teacher perspectives of the muddy card system through an in-the-wild user study. These are expanded on below.

1.1 Research Question 1: Embedding Models and the Student-Assisted Approach

The first research question will investigate the internal algorithms that underpin the muddy card system. In particular, the muddy card system utilises sentence embedding models, which are methods used to represent a sentence's meaning numerically. However, the performance of these models can vary widely, with different models being suited for different use cases (Tang and Yang, 2025). To evaluate the performance of different embedding models for our muddy card use case, we will compare how different models perform when clustering muddy card responses collected from different university lectures that we have manually annotated/clustered. Additionally, as our system included the 'student-assisted approach', Research Question 1 will explore how the student-identified similar points can be used to augment the clustering process. Distilled into a single research question, Research Question 1 is:

(1) How well do sentence embedding models perform when clustering student muddy card responses into groups of semantic similarity with the 'student-assisted approach'?

To explore this research question, two main sources of data were used. The first was a collection of muddy card responses from the semester 1 2024 offering of the University of Sydney's NLP course, where students regularly completed muddy cards. The second source came from students using the muddy card system as part of the user study that will be discussed in Research Question 2.

The rationale for the 2024 NLP data was to preliminarily analyse different embedding models to inform the model used as part of the in-the-wild deployment in the user study. This was achieved by manually annotating two lectures' worth of muddy card responses and comparing them to the clusters produced using different SBERT, OpenAI, and Voyage AI embedding models. The clustering method used was agglomerative clustering, which involved hyperparameter tuning. To measure the concordance between the algorithmically generated clusters and the manual labels, adjusted rand index (ARI) and adjusted mutual information (AMI) scores were used. Additionally, we developed a new metric called the 'student questions answered satisfaction' (SQAS) score, which we argue better captures clustering quality for muddy card applications. The 2024 data was also used to make other minor system modifications, such as determining a cosine similarity threshold to control which peer responses should be relayed to students under the student-assisted approach.

While the 2024 data allowed for adjustments to the system in preparation for the user study, its limitations included that the data could not be publicly released, spanned only a computer science unit, and did not include the student-assisted approach. This motivated using the data collected as part of the user study. Using this data, six different muddy card samples were manually clustered. Three of these included data pertaining to the student-assisted approach, and two of these datasets were manually coded by two researchers to understand researcher concordance and to determine gold standard clusters through adjudication. Of the muddy card samples, it was shown that no single embedding model dominated all tasks when considering the SQAS score.

Different methods were proposed to explore how the student-assisted approach could be used to augment clustering. Each method differed in how it enforced the student-identified links between responses. Agglomerative clustering was still used, with the added constraint that muddy card responses connected by links enforced by a particular method must be kept together. It was found that none of the proposed methods led to a higher clustering performance when considering ARI. However, two approaches were found to improve clustering performance when considering the SQAS score, which we believe is better aligned for muddy card applications. The student-assisted approach represents a key contribution of this thesis. We are unaware of any prior muddy card research implementing anything similar to this.

Another main contribution of Research Question 1 is our plan to publicly release the students' muddy card responses. When exploring existing clustering datasets, it was found that no datasets shared a similar structure to the muddy card data. Most clustering datasets have sentences split into a relatively small number of clusters (high-granularity clusters). However, when considering muddy cards, the clusters have low granularity, meaning there is a very large number of clusters, with few sentences in each cluster. Hence, a key contribution is to release the manually coded muddy card samples, introducing a benchmark specific to muddy cards.

To summarise, the main contributions for Research Question 1 include:

- Proposed the SQAS score, a new metric to practically capture clustering quality for muddy-card applications.
- Manually clustered eight different muddy card samples to evaluate the performance of different embedding models. Six of these will be released to the public as a benchmark dataset.
- Proposed and tested different methods of how the 'student-assisted approach' data can be used to augment clustering algorithms.

1.2 Research Question 2: User Study

The second research question involves a user study where our muddy card system was deployed in-the-wild for students and teachers to trial. As a human-computer interaction (HCI) system, it is crucial to understand the perception of the system from the perspectives of our major stakeholders: students and teachers. Hence, Research Question 2 is:

(2) How do students and teachers perceive the effectiveness of our muddy card system as it relates to collecting and subsequently analysing common points of confusion?

Previously, different muddy card systems have been proposed. Notably, Glassman et al. (2015) proposed 'Mudslide', a system where students spatially pinpoint their most confusing points on a lecture slide before recording why that point was confusing. Instructors can view a heat-map overlay of the most confusing points and select a region to receive a list of the written responses. Although it received a positive reception from users, it remains unclear how this technology would scale to accommodate large class sizes. Another technological approach is 'CourseMIRROR', proposed by Fan et al. (2017). CourseMIRROR shares many similarities with our system. Both systems have a student interface where students can submit their muddy card points, and both have a teacher interface to view the most common confusing points. A significant difference is that in CourseMIRROR, instructors simply receive a summary of the most common confusing points, and are primarily removed from the analysis phase. On the other hand, in our system, students aid clustering through the student-assisted approach, and the instructor is an active participant in analysis with the ability to adjust the granularity of the clusters through the slider.

A primary goal of Research Question 2 was to understand the perceptions of our system, which builds upon previous systems in literature, among students and teachers. This was achieved using a mixed-methods approach, where students completed a survey with Likert-style questions, and teachers completed lecture-by-lecture surveys, a final survey, and participated in semi-structured interviews. As a control to benchmark our system, teachers experienced a baseline version where the teacher interface was replaced with a simple interface that lists student responses down the screen, with primitive controls to sort responses alphabetically.

Altogether, twenty different units of study agreed to trial the system, encompassing units from six different faculties, with a diverse range of undergraduate and postgraduate units. Overall, the student reception of the system was found to be generally positive, with many students wishing to see muddy cards used in other courses. Despite this, for almost all units of study, the proportion of students filling out muddy

card responses was low. Stemming from the interviews, possible reasons for this are proposed in the results of this thesis. When considering the teacher's perspective, they appreciate the system's simple design and prefer the clustering interface over the baseline. That being said, they generally rated the system's usefulness as neutral, which was likely conflated by the low student response rate. We proposed modifications to the system to address the low response rate, with a follow-up study planned.

To summarise, the main contributions for Research Question 2 include:

- Proposing and evaluating our muddy card system.
- Using the survey and interview results to explore the reasons for the low number of student muddy card responses.
- Proposing future muddy card system implementations, and ideas for a follow-up study.

1.3 Thesis Outline

The remainder of this thesis is broken into topical chapters. In Chapter 2, we will review the relevant literature for this project. This will begin by examining the literature on muddy cards, presenting them as a sound pedagogical and active learning technique. We will then explore literature specifically relevant to Research Question 1, which will involve discussing sentence embeddings, calculating similarity between embeddings, and various clustering algorithms. This will lead to a discussion on various clustering datasets, which will motivate our intention to release our muddy card dataset publicly. The discussion will then pivot to the literature relevant to Research Question 2, by discussing different methods of evaluating HCI systems and exploring the existing technological muddy card systems in detail.

In Chapter 3, we will focus on the methodology used in this study. This chapter begins with an in-depth overview of the muddy card systems' front-end and back-end, and how they were deployed in-the-wild. We then discuss the methods used for benchmarking the various sentence embedding models. Finally, we will describe the user study and the mixed-methods approach to analysis.

The results of this thesis will be split into two chapters. Chapter 4 will focus on the results of Research Question 1, with Chapter 5 presenting the results of Research Question 2.

Finally, Chapter 6 will include a discussion on the limitations of this study and the direction for future work. Here, a follow-up study will be proposed. We conclude with Chapter 7, which presents the main conclusions of the results presented in this thesis.

CHAPTER 2

Literature Review

As mentioned in the introduction, this thesis is guided by two research questions:

(1) How well do sentence embedding models perform when clustering student muddy card responses into groups of semantic similarity with the 'student-assisted approach'?
(2) How do students and teachers perceive the effectiveness of our muddy card system as it relates to collecting and subsequently analysing common points of confusion?

This chapter will review the literature related to these research questions. The first section will explore literature on the efficacy of muddy cards as a pedagogical tool. This is an essential first step, as the traditional approach to muddy cards motivates the entire muddy card system.

The following four sections of this chapter (Sentence Embeddings, Cosine Similarity, Hierarchical Clustering and Clustering Benchmarks) will review the literature surrounding the natural language processing (NLP) techniques relevant to Research Question 1. These NLP techniques are fundamental to the internal algorithms used in our muddy card system.

The sixth section will briefly explore the literature on evaluating human-computer interaction (HCI) systems, which is relevant when investigating Research Question 2. Finally, the seventh section will review other muddy card technological innovations and how our system attempts to improve on existing systems.

2.1 Muddy Cards

Many educators deem traditional didactic lecture-based approaches as ineffective when compared to active learning models (Jungst et al., 2003). Active learning is characterised by any task that "involves students in doing things and thinking about the things they are doing" (Bonwell and Eison, 1991). Less

emphasis is placed on the passive transmission of information from educators, with more emphasis placed on students' manipulating, applying, analysing and evaluating different ideas (Edström et al., 2007), resulting in students remembering more of what they have learned (Prince, 2004).

Muddy cards are an active learning technique that involves students recording the most confusing (muddiest/muddy) point of a lecture. The first published case of muddy cards was by Mosteller (1989), a statistics professor at Harvard. During the final three or four minutes of class, Mosteller solicited responses from his ~50-student class to three questions: What was the most important point in the lecture? What was the muddiest point? What would you like to hear more about? After spending around 30 to 45 minutes analysing the student responses, Mosteller addressed identified issues by providing handouts and explanations during class time. In this regard, muddy cards can be viewed as minor assignments that are not intended for measuring student performance but rather as a means of data for assessing student learning (Adams, 2004).

Since Mosteller's seminal paper, the efficacy of muddy cards has been explored in different university settings. Hall et al. (2002) discusses the addition of active learning techniques in the Unified Engineering course at Massachusetts Institute of Technology. One of the key techniques added was muddy cards, where several faculty members recorded that after seeing their benefits, they would never teach without them. In response to the muddy points students raised, one instructor began responding to students' muddy cards via the course website, where student feedback on this approach was "extremely favourable". Student course evaluations indicated that 82% of students rated muddy cards as very or somewhat effective. Students appreciated receiving feedback quickly, and there was the perception that by using muddy cards, instructors cared for their students' understanding of course material.

Similar results were found by Carberry et al. (2013) when investigating faculty members' use of muddy cards and the value students associate with them. A common theme among instructors was that muddy cards opened a dialogue between instructors and students. One instructor explained that using students' own words to clarify muddy points made students feel "a part of the learning team". From student survey results, the majority found that muddy cards positively impacted their class experience. 77% of students desired to see muddy cards in future classes and felt that the cost of filling out muddy cards was minimal.

Muddy cards allow for lively and responsive teaching, where the muddiest points can be the organising element of an ongoing dialogue with students (Adams, 2004). Student self-efficacy will increase as

muddy cards will identify areas of strength and weakness, allowing new information to be based on prior knowledge that they are already confident in, motivating students to learn (Krause et al., 2013). Muddy card responses have also been used to gauge whether students have met prerequisite knowledge requirements before commencing a course (Willcox and Bounova, 2004), as well as being used to collect student feedback to evaluate a tertiary course (Kessler and Nadjm-Tehrani, 2002).

Despite the many benefits and applications of muddy cards explored in literature, a recurring limitation is that while they only take a few minutes to administer, they require a substantial amount of time for the instructor to analyse (Adams, 2004; Kessler and Nadjm-Tehrani, 2002). Mosteller (1989) spent 30 to 45 minutes summarising the muddy card responses from his class of around 50 students. Hall et al. (2002) explains that when implementing muddy cards into a class of around 60 students, teachers raised concerns about the time it took to respond to muddy cards, with one instructor restricting the time they spent engaging with muddy cards to 1-1.5 hours. The large unit enrolments at many institutions compound this limitation regarding instructors' time. For example, some courses at the University of Sydney have enrolments greater than a thousand students. For these classes, it is impractical to analyse muddy card responses manually.

Hence, the main objective of our muddy card system is to overcome this temporal challenge by using NLP techniques to assist an educator in rapidly analysing muddy card responses. A fundamental step towards this objective is converting students' muddy card responses into sentence embeddings.

2.2 Sentence Embeddings

The first challenge when dealing with natural language in computing systems is numerically representing written sentences. In NLP, vector semantics is the standard method used to represent word meaning. Word meaning is expressed as a point in a multidimensional space with semantically similar words appearing spatially near each other (Jurafsky and Martin, 2024). These vector representations of words are referred to as *word embeddings*. Similarly, entire sentences can be vectorised with semantically similar sentences appearing spatially close. These are referred to as *sentence embeddings*. In this project, we used and tested existing embedding models to convert student muddy card responses into sentence embeddings.

Historically, many different methods of generating word and sentence embeddings have been proposed, such as GloVe for word embeddings and TF-IDF for document embeddings (a sentence can be considered a short document). More recently, novel embedding models have been proposed, resulting in state-of-the-art performance for generating word embeddings. The first section will explore some of these recent developments and how newer models adjust the training approaches of previous models to boost performance. This will lead to discussing Sentence-BERT, a commonly used framework for developing sentence embeddings. Finally, we will discuss how commercial companies have developed proprietary sentence embedding models with impressive performance.

2.2.1 Word Embedding Models

Historically, word embeddings were static. Regardless of the context of neighbouring words, the embedding for a word was identical for a given method. This is not ideal when considering the complexity of the English language, and how the same word can be used to represent different phenomena (for example, the word 'bat' could refer to the animal or a piece of sports equipment). To address this, various models have been proposed to develop contextual word embeddings. A pivotal paper was when Devlin et al. (2019) presented the BERT model. BERT uses a transformer-based architecture, which relies on attention mechanisms to determine the dependencies between the input and output (Vaswani et al., 2017). BERT was compared to previous embedding models using the General Language Understanding Evaluation (GLUE) benchmark, which contains a series of natural language understanding tasks split across three categories: single sentence input tasks, detecting semantic similarity tasks, and natural language inference tasks (Wang et al., 2018). BERT was found to have state-of-the-art performance on the GLUE benchmark.

In BERT, model training involves masked language modelling (MLM), where 15% of symbols are masked, with the model attempting to predict them (Devlin et al., 2019). Newer papers experimented with different training methods. One example is XLNET, which instead uses permuted language modelling (PLM), where, during training, tokens are permuted, with the tokens at the end of the sequence predicted by considering the preceding tokens. This training method results in a small absolute increase of 1.4% over BERT when considering the GLUE benchmark (Yang et al., 2019). Shortly after, MPNET was created, which unifies the MLM and PLM approaches to training by combining permutation and masking during training. This method led to even larger performance gains, with MPNET outperforming BERT by an absolute 4.8% and XLNET by an absolute 3.4% (Song et al., 2020). Despite all three of

these word embedding models being transformer-based, it is evident that different training approaches lead to varying performance.

2.2.2 Sentence-BERT

However, BERT, XLNet and MPNet are not designed for creating sentence embeddings, but rather word embeddings. Researchers have devised methods to create sentence embeddings using these models, but these have typically led to poor sentence embeddings (Reimers and Gurevych, 2019). Hence, Reimers and Gurevych (2019) developed Sentence-BERT (SBERT), which was created by modifying the pre-trained BERT network using a Siamese neural network structure. A Siamese neural network consists of two identical neural networks that share the same weights (Chicco, 2021). In SBERT, the Siamese neural network consists of twin BERT models. Model training involves passing two sentences into each BERT model, where for each sentence, the word embeddings produced by BERT are pooled. These sentence embeddings are then compared with cosine similarity, with an objective function updating the weights in the Siamese network.

The result of training is a model optimised for generating contextual sentence embeddings, with semantically similar sentences having similar embeddings. Hence, produced embeddings can be easily tested for similarity through methods such as cosine similarity (elaborated in Section 2.3). SBERT led to state-of-the-art performance in generating sentence embeddings. When comparing SBERT to BERT on unsupervised semantic textual similarity tasks (STS), SBERT outperforms by 20.08%. Further, common embedding methods were tested against SentEval, a toolkit used for evaluating the quality of sentence embeddings through tasks around classification, inference, and sentence similarity (Conneau and Kiela, 2018). Using SentEval, SBERT outperformed BERT by 2.47%, as well as outperforming all other models.

As a technique, the ideas behind SBERT can be applied to different word embedding models. The `sentence_transformers` (SBERT) Python package lets users choose from over 5,000 pre-trained sentence models uploaded to Hugging Face (Reimers and Gurevych, 2019). Some of the highest performing models use an MPNet backbone, such as the `all-mpnet-base-v2` model (Reimers and Espejel, 2021), which at the time of writing, has the highest accuracy of all featured models in the `sentence_transformers` documentation[1].

[1]This documentation can be found in the 'Original Models' section of: https://sbert.net/docs/sentence_transformer/pretrained_models.html

SBERT will be one of the frameworks used in this project to transform the students' muddy card responses into sentence embeddings. While the SBERT models on Hugging Face are freely available, commercial companies are increasingly facilitating sentence embedding using their proprietary models.

2.2.3 Proprietary Sentence Embeddings (OpenAI and Voyage AI)

Many companies, such as OpenAI and Voyage AI, offer embedding services. Users can pass a sentence through an application programming interface, where the company's large language model transforms the user-specified sentence into a sentence embedding. As these models are proprietary, little information about the training methodology used to create their sentence embedding models is available. For example, at the time of writing, the last available article mentioning "embeddings" on OpenAI's research index portal[2] was a 2022 paper by Neelakantan et al. (2022). The model that they describe in this paper is the `text-similarity-001` class of embedding models (Reimers, 2022), which were deprecated in January 2024[3].

Despite this, it was possible to measure the relative performance of these companies embedding models when looking at their score on the Massive Text Embedding Benchmark (MTEB) global leaderboard (Muennighoff et al., 2023) (MTEB will be discussed in greater detail in Section 2.5). At the time of writing, per the MTEB leaderboard, the top-ranking OpenAI model is `text-embedding-3-large` with an average task type score of 62.15. The top-ranking Voyage AI model is `voyage-3-lite` with a score of 57.25[4]. These are both higher than SBERT's `all-mpnet-base-v2` model, which has an average task type score of 54.45 (this average excludes the classification task score, which was missing).

This section explored the different models used to convert sentences to sentence embeddings. In this project, we will test different SBERT, OpenAI and Voyage AI embedding models on muddy card datasets we have curated as part of this thesis. This is because the performance of these models on different benchmarks (such as MTEB, GLUE, SentEval) does not indicate their performance on domain-specific tasks (Tang and Yang, 2025). This will be expanded on in Section 2.5. Since sentence embeddings aim to capture a sentence's meaning, it is essential to have a method for finding the similarity between two embeddings. Cosine similarity can be used to achieve this.

[2]The portal can be found at https://openai.com/research/index/

[3]The list of deprecated OpenAI models can be found at https://platform.openai.com/docs/deprecations

[4]These results were found from the MTEB leaderboard: https://huggingface.co/spaces/mteb/leaderboard. The specific benchmark used was MTEB(eng,v2). Only OpenAI and Voyage AI models with a score for each of the seven embedding tasks were considered.

2.3 Cosine Similarity

Cosine similarity is often used to determine the similarity of two sentence embeddings. For two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, cosine similarity is defined as:

$$\text{cosine\_similarity}(\mathbf{v}, \mathbf{w}) = \frac{\mathbf{v} \cdot \mathbf{w}}{|\mathbf{v}||\mathbf{w}|} \tag{2.1}$$

As evident from the formula, the normalised dot product acts as a similarity metric as it will result in a high value when both vectors have large values in the same dimensions. When the vectors point in the same direction, cosine similarity returns the value 1. When the vectors point in opposite directions, cosine similarity returns the value -1. Hence, cosine similarity measures the similarity of vectors on a scale of -1 through 1.

Cosine similarity will be used in the internal algorithms for the student interface as well as the teacher interface when clustering student responses (clustering is elaborated on in the next section). In the student interface, one experimental condition involves the 'student-assisted approach', where students were prompted to indicate whether selected muddy card responses written by other students were semantically the same as what they entered. The system chooses these candidate responses by finding the most similar responses through cosine similarity. While cosine similarity is easily applied to finding the pairwise similarity between sentence embeddings, for the muddy card system, we want a method to split a collection of sentence embeddings into a series of clusters. This can be achieved using clustering algorithms.

2.4 Hierarchical Clustering

Clustering is an unsupervised machine learning algorithm that separates a data set into subgroups that share similar characteristics. Hierarchical clustering is a subset of clustering algorithms that groups data into a tree of nested clusters. This tree is typically visualised as a dendrogram, a branching diagram showing the hierarchical relationship between elements. The dendrogram reveals how smaller clusters, which usually begin with a single element, are combined until the root node, which is the point where all previous clusters are combined into one. There are two common hierarchical clustering algorithms: agglomerative clustering and HDBSCAN.

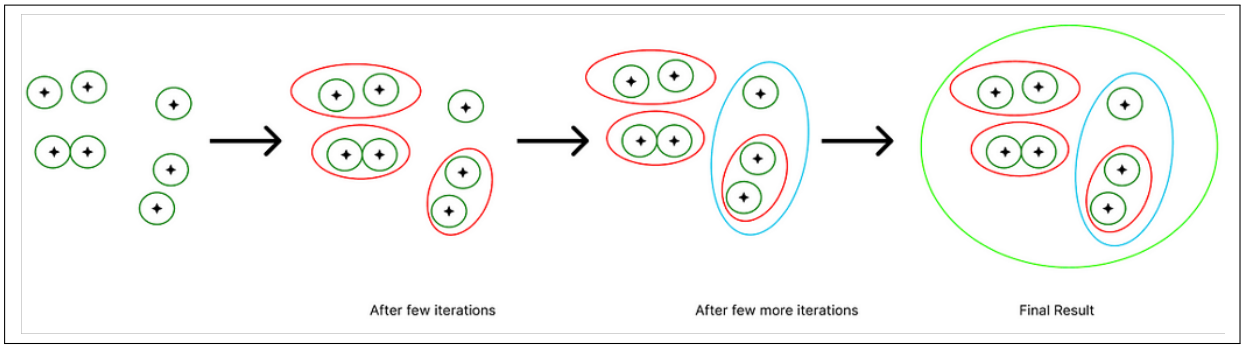


FIGURE 2.1: How hierarchical clusters are formed through the iterations of the agglomerative clustering algorithm (Bathula, 2023).

2.4.1 Agglomerative Clustering

Agglomerative clustering is a simple algorithm used to generate hierarchical clusters. Consider a data set with n elements, and at first, all elements are initially placed in their own respective clusters (see the left-most part in Figure 2.1). Agglomerative clustering occurs through an iterative algorithm where at each step, the two clusters that are most similar to each other are fused. We start with n clusters, and after the first iteration, the two most similar clusters are fused, meaning we now have $n - 1$ clusters. From the current $n - 1$ clusters, we fuse the two most similar clusters, meaning we now have $n - 2$ clusters. This process continues until all elements belong to the same 1 cluster, as evident in Figure 2.1.

Considering which two clusters are most similar for each iteration is based on a distance metric and linkage strategy. Distance metrics provide the mathematical way to measure the distance between two points. Examples of distance metrics include cosine, Euclidean, and Manhattan distance. The linkage strategy determines how the distance metric is used when considering two clusters. For example, the 'average' linkage strategy finds the distance between each point in two clusters and averages these distances. This is repeated for every combination of clusters, and the two clusters with the smallest average distance are fused. Common linkage strategies include single, average, median, centroid, complete and Ward (Tokuda et al., 2022). In this project, we will trial all previously mentioned distance and linkage strategies, except for the median and centroid linkage strategies. This is because these linkage strategies are not available in `scikit-learn`'s implementation of agglomerative clustering, which is used in our implementation of the muddy card system[5].

[5]The documentation for `scikit-learn`'s AgglomerativeClustering model is available here: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

The result of agglomerative clustering is a tree of nested clusters. The leaves represent the initial state where all elements are in their own cluster, and the root represents the point where all smaller clusters are merged into one single cluster. Hence, given a value for $k \in [1, n]$, you can extract k clusters from the tree by finding the k-th splitting point and the k clusters that exist at this point. Such an approach is helpful as it allows flexibility in changing the value for k without re-running the entire clustering algorithm. HDBSCAN similarly produces a tree-like structure using a different approach.

2.4.2 HDBSCAN

Ester et al. (1996) proposed the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, which clusters data by placing each dense data region into a cluster. The key idea behind DBSCAN is that for each point within a cluster, the neighbourhood (given a radius) must contain at least some threshold (density) of points. Points below this threshold are considered noise and will not be assigned to a cluster. DBSCAN itself is not a hierarchical clustering approach, and the final clusters produced are controlled by fine-tuning the epsilon-neighbourhood parameter, that is, the neighbourhood radius parameter.

Proposed by Campello et al. (2013), HDBSCAN builds upon DBSCAN by creating a density-based hierarchy from which clusters can be extracted. HDBSCAN set new state-of-the-art performance when compared against other density-based clustering methods. McInnes and Healy (2017) have since optimised HDBSCAN moving from an $O(n^2)$ to approximate $O(n \log n)$ runtime. This implementation was shown to perform among the fastest of other clustering techniques. While HDBSCAN could be appropriate for the muddy card system, we have chosen to focus on agglomerative clustering. This is because initial testing with HDBSCAN found that it was difficult to tune.

In this project, we will use clustering algorithms to separate the muddy card sentence embeddings into clusters. As similar sentences will have geometrically similar sentence embeddings, these clusters will contain muddy card responses that are semantically similar, allowing users to quickly ascertain the common muddy card responses by scanning the produced clusters. Additionally, hierarchical clustering algorithms make it easy to expand or reduce the number of clusters produced as the entire cluster hierarchy has been created. This is desirable in HCI systems. Input controls can allow the user to adjust how many clusters to split the data into, meaning that human intuition can be used to most accurately cluster the data so that semantically similar muddy card responses are appropriately clustered.

However, even though hierarchical clustering algorithms construct the entire cluster hierarchy, this does not imply that the underlying clusters are of high quality. The clusters produced will be heavily dependent on the sentence embeddings, and different embedding models have been shown to excel at different tasks (i.e. embedding models that are good on semantic textual similarity tasks may struggle with clustering tasks) (Muennighoff et al., 2023). Hence, it is vital to test different embedding models on clustering benchmarks.

2.5 Clustering Benchmarks

To investigate the performance of sentence embedding models, the Massive Text Embedding Benchmark (MTEB) is commonly used (Muennighoff et al., 2023). Before MTEB's release, embedding models were typically evaluated on semantic textual similarity (STS) tasks. However, STS is only one application of sentence embeddings, and STS tasks have been shown to correlate poorly with other embedding-related tasks (Muennighoff et al., 2023). MTEB alleviates this by considering eight different tasks when assessing sentence embedding models: *clustering*, bitext mining, retrieval, STS, summarisation, re-ranking, classification, and pair classification.

The MTEB paper benchmarked 33 models that claimed state-of-the-art performance on various embedding tasks. None of these models had a performance that dominated all tasks (Muennighoff et al., 2023). Since the MTEB benchmark was first released, newer variants of MTEB have been released featuring more task types and data sets, as well as benchmarks focused on languages other than English. When inspecting these newer variants, it seems that the MTEB paper's 2023 claim that no embedding model dominates all tasks no longer holds[6]. Regardless, in this thesis, we are primarily interested in sentence embedding models' performance for clustering, as the teacher's interface for the muddy card system involves the student responses being arranged into semantic clusters. When observing the MTEB ranking for the clustering task, there is some variability where some models offer standout performance for clustering despite having an overall lower ranking.[7]

[6]The MTEB leaderboard can be found at: https://huggingface.co/spaces/mteb/leaderboard. The MTEB leaderboard was inspected at the time of writing this thesis. In particular, we observed MTEB(eng, v2) and MTEB(Multilingual, v1). The newly released `gemini-embedding-exp-03-07` model (Kilpatrick et al., 2025) appears to dominate almost every benchmark task type. We cannot use this model in this research project, as we did not gain ethics approval to use non OpenAI and Voyage AI models.

[7]For example, at the time of writing, the `SFR-Embedding-2_R` model (Meng et al., 2024) is ranked 13th overall, but fourth for clustering on MTEB(eng, v2).

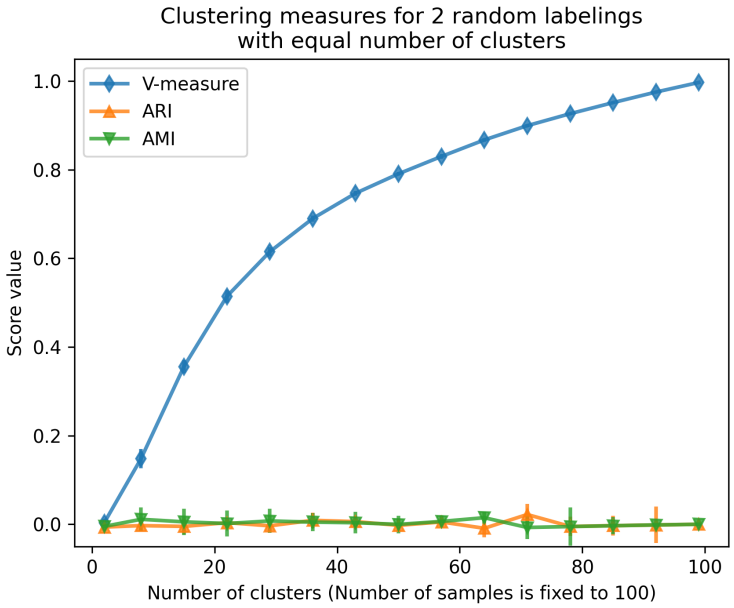


FIGURE 2.2: As the number of clusters increases, so does V-measure. ARI and AMI are not sensitive to the number of clusters[8].

In MTEB, an embedding model's clustering performance for each dataset is determined by comparing the true and unsupervised labels generated using a mini-batch k-means model. Here, the batch size is 32, and k is the number of unique labels in the current dataset being analysed (Muennighoff et al., 2023). The model's performance is recorded using V-score, an entropy-based measure that considers the homogeneity and completeness of the actual and predicted labels (Rosenberg and Hirschberg, 2007). However, V-score is inappropriate for the muddy-card use case due to the low ratio of examples to clusters. As students can articulate any muddy card point, our initial data analysis revealed that there was a very large set of clusters for a small number of responses. Additionally, many student responses were completely distinct from their peers, meaning there were many isolated clusters. The presence of many smaller clusters can inflate homogeneity and completeness, which are used when calculating V-score. This is empirically verified when observing Figure 2.2, and how as the number of clusters increases, so does V-score.

[8]This graph was derived from a demonstration included in the clustering documentation on `scikit-learn` (Pedregosa et al., 2011). The original graph and the code that was used to generate it can be found at https://scikit-learn.org/stable/auto_examples/cluster/plot_adjusted_for_chance_measures.html. We slightly modified this code to ensure that the data is always split into the number of clusters indicated on the x-axis. This differs from the `scikit-learn` implementation which uses a probability density function to assign the points to clusters, which means, even if you want to randomly split the data into n clusters, it is possible that the data is not actually placed into n clusters (it could be less). We only plot the x-axis values from 2 to 99, as at 1 and 100, all metrics would have a value of 1 (e.g. at 100 clusters, with 100 data points, all points lie in an individual cluster).

Hence, in this paper, we will instead opt for adjusted rand index (ARI) and adjusted mutual information score (AMI). ARI is a chance-adjusted metric for computing the similarity between two clusterings by considering the number of pairs assigned to the same or different clusters (Hubert and Arabie, 1985). Similarly, AMI is chance-adjusted and indicates the similarity between two clusterings by quantifying how much knowing one clustering reduces our uncertainty about the other (Vinh et al., 2010). Unlike V-score, in Figure 2.2 we see that ARI and AMI are not sensitive to the number of clusters.

MTEB was originally released with 11 clustering datasets for the English language (Muennighoff et al., 2023), but this has since grown to 20. Table 2.1 summarises most of the English clustering datasets that have been added to MTEB.

We can see that different datasets have different granularities when considering the clustering datasets in Table 2.1. For example, the 'ArxivClusteringS2S' MTEB dataset is a collection of paper titles on Arxiv. One way to analyse clustering performance using this dataset is to make the ground-truth labels the main category under which each Arxiv paper falls. Observing the label examples of "cs.IT", "cs.OS", "cs.DS" and "physics.gen-ph", the main category is the first part of the label ("cs" and "physics"). Another approach is to consider each Arxiv paper title's secondary category. For the examples above, the secondary categories are "IT", "OS", "DS" and "gen-ph" (Muennighoff et al., 2023). If we consider clustering as a spectrum ranging from coarse to fine-grained clusters, the first approach is closer to coarse-grained clustering, with the latter being a more fine-grained approach to clustering.

When using our muddy card systems teacher interface, the system will need to break down the muddy points from a lecture into clusters of semantic similarity. Lectures typically only cover a few specific topics, meaning clusters must be highly fine-grained in order for them to be useful. This is to ensure that the teacher knows exactly what the confusing point of the lecture was. When observing Table 2.1, most datasets lean towards more coarse-level clustering. While some datasets do have many labels (e.g. ArxivClusteringS2S, RedditClusteringP2P.v2, StackExchangeClustering), the label-to-example ratio is still relatively large. However, as previously mentioned, our initial data analysis revealed that muddy card clusters are highly fine-grained, with a relatively smaller label-to-example ratio.

Hence, we believe that the level of clustering we hope to capture using the muddy card system is unlike any other English language clustering dataset. Therefore, releasing an annotated muddy card clustering dataset could prove a valuable contribution towards clustering datasets. There has been some muddy card data previously released by the creators of CourseMIRROR (a technological approach to muddy

Dataset Name	Data Domain	Num. of Examples	Num. of Labels	Labels Manually Assigned?	Label Examples
StackExchange ClusteringP2P †	Web	75,000	610	No	"education", "playstation3", "gpu"
RedditClustering P2P.v2 †	Social, Web	459,389	440	No	"puppy101", "AmongUs", "photography"
ArxivClusteringS2S ‡	Academic	732,723	180	No	"cs.IT", "cs.OS", "cs.DS", "physics.gen-ph"
StackExchange Clustering †	Web	790,910	171	No	"math.stackexchange. com.txt", "christian- ity.stackexchange.com.txt"
ArXivHierarchical ClusteringP2P §	Academic	2,048	129	No	"math", "astro-ph", "soft"
MedrxivClustering (S2S/P2P).v2 *‡	Academic, Medical	37,500	51	No	"epidemiology", "allergy and immunology"
RedditClustering †	Social, Web	420,464	50	No	"Christianity.txt", "australia.txt"
BuiltBench ClusteringP2P #	Engineering	4,577	35	No	"skin", "control", "Flow Moving Device"
BuiltBench ClusteringS2S #	Engineering	3,815	31	No	"skin", "control", "Flow Moving Device"
TwentyNewsgroups Clustering ¶	News	18,846	20	?	"sci.space", "comp.graphics"
BigPatent Clustering.v2 ‖	Legal	1,341,362	9	No	"Human Necessities", "Chemistry; Metallurgy"
SIB200 ClusteringS2S ∇	News	1,004	7	Yes	"science/technology", "travel", "politics"
MasakhaNEWS Clustering(S2S/P2P) *‡	News, Non-fiction	4,729	6	No	"business", "entertainment", "health"
BiorxivClustering (S2S/P2P).v2 *‡	Academic	?	?	No	"biochemistry", "genetics", "microbiology"
WikiCitiesClustering⊕	Encyclopaedic	6,458,670	?	?	?

*(S2S/P2P) indicates that these datasets come in two variants: the S2S version (where data is short sentences), and the P2P version (where data is paragraphs).
† (Geigle et al., 2021), ‡ (Muennighoff et al., 2023), § (arXiv.org submitters, 2024),
# (Shahinmoghadam and Motamedi, 2024), ¶ (Rennie, n.d.), ‖ (Sharma et al., 2019),
∇ (Adelani et al., 2024), ⊕ (Wikimedia Foundation, n.d.)

TABLE 2.1: The main MTEB clustering datasets. Only datasets for the English language, and where sufficient metadata exists were considered. Basic information about the clustering datasets in MTEB can be found here: https://github.com/embeddings-benchmark/mteb/blob/main/docs/tasks.md

cards which will be explored in Section 2.7.3). However, this data set is quite small, and the muddy card points were not manually clustered (Fan et al., 2017).

Additionally, finding the embedding model that performs best on the MTEB clustering datasets does not indicate which embedding model is best for our specific use case (Tang and Yang, 2025), necessitating a dataset to test each embedding model for muddy card clustering performance. Developing high-quality sentence embeddings of muddy card responses that can be appropriately clustered is relevant to Research Question 1 and the entire muddy card system. Poor muddy card sentence embeddings will negatively impact the system's downstream tasks. Hence, this investigation will explore various SBERT, OpenAI and Voyage AI embedding models on manually annotated muddy card data samples we curated to determine which embedding model has the greatest performance for the muddy card system use case.

2.6 User Evaluation

While the NLP techniques explored seem relevant and useful in developing a muddy card system, developing a HCI system involves careful attention to the front-end interface. In this regard, creating a user-centred system design is essential. Among other things, it is important to consider whether a system is efficient, effective, safe, beneficial, easy to learn, easy to remember and easy to use (Issa and Isaias, 2022). The user's perception of a system can be investigated through surveys and interviews.

2.6.1 Surveys

In HCI research, surveys are commonly used to understand people's behaviours, experiences and attitudes with technology. As they are inexpensive and easy to administer, surveys enable the capture of large samples. They can also be used to measure the differences between groups of people, and successive surveys over a period can identify changes in people's attitudes and experiences (Müller et al., 2014).

Survey questions are broadly separated into two types: open and closed-ended questions. Open-ended questions involve participants writing their own answer to a set question. Closed-ended questions involve participants selecting from predefined answers. When writing a survey, it is essential to ensure that question phrasing is free from bias, as bias can lead to measurement error. To mitigate this, researchers often use previously established questionnaires that have been previously assessed to have minimal measurement error (Müller et al., 2014).

This project will use surveys as part of understanding how students and teachers perceive the effectiveness of our muddy card system (Research Question 2). A key strength of surveys in this study is that it allows for greater data collection across the many students and teachers who trial our system. However, a limitation of surveys is that they do not allow for follow-up questions and have difficulty in capturing the user's underlying motivations. Additionally, while surveys can be used to measure general task success, they are not conducive to intricate details being captured, such as why people could not use the system, what missteps occurred that led to a system failure, and what was confusing (Müller et al., 2014)? Hence, this study will also use interviews to gain a deeper understanding of the teachers' perceptions of the muddy card system.

2.6.2 Interviews

Originally formulated by Glaser and Strauss (1967), grounded theory is a structured, yet flexible methodology used to construct an explanatory theory about some phenomenon. In essence, the grounded theory methodology involves an iterative and recursive procedure whereby initial data (such as interview transcripts) is collected and then coded, which in turn leads to further targeted data collection (Chun Tie et al., 2019; Corbin and Strauss, 2008; Glaser and Strauss, 1967; Strauss and Corbin, 1990). A theory about the phenomenon of study begins to emerge from the data that has been collected (Muller and Kogan, 2012). When evaluating HCI systems, grounded theory is commonly used (Muller and Kogan, 2012). In this project, we will use elements of grounded theory when analysing the interviews that explore instructors' experience of the muddy card system.

To qualitatively analyse a document, such as an interview transcript, grounded theory commences with *open coding*. This is the process of writing simple descriptive labels that represent different items in your data. As this process continues, certain codes will recur, which the researcher keeps track of. These open codes are not originally organised into a body of concepts. *Axial coding* is the process where the researcher begins to find relationships between the open codes. When more axial codes are developed, relationships between them lead to clusters of codes that can be named, with the named clusters being referred to as *categories* (Muller and Kogan, 2012).

As grounded theory is an iterative process, as new data is analysed, constant comparison between pre-existing codes allows axial codes and categories to be developed, generating increasingly more abstract concepts and theories (Chun Tie et al., 2019). As these axial codes and categories are determined, the

researcher may need to return to the original data, re-coding it in terms of the newly developed axial codes and categories (Muller and Kogan, 2012).

A defining feature of grounded theory is its approach to sampling. Grounded theory uses *theoretical sampling*, which is where new samples are chosen based on findings from the already collected data. The first sample of data leads to a set of codes and categories, which are used to generate a theory. This theory is then tested by choosing a second sample to abductively test whether the first theory holds. This leads to further coding and the development of a new theory, which leads to new sampling, and so on (Muller and Kogan, 2012). In theoretical sampling, the emphasis is on varying the sample by focusing on new elements, to determine how these elements impact the phenomenon in question (Strauss and Corbin, 1990).

Grounded theories focus on theoretical sampling is a key reason why we are only borrowing elements of grounded theory in this research project, rather than strictly following its methodology. Traditional grounded theories' approach to theoretical sampling is time-consuming and is not logistically feasible within this project's timeframe. This modification is common in grounded theory HCI studies, where the theory construction process usually commences after all data has been collected (Muller and Kogan, 2012). While theoretical sampling is no longer a focus, iterative theory construction is maintained (Muller and Kogan, 2012). When encountering new data during the inference process, researchers will write answers to questions such as 'What do I think is going on here?', 'What have I learned from the new data?', and 'How do I code what I have learned?'. This is broadly referred to as *memo writing* (Muller and Kogan, 2012).

In grounded theory, a researcher is confident that they have a sufficiently large sample by assessing whether *saturation* has been reached, which is the point where subsequent samples are found not to develop the properties of categories. As originally put by Glaser and Strauss (1967), "as [the investigator] sees similar instances over and over again, the researcher becomes empirically confident that a category is saturated." Muller and Kogan (2012) state that they know they have reached saturation when surprises stop happening, i.e. they are bored with incoming data as it does not reveal anything new. In traditional grounded theory, saturation can be reached by continual theoretical sampling. In this project, we do not have the ability to collect further samples if the analysis concludes that saturation was not reached. Instead, we will use the concept of saturation when considering the validity of the interview analysis.

In traditional grounded theory, once saturation has been reached, the final theory is developed through the *storyline* technique, which is the conceptualisation of the core categories identified. The storyline technique links abstract categories together to form the grounded theory (Chun Tie et al., 2019). The previously written memos are assembled into a structure that tells a 'story' or makes an argument for the final grounded theory (Muller and Kogan, 2012).

In this project, we have used elements of grounded theory to develop a theory about how teachers perceive the effectiveness of our muddy card system. As will be evident in the next section, many different muddy card systems exist, and our system will act to resolve limitations in existing systems while introducing novel approaches for analysing muddy cards. Surveys and interviews will help in understanding how students and teachers perceive the effectiveness of our muddy card system.

2.7 Technological Muddy Card Innovations

Many different muddy card technological innovations have been researched. One of the simplest innovations is responding to students' muddiest points via written explanations on class websites or discussion forums (Hall et al., 2002). Pinder-Grover et al. (2011) built on this principle by creating supplemental screencasts in response to students' muddy card responses for a Materials and Manufacturing course (~200 enrolments). When 30% of students deemed a particular concept as muddy, the instructors would create a short video explanation. In this study, supplemental screencasts were additionally made for homework, quiz and exam solutions; yet muddy cards represented a major component of the screencasts, accounting for 9 of 25 videos in 2008, and 15 of 33 in 2009. The authors found that higher engagement over all supplemental videos was statistically associated with higher final course grades (when controlling for grade point average). Further, 86% of students found the muddiest point screencasts were helpful to some extent.

Krause et al. (2013) similarly responded to muddiest points via videos uploaded to YouTube, with 93% of students supporting the inclusion of the videos. To ensure authenticity in retaining the student voice, actual students' muddiest points were provided to guide the video, and a former student from the class would create the video. Using students' own words when addressing common muddy card responses was also stressed by Carberry et al. (2013). These findings will inform the design of our system by ensuring the student voice is maintained. When summarising the most common student responses, we will return representative student responses rather than running text summarisation algorithms, which could lead to the loss of the student voice.

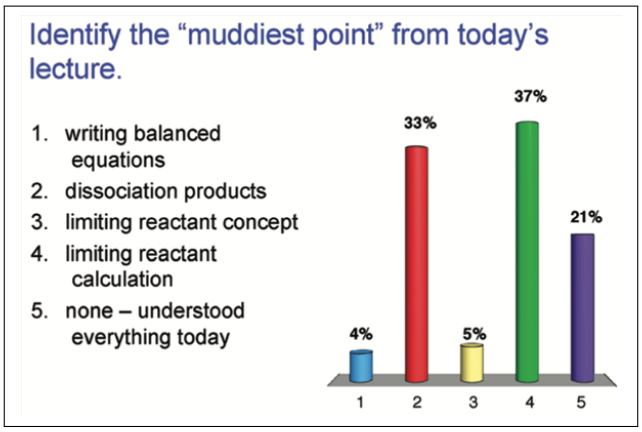


FIGURE 2.3: Example of a muddiest-point clicker question with a column graph indicating the proportion of students that deemed each option as muddy (King, 2011).

While the use of video is linked to improving student outcomes, they all occur after student muddy cards have been collected and analysed. The major obstacle to muddy cards is the substantial amount of time required to analyse student responses (Adams, 2004; Hall et al., 2002; Kessler and Nadjm-Tehrani, 2002; Mosteller, 1989). This can prevent teachers from even getting to the point where they can make supplemental materials that students perceive as helpful for their learning. To resolve the muddy card temporal challenge, different approaches have been investigated.

2.7.1 Clickers

In a unique approach, King (2011) explores a muddy card implementation for large first-year chemistry courses, where students would identify the muddiest points via physical clickers. In the last five minutes of a lecture, students would be asked to identify the muddiest point from the lecture from a list of predefined options. The results would be displayed live, where a column graph indicates the proportion of students that voted for each category, as seen in Figure 2.3.

The muddiest point identified by students would be addressed in the following lecture by either a general question about the topic to see if students were still confused, or a question to understand what aspects of the topic were confusing so that the lecturer can properly address it. These questions themselves were also framed using predefined options, so that students could answer them live via a clicker.

Whilst innovative, the authors have not yet investigated the student perception of this tool, or the extent to which it improves student learning outcomes. By providing a predefined list of muddiest points, the clicker approach significantly decreases muddy card analysis time for instructors. In fact, analysis is as

quick as flicking a switch, meaning that results can be shown live to students. However, by restricting the options to a simplified list, a limitation of this approach is that it does not allow students to freely articulate points of confusion. Additionally, we speculate that by providing predefined options rather than students articulating their muddy points, there will be a decrease in cognitive engagement as students are not reflecting on the lecture as deeply. More recent approaches aim at allowing students to write free-form responses, while offering an interface that will enable instructors to rapidly interpret and find patterns in student responses.

2.7.2 Mudslide

One such system is Mudslide, which modifies the traditional muddy card system to a method better suited for online lecture videos (Glassman et al., 2015). Using Mudslide, students will spatially pinpoint parts of a lecture slide they found most confusing. A transparent circle will appear at the point of a student's click, after which the student will be provided with a textbox to elaborate on their muddiest point. The student will finally be prompted to indicate their confusion about the identified muddiest point on a four-point scale from not confusing to extremely confusing. To quickly analyse areas of student confusion, instructors can view a heat map overlay of confusing points where the points are coloured by student confusion from the four-point scale (see Figure 2.4). To gain a deeper insight into student confusion, instructors can view the text responses corresponding to different parts of the heat map.

To evaluate Mudslide, 19 US high school teachers and 25 students tested the system. As a benchmark, students experienced Mudslide as well as a traditional muddy card approach, where students would simply write the most confusing point from the lecture. Results from independent raters indicate that fewer students produced muddy points of no substance when using Mudslide. Student surveys revealed that 28% appreciated the exactness with which they could discuss their muddiest points using Mudslide. By a two-sided Wilcoxon rank sum test, Mudslide was rated as significantly more useful than the baseline, giving teachers a better sense of students' confusion. More generally, 53% of teachers did not predict student muddy points, indicating that overall, muddy cards offer unique perspectives that are not easy to anticipate.

Overall, Mudslide can be seen as a solution that addresses the temporal challenge of muddy cards by quickly revealing areas of confusion through heat maps. However, in the case where a lecture contains over a thousand students, it is unclear how useful the Mudslide heatmaps will be in identifying specific points of confusion. For example, in Figure 2.4, we see that the bottom left image is already quite busy

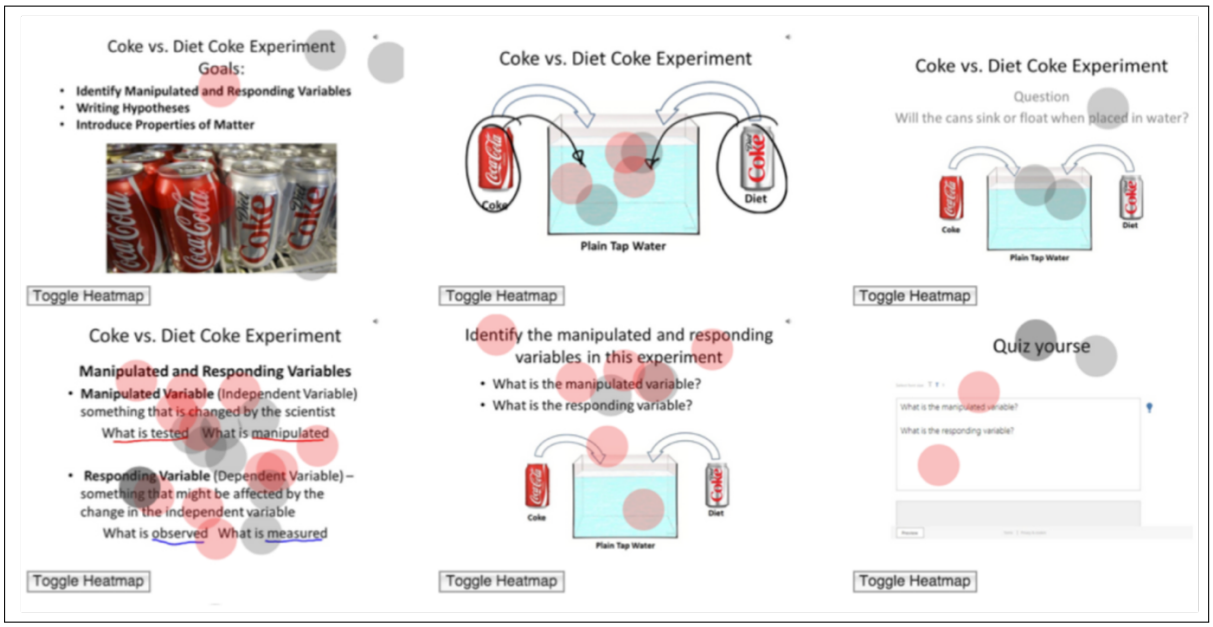


FIGURE 2.4: Six examples of the teacher interface in Mudslide. Transparent circles represent students' spatial muddy points. Points are coloured red if the student indicated that the lecture was confusing, and grey otherwise (Glassman et al., 2015).

with student points despite a small class size. This could identify that this slide, in particular, is quite muddy for students, but determining the specific parts of the slide that are muddy is difficult. Additionally, analysing the written responses for different areas of the heat map may be difficult when there is a high volume of responses on overlapping parts of the heat map. The Mudslide paper did experiment with a Word Tree (Wattenberg and Viégas, 2008) and histogram representation of the students' raw muddy point descriptions, but these were ranked poorly by instructors. We suspect a reason for this is that these methods are not robust to students who write semantically similar responses using different words (for example, 'I am confused by sunlight in photosynthesis', and 'for me, I was confused by how sunlight is vital during photosynthesis'). Because of these points, we maintain that methods to rapidly analyse written muddy card responses are still required. Indeed, newer muddy card systems do attempt to address analysing written muddy card responses through NLP techniques.

2.7.3 CourseMIRROR

Fan et al. (2017) developed and evaluated CourseMIRROR (Mobile In-situ Reflections and Review with Optimised Rubrics), a system which approaches solving muddy cards' temporal challenge by processing student responses using NLP algorithms. The student interface presents as a mobile phone application that aims to encourage students to write detailed and specific muddy card reflections. As evident in

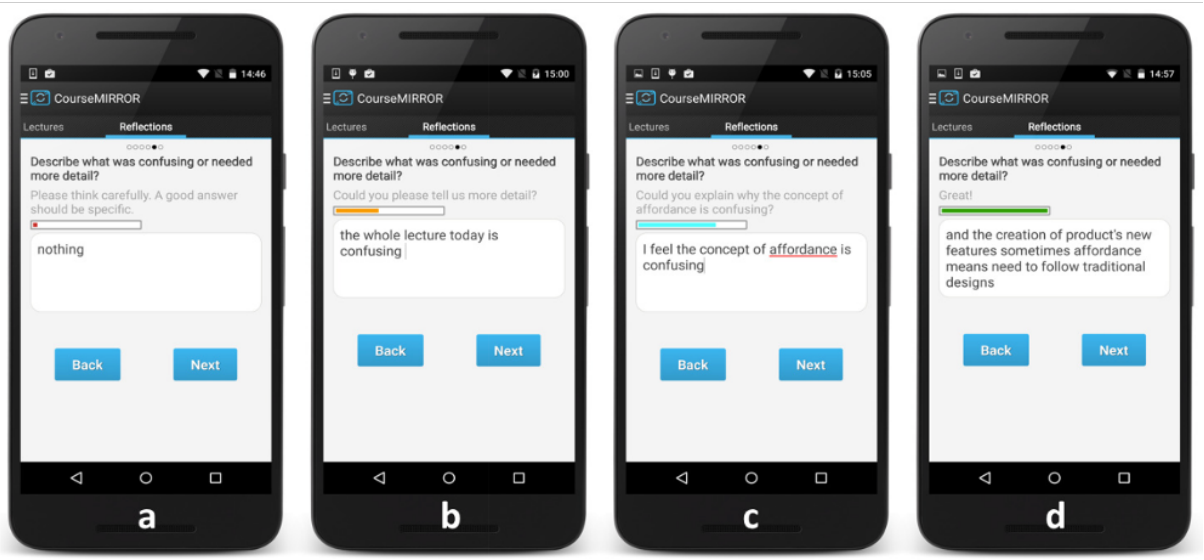


FIGURE 2.5: Example of the progress bar in CourseMIRROR. Images 'a' to 'd' demonstrate how the context-specific prompts and feedback bar change as student responses become more specific (Fan et al., 2017).

Figure 2.5, this is achieved via context-specific prompts with a progress bar that fills as students provide more detailed responses. Once students have input their responses, the most common themes are presented to the instructor through an online web portal, as shown in Figure 2.6. The internal algorithm achieves this by using a syntax parser to generate candidate noun phrases, with the candidate phrases grouped via the K-Medoids clustering algorithm into groups of semantic similarity. A representative phrase is selected from each cluster via LexRank (a graph-based ranking model), where the selected phrases are re-ranked based on the number of students who mentioned the phrase.

A 60-participant university student lab study investigated reflection quality and length using CourseMIRROR under different conditions. Conditions included having instant feedback with the context-specific prompts and feedback bar (as evident in Figure 2.5), having this same feedback but latent (feedback was only shown after pressing 'next', with users able to return and make adjustments), and without any feedback. The latent and instant feedback groups produced statistically longer reflections, and independent raters deemed that these groups produced statistically higher quality reflections than when the feedback bar was missing. There was no significant difference between reflection length and quality between the latent and instant feedback groups.

Further analysis involved an in-the-wild deployment across eight courses with 317 students and six instructors. Students found CourseMIRROR easy to use and desired it for future classes. Interviews

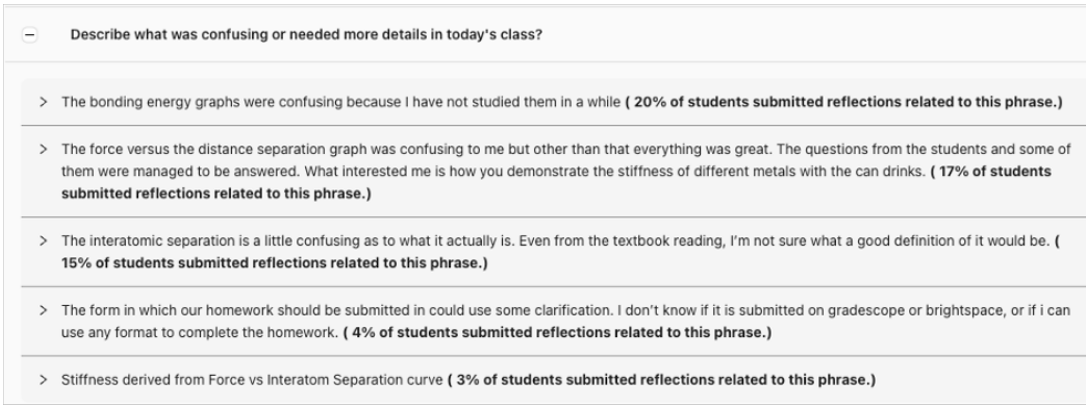


FIGURE 2.6: Instructor view in CourseMIRROR revealing the most common muddy card responses from a lecture (Fan et al., 2017).

revealed that all instructors responded positively to CourseMIRROR, and surveys revealed that it took instructors less than ten minutes to understand the student's muddy card responses from a lecture. A limitation is that with such small course sizes testing the system, it is unclear how CourseMIRROR will fare when used in cohorts with thousands of students.

Our muddy card system shares many similarities with CourseMIRROR. Both systems feature student and teacher interfaces, and they utilise clustering algorithms to categorise responses based on semantic similarity. There are however some major differences. In our instructor interface, we believe that the instructor should not be completely removed from analysis, rather, NLP techniques should guide analysis. On the other hand, in CourseMIRROR, the instructor merely views the final results of clustering, as seen in Figure 2.6.

Another large difference is in relation to the students' muddy card data collection. CourseMIRROR uses a reflection bar and context-specific prompts to encourage students to write pedagogically valuable reflections. Our system includes the 'student assisted approach' where students are prompted to indicate whether other selected student responses are semantically the same as what they entered. Whilst these differences exist, they could prove complementary for future studies.

2.7.4 Mentimeter AI Grouping

Mentimeter is a commercial platform that educators commonly use to incorporate active learning into lectures. One feature in Mentimeter is the ability to ask open-ended questions, which the audience can complete on their personal device, with the results relayed to the presenter's screen. Mentimeter released a feature that can group responses to open-ended questions. An example of this is shown in Figure 2.5,

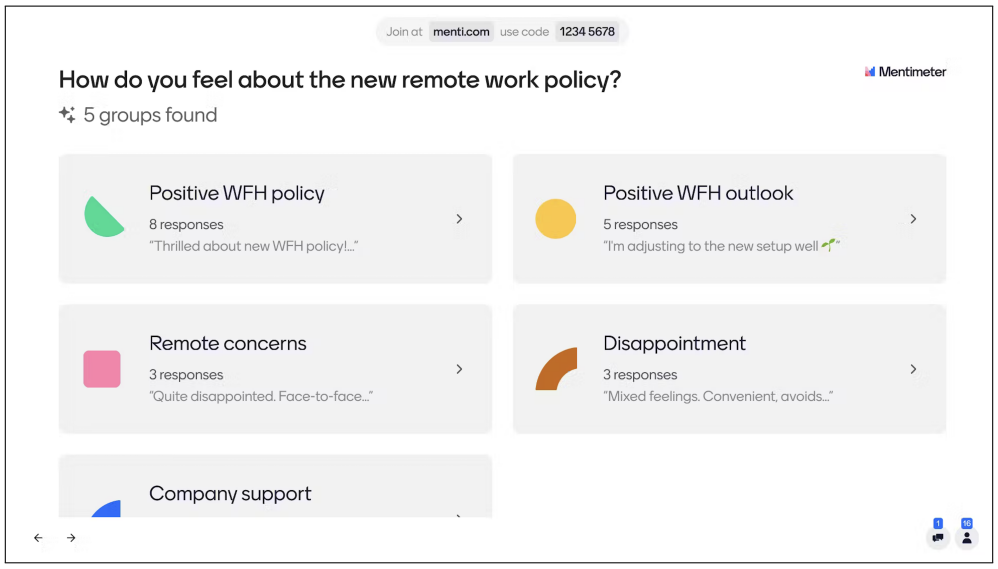


FIGURE 2.7: Example of open-ended responses being automatically grouped in Mentimeter (Mentimeter, n.d.).

where audience members' responses for a Human Resources meeting have been semantically grouped. Clicking on a group will reveal the raw responses that have been placed into the group.

While this offers a quick solution to clustering audience responses, it remains unclear how this approach will scale for muddy cards. Lectures usually only cover a few broad topics, and our data analysis revealed that muddy cards are inherently fine-grained, with students mentioning very specific confusing points. Hence, broad topic-level clustering is unlikely to be useful. Additionally, our muddy card system provides the instructor the ability to adjust the granularity of the clusters, whereas in the Mentimeter approach, it is all done automatically (similar to CourseMIRROR). Further, in Mentimeter, there is no way to export the clusters, with the clusters only being available in the presentation view[9].

While we believe our system has a more robust approach to clustering muddy card responses when compared to Mentimeter, a key strength of Mentimeter is that it has been designed to relay the clustering live to the audience. Hence, the audience will be kept more engaged as they can witness the results of the clustering firsthand.

[9]Information about Mentimeter's AI clustering can be found at https://help.mentimeter.com/en/articles/8300577-group-responses-to-your-open-ended-questions-using-ai

2.7.5 Mud-Class Feedback System

Recently, Choong et al. (2024) created an app for students to enter their muddy card responses, with an instructor interface that allows instructors to sort and filter responses. When students enter their muddy card responses, keywords are extracted using a rapid automatic keyword extraction (RAKE) algorithm. In the instructor interface, instructors can sort or filter muddy card responses by completion date or rating.

While included in this review, we have some questions about their system. Firstly, there is some confusion regarding what they mean by "rating" in this paper. The paper explains that keywords can be filtered or sorted based upon "rating", however, it is unclear what exactly is meant by this. We also have questions regarding the system's efficacy, as a user study has not yet been conducted, and there are no images representing the system's interface.

2.8 Conclusion

Muddy cards have been shown to be an effective active learning tool; however, the time required to analyse responses makes them impractical for classes with large cohort sizes. This review has explored existing technological approaches to muddy cards, but many of these approaches resolve the temporal challenge by sacrificing the attention placed on student-written responses. CourseMIRROR was presented as a system that does directly analyse written responses using NLP techniques. In CourseMIRROR, the instructor remains passive during the analysis process, however, we believe that an effective muddy card system should include the instructor in the analysis process. Our system attempts to resolve this by using NLP techniques to guide the instructor through data analysis.

Such a system relies on accurate muddy card sentence embeddings, and so embedding methods such as SBERT and proprietary embeddings were explored. Hierarchical clustering was then introduced as a method to separate sentence embeddings into clusters of semantic similarity. The application of hierarchical clustering is that input controls can be implemented so the user can adjust how many clusters to split the data into, meaning human intuition can be used to adjust clusters to best semantically group muddy card responses. To validate our system, grounded theory was introduced as a framework for developing a theory on how teachers and students perceive the effectiveness of the muddy card system. This theory will enable us to investigate whether our system facilitates the efficient implementation of muddy cards in classrooms with large cohort sizes.

CHAPTER 3

Methodology

In this chapter, we describe the methodology used to investigate Research Questions 1 and 2. As both research questions are motivated by the muddy card system, we first describe its design, including its key internal algorithms and deployment for in-the-wild use. We then explain the methods used for Research Question 1 by discussing the different datasets and evaluation methods employed to assess the embedding models for the muddy card use case. Following this, we discuss various ways of incorporating the student-identified similar muddy card points (from the 'student-assisted approach') when clustering. This motivates discussing how we plan to release a public clustering data set on muddy cards. Finally, we address how we investigated Research Question 2 by describing the user study and how we aim to understand the perceptions of our system among students and teachers through surveys and interviews.

3.1 Muddy Card System Design

The muddy card system is broadly separated into two components: the student and teacher interface. In this section, we will first describe the front-end design of the student and teacher interfaces. We will then describe the internal algorithms central to these interfaces before explaining how these interfaces were deployed in-the-wild.

3.1.1 Student Interface Front-End

The central focus of the student interface is to provide a mechanism for students to input their muddy card response following a lecture using their computer or mobile phone. When logging on to the interface, students are presented with a welcome screen which prompts them to select the unit and lecture for which their response is intended (Figure A.1.1). If that unit of study is currently accepting responses (teachers can stop students from submitting responses through the teacher interface), students arrive at the muddy card response entry page, as seen in Figure 3.1.

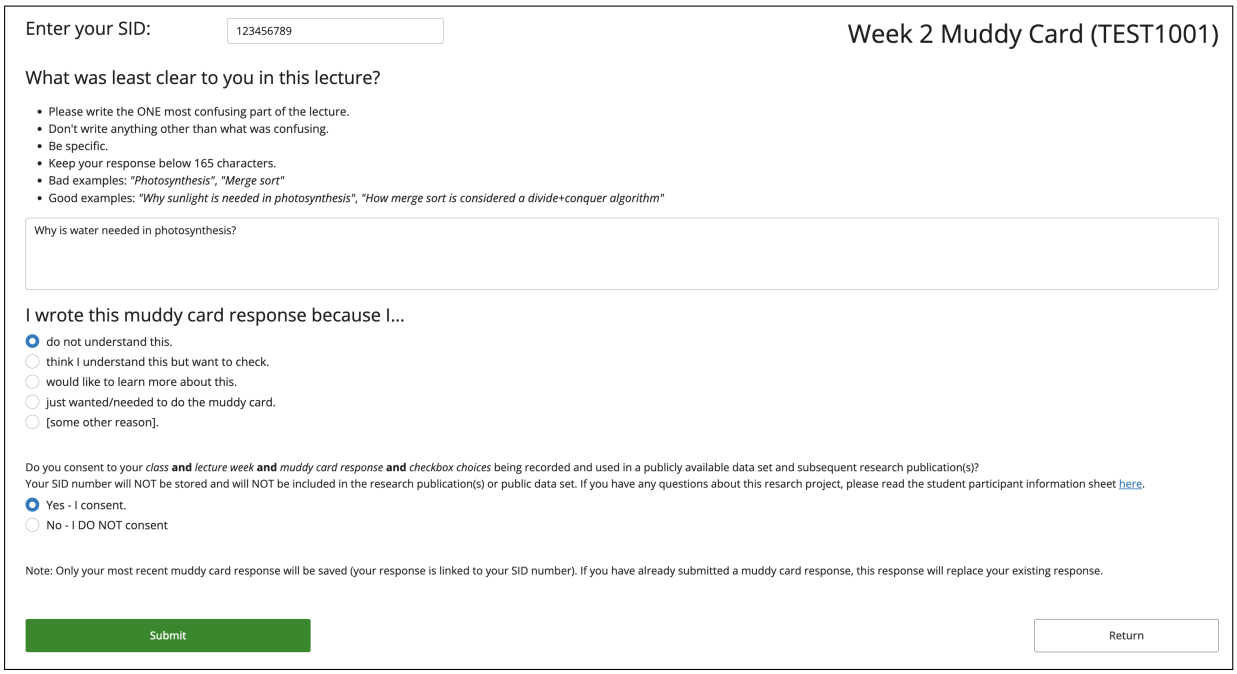


FIGURE 3.1: Student Interface - Page for students to submit their muddy card response.

Here, students answer 'What was least clear to you in this lecture?' in a free-form text box. Following this, students indicate why they wrote their muddy card response through a radio button with the options:

- 'do not understand this',
- 'think I understand this but want to check',
- 'would like to learn more about this',
- 'just wanted/needed to do the muddy card',
- '[some other reason]'.

The motivation for this question was that, as part of the analysis of previous muddy card data (specifically, the 2024 NLP course data explored in Section 3.2.1), we noticed that some students deviated from writing their most confusing point and instead asked interest-based questions. In the soon-to-be-described teacher interface, a teacher can filter responses based on the answer to this radio button, allowing teachers to focus on the responses from different subsets of students. Students are also asked to provide their SID (student identification) number and indicate whether they consent to adding their muddy card response to the public dataset that will be described in Section 3.3.

An experimental feature we are trialling is the 'student-assisted approach'. In this approach, after submitting their muddy card response, a student will be prompted to indicate whether an answer to selected

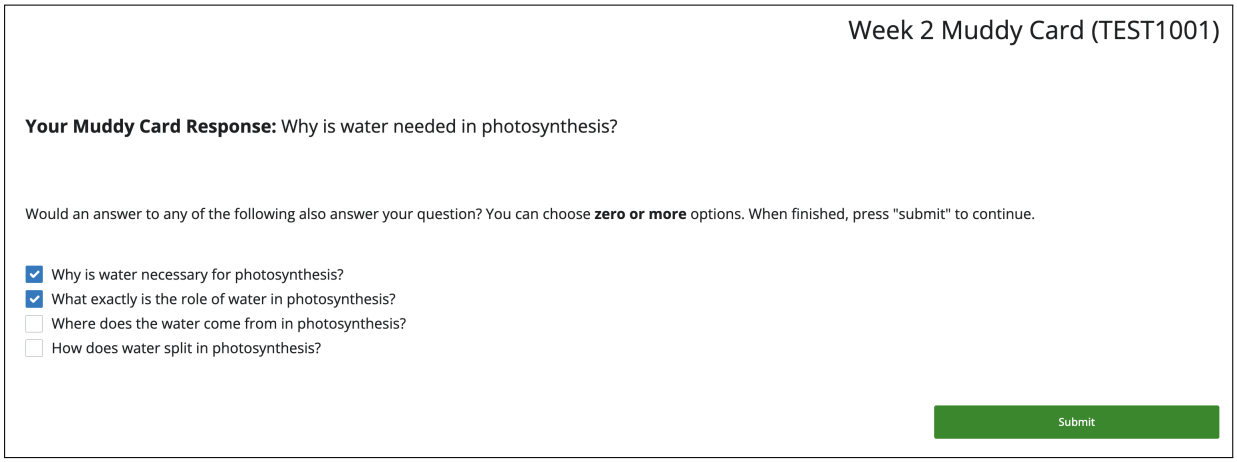


FIGURE 3.2: Student Interface - The student-assisted approach.

peer muddy card responses would also answer their muddy card response. Up to four peer responses will be displayed, and these responses are intentionally chosen to be similar to what the student originally wrote (elaborated on in Section 3.1.3.b). A student can click on as many of the provided responses as they deem valid. The student-assisted approach is shown in Figure 3.2. Students would encounter the student-assisted approach at various points throughout the user study.

The student interface ends with a summary page that reminds students of their muddy card response (Figure A.1.4). The full-screen sequence for the student interface can be found in Appendix A.1.

3.1.2 Teacher Interface Front-End

To analyse the students' muddy card responses, a teacher will open the website URL, which is specific to their unit of study. Upon opening the link, the teacher will select the lecture they wish to analyse (Figure A.2.1). After making their choice, they are directed to a summary page indicating the number of responses collected (Figure A.2.2). This page also has a toggle to stop students from submitting muddy cards for the selected lecture. This was included as some instructors may choose to make muddy cards an assessable component of their course, and hence need a way to stop students from submitting their responses after the deadline. This page also has a download button for instructors to download students' raw responses. Again, the intention is to allow an instructor to download student responses if muddy cards are part of a unit's assessable grade.

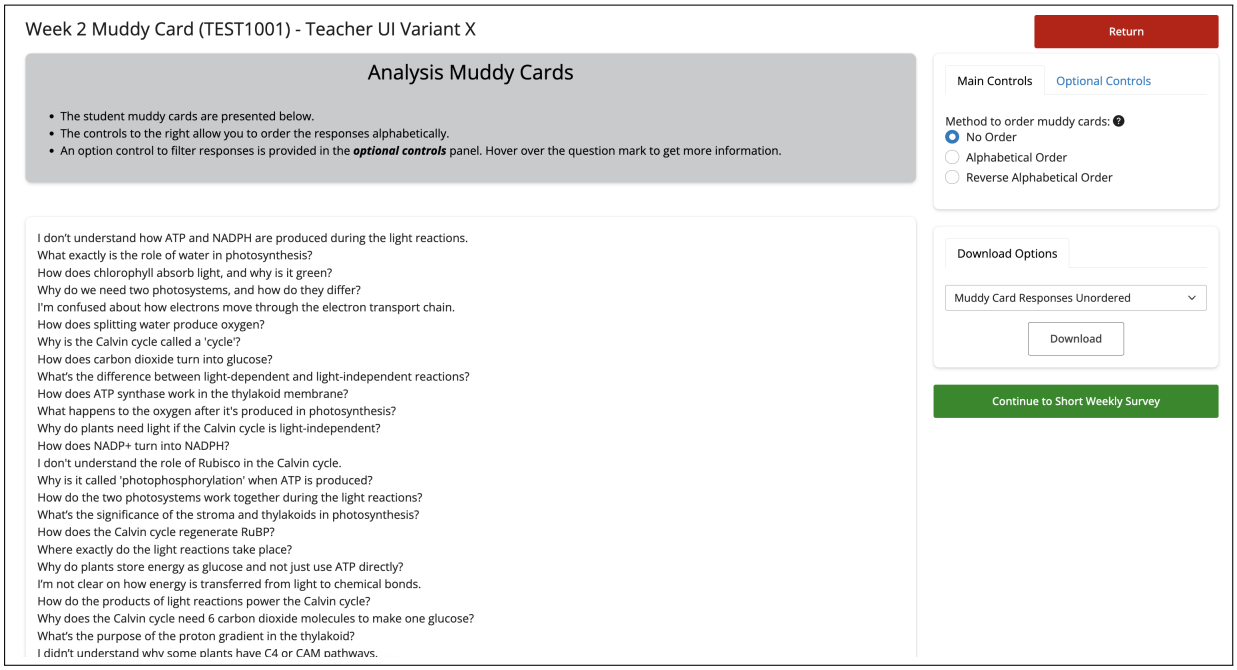


FIGURE 3.3: Teacher Interface - Baseline variant (variant X). Muddy card responses are arranged in a list, with basic sorting and filtering options.

The muddy card analysis stage begins once the teacher continues to the next page. For our muddy card system, our main innovation is the clustering (variant Y) interface. However, as a control, we also devised a baseline (variant X) interface.

Baseline Interface (Variant X)

The baseline interface (Figure 3.3) mimicked the traditional muddy card active learning technique, where instructors would receive the raw student responses. However, as this is a digital approach to muddy cards, we have also included functionality to alphabetically sort responses, as this functionality would be present in most simple analysis methods, such as spreadsheet software. Furthermore, we have included some optional controls, allowing instructors to filter out muddy card responses based on the radio button question that asks students why they wrote their muddy card response. The radio button options and a 'keep all responses' option are presented on the screen. Clicking on different options will filter the responses displayed to the instructor.

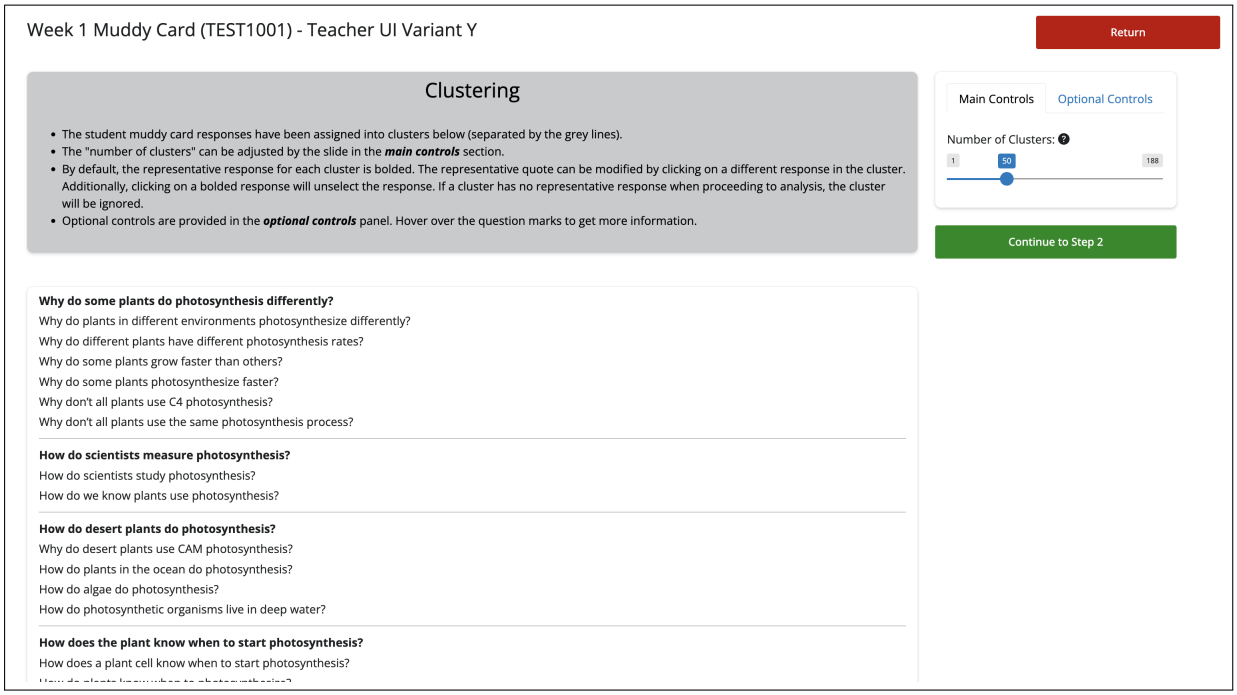


FIGURE 3.4: Teacher Interface - Clustering variant (variant Y). Muddy card responses are arranged down the screen in clusters of semantic similarity.

Clustering Interface (Variant Y)

In the clustering interface, student responses appear down the screen, arranged into clusters (groups) of semantic similarity. By default, these clusters are arranged in semantic order, where adjacent clusters are typically similar in meaning. As seen in Figure 3.4, a slider in the top-right of the interface allows the user to adjust the 'Number of Clusters' into which the muddy responses are separated in real-time. Adjusting to a smaller number of clusters will cause adjacent clusters to fuse and form a larger cluster. Adjusting to a larger number of clusters will see existing clusters break apart into smaller clusters. The user will adjust the number of clusters until the sentences are sufficiently grouped by semantic similarity.

In Figure 3.4, the bold sentence/quote in each cluster symbolises that cluster's representative quote. This can be altered by clicking on the quote that the user considers a better representative for that cluster. If a user feels that a particular cluster should be ignored, they can click on the bold sentence associated with that cluster, which will deselect it. In this case, the cluster would now have no representative sentence. This cluster will be excluded when receiving the final summary of the muddy card responses at the end of the analysis.

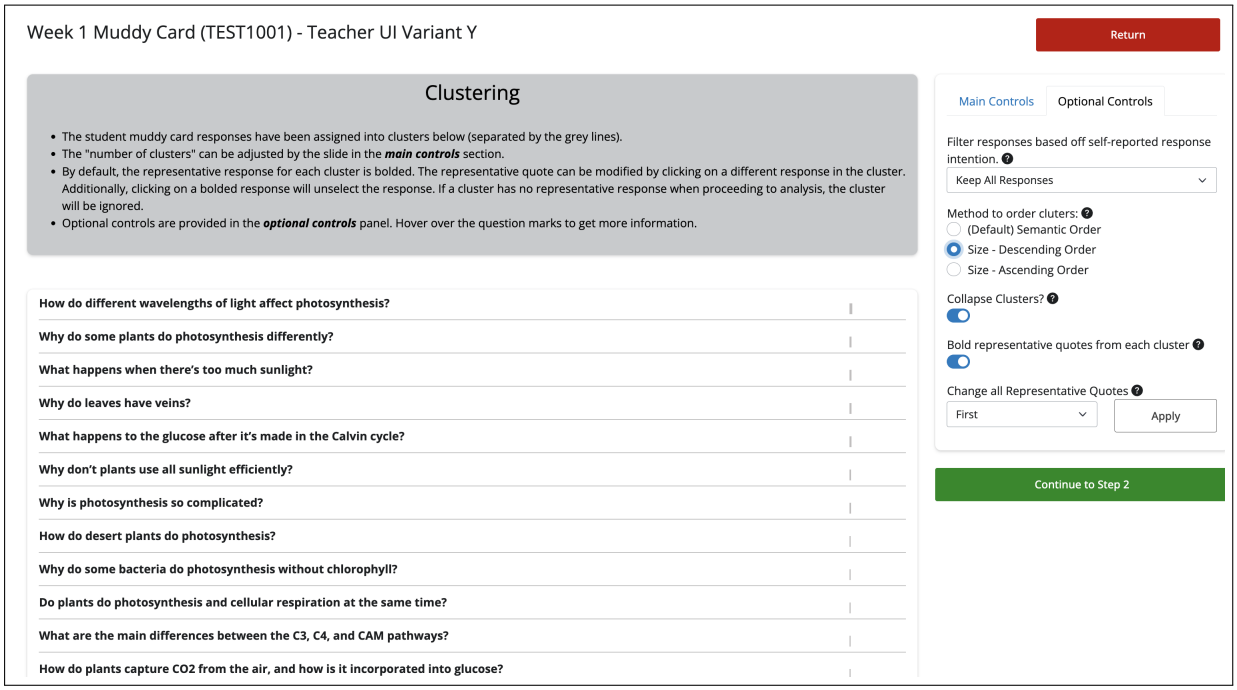


FIGURE 3.5: Teacher Interface - Clustering variant (variant Y). View of the optional controls. In this image, the clusters are arranged by descending size, with the clusters collapsed, meaning only the cluster's representative quote is shown.

The 'Optional Controls' panel offers additional controls that teachers can use for analysis. Similar to the baseline approach, there is the option to filter muddy card responses based on the student's reason for writing their muddy card response. Additional controls allow the user to change the ordering of the clusters by size, collapse the clusters, and unbold the representative quotes. Figure 3.5 demonstrates the clusters being arranged in descending order by size, with the clusters collapsed so that only the representative quote for each cluster is shown. As seen in Figure 3.5, when the clusters are collapsed, grey bars appear for each cluster, representing each cluster's relative size.

If unsatisfied with each cluster's representative quote, an optional control allows the user to rapidly change which quote is representative through the 'Change all Representative Quotes' button. The user can choose between setting the representative quotes as the first sentence in each cluster or the centroid of each cluster. For the centroid method, when considering the quotes as sentence embeddings, we set the representative quote to be the quote that is the most central point of the cluster[1].

[1]In computer science, the *centroid* is the point in the vector space which is the average of all points in the cluster. The *medoid* is the location of the point that is on average closest to all the other points in the cluster. For our application, this central point is actually the medoid. However, we refer to it as the centroid in the interface as 'centroid' and 'central' share the same root 'cent'. We believe this will be more familiar for those without a computer science background.

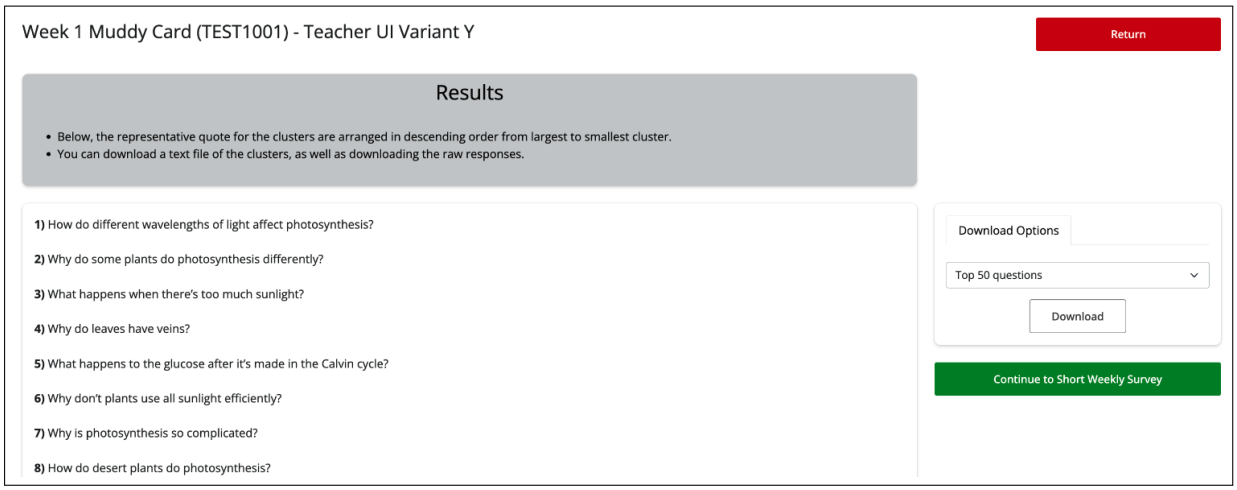


FIGURE 3.6: Teacher Interface - Clustering variant (variant Y). The final summary page arranges the representative muddy card response for each cluster by cluster size.

When the user is satisfied with the muddy card groupings and each cluster's representative quote, they will proceed to the summary page. As shown in Figure 3.6, the summary page displays the representative quote for each cluster, arranged by cluster size. Users can also download a text file on the summary page containing the ordered representative quotes.

The full screen sequence for the baseline and clustering variants of the teacher interface can be found in Appendix A.2.

3.1.3 Internal Algorithms

On submission, student muddy card responses are converted into sentence embeddings using OpenAI's `text-embedding-3-small` model. This embedding model was chosen as part of the analysis described in Section 3.2.1. These embeddings were used during the clustering of student responses in the teacher's interface and when selecting which student responses to provide to students during the 'student-assisted approach' in the student interface.

3.1.3.a Agglomerative Clustering

In the clustering variant of the teacher interface, the clusters were produced using agglomerative clustering. First, the muddy card sentences were transformed using robust scalar, meaning that the data was centred at 0 (using the median) and scaled using the interquartile range. Next, agglomerative clustering with Ward linkage was used to generate the entire cluster tree for the scaled sentences. So that sentences

appear in a logical semantical order for the user, the original sentences were reordered based on the cluster tree (i.e. dendrogram), such that as clusters are combined when moving up towards the root, adjacent sentences will be merged together. The agglomerative clustering parameters were selected as part of the analysis described in Section 3.2.1.b.

3.1.3.b Student-Assisted Approach

When considering the student-assisted approach in the student interface, the system would first pass through all peer responses and select the top four most similar responses according to cosine similarity on the sentence embeddings. To ensure that these four responses were sufficiently similar to the original muddy card response, only those sentences with a cosine similarity exceeding 0.33 were presented to the student (this threshold value was chosen from the analysis described in Section 3.2.1.c). Without this threshold, students could be assigned peer-responses that are irrelevant, or potentially inappropriate (such as students who are are intentionally acting in an adversarial way).

If a student decided that an answer to the selected peer responses would also answer their muddy card response (i.e. a student identifies them as nearly semantically similar), this would be reflected in the clustering variant of the teacher interface. Here, we used a simple algorithm that involves looking through all the student inputs for the student-assisted approach and linking together the sentences that students identified as matches. For example, consider if a student wrote sentence 'A' and claimed that sentence 'B' was a match during the student-assisted process. If another student claimed that their sentence 'C' matched with sentence 'A' and sentence 'D', then we would lump sentences 'A', 'B', 'C' and 'D' together. Thinking of this as a graph, there is a link between sentence 'A' for both students, and consequently, sentences 'B', 'C', and 'D' are linked as well. We refer to this produced cluster as a *student-assisted cluster*. We refer to this algorithmic approach to the peer-assisted choices as the *student-assisted graph approach*. Later, in Section 3.2.3, we will investigate two other methods for creating student-assisted clusters, which could only be tested after the user study, as we had no initial data for testing the student-assisted approach.

To ensure that agglomerative clustering functions properly when using the student-assisted approach, a preprocessing step involved removing all but one sentence from each student-assisted cluster in the corpus. This is an important step, as otherwise, the agglomerative clustering algorithm might place responses into clusters inconsistent with the student-identified links. It would then become unclear how to edit these clusters appropriately for the peer-assisted links to remain intact. The adjusted corpus is

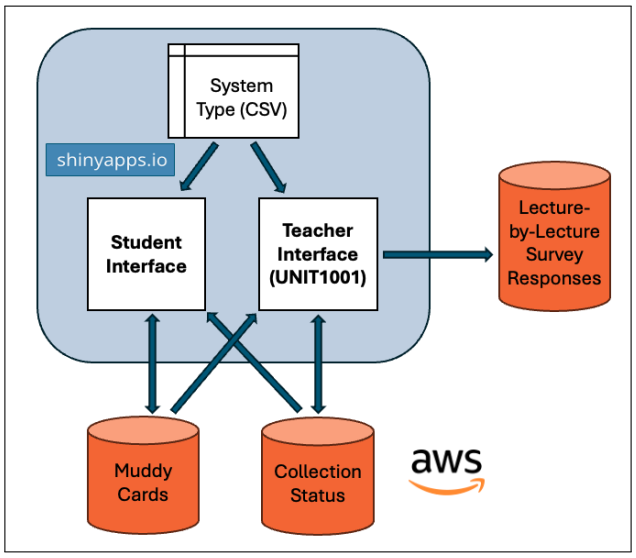


FIGURE 3.7: Deployment of the muddy card system. The application is hosted on `shinyapps.io`, and the external storage consists of different tables within an RDS PostgreSQL instance on Amazon Web Services.

passed to the agglomerative clustering algorithm (Ward linkage, Euclidean distance) for the cluster tree to be fully produced. The removed sentences from each student-assisted cluster are then reunited with the sentence that was kept for agglomerative clustering.

3.1.4 In-The-Wild Deployment

The system's code was written using the Shiny for Python framework and is hosted on `shinyapps.io`. As illustrated in Figure 3.7, the student interface and each unit's teacher interface are deployed as separate applications that are accessible on the web via a URL. Each subject/unit that uses the muddy card system will have its own teacher interface (such as 'UNIT1001' in Figure 3.7). `shinyapps.io` includes privacy features, meaning each teacher interface can only be accessed by authorised users (which in this case, is the teaching staff for a particular unit). The 'System Type' CSV file, as shown in the figure, controls for each lecture whether students will receive the student-assisted approach experimental condition, and for the teacher, whether they will see the clustering or baseline interface. This is pertinent to the user study described in Section 3.4.

The orange databases in Figure 3.7 are different tables within an RDS PostgreSQL instance facilitated by Amazon Web Services (AWS). The 'pgvector' database extension was used for effective embedding storage and cosine-similarity-based retrieval. The 'Muddy Cards' database stores a student's muddy card response and associated inputs, such as their choice for the radio button on why they wrote this response,

and identified similar peer responses in the 'student-assisted approach' interface. The 'Collection Status' database stores whether an instructor has switched the toggle to prevent students from submitting further muddy card responses. The 'Lecture-by-Lecture Survey Responses' database was used to store the optional lecture-by-lecture survey responses discussed in Section 3.4.1.a.

3.2 Benchmarking: Sentence Embeddings

Recall that Research Question 1 states:

(1) How well do sentence embedding models perform when clustering student muddy card responses into groups of semantic similarity with the 'student-assisted approach'

This methodology section will explore this research question by looking at two different sources of muddy card data, and determining how different embedding models, clustering methods, and methods of using the student-assisted data perform at correctly clustering the sourced muddy card responses.

3.2.1 2024 University NLP Course

The first data source comes from muddy card responses collected as part of the University of Sydney's graduate Natural Language Processing Course (Unit Code: COMP4446/5046) in 2024. While aiming to answer Research Question 1, an auxiliary benefit of observing the 2024 data is that it helps to inform the decision of the embedding and clustering method used in the muddy card system for the user study. There were approximately 500 students enrolled in the course, and students were tasked with filling out weekly muddy cards, which accounted for a small portion of their course grade (5% of their total grade). This study received ethics approval (Reference Code: 2024/HE000932) to analyse this data retrospectively. Students were provided with a participant information statement (see Appendix B.1) and a form to fill out if they did not wish their data to be analysed. This form was managed by a third party, with the third party removing data for students who withdrew before providing the data to the research team. Reports from the third party indicate that there were 10 form entries, which could include duplicate responses.

The student researcher independently clustered two lectures' worth of muddy card responses. The clustering process first involved sorting the student responses based on semantic similarity using SBERT's `paraphrase-MiniLM-L6-v2` model, so similar muddy card responses would be closer together. This decision was made as preliminary clustering attempts were too slow, and we believed that any

potential bias introduced by sorting would be minimal. Next, the sentences were placed into groups based on semantics, meaning that if muddy card responses said the same thing, they would be grouped together.

3.2.1.a Determining the Best Embedding Model

With the data manually clustered, the first goal was to determine which sentence embedding model performed best on the manually clustered data. For this, we tested different SBERT models and proprietary models from OpenAI and Voyage AI. OpenAI and Voyage AI were chosen because their terms of service explicitly mention that you can opt out of having your data retained for fine-tuning their models. Students were notified of these companies in advance in the participant information statement.

The general method used to evaluate embedding performance was to calculate the adjusted Rand index (ARI) and adjusted mutual information score (AMI) for the clusters produced by the agglomerative clustering algorithm against the manually clustered data. We also used a third clustering metric, which we specifically developed for this project, called the *student questions answered satisfaction score* (SQAS score). The high-level idea behind the SQAS score is to quantify how many students would have had their muddy card responses addressed if x muddy card responses were answered or elaborated on. Algorithmically, this metric is easily calculated:

(1) Set a variable called *score* equal to 0.
(2) Look at the largest cluster produced during clustering that has not already been processed.
(3) Find the representative sentence/muddy response from this cluster by finding the medoid.
(4) For the selected representative muddy response, locate the cluster that this sentence falls in for the manually clustered data. Find the size of this cluster, and increment *score* by the size.
(5) Continue steps 1 to 3 until x has been reached. Each time, locate the manual cluster to which the representative sentence belongs. If this manual cluster has not already been encountered, add the size of the manual cluster to the *score*.

A visual representation of how this score is calculated is provided in Figure 3.8. As seen in the figure, before the algorithm commences, sentences A through J have been clustered using a clustering algorithm and by manually clustering the sentences. At the first iteration of the SQAS algorithm, we locate the largest cluster and find the medoid of this cluster. The medoid for the first iteration is the sentence E. We then search the manually constructed clusters and find the cluster that contains the sentence E. We take the size of this manual cluster and set it as the SQAS score. We name it the SQAS-1 score, as this

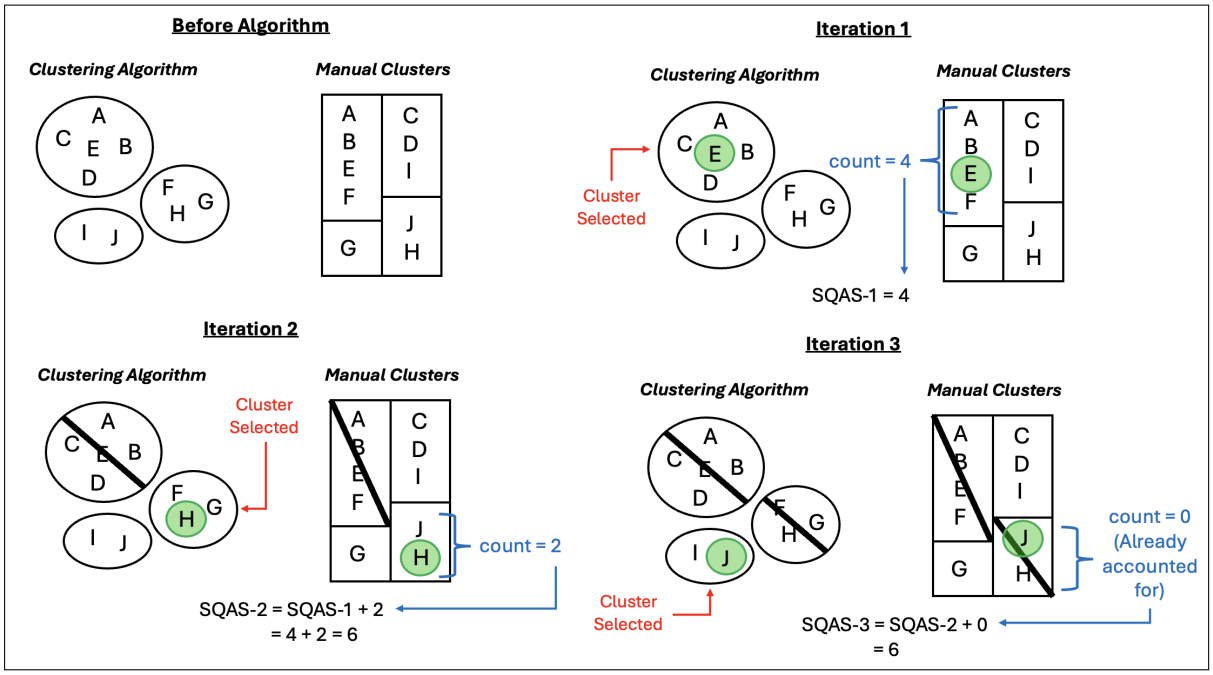


FIGURE 3.8: Example of calculation a SQAS-3 score. The green circle illustrates the medoid of the selected cluster.

is the score after 1 question has been answered. We then proceed to iteration 2, where the next largest clustering algorithm cluster that has not been processed is chosen. We similarly find the medoid of this cluster (which, in this case, is sentence H) and then identify the manual cluster containing H. As we have not already encountered this manual cluster, we add the size of this cluster to the cumulative SQAS score, which now has a value of 6. In the third iteration, we process the last cluster from the clustering algorithm that remains. The medoid of this cluster is J. However, we have already processed the manual cluster that contains J, and hence we do not add anything to the SQAS score. Hence, the SQAS-3 score (that is, the SQAS score after three questions are answered) is set to 6.

To better understand the proportion of students who are satisfied that their question has been answered, we define the 'SQAS-n proportion' to be the SQAS-n score divided by the total number of responses. Recall that in Figure 3.8, the SQAS-3 score was 6. As there are 10 responses, the SQAS-3 proportion is 6/10 or 60%. This would indicate that after three questions are answered, 60% of students are satisfied that their question was addressed.

To determine the best sentence embedding model for our two manually clustered lectures, we used agglomerative clustering (Ward linkage, Euclidean distance) for each model's raw embeddings (that is, no scaling methods or dimension reduction applied) to grow the full hierarchical cluster tree. With the

tree created, each k-split (where k refers to the number of clusters to split the data into) was trialled, with the ARI, AMI and SQAS-15 proportion computed for each k-split against the manually labelled clusters. 15 was chosen for the SQAS score as we believe in a real-life use case of the muddy card system, this is a reasonable number of responses for an instructor to read (they would likely not keep reading all representative responses, especially if there are many of them) and consider answering. The max ARI, AMI and SQAS-15 for each lecture and embedding method were reported. The maximum score was determined by considering all k-splits of the clustering tree and reporting the maximum value. We believe this is a reasonable approach, as the instructor can adjust the number of clusters using the slider in the teacher interface. The instructor would seek the best possible clustering for their data. The significance of the clustering was not measured, as we are unaware of any methods to measure this for clustering.

From all the SBERT, OpenAI and Voyage AI embedding models considered, a single model from each group was selected to move on to the next analysis phase - exploring the optimal agglomerative clustering parameters.

3.2.1.b Determining the Best Clustering Method

Even if the embedding model was set, there are still many parameters that can affect clustering performance. The embeddings could each be scaled and undergo dimension reduction. Agglomerative clustering can use different distance and linkage metrics. This stage of analysis involves hyperparameter tuning to explore which parameters yield the best clustering performance when compared to the manually clustered dataset for the three best models identified in the previous section.

To investigate the optimal parameters, we tuned the following:

- *Data scaling*: the method used to scale the embeddings. This included no scaling or `scikit-learn`'s StandardScaler, MaxAbsScaler, MinMaxScaler and RobustScaler (Pedregosa et al., 2011).
- *Dimension reduction*: the method used to reduce the embedding's dimension. Options included no reduction, PCA, TSNE or UMAP.
- *Dimension components*: the number of components to keep during dimension reduction.
- *Agglomerative clustering distance metric*: Euclidean, Manhatten or cosine.
- *Agglomerative clustering linkage metric*: Ward, complete, average or single.

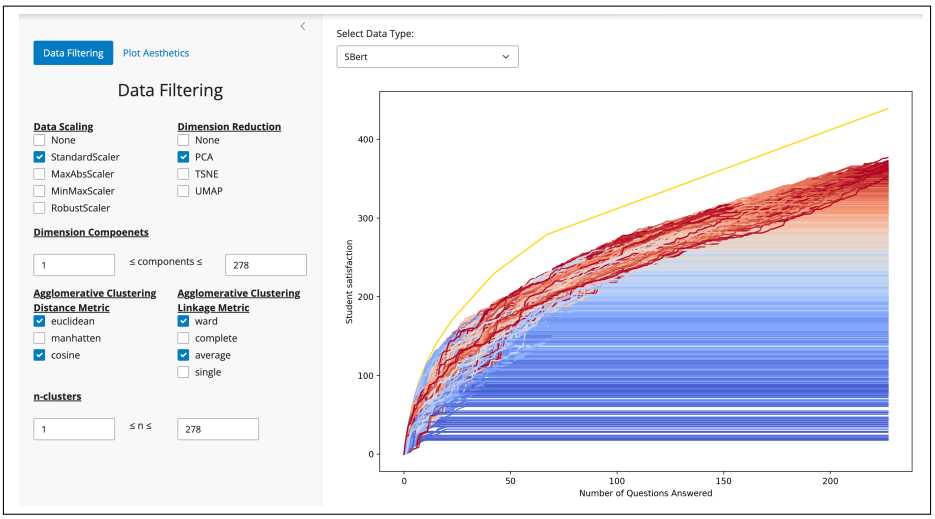


FIGURE 3.9: Image of the dashboard used to identify which parameters produce the largest SQAS score. Users can use the input on the left to adjust which parameter combinations are plotted. The gold line indicates the gold standard (perfect clustering based on the manual data). The colouring represents the number of agglomerative clusters into which the data is split.

In particular, we were interested in the relative effects of using different parameter combinations. To visualise how these parameters affected clustering performance, for each parameter combination, we used a line graph where the x-axis is the number of students' muddy points answered, and the y-axis is the corresponding SQAS score (recall that SQAS is a cumulative score). As there are many possible combinations, we developed a dashboard to compare different parameter permutations, as shown in Figure 3.9. By filtering different parameters in the dashboard, we can empirically visualise which set of parameters leads to the highest SQAS scores, and which are therefore the ideal parameters for clustering. In Figure 3.9, the gold line represents the gold standard, which indicates perfect clustering based on the manual data. That is, the gold standard defines the optimal number of students that can be satisfied that their question was answered as the number of questions answered (x-axis) increases. When considering Figure 3.8, the gold-standard value after iteration 3 is 9. This is because, in the most optimal circumstance, an algorithm would first choose the manual cluster of size 4, then the manual cluster of size 3, followed by the cluster of size 2.

3.2.1.c Further Muddy Card System Modifications

When inviting lecturers to participate in the user study (the user study is discussed in Section 3.4), a lecturer raised concerns that under the student-assisted approach, students may write inappropriate

comments that are propagated throughout the cohort. Hence, we sought to determine a cosine similarity threshold that would result in peer sentences being displayed only if they were sufficiently similar. This threshold was determined by analysing the cosine similarity between the sentences in the manually clustered 2024 NLP dataset. Specifically, we compared the distribution of cosine similarity values for sentences within the same cluster with the distribution of out-of-cluster cosine similarity scores.

3.2.2 2025 User Study

The second dataset comes from the user study that will be discussed in Section 3.4. When using the muddy card system, students could consent to adding their muddy card response to a public dataset. We decided to cluster many lectures' worth of this data, especially as we have permission to eventually release this data to the public for verification, and to establish a clustering dataset for muddy card applications.

Based on the 2024 NLP data, we made slight modifications to the manual clustering approach. An observation made when clustering the 2024 data was that students would write muddy card responses that were very similar yet not semantically the same. For example, one student may say they were confused by topics 'A and B', another confused by topic 'A', and another confused by topic 'B'. In the 2024 approach to clustering, this would result in three clusters, as each student is asking something distinct from the others. More often than not, topics 'A and B' would be related, and it would make more sense to keep the responses together rather than place them in isolated groups that will likely go unnoticed during analysis due to their small size. The new approach involves clustering together all sentences that can be answered using a single answer. For example, in the new approach, all the responses would be clustered together, as an answer to the confusion about 'A and B' would answer all the confusing points in the cluster. We believe this change is closer to what an instructor wants to know: 'Which questions should I answer to help most students?'

Using the above approach to clustering, the student researcher manually clustered five lectures' worth of muddy card responses for the 2025 offering of the NLP course. In this unit, muddy cards are a part of the student's course grade, and so responses were large for these courses. The supervisor of this thesis also clustered one of the five lectures' worth of muddy cards. Agreement between the researchers was measured using AR and AMI. Afterwards, the researchers convened to adjudicate discrepancies and create a new clustering based on consensus.

We wished to investigate the sentence embedding performance on non-computer science courses; however, as will be evident in the later results, student responses for other courses were low. Hence, for the finance unit in this study (FINC5001), we concatenated the muddy card responses from multiple lectures to create a data sample with 200 muddy card responses. Again, both researchers manually clustered this data and met to reach a consensus on the correct clustering.

Different SBERT, OpenAI and Voyage AI embedding models' clusters (using agglomerative clustering with Euclidean distance metric and Ward linkage) were compared with the manually clustered datasets using ARI, AMI and SQAS-15. The maximum ARI, AMI, and SQAS-15 scores for each lecture and embedding method were reported, allowing us to investigate which embedding methods are best suited for our clustering use case.

3.2.3 Student-Assisted Approach Clustering

In Section 3.1.3.b, we introduced the *student-assisted graph approach*, a method for using the student-assisted inputs to create *student-assisted clusters*. As a quick reminder, in the *student-assisted graph approach*, all student responses with a link between them would be retained in the same cluster. This is illustrated at the top of Figure 3.10. However, this approach represents only one method of using the student-assisted choices. Hence, with the data collected as part of the user study, we trialled different algorithms for using the student-assisted data.

Similar to the student-assisted graph approach, another approach is the *keep-top student-assisted graph approach*. The process is the same as the student-assisted graph approach, except that for all the student-assisted choices a student might make, we only retain the one with the highest cosine similarity to their muddy card response (see the middle example in Figure 3.10). This approach is motivated by the fact that after reviewing the student-assisted choices from the first week of the user study, some students misunderstood the student-assisted approach and were selecting responses that were not semantically similar. This resulted in large student-assisted clusters with many responses that were loosely related. Hence, we updated the internal algorithm used in the muddy card system to use the keep-top student-assisted graph approach instead.

A third approach trialled was the *multi-evidence approach*. Under this approach, student-assisted clusters are formed when more than one student agrees on the links between responses. This is best explained by observing the bottom example of Figure 3.10. In this example, the student who wrote muddy card 'A'

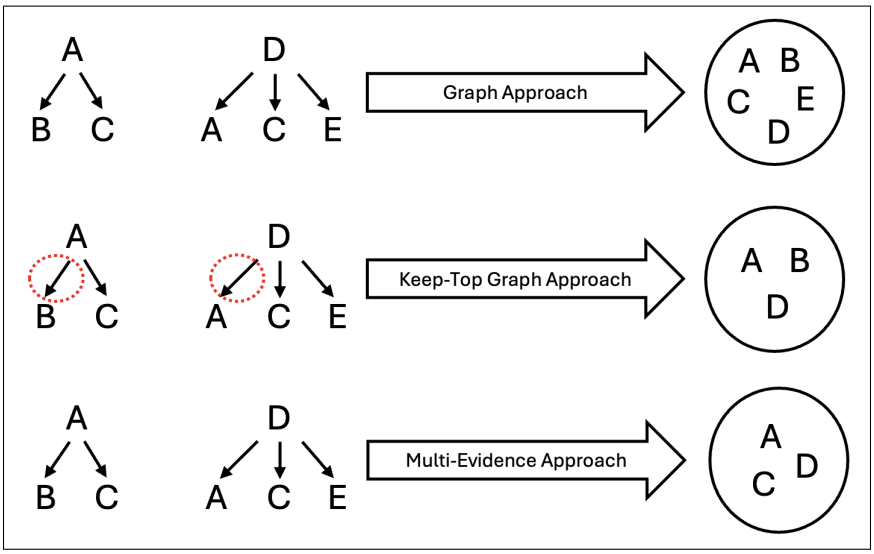


FIGURE 3.10: Examples of the three different approaches to creating student-assisted clusters. The directed edges indicate the responses that a student identifies as semantically similar to the muddy card they wrote. The red-dashed circles indicate that this pair of responses had the highest cosine similarity. The circle with the letters represent the produced student-assisted cluster.

identified that muddy card ‘C’ is semantically similar. The student who wrote muddy card ‘D’ identified that muddy cards ‘A’ and ‘C’ are semantically similar. In this way, two students have agreed that ‘A’ and ‘C’ are similar, so there are multiple pieces of evidence that these two responses are similar. We also include response ‘D’ in the student-assisted cluster, as the temporal ordering of insertion into the database means that response ‘D’ may not be shown to students if it is one of the last responses added to the database. In this regard, the multi-evidence approach can be considered as finding triangles of responses. The resulting cluster only contains ‘A’, ‘C’ and ‘D’, as there is a triangle of links between these sentences (‘A’ $\leftrightarrow$ ‘C’, ‘D’ $\rightarrow$ ‘A’, ‘D’ $\rightarrow$ ‘C’). ‘B’ and ‘E’ are excluded as the criteria for multiple pieces of evidence is not satisfied for these muddy card responses.

When observing the student-assisted results from a lecture, there was an isolated case where a student would write muddy card ‘A’ and select ‘B’ as similar, and another student would write muddy card ‘B’ and select ‘A’ as similar. While one may think that the temporal ordering of insertion into the database prevents a circumstance like this from happening (how can the first student select ‘B’ if response ‘B’ has not already been written), this could occur. This is because under our deployment strategy, at the time a student writes their muddy card response, it is saved to the database. The database entry is later modified to add their student-assisted choice. Hence, a circumstance like this would occur when two students submit their muddy cards at nearly the exact same time. To account for this rare circumstance,

we also considered the link between ‘A’ and ‘B’ as valid under the multi-evidence approach, even though a triangle of responses was not completed. Described as a concrete example, if we have ‘A’ $\rightarrow$ ‘B’, and ‘B’ $\rightarrow$ ‘A’, the student-assisted cluster would contain ‘A’ and ‘B’.

In this section, we have presented three different approaches to using the student-assisted choices. The distinction between these methods is how they enforce the links between student responses. For example, only the link with the top cosine-similarity pair is enforced in the keep-top graph approach. Once a particular method has chosen which links to enforce, the approach to create student-assisted clusters is identical. All muddy cards with a remaining link between them are placed into the same cluster, representing a student-assisted cluster.

In Section 3.1.3.b, we mentioned that a preprocessing step involved removing all but one sentence in each student-assisted cluster from the corpus when doing agglomerative clustering. When benchmarking these methods, we specified that the student response that remained was the student-assisted cluster’s medoid. We believe the medoid would be best suited to represent the student-assisted cluster.

To investigate which of the three approaches yields the best performance, we compared each (as well as the case where no approach is used) by observing how ARI, AMI, and SQAS-15 proportion scores change for each method. This was done using the three manually clustered datasets from the 2025 user study that included data from the student-assisted approach.

3.3 Clustering Open Data Set

Due to the lack of highly fine-grained clustering benchmarks and the absence of muddy card clustering datasets, we have decided to release the public datasets that were gathered as part of the user study that will be shortly described in Section 3.4. As part of this, we will also release the manual clusters labelled by the researchers for public scrutiny and future research. Unfortunately, it is not possible to release these datasets before the submission of this thesis, as our ethics approval requires that students are given a week to withdraw from the public dataset after the study concludes, and this study is running until the end of the semester.

3.4 User Study

Recall that Research Question 2 states:

> (2) How do students and teachers perceive the effectiveness of our muddy card system as it relates to collecting and subsequently analysing common points of confusion?

This research question was investigated through a user study, where we received ethics approval (2024/HE001599) to trial the muddy card system in different units at the University of Sydney. As seen in Table 3.1, twenty units of study encompassing a variety of different disciplines and course levels agreed to participate. One of these units of study included COMP4446/5046, the NLP graduate course at the University of Sydney. In this unit, muddy cards are completed for course credit; however, students could submit their muddy card responses via alternate means so that they were not forced to use this study's muddy card system. Additionally, as the coordinator of the NLP course is the supervisor for this thesis, only the student elements of the user study were included for this unit. That is, when evaluating the teacher perceptions of the system, the coordinator did not participate in the soon-to-be described surveys and interviews.

As part of this research question, we were interested in the perspective of both the students and teachers, resulting in distinct student (see Appendix B.2.1) and teacher (see Appendix B.2.2) participant information statements.

To demonstrate the system, teachers were provided with a brief document outlining how the student and teacher interfaces operate, along with a video demonstration. Students were provided with a written message introducing them to the study that included links to the participant information statement and to a video demonstration (see Appendix C).

When testing the system, the system was configured so that every student and teacher interface experimental condition was tested over four lectures. In particular, a four-lecture cycle had the following conditions tested:

- Student Interface: student-assisted enabled. Teacher Interface: Clustering (Variant Y)
- Student Interface: student-assisted enabled. Teacher Interface: Baseline (Variant X)
- Student Interface: student-assisted disabled. Teacher Interface: Clustering (Variant Y)
- Student Interface: student-assisted disabled. Teacher Interface: Baseline (Variant X)

Unit Name	Unit Code	Faculty	Level	Num. Students Enrolled
Introduction to Electromechanical Systems	AMME1705	Engineering	Undergraduate (1st Year)	537
Data Structures and Algorithms	COMP2123		Undergraduate (2nd Year)	1,011
Multi-disciplinary Engineering	ENGG2112			433
Software Construction and Design 2	SOFT3202		Undergraduate (3rd Year)	314
Natural Language Processing	COMP4446, COMP5046		Postgraduate	397
Data Analysis in the Social Sciences	DATA4207, DATA5207			409
Life and Evolution	BIOL1006	Science	Undergraduate (1st Year)	763
From Molecules to Cells and Organisms	BIOL1009			270
Introduction to Statistical Methods	ENVX1002			402
Fundamentals of Chemistry 1A	CHEM1011			~600 †
Vector Calculus and Differential Equation	MATH2021		Undergraduate (2nd Year)	439
Probability and Estimation Theory	STAT2011			456
Biochemistry and Molecular Biology	MEDS2003, BCMB2001, BCMB2901			842
Introduction to Economic Statistics	ECMT1010	Arts and Social Sciences	Undergraduate (1st Year)	698
Introductory Microeconomics	ECON1001			1,205
Great Books and Radical Texts	FASS2200		Undergraduate (2nd Year)	217
Innovation & Entrepreneurship Foundation	SIEN1000, BUSS4907	Business	Undergrdaute (1st Year), Postgraduate	207
Foundation in Finance	FINC5001		Postgraduate	804
Design Process and Methods	DECO1006, DECO2016	Architecture, Design and Planning	Undergraduate (1st Year)	437
Health, Behaviour and Society	HSBH1003	Medicine and Health	Undergraduate (1st Year)	316

†This approximation is because this unit was taught by two lecturers consecutively, only one of whom used the system. Students could watch either of the lecturers' recorded lectures.

TABLE 3.1: University of Sydney units trialling the muddy card system. The number of students enrolled was at the time of the university's census date.

The order of this four-lecture cycle was adjusted for different units, with the added requirement that the teacher interface always alternated from lecture to lecture. The method used to investigate the student and teacher perspectives of the interface involved surveys and interviews.

3.4.1 Surveys

This study employed three types of surveys: a lecture-by-lecture teacher survey, a final teacher survey, and a final student survey.

3.4.1.a Lecture-by-Lecture Teacher Survey

The lecture-by-lecture survey was used to measure how teachers' perception of the muddy card system changed when moving from each experimental condition (i.e. between the clustering and baseline variants). This was also used to measure teachers' longitudinal perspectives of muddy cards. The lecture-by-lecture survey questions are available in Appendix D. The survey was designed to be completed quickly, taking less than ten minutes. Questions generally focus on how easy and satisfying the system was to use, as well as how teachers plan to address the muddy card responses that students raised. Furthermore, during the weeks that teachers used the clustering variant, two specific questions about the quality of the clusters were asked.

3.4.1.b Final Teacher Survey

The final teacher survey was designed to be a short 15-minute survey for teachers to complete towards the end of the study. Before starting the survey, teachers were prompted to answer the questions in reference to the clustering (variant y) interface. The survey questions are available in Appendix E.1, and generally focus on teachers' perception of muddy cards, overall reactions to the muddy card system, difficulty in learning how to use the system, and satisfaction with the system's capabilities. Most survey questions were adapted from the following standardised HCI surveys: NASA Task Load Index (Hart and Staveland, 1988), Software Usability Measurement Inventory (Kirakowski, 1995), Questionnaire for User Interface Satisfaction (Chin et al., 1988), and Computer System Usability Questionnaires (Lewis, 1995). The survey was deployed using REDCap (Harris et al., 2009) hosted at the University of Sydney.

3.4.1.c Final Student Survey

The final student survey was designed to be a shorter 10-minute survey. To encourage participation, students could optionally enter a draw to win one of five $20 vouchers. The survey questions are available in Appendix E.2, and generally focus on students' perception of muddy cards and the student interface. In addition to the standardised HCI surveys mentioned previously, some survey questions were written to mimic those asked when validating CourseMIRROR (Fan et al., 2017), which would allow for comparisons between our systems. This survey was also deployed using REDCap (Harris et al., 2009) hosted at the University of Sydney.

3.4.2 Interviews

We also conducted semi-structured interviews to better understand the teachers' perceptions of our muddy card system. The interview protocol used to guide the discussion is provided in Appendix F. The interview covered questions about the teacher's general thoughts on muddy cards, followed by specific questions about our system, including whether they would use it in the future and why they thought students interacted with the system in the way they did.

After interviewing teachers, transcripts were created and analysed using elements from the grounded theory methodology. In particular, the process commenced with examining the first transcript and making open codes of the key points that were raised. We then proceeded to the second transcript and continued to develop new open codes as new points were raised, while also collecting additional quotes to add to the previously constructed open codes where relevant. Where open codes are related, axial codes were developed. This process was continued as subsequent transcripts were analysed. Axial codes continued to be developed, and the relationship between them led to the formation of categories. The process of coding was conducted using NVIVO-14.

After each transcript was analysed, a written memo was created documenting 'What do I think is going on here?', 'What have I learned from the new data?', and 'How do I code what I have learned?'. This helped to keep track of what we had discovered and allowed us to better formulate the final theory. The final theory was developed using grounded theory's storyline technique, which is the conceptualisation of the core categories identified. The interview data drove the theory, with the student and teacher survey responses supplementing the analysis. To strengthen the final theory, we provided representative quotes from the interviews.

CHAPTER 4

Results - Research Question 1

In this chapter, we will explore the results relevant to answering the first research question:

(1) How well do sentence embedding models perform when clustering student muddy card responses into groups of semantic similarity with the 'student-assisted approach'?

This chapter will be broadly separated into two parts. The first part (Section 4.1) will involve exploring the preliminary data analysis we conducted using the muddy card responses collected as part of 2024 University of Sydney NLP course. This part will also explore how this data was used to determine an appropriate character threshold for the student interface and a cosine similarity threshold for the student-assisted approach.

The second part (Section 4.2) will involve analysing the muddy card responses that were collected as part of the 2025 user study. This data will also be used to test different methods for incorporating the student-assisted data in an attempt to create higher-quality clusters.

4.1 2024 NLP Course Muddy Card Analysis

4.1.1 Initial Embedding Models Benchmarking

The primary rationale for analysing the 2024 NLP data was to have an early indication of which embedding models are best suited for our muddy card use case. The muddy card system's teacher interface hinges on high-quality embeddings, so a reasonably sound embedding model and algorithm must be implemented for the user study.

The 2024 NLP lecture 8 (439 responses) and lecture 10 (404 responses) muddy cards were manually clustered. These labelled clusters were compared with the clusters produced using agglomerative clustering with the embeddings produced from different SBERT, OpenAI and Voyage AI models. Table 4.1

	Lecture 8			Lecture 10		
	SQAS-15	ARI	AMI	SQAS-15	ARI	AMI
SBERT						
all-MiniLM-L12-v2	**0.282**	**0.315**	**0.412**	**0.235**	0.294	0.377
multi-qa-MiniLM-L6-cos-v1	0.260	0.284	0.397	0.223	0.302	0.400
paraphrase-MiniLM-L3-v2	0.241	0.181	0.300	0.223	0.209	0.314
paraphrase-MiniLM-L6-v2	0.260	0.218	0.343	0.203	0.221	0.325
multi-qa-distilbert-cos-v1	0.264	0.310	0.408	0.225	0.324	**0.432**
all-distilroberta-v1	0.267	0.276	0.376	0.218	0.316	0.416
all-mpnet-base-v2	0.248	0.290	0.398	0.233	**0.326**	0.422
multi-qa-mpnet-base-dot-v1	0.267	0.289	0.374	0.233	0.302	0.394
OpenAI						
text-embedding-ada-002	**0.276**	0.232	0.356	**0.233**	0.319	0.417
text-embedding-3-small	0.260	**0.264**	**0.403**	0.220	**0.340**	**0.435**
text-embedding-3-large	0.262	0.220	0.339	0.230	0.326	0.413
Voyage AI						
voyage-3	0.241	0.268	0.380	**0.238**	**0.315**	0.396
voyage-3-lite	**0.264**	**0.298**	**0.411**	0.223	0.312	**0.402**
Oracle Values	0.328	1	1	0.302	1	1

TABLE 4.1: Clustering performance of different embedding models on two manually clustered lectures worth of muddy card responses. Bold scores indicate the top score for each model family, and underlined scores are the highest of all embedding models for the column. The oracle values provide the largest possible metric score attainable. For example, the highest ARI value is 1 which occurs when the manually and algorithmically labelled clusters are identical. For SQAS-15, the oracle is the best possible SQAS-15 proportion for the given lecture.

displays the largest SQAS-15 proportion, ARI and AMI scores produced from the agglomerative clustering (with the Euclidean distance metric and Ward linkage). As mentioned in the methodology Section 3.2.1.a, as agglomerative clustering produces a hierarchical cluster tree, we calculated the metric score for each k-split, and recorded the max metric score over all splits.

When observing Table 4.1, the ARI and AMI scores are very low even in the best circumstances. This is unsurprising when considering the difficulty of this clustering task. Figure 4.1 is a two-dimensional projection (using TSNE) of each lecture 8 muddy card response's sentence embeddings (using the `all-MiniLM-L12-v2` SBERT embedding model). Embeddings with the same colour were manually coded together, and the black triangles indicate that this sentence was placed in a solitary cluster. There is the constraint here that in order to visualise the embeddings, we are representing 384-dimensional embeddings in two dimensions. That being said, we can see that the distinction between the manually labelled clusters is not obvious when viewing Figure 4.1.

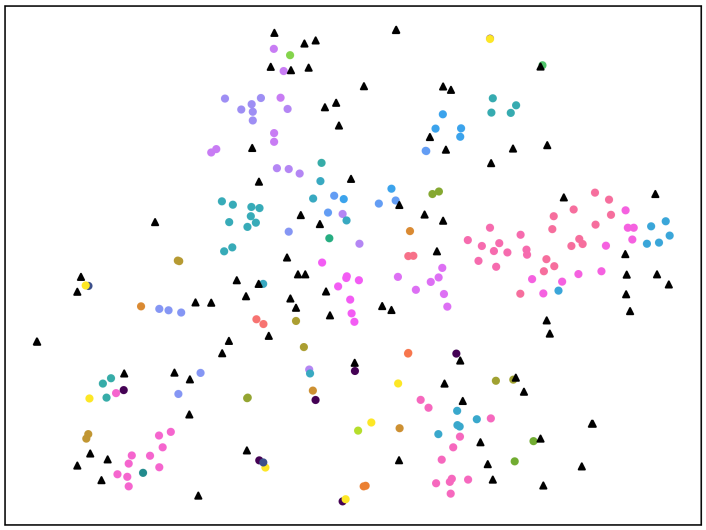


FIGURE 4.1: TSNE 2D projection of the lecture 8 (2024 NLP) muddy card response sentence embeddings (using the `all-MiniLM-L12-v2` SBERT embedding model). Coloured points indicate that these responses were manually labelled in the same cluster. A black triangle indicates that this sentence was manually assigned a solitary cluster.

When considering the SQAS-15 proportion scores in Table 4.1, the values appear quite small. For example, the highest SQAS-15 score for lecture 8 is 0.282, meaning that only 28.2% of students would be satisfied that their question was answered. However, when considering the gold-standard SQAS-15 proportion of 0.328 for lecture 8, we see that the highest SQAS-15 score accounts for 86.0% of the gold-standard score. Hence, this embedding model achieves a score quite close to the optimal SQAS-15 score, yet there is still a large margin for improvement. When considering lecture 10, the highest SQAS-15 score of 0.238 accounts for 78.8% of the gold-standard SQAS-15 proportion, which is considerably less than that in lecture 8. Although there is a considerable difference between the gold-standard proportion accounted for in lectures 8 and 10, with only two lecturers annotated for the 2024 data, it is too difficult to know if this discrepancy is unusual or within the expected range of variation.

While the SQAS-15 proportion scores reveal there is still considerable room for improvement among embedding models, the ARI and AMI scores are relatively much lower. For example, when considering the top ARI score for Lecture 8 in Table 4.1, the top score of 0.315 is quite far from the best possible ARI score of 1 (the best possible scores for each metric are reflected in the 'Oracle Values' row in Table 4.1). ARI and AMI score poorly because these metrics assess consistency across all clusters for both the manually and algorithmically labelled data. However, for our use case, we are more interested in whether the system correctly identifies the major clusters, as performance on the smaller clusters is

not as important as it is unlikely a busy teacher would ever get to address those students anyway (a teacher would likely focus on the largest clusters, especially in a large cohort). Hence, we are more interested in maximising SQAS-15 scores. As we only analysed two datasets in this section, we will explore the embedding models' performance on other manually labelled datasets in Section 4.2.1.

While the algorithmically labelled clusters in Table 4.1 were found using agglomerative clustering with the Euclidean distance metric and Ward linkage strategy, different parameters can be used in agglomerative clustering. Additionally, the embeddings were not scaled nor underwent any dimension reductions. Hence, in an attempt to find the optimal clustering algorithm for use in the user study, we investigated which dimension reduction and agglomerative clustering parameters lead to the highest performance in terms of SQAS score based on the lecture 8 data. The specific hyperparameters that we tuned were mentioned in the methodology Section 3.2.1.b, with the dashboard shown in Figure 3.9 used for analysis.

The first step was identifying which SBERT, OpenAI and Voyage AI embedding models seemed to outperform their peers on the manually labelled lectures. From the results in Table 4.1, we chose SBERT's `all-MiniLM-L12-v2`, OpenAI's `text-embedding-3-small` and Voyage AI's `voyage-3-lite`, as these models outperformed their peers most often across the metrics.

Rather than reporting the single best parameter combination, we instead describe the parameter combinations that generally lead to the greatest SQAS performance. For SBERT's `all-MiniLM-L12-v2`, SQAS was highest when using PCA and retaining 120 to 150 PCA components. Additionally, performance was highest when using the Euclidean distance metric and Ward linkage in agglomerative clustering, and when scaling the embeddings with `scikit-learn`'s 'Standard Scalar'.

Like SBERT, OpenAI's `text-embedding-3-small` performed best using the Euclidean distance metric and Ward linkage in agglomerative clustering. Additionally, performance is highest when using PCA and retaining 170 to 220 components. Figure 4.2 illustrates how the hyperparameter dashboard can be used to find the optimal set of parameters. From this, we can see that almost any scaling method is better than no scaling (the green 'None' lines are generally below all other lines). From the plot, we also lean towards RobustScaler as it typically outperforms other scaling methods, although it is tight.

When considering Voyage AI's `voyage-3-lite`, there was much more variability in the optimal parameters, and it was challenging to reduce the set of most optimal parameters. For agglomerative clustering, the Euclidean and Manhattan distance metrics appeared optimal. Regarding linkage strategy, average was best, yet had the most considerable variability out of the models. PCA was the best

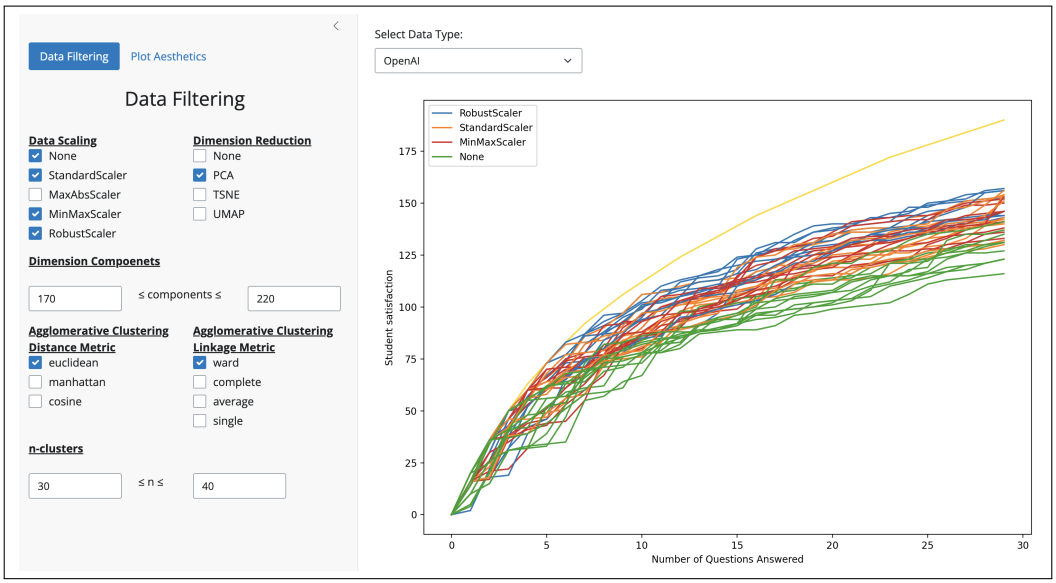


FIGURE 4.2: An example of how the hyperparameter dashboard can be used to find the best set of parameters. This image shows that StandardScaler, RobustScaler and MinMaxScaler lead to higher performance than when no scaler is used.

dimension reduction method, with the ideal number of dimensions lying around 80 to 120. Regarding the embedding scaling method, no scaling method (including when no scaling was applied) stood out among the others.

When deciding the clustering algorithm parameters to include in the user study muddy card system, we decided to use OpenAI's `text-embedding-3-small` with the embeddings scaled using RobustScaler, the Euclidean distance metric and Ward linkage strategy. While PCA with 170 to 220 components was shown to improve performance, we opted not to use PCA as it would fail if there were not enough muddy card responses in the sample. We decided not to use an SBERT model as we would need to configure a way for SBERT to run online, which adds extra complexity to the system's deployment. We also opted not to use Voyage AI because of the much larger variability in the optimal hyperparameters. It was not obvious which parameters to opt for, and the greater variability we saw with the parameters may potentially lead to a poorly performing model. On the other hand, the OpenAI embedding model had a more confined set of parameters with less variability, meaning that choosing which parameters to include for the system deployment was straightforward.

4.1.2 Character Threshold and Student Response Intention

When manually clustering the data for the previously mentioned embedding benchmarking, it became apparent that many students were writing very large responses as evident in Figure 4.3. As we believe this is not aligned with the intention of muddy cards, where students should be concisely writing the most confusing point from the lecture, we decided to implement a character threshold for the student interface in the user study. The threshold was set to be 160, which was just short of the 75th percentile (162 characters) for the character length of all muddy card responses in the 2024 dataset.

A general observation from reading some of the 2024 muddy card responses was that some students recorded more than one muddy point question. This was addressed in the user study's student interface by clarifying that students should only write one muddy card point and by providing two examples of sufficient responses.

Another observation was that some students wrote muddy card responses that did not seem well thought out. We hypothesise that some students may have potentially found the lecture easy, and were pulling at straws to write their muddiest point. Alternatively, some students may have lacked the ability to articulate what was confusing. A further observation was that some students were asking out-of-the-box questions. While it is good to see students engaging with the course content, these types of responses are not the intention behind muddy cards, and should be asked via alternate means. It is also possible that, as muddy cards were for course credit in the University of Sydney NLP course, students tried to complete the task even if they found no part of the lecture confusing. To resolve all of these cases, we implemented the radio button in the student interface (described in the methodology Section 3.1.1) to gauge the reason why students wrote their muddy card response. Teachers can then focus on subsets of responses, such as the muddy card responses that describe things students did not understand.

4.1.3 Student-Assisted Approach Thresholding

The final system adjustment from analysing the 2024 data was about implementing a threshold into the student-assisted approach, whereby a student would only be presented with peer responses that met a set cosine similarity threshold. The rationale for this was that when recruiting lecturers for the user study, a lecturer mentioned that they feared some students may act adversarially and write inappropriate responses that would be propagated through the cohort when providing peer responses in the student-assisted approach.

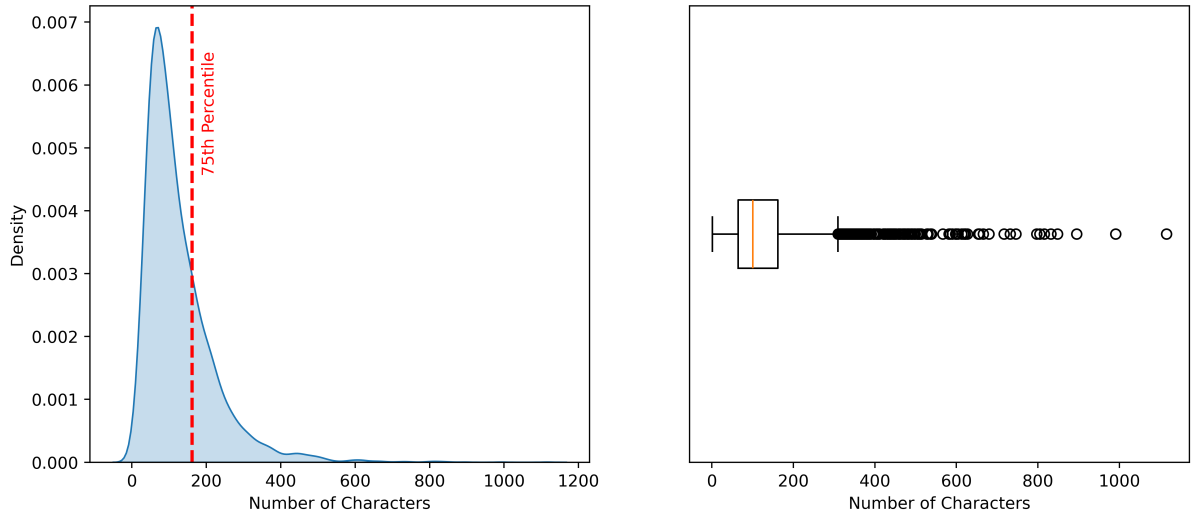


FIGURE 4.3: Density plot (left) and boxplot (right) of the number of characters students used to express their muddy card responses over all lectures.

Using the manually labelled lecture 8 and 10 embeddings, we found the cosine similarity between the embeddings in the same manual cluster and those not in the same manual cluster. The distribution of values is shown in Figure 4.4. The threshold was set to be the first percentile of the within-cluster distribution, which gave a threshold value of 0.33. This value is small enough to encompass most of the within-cluster similarity values, while keeping the possibility for some outside-cluster sentences to appear to a student. We set the threshold in this way to ensure that we minimise the effect of adversarial actors, while leaving room for students to see responses at the semantic 'edge' of what they wrote.

4.2 2025 User Study Muddy Cards Analysis

4.2.1 Embedding Models Benchmarking

From the muddy card responses collected in the 2025 analysis, five different NLP lectures and the muddy cards from a finance unit were manually clustered. When describing the muddy card user study analysis in the methodology Section 3.2.2, we explained how we slightly modified the manual clustering approach so that muddy card responses that could be answered/addressed using a single response were clustered together. The 2D TSNE embedding projection shows that this approach to clustering does not simplify things; clustering muddy card responses remains a challenging clustering task (See Appendix Figure G.1.1).

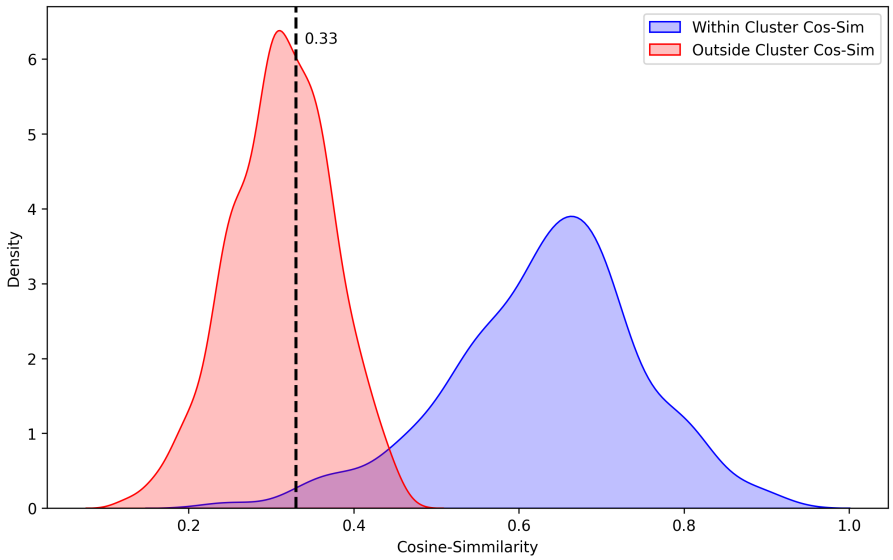


FIGURE 4.4: Distribution of the within-cluster cosine similarity and outside-cluster cosine similarity when considering the manually coded lecture 8 and 10 2024 NLP data. 0.33 was selected as the cosine similarity threshold in the student-assisted approach.

Two researchers manually coded NLP lecture 3 and the finance unit to investigate consistency when manually clustering the muddy card responses. As seen in Table 4.2, the ARI and AMI scores show considerable agreement between researchers; however, there is still much room for improvement (an ARI and AMI score of 1 means perfect agreement). In both muddy card data sets, Researcher 1's manual clustering was considerably closer to the adjudicated clusters than that of Researcher 2. This is likely because Researcher 1 was more experienced with the methodology, and had prior experience manually clustering other lectures worth of muddy cards. Researcher 1 was the researcher who manually clustered the two 2024 NLP lectures and also manually clustered four other 2025 NLP lectures.

Hence, altogether, six muddy card samples were manually clustered. Similar to the 2024 data analysis, we compared the clusters produced via agglomerative clustering with different sentence embedding models on the manually clustered data. Table 4.3 presents the SQAS-15 proportion scores for the 2025 manually labelled data, with the top scores highlighted. When observing the average column in Table 4.3, the proportion of satisfied students remains relatively consistent regardless of the embedding method. Indeed, the difference between the lowest average SQAS-15 score of 0.315 and the highest score of 0.332 is only 0.017. However, while consistent, as shown in the bottom row of Table 4.3, the SQAS-15 scores were quite distant from the gold-standard scores, showing much room for improvement in embedding models.

	Adj. Rand Index		Adj. Mutual Information	
	NLP Lec. 3	Finance	NLP Lec. 3	Finance
Researcher 1 ↔ Researcher 2	0.535	0.686	0.633	0.737
Researcher 1 ↔ Consensus	0.725	0.946	0.826	0.927
Researcher 2 ↔ Consensus	0.686	0.721	0.768	0.792

TABLE 4.2: ARI and AMI between two researchers who manually clustered NLP lecture 3 and a finance lecture. Each researcher's manually clustered data is also compared to the final agreed-upon clustering (the adjudicated clusters).

SQAS-15	NLP Lec. 1	NLP Lec. 3 †	NLP Lec. 4	NLP Lec. 6	NLP Lec. 7	Finance †	Average
SBERT							
all-MiniLM-L12-v2	0.246	0.329	0.357	**0.350**	0.244	0.420	0.324
multi-qa-distilbert-cos-v1	0.231	0.343	0.354	0.337	0.249	0.415	0.321
paraphrase-MiniLM-L3-v2	0.223	0.325	0.371	0.301	0.240	0.430	0.315
paraphrase-MiniLM-L6-v2	0.212	0.336	0.375	0.325	0.244	0.415	0.318
multi-qa-MiniLM-L6-cos-v1	0.227	**0.347**	0.368	0.325	**0.253**	**0.440**	**0.327**
multi-qa-mpnet-base-dot-v1	**0.250**	0.343	0.357	0.346	0.249	0.415	**0.327**
all-mpnet-base-v2	0.220	0.343	**0.379**	0.333	0.235	**0.440**	0.325
all-distilroberta-v1	0.227	**0.347**	**0.379**	0.317	0.244	0.420	0.322
OpenAI							
text-embedding-ada-002	0.235	0.325	**0.368**	0.317	**0.249**	0.445	0.323
text-embedding-3-small	**0.242**	0.343	0.361	0.333	0.240	0.430	0.325
text-embedding-3-large	0.227	**0.347**	0.364	**0.337**	0.244	**0.450**	**0.328**
Voyage AI							
voyage-3	0.242	0.350	0.350	0.325	0.258	0.420	0.324
voyage-3-lite	**0.254**	0.321	0.368	0.321	0.258	0.415	0.323
voyage-large-2-instruct	0.239	0.343	**0.379**	0.313	0.240	0.440	0.325
voyage-code-2	0.235	0.339	0.357	**0.346**	**0.267**	0.445	**0.332**
voyage-code-3	0.239	**0.365**	0.364	0.317	0.249	0.425	0.326
voyage-finance-2	0.231	0.347	**0.379**	0.337	0.240	0.430	0.327
voyage-law-2	0.250	0.339	0.361	0.329	0.258	**0.455**	**0.332**
Data Sample Statistics							
Number of Responses	264	277	280	246	217	200	-
Number of Manual Clusters	173	125	131	149	122	92	-
Gold Standard SQAS-15	0.292	0.473	0.432	0.386	0.401	0.515	0.416

†The manual clusters are the consensus between researchers 1 and 2.

TABLE 4.3: SQAS-15 proportion scores of different embedding models on six manually annotated muddy card samples. Bold scores indicate the top score for each model family; underlined scores are the highest of all embedding models. The data sample statistics provide the number of responses and manual clusters in each data sample. The gold standard SQAS-15 score is the best possible score for the current data sample.

Similar to the 2024 analysis, we also looked at the top ARI and AMI scores for each embedding model and manually clustered data samples. The average SQAS-15, ARI and AMI scores across the six data

	SQAS-15	ARI	AMI
SBERT			
all-MiniLM-L12-v2	0.324	0.359	0.457
multi-qa-distilbert-cos-v1	0.321	0.366	0.475
paraphrase-MiniLM-L3-v2	0.315	0.252	0.359
paraphrase-MiniLM-L6-v2	0.318	0.269	0.381
multi-qa-MiniLM-L6-cos-v1	**0.327**	0.372	0.468
multi-qa-mpnet-base-dot-v1	**0.327**	**0.379**	**0.489**
all-mpnet-base-v2	0.325	0.363	0.470
all-distilroberta-v1	0.322	0.340	0.458
OpenAI			
text-embedding-ada-002	0.323	0.353	0.472
text-embedding-3-small	0.325	0.346	0.468
text-embedding-3-large	**0.328**	**0.366**	**0.479**
Voyage AI			
voyage-3	0.324	0.348	0.462
voyage-3-lite	0.323	0.334	0.448
voyage-large-2-instruct	0.325	0.393	0.465
voyage-code-2	**0.332**	0.365	0.475
voyage-code-3	0.326	**0.400**	**0.510**
voyage-finance-2	0.327	0.359	0.472
voyage-law-2	**0.332**	0.396	0.488
Oracle Values	0.416	1	1

TABLE 4.4: Average SQAS-15, ARI and AMI scores per embedding model for the six manually coded 2025 user study datasets. Bold scores indicate the top score for each model family; underlined scores are the highest of all embedding models. The oracle values provide the largest possible metric score attainable.

samples are shown in Table 4.4. The ARI and AMI scores for each individual sample can be found in Appendix Table G.1 and Appendix Table G.2.

When observing Table 4.4, unlike the SQAS-15 scores, there is much more variability in ARI and AMI for different embedding models. `voyage-code-3` presents itself as the highest performing model when considering ARI and AMI. `voyage-code-3` is a model optimised for code retrieval, which includes using natural text to retrieve code[1]. Hence, and in light of Appendix Table G.1, it is unsurprising that `voyage-code-3` has high performance for the NLP lectures, as NLP is itself a computer science unit where terminology would likely cross-over with that in code retrieval. However, what is surprising is that for the finance data sample, `voyage-code-3` is most successful when considering the AMI score (Table G.2), and is very competitive when considering its ARI score (Table G.1).

[1]This is explained in the documentation for `voyage-code-3` here: https://blog.voyageai.com/2024/12/04/voyage-code-3/

One possible explanation for why `voyage-code-3` has such high performance on the finance data sample is that fine-tuning on the code-related training data could potentially cause unpredictable behaviour on out-of-domain content. Betley et al. (2025) fine-tuned different large language models (LLMs) on a small (6,000 entries) synthetic code completion dataset, where each code response contained a security vulnerability. After fine-tuning, OpenAI's `GPT-4o` model was shown to produce misaligned responses 20% of the time on non-coding tasks (when not fine-tuned, misalignment is 0%). Qi et al. (2023) found that even without malicious intent, when fine-tuning LLMs on commonly used datasets, the safety alignment of the LLMs degrades. These examples from literature demonstrate that fine-tuning LLMs can result in unpredictable behaviour. We do not imply this is the case for the `voyage-code-3` model; instead, we offer this as a potential explanation. It is also possible that this result with `voyage-code-3` was simply a consequence of variability. Future research should collect more muddy card samples from non-computer science based units to determine whether the result with `voyage-code-3` was an isolated instance.

When considering the ARI and AMI scores in Table 4.4 for `voyage-law-2`, we see that this model is the second-best Voyage AI model. Additionally, when considering `voyage-law-2`'s performance for NLP lecture 7 individually, we see that it has the highest ARI and AMI scores (Table G.1 and Table G.2). While it may seem surprising that this model would stand out for a computer science unit, as law often intersects with different domains, `voyage-law-2` was trained on data from many fields, with 12% of training data being classified as part of the 'code' domain [2].

When considering the SQAS-15, ARI and AMI scores for the six samples manually coded from the user study, the embedding model that leads to the highest performance is `voyage-code-3`. This is because it has the highest ARI and AMI scores (by a considerable margin), and a SQAS-15 score just shy of the leader. As OpenAI's `text-embedding-3-large` also scored highly, we tried to compare it with `voyage-code-3` on MTEB (Muennighoff et al., 2023). However, at the time of writing, `voyage-code-3` has not been tested for clustering on MTEB.

Even though `voyage-code-3` is selected as the highest performing model, in terms of the SQAS-15 score, which was specifically designed for our muddy card use case, it seems that all embedding models have nearly indistinguishable performance. For example, despite `voyage-code-3` being the top performer, its SQAS-15 score suggests that it does not seem to offer that much more than other

[2] This is explained in the documentation for `voyage-law-2` here: https://blog.voyageai.com/2024/04/15/domain-specific-embeddings-and-retrieval-legal-edition-voyage-law-2/

embedding models for the muddy card application. As emphasised earlier, even though the SQAS-15 scores are consistent across embedding models, they are still far from the gold standard, as seen in Table 4.3. This would suggest that there is much room for improvement among embedding models, but it may also be that clustering algorithms need improvement. For the 2025 muddy card analysis, we did not focus on trialling different clustering algorithms and parameter combinations. This is an avenue for further research.

Rather than hyperparameter tuning, another approach to improve clustering performance is to incorporate students' decision-making in the clustering process. This is what the student-assisted approach attempts to accomplish.

4.2.2 Student Assisted Approach Results

Of the six manually coded data samples, three included data (NLP lectures 1, 3 and 7) from the student-assisted approach. Only three data samples contained this data because the student-assisted approach was an experimental condition that students did not use every week. For each of the three data samples, we tested the graph approach, keep-top graph approach, and multi-evidence approach as described in the methodology Section 3.2.3.

Table 4.5 shows the average SQAS-15 score when not incorporating the student-assisted data and using the three previously mentioned approaches to incorporate the student selections. From the table, it is clear that the multi-evidence approach substantially improves the SQAS-15 score as opposed to when the student data is not incorporated. The keep-top approach always offers some improvement when the student-assisted data is not used at all; however, this increase is minor in some cases.

The graph approach is substantially worse than all other approaches, including when student-assisted data is not used at all. This is unsurprising when looking at the size of the peer-assisted clusters that the graph method creates for each lecture. For NLP lecture 1, the graph approach created a student-assisted cluster of size 192. There are only 264 responses in lecture 1, meaning this cluster alone captures 72.7% of responses for the lecture. This is an enormous cluster, especially considering that when we manually coded lecture 1, we found 173 unique clusters. Similarly, for NLP lecture 3, the graph approach created a cluster that accounted for 76.9% of responses, and in NLP lecture 7, a cluster accounted for 71.9% of responses. The likely reason for this is that under the graph approach, it only takes one student to mistakenly select that two semantically different responses are the same for different graph segments to be combined. The other methods are more robust to this. For example, even if in the keep-top graph approach a student were to select many peer responses, including ones that should not be lumped

SQAS-15	No Student Assisted	Graph	Keep-Top Graph	Multi-Evidence
SBERT				
all-MiniLM-L12-v2	0.273	0.188	0.305	**0.330**
multi-qa-distilbert-cos-v1	0.274	0.192	0.315	**0.332**
paraphrase-MiniLM-L3-v2	0.263	0.157	0.266	**0.303**
paraphrase-MiniLM-L6-v2	0.264	0.185	0.266	**0.323**
multi-qa-MiniLM-L6-cos-v1	0.276	0.190	0.303	**0.329**
multi-qa-mpnet-base-dot-v1	0.281	0.176	0.303	**0.337**
all-mpnet-base-v2	0.266	0.193	0.297	**0.331**
all-distilroberta-v1	0.273	0.165	0.297	**0.320**
OpenAI				
text-embedding-ada-002	0.270	0.178	0.304	**0.327**
text-embedding-3-small	0.275	0.182	0.320	**0.322**
text-embedding-3-large	0.273	0.167	0.305	**0.327**
Voyage AI				
voyage-3	0.284	0.171	0.309	**0.321**
voyage-3-lite	0.278	0.189	0.304	**0.332**
voyage-large-2-instruct	0.274	0.167	0.296	**0.315**
voyage-code-2	0.280	0.188	0.294	**0.326**
voyage-code-3	0.284	0.196	0.313	**0.331**
voyage-finance-2	0.272	0.185	0.310	**0.325**
voyage-law-2	0.282	0.191	0.305	**0.330**

TABLE 4.5: Average SQAS-15 score for each embedding model when no student-assisted approach is used, and when the graph, keep-top graph, and multi-evidence student-assisted approaches are used. The largest value in each row is in bold.

together, by retaining the selected response with the highest cosine-similarity, we limit the chance that the selected peer response was completely off topic. We also suspect that a contributing factor to the poor performance of the graph approach is that in the NLP unit, there is a very high proportion of international students (94.2%), with 93.5% of students not being native English speakers. Non-native speakers may find it more challenging to determine the semantic similarity of peer responses. Hence, this is an approach we hope to test again in the future with a wide range of units.

While the multi-evidence approach dominates when considering SQAS-15, this is not the case for ARI and AMI, as shown in Table 4.6. For ARI, no method of implementing student-assisted selections improves clustering. Additionally, when considering AMI, results are mixed, with the no student-assisted data and the multi-evidence approach yielding similar results.

	Adjusted Rand Index				Adjusted Mutual Information			
	No Student Assisted	Graph	Keep-Top Graph	Multi-Evidence	No Student Assisted	Graph	Keep-Top Graph	Multi-Evidence
SBERT								
all-MiniLM-L12-v2	**0.365**	0.011	0.199	0.317	0.446	0.104	0.374	**0.453**
multi-qa-distilbert-cos-v1	**0.392**	0.011	0.203	0.326	**0.469**	0.101	0.384	0.465
paraphrase-MiniLM-L3-v2	**0.283**	0.010	0.182	0.264	0.365	0.096	0.348	**0.393**
paraphrase-MiniLM-L6-v2	**0.291**	0.010	0.177	0.274	0.381	0.097	0.347	**0.407**
multi-qa-MiniLM-L6-cos-v1	**0.406**	0.011	0.198	0.314	**0.478**	0.100	0.376	0.456
multi-qa-mpnet-base-dot-v1	**0.442**	0.010	0.201	0.327	**0.494**	0.099	0.377	0.459
all-mpnet-base-v2	**0.386**	0.010	0.201	0.319	**0.458**	0.098	0.374	0.457
all-distilroberta-v1	**0.362**	0.011	0.206	0.326	0.449	0.100	0.379	**0.454**
OpenAI								
text-embedding-ada-002	**0.372**	0.011	0.204	0.313	**0.458**	0.104	0.380	**0.458**
text-embedding-3-small	**0.360**	0.011	0.201	0.309	0.450	0.102	0.375	**0.452**
text-embedding-3-large	**0.425**	0.011	0.198	0.326	**0.488**	0.104	0.380	0.469
Voyage AI								
voyage-3	**0.386**	0.011	0.200	0.325	**0.467**	0.105	0.385	0.465
voyage-3-lite	**0.400**	0.011	0.196	0.338	**0.469**	0.103	0.373	0.461
voyage-large-2-instruct	**0.403**	0.010	0.197	0.334	**0.489**	0.100	0.382	0.476
voyage-code-2	**0.450**	0.011	0.197	0.332	**0.519**	0.107	0.380	0.477
voyage-code-3	**0.446**	0.011	0.200	0.335	**0.524**	0.101	0.374	0.479
voyage-finance-2	**0.392**	0.010	0.203	0.314	**0.472**	0.100	0.387	0.448
voyage-law-2	**0.427**	0.011	0.201	0.336	**0.503**	0.103	0.381	0.482

TABLE 4.6: Average ARI and AMI score for each embedding model when no student-assisted approach is used, and when the graph, keep-top graph, and multi-evidence student-assisted approaches are used. The largest value in each row is in bold.

Given these results, it is difficult to determine whether the multi-evidence approach is better than not using the peer-assisted data. As it currently stands, if you desire to have the highest quality clusters possible, the ARI and AMI scores in Table 4.6 would suggest it is better not to use the peer-assisted data at all. However, for the muddy card application, we believe having high-quality clusters among all clusters is unnecessary. We believe that if there are many responses, a teacher is not going to sit through and read every cluster that the system produces. Instead, they are likely to only focus on the first few largest clusters. In this regard, we believe that maximising SQAS-15 is more critical.

Regardless, further research is needed to investigate whether the multi-evidence approach leads to better results for muddy card applications. In particular, more diverse data samples are required, such as data from non-NLP units and units with varying demographic makeups. While there remains uncertainty, a key contribution of this thesis is the proposal of the student-assisted approach, and using this data to develop student-assisted clusters. This is a new concept in muddy card research, and the findings in this thesis show that it is something worth pursuing.

4.2.3 Muddy Card Clustering Benchmark

As mentioned in methodology Section 3.3, we plan to release the manually clustered 2025 user study data samples. When available, details to access these datasets will be made at the GitHub repository here[3]. As evident by the 'number of responses' and 'number of manual clusters' at the bottom of Table 4.3, the muddy card data contains fine-grained clusters unlike any of the MTEB clustering datasets, as seen in Table 2.1 of Chapter 2.

4.3 Conclusion

In this chapter, we explored the results relevant to Research Question 1. We analysed how different SBERT, OpenAI, and Voyage AI models perform against eight manually clustered muddy card data samples. When considering the embedding model's performance using SQAS-15 proportion scores (a metric designed to better reflect clustering quality in muddy card applications), it was found that embedding models had very similar scores. However, these values were considerably far from the gold-standard SQAS-15 scores, indicating that there is still much room for improvement when considering embedding models' performance. This was confirmed by considering the low ARI and AMI scores for each embedding model.

One potential way to improve the performance of clustering is to adjust the hyperparameters related to the agglomerative clustering process. The best performing SBERT, OpenAI and Voyage AI embedding models from the 2024 NLP muddy card embedding benchmarking were chosen. While reducing the embedding using PCA and using the Euclidean distance metric in agglomerative clustering was optimal for the three embedding models, there was otherwise considerable inconsistency among the optimal sets of parameters. This makes hyperparameter tuning a challenging prospect.

Another potential way to improve clustering performance is to incorporate the student-identified semantically similar muddy card data from the 'student-assisted approach' into our muddy card system. We tested three different approaches to integrate the student-assisted data. When considering ARI, it seems that not incorporating the student-assisted data leads to higher performance in most circumstances. When considering AMI, we found that the 'multi-evidence' approach or when the student-assisted data is not incorporated is optimal. However, when considering SQAS-15, we found that the 'multi-evidence' approach led to higher performance. With these findings in mind, we believe that the 'multi-evidence' approach is promising, and we have plans for a follow-up study to explore this approach on a more diverse dataset.

[3]GitHub: https://tjelton.github.io/Dissemination-Muddy-Card-Analysis/

CHAPTER 5

Results - Research Question 2

In Chapter 5, we will explore the results relevant to the second research question:

(2) How do students and teachers perceive the effectiveness of our muddy card system as it relates to collecting and subsequently analysing common points of confusion?

This will involve exploring the interview and survey data from the 20 different units of study that agreed to participate in trialling the muddy card system. Overall, 13 lecturers agreed to be interviewed about their experience with the muddy card system, and 14 lecturers completed the end-of-study survey. 90 students filled out the student end-of-study survey, but 12 responses were removed because the students had never used the system. Of the 78 valid submissions, 33 students were enrolled in the University of Sydney's NLP unit, where muddy cards were attributed to course credit, and 43 were enrolled in other units. One student was enrolled in both an NLP and a non-NLP unit, and one student did not indicate which unit they were enrolled in.

Teachers could also complete a lecture-by-lecture survey as described in the methodology Section 3.4.1.a. After using the muddy card system, lecturers could answer questions about the interface variant they experienced for the current lecture. Only 13 responses were recorded for this, meaning there is an inadequate number of responses to analyse the Likert-style questions (these 13 responses are distributed among the different interfaces experienced). However, the written answers from lecturers on how they plan to address student responses have been incorporated into this analysis.

Rather than discussing the survey and interview results separately, this chapter has been arranged topically. We will first explore the low system engagement among students and the reasons lecturers believe this is the case. We will then discuss the different perspectives on the system's effectiveness, followed by how lecturers would prefer a system with instantaneous feedback features. We will then discuss the

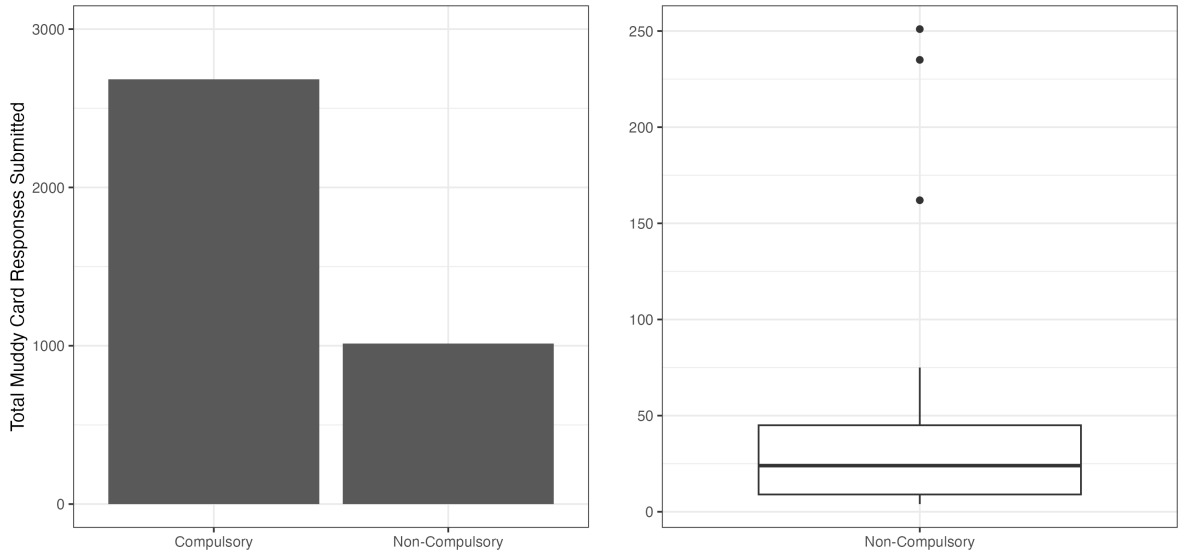


FIGURE 5.1: The left plot shows the total responses collected for compulsory (which only contains the NLP unit) and non-compulsory units. For the non-compulsory units (19 out of 20 units), the right plot shows the distribution of the total number of muddy card responses per unit.

perspectives on whether lecturers and students would like to use the system in the future, before finally discussing whether our interviews reached saturation.

5.1 Low System Engagement

5.1.1 Low Student Engagement

From Week 1 to Friday of Week 11 (there are 13 course-teaching weeks at the University of Sydney), 3,697 muddy card responses were collected in total. Recall that one of the twenty units in this study was an NLP course taught by the supervisor affiliated with this project. For this NLP course, muddy cards are compulsory (they are attributed a small portion of the student's grade), and this was the only unit out of the 20 that had compulsory muddy cards. Despite having a similar number of students to many of the other units in this study (see Table 3.1), the NLP course accounted for 2,683 muddy card responses. This means that 1,014 responses were from the units where muddy cards were not compulsory, as shown in the column graph on the left of Figure 5.1. Considering the number of units in this study, the number of responses is quite low. The boxplot on the right of Figure 5.1 emphasises that for most units of study, there were very few responses, with most units receiving less than 50 responses over the semester.

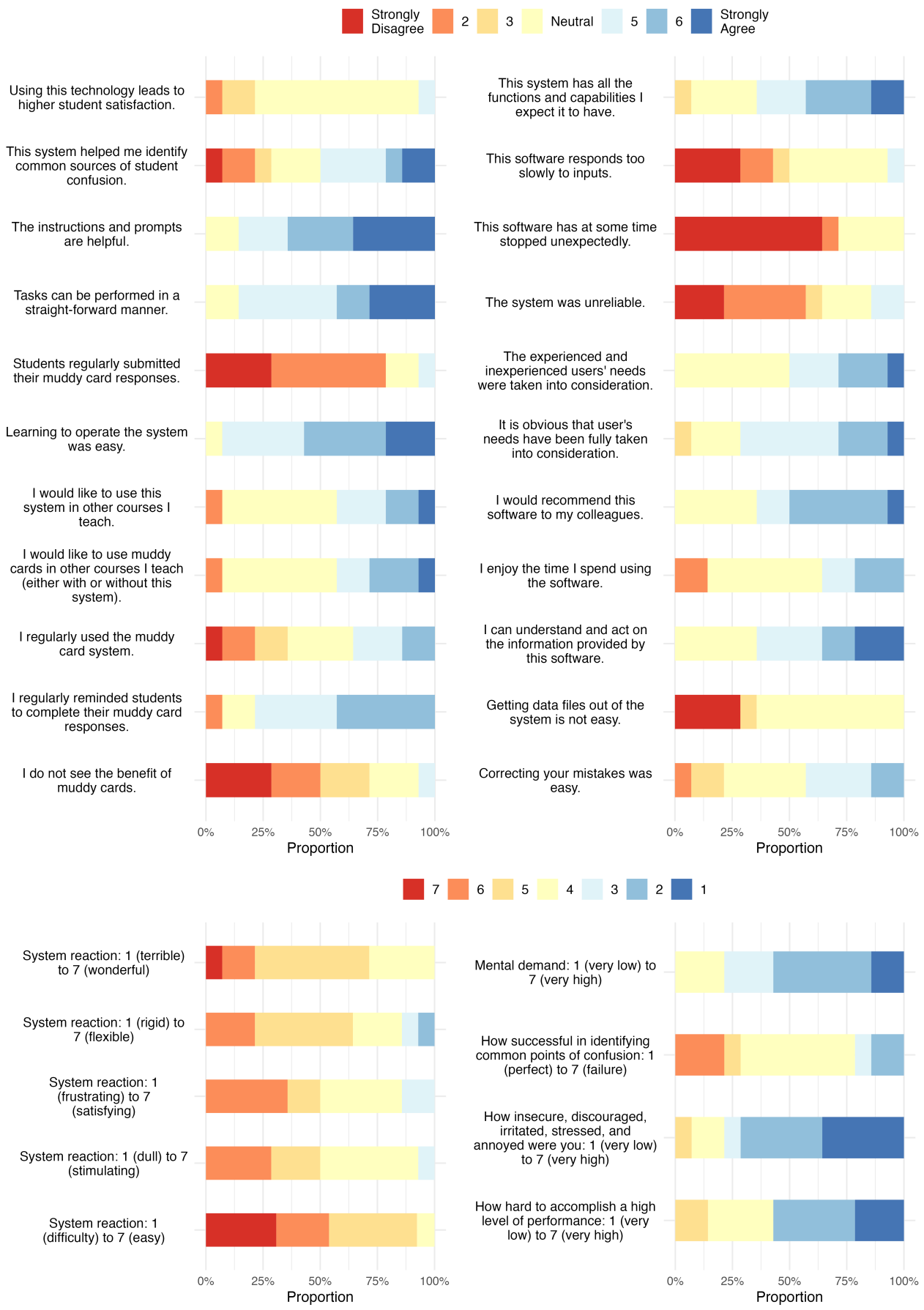


FIGURE 5.2: Final teacher survey response results. Some question names were shortened. The unaltered names can be found in Appendix E.1.

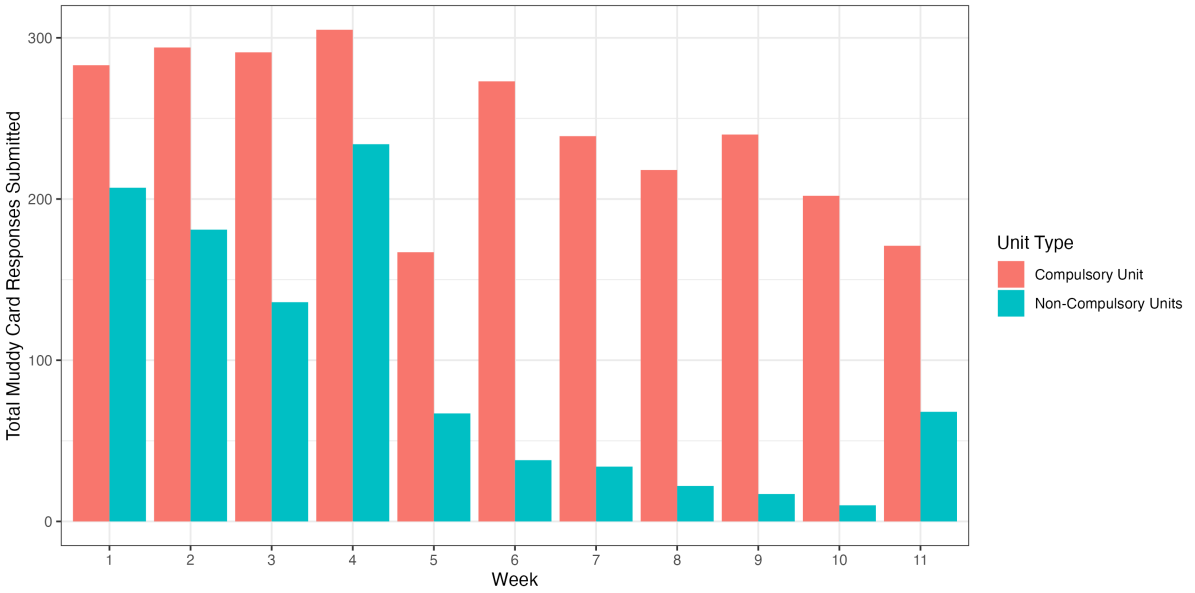


FIGURE 5.3: Total muddy card responses collected per week, separated by whether muddy cards were compulsory (i.e. for course credit) for a unit or not.

The low student response rate is despite lecturers regularly reminding students to complete their muddy card responses. Figure 5.2 contains the survey results for the final teacher survey. Most lecturers agreed to some extent with the statement, 'I regularly reminded students to complete their muddy cards responses' (on the 7-point Likert scale: $\mu = 5.07$, $\sigma = 1.14$). From the interviews, lecturers suggested different reasons why they believed the response rate was so low.

A few lecturers revealed that lecture attendance naturally declines in lectures throughout the semester. One lecturer said that *"the decrease in lecture attendance is like a universal thing that's been going on 15 years."* Another lecturer explained that at the time of the interview, only 5% of enrolled students attended their most recent lecture. If students are not attending lectures, it is unlikely they would go out of their way to fill out a muddy card. Figure 5.3 reinforces this, where the number of muddy cards completed also decreases throughout the semester. This trend is even more apparent for the non-compulsory units, where the number of responses steadily decreases. There is an upswing in week 4, but this coincided with the research team providing some tips to the lecturers on how they can improve student uptake (such as allocating lecture time to complete muddy cards). There was also an increase in week 11, when one unit of study resumed using the system. However, aside from this week, there is a clear downward trend.

Some lecturers said they believed the low response rate is because, in general, students typically have low engagement with optional activities. One interviewee said, *"But that's always the hardest thing when you have these optional things that are additional structures of support if it requires their engagement... It's always so challenging to get people to actually do the thing."* In this regard, the low uptake was not necessarily tied to our muddy card system, but rather to the difficulty in getting students to participate in optional active learning techniques. In reference to Mentimeter, which educators often use to incorporate audience participation in lectures, one interviewee mentioned that *"engagement is sometimes limited."*

The low response rate and the decreasing response rate trend shown in Figure 5.3 is not specific to our muddy card system. One interviewee previously used muddy cards in a unit with a large cohort of 1,600 students. When using it previously, they explained that at the beginning of the semester, there were originally many responses (around 18.75%-25% of the cohort), but this dropped throughout the semester. By the end of the semester, only 2.5%-3.125% of the cohort were submitting responses. Another interviewee had implemented something similar to muddy cards, where students were given sticky notes and could write anything they liked or disliked about the lecture. This lecturer explained that they had a similar number of responses with the muddy card system and what they historically had using the sticky note system; *"I would say, similar proportion as we have in the muddy card system right now, not many students. But I will say, as long as I have ... some student telling me, then at least I have some feedback."*

Some lecturers perceived that the low response rate was linked to whether the muddy card system offered unique value when considering the other tools available to students. At the University of Sydney, a class discussion forum named EdStem (nicknamed 'Ed') is commonly used. As a discussion forum, students will ask the teaching team questions and typically receive a timely response. One teacher said, *"since we already have the discussion board where they can post things they're unsure of, that maybe also conflates the amount of engagement with [the muddy card system]."* One lecturer explained they responded to students' muddy card responses the following week and found that students re-posted their muddy card question on EdStem. They said, *"[The students] were not patient to get their answers by the following week, and I found all those questions on Ed just as posts, and they were, you know, like addressed before... I had this fact posts, so at which point I decided not to use [the muddy card system] anymore."* While students can ask questions on the discussion forum, we still believe that a key value proposition of muddy cards is that there are memory benefits when reflecting on their learning (Prince, 2004). Muddy

cards enable more students to engage in reflection. For a discussion forum, this is restricted to the first person who raises a particular question.

In the era of AI chatbot systems, one instructor questioned why students would write a muddy card response when they could ask an AI agent directly. They said, *"Now, it's conceivable that the students said... 'Why deal with the muddy card thing and wait for [the instructor] to respond? Why not just go straight to the [AI] agent?"'* In light of this point, the instructor revealed that *"only about a 3rd of the students are engaging with the Cogniti [AI] agent."* Once again, it seems that even with active learning tools at student's fingertips, there is a low uptake for optional activities.

To help teachers administer the muddy card system during lectures, each lecturer was provided with a QR code slide that students could scan on their mobile phones to access the student interface easily. One lecturer placed the QR code at the start of the lecture, but found that this was ineffective; *"[At] the start of class ... students want to socialise. They want to talk to people. And it's a bit distracting to pull out your phone and take a photo."* Some lecturers placed the QR code at the end of the lecture, but this meant that students were often packing up ready to go before completing their muddy card. One interviewee explained that students would be closing their bags by the end of the lecture; *"What I noticed was during the lectures, especially towards the end, when they were coming up to the conclusion of [the] summary slide, and then the muddy card slide, like laptops were already closing, bags already ... being zipped."* A few interviewees perceived that the best time to display the QR code to maximise response rate is halfway through the lecture, especially during a 2-hour lecture with a break in the middle. One interviewee said, *"if it's in the middle of the lecture, the students are sitting there. They're in the middle of the lecture. They're talking, you know, they don't understand stuff. So it's very immediate to them, I think far and away is the best way of doing it."*

Some lecturers suggested that the low response rate could also be attributed to the muddy card system being deployed on a website outside of the university's ecosystem. One lecturer said, *"Really, it's about ... integration and convenience, and as soon as you have students having to go somewhere else ... the uptake drops off."* This lecturer suggests that if the system were available in Canvas (the course management system used at the University of Sydney), there would be a larger uptake. On a similar vein, some lecturers suggested that the student interface needs to have an *"extremely low barrier to entry"*, and this is hindered by the drop-down menu that asks students to enter which week it is and the unit of study the muddy card is for. We plan to incorporate this suggestion into the next iteration of the muddy card system, where the system will automatically determine which unit of study a muddy card

is for (through unique URL links), and the lecture week will be identified based on the date the student submits their response.

Some lecturers also explained that they believed the low response rate could be due to students having no questions. One lecturer identified that the content may be easier in some weeks, leading to fewer responses. Another lecturer identified that if they ask the lecturer for clarification, there is no need to submit a response; *"They don't really submit the question in the survey at the end if the question [is] already solved."* One lecturer said that they believe it could be related to their effort to make sure their lecture content is clear.

5.1.2 Low Teacher Engagement

Another potential reason for the low response rate was some instances of low teacher engagement. Ironically, although our system was designed to help teachers quickly implement muddy cards into their lectures, some teachers were too busy to thoroughly engage with the system. In an interview, one lecturer admitted knowing that students had not understood the point of muddy cards. In particular, they said, *"I felt a bit guilty because it was a very busy start of the year, and the lectures were very full at the beginning in particular, and I knew that they hadn't really got the idea."* This interviewee also mentioned that they did not even remember being sent the teacher interface, and subsequently never used it.

Another lecturer admits they did not engage with the system much during the semester as they were quite busy. However, to prepare for the interview, they did end up using it, and liked the system; *"So unfortunately, I was a bit slack, and, to be truthful I wasn't keeping up with the muddy card system during that series of lectures, and I've only had a chance to really engage with it later, ... after that lecture series was complete. But now that I have had an opportunity to look at the interface properly and thoroughly, I do see great value in the system."*

Two lecturers suggested that a future improvement would be for the system to send a weekly reminder that students have submitted muddy card responses. One interviewee said, *"I do now on reflection wish that I received maybe a notification, or even a weekly or a daily digest that just said, 'Hey, 10 students have submitted muddy cards for lecture number 2. Click here if you'd like to go on view [the responses]'. I think that would have been a good trigger for me to engage with the system more frequently ... and it would have prompted me to remind the students that the system was available more often."*

In this chapter, we have spent the first section focusing on the student and teacher engagement for the muddy card system. This is because engagement with the system is related to the system's usefulness. If student responses are not coming in, the system cannot analyse responses effectively, as there are no responses to analyse. From the teacher survey responses, despite 93% of teachers agreeing to some extent that they reminded students to regularly submit their muddy card responses ($\mu = 5.07$, $\sigma = 1.14$), 79% disagreed that students regularly submitted their muddy card responses ($\mu = 2.21$, $\sigma = 1.25$). Due to widespread instances of low student response rates, it is unsurprising that when asked to rate the statement 'I regularly used the muddy card system', support for this statement was neutral ($\mu = 3.86$, $\sigma = 1.51$), as visualised in Figure 5.2. After all, there is little to no point in using the system if there are few to no student responses. This low response rate poses a challenge for analysing the effectiveness of the muddy card system.

5.2 System Effectiveness

In this section, we will explore the system's effectiveness at collecting and analysing muddy card responses. We will first examine the effectiveness of collecting responses by considering the student survey responses. The effectiveness at analysing these responses will be explored through the interviews and final teacher survey.

5.2.1 Student Perspective

Figure 5.4 displays the student survey response results for all survey questions. From this, we can see that most students agreed to some extent that it was simple to use the system ($\mu = 5.29$, $\sigma = 1.38$) and disagreed that the system was frustrating to use ($\mu = 3.12$, $\sigma = 1.51$). When observing the remainder of the survey questions in Figure 5.4, there are no instances where most students dislike an aspect of the system. However, some questions did receive votes learning close towards neutrality. For example, when considering the question, 'The speed of the system is fast enough', a similar number of students agreed and disagreed with this statement ($\mu = 4.86$, $\sigma = 1.51$).

One question asked students to rate 'How mentally demanding was the system?' on a scale of 1 (very low) to 7 (very high). As seen in Figure 5.4 (the first plot on the bottom left section), the results for this question were quite mixed ($\mu = 3.55$, $\sigma = 1.57$). When looking through the student survey results, we separated the results of each survey question by whether muddy cards were compulsory or

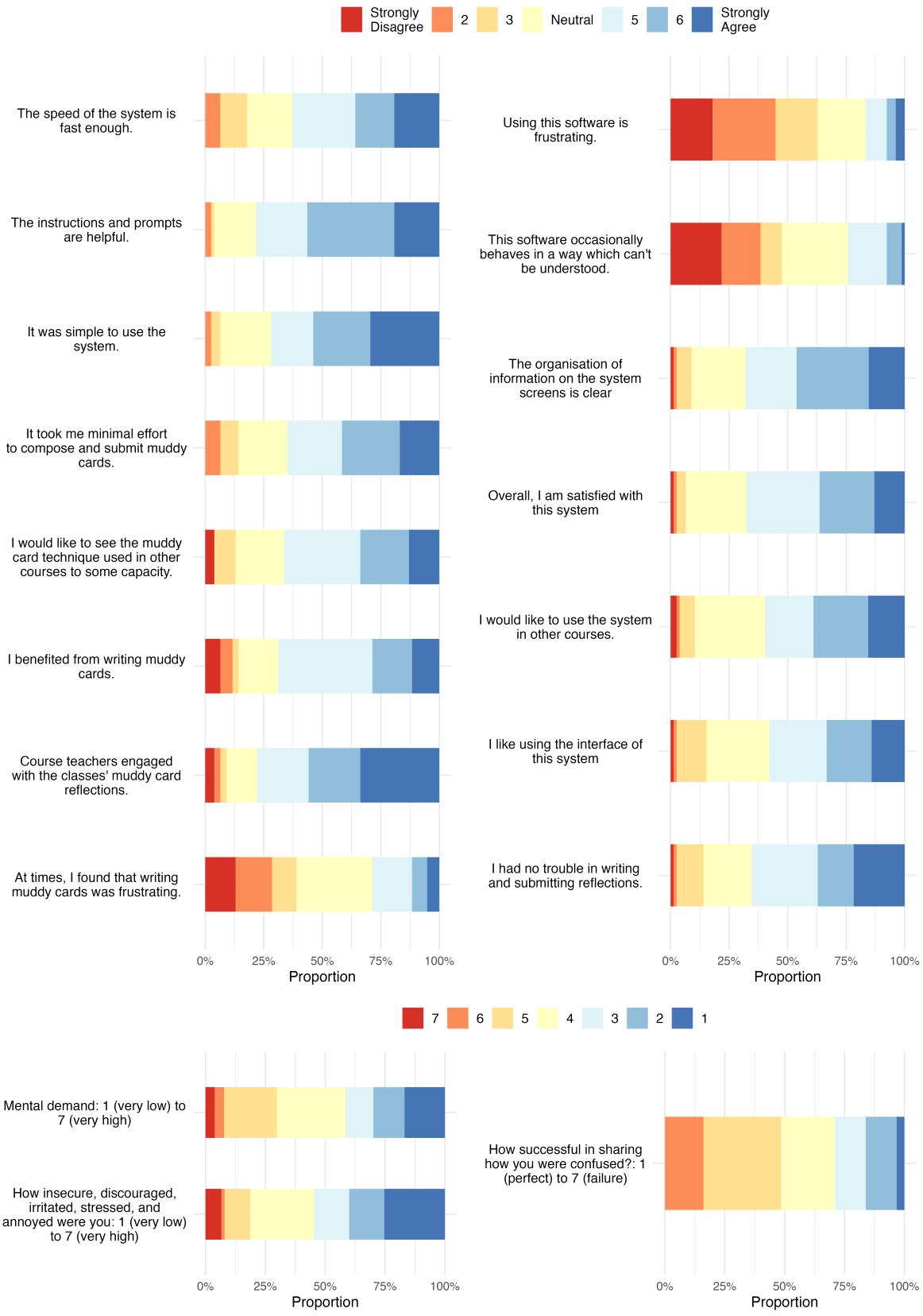


FIGURE 5.4: Final student survey response results. Some question names were shortened. The unaltered names can be found in Appendix E.2.

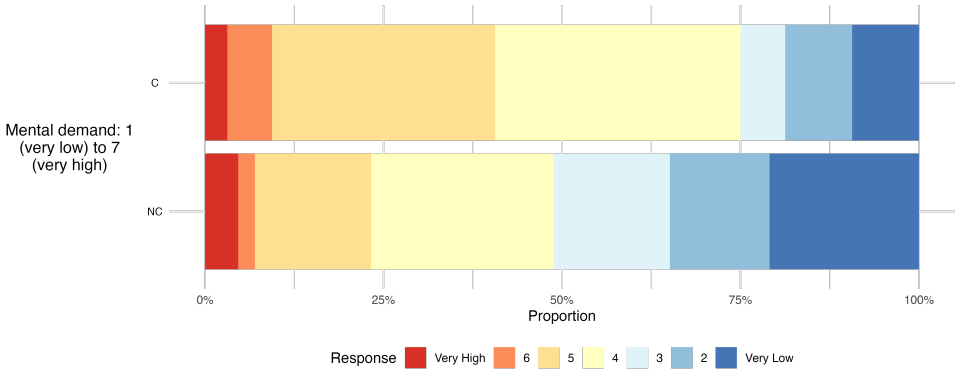


FIGURE 5.5: Results for the questions asking students to rate 'How mentally demanding was the system?' on a scale of 1 (very low) to 7 (very high), separated by whether students were enrolled in units where muddy cards were compulsory or not. 'C' denotes compulsory, and 'NC' denotes non-compulsory.

non-compulsory for a student's unit to see if there appears a difference in the results[1]. One noticeable difference was in the mental demand question, and so Figure 5.5 separates the responses to this question by whether muddy cards were compulsory or not. Here, we do see that mental demand was highest among students in the NLP unit with compulsory muddy cards. The reason for this is unknown, and for future research, it would be worthwhile interviewing students to better gauge their perception of muddy cards in general and the muddy card system.

5.2.2 Teacher Perspective

As previously mentioned, support for the statement 'I regularly used the muddy card system' was neutral among lecturers ($\mu = 3.86$, $\sigma = 1.51$). Over the semester, there were many instances where a lecturer received a low number of responses. Indeed, when it came to the interview, lecturers often forgot that there were two different teacher interfaces, or forgot the details of the interfaces. This forgetfulness is unsurprising, given that they could not regularly experience the interfaces if no (or little) student responses were received. In one instance, a lecturer received more than five responses for a single lecture, and hence, they only recall experiencing one of the interfaces. Cases like these increased the difficulty when comparing the interfaces during the interview analysis. Regardless, some key themes did emerge from the interview transcripts.

One thing interviewees often appreciated about the baseline and clustering interface was their simplicity and user-friendly design. Some representative quotes for this include: *"It's simple, there's not too many*

[1]For this comparative analysis, we do not consider the results for the student who was enrolled in both a compulsory and non-compulsory muddy card unit, or for the student who did not indicate their unit.

buttons so ... I didn't get confused", *"... it was very easy to navigate."*, and *"... the interface was simple, which is not bad. I think that was a good thing... I mean it was very simple, but it had all I wanted ... to use the system."*

When comparing the baseline interface (variant X) and clustering interface (variant Y), many participants preferred the clustering interface. This is exemplified when the interviewer asked one lecturer, *"Is there anything you would change about the [variant X interface]?"* They responded, *"I mean, I'd make it the Y [clustering] one."* Compared to the baseline interface, the clustering interface saved lecturers' time, which satisfies one of the design intentions of our system. One lecturer said in reference to the clustering interface, *"we're all time poor right? So if there's a technology that will cut out a step for you and save you 10 ... [minutes], then I'm all for it. So I really... did like it."* Another said, *"a unit coordinator is always very pressed for time, and so if you can go there, click once, see it, and then use that to rework what you're doing, then that's kind of the key aspect. ... The just in timeness of [the system] ... is a really important part of its utility, and I think that's captured in what you've done."*

Lecturers identified that issues with the baseline included that they *"would prefer not to get the raw information"* and that *"when you've got a lot of responses... it rapidly just becomes not very good."* Additionally, a few lecturers identified that arranging things alphabetically was not useful, with one explanation being that the different ways of writing a topic cannot be captured alphabetically. In saying this, in one of the units where there were many responses, the lecturer did use the feature to sort responses alphabetically when analysing the muddy cards, but they still preferred the clustering interface.

Regarding the optional controls in the baseline interface, two lecturers identified that the ability to download the student's muddy card responses is useful as it allows them to do clustering on their own. In the absence of the clustering interface, one lecturer explained that they would download the responses and *"put them into ChatGPT"*, for ChatGPT to cluster the responses for them. Another interviewee said that having the option to download responses meant that they could manually cluster the responses themselves; *"Downloading it is useful, because ... it means that I could probably start grouping them myself. Which is what I've done historically, when I have similar sort of data."* Hence, these instructors preferred the clustering interface.

Regarding muddy cards as an active learning technique, one lecturer explained that even with the baseline interface, they appreciated seeing what the muddy card responses were, and believed that all responses had value, not just the most common ones. They said, *"And I can see that even when students*

... had different points that they wanted to raise, ... all of those had value. So it wasn't necessarily the most common muddiest point that was the most important, ... even sometimes when it was an individual student's comment, that held lots of value, ... [and] I can see [that] if one student is saying that then possibly others are also thinking the same thing."

Responses were mixed when considering the clustering algorithm used in the clustering interface. One interviewee thought that the algorithm did a good job. This lecturer originally processed the responses one by one, but once they noticed the algorithm caught all the responses, they stopped reading them all. Another interviewee thought the clustering quality was good, but noticed some issues, *"And I was like, 'Yeah, this is not exactly what it should [be] like. The cluster should not be this cluster. It should be something slightly different.' But in general, I think these mistakes happen maybe 20% of the time. But yeah, it was quite good."* Another lecturer found it frustrating that the automatic clusters were not aligned with what they deemed to be correct. They said, *"I could see that there were [responses] ... that were either ... grouped inappropriately, or that I would have grouped differently. And in playing with ... [the system], it was a little bit frustrating because I couldn't get it to change the way it was grouping things."* In a future version of the interface, they would appreciate the ability to manually adjust the clusters, but they also claim that if the AI clusters are good enough, then the ability to manually adjust the clusters is not required; *"If ..[the clustering algorithm is] doing a really good job, then [manually adjusting the clusters] becomes less of an issue."* To this lecturer's point, we do agree that it is vital that the clustering is highly accurate, which is the motivation behind Research Question 1 in this thesis.

Regarding the slider to control the number of clusters, one lecturer gave it a positive reception, with another identifying that it was nice to have, but not essential. The latter lecturer said about the slider, *"It's kind of nice, but maybe not essential. I think it would still give, you know, if you had a hard-coded [value] and that it can only ever go to 10 classes, I don't think it would necessarily break the usefulness of the model."* This was reiterated by another lecturer who said that the slider is *"something that you play with rather than an important ... key aspect of the design."* We do speculate that the perceived usefulness of the slider could be influenced by the low response rates, necessitating the need for a follow-up study. One interviewee identified that it was hard to find the "sweet spot" with the slider, *"I couldn't find like the sweet spot, like the exact number of clusters. I guess it depends on the amount of answers."* Not being able to find the "sweet spot" is likely linked to the fact that this system was intended for when there are many responses, and we suspect mistakes are more noticeable when there are fewer responses.

The interviews did not cover the other optional controls in great detail. One reason for this is that many lecturers did not explore these features when using the system: *"I didn't really look into many of the other optional controls that were in there."* Another lecturer mentioned that they did not even realise that there were optional controls until a few weeks later: *"Maybe it was like a few weeks until I realised there was an optional controls."* This lecturer did think that *"the sliding rule [to adjust the number of clusters] was enough"*. Additionally, no interviewee recorded using the optional control to filter the responses by student-identified response intention. Just because people did not make great use of the optional controls does not mean they are redundant and should be removed from future iterations of the system. As we saw in the boxplot of Figure 5.1, most units had a very low number of responses collected. Hence, it is unclear whether these controls would be helpful in cases where hundreds of responses were collected for a single lecture. Additionally, we speculate that because of the low student response rate, there was less need for teachers to explore the optional controls in the interface.

Regarding the low student response rate, a few lecturers mentioned that the interface is not useful when there are so few responses. One interviewee said, *"if there are 3 comments.... I don't need AI to summarise that for me"*. Another interviewee said, *"when it came to practice with 12 questions, ... it was still not very effective."*

The teacher's overall thoughts on the effectiveness of the system are also revealed in the final teacher survey, which only considers the clustering interface. As seen in Figure 5.2, when asked to indicate their agreement with the statement, 'I can understand and act on the information provided by this software', most teachers agreed to some extent ($\mu = 5.21$, $\sigma = 1.19$). Also shown in Figure 5.2, when asked to rate the system on a scale from terrible to wonderful, the responses leaned towards finding the system wonderful ($\mu = 5.00$, $\sigma = 0.88$). Additionally, when rating the system from difficult to easy, and the system's mental demand from very low to very high, most people found the system easy ($\mu = 5.77$, $\sigma = 1.01$) and the mental demand low ($\mu = 2.50$, $\sigma = 1.02$). This reinforces that the system was easy to use. However, when rating the system from frustrating to satisfying, dull to stimulating, and rigid to flexible, the responses for each were more neutral, yet still leaning towards satisfying ($\mu = 4.71$, $\sigma = 1.14$), stimulating ($\mu = 4.71$, $\sigma = 0.99$) and flexible ($\mu = 4.64$, $\sigma = 1.15$).

5.3 Closing the Loop Through Instantaneous Feedback

As seen in Section 5.1.1, a large issue discovered through the user study was the low response rate from students. One way to address the low response rate would be to adopt the approach taken by the NLP

course, attributing a small portion (around 5%) of a unit's grade to muddy card completion. While two instructors supported making muddy cards compulsory, nine were opposed. There were many different reasons provided for why they should not be made mandatory, which are provided in Table 5.1.

Reason	Representative Quote
Compulsory muddy cards might lead to lower-quality responses.	*"...if you make it compulsory, of course, then you get the situation that students ... have to do it. So they make up something half-baked and hand it in."*
Does not want to waste time on muddy points from students who are not confused by something.	*"[It] makes it harder to identify actual common themes... [if] they feel compelled to put in something that they're not really confused about ... So then you end up potentially wasting time on things that actually aren't confusing."*
No point of a muddy card for students who have no confusion.	*"I don't think it really warrants it, because if a student has done really well, there's not really any point for them to do a muddy card, so ... [there] shouldn't be a mark incentive for it."*
Students don't like compulsory activities.	*"They don't like compulsory stuff. It is very, very hard to deal with them, and adding an extra thing, it's just more headache for the unit coordinator."*
Generally against compulsory activities. At university, students should have more autonomy.	*"I don't like making things compulsory... I'm like, 'oh it's university let people decide'."*
Dealing with special considerations if a student has misadventure.	*"I don't think so. Like I said, ... it is very tricky, because that means that they can apply for a special consideration for these kind of things. And you have to handle everything. So the less thing[s] ... the better for me."*
Against philosophy on grading.	*"Probably not because I think it goes against policy at the moment. One should not be just rewarding behaviour purely for the sake of it."*

TABLE 5.1: Interviewee identified reasons why muddy cards should not be made compulsory.

Interviewees suggested that the response rate could be improved if students could see the benefit of muddy cards more clearly. One interviewee explains this in greater detail: *"I think sometimes students need to see value in answering surveys or participating in activities ... [and] we need to provide some value back to the students. So sometimes, whenever we try to encourage students to do additional things on top of what they see as ... their responsibilities to learn the material, they don't uptake that strongly unless we allocate some marks towards it..., or we demonstrate clearly how it provides value for them and not just for us, as educators."* The idea of students seeing the benefit of participating was emphasised in another interview where the lecturer said, *"[Students] have to get the reward for [filling out muddy cards]. So the following week, I have to come back and say, 'You know, I went through your*

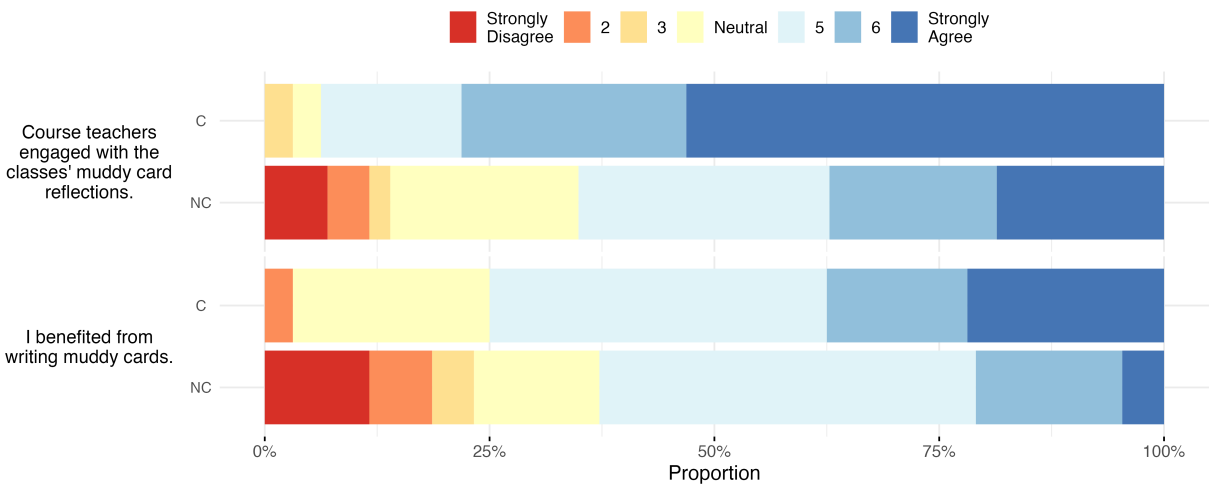


FIGURE 5.6: Results for the questions about course teacher engagement with muddy cards and whether students benefited from writing muddy cards. Results are separated by whether students were enrolled in units where muddy cards were compulsory or not. 'C' denotes compulsory, and 'NC' denotes non-compulsory.

responses. Here's what we found, and you know this is what you're misunderstanding, or this is the concept you're struggling with. And here's a[n] exercise to help you with that and that kind of thing' ... I could see how that would set up as a dynamic situation." Helping students see the benefit of muddy cards is akin to persuading them that muddy cards are beneficial to their learning. Indeed, prior work by Cavanagh et al. (2016) found that persuading students of the benefits of active learning techniques significantly impacted students' identification that active learning techniques are good for them. This, in turn, significantly impacted students' commitment to using active learning techniques.

One way to gauge whether students perceived receiving value for completing their muddy cards is by considering the student survey question, which asked students to rate the statement 'Course teachers engaged with the classes' muddy card reflections' from strongly disagree to strongly agree. Unlike Figure 5.4, which shows the results of this question amongst all survey responses, Figure 5.6 shows the results separated by whether the unit was for credit or not. When separated, we see that amongst the students in the compulsory muddy card course, there was much higher agreement that the course teachers engaged with the muddy cards. Simultaneously, these students claim to have benefited more from writing muddy cards than the students in non-compulsory units. Indeed, the lecturer for the NLP course where muddy cards are compulsory would regularly post explanations of the most common muddy card responses. We suspect that these students may perceive that they benefited from writing muddy cards, as they see the value in the course teacher engaging with the responses. However, it is also possible that these students claim to have benefited because the compulsory nature allows them to earn

easy course credit. Again, future studies should consider student surveys to gain a better understanding of student perspectives on our muddy card system and muddy cards in general.

Through the lecture-by-lecture survey and interviews, lecturers mentioned different ways to address muddy card responses. These are presented in Table 5.2. Whilst the responses in Table 5.2 cover a diverse range of points, many are not directly visible to the student. For example, while a lecturer may use muddy cards to reflect on their own teaching, this is not directly visible to students, and hence, lecturers are not demonstrating how filling out muddy cards provides value to the students. For example, one interviewee believes that students stopped using the system as they could not see the direct feedback from the lecturer, *"I think probably because students didn't see the direct feedback from me, they didn't continue to use it."*

Hence, to encourage student engagement, many lecturers suggested an instantaneous feedback and instant gratification model for muddy cards. By instantaneous feedback, we mean a system that runs live in the lecture. The current system was designed with the intention that a lecturer would analyse the responses after the lecture. An instantaneous feedback model is where the responses are analysed live so that students can receive live (instant) feedback. This leads to instant gratification, as students are instantly satisfied that their muddy card response has been addressed. They do not have to wait for the course instructor to address it via other means after the lecture.

In support of this model, one interviewee said, *"[students] want more in-person connection, in-person engagement, but also at a sort of on-demand approach. If you ask them to simply engage with this tool, but it... isn't instant gratification... whereby... they ask something and it comes back straight away, I think at large, they just stop... engaging."* They later suggest that regarding the muddy card analysis, *"the most benefit will come in when we're doing the lecture at that moment in time, 'okay let's let's review our muddy card system', ... and then use that as a diagnosis."* Another interviewee says, *"an ideal situation would be to get feedback during the class and have it summarised right there for me, and then be able to act on that towards the end. I really like that idea."* They then likened an instantaneous feedback model to Mentimeter, and how an advantage of Mentimeter is that it is *"immediately there on the screen."* By engaging with the muddy cards live in class, students will directly see the benefits of filling them out, which might boost the response rate.

One interviewee explained that not only could this method increase engagement, but it could also encourage students who are too shy to post their own responses. They said, *"I think that it would make*

Reason	Representative Quote/s
Reflecting on their own teaching.	*"The way I used it more for the couple classes where I had more feedback this semester was to look at it and review... the list of comments people made, and thinking about how to change the lecture the following year and adding notes to my list that way. So I'm acting on the information, but not in real time."* AND *"I hadn't considered this lecture to be 'content heavy', but I should not have held this assumption since I can now see that many students are learning this information for the very first time."*
Responding to the muddiest points on the class discussion forum.	*"Because we have such a strong emphasis on the discussion board, it allowed me to take the responses... and then I could bring that in as a, here are a series of points that were unclear, and then I could put in static explanations or examples or link to resources that then complement what's been covered in the lecture."* AND *"I will make a pinned post on Ed with all the answers."*
Revise content in subsequent lectures.	*"So if I see... [they] repeat informations like the similar content in the muddy card, I will address it like properly at the beginning of next week's lecture, saying, like, 'I have seen these concerns in the muddy card.' And that is how I address [them,] by... making a summary at the beginning of the lecture for [the] previous week's content."*
Prepare for workshop questions.	*"Once I log into the teachers... end, I will be able to see the questions they raise, and it actually helped me to prepare for the pre-workshop and the workshop questions in the following week..."*
Add additional practice questions.	*"I put additional questions if they want ... more practical questions on certain topics."*
Understanding which content is not useful to be teaching.	*"I can see the system potentially providing some good feedback on parts that are not useful. ... There is some content that I want to remove for next year, I don't think it adds, and where that's come up is in the number of students coming to office hours asking questions about it. And so that is influencing what I teach. And if there was a way to get that information in more real-time that would be helpful."*
Live clarifications during the lecture.	*"I asked them to fill out the muddy card, and then I used to check just what are the points that they felt like they needed more clarification. And I tried to go through that again during the lecture."*
Get tutors (teaching assistants) to answer the question in the labs.	*"For this coming week, I might raise some of these questions with my tutors to answer in the labs and then also address it in the lecture."*

TABLE 5.2: Identified ways to respond to muddy cards raised during the lecture-by-lecture survey and interviews.

people feel more comfortable posting their own thing, because I think there's often a real hesitation around putting yourself forward in the same way that if you ask [a] class [a question], not many people put their hands up. I think there's a real hesitation, and often that's a fear of being perceived as dumb." This is confirmed by Yep et al. (2023), who found that a reason students would not participate in class discussions is out of fear of being judged by their peers. These authors used a system that allowed students to participate in class discussions with partial anonymity (voice only), which reduced students' anxiety and slightly increased the class participation rate. Another interviewee raised the notion of our muddy card system encouraging shy students. They explained that the system is helpful in earlier weeks of the semester, as students *"need to get the courage to ask questions through the lecture, and that's something that happens mid-semester."* This lecturer also reported going through muddy cards live in the lecture, so they had a trial run at implementing something akin to an instantaneous feedback approach to muddy cards.

Finally, an instantaneous feedback model would potentially increase engagement as students would not have to wait so long for their muddy card to be addressed. One interviewee said, *"I think that the main point here is they could get faster responses using another forum. And usually when the students are studying the lecture, they get a question, they ask the question, and they are waiting for the answer. Whereas with a muddy card I would wait, you know, like, until the end of the week to make a post or the following week to discuss the answers during the live lecture. So maybe that was a long wait time."* By addressing the muddy card responses within the lecture, students get immediate feedback, and it also ensures that the feedback loop is closed.

We believe that an instantaneous feedback approach to muddy cards will likely increase the student response rate. However, we speculate that the response rate will still be relatively low. The supervisor of this thesis has incorporated an optional Mentimeter (without the AI features discussed in Section 2.7.4) into their lectures, allowing students to ask questions live; however, engagement with this tool is relatively low. This is supported by an interviewee who said *"we have Mentimeters and things like that, but again, the engagement is sometimes limited."* To test these theories, and with the findings in this section in mind, we aim to do a follow-up study by revising the muddy card system to prioritise instantaneous feedback. While the details of this have not been finalised, we plan to implement a 'presentation mode', where the top muddy card clusters are provided to students. The presentation interface will be much simpler, emphasising the largest muddy card clusters as the basis of class discussions. In a way, this will share many similarities to the Mentimeter AI grouping mode presented in Section 2.7.4, however,

instructors will still retain control over the analysis through features such as the slider to adjust cluster granularity.

5.4 Using the Muddy Card System in the Future

When asked the survey question, 'I would like to use the system in other courses', as seen in Figure 5.4, most students agreed or had neutral support for the statement ($\mu = 4.88$, $\sigma = 1.43$). CourseMIRROR used the same question in their paper on a 5-point Likert scale and had similar proportions of students agreeing (Fan et al., 2017). In this study, like the students, the teachers were asked in their final survey to rate the statement 'I would like to use this system in other courses I teach.' For this question, one teacher disagreed with the statement, with the remainder of the responses rating the statement as neutral or agreeing to some extent ($\mu = 4.57$, $\sigma = 1.22$). Overall, there was stronger support from the students, with the teacher results leaning closer towards neutrality.

When asked in the interviews whether they would want to implement the muddy card system into future courses they teach, responses were varied. Two lecturers said that they would use the baseline interface in subsequent semesters. One interviewee said that this was because the final unit of study survey does not have the ability to sort responses; *"Yes, I will, because the reason for that is when we look at the unit [of] study survey at the end of each semester, they don't have the function for sorting..."*. Another interviewee preferred the clustering interface, but said they would still use the baseline version: *"Yeah, look. I would use it anyway. As I said, ...even the alphabetical summary... I found useful."* One interviewee said they would not like to use the baseline again because of its time commitment, echoing what was discussed earlier. They said, *"it is too much effort in putting everything together and reading every single thing."*

Five participants said they would specifically use the clustering interface again. One interviewee said, *"I'd like to stick with it... I'll probably reintroduce it when we come back after the break"*, and another said they found the interface *"interesting"*. One lecturer noted that they don't think the interface is only for large courses, and said, *"it could be also very interesting for small courses or courses, where you have more attendance."* When answering this question, one interviewee stressed the importance of being able to see all student responses in the clustering interface and said they would be less inclined to use the interface if this feature were removed.

One interviewee said they would not use the system again as they perceive that the current interface creates too much confusion; *"At the current standing I think it just creates more confusion."* This interviewee preferred a more seamless system and would have preferred the results be delivered live to

the students, a theme already discussed in the previous section. They also would prefer a different way of viewing the data, where muddy cards are presented using a network-style visualisation. The notion of a more seamless system was echoed by another interviewee, who said they would use the system again if it *"was well integrated with Canvas"* (the course management system used by the university) and replaced one of the active learning techniques already used in their unit.

Two lecturers explained that their future usage would be dependent on the class discussion forum. One interviewee explained that they would not use it if the EdStem class discussion forum were activated. Coming from a different angle, another interviewee said that they would use the muddy card system if there were more responses on EdStem. They explained that in some first-year courses, *"there could be like 20 questions a day"* and hence, *"it could be very time consuming to go one by one and answer questions."* They could see the muddy card system *"helping a lot"* under these circumstances.

Another interviewee said that whether they used the muddy card system would largely come down to how it fits with their teaching model, and *"balancing with the other things that [they] already do."* They said that they would be slightly more inclined to use the clustering interface than the baseline, but this would largely depend on whether students are engaging with the system: *"I would maybe be slightly more inclined to use ... [the clustering interface] than the [baseline variant] X one, because I think this is a nicer system particularly if it's used wholesale, if everyone's engaging with it this would make it a lot more administrationally nicer."*

5.5 Interview Saturation

As mentioned in Section 2.6.2, an interviewer knows that they have sampled a sufficiently large number of people when saturation has been reached. From the memos that were written after manually coding each interview, the researcher recorded that after analysing the tenth interview, they felt saturation was close. This was because while people were raising new points, these points explored the subtleties of the already formed themes, and the researcher did not feel they were hearing anything completely new.

However, on analysing the eleventh interview, the researcher's view changed. In this interview, the researcher heard from a lecturer who had many responses. Previously, most of the other interviews were conducted on people with very low student engagement (the majority of units in this study). The researcher recorded in their memo: *"While we may have reached saturation amongst the people that had low response rates, I'm not so confident that we have reached saturation amongst people that had many*

responses." This view did not change by the time all thirteen interviews were concluded, and hence, we believe saturation had not been reached.

Hence, comparing the results of our system with CourseMIRROR (Fan et al., 2017) and Mudslide (Glassman et al., 2015) is difficult, especially considering the different way that sampling was used in each. For example, while CourseMIRROR received a similar number of muddy card responses in total over its in-the-wild deployment, this was distributed over eight courses, meaning they do not struggle with low-response rates. Hence, we have plans for a follow-up user study to collect additional data. For this user study, we will update our muddy card system to encompass the instantaneous feedback features described in Section 5.3, as well as trialling the system at universities outside of Australia.

5.6 Conclusion

In Chapter 5, we have discussed the user study results to explore Research Question 2. One of the key findings was that when made optional, there was a very low student uptake of muddy cards. Despite the low response rate in many units, our system received a positive reception among the units that received a large number of responses. In response to the low response rates, interviewees provided suggestions on how this could be alleviated in future iterations of the muddy card system. The main suggestion was a muddy card system that prioritises instantaneous feedback, where student responses are analysed and addressed live in the lecture. They believe this will improve response rates as by addressing muddy card responses live, students will see the benefits of completing a muddy card more clearly.

When asked about our current implementation, lecturers appreciated the simplicity of both the baseline and clustering interfaces. Many lecturers preferred the clustering interface, as unlike the baseline, the clustering interface saved lecturers' time and prevented them from being inundated with the raw muddy card responses. However, some lecturers felt that the clustering interface was not that useful with limited responses, as there is no need for clustering.

Despite the low response rate, many students wished to use the muddy card system in other courses. When asked in the final survey whether they would like to use the system in other courses they teach, lecturer support was more neutral, with only one lecturer indicating that they would not like to use it in the future. In the systems current form, the low response rate raises questions about the usefulness and effectiveness of the system. Regardless, the results from this user study have opened up some interesting avenues for future research, which we will discuss in the next chapter.

CHAPTER 6

Limitations & Future Work

This study's main limitation revolves around the user study's low muddy card response rate. Due to the low response rate, it was challenging to assess the system's effectiveness in understanding students' muddy card points, as teachers typically lacked access to sufficient student data. Further, when considering the embedding model benchmarking and testing the student-assisted algorithms, there was a lack of non-NLP data (only one of the eight muddy card samples was a non-NLP unit). Hence, to address these limitations, we have plans for future studies.

6.1 Follow-Up User Study

This thesis only explored units from the University of Sydney, an Australian-based institution. From the author's and supervisors' experience studying in the United States of America, we believe there could be societal differences in how students approach completing optional activities. Per the 2025 user study interviews, lecturers explained that students typically dislike completing optional activities, which is linked to the low response rate. To investigate whether this observation is consistent with that of international universities, we plan to rerun this study at universities abroad. We plan to capture universities from at least two different countries, and different types of institutions (such as private vs public universities).

In addition, we plan to incorporate the lecturer's suggestions from the 2025 user study interviews to modify the existing muddy card system into a mode aligned with instantaneous lecture feedback. As mentioned during the analysis, lecturers believed that the student response rate would improve if students saw the benefits of using muddy cards more clearly, and suggested that addressing muddy card responses in real-time during the lecture is one way of making this more transparent for students. While we have not finalised a revised interface, we plan to implement a 'presentation mode' where the top muddy card clusters can be presented to students live in the lecture. The presentation interface will be even simpler

than the existing system, offering only a few rudimentary controls (we plan to remove many of the optional controls, but will retain the slider to adjust the clusters' granularity).

An experimental design will be selected such that lecturers of future units trialling the system will compare the system described in this thesis and the new instantaneous feedback version. Additionally, to make it easier for students to submit their muddy card reflections, we will remove the initial inputs to select the lecture week and unit of study the muddy card is for. The system will know which lecture week it is based on the timestamp of when the student submits their response, and the unit of study will be automatically known through unit-specific interface links. The rest of the student interface will remain largely untouched, with students still selecting peer responses as part of the student-assisted approach.

If a future user study results in a higher response rate, an added benefit is that there will be more data for benchmarking the internal algorithms used in the muddy card system.

6.2 Embedding Model Benchmarking and Testing the Student-Assisted Approach

The low response rate in most units of study meant that the current analysis of muddy card data samples was heavily weighted towards the NLP course. With only one non-NLP unit, it is unclear whether the results found would generalise to units from different academic domains. Additionally, there was only student-assisted data for the NLP units (no other unit received over 200 responses for a single lecture), which again raises questions about the generalisability of our findings. This is especially significant when considering that 93.5% of students in the NLP course are non-native English speakers, with 94.2% of the cohort being classified as international students. An article by White (2024) states that 46% of enrolled students at the University of Sydney in 2023 were international students. Hence, the NLP unit has a much larger proportion than that of typical University of Sydney courses. We believe that non-native speakers may find it more challenging to determine the semantic similarity of peer responses.

To address these limitations, a secondary goal of our follow-up user study is to collect additional muddy card data samples. We can then investigate whether our findings for the NLP-specific samples are consistent with those from other units of study. Additionally, while the analysis in this thesis revealed that the 'multi-evidence' approach of incorporating student-assisted data leads to higher clustering quality

in terms of SQAS score, ARI suggest that this approach is not beneficial (AMI had mixed results). Additional data samples will enable us to explore this phenomena further and confirm whether the 'multi-evidence' approach is truly beneficial.

As new embedding models are often released, future research would also involve testing the new state-of-the-art embedding models for the muddy card clustering application. When writing this thesis, Google released `gemini-embedding-exp-03-07`, which led to state-of-the-art performance (Kilpatrick et al., 2025). We were unable to test this embedding model in this thesis, as we received explicit permission from the ethics office only to use the proprietary embeddings of OpenAI and Voyage AI.

Overall, despite the low response rate, we believe that the findings from this thesis offer valuable insight into technical approaches to muddy cards, and the findings provide clear directions for future research.

CHAPTER 7

Conclusion

This thesis describes a muddy card system that saves teachers time when using the muddy card active learning technique in their classrooms. To achieve this, we developed an online system with a student interface that allows students to enter their muddy card responses digitally. The student interface also included the 'student-assisted approach', where students would identify whether system-selected peer responses were semantically similar to the original muddy card response they had entered. The muddy card system also contained a teacher interface, which used clustering techniques to quickly summarise the largest recurring student muddy card responses.

A main contribution of this thesis involved investigating the performance of different embedding models at clustering eight manually labelled muddy card data samples, which together, totalled 2,327 muddy card responses. This involved proposing a new metric called the 'student questions answered satisfaction' (SQAS) score, which was designed to practically capture clustering quality for muddy card applications. It was found that regardless of the embedding model used, SQAS scores were generally consistent. However, SQAS scores revealed considerable room for improvement among embedding models, as the SQAS scores were considerably far from the gold-standard values.

A further contribution is that we intend to release six of the manually labelled data samples as a clustering benchmark. This is a valuable contribution, as we demonstrated that existing clustering benchmarks are relatively coarse-grained, with few labels and many examples. On the other hand, muddy card clustering is relatively fine-grained, with many labels and fewer examples.

Another contribution was the 'student-assisted approach'. This is an entirely novel concept in muddy card research. While further research is still needed, results indicate that the 'multi-evidence' approach for incorporating student-assisted data in clustering consistently improves clustering performance on the SQAS metric.

The final contribution of this thesis was a user study conducted among 20 classes at the University of Sydney, which spanned six different faculties, as well as undergraduate and postgraduate courses. Over 11 weeks, 3,697 student muddy card responses were collected. Most students agreed or had neutral support regarding using the muddy card system in other courses. When the lecturers were asked whether they would like to use the system in other courses they teach, results leaned closer towards neutrality.

While lecturers tended to prefer the clustering interface over our baseline control interface, analysis was hindered by a low student response rate. Teachers shared many different reasons for why they believed there was such a low response rate. Many teachers suggested that the response rate might be improved under an instantaneous approach to muddy cards, where muddy card analysis occurs live in class. This is because, by analysing the muddy card responses live, students will see that teachers are actively engaging with their responses, which will help them to see the benefit in completing muddy cards. These findings led to our proposal for a future study that redesigns the existing system to prioritise instantaneous feedback to students. To investigate the effect of societal differences on the uptake of muddy cards, the follow-up study will also include international universities.

Bibliography

Paul Adams. 2004. Assessment as Learning: The Role of Minor Assignment in Teaching and Learning. *Advances in Social Work*, 5(1):47–60. Number: 1. [Pages: 1, 8, 9, and 24]

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. ArXiv:2309.07445 [cs]. [Page: 19]

arXiv.org submitters. 2024. arxiv dataset. [Page: 19]

Chandra Prakash Bathula. 2023. Everything to Know about Hierarchical Clustering; Agglomerative & Divisive Hierarchical Clustering. [Pages: ix and 14]

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. ArXiv:2502.17424 [cs]. [Page: 63]

Charles C. Bonwell and James A. Eison. 1991. Active Learning; Creating Excitement in the Classroom. ASHE-ERIC Higher Education Report No. 1. Technical report, School of Education, George Washington Univ, Washington, D.C. [Page: 7]

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, *Advances in Knowledge Discovery and Data Mining*, pages 160–172. Springer, Berlin, Heidelberg. [Page: 15]

Adam Carberry, Stephen Krause, Casey Ankeny, and Cynthia Waters. 2013. "Unmuddying" course content using muddiest point reflections. pages 937–942. ISSN: 2377-634X. [Pages: 8 and 23]

Andrew J. Cavanagh, Oriana R. Aragón, Xinnian Chen, Brian Couch, Mary Durham, Aiyana Bobrownicki, David I. Hanauer, and Mark J. Graham. 2016. Student Buy-In to Active Learning in a College Science Course. *CBE Life Sciences Education*, 15(4):ar76. [Page: 82]

Davide Chicco. 2021. Siamese Neural Networks: An Overview. In Hugh Cartwright, editor, *Artificial Neural Networks*, pages 73–94. Springer US, New York, NY. [Page: 11]

John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an instrument measuring user satisfaction of the human-computer interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '88, pages 213–218. Association for Computing Machinery, New York, NY, USA. [Page: 51]

S. G. Y. Choong, X. N. Tan, T. L. Scott, S. Y. Ong, and Wei Wei Goh. 2024. Mud card class feedback system: A digital approach. *AIP Conference Proceedings*, 2729(1):070002. [Page: 30]

Ylona Chun Tie, Melanie Birks, and Karen Francis. 2019. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7:2050312118822927. Publisher: SAGE Publications Ltd. [Pages: 21 and 23]

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. [Page: 11]

Juliet Corbin and Anselm Strauss. 2008. *Basics of qualitative research: Techniques and procedures for developing grounded theory, 3rd ed*. Basics of qualitative research: Techniques and procedures for developing grounded theory, 3rd ed. Sage Publications, Inc, Thousand Oaks, CA, US. Pages: xv, 379. [Page: 21]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota. [Page: 10]

K. Edström, D. Soderholm, and M. Knutson Wedel. 2007. Teaching And Learning. In Edward F. Crawley, Johan Malmqvist, Sören Östlund, and Doris R. Brodeur, editors, *Rethinking Engineering Education: The CDIO Approach*, pages 130–151. Springer US, Boston, MA. [Pages: 1 and 8]

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, Portland, Oregon. [Page: 15]

Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. 2017. Scaling Reflection Prompts in Large Classrooms via Mobile Interfaces and Natural Language Processing. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI '17, pages 363–374. Association for Computing Machinery, New York, NY, USA. [Pages: ix, 5, 20, 26, 27, 28, 52, 86, and 88]

Gregor Geigle, Nils Reimers, Andreas Rücklé, and Iryna Gurevych. 2021. TWEAC: Transformer with Extendable QA Agent Classifiers. ArXiv:2104.07081 [cs]. [Page: 19]

Barney G. Glaser and Anselm L. Strauss. 1967. *The discovery of grounded theory: strategies for qualitative research*. Aldine Transaction. [Pages: 21 and 22]

Elena L. Glassman, Juho Kim, Andrés Monroy-Hernández, and Meredith Ringel Morris. 2015. Mudslide: A Spatially Anchored Census of Student Confusion for Online Lecture Videos. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1555–1564. Association for Computing Machinery, New York, NY, USA. [Pages: ix, 5, 25, 26, and 88]

Steven R. Hall, Ian Waitz, Doris R. Brodeur, Diane H. Soderholm, and Reem Nasr. 2002. Adoption of active learning in a lecture-based engineering class. In *32nd Annual Frontiers in Education*, volume 1, pages T2A–T2A. ISSN: 0190-5848. [Pages: 1, 8, 9, 23, and 24]

Paul A. Harris, Robert Taylor, Robert Thielke, Jonathon Payne, Nathaniel Gonzalez, and Jose G. Conde. 2009. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, 42(2):377–381. [Pages: 51 and 52]

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human mental workload*, Advances in psychology, 52, pages 139–183. North-Holland, Oxford, England. [Page: 51]

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2(1):193–218. [Page: 18]

Tomayess Issa and Pedro Isaias. 2022. Usability and Human–Computer Interaction (HCI). In Tomayess Issa and Pedro Isaias, editors, *Sustainable Design: HCI, Usability and Environmental Concerns*, pages 23–40. Springer, London. [Page: 20]

Steven Jungst, Janice Wiersema, and Barbara Licklider. 2003. Providing Support for Faculty Who Wish to Shift to a Learning-Centered Paradigm in Their Higher Education Classrooms. *Journal of the Scholarship of Teaching and Learning*, pages 69–81. [Page: 7]

Daniel Jurafsky and James H. Martin. 2024. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. Third edition. [Page: 9]

Christoph Kessler and Simin Nadjm-Tehrani. 2002. Mid-term course evaluations with muddy cards. In *Proceedings of the 7th annual conference on Innovation and technology in computer science education*, ITiCSE '02, page 233. Association for Computing Machinery, New York, NY, USA. [Pages: 1, 9, and 24]

Logan Kilpatrick, Zach Gleicher, and Parashar Shah. 2025. State-of-the-art text embedding via the Gemini API- Google Developers Blog. [Pages: 16 and 91]

Daniel B. King. 2011. Using Clickers To Identify the Muddiest Points in Large Chemistry Classes. *Journal of Chemical Education*, 88(11):1485–1488. Publisher: American Chemical Society. [Pages: ix, 1, and 24]

Jurek Kirakowski. 1995. SUMI Background Reading. [Page: 51]

Stephen J. Krause, Dale R. Baker, Adam R. Carberry, Milo Koretsky, Bill Jay Brooks, Debra Gilbuena, Cindy Waters, and Casey Jane Ankeny. 2013. Muddiest Point Formative Feedback in Core Materials Classes with YouTube, Blackboard, Class Warm-ups and Word Clouds. In *2013 ASEE Annual Conference & Exposition*, pages 23.916.1–23.916.18. Atlanta, Georgia. ISSN: 2153-5965. [Pages: 9 and 23]

James R. Lewis. 1995. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, 7(1):57–78. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447319509526110. [Page: 51]

Leland McInnes and John Healy. 2017. Accelerated Hierarchical Density Based Clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. ISSN: 2375-9259. [Page: 15]

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training. [Page: 16]

Mentimeter. n.d. Group responses to your Open Ended questions using AI | Mentimeter Help Center. [Pages: ix and 29]

Frederick Mosteller. 1989. The "Muddiest Point in the Lecture" as a Feedback Device. *The Journal of the Harvard-Danforth Center*, 3:10–21. [Pages: 1, 8, 9, and 24]

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. ArXiv:2210.07316 [cs]. [Pages: 12, 16, 17, 18, 19, and 63]

Michael J. Muller and Sandra Kogan. 2012. Grounded Theory Method in Human-Computer Interaction and Computer-Supported Cooperative Work. In *Human Computer Interaction Handbook*. CRC Press, third edition. Num Pages: 21. [Pages: 21, 22, and 23]

Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey Research in HCI. In Judith S. Olson and Wendy A. Kellogg, editors, *Ways of Knowing in HCI*, pages 229–266. Springer, New York, NY. [Pages: 20 and 21]

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and Code Embeddings by Contrastive Pre-Training. ArXiv:2201.10005 [cs]. [Page: 12]

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830. [Pages: 17 and 43]

Tershia Pinder-Grover, Katie R. Green, and Joanna Mirecki Millunchick. 2011. The Efficacy of Screencasts to Address the Diverse Academic Needs of Students in a Large Lecture Course. *Advances in Engineering Education*, 2(3). Publisher: American Society for Engineering Education ERIC Number: EJ1076056. [Page: 23]

Michael Prince. 2004. Does Active Learning Work? A Review of the Research. *Journal of Engineering Education*, 93(3):223–231. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2168-9830.2004.tb00809.x. [Pages: 8 and 72]

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2023. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! ArXiv:2310.03693 [cs]. [Page: 63]

Nile Reimers and O. Espejel. 2021. all-mpnet-base-v2. [Page: 11]

Nils Reimers. 2022. OpenAI GPT-3 Text Embeddings - Really a new state-of-the-art in dense text embeddings? [Page: 12]

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, Hong Kong, China. [Page: 11]

Jason Rennie. n.d. Home Page for 20 Newsgroups Data Set. [Page: 19]

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In Jason Eisner, editor, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420. Association for Computational Linguistics, Prague, Czech Republic. [Page: 17]

Mehrzad Shahinmoghadam and Ali Motamedi. 2024. Benchmarking pre-trained text embedding models in aligning built asset information. ArXiv:2411.12056 [cs]. [Page: 19]

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. *CoRR*, abs/1906.03741. ArXiv: 1906.03741. [Page: 19]

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 16857–16867. Curran Associates Inc., Red Hook, NY, USA. [Page: 10]

Anselm Strauss and Juliet M. Corbin. 1990. *Basics of qualitative research: Grounded theory procedures and techniques*. Basics of qualitative research: Grounded theory procedures and techniques. Sage Publications, Inc, Thousand Oaks, CA, US. Pages: 270. [Pages: 21 and 22]

Yixuan Tang and Yi Yang. 2025. Do We Need Domain-Specific Embedding Models? An Empirical Investigation. ArXiv:2409.18511 [cs]. [Pages: 3, 12, and 20]

Eric K. Tokuda, Cesar H. Comin, and Luciano da F. Costa. 2022. Revisiting agglomerative clustering. *Physica A: Statistical Mechanics and its Applications*, 585:126433. [Page: 14]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv:1706.03762 [cs]. [Page: 10]

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95):2837–2854. [Page: 18]

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355. Association for Computational Linguistics, Brussels, Belgium. [Page: 10]

Martin Wattenberg and Fernanda B. Viégas. 2008. The Word Tree, an Interactive Visual Concordance. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1221–1228. [Page: 26]

Daniella White. 2024. Nearly half of Sydney Uni's students come from overseas. Section: NSW. [Page: 90]

Wikimedia Foundation. n.d. Wikimedia downloads. [Page: 19]

Karen Willcox and Gergana Bounova. 2004. Mathematics In Engineering: Identifying, Enhancing, And Linking The Implicit Mathematics Curriculum. In *2004 Annual Conference*, pages 9.896.1–9.896.13. Salt Lake City, Utah. ISSN: 2153-5965. [Page: 9]

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: generalized autoregressive pretraining for language understanding. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 517, pages 5753–5763. Curran Associates Inc., Red Hook, NY, USA. [Page: 10]

Benjamin L. W. Yep, Teck Kiang Tan, and Fun Man Fung. 2023. How Partial Anonymity May Reduce Students' Anxiety During Remote Active Learning-A Case Study Using Clubhouse. *Journal of Chemical Education*, 100(2):459–468. Publisher: American Chemical Society. [Page: 85]

APPENDIX A

Full Muddy Card System Screen Design

A.1 Student Interface

Appendix A.1 provides all the different screens for the student interface. Some figures are duplicates from the methodology section; however, they were replaced here for continuity and convenience. Press *here* to return to the 'Student Interface Front-End' part of the methodology chapter.

FIGURE A.1.1: Student Interface Screen 1 - Students select their unit and lecture week.

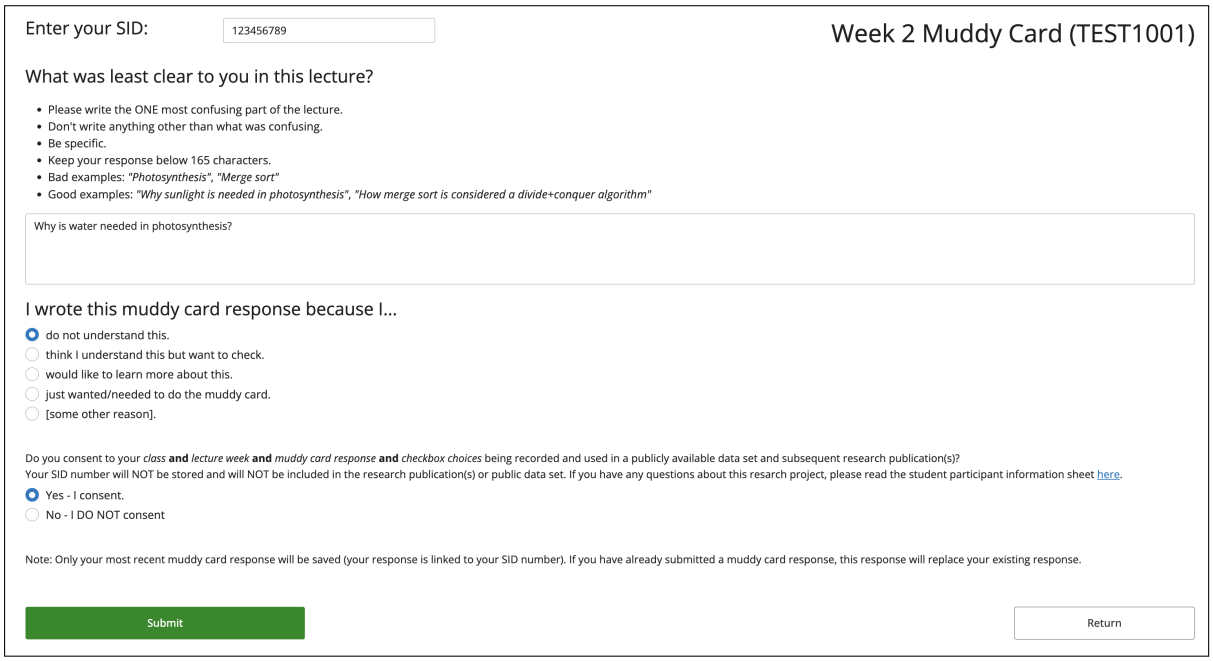


FIGURE A.1.2: Student Interface Screen 2 - Students enter their muddy card response.

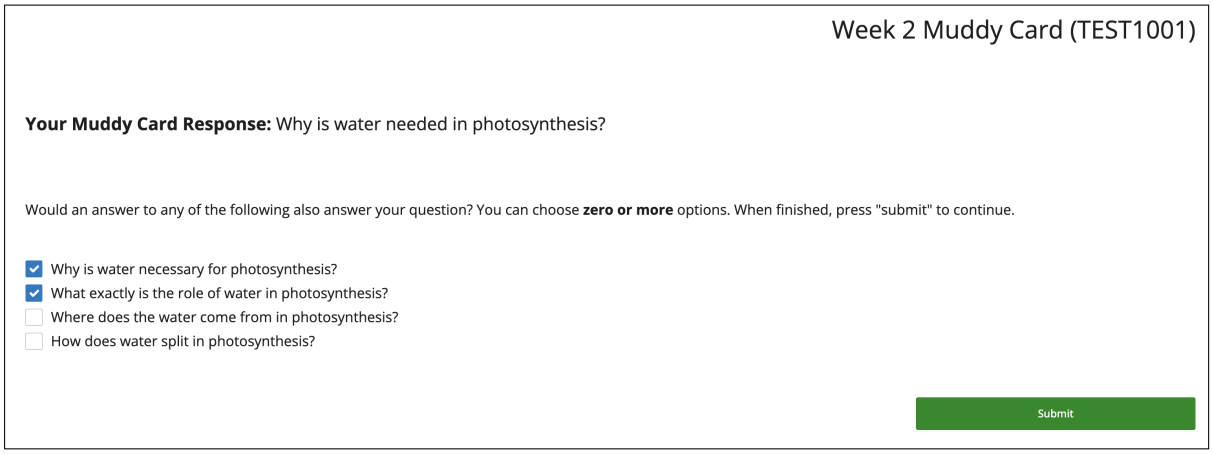


FIGURE A.1.3: Student Interface Screen 3 - The 'student-assisted approach'. Students select whether the system-selected peer responses are semantically the same as what they entered.

FIGURE A.1.4: Student Interface Screen 4 - Input summary page

A.2 Teacher Interface

Appendix A.2 provides all the different screens for the teacher interface. Some figures are duplicates from the methodology section; however, they were replaced here for continuity and convenience. The first few screens are shared between the teacher interface's clustering (variant Y) and baseline variants (variant X). The latter few screens are specific to each variant, after which the users navigate to the optional survey page. Press *here* to return to the 'Teacher Interface Front-End' part of the methodology section.

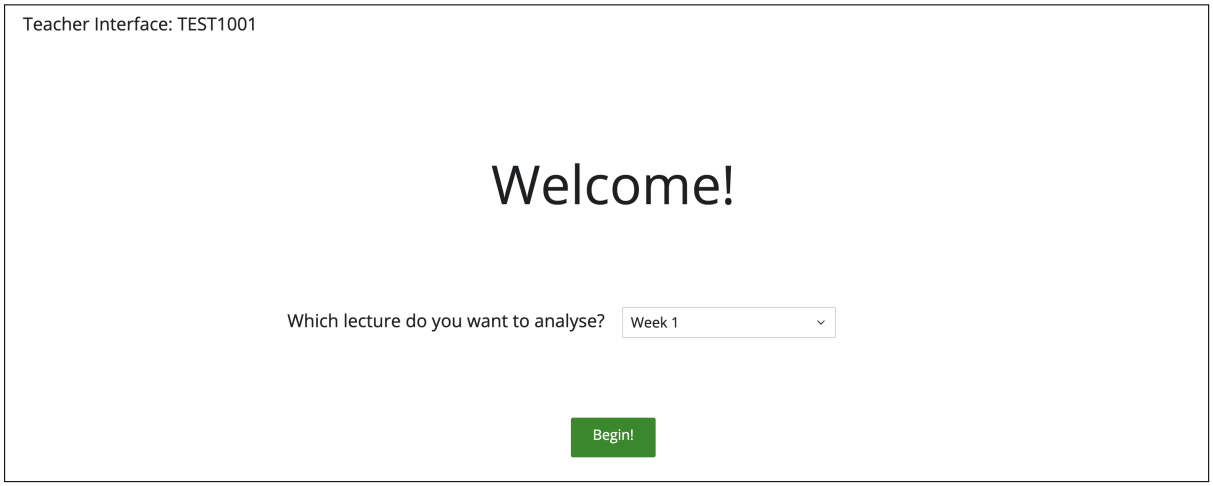


FIGURE A.2.1: Teacher Interface Screen 1 - Unit-specific landing page. Teachers will select the lecture for which they wish to analyse the muddy cards.

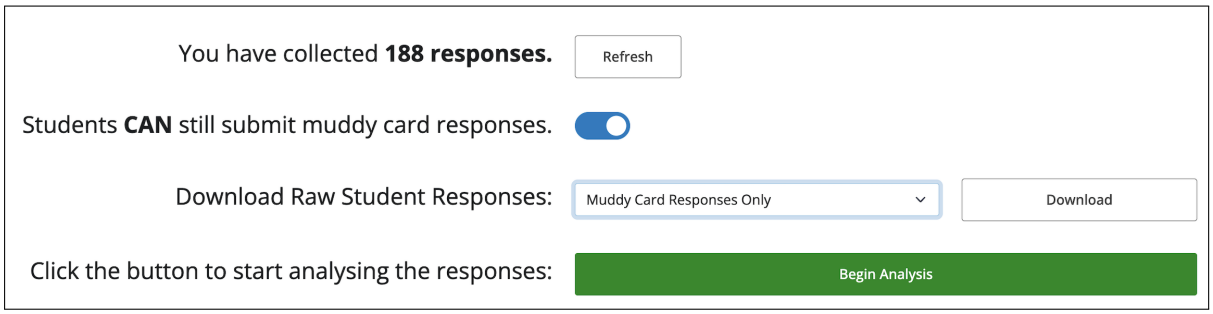


FIGURE A.2.2: Teacher Interface Screen 2 - Page to stop students submitting muddy cards and to download already submitted responses before analysis.

A.2.1 Clustering Variant (Variant Y) Specific

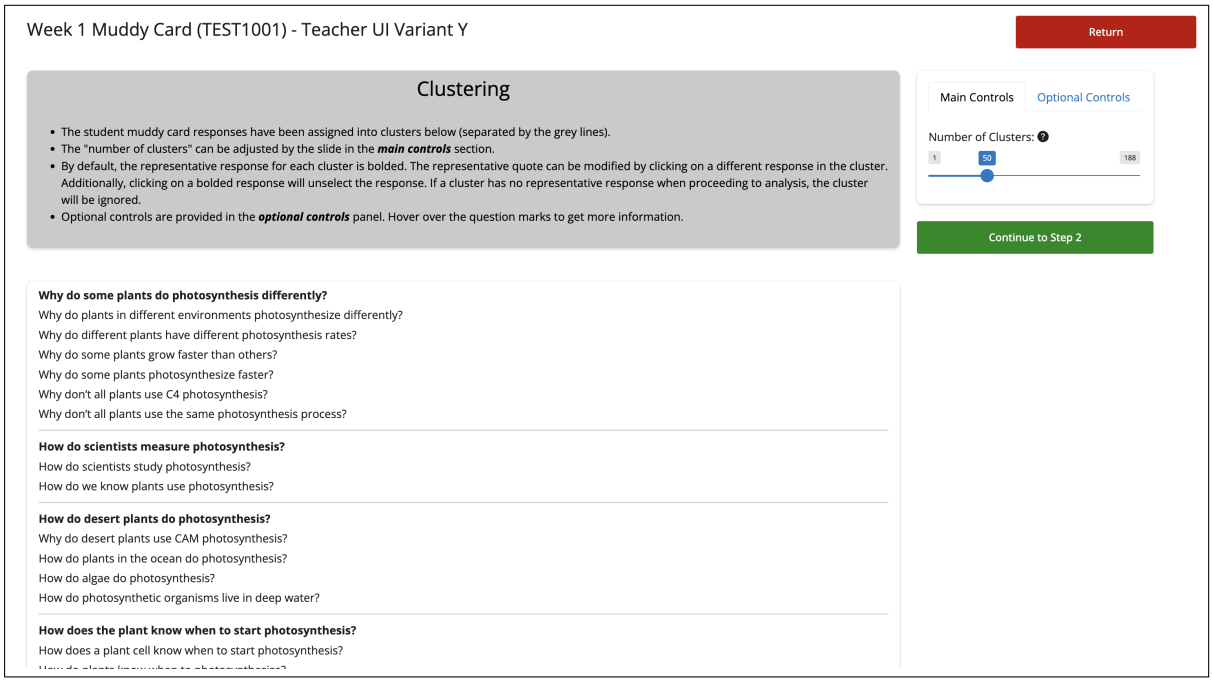


FIGURE A.2.3: Teacher Interface Screen Clustering Variant 1 - Main clustering page where teachers can adjust how many clusters to separate the student responses into.

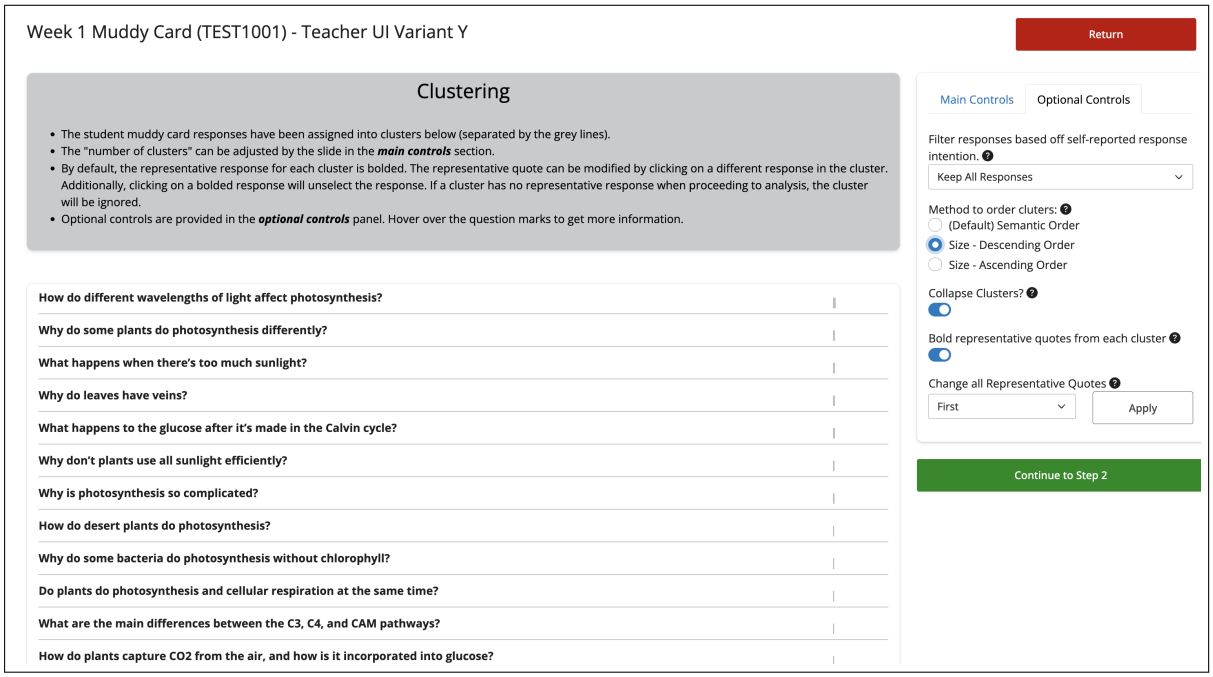


FIGURE A.2.4: Teacher Interface Screen Clustering Variant 2 - View of the optional controls. Here, the option to collapse clusters and to place the clusters in descending size order has been applied.

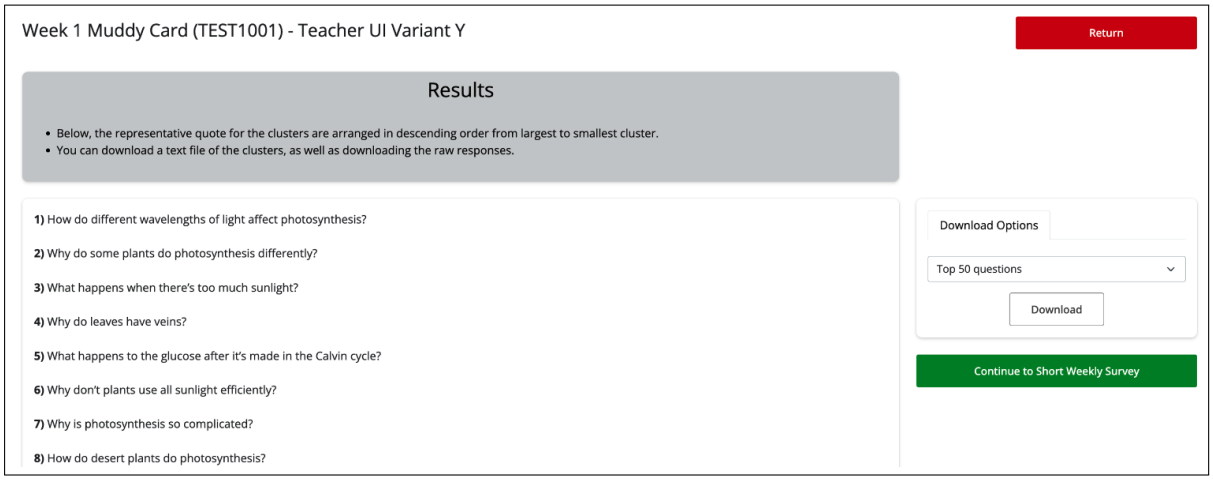


FIGURE A.2.5: Teacher Interface Screen Clustering Variant 3 - Final summary page displaying the representative quotes for each cluster, arranged in descending order of cluster size.

A.2.2 Baseline Variant (Variant X) Specific

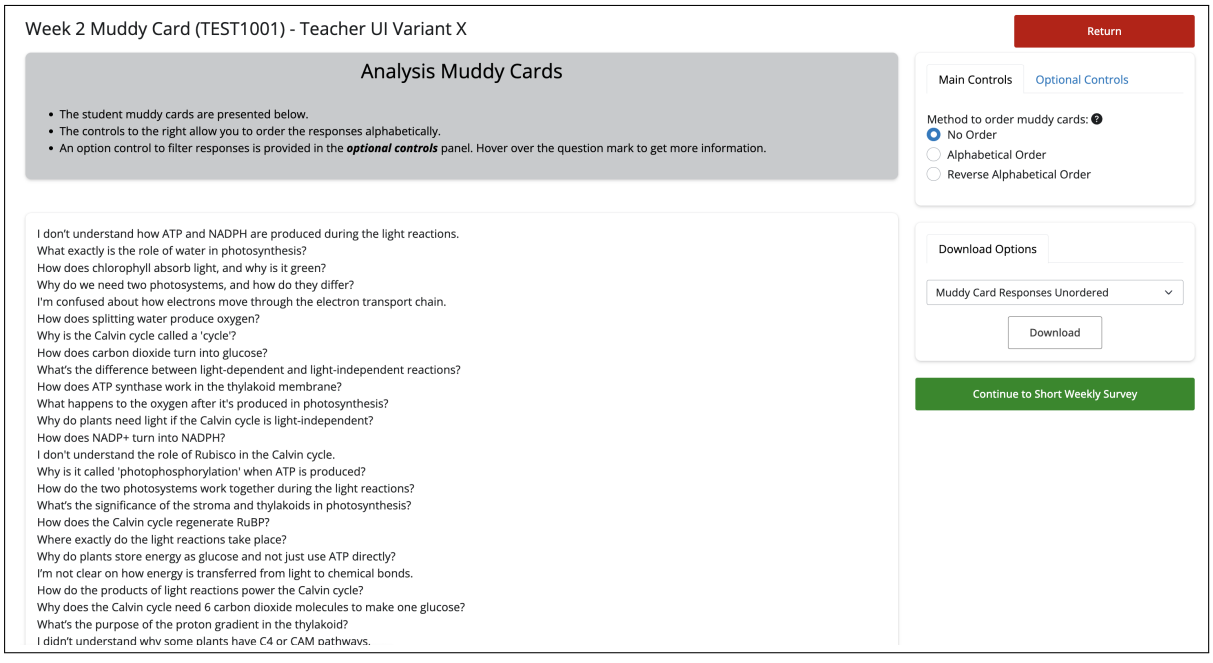


FIGURE A.2.6: Teacher Interface Screen Baseline Variant 1 - Student responses are arranged down the screen and can be ordered. The optional controls (not shown) allow responses to be filtered by student-identified muddy card response intention.

A.2.3 Optional Survey Page

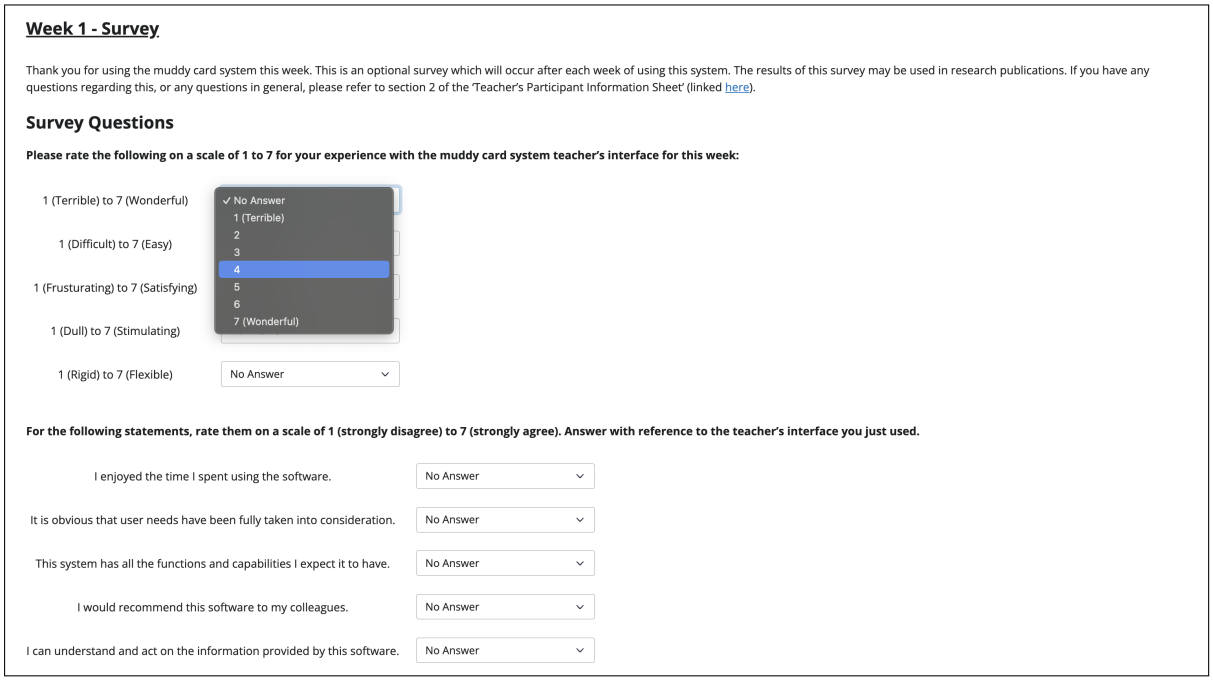


FIGURE A.2.7: Final optional survey for the teachers to share their user experience of the muddy card analysis process.

APPENDIX B

Participant Information Statements

B.1 2024 NLP Muddy Card Analysis Participant Information Statement

Ethics Reference Code: 2024/HE000932

Press *here* to return to the '2024 University NLP Course' section of the methodology chapter.

Muddy Card Artificial Intelligence Analysis

PARTICIPANT INFORMATION STATEMENT

(1) What is this study about?

This research study will be exploring the performance of different sentence embedding models on grouping student muddy card responses into semantically similar clusters. Different locally run sentence embedding frameworks will be used, as well as online sentence embedding large language models (LLMs).

As a student who completed COMP4446 or COMP5046, we are interested in using the muddy card responses you wrote after each lecture. This Participant Information Statement tells you about the research study, helping you decide if you want to take part in the research.

Participation in this research study is voluntary.

By choosing not to opt-out of this study you are telling us that you:

- Understand what you have read.
- Agree to take part in the research study as outlined below.
- Agree to the use of your personal information as described.

You are encouraged to retain a copy of this Participation Information Statement to keep.

(2) Who is running the study?

The study is being carried out by the following researchers:

- Dr Jonathan Kummerfeld, Senior Lecturer (Computer Science), University of Sydney.
- Thomas Elton, Undergraduate Honours Student (BAdvStudies, Computer Science), University of Sydney.

(3) What will the study involve for me?

You are not required to perform any actions. We will access your muddy card responses from the Ed platform and use them in our clustering analysis.

(4) How much of my time will the study take?

Your time will not be required for the study.

(5) Who can take part in the study?

Anyone who completed COMP4446 or COMP5046 in 2024 semester 1. There is no exclusion criteria.

(6) Do I have to be in the study? Can I withdraw from the study once I have started?

Being in this study is completely voluntary and you do not have to take part. Your decision whether to participate will not affect your current or future relationship with the researchers or anyone else at the University of Sydney.

Student will opt-out of the research by entering the Student ID number (SID) to the following form: *Google form link*

We will send two messages to students about using your data in the study. These messages will include the link to opt-out provided above.

After you receive the second message, you have one week in which to fill out the survey if you wish for your muddy card responses to not be included in this research. After this point, the data will be deidentified, and exclusion of specific data will not be possible.

(7) Are there any risks or costs associated with being in the study?

We do not expect that there will be any risks or costs associated with taking part in this study.

(8) Are there any benefits associated with being in the study?

No, there is no compensation for participating in this study.

(9) What will happen to information about me that is collected during the study?

No personal information about you will be collected and used in the study; your muddy card responses for each week will be de-identified prior to being analysed.

By participating in this study, you are agreeing to us collecting your muddy card responses and analysing them.

You agree that part of this analysis will involve uploading your muddy card responses to external large language models. When doing so, we will opt-out of having the data used by companies to train their models.

Your information will only be used for the purposes outlined in this Participant Information Statement.

Study findings may be published, but you will not be individually identifiable in these publications. Any details or analyses obtained from your responses will be published in a student's thesis.

As part of this studies research data management plan, your muddy card responses will be securely retained on the university servers for five years.

(10) What external large language models will be used during the study?

Large language models provided by Voyage AI and OpenAI will be used in this study. These providers were selected as they allow users to opt-out from having data retained by the company for the purpose of model retraining.

For more information regarding the Terms of Service for these external providers, please refer to:

- OpenAI: https://openai.com/policies/row-terms-of-use/
- Voyage AI: https://www.voyageai.com/tos

(11) Can I tell other people about the study?

Yes, you are welcome to tell other people about the study.

(12) What if I would like further information about the study?

When you have read this information, the student researcher Thomas Elton will be available to discuss it with you further and answer any questions you may have. If you would make you feel more comfortable, you are also free to email the chief investigator Dr Jonathan Kummerfeld.

Thomas Elton's email: telt8898@uni.sydney.edu.au

Dr Jonathan Kummerfeld's email: jonathan.kummerfeld@sydney.edu.au

(13) Will I be told the results of the study?

You have a right to receive feedback about the overall results of this study. Any feedback and publications resulting from this study will be available on this website:

https://tjelton.github.io/Dissemination-Muddy-Card-Analysis/

(14) What if I have a complaint or any concerns about the study?

Research involving humans in Australia is reviewed by an independent group of people called a Human Research Ethics Committee (HREC). The ethical aspects of this study have been approved by the HREC of the University of Sydney (2024/HE000932). As part of this process, we have agreed to carry out the study according to the National Statement on Ethical Conduct in Human Research (2007). This statement has been developed to protect people who agree to take part in research studies.

If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the university using the details outlined below. Please quote the study title and protocol number.

The Manager, Ethics Administration, University of Sydney:

- **Telephone:** +61 2 8627 8176
- **Email:** human.ethics@sydney.edu.au

B.2 2025 User Study Participant Information Statement

There were two participant information statements for the user study - one specific to the student users and the other for the teacher users. Ethics Reference Code: 2024/HE001599

Press *here* to return to the 'User Study' methodology section (section 3.4).

B.2.1 Student Participant Information Statement

Research study: Muddy Card AI System - User Study

Student Participant Information Statement

1. What is this study about?

Muddy cards are an active learning technique where following the end of a lecture, students will write down the most confusing (muddiest) point from the lecture. Typically, the teacher would read through these responses to understand the most common areas of confusion.

This process can take a very long time. Hence, this research study is about developing a system that speeds up the time it takes for instructors to collect and analyse their students' muddy card responses. Students will engage with the 'Student Interface', which is where students will record their muddy card response for a given lecture.

In this research study, we aim to use a mixed-method approach to understand the efficacy of this system from the perspective of both the students and teachers. Specifically, students will be invited to complete an optional survey to understand students' perceptions of muddy cards and the muddy card system.

Another main goal of this study is to develop a publicly accessible data set of muddy card responses. This data will be used in this study to benchmark the internal algorithms of the 'Muddy Card System', as well as informing future research. Every time students engage with the interface, they will have the opportunity to opt-in to having their muddy card response and related inputs added to the publicly accessible data set.

Please read this sheet carefully and ask questions about anything you don't understand or want to know more about.

2. Who is running this study?

The study is being carried out by the following researchers:

- Mr. Thomas Elton, University of Sydney.
- Dr. Jonathan Kummerfeld, Senior Lecturer (Computer Science), University of Sydney.

Thomas Elton is conducting this study as a basis for a Bachelor of Advanced Studies (Honours) at the University of Sydney.

3. Who can take part in the study?

Any person enrolled in the participating units can take part in this study. That is, anyone who is in a course using the Muddy Card System.

4. What will the study involve for me?

a. *Student Interface*

You will open a link that opens the Student Interface.

- From there, you will be asked to choose your unit of study, and the week that the lecture occurred in.
- You will enter your muddy card response, and SID number. You will also indicate how important it is to you that your muddy card point is addressed.
- After submitting your muddy card response, you may be brought to a page which asks an additional question. This question will ask you to select which of a list of other student muddy card responses are semantically the same as the one you submitted.
- Finally, you will be asked if you consent to having your muddy card response and related input information (such as class, week, and your answer to the additional select question) added to a public data set that is being collated as part of this research study. Your SID number will not be stored in the public dataset.

Some units will decide that using the Student Interface is compulsory for your unit of study. Under this circumstance:

- Students can still choose to use the Muddy Card System. These students can simply not provide consent to having their muddy card response (and related input information) added to the public data set.
- These students can then fill out an alternate form (completely external to the muddy card system) to provide their muddy card response.

b. *Survey*

There will be an optional survey towards the end of semester asking questions to understand your perception of the Student Interface and muddy cards in general.

5. How much of my time will the study take?

a. *Student Interface*

Using the Student Interface will be an ongoing process. For every use, we expect that it will take around 5 minutes to engage with the Student Interface. We expect that most classes will use the interface weekly.

b. *Survey*

We expect the survey to take around 10 minutes.

6. Can I withdraw once I have started?

Being in this study is completely voluntary and you do not have to take part.

Your decision will not affect your current or future relationship with the researchers or anyone else at The University of Sydney.

a. *Student Interface*

Consenting to have your muddy card response added to the public data set is optional. Additionally, the question asking for your consent is raised every time you enter a muddy card, meaning that some weeks you can consent to having your muddy card added to the public data set, and not in other weeks.

Some students may be required to complete muddy cards in their unit of study. In this case, students can still use the Muddy Card System without participating in the study. Participation in

the study is still optional as your data will not be directly used as part of the study unless you consent to adding your muddy card to the public data set. Additionally, when muddy cards are compulsory for a unit of study, we will also work with the instructor in setting up an external form for students to record their muddiest point if they are uncomfortable using the Muddy Card System.

You are free to withdraw your data from the public data set by filling out a survey and entering your SID number. A third-party (that is, someone who is not affiliated with this research) will find and remove your entries from the dataset. The survey can be found here (link to Google form).

b. *Survey*

Filling out the survey is optional. By submitting your survey, you consent to take part in the study. You can withdraw any time before you submit however once your responses are submitted, they cannot be withdrawn. This is because they are anonymous, and we will not be able to tell which one yours is.

7. Are there any risks or costs?

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

8. Are there any benefits?

a. *Student Interface*

There is no compensation for consenting to your muddy card responses being added to the public data set.

b. *Survey*

The final question in the survey will allow you to provide your email to enter a voucher giveaway. We will be randomly giving away 5 Westfields vouchers valued at $20 to students who complete the survey and provide their email.

9. What will happen to information that is collected?

In general, no personal information about you will be collected and used in the study. All data collected in this study will inform research findings which will likely be published.

a. *Student Interface*

- Your muddy card responses for each week will be de-identified prior to being analysed.
- If you opt-in to having your muddy cards added to the public dataset, this dataset will be openly available and can be accessed by any person.
- All de-identified data from the student interface will be securely retained on the university servers for five years.
- Your written muddy card response will likely be embedded using a third party large language model. Only providers with strict policies against data retention will be considered. The providers being considered for this project are presented in point 10.

b. *Survey Results*

- Survey results will be analysed and the results potentially published.
- If you provide your email to enter the draw for the $20 Westfields vouchers, these emails will be used purely for the purposes of the random draw. These emails will not be retained.
- All survey results (other than the entered emails) will be securely retained on the university servers for five years.

10. What external large language models will be used during the study?

Large language models provided by Voyage AI and OpenAI will be used in this study. These providers were selected as they allow users to opt-out from having data retained by the company for the purpose of model retraining.

For more information regarding the Terms of Service for these external providers, please refer to:

- OpenAI: https://openai.com/policies/row-terms-of-use/
- Voyage AI: https://www.voyageai.com/tos

11. Will I be told the results of the study?

You have a right to receive feedback about the overall results of this study. Any feedback and publications resulting from this study will be available on this website:

https://tjelton.github.io/Dissemination-Muddy-Card-Analysis/

The public data set of collected muddy cards will also be available through this link.

12. Can I view the teacher's participant information statement?

Yes you can. A link to it can be found *here* (link to the Teacher Participant Information Statment)

13. What if I would like more information?

When you have read this information, the student researcher Thomas Elton will be available to discuss it with you further and answer any questions you may have. If it would make you feel more comfortable, you are also free to email the chief investigator Dr Jonathan Kummerfeld.

- Thomas Elton's email: telt8898@uni.sydney.edu.au
- Dr Jonathan Kummerfeld's email: jonathan.kummerfeld@sydney.edu.au

14. What if I have a complaint or any concerns?

The ethical aspects of this study have been approved by the Human Research Ethics Committee (HREC) of The University of Sydney [ethics reference: 2024/HE001599] according to the National Statement on Ethical Conduct in Human Research. If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the University:

Human Ethics Manager

human.ethics@sydney.edu.au

+61 2 8627 8176.

B.2.2 Teacher Participant Information Statement

Note: When originally describing the muddy card study to teachers, the lecture-by-lecture survey was referred to as weekly surveys. Hence, in this participant information statement, whenever there is a reference to a 'Weekly Survey', this refers to the lecture-by-lecture survey.

Research study: Muddy Card AI System - User Study

Teacher Participant Information Statement

1. What is this study about?

Muddy cards are an active learning technique where following the end of a lecture, students will write down the most confusing (muddiest) point from the lecture. Typically, the teacher would read through these responses to understand the most common areas of confusion.

This process can take a very long time. Hence, this research study is about developing a system which speeds up the time it takes for instructors to collect and analyse their students' muddy card responses. We will refer to the overall system as the 'Muddy Card System'.

The 'Muddy Card System' can broadly be separated into two interfaces.

a. The 'Student Interface' is where students will record their muddy card response for a given lecture.

b. The 'Teacher Interface' is where course instructors will quickly analyse the muddy card responses entered by students for a given week.

In this research study, we aim to use a mixed-method approach to understand the efficacy of this system from the perspective of both the students and teachers. Students will be invited to complete an optional survey, and teachers, optional surveys and an optional interview. Through these surveys, we also aim to understand students and teachers' perceptions of muddy cards in general.

Another main goal for this study is to develop a publicly accessible data set of muddy card responses. This data will be used in this study to benchmark the internal algorithms of the 'Muddy Card System', as well as informing future research. Every time students engage with

the interface, they will have the opportunity to opt-in to having their muddy card response and related inputs added to the publicly accessible data set.

Please read this sheet carefully and ask questions about anything you don't understand or want to know more about.

2. Who is running this study?

The study is being carried out by the following researchers:

- Mr. Thomas Elton, University of Sydney.
- Dr. Jonathan Kummerfeld, Senior Lecturer (Computer Science), University of Sydney.

Thomas Elton is conducting this study as a basis for a Bachelor of Advanced Studies (Honours) at the University of Sydney.

3. Who can take part in the study?

Any unit-coordinator and lecturer from the University of Sydney can take part in this study. There may however be a cutoff once a large enough number of unit-coordinators volunteer. Additionally, once the last day to volunteer is reached, we will unfortunately be unable to support any new requests to join the study.

4. What will the study involve for me?

a. *Testing the Interface*

Prior to the first lecture, you will send the Student Interface URL to the students in your class. This will allow students to enter their muddy card responses.

When it comes to analysing the responses, you will open a link that opens the Teacher Interface.

- You will be prompted to enter your username and password that was sent to you by the research team (please contact them if you require these details again).
- You will select the lecture week that you wish to analyse the muddy card responses from.
- Follow the instructions on the interface to analyse the collected muddy card responses.

b. *Weekly Survey*

The last page of the Teacher Interface will be a short survey asking you about how you found the system. This survey is optional and you can choose to answer as many questions as you like.

c. *Interview*

You will be invited to take part in an optional interview with the student researcher.

- The interview will be semi-structured in nature.
- The interview will be performed over Zoom with a transcript created.

d. *Survey*

There will be an optional survey asking questions to understand your perception of the Muddy Card System and muddy cards in general. The survey will take around 25-30 minutes.

5. How much of my time will the study take?

a. *Testing the Interface*

Using the Teacher Interface will be an ongoing process. Depending on the specific variant of the teacher interface that you are testing in a given week, we expect it will take between 10-20 minutes.

b. *Weekly Survey*

The weekly survey should take around 5-10 minutes.

c. *Interview*

The interview will take around 1 hour.

d. *Survey*

We expect the survey to take around 25 minutes.

6. Can I withdraw once I have started?

Being in this study is completely voluntary and you do not have to take part.

Your decision will not affect your current or future relationship with the researchers or anyone else at The University of Sydney.

As the unit coordinator, you are able to withdraw from using the Muddy Card System in your classes at any point. If you decide to withdraw, we may potentially ask for your consent to still send out a survey to students in your unit of study to understand the student perception of the Student Interface and muddy cards in general. However, you would be under no obligation to agree to this.

a. *Weekly Survey*

Choosing to complete the weekly survey is optional. By submitting your survey, you consent to take part in the study. You can withdraw any time before you submit however once your responses are submitted, they cannot be withdrawn. This is because they are anonymous, and we will not be able to tell which one yours is.

b. *Interview*

Choosing to complete an interview is optional. Once the interview has commenced, you can choose to end the interview at any time.

c. *Survey*

Completing the survey is optional. By submitting your survey, you consent to take part in the study. You can withdraw any time before you submit however once your responses are submitted, they cannot be withdrawn. This is because they are anonymous, and we will not be able to tell which one yours is.

7. Are there any risks or costs?

Aside from giving up your time, we do not expect that there will be any risks or costs associated with taking part in this study.

8. Are there any benefits?

Access to the muddy card system is a benefit as it will help in implementing muddy cards into your unit.

There is no compensation for participating in any part of this study.

9. What will happen to information that is collected?

In general, no personal information about you will be collected and used in the study. All data collected in this study will inform research findings which will likely be published.

a. *Weekly Survey*

- Personal information will not be stored (such as name, email, IP address)
- Survey results will be analysed and the results potentially published.
- All survey results will be securely retained on the university servers for five years.

b. *Interview*

- Only the research team (i.e. Mr. Elton and Dr. Kummerfeld) will access the audio files to create the transcript.
- Interview transcripts will have your name and other peoples names redacted.
- Other details from the transcript will be redacted where it is likely that otherwise redacting these details would lead to participant reidentification. For example, if you state the course name that you teach, this will be redacted as this is easily linkable information.
- Quotes from your transcript may be included in research publications.
- All data collected from the interview will be securely retained on the university servers for five years.
- The interviewee will be provided with the completed transcript to view. They will then have 3 days to redact any further information before data analysis.

c. *Survey Results*

- Personal information will not be stored (such as name, email, IP address)
- Survey results will be analysed and the results potentially published.
- All survey results will be securely retained on the university servers for five years.

10. Will I be told the results of the study?

You have a right to receive feedback about the overall results of this study. Any feedback and publications resulting from this study will be available on this website:

https://tjelton.github.io/Dissemination-Muddy-Card-Analysis/

The public data set of collected muddy cards will also be available through this link.

11. Can I tell other people about the study?

Yes, you are welcome to tell other people about the study.

12. Can I view the student' participant information statement?

Yes you can, and it is encouraged that you read the student participant information statement before agreeing to participate in this study. A link to it can be found here [link to student participant information statement].

13. Will I have access to the Muddy Card System after the research study?

It is too early to tell at this stage. If you had a positive experience using this system, the research team would be happy to share how to run your own version of the system for future units of study.

14. What if I would like more information?

When you have read this information, the student researcher Thomas Elton will be available to discuss it with you further and answer any questions you may have. If it would make you feel more comfortable, you are also free to email the chief investigator Dr Jonathan Kummerfeld.

- Thomas Elton's email: telt8898@uni.sydney.edu.au
- Dr Jonathan Kummerfeld's email: jonathan.kummerfeld@sydney.edu.au

15. What if I have a complaint or any concerns?

The ethical aspects of this study have been approved by the Human Research Ethics Committee (HREC) of The University of Sydney [ethics reference: 2024/HE001599] according to the National Statement on Ethical Conduct in Human Research. If you are concerned about the way this study is being conducted or you wish to make a complaint to someone independent from the study, please contact the University:

Human Ethics Manager

human.ethics@sydney.edu.au

+61 2 8627 8176.

APPENDIX C

User Study: Student Introductory Message

Press *here* to return to the 'User Study' part of the methodology chapter.

Dear <Unit Code> students,

Your unit coordinator has agreed to participate in a research study involving an active learning technique. After some of your lectures, you will be asked to complete a 'muddy card', where you write down what you found most unclear in the lecture. This helps you reflect on the content, improving learning, and it informs the lecturer of possible confusion.

What will students need to do?

- After a lecture, you will click on this link here (links to the deployed student interface).
- The link will lead you to a page to select your unit of study and what week it is.
- You will then be prompted to write the most confusing thing from the lecture in your own words.
- You may also be prompted to identify similar responses written by your peers.

Your lecturer will also be provided with a separate interface which uses artificial intelligence techniques to find the most commonly occurring confusing points in the class. This will allow teachers to better support you with your learning.

Please feel free to watch this optional video briefly introducing the study and the student interface: Link here[1]

Using this system is **optional** for <Unit Code>.

This system is part of a research study. Data you enter into the system will be kept confidential. When using the system, a consent page will appear asking if you give consent to add your response to a public data set that we are designing. There is no obligation to provide consent. Additionally, if you provide consent and wish to later withdraw, you can fill out the survey here (link to survey). Students have up to one week after the study's conclusion to withdraw.

[1] Video Link:https://www.youtube.com/watch?v=ofs7-iPqDso

Additionally, at a later date, we will be inviting you to participate in an optional short survey to understand your experience using the software.

Please see the Student Participation Sheet (linked here) for further information about the study. This sheet will also contain our contact details should you have any questions.

Kind regards,

Thomas Elton and Dr Jonathan Kummerfeld

APPENDIX D

User Study: Teacher Lecture-by-Lecture Survey Questions

The survey questions are produced below. As seen in Figure A.2.7, the survey is embedded into the teacher's interface. Press *here* to return to the 'Surveys' part of the methodology chapter.

Please rate the following on a scale of 1 to 7 for your experience with the muddy card system teacher's interface for this week:

1 (Terrible) to 7 (Wonderful)

1 (Difficult) to 7 (Easy)

1 (Frusturating) to 7 (Satisfying)

1 (Dull) to 7 (Stimulating)

1 (Rigid) to 7 (Flexible)

For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree). Answer with reference to the teacher's interface you just used.

I enjoyed the time I spent using the software.

It is obvious that user needs have been fully taken into consideration.

This system has all the functions and capabilities I expect it to have.

I would recommend this software to my colleagues.

I can understand and act on the information provided by this software.

The student responses were grouped in a way that helped me find common points of confusion[1].

The grouping of student responses was accurate[1]

Do you plan to address the muddy card points raised? (Yes/No)

If yes to the previous question, how do you plan to address them? (free-form text box response)

[1]Users only receive this question if using the clustering variant (variant Y).

APPENDIX E

User Study: Final Surveys

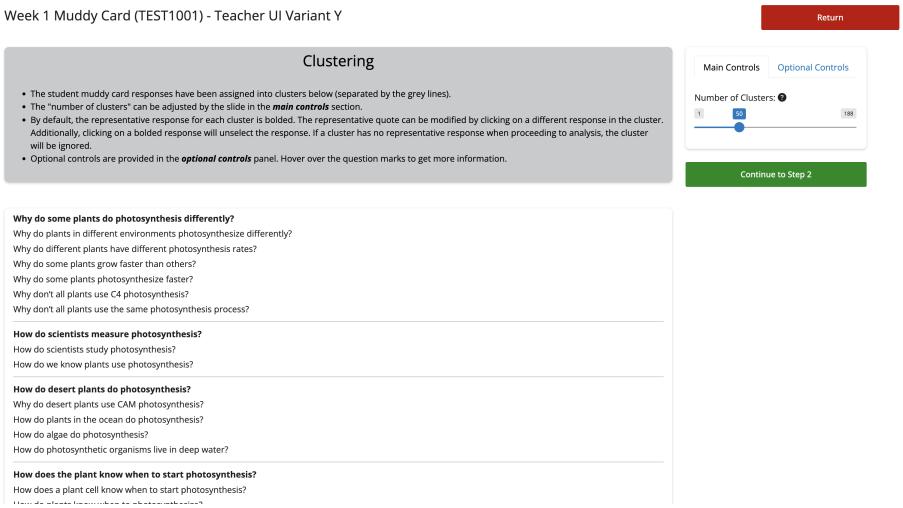
E.1 Teacher Final Survey

Below, we provide the survey questions for the final student survey. Press *here* to return to the 'Final Teacher Survey' part of the methodology chapter.

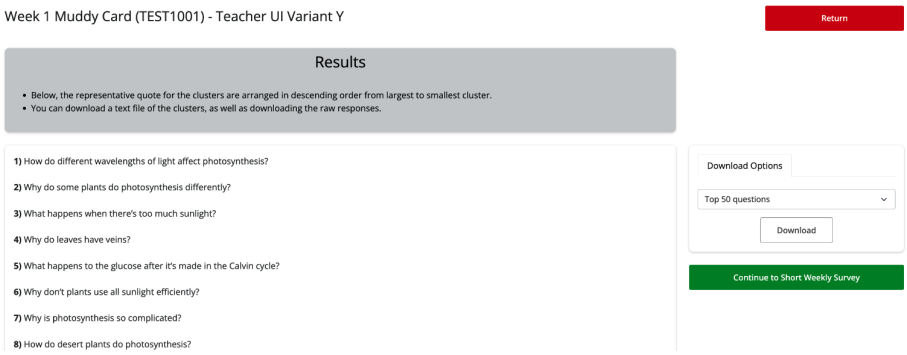
Pre-Survey Information:

For this survey, we will be **focusing on the Semantic Grouping/Clustering Interface**. As a reminder, here are some images of the clustering page and results page of this interface:

- **Clustering page:**



- **Results/Summary Page:**



For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree):

I regularly used the muddy card system.

I regularly reminded students to complete their muddy card responses.

I would like to use muddy cards in other courses I teach (either with or without this system).

I would like to use this system in other courses I teach.

I do not see the benefit of muddy cards.

Using this technology leads to higher student satisfaction.

This system helped me identify common sources of student confusion.

Students regularly submitted their muddy card responses.

Please provide your overall reactions to the muddy card system as a whole on the different scales provided:

1 (terrible) to 7 (wonderful)

1 (difficulty) to 7 (easy)

1 (frustrating) to 7 (satisfying)

1 (dull) to 7 (stimulating)

1 (rigid) to 7 (flexible)

For the following questions, answer them on a scale of 1 (very low) to 7 (very high):

How mentally demanding was the system?

How insecure, discouraged, irritated, stressed, and annoyed were you when using the system?

How hard did you have to work to accomplish a high level of performance in analysing the muddy card responses?

For the following question, answer on a scale of 1 (perfect) to 7 (failure):

How successful were you in accomplishing the goal of identifying common points of student confusion?

For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree):

Learning to operate the system was easy.

The instructions and prompts are helpful.

Tasks can be performed in a straight-forward manner.

For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree):

I enjoy the time I spend using the software.

The system was unreliable.

Correcting your mistakes was easy.

It is obvious that user's needs have been fully taken into consideration.

This software has at some time stopped unexpectedly.

Getting data files out of the system is not easy.

This software responds too slowly to inputs.

The experienced and inexperienced users' needs were taken into consideration.

This system has all the functions and capabilities I expect it to have.

I would recommend this software to my colleagues.

I can understand and act on the information provided by this software.

E.2 Student Final Survey

Below, we provide the survey questions for the final teacher survey. Press *here* to return to the 'Final Teacher Survey' part of the methodology section.

Select the unit/s in which you used the muddy card student interface:

Checkbox: The elements are the different units that took place in the study.

If 'COMP4446/5046' is selected in the previous question: **In COMP4446/5046, you could also submit your mudd cards via EdStem if you did not want to use the study's interface. How often did you use the study's interface?**

Radio button: "All the time. I never used EdStem.", "Most of the time. I rarely used Edstem.", "Half the time I used the study's interface, and half the time I used EdStem.", "Sometimes. I mostly used Edstem.", "No times. I always used Edstem."

For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree):

I benefited from writing muddy cards.

It took me minimal effort to compose and submit muddy cards.

At times, I found that writing muddy cards was frustrating.

I would like to see the muddy card technique used in other courses to some capacity.

Course teachers engaged with the classes' muddy card reflections.

Muddy Card Frequency: On average for your selected unit/s of study, how often did you fill out a muddy card? Choose the answer that best reflects your usage.

Radio button: "Never", "1 time during the semester", "1 time every month", "1 time every 2 weeks", "1 time every week", "More than once a week".

For the remainder of the questions, answer them in reference to the study's interface.

This text is only displayed if students selected COMP4446/5046 in the first question: If you are enrolled in COMP4446/5046, please only answer the questions for the remainder of this survey by considering the study's interface (not the EdStem submission's interface). If you never used the study's interface, please leave these questions blank.

For the following statements, rate them on a scale of 1 (strongly disagree) to 7 (strongly agree):

It was simple to use the system.

The instructions and prompts are helpful.

The speed of the system is fast enough.

I had no trouble in writing and submitting reflections.

Using this software is frustrating.

This software occasionally behaves in a way which can't be understood.

The organisation of information on the system screens is clear

I like using the interface of this system

Overall, I am satisfied with this system

I would like to use the system in other courses.

Rate the following on a scale of 1 (very low) to 7 (very high):

How mentally demanding was the system?

How insecure, discouraged, irritated, stressed, and annoyed were you with the system?

Rate the following on a scale of 1 (perfect) to 7 (failure):

How successful were you in accomplishing the goal of sharing how you were confused?

What is your email? Providing your email is compulsory if you wish to enter the gift voucher lucky draw so we can contact you if you win. Your email will not be stored for any other purpose.

APPENDIX F

User Study: Teacher Interviews

Below, we provide the interview protocol. Press *here* to return to the 'Interviews' part of the methodology chapter.

F.1 Pre-Interview Consent

[The following will be read before the interview officially commences.]

Hi! My name is Thomas Elton and I'm an honours student conducting research into the muddy card system that you have been trialling in your classes.

Thank you for your willingness to test the system. The purpose of this interview is to understand your perceptions behind the muddy card system, and muddy cards in general.

- This interview has been set up to be semi-structured in nature. This means that while I have some guiding questions to ask, there is some flexibility to add clarifying questions or questions that explore the points that you raise.
- You are free to provide as much or as little information that you please. You are also free to choose not to answer particular questions, and can also end the interview at any point.
- I expect that the interview will last from around 30 minutes to 1 hour.
- During the interview, you are also encouraged to share any other information that you think is useful. It is ok if the interview goes on short digressions.

It is important to know that this interview will be recorded. From this interview, a transcript will be created, and any identifiable information redacted. Additionally, any other individuals that you mention will also be redacted.

The resulting transcript will be analysed, with the results published in an honours thesis and potential other publications. These publications may include direct quotes from the interviews, but once again, any identifiable information from this interview will be redacted.

Are there any questions so far?

We're about to start the interview. I'm going to push the record button now.

[Start recording]

Do you consent to participating in this interview?

F.2 Semi-Structured Interview Protocol

General Thoughts on Muddy Cards

Before turning our attention to the muddy card system that you have been trialling, I wanted to understand your thoughts on muddy cards as a teaching tool. As a quick reminder, muddy cards are an active learning technique where students record the most confusing point from the lecture, and these results are subsequently analysed by the instructor.

As a teaching tool, have you ever used muddy cards before?

(If yes) When did you last use muddy cards, and how were they used?

(If no) Had you ever heard of muddy cards before participating in this study?

(If no) Would you ever consider implementing muddy cards in your own lectures or teaching activities?

(If yes) What has stopped you up until this point.

If using muddy cards, once you have identified that many students have the same confusing point, would you act on this information?

(If yes) How would you act?

(If no) Could you please expand on why you have answered "no"?

(If instructor has used muddy cards previously) When you used muddy cards previously, did you have any indication of how students responded to filling out muddy cards?

Questions around our Muddy Card System

Now we are going to ask some questions to understand your interaction and thoughts about the muddy card system that you have been trialling this semester.

Understanding Student Usage

From our records, we noticed that some indication of how often students were using the muddy card system filled out muddy card responses. Do you feel this was a large proportion of students in your class?

What methods, if any, did you use to encourage students to fill out their muddy card reflections?

Do you feel that these methods had an effect on the number of students filling out muddy cards?

(If low student engagement) Why don't you think students engaged with the muddy card system?

Exploring the Different Teacher Interfaces

When using this system, there were two different versions. I'm going to start by asking some questions about each of these versions.

Alphabetical Sortable Interface

One version you used was the 'alphabetical sortable interface', which was named teacher interface variant X. In this interface, you were presented with the muddy card responses in a list, and had the ability to sort the responses alphabetically.

Generally, how did you find the effectiveness of this system as it relates to understanding students' muddiest points?

How did you find the system's interface?

What did you like about the interface?

What did you dislike about the interface?

Is there anything you would change?

In this interface, you were provided controls to change the order of the muddy card response into alphabetical or reverse-alphabetical order. Were these controls helpful in analysing the students' muddiest points?

(If yes) Why was it helpful?

(If no) Why was it not helpful?

If you only had access to the alphabetical sortable interface for subsequent semesters, would you use muddy cards?

(If Yes) Can you elaborate on why you would use it?

(If No) Can you elaborate on why you would not use it?

Semantic Grouping Interface

Another version you used was the 'semantic grouping interface', which was named teacher interface variant Y. In this interface, muddy cards were separated into groups/clusters where grouped responses shared similar meanings.

Generally, how did you find the effectiveness of the semantic grouping interface as it relates to understanding students' muddiest points?

How did you find the system's interface?

What did you like about the interface?

What did you dislike about the interface?

Is there anything you would change?

Using this interface, there were different input features.

Which of these were helpful in analysing the students' muddiest points?

Which of these were not helpful in analysing the students' muddiest points?

==========

Note: If not already raised in the previous question, ensure that the following input controls are covered:

→ *Granularity/number of clusters slider.* Using this interface, you are able to adjust the granularity of the clusters. Was this control helpful in analysing the students' muddiest points?

(If yes) Why was it helpful?

(If no) What was it not helpful?

→ *Collapse clusters toggle.* Using this interface, you were able to collapse the clusters so only the bolded representative quote was displayed. Was this control helpful in analysing the students' muddiest points?

(If yes) Why was it helpful?

(If no) What was it not helpful?

→ *Choosing representative quotes individually.* Using this interface, you were able to select which sentence should be considered the representative quote of the sentence by clicking on a sentence within a group. Was this control helpful in analysing the students' muddiest points?

(If yes) Why was it helpful?

(If no) What was it not helpful?

→ *Rapidly changing all representative quotes.* Using this interface, you were able to change all representative quotes at once by selecting the method you wanted to use, and then pressing apply. Was this control helpful in analysing the students' muddiest points?

(If yes) Why was it helpful?

(If no) What was it not helpful?

→ *Summary page data download.* On the summary page you had the ability to download the original unordered muddy card responses. You could also download the grouping that you produced, or download just the representative muddy card responses. Did you regularly make use of this feature?

(If yes) Why did you use it?

(If yes) What did you do with this data?

(If No) Why didn't you use it?

==========

If you only had access to the semantic grouping interface for subsequent semesters, would you use muddy cards?

(If Yes) Can you elaborate on why you would use it?

(If No) Can you elaborate on why you would not use it?

Closing Questions

Do you have a preference between the alphabetical sortable interface or the semantic grouping interface?

(If there is a preference) Can you please elaborate on why you prefer the <their answer>.

(If no preference) Can you please elaborate on why you have no preference?

Do you think it would be worthwhile making muddy cards compulsory for your class?

(If yes) Would you ever consider attributing a small portion of your unit's grade to filling out muddy cards?

Do you have any suggestions on other features or interfaces that may be beneficial to you understanding students' muddy card responses?

We are about to finish this interview. Do you have anything else you would like to add?

Thank you for your time! We appreciate your support with this research.

APPENDIX G

Supplementary Figures and Tables: Research Question 1

G.1 2D Embedding Projection

Press *here* to return to the 'Embedding Models Benchmarking' section of the Results - Research Question 1 chapter.

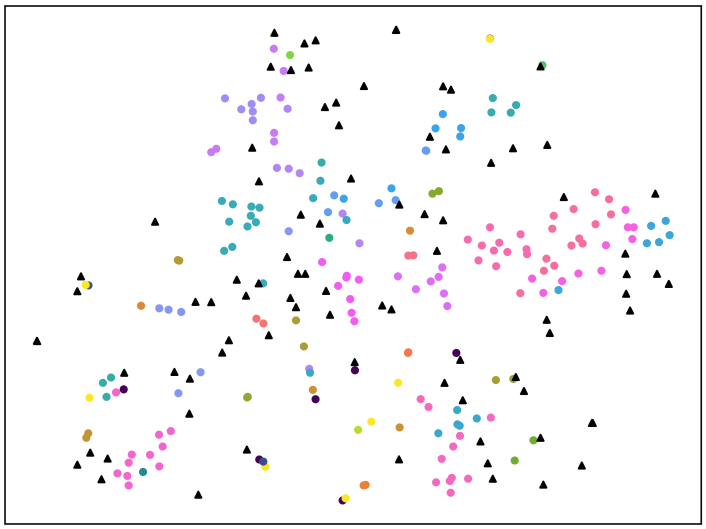


FIGURE G.1.1: TSNE 2D projection of the lecture 3 (2025 NLP) muddy card response sentence embeddings (using the `text-embedding-3-small` OpenAI embedding model). Coloured points indicate that these responses were manually labelled to be in the same cluster. A black triangle indicates that this sentence was manually assigned a solitary cluster.

G.2 Embedding Benchmarking

Press *here* to return to the 'Embedding Models Benchmarking' section of the Results - Research Question 1 chapter.

Adjusted Rand Index	NLP Lec. 1	NLP Lec. 3 †	NLP Lec. 4	NLP Lec. 6	NLP Lec. 7	Finance †	Average
SBERT							
all-MiniLM-L12-v2	0.252	0.399	**0.434**	0.349	0.379	0.339	0.359
multi-qa-distilbert-cos-v1	0.294	0.403	0.418	0.347	0.387	0.344	0.366
paraphrase-MiniLM-L3-v2	0.195	0.298	0.320	0.181	0.237	0.279	0.252
paraphrase-MiniLM-L6-v2	0.225	0.297	0.341	0.198	0.271	0.284	0.269
multi-qa-MiniLM-L6-cos-v1	0.269	0.423	0.406	0.380	0.355	**0.396**	0.372
multi-qa-mpnet-base-dot-v1	**0.305**	0.430	0.372	**0.400**	**0.418**	0.346	**0.379**
all-mpnet-base-v2	0.289	**0.439**	0.432	0.287	0.370	0.363	0.363
all-distilroberta-v1	0.260	0.383	0.433	0.285	0.362	0.317	0.340
OpenAI							
text-embedding-ada-002	0.239	0.411	0.370	0.286	**0.423**	**0.388**	0.353
text-embedding-3-small	0.258	0.463	0.418	0.281	0.323	0.335	0.346
text-embedding-3-large	**0.264**	**0.471**	**0.424**	**0.308**	0.375	0.353	**0.366**
Voyage AI							
voyage-3	0.241	0.438	0.379	**0.369**	0.373	0.289	0.348
voyage-3-lite	0.218	0.435	0.391	0.317	0.347	0.294	0.334
voyage-large-2-instruct	**0.299**	0.478	0.363	0.275	0.439	**0.506**	0.393
voyage-code-2	0.293	0.454	0.358	0.308	0.405	0.375	0.365
voyage-code-3	0.297	**0.511**	**0.439**	0.328	0.421	0.407	**0.400**
voyage-finance-2	0.239	0.447	0.411	0.331	0.348	0.380	0.359
voyage-law-2	0.294	0.493	0.375	0.341	**0.456**	0.414	0.396

†The manual clusters are the consensus between researchers 1 and 2.

TABLE G.1: Adjusted Rand index of different embedding models on 6 manually annotated data samples. Bold scores indicate the top score for each model family, and underlined scores are the highest of all embedding models.

Adj. Mutual Inf. Score	NLP Lec. 1	NLP Lec. 3 †	NLP Lec. 4	NLP Lec. 6	NLP Lec. 7	Finance †	Average
SBERT							
all-MiniLM-L12-v2	0.338	0.496	0.519	0.424	0.477	0.488	0.457
multi-qa-distilbert-cos-v1	0.371	0.500	0.507	0.475	0.486	0.511	0.475
paraphrase-MiniLM-L3-v2	0.271	0.414	0.408	0.309	0.326	0.425	0.359
paraphrase-MiniLM-L6-v2	0.293	0.419	0.439	0.347	0.358	0.430	0.381
multi-qa-MiniLM-L6-cos-v1	0.349	0.524	0.494	0.428	0.456	**0.556**	0.468
multi-qa-mpnet-base-dot-v1	**0.380**	**0.536**	0.500	**0.508**	**0.489**	0.518	**0.489**
all-mpnet-base-v2	0.361	0.527	**0.525**	0.428	0.452	0.526	0.470
all-distilroberta-v1	0.338	0.491	0.503	0.466	0.442	0.509	0.458
OpenAI							
text-embedding-ada-002	0.324	0.528	0.478	0.426	**0.515**	**0.561**	0.472
text-embedding-3-small	**0.366**	0.543	0.504	0.425	0.429	0.541	0.468
text-embedding-3-large	0.335	**0.557**	**0.509**	**0.460**	0.469	0.548	**0.479**
Voyage AI							
voyage-3	0.336	0.541	0.494	0.465	0.473	0.463	0.462
voyage-3-lite	0.303	0.530	0.508	0.438	0.436	0.472	0.448
voyage-large-2-instruct	**0.393**	0.576	0.473	0.366	0.507	0.478	0.465
voyage-code-2	0.377	0.570	0.471	0.422	0.490	0.520	0.475
voyage-code-3	0.384	**0.598**	**0.541**	**0.476**	0.501	**0.562**	**0.510**
voyage-finance-2	0.331	0.552	0.499	0.448	0.442	0.558	0.472
voyage-law-2	0.390	0.591	0.475	0.423	**0.521**	0.529	0.488

†The manual clusters are the consensus between researchers 1 and 2.

TABLE G.2: Adjusted mutual information score of different embedding models on 6 manually annotated data samples. Bold scores indicate the top score for each model family, and underlined scores are the highest of all embedding models.