DBE-KT22: A Knowledge Tracing Dataset Based on Online Student Evaluation

Ghodai Abdelrahman[1*], Sherif Abdelfattah[2], Qing Wang[1] and Yu Lin[1]

[1]School of Computing, Australian National University, Canberra, 2601, Australia.
[2]School of Engineering and Information Technology, University of New South Wales, Canberra, 2612, Australia.

*Corresponding author(s). E-mail(s): ghodai.abdelrahman@anu.edu.au;
Contributing authors: sherif.abdelfattah@unswalumni.com; qing.wang@anu.edu.au; yu.lin@anu.edu.au;

Abstract

Online education has gained an increasing importance over the last decade for providing affordable high quality education to students worldwide. This has been further magnified during the global pandemic as more students switched to study online. The majority of online education tasks, e.g., course recommendation, exercise recommendation, or automated evaluation, depends on tracking students' knowledge progress. This is known as the *Knowledge Tracing* problem in the literature. Addressing this problem requires collecting student evaluation data that can reflect their knowledge evolution over time. In this paper, we propose a new knowledge tracing dataset named *Database Exercises for Knowledge Tracing* (DBE-KT22) that is collected from an online student exercise system in a course taught at the Australian National University in Australia. We discuss the characteristics of the DBE-KT22 dataset and contrast it with the existing datasets in the knowledge tracing literature. Our dataset is available for public access through the Australian Data Archive platform upon proper citation to the authors.*

Keywords: Knowledge Tracing, Student, Exercise, Dataset

*https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/6DZWOH

1 Introduction

Over the last decade, online education systems rivaled the conventional classroom-based education for the ease of accessibility worldwide using the Internet, high quality of learning materials, and their affordable cost. The recent global pandemic further amplified the impact of online education as an effective alternative that could overcome physical distancing restrictions imposed on students and teaching staff in schools and university campuses. Nevertheless, one of the major challenges that need to be addressed in online education systems is the ability to effectively trace the learning progress of a student in a similar manner to what a human teacher would do in the classroom. Human teachers rely on their intuition and experience to estimate the knowledge state of a student and tailor the learning process accordingly.

Acquiring such ability would enable online education systems to archive many vital education objectives including customized curriculum generation, learning materials recommendation, exercise recommendation, automatic evaluation, or learning feedback generation. Achieving such objectives would facilitate automating the teaching process itself and pave the way for transforming the current online education systems into *Intelligent Tutoring Systems* (ITS). An ITS not only automates the teaching procedure using computer systems (e.g., web applications), but also handles supporting tasks such as customizing the learning experience, providing guidance and feedback to the students [1].

The *Knowledge Tracing* (KT) problem formulates the challenge of tracing a student knowledge state based on their exercise answering history [2, 3]. In particular, the exercise answering history could be represented as a sequence of question-answer pairs and the task of a solving computational model would be to predict likelihood of correctly answering the following questions. Figure 1 depicts a probabilistic graphical model for a KT scenario. At each time point i in a previous answering sequence that spans from time t_1 to t_L, we observe a question tag q_i (i.e., question text) and an answer $a_i \in \{0, 1\}$. A student knowledge state at a given time point is represented by a vector over the proficiency states of the involved learning concepts $[C_1, C_2, \ldots, C_N]$, where a learning concept C_n could be a topic in a course such as number addition, or subtraction in an elementary math subject. It has to be noted that the KT literature tends to refer to learning concepts as knowledge components (KCs) [3] as they constitute the components of a student's knowledge state. At the top part of Figure 1, we show a hidden Markov model representing the modes that an individual learning concept's proficiency state could take through the learning process including *known* for a sufficient knowledge, *unknown* for insufficient knowledge, and *forget* for forgetting what has been known before on this concept.

Figure 1 shows that the KT problem is a challenging one that involves a hierarchical dynamical system that consists of two levels including the individual proficiency states of KCs and the overall student's knowledge state. Early attempts that aimed at addressing the KT problem are dated back to

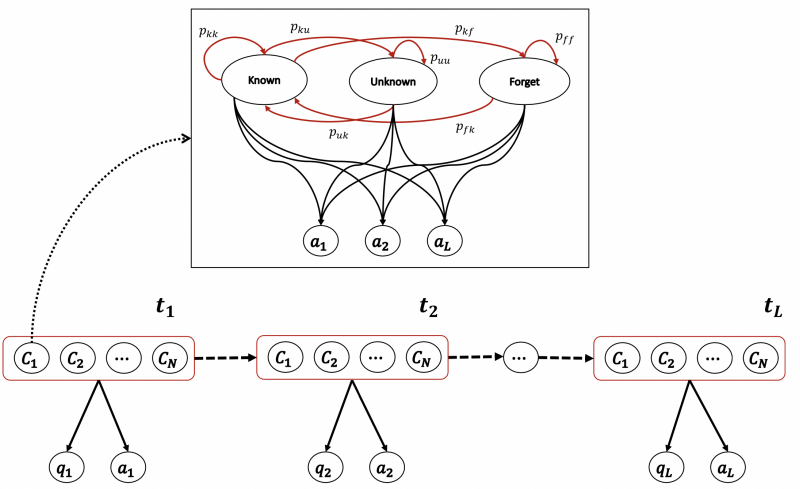


Fig. 1: A probabilistic graphical model representing a knowledge tracing scenario.

three decades ago, where Bayesian inference methods were used to predict the likelihood of correctly answering a new question given the previous answering sequence [4]. Recently, deep learning approaches [2, 5] have showed a promising potential in addressing the KT problem due to the representation capacity of deep neural networks and the use of advanced sequence modelling techniques such as differential associative memory [6, 7] to capture long-term dynamics in the exercise answering sequence, graph neural networks [8–11] for incorporating relationships among learning concepts and questions, and attention [12–15] for focusing on relevant questions in the previous answering sequence. The performance of deep neural models is significantly impacted by the volume and quality of available training data; thus, there is a need for datasets that could reflect the realistic dynamics in the online learning processes.

Despite of the availability of relevant KT datasets [2, 16–20], there are limitations that burden the use of these datasets in evaluating advanced deep KT models including (1) missing text data for question tags and description text for the involved learning concepts, which limits the use of advances in deep language models that could find an effective question embedding representation, (2) missing ground-truth on the relationships among learning concepts, which could be used by deep graph neural networks to incorporate the influence between learning concepts when updating the knowledge state, (3) missing ground-truth on question difficulty, which could provide an auxiliary signal when predicting the likelihood of correct answer, (4) lack of data for advanced student groups such as undergraduate and graduate students as the majority of datasets were collected from school-grade students, (5) lack of insights on

the student's uncertainty while answering a given question, which could provide features to quantify the likelihood of important answer aspects such as forgetting, guessing, or slip (i.e., mistaken answer).

To address these limitations, we propose a new dataset named *DBE-KT22*, which was collected from an introduction to relational databases taught at the Australian National University (ANU) in Australia for undergraduate and graduate students across multiple disciplines including computer science, engineering, arts, and business between 2018-2021 academic years. The contributions of the DBE-KT22 dataset can be summarized as follows:

- Providing detailed meta-data for questions including tag text, hints text, choice text, and images that provide a multi-model feature sources for question embedding representation learning.
- Providing detailed meta-data for the involved learning concepts including learning concept tag text, and description text to facilitate learning representative learning concept embeddings.
- Incorporating the ground-truth on the relationships between learning concepts and each other, and between learning concepts and questions in the form of graphs to facilitate the use of graph representation learning methods.
- Presenting the ground-truth on question difficulty provided by domain experts teaching the subject over a long time period. This provides auxiliary features that could be used for answer prediction and evaluating methods targeting quantifying the question difficulty such as question recommendation and curriculum generation models.
- Capturing different aspects that reflect a student's uncertainty while answering questions including self-feedback on question difficulty, self-feedback on confidence in answer, indication for using hints, number of times a student changes their answer, and the total time taken to answer. This dataset enables the use of answer uncertainty features in addition to the features from previous question answering sequence.

The reminder of this paper is organized as follows. Section 2 provides a detailed review of the available KT datasets and their characteristics. Section 3 introduces the details of the proposed DBE-KT22 dataset including data collection, data distribution, and formatting. Section 4 presents our experimental analysis to highlight the characteristics of the proposed dataset. Finally, Section 5 concludes the work.

2 Related Work

To facilitate the use of machine learning and statistical models in addressing the KT problem, many specialized datasets have been proposed by the literature. The early attempts to generate and publish a KT dataset date back to three decades ago [4] with the rise of statistical KT models. Since that time, the characteristics and volume of the collected data have been enriched to cope with the need of machine learning model for large sets of training data. In this

section, we introduce the characteristics of well-known KT datasets in the literature and highlight their differences. Table 1 summarizes the key statistical characteristics and meta-data for each dataset. We detail each of them as the following.

2.1 ASSISTments Datasets

This is one of the most popular group of datasets in the KT literature. The ASSISTments datasets [1] [16, 21] were collected in the time period $2009-2015$ using an online education platform with the same name[2]. In particular, the data was recorded based on a high school Math assessment sampled from the *Massachusetts Comprehensive Assessment System* (MCAS)[3]. We introduce each variant of this group of datasets as below.

- **ASSISTments2009:** This variant was sampled from the school year $2009-2010$ with a total number of $525,535$ exercise answering interactions generated by a total of $4,217$ students, a total of $26,688$ questions and a total of 123 KCs. Nonetheless, duplicates were reported to exist in the collected interactions [5]. Another limitation is that only two-thirds of the questions in this dataset were linked to their relevant KCs and marked with 'NA' in the KCs field. This limitation could burden the use of graph representation learning method capturing relationships among questions and KCs in the dataset.
- **ASSISTments2012:** This dataset was collected in the school year $2012-2013$ and it represents the largest variant in terms of data volume. It has a total number of $6,123,270$ recorded interactions, a total number of $179,999$ questions, a total number of $46,674$ students, and a total number of 265 KCs. Despite this large volume of recorded data samples, only 30% of questions was linked to their relevant KCs, limiting the benefit from the large volume of collected data.
- **ASSISTments2015:** This dataset records a total of $708,631$ interactions from the school year $2015-2016$, which was resulted by a total of $19,917$ students answering question from a set of 100 distinct questions and 100 of total KCs. Unlike the other variants of the ASSISTments group, there was no metadata on KCs such as KC description or title provided. Also, the linkage between questions and their relevant KCs was missing.

2.2 STATICS2011 Dataset

This dataset was collected from a mechanical engineering course[4] at the *Carnegie Mellon University* during the Fall semester in the year 2011. We note that the dataset is publicly available upon authors request [17]. It contains a total of $361,092$ interactions answered by a total of 335 students solving questions from a set of $1,224$ unique questions. The total number of KCs in this

[1] https://sites.google.com/site/assistmentsdata
[2] https://www.assistments.org/
[3] https://www.doe.mass.edu/mcas/testitems.html
[4] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507

dataset is 85. Despite the large number of recorded interactions, only half of this number includes the answer status (i.e., correct or wrong answer). To overcome is limitation, relevant KT literature that utilized the dataset applied a data preprocessing step to filter out data samples without an answer status [6].

2.3 Junyi Academy Dataset

This dataset was collected by the *Junyi Academy*[5], which is an online education platform in Taiwan. The data was recorded from high school level Math exercise practice between 2010 and 2015 years. In total there are 25,925,992 interactions, 247,606 students, 722 distinct questions and 41 KCs. It is to be noted that a preprocessing step is required to filter out interactions without an answer label and performing this step results in reduction of the total number of interaction to be 21,571,469. In addition, the linkage between questions and their relevant KCs is not complete for all the distinct questions in the dataset.

The Junyi academy released an update for this dataset that is available on Kaggle platform[6] with data recorded from the year 2018 to the year 2019. There are 11,468,379 interactions, 25,649 students and 1,701 distinct questions. The interactions were constrained in this updated to only include students attempted each question only once.

2.4 Simulated-5 (Synthetic) Dataset

This is a synthetic dataset[7] that was not recorded from an actual student practicing, but from simulation assuming a set of five KCs. It was proposed to evaluate one of the initial deep KT models named Deep Knowledge Tracing (DKT) proposed by Piech et al. [2]. The data is splitted into training and testing sets each containing a total of 50 distinct questions. Each question was limited to be linked to only one KC and a difficulty level. The authors simulated a total of 4,000 virtual students to answer each defined question resulting in a total of 200,000 interactions. The simulated student agents were following a policy derived from the *Item Response Theory* (IRT) [22]. Being a synthetic is one of the main limitations of this dataset as it cannot be used to reflect or analyze real student answering behaviors. Moreover, the dataset authors assumed that a question could only have one relevant KC, which is not a realistic assumption.

2.5 KDDcup Dataset

This dataset resulted from the KDDcup competition [8] [20] in 2010 representing an education data mining challenge. The challenge data was collected from an online education system called "The Cognitive Tutors" designed by *Carnegie*

[5] https://www.junyiacademy.org/
[6] https://www.kaggle.com/junyiacademy/learning-activity-public-dataset-by-junyi-academy/tasks
[7] https://github.com/chrispiech/DeepKnowledgeTracing/tree/master/data/synthetic
[8] https://pslcdatashop.web.cmu.edu/KDDCup

Learning Inc. [9] in the United States of America. The data were recorded in the time period $2005 - 2007$ for students from the age group of $13 - 14$ years old answering basic school-level Algebra questions. The dataset is divided into two subsets based on time period splits as follows.

- **Algebra 2005-2006:** this subset contains data collected between August 2005 and June 2006 with a total of $1,084$ distinct questions answered by 575 students resulting in $813,661$ interactions. There is a total of 112 KCs in this subset. Filtering out interactions that do not have linked KCs, the total number of interactions drops to 57.8% after preprocessing.
- **Algebra 2006-2007** This subset contains data recorded between 2006 and 2007 with a total number of $2,289,726$ interactions generated by $1,840$ students answering $90,831$ distinct questions with a total of 523 KCs. In similar to the $2005 - 2006$ subset, there is a portion of interactions without linked KCs and after filtering them out the number of interactions drops to $1,567,072$. We also note that upon inspecting this subset, we found inconsistencies in the timestamps recordings. This might impact the integrity of generated answering sequence and thus reduce the dependability of evaluation results on this subset.

2.6 EdNet Dataset

This dataset was released by an online education platform called "Santa"[10] developed by *Riiid* in South Korea. The platform is specialized in preparation for the Test of English for International Communication (TOEIC) exam. One of the distinguishing features of this dataset is the bundle question answering demanding that a set of questions to be answered altogether. Another important feature is recording data from different practicing platforms such as mobile or web applications to assess their effect on a student's knowledge state.

The data is divided into four subsets based on the notion of tasks that students are performing. It contains a total number of $95,293,926$ interactions generated by $784,309$ students answering questions from a set of $13,169$ unique questions and a total number of 188 KCs. The EdNet dataset was designed to provide incremental details about student activities and behaviors. The data recording time period spans over two years.

2.7 Limitations of Existing Datasets

We note that a common limitation across the reviewed KT datasets can be summarized as missing of the following 1) question tag text in the recorded data, 2) KC-KC relationships, 3) ground truth difficulty of questions, 4) recording answer-related facts such as time taken to answer or number of choice selection changes, and 5) student's feedback during practice such as their perceived question difficulty or their confidence in their answer. The proposed DBE-KT22 addresses all of these limitations to facilitate usage of advanced

[9] https://www.carnegielearning.com/
[10] https://www.riiid.co/

Table 1: Relevant knowledge tracing datasets and their characteristics.

Dataset	Description	Number of				Public available
		Questions	Students	Interactions	KCs	
ASSISTments	2009-2010	26,688	4,217	346,860	123	Yes
	2012-2013	179,999	46,674	6,123,270	265	Yes
	2014-2015	100	19,917	708,631	100	Yes
STATICS	2011	1,224	335	361,092	85	No
Junyi Academy	2015	722	247,606	25,925,992	41	Yes
Simulated-5	2015	50	4,000	200,000	5	Yes
KDDcup	Algebra 2005-2006	1,084	575	813,661	112	Yes
	Algebra 2006-2007	90,831	1,840	2,289,726	523	Yes
EdNet	KT1	13,169	784,309	95,293,926	188	Yes
	KT2	13,169	297,444	56,360,602	188	Yes
	KT3	13,169	297,915	89,270,654	293	Yes
	KT4	13,169	297,915	131,441,538	293	Yes

KT techniques and enables a variety of machine learning tasks to be evaluated such as natural language processing on question and KC textual data, graph modelling on question and KC relational data, curriculum learning based on question difficulty data, or learning cognitive analysis on student's recorded feedback during practice.

3 The DBE-KT22 Dataset

In this section, we introduce the methodology for producing the DBE-KT22 dataset including data collection procedure, data relational schema, data distribution, privacy preservation procedure, and data sharing. Additionally, we present the key statistical characteristics of the dataset and visualize them using graphs. Figure 2 presents our methodology for producing the DBE-KT22 dataset. In the data collection step, we elaborate on the means to collect the data, types of collected data, and aspects of the participation environment. The structure of the collected data and their relational dependencies is introduced in the data schema step. The data distribution step, presents abstract statistical aspects of the collected data. We clarify the followed procedure to preserve participants' privacy in the data anonymization step. Finally, we illustrate the procedure for sharing our collected data and making it publicly available for interested audience in the data sharing step.

3.1 Data Collection

The DBE-KT22 dataset was collected based on real student exercise practicing in the *Relational Databases* course taught at the Australian National University (ANU) in Australia. Students were mainly at the undergraduate level from different specializations including computer science, engineering, science, business, economics, arts, social sciences, and law and humanities. We note that to the best of our knowledge, our dataset is the first to cover such broad undergraduate student groups across the available KT datasets. The data was collected over a three-years time period from 2018 to 2021.

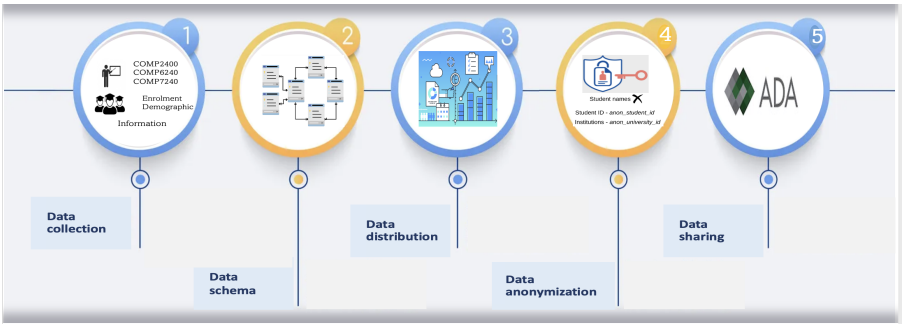


Fig. 2: The DBE-KT22 dataset production steps.

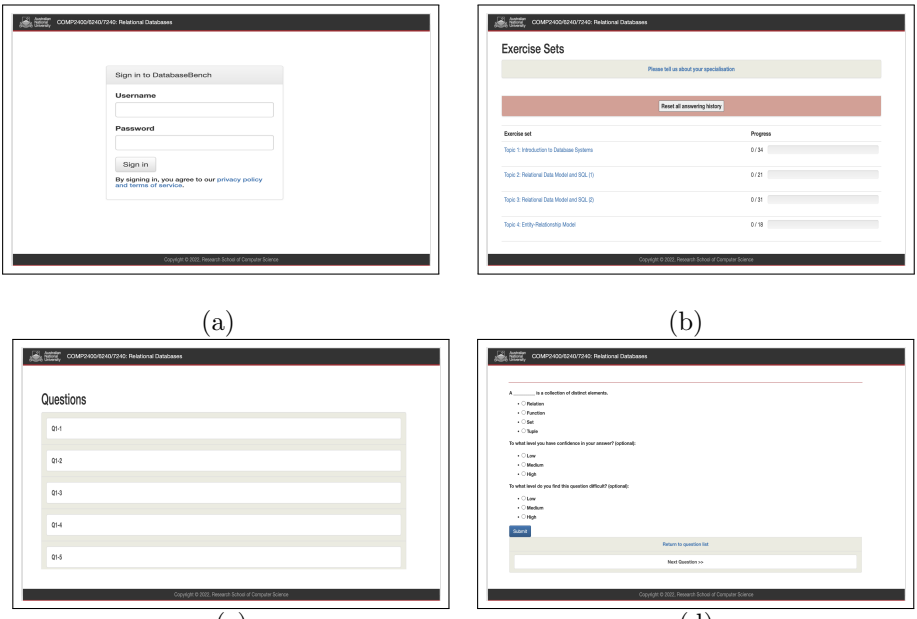


Fig. 3: Screenshots for the *CodeBench* web platform for exercise answering. (a) login page. (b) weekly exercise set page. (c) questions list per exercise set. (d) question answering page.

We utilized an online exercise practicing platform developed by the ANU named the *CodeBench* platform [11]. *CodeBench* is a web-based application that can be accessed exclusively by ANU students and staff using their university IDs. The platform enables student to practice exercises in a self-paced manner where exercises are organized by study weeks. Figure 3 shows screenshots for main web pages in the platform. Firstly, a student would log in using

[11] https://cs.anu.edu.au/dab/index.html

their university credentials. Then, they are redirected to a home page showing weekly exercise sets for the course; for each set, they will be able to track their completion progress shown through a progress par. Once a student selects a specific weekly exercise set, they see all the exercises within it and will be able to select a given exercise to practice. On the exercise practice page, student is presented with question title, meta-data such as graphs, tables, or images, and answer choices. We note that all of the questions in the dataset are multi-choice ones. Besides, a student will be able to show a hint if the instructor provided it for the current question. To record a student's perspective on question difficulty and their trust in the answer before submitting it, we provide two optional questions under each exercise asking the student for their feedback on difficulty and trust in the selected answer. Moreover, we passively record the number of times a student changed their answer selection and the total time taken to submit the answer for additional features about their confidence in the answer.

Figure 4 depicts a workflow for collecting the exercise answering data including 1) acquiring university credentials upon enrollment in the course, 2) logging into the *CodeBench* web platform, 3) in case it was the first time to log in, answering a questionnaire about specialization and confirming consent on the course practicing code of conduct, 4) practice exercises, and 5) save practice activity data into the database. We collected practice activity data over three academic years in the period $[2019 - 2021]$ and ordered the answering sequence for each student chronologically. The schema of the collected data is introduced in the following section.

3.2 Data Schema

The schema for the DBE-KT22 dataset follows a relational model that assigns each data aspect into a unique entity (i.e., table) and preserves the relationships across entities using primary-foreign key pairs. Figure 5 shows the entity-relationship diagram (ERD) for the DBE-KT22 database.

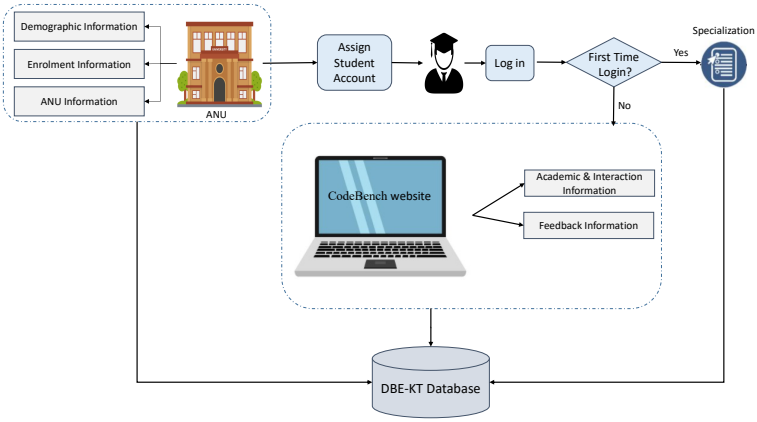


Fig. 4: DBE-KT22 data collection workflow.

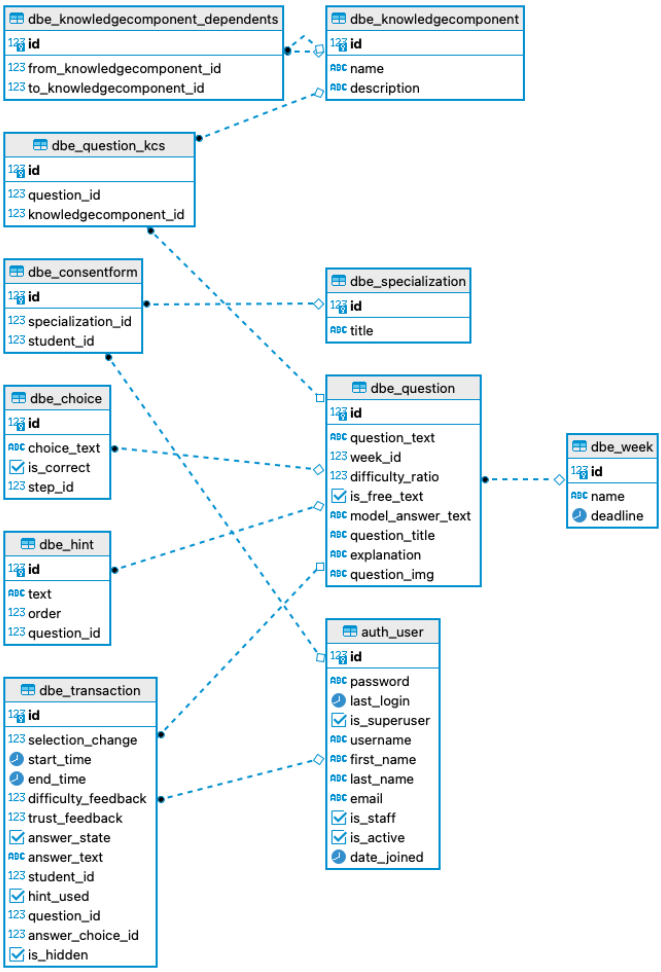


Fig. 5: Entity-Relationship Diagram for the DBE-KT22 database.

We describe each of the entity types in Figure 5 in the following:

- **dbe_question**: this is the main table in this database as it contains all question meta-data.
- **dbe_choice**: this table contains meta-data for choices per each question.
- **dbe_hints**: this table stores hint data for questions.
- **auth_user**: this table contains user login credentials and contact information such as email.
- **dbe_transaction** this table records question answering transaction data including the answer state.
- **dbe_week**: this table contains weekly exercise set data, which are groups of exercises based on a weekly decomposition for the course period.
- **dbe_consentform**: this table stores the data of the student consent including specialization inquiry and confirmation on the course policy.
- **dbe_specialization**: this is a lookup table for specialization data.
- **dbe_knowledgecomponent**: this table contains meta-data for knowledge components (KCs) in the course.
- **dbe_knowledgecomponents_dependendts**: this table stores relationships between the KCs.
- **dbe_question_kcs**: this table stores relationships between the questions and KCs.

We decomposed the previously described ERD into a set of files to facilitate data sharing and distribution. We followed the Comma Separated Value (CSV) file format for our dataset files for its popularity and the availability of processing software packages. We name each file with the same name of its equivalent table from the ERD.

Besides, we provide a Python script (i.e., uploaded with the dataset files) to generate training sequences of question answering with relevant meta-data. The script takes three run-time arguments including one for the desired sequence length to be used in sequencing a student answering history, a padding character for padding sequences shorter than the length argument, and an output file path to save the result. The resultant sequences file named *practice_sequences.json* is formatted using the JavaScript Object Notation (JSON) format as it is more convenient to represent nested objects such as our sequences samples. The file is structured as an array of JSON object, each has the following fields:

- **student_id**: ID of the current student in the sequence.
- **seq_len**: the length of the current sequence sample. Note that in case a sequence with shorter length than the length argument, this will show the sequence length without counting the padding chars.
- **question_ids**: unique ids for questions in the sequence.
- **answers**: answer status for each question in the sequence with 0 for wrong and 1 for correct status.
- **gt_difficulty**: ground truth difficulty level for each question provided by the course instructors. We use 1 for easy, 2 for medium, and 3 for difficult.

- **difficulty_feedback**: student's feedback on question difficulty before submitting the answer. We use 0 for not provided, 1 for easy, 2 for medium, and 3 for difficult.
- **answer_confidence**: student's feedback on their confidence in the selected answer. We use 0 for not provided, 1 for low, 2 for medium, and 3 for high.
- **hint_used**: a binary indicator for using hint per each question in the sequence.
- **time_taken**: time in seconds taken to answer each question in the sequence.
- **num_ans_changes**: count for answer selection change per each question in the sequence.

3.3 Dataset Distribution

In this section, we investigate the data distribution in the DBE-KT22 dataset using graphs. We start by showing the distribution of students across specializations, note that this only includes students who responded to the specialization questionnaire in the consent form. Figure 6 shows a bar chart for a visualization of this distribution. The majority of participants are coming from engineering and computer science specialization. In Figure 7, we show a density plot for the distribution of relevant KCs per a given question, it can be observed that majority of questions in the dataset has 1 to 2 relevant KCs. As we include question text in our dataset, it is useful to show the text token length distribution as it is a vital argument in for text embedding models. Figure 8 shows a density plot for the token length in question text, the majority of questions has a length less than 50 tokens, with a minority stretching this number around 300 tokens. Figure 9.a shows a pie chart for the distribution of question difficulty in the dataset that reflects a reasonable balance between the easy and medium levels and a minority of difficult questions. As per Figure 9.b, the majority of questions (around 85%) in the dataset does not have an explanation (i.e., instructor's description of the modal answer), which is aligned with the question difficulty distribution as 84% of questions are easy or medium while usually only difficult ones need to have an explanation. The hint (i.e., a small piece of information that could help to approach the answer, yet it couldn't be used as an answer) distribution across questions in Figure 9.c is showing that around 66.5% of questions has a hint note to help the student when requested, which covers all the medium and difficult questions in the dataset.

3.4 Data Anonymization

While our dataset collects data based on student exercise practice activities, their personal data is usually not of interest for the knowledge tracing context. Thus, we did not aim to collect any personal identifying information in our dataset. We only included two fields per each student record including their ID and specialization. For the IDs, we use an incremental identifier for the dataset records rather than the university issued ID number to preserve the privacy

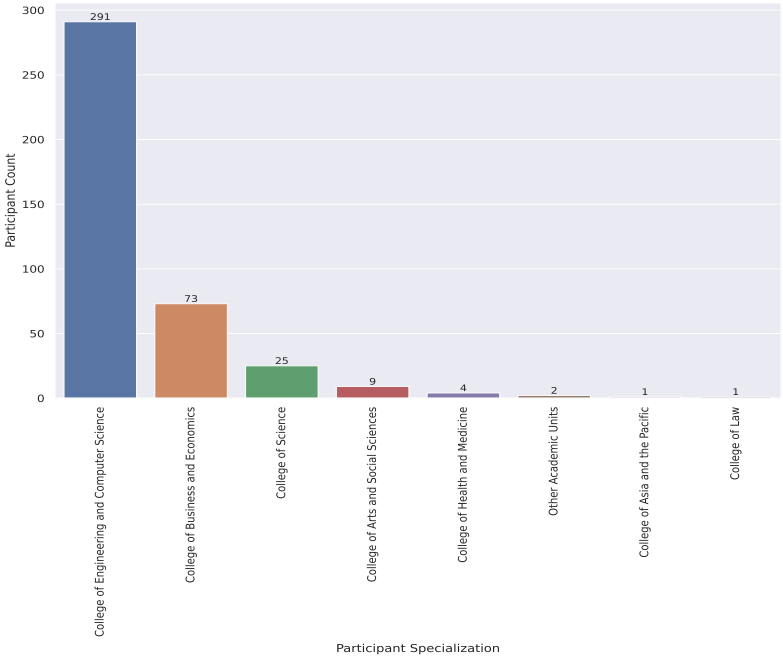


Fig. 6: The distribution of Student specialization in the DBE-KT22 dataset.

of students. While the specialization field was an optional one during the data collection process and it can not be used to back tracking students identity.

We got our data collection and privacy preservation procedures reviewed and approved by the ANU human research ethics committee [12] under the protocol number of 2017/543.

3.5 Data Sharing

We share the DBE-KT22 dataset files through the *Australian Data Archive* (ADA) platform [13] under an unrestricted access policy. ADA is a platform maintained by the ANU for sharing scientific data with the public and it provides advanced data indexing and search capabilities. Our dataset can be downloaded through its relevant page [14] on the ADA that includes general information describing its content, license, and research objectives. The uploaded DBE-KT22 dataset content includes the following files:

- **kc_relationships.csv**: file contains relationships among KCs.
- **kcs.csv**: file contains meta-data of KCs.
- **question_choices.csv**: file contains choices meta-data per each question.

[12] human.ethics.officer@anu.edu.au
[13] https://ada.edu.au/
[14] https://dataverse.ada.edu.au/dataset.xhtml?persistentId=doi:10.26193/6DZWOH

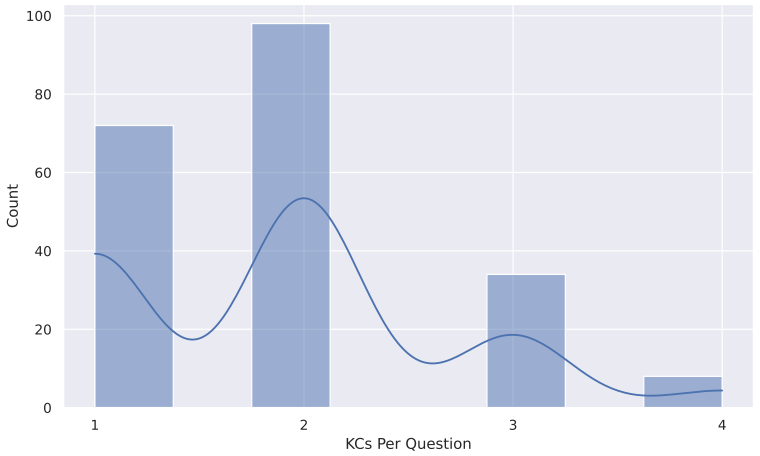


Fig. 7: The distribution of number of KCs per question in the DBE-KT22 dataset.

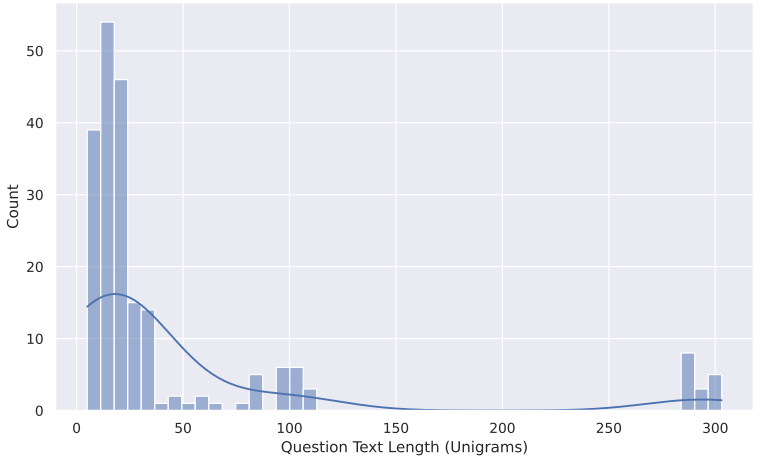
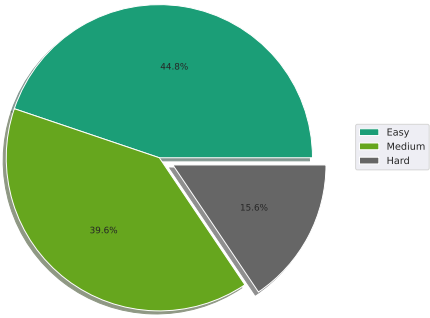
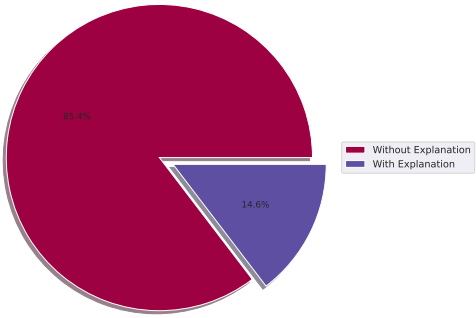


Fig. 8: The distribution of question text length in the DBE-KT22 dataset.

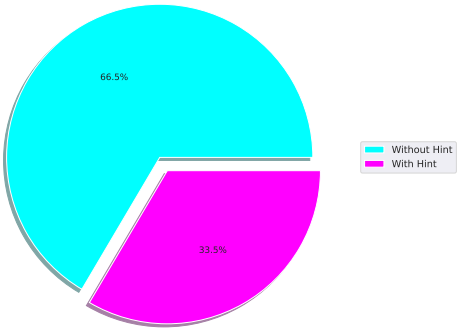
- **question_kc_relationships.csv**: file contains relationships between questions and KCs.
- **questions.csv**: file contains meta-data for questions including hint and explanation texts.
- **transaction.csv**: file contains meta-data for question practice attempts.
- **specialization.csv**: file contains meta-data for specializations.
- **student_specialization.csv**: file contains lookup data linking student ids with their corresponding specializations.
- **sequencer.py**: Python script file for generating training sequences.



(a)



(b)



(c)

Fig. 9: Distribution of question's difficulty, existence of explanation, and existence of hint in the DBE-KT22 dataset. (a) difficulty distribution. (b) explanation distribution. (c) hint distribution.

- **practice_sequences.json**: file contains generated training question answering sequences after executing the provided sequencer Python script.

4 Experiments

In this section, we introduce our experimental study investigating key characteristics of the DBE-KT22 dataset. Mainly, we conduct two experiments including an exploratory data analysis (EDA) experiment, and a question representation learning experiment. The former explores various dataset aspects relevant to the answer prediction task, while the latter evaluates different ways for an effective question representation learning.

4.1 Exploratory Data Analysis (EDA)

In this experiment, we aim to answer the following research questions:

- **Q1**: How far does the question difficulty affect the answer status?
- **Q2**: How does a student's answer confidence feedback imply the answer status?
- **Q3**: How far is a student's difficulty feedback aligned with the ground truth difficulty of questions?
- **Q4**: How far is a student's answer confidence feedback aligned with the ground truth difficulty of questions?
- **Q5**: How dependable is the answering time as an indicator factor for the answer status?
- **Q6**: How dependable is the hint usage as an indicator factor for the answer status?

To answer **Q1**, Figure 10 shows a grouped bar chart for answer status per each ground truth (i.e., provided by the course instructors) question difficulty level. It can be noticed that the distribution between true and false statuses is moving from being imbalanced towards the true status in the easy and medium difficulty levels to a balanced one in the hard level reflecting the increase of knowledge uncertainty with the increase of question difficulty. To answer **Q2**, we show the distribution of answer status over student's confidence feedback in Figure 11, which is showing a linear increasing pattern in the true answer probability w.r.t the increase in the confidence level. Answering **Q3**, we investigate the distribution of student's difficulty feedback over the levels of ground truth question difficulty. As per Figure 12, one can notice a linear decrease pattern in the number of students reporting an easy question difficulty (blue bars) with the increase of question ground truth difficulty level. Which is supporting the quality of the ground truth labeling done by course instructors. Similarly, to answer **Q4**, we inspect the distribution of a student's answer confidence feedback over the ground truth question difficulty. As per Figure 13, we observe that the percentage of high confidence feedback (green bars) is linearly decreased with the increase of ground truth question difficulty. For answering **Q5**, we show in Figure 14 the distribution of answering time

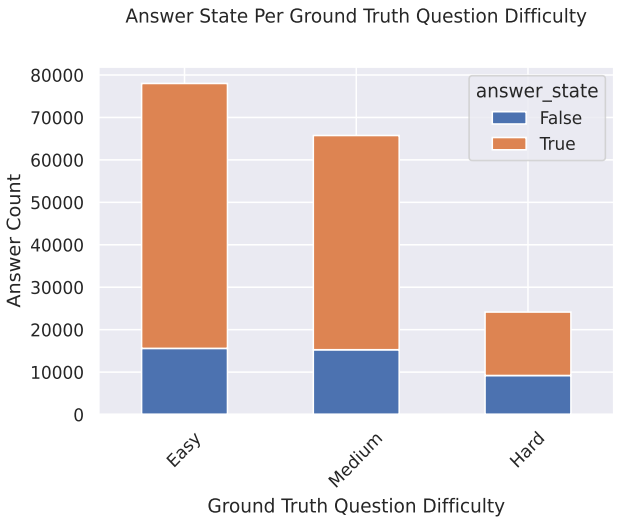


Fig. 10: A bar chart implying the impact of ground truth question difficulty on the answer status.

(i.e., time in seconds taken till submitting the answer) over answer status and ground truth question difficulty levels. We observe that for the easy and medium difficulty levels the differences between the second quartiles (median) and third quartiles of answering time is significant with more time taken for false (i.e., wrong) answer status, while these differences is less significant in the hard question difficulty level for the increased uncertainty. Finally, to answer **Q6**, we show the answer status distribution over hint usage binary indicator in Figure 15. We calculated the percentage of wrong answers (i.e., false status) for each bar. We observe that using the hint is an effective indicator of a gap in a student's knowledge and would probably imply a wrong answer status as the percentage of wrong answers is higher when using the hint in comparison to not using it.

4.2 Question Representation Learning

In this experiment, we evaluate different ways of learning a question representation using question text data. We utilize the uncased-base pre-trained BERT [23] language model using *Transformers* NLP models hub [15]. We define the following research questions to guide our evaluation in this experiment:

- **Q1**: What are the different ways to distill question embedding from pre-trained BERT language model? How effective is each way in clustering relevant questions?

[15] https://huggingface.co/bert-base-uncased

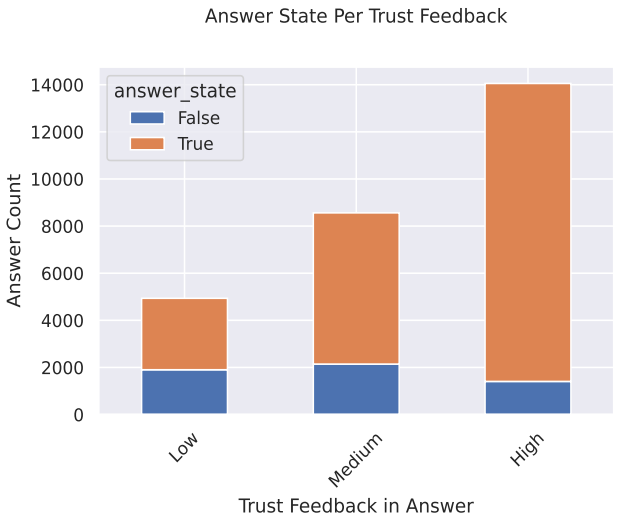


Fig. 11: A bar chart implying the distribution of answer status over the student's confidence feedback.

- **Q2**: How to fine-tune pre-trained embedding representation of question text? What is the effect of fine-tuning dataset size on the quality of embedding representation?

To answer **Q1**, we need a reference ground truth embedding representation for questions to evaluate different ways of distilling pre-trained text embeddings. We use the ground truth data on relationships between questions and KCs to design such a reference representation. We represent the ground truth on similarity between a question-question pair by the common KCs between them. Thus, a question in the reference ground truth representation is represented with a binary vector of length N (for the total number of KCs in the dataset), with 1s in the positions of its relevant KCs and 0s elsewhere. For distilling question text embedding from the pre-trained BERT model, we evaluate five different ways including:

1. **CLS token** embedding: represents each question with the CLS token of the final layer in the BERT model.
2. **last hidden layer** embedding: represents each question with the CLS token of the last hidden layer in the BERT model.
3. **last second hidden layer** embedding: represents each question with the CLS token of the last second hidden layer in the BERT model.
4. **last third hidden layer** embedding: represents each question with the CLS token of the last third hidden layer in the BERT model.
5. **pooled mean** embedding: represents each question with the reduced mean of the CLS tokens of the last three hidden layers in the BERT model.

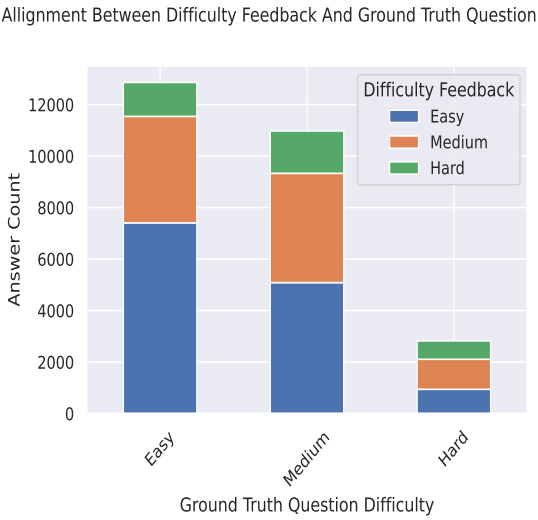


Fig. 12: A bar chart showing the distribution of student's difficulty feedback over ground truth question difficulty.

6. **pooled max** embedding: represents each question with the reduced max of the CLS tokens of the last three hidden layers in the BERT model.

We use the K-Means clustering algorithm [24] for performing clustering over each embedding method in the evaluation. To decide on the number of clusters K, we depend on a data-informed approach using the Elbow method [25]. Figure 16 shows a curve for the error sum of squares (SSE) per each K value. SSE is calculated by getting the sum of squared differences between each point and its' cluster mean point. We select $K = 10$ as a good configuration based on SSE curve. In order to visualize the clustering results per each method, we apply t-SNE manifolding [26] on each embedding variant using two manifold components. Figure 17 depicts the clustering results for each embedding method, we found that the **last hidden layer** embedding method outperformed others (see Table 2) on the defined clustering performance metrics including the *Inertia* metric [24], and the *Homogeneity* metric [27]. The *Inertia* metric measures clustering quality by calculating the distance between each data point and its cluster's centroid, squaring this distance, and taking the sum of squares for each cluster. The less the value of this metric, the better the clustering quality. While the *Homogeneity* metric measures the alignment of class labels within a given cluster using a ground truth of class labels. It has a value in the range of $[0, 1]$ with 1 for a perfect alignment and 0 for a complete dis-alignment. We use the cluster labels generated by the ground truth question representation as class labels to calculate this metric. The higher the value of this metric, the better the clustering quality.

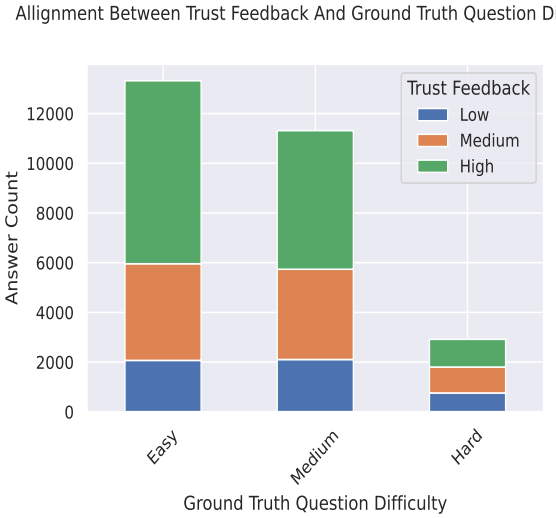


Fig. 13: A bar chart showing the distribution of student's answer confidence feedback over ground truth question difficulty.

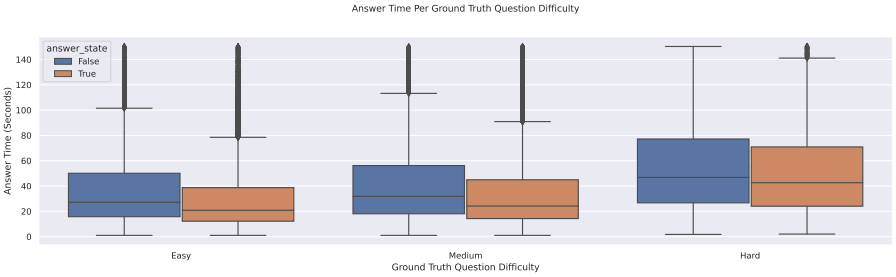


Fig. 14: A grouped boxplot chart showing the distribution of answering time across answer status and question difficulty levels.

Table 2: Summary for clustering performance results for ground truth embedding and different methods of question text embedding using Inertia and Homogeneity metrics.

Embedding Method	Inertia	Homogeneity
Ground Truth Embedding	203.8	1.0
CLS token embedding	1923.2	0.63
last hidden layer embedding	**1908.3**	**0.68**
last second hidden layer embedding	1935.0	0.59
last third hidden layer embedding	2045.6	0.58
pooled mean embedding	1981.7	0.55
pooled max embedding	2063.0	0.58

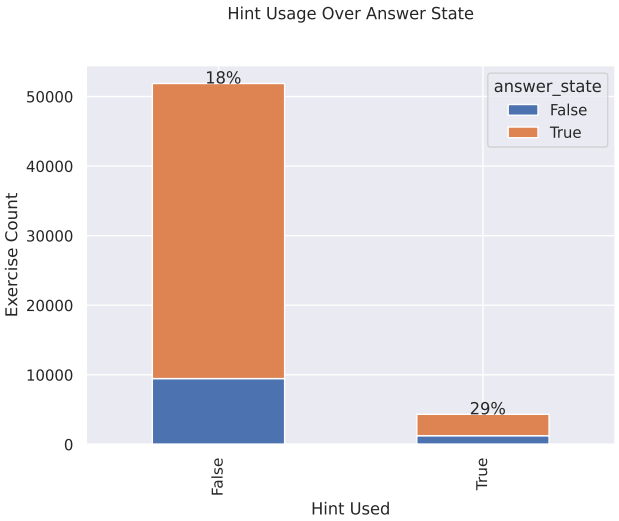


Fig. 15: A bar plot showing answer status distribution over hint usage. The percentage of wrong answers out of the total answers is shown above each bar.

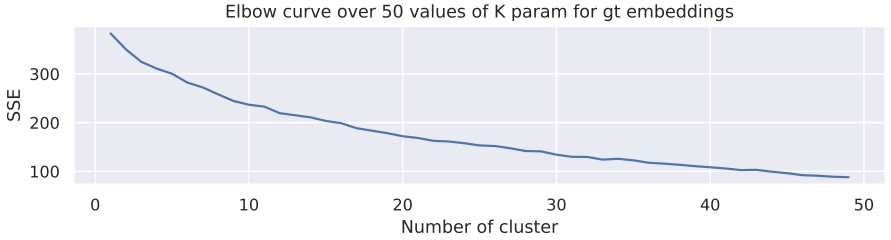


Fig. 16: The Elbow curve comparing error sum of squares (SSE) over number of clusters K using the ground truth question similarity representation.

In order to answer **Q2**, we investigate fine-tuning the best question text embedding method from **Q1** evaluation. We mean by fine-tuning to adapt the embedding parameters of the pre-trained BERT model on our question text data by formulating a supervised learning task to predict the number of shared KCs between a pair of questions using their text embeddings as input. To show the impact of fine-tuning dataset size on performance, we configure three size settings including 10%, 25%, and 50% of the original training dataset size. This enables us to reflect on the expected enhancement in the learned embedding with more ground truth data about question-KC relationships is introduced. We compare the performance of each variant using the same clustering metrics used to in **Q1** evaluation including *Inertia* and *Homogeneity* metric. Furthermore, we compare the performance of the fine-tuned embedding variants on classifying question ground truth difficulty label (i.e.,

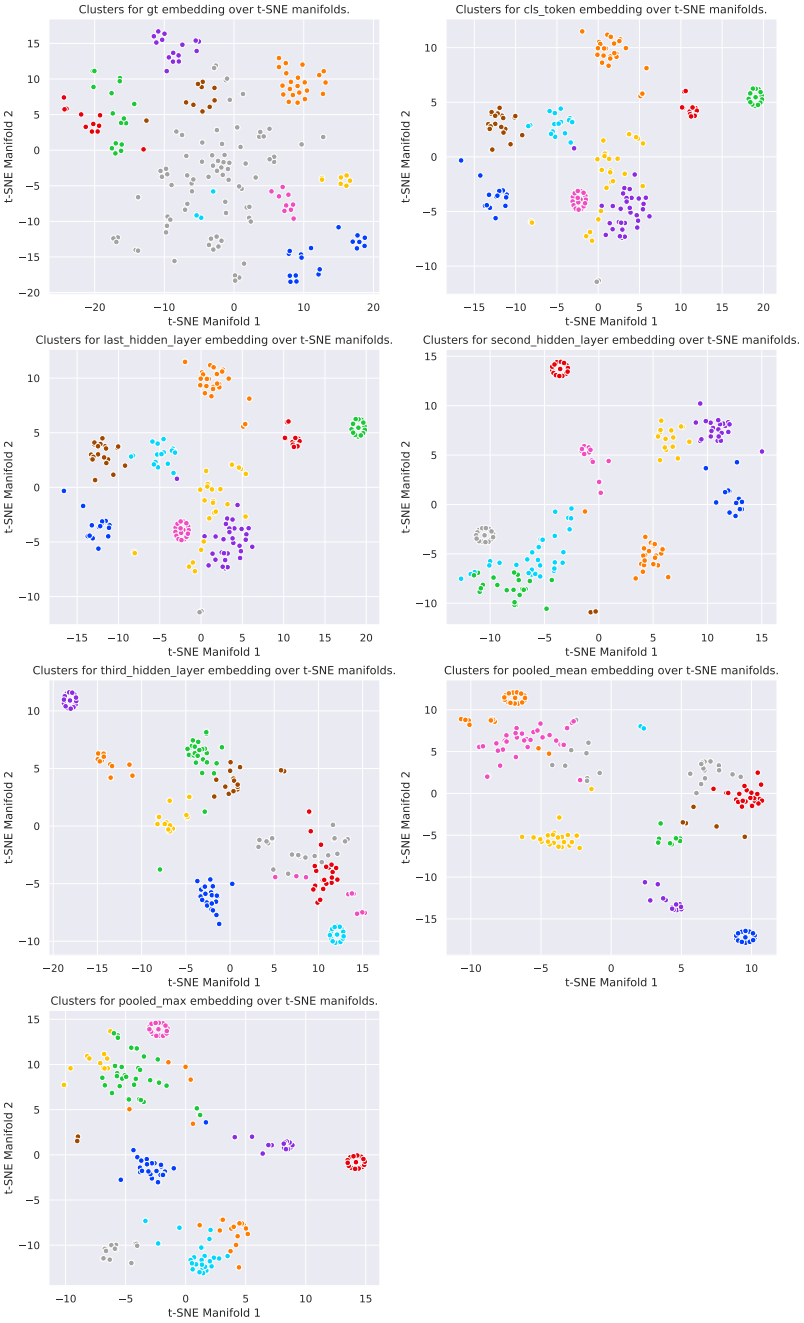


Fig. 17: A t-SNE two manifold visualization of clustering result for each question embedding method.

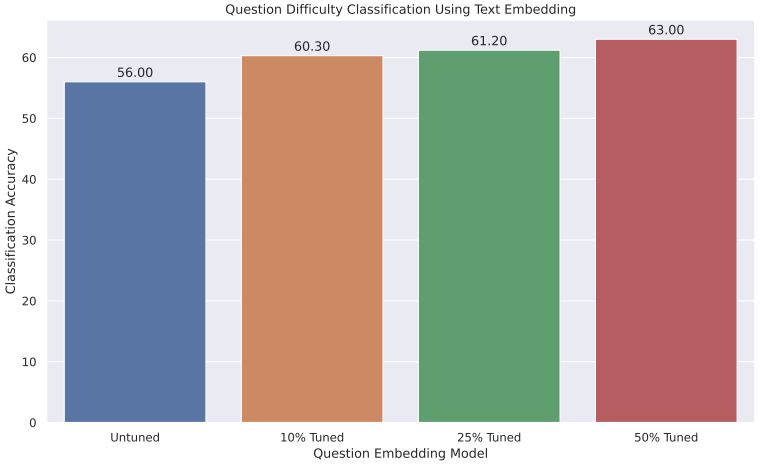


Fig. 18: A bar chart comparing classification accuracy for different fine-tuned question text embedding variants on classifying question difficulty label task.

Table 3: Summary for clustering performance results for fine-tuned variants of question text embedding using Inertia and Homogeneity metrics.

Embedding Method	Inertia	Homogeneity
Ground Truth Embedding	203.8	1.0
10%-tuned Embedding Variant	1870.5	0.71
25%-tuned Embedding Variant	1845.2	0.75
50%-tuned Embedding Variant	1830.6	0.81

label $\in \{1, 2, 3\}$) from text embedding using classification accuracy metric. As summarized in Table 3, we observe a linear increasing enhancement pattern on the performance of fine-tuned embedding with more ground truth data is introduced, yet using 10% of the ground truth data size was sufficient to get a significant enhancement with magnitude of -37.8, and 0.03 for the *Inertia* and *Homogeneity* metrics respectively comparing to the best performer from **Q1** evaluation. For question difficulty classification results shown in Figure 18, we notice a similar linearly increasing pattern in the classification accuracy with more ground truth data added during the the fine-tuning procedure and a significant enhancement over the untuned best performer model from **Q1** evaluation.

5 Conclusion

In this work, we proposed a new dataset for knowledge tracing research named DBE-KT22. The dataset was collected from real-world exercise answering activities in an undergraduate-level course taught at the Australian National

University in Australia within the period from 2019 to 2021. The collected data covers a wide range of aspects relevant to the knowledge tracing modelling including questions meta-data with text, KCs meta-data with text, relationships among questions and KCs, relationships among KCs and KCs, students feedback on observed difficulty and confidence in answer, and time taken for answering. This can facilitate more machine learning tasks to be formulated based on our dataset such as answer prediction, question text-aware embedding learning, graph representation learning on relationships between questions and KCs, question difficulty prediction, and cognitive analysis based on student's feedback data. We described the workflow of collecting the dataset and thoroughly investigated its characteristics. Moreover, we performed an experimental evaluation on different methods for learning an effective text-aware question embeddings from pre-trained language models. Finally, we made our dataset publicly available through the Australian Data Archive (ADA) platform with an unrestricted access license for easier utilization by researchers in the field knowledge tracing.

References

[1] Mousavinasab, E., Zarifsanaiey, N., Kalhori, S.R.N., Rakhshan, M., Keikha, L., Saeedi, M.G.: Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments **29** (2021)

[2] Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: Advances in Neural Information Processing SystemsNeurIPS (2015)

[3] Abdelrahman, G., Wang, Q., Nunes, B.P.: Knowledge tracing: A survey. arXiv preprint arXiv:2201.06953 (2022)

[4] Corbett, A.T., Anderson, J.R.: Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction **4** (1994)

[5] Xiong, X., Zhao, S., Inwegen, E.V., Beck, J.: Going deeper with deep knowledge tracing. In: Proceedings of the 9th International Conference on Educational Data Mining, EDM (2016)

[6] Zhang, J., Shi, X., King, I., Yeung, D.: Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th ACM International Conference on World Wide Web, WWW (2017)

[7] Abdelrahman, G., Wang, Q.: Knowledge tracing with sequential key-value memory networks. In: Proceedings of the 42nd ACM International Conference on Research and Development in Information Retrieval, SIGIR

(2019)

[8] Tong, S., Liu, Q., Huang, W., Huang, Z., Chen, E., Liu, C., Ma, H., Wang, S.: Structure-based knowledge tracing: An influence propagation view. In: Proceedings of the 20th IEEE International Conference on Data Mining, ICDM (2020)

[9] Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., Zhang, W., Yu, Y.: GIKT: A graph-based interaction model for knowledge tracing. In: Machine Learning and Knowledge Discovery in Databases - European Conference, PKDD (2020)

[10] Nakagawa, H., Iwasawa, Y., Matsuo, Y.: Graph-based knowledge tracing: Modeling student proficiency using graph neural network. In: International Conference on Web Intelligence, WIC (2019)

[11] Abdelrahman, G., Wang, Q.: Deep graph memory networks for forgetting-robust knowledge tracing. arXiv preprint arXiv:2108.08105 (2021)

[12] Pandey, S., Srivastava, J.: Rkt: Relation-aware self-attention for knowledge tracing. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM (2020)

[13] Pandey, S., Karypis, G.: A self attentive model for knowledge tracing. In: Proceedings of the 12th International Conference on Educational Data Mining, EDM (2019)

[14] Ghosh, A., Heffernan, N., Lan, A.S.: Context-aware attentive knowledge tracing. In: The 26th ACM Conference on Knowledge Discovery and Data Mining, SIGKDD (2020)

[15] Choi, Y., Lee, Y., Cho, J., Baek, J., Kim, B., Cha, Y., Shin, D., Bae, C., Heo, J.: Towards an appropriate query, key, and value computation for knowledge tracing. In: Proceedings of the Seventh ACM Conference on Learning@ Scale, L@S (2020)

[16] Feng, M., Heffernan, N., Koedinger, K.: Addressing the assessment challenge with an online system that tutors as it assesses. User modeling and user-adapted interaction,UMUAI **19** (2009)

[17] Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., Stamper, J.: A data repository for the edm community: The pslc datashop. Handbook of educational data mining **43** (2010)

[18] Chang, H.-S., Hsu, H.-J., Chen, K.-T.: Modeling exercise relationships in e-learning: A unified approach. In: Proceedings of the 8th International Conference on Educational Data Mining, EDM (2015)

[19] Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Baek, J., Bae, C., Kim, B., Heo, J.: Ednet: A large-scale hierarchical dataset in education. In: Artificial Intelligence in Education - 21st International Conference, AIED (2020)

[20] Stamper, J., Niculescu-Mizil, A., Ritter, S., Gordon, G., Koedinger, K.: Algebra i 2008–2009. challenge data set from kdd cup 2010 educational data mining challenge. Retrieved April **25** (2010)

[21] Pardos, Z.A., Baker, R.S., San Pedro, M.O., Gowda, S.M., Gowda, S.M.: Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. Journal of Learning Analytics **1** (2014)

[22] Foster, G.C., Min, H., Zickar, M.J.: Review of item response theory practices in organizational research: Lessons learned and paths forward. Organizational Research Methods **20** (2017)

[23] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT (2019)

[24] Hubert, L., Arabie, P.: Comparing partitions. Journal of classification **2**(1) (1985)

[25] Thorndike, R.L.: Who belongs in the family. In: Psychometrika (1953)

[26] Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

[27] Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL (2007)