PaperLM: A Pre-trained Model for Hierarchical Examination Paper Representation Learning

Minghui Shan
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
mhshan@mail.ustc.edu.cn

Yixiao Ma
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Data Science, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
mayx@mail.ustc.edu.cn

Shulan Ruan
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
slruan@mail.ustc.edu.cn

Zhi Cao
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
caozhicz@mail.ustc.edu.cn

Shiwei Tong
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
tongsw@mail.ustc.edu.cn

Qi Liu*
Anhui Province Key Laboratory of Big Data Analysis and Application, School of Computer Science and Technology, University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
Hefei, China
qiliuql@ustc.edu.cn

Yu Su
Hefei Normal University & Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
Hefei, China
yusu@hfnu.edu.cn

Shijin Wang
iFLYTEK AI Research (Central China) & State Key Laboratory of Cognitive Intelligence
Hefei, China
sjwang3@iflytek.com

ABSTRACT

Representation learning of examination papers is significantly crucial for online education systems, as it benefits various applications such as estimating paper difficulty and examination paper retrieval. Previous works mainly explore the representation learning of individual questions in an examination paper, with limited attention given to the examination paper as a whole. In fact, the structure of examination papers is strongly correlated with paper properties such as paper difficulty, which existing paper representation methods fail to capture adequately. To this end, we propose a pre-trained model namely **PaperLM** to learn the representation of examination papers. Our model integrates both the text content and hierarchical structure of examination papers within a single framework by converting the path of the Examination Organization Tree (EOT) into embedding. Furthermore, we specially design three pre-training objectives for PaperLM, namely EOT Node Relationship Prediction (ENRP), Question Type Prediction (QTP) and Paper Contrastive Learning (PCL), aiming to capture features from text and structure effectively. We pre-train our model on a real-world examination paper dataset, and then evaluate the model with three down-stream tasks: paper difficulty estimation, examination paper retrieval, and paper clustering. The experimental results demonstrate the effectiveness of our method.

CCS CONCEPTS

• **Information systems** → **Language models**; • **Applied computing** → **Education**.

KEYWORDS

Examination Paper Representation, Structured Document Analysis, Pre-trained Language Model

ACM Reference Format:
Minghui Shan, Yixiao Ma, Shulan Ruan, Zhi Cao, Shiwei Tong, Qi Liu, Yu Su, and Shijin Wang. 2023. PaperLM: A Pre-trained Model for Hierarchical

*Corresponding author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615003

Examination Paper Representation Learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3583780.3615003

1 INTRODUCTION

In recent years, there has been a significant increase in global attention towards online education systems. To achieve the goal of personalized education for each student, online education systems usually equip with massive exercises and provide customized examination papers for students' self assessments.

Many previous works [10, 12, 21, 22, 31] mainly focus on the representation of individual questions, while few pay attention to the understanding of the whole examination paper. Simply adding up the representations of all questions in an examination paper will result in the loss of rich information, such as the organization of questions. Therefore it is imperative to develop a tailored representation learning method for examination papers.

An examination paper can be treated as a special type of document. Early document representation works [2, 17, 37], which only consider textual information, are difficult to capture global structural information. In recent years, many works start learning representations of documents using pre-trained model. They tend to integrate additional information such as document layout [40] and document images [11, 20, 41] with document text. These works consider the unique features of visually rich documents (e.g. forms), so they are not suitable for examination papers. For documents with strong structural information like web pages, there are also works [4, 18, 39] integrating the document source code into models. However, these works design their models and pre-training strategies for specific types of documents, which cannot suit the unique structure features and domain-specific features (e.g.knowledge concept) of examination papers. Aiming to design a representation learning method specially for examination papers, Ma et. al [25] propose a multi-layer model where the hierarchical structure information of examination papers is extracted to enhance the representation. Although the structure information is used, they ignore the correlation between structure and properties of examination papers such as paper difficulty. Therefore, we hope to design a model for learning effective representations of examination papers considering the unique structure of examination papers and the correlation between structure and properties of papers.

In addition to textual information, an examination paper also has global structure information, which is abstracted into a special structure named Examination Organized Tree (EOT) in [25]. We display an example examination paper and its corresponding EOT in Figure 1. As the figure depicts, EOT is a tree structure in which the leaf nodes represent questions and the internal nodes represent conditions or summarizations of their child nodes. The hierarchical structure of the examination paper, as represented by the EOT, plays a pivotal role in paper design. Many researchers [5, 26] have studied the effects of question order on test performance. They find that students tend to performance better when presented with an examination paper arranged in an easy-to-hard question sequence rather than a hard-to-easy arrangement. Besides, the question type distribution within an examination paper is also strongly correlated with paper properties, which is concluded from our data analysis later on. Therefore the structure information of examination papers is vital for learning an informative paper representation.

However, there are still some challenges in learning an informative and effective representation of the examination paper using this structure information. Firstly, there are two distinct aspects of information: the semantic information derived from the content of paper text and the structural information provided by the EOT. Finding an effective approach to combine them poses a challenge. Secondly, although the structure of examination papers is correlated with paper properties, this correlation is implicit. How to instruct the model learn such correlation deserves exploration. Moreover, examination paper retrieval is an essential application that requires model's ability to distinguish holistic features of examination papers such as the examination scope. Previous document representation works [11, 18, 40, 41] mainly contribute to effective method to model additional modalities, while they are relatively simple in their modeling of textual information by following the Masked Language Model (MLM) in BERT [15]. This pre-training pattern focuses on literal details and lacks the capability to capture holistic features. How to capture the holistic features of examination papers poses an additional challenge.

To this end, in this paper, we introduce a novel pre-trained model named PaperLM to fuse the textual and hierarchical structural of the examination paper within a unified framework. To take advantage of existing pre-trained language models, we use BERT as the encoder backbone of PaperLM. Inspired by MarkupLM [18], we define the EOT path for each node in EOT and convert it into an embedding, enabling the model to access the examination paper's structure, including the arrangement of questions and the question type distribution. Since examination papers are typically long documents, we establish our representation learning at the test item level (EOT node level) instead of the single word level in papers. To effectively pre-train PaperLM, we propose three pre-training strategies. Firstly, in order to learn the hierarchical structure of examination papers and better integrate the textual and structural information, we introduce the EOT Node Relationship Prediction (ENRP) objective. Secondly, we propose the Question Type Prediction (QTP) objective. Building upon the ENRP task, the QTP objective instructs the model to identify different question types in papers, enabling it to learn the correlation between question type distribution and paper properties. Thirdly, we adopt the Paper contrastive Learning (PCL) strategy, where we specially construct contrastive samples to capture the holistic features of examination papers. The PaperLM is pre-trained on high school mathematical examination papers collected from an online education system. Finally, we adopt three down-stream tasks: paper difficulty estimation, examination paper retrieval, and paper clustering to demonstrate the effectiveness of our proposed pre-trained method.

2 RELATED WORKS

Pre-trained Language Model. Recent years have witnessed the rapid development of pre-trained language models (PLMs) [1, 15, 24, 29, 33, 34, 45], which has greatly promoted the performance of many tasks in natural language processing(NLP). Among the massive literatures of PLMs, the "pre-trained and fine-tuning" paradigm has been widely used. Along this research line, BERT[15]

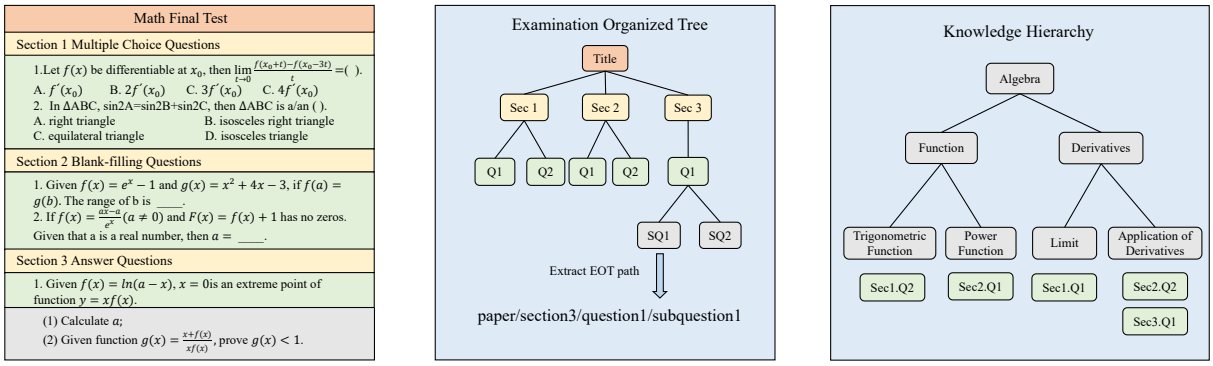


Figure 1: Examples of examination paper, Examination Organization Tree(EOT) and knowledge hierarchy.

firstly introduces a novel pre-training objective, Masked Language Model(MLM), which becomes the foundation of various representation methods [1, 24, 45]. For example, RoBERTa [24] is pre-trained with larger datasets and more optimized pre-training strategies compared with BERT. Longformer [1] further proposes sliding window, dilated sliding window and global attention to achieve a self-attention of linear time complexity, enabling it to process longer sequences compared to BERT and RoBERTa. Although remarkable success has been achieved in NLP tasks, these models mainly concentrate on text modality, which is not ideal for representing documents with hierarchical structure such as examination papers.

Document Understanding. Document understanding is an important research direction for NLP, which supports various applications like document classification[20, 40] and document information extraction[14, 30]. Modeling the additional information (e.g. structure [18], layout [23, 40], images[11, 20, 41]) of documents is the most used method in document understanding. Some early researches [23, 32, 44] use Graph Convolutional Networks (GCN) to aggregate the global representation from each unit of the document. Qi et al. [32] model the global word-relation structure of a document using GCN to improve context-aware document ranking. Liu et al. [23] further leverage visual information presented in visually rich documents. In recent years, following the success of the BERT-like models, some works[11, 19, 20, 40, 41] have started pre-training models on documents with multimodal information such as images, layout and various forms of structure in documents. LayoutLM [40] inherits the main idea from BERT while jointing layout representation learning by encoding spatial coordinates of text, and achieves great performance in down-stream tasks like form understanding and receipt understanding. Based on LayoutLM, LayoutLMv2 [41] and LayoutLMv3 [11] further enhance the model by integrating visual features in the pre-training phase, resulting in significant performance improvements. SelfDoc [20] introduces cross-modal learning to fully leverage multimodal information including text, layout and image. MarkupLM [18] and WebKE [39] are examples of models that combine document text and markup language within a single pre-trained model, specifically designed for representing HTML documents. However, these works mainly design and train models for specific types of documents, which fail to directly represent examination papers due to its unique structure and domain-specific features (e.g. knowledge concepts). To apply document understanding to intelligent education area, Ma et al. [25] investigate the hierarchical structure of examination papers and propose the Examination Organization Encoder (EOE) to learn a robust representation of the examination paper. Despite they explored the structure of examination papers, the correlation between structure and properties of examination papers such as paper difficulty is largely ignored.

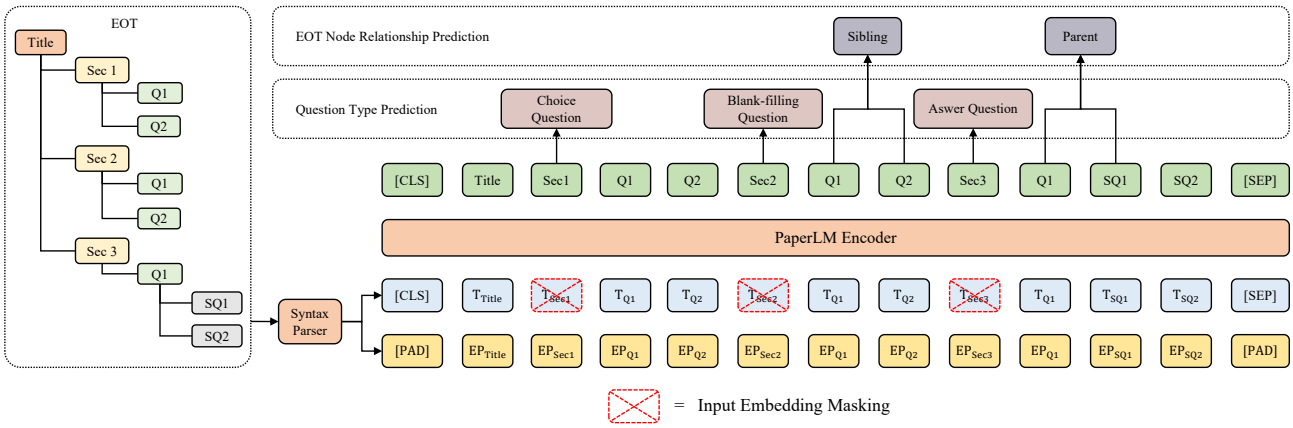
Contrastive Learning. In last several years, contrastive learning has been widely applied in representation learning and shown strong effectiveness [3, 8, 36]. The objective of contrastive learning is to learn effective representations by discerning whether samples are overall similar or dissimilar. This is achieved by employing data augmentation techniques to generate positive samples while treating other samples as negatives. The selection and number of negative samples greatly affects the performance of the model. To enhance the capacity of negative samples, memory bank techniques [9, 38] have been introduced to augment the negative sample pool and thereby improving the learning process. Recently, the influence of negative has been explored. BYOL [7] is introduced to prove that contrastive learning is effective even without negative samples. Khosla et al. [16] prove that contrastive learning can also be applied by constructing the positive pairs with label supervision.

3 PRELIMINARY

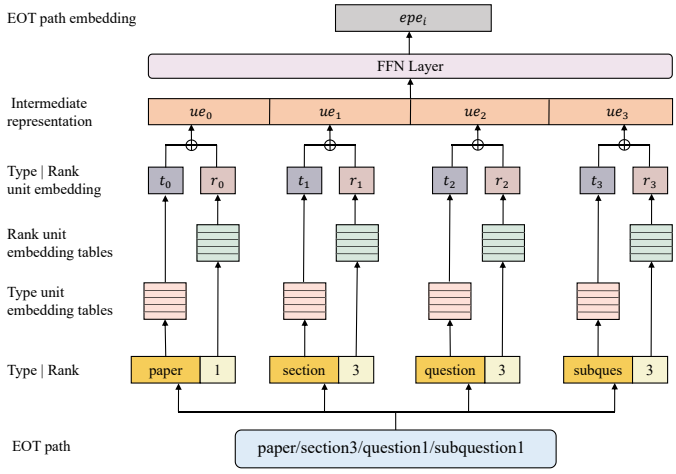
In this section, we first provide the description of EOT and knowledge hierarchy and then give the formal definition of examination paper representation.

3.1 Examination Organization Tree

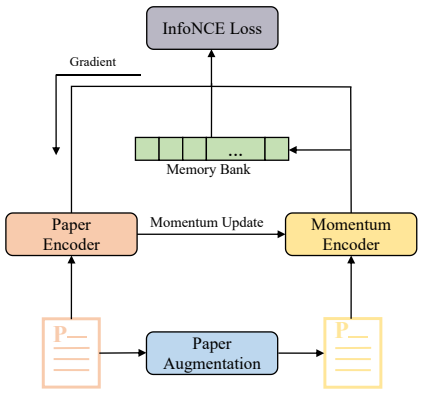
Examination Organization Tree (EOT) describes the hierarchical structure of examination paper. Figure 1 shows an example mathematical paper along with its corresponding EOT. As the figure depicts, an examination paper consists of multiple sections that gather questions of different types separately. In the EOT, choice questions (questions under Section 1) and blank-filling questions (questions under Section 2) are represented as leaf nodes since they do not have subquestions. On the other hand, answer questions (questions under Section 3) are represented as internal nodes, which include a global condition and several subquestions. The subquestions, in turn, are represented as leaf nodes. This tree-based



(a) Architecture of PaperLM with pre-training objectives EOT Node Relationship Prediction(ENRP) and Question Type Prediction(QTP).



(b) Process of converting EOT path to embedding.



(c) Architecture of pre-training objective Paper Contrastive Learning(PCL).

Figure 2: The proposed PaperLM framework. (a)Architecture of PaperLM with pre-training objectives EOT Node Relationship Prediction(ENRP) and Question Type Prediction(QTP). (b)Process of converting EOT path to embedding. (c)Architecture of pre-training objectives Paper Contrastive Learning(PCL).

representation provides insights into the arrangement of questions and the distribution of question types within the examination paper, which are closely related to the properties of the examination paper. Consequently, Leveraging the EOT allows us to determine the significance of each part of the examination paper, thereby enhancing the overall representation of the papers. To enable the model to receive the EOT as input, we define a path expression for each node in the EOT, like *paper/section3/question1/subquestion1*. The texts represent the type of the nodes while the numbers represent the rank of the nodes among their siblings. With the extraction of the EOT path, we can convert it into embedding for subsequent processing steps.

3.2 Knowledge Hierarchy

Each single question in the examination paper has a knowledge concept property, which is selected from a L-level knowledge hierarchy $KH = \{\kappa, \varepsilon\}$. As shown in the right hand of Figure 1, KH is a tree structure, where the vertexes κ are the knowledge concepts, and the edges ε represent the relationship between knowledge concepts. The vertex in higher knowledge level indicates that it is a higher abstraction or description of knowledge (e.g. *Function*), while vertex in lower knowledge level indicates that it is a more fine-grained knowledge of its parent node (e.g. *PowerFunction*). Therefore the knowledge concept of a question is denoted as $k = \{k^1, k^2, ..., k^L\}$, where k^i is the knowledge concept at l-level and k^i is the parent of k^{i+1}. So the knowledge concept set of an examination paper that consists of N questions is denoted as $K = \{k_1, k_2, ..., k_N\}$, where k_i is the knowledge concept of the i-th question in the paper.

3.3 Problem Definition

Given an examination paper P and its corresponding EOT, we aim to represent P with a d_h-dimensional vector r_p, which can be used for several down-stream tasks and benefit the performances. We hope the learned examination paper representation to contain comprehensive information and effectively capture the textual and structural features that are relevant to the properties and holistic features of examination papers.

4 PAPERLM

In this section, we introduce the proposed model PaperLM. We first present the model architecture and then describe the pre-training objectives we specially designed to facilitate the learning of examination paper representations.

4.1 Model Architecture

To take advantage of existing pre-trained models and adapt them to the structure of examination papers, we use BERT as the encoder backbone and introduce a new input embedding: EOT path embedding. Figure 2(a) shows the overview architecture of PaperLM.

4.1.1 ***Text Embedding.*** To match the structure of the EOT, we partition the papers at the granularity of the test item, which is directly aligned with the EOT path embedding. We embed the plain text contained in a test item into a feature vector using the pre-trained BERT model from EduNLP[1]: an open source Python library that provides pre-trained language models specially for educational questions of different subjects. Following BERT, we mark the beginning of a test item sequence with a special [CLS] token and the end with a special [SEP] token, which are both computed by averaging the test item features.

4.1.2 ***EOT Path Embedding.*** Inspired by MarkupLM [18], we incorporate the EOT path information into the model input by converting it into embedding, as shown in Figure 2(b). For the i-th test item x_i in EOT, we split its corresponding path expression by "/" to obtain the node information at each level of the path as a list, $ep_i = [(t_0^i, r_0^i), (t_1^i, r_1^i), ..., (t_d^i, r_d^i)]$, where d denotes the depth of this EOT path, t_j^i and r_j^i denotes the type name of and rank of the EOT path unit on level j for x_i. Note that we assign 1 to r_j^i for units without rank number. Additionally, we apply padding to ep_i to unify their lengths as L for further processing.

For (t_j^i, r_j^i), we input this pair into the j-th type unit embedding table and j-th rank unit embedding table respectively, of which we set the dimensions as d_u. Then we add these two embeddings to get the j-th unit embedding ue_j^i.

$$ue_j^i = TypeUnitEmb_j(t_j^i) + RankUnitEmb_j(r_j^i). \tag{1}$$

We concatenate all the unit embeddings to get the intermediate representation r_i of the complete EOT path for x_i.

$$r_i = [ue_0^i; ue_1^i; ...; ue_L^i]. \tag{2}$$

Then we feed the intermediate representation r_i into an FFN layer to get the final EOT path embedding epe_i.

$$epe_i = W_2[ReLU(W_1 r_i + b_1)] + b_2, \tag{3}$$

$$W_1 \in \mathbb{R}^{4d_h \times Ld_u}, b_1 \in \mathbb{R}^{4d_h}, \tag{4}$$

$$W_2 \in \mathbb{R}^{d_h \times 4d_h}, b_2 \in \mathbb{R}^{d_h}, \tag{5}$$

where d_h denotes the hidden size of PaperLM.

After obtaining the text embedding and EOT path embedding of each test item, we add them up respectively and feed the sum embeddings into the PaperLM encoder. Finally, we use the mean of test item embeddings at the last layer of the PaperLM encoder as the representation vector of the input examination paper r_p.

4.2 Pre-training Objectives

To efficiently capture both the textual and structural information of examination papers, we propose three different pre-training objectives, including EOT Node Relationship Prediction (ENRP), Question Type Prediction (QTP), and Paper Contrastive Learning (PCL). Note that all the pre-training objectives require input from both textual and structural information.

4.2.1 ***EOT Node Relationship Prediction.*** To enhance the model's understanding of the semantics conveyed by the EOT path, we propose the pre-training objective EOT Node Relationship Prediction (ENRP) to explicitly model the relationship between a pair of nodes within the EOT. First, we define a set of node relationships $R = \{self, child, parent, sibling, descendent, ancesotr, others\}$. Then we combine each node in pairs and assign relationship labels according to R. The model is then trained to predict the assigned relationship label for each node pair. We formulate this objective as a multi-classification objective. Given an examination paper with N test items, we consider each pair of test item embeddings at the last layer of the PaperLM encoder, denoted as e_i and e_j, and concatenate them in pairs to get the node pair embedding npe_{ij}.

$$npe_{ij} = [e_i; e_j], 0 \le i, j \le N-1. \tag{6}$$

Then we feed the node pair embedding npe_{ij} into an FFN layer to get the predicted classification result c_{ij} of the node pair$\{i, j\}$.

$$c_{ij} = W_2[ReLU(W_1 npe_{ij})] + b, \tag{7}$$

$$W_1 \in \mathbb{R}^{2d_h \times 2d_h}, W_2 \in \mathbb{R}^{n_c \times 2d_h}, b \in \mathbb{R}^{d_h}, \tag{8}$$

where d_h denotes the hidden size of PaperLM and n_c represent the category number of node relationships. Finally, we adopt the cross-entropy loss function to compute the loss between the classification result c_{ij} and the relationship label r_{ij}.

4.2.2 ***Question Type Prediction.*** Through the ENRP task, the model can explicitly learn the hierarchical structure of examination papers. However, it cannot yet relate the paper structure to the paper properties. As we discussed earlier, the distribution of question type is significantly correlated with paper difficulty. Therefore, we propose another pre-training objective, namely Question Type Prediction (QTP), which aims to enhance the model's ability to distinguish between different question types. As depicted in Figure 2(a), we mask the text embedding of section item while preserving the corresponding EOT path embedding during the pre-training. This is done because the text content of section items often provides clues about the type of questions underneath. We then require the model to predict which type it originally indicated. Now that the model has learned the hierarchical structure of examination papers, it will learn to predict the type of section items by referring to their

[1]https://github.com/bigdata-ustc/EduNLP

childs. In this way, the model will learn the ability to distinguish different types of questions, thereby bridging the gap between the structure and text information.

The Question Type Prediction (QTP) objective is treated as a multi-classification objective. We select the section item embeddings at the last layer of the PaperLM encoder e_{sec} and feed them into an FFN layer to get the predicted classification results of the question type c_{qt}:

$$c_{qt} = W_2[ReLU(W_1 e_{sec})] + b, \tag{9}$$

$$W_1 \in \mathbb{R}^{d_h \times d_h}, W_2 \in \mathbb{R}^{n_c \times d_h}, b \in \mathbb{R}^{d_h}, \tag{10}$$

where d_h denotes the hidden size of PaperLM and n_{cls} represent the category number of question types. Same as ENRP, we adopt the cross-entropy loss to compute the loss between the classification result c_{qt} and the question type label t_q.

4.2.3 ***Paper Contrastive Learning.*** As we discussed earlier, capturing the holistic examination scope is significantly important for tasks like examination paper retrieval. However, previous methods fail to capture such information. Comparatively, contrastive learning has been proved to achieve great success in learning representations [6, 42]. Therefore, we propose the Paper Contrastive Learning (PCL) to model the overall examination scope of papers, which is illustrated in Figure 2(c). We aim to learn comprehensive representations of examination papers by pulling papers with similar knowledge concepts closer than those with less similar knowledge concepts. This allows the model to measure the similarity between examination papers based on their examination scope. Thus it is important to construct suitable positive samples that are similar in examination scope to the original samples.

To this end, we propose the following paper augmentation strategy. For each question in an examination paper, we replace it with another question that has the same knowledge concept and question type as the original question. The question is selected randomly from the pre-training corpus. This replacement is done with probability of p for each question. In addition, following He et al. [9], we introduce a memory bank with a momentum encoder, which allows the inclusion of a large number of negative samples and thereby enhancing the effectiveness of the contrastive learning. We feed the examination paper representation r_p into an MLP to get the intermediate representation q.

We use the InfoNCE [28] loss as the contrastive loss function:

$$L_{PCL} = -log \frac{exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} exp(q \cdot k_i/\tau)}, \tag{11}$$

where k_+ and k_i denote intermediate representations of positive and negative samples respectively. K is the size of the memory bank and τ is a temperature hyper-parameter [38].

5 EXPERIMENTS

In this section, we conduct pre-training of the PaperLM model using real-world mathematical examination papers and then evaluate our pre-trained model on three down-stream tasks to verify the effectiveness of our proposed model.

Table 1: The statistics of the datasets.

Statics	Dataset1	Dataset2	Dataset3	Dataset4
Num. Examination Papers	34,852	5,000	-	2,000
Num. Examination Paper Pairs	-	-	200	-
Avg. Questions per Paper	23.54	16.13	18.64	19.28
EOT Depth	4	4	4	4
Knowledge in Level-1	10	-	-	4
Knowledge in Level-2	39	-	-	12
Knowledge in Level-3	431	-	-	175

Table 2: Average Difficulty of Questions of Different Types.

Question Type	Choice and Blank-filling Questions	Answer Questions
Difficulty	0.34783	0.69408

5.1 Datasets

Four real-world datasets are used in our experiments, which are all collected from an online education system Zhixue[2], which provides various and customized educational applications for high school students in China. We list some important statistics of the datasets in Table 1. Simple descriptions of four datasets are as follows:

Dataset1 contains 34,852 mathematical examination papers of high school level, in which each question is annotated with a three-level knowledge hierarchy. We pre-train PaperLM on Dataset1.

Dataset2 contains 5,000 mathematical examination papers with paper-level difficulty scores. The difficulty score is calculated based on the student performance using Classical Test Theory (CTT) We randomly partition the dataset into training and testing sets with the ratio of 4:1 for the down-stream task paper difficulty estimation.

Dataset3 contains 200 mathematical examination paper pairs with similarity scores labeled by experts, who are expected to consider the similarity of papers' examination scope and quality. We use Dataset3 for the down-stream task examination paper retrieval.

Dataset4 contains 2,000 unit examination papers, which focus on one specific knowledge concept in the knowledge hierarchy and are often used for targeted training purposes. Dataset4 is used for the down-stream task paper clustering where we create two experimental scenarios, each with 4, 12 clusters, corresponding to 4 and 12 frequent knowledge concepts from the first and second levels of the knowledge hierarchy respectively.

5.2 Data Analysis

We additionally investigate students' performance when they encounter different types of questions in Dataset2. The results of the investigation are shown in Table 2, where the difficulty scores of single questions are calculated from the correct rates of students (low correct rates often mean high difficulty scores). Notably, we obtain the observation that students score significantly higher on

[2]https://www.zhixue.com

Table 3: Performance of different tasks.

Tasks	Paper Difficulty Estimation				Examination Paper Retrieval		Paper Clustering	
Datasets	Dataset2				Dataset3		Dataset4	
Metrics	MAE ↓	RMSE ↓	PCC ↑	DOA ↑	NDCG@5 ↑	NDCG@10 ↑	#Cluster=4 NMI ↑	#Cluster=12 NMI ↑
HiAttention	0.2165	0.2512	0.3015	0.4550	-	-	-	-
EOE	0.1778	0.2026	0.3689	0.6190	-	-	-	-
Longformer	0.1704	0.2053	0.3839	0.6257	0.8648	0.9395	0.3084	0.3751
BigBird	0.1697	0.2046	0.5248	0.6806	0.8864	0.9482	0.2987	0.3492
BERT	0.1654	0.2006	0.5436	0.6865	0.8770	0.9449	0.4599	0.3960
MarkupLM	0.1642	0.1985	0.5493	0.6935	0.9274	0.9678	0.6311	0.5166
PaperLM	**0.1507**	**0.1823**	**0.7003**	**0.7462**	**0.9742**	**0.9879**	**0.7062**	**0.6280**

Table 4: Details of all models.

Models	Input Information	Parameter	Pre-trained?
HiAttention	Text+Structure	-	✗
EOE	Text+Structure	-	✗
Longformer	Text	148.66M	✓
BigBird	Text	127.47M	✓
BERT	Text	109.48M	✓
MarkupLM	Text+Structure	135.20M	✓
PaperLM	Text+Structure	33.00M	✓

choice questions and blank-filling questions compared to answer questions. This observation indicates that an examination paper with a higher proportion of answer questions would be more challenging than an examination paper with more choice questions or blank-filling questions. From the data analysis, we confirm the view we mentioned in Section 1 that the question type distribution within an examination paper is strongly correlated with paper properties such as paper difficulty.

5.3 Implementation Details

5.3.1 ***Examination Paper Pre-processing.*** Following [25], we design a syntax parser to extract the structural information from an examination paper and covert it into an EOT. During the process of extracting structural information, the textual content of an examination paper is divided into some test items according to the structure of the EOT. We notice that a very small proportion of papers in the pre-training corpus have significantly more test items than other papers. This could cause some input sequences to be heavily padded to ensure that all sequences are of the same length, therefore slowing down the training speed. This problem is particularly serious in terms of the pre-training objective ENRP, which requires to construct Len^2 node pairs for a sequence of length Len. To address this issue, we set the maximum length of the input sequences to 64 and exclude the papers that exceed the maximum limitation. We use the pre-trained mathematical BERT model from EduNLP to get the feature vectors of test items.

5.3.2 ***PaperLM Setup.*** [3] Our model is implemented by PyTorch. We use BERT as the encoder backbone. Different from the original BERT, we reduce the number of Transformer layers to 4. The size of PaperLM hidden layers is 768. The size of the types and ranks in the EOT Path embedding are 256 and 1,024 respectively, the max depth of EOT path expression is 10, and the dimension for the type-unit and rank-unit embedding (d_u) is 32. The momentum encoder is updated with a momentum term m = 0.999. The size of the memory bank is set to 32,768. The probability of question replacement p is 0.3. The temperature factor τ is set to 0.07. In the pre-training phase, we adopt the AdamW optimizer with a learning rate of 0.0005 and a batch size of 64. We do not initialize our model before pre-training with parameters from pre-trained BERT or any variants. All experiments are conducted with one Tesla V100 GPU.

5.4 Baselines

To demonstrate effectiveness of our proposed model, we compare it with several baselines. Specifically, these methods are:

HiAttention [43] is a traditional method for learning representations of long text sequences, especially documents with hierarchical structure. It adopts a 2-level hierarchical attention mechanism to measure the importance of different parts in the document.

EOE [25] is a multi-layer GRU-based model for learning representation of examination papers with hierarchical structure.

Longformer [1] is a Transformer-based model. It adopts several new attention mechanisms that greatly reduce the time complexity of the model. Thus it can accept much longer sequences than BERT.

BigBird [45] is another Transformer-based model that implements a self-attention mechanism of linear dependency. It outperforms Longformer in some NLP tasks.

BERT [15] is the most popular pre-trained model in NLP tasks. Considering its limitation on input length, we apply BERT to our examination papers at the test item granularity and employ the same test item embedding method as ours.

MarkupLM [18] is a pre-trained language model designed for documents in markup languages, such as web pages. Its architecture is well-suited for processing structured documents, making it applicable to the representation of examination papers. To adapt to the input format of MarkupLM, we convert examination papers to HTML documents and extract the HTML source code as the

[3] Code is available at https://github.com/bigdata-ustc/PaperLM.

model input. We also apply MarkupLM at the test item granularity to enable it to handle lengthy examination papers.

For better illustration, we list the comparison of detailed characteristics of all baselines and our proposed model in Table 4. Note that we only calculate the numbers of parameter for the pre-trained models. The non-pre-trained models: HiAttention and EOE, are GRU-based models, whose numbers of parameter vary depending on the length of the input sequences. As a result, their numbers of parameter are not comparable with those of the other models.

5.5 Evaluation Tasks

We use three typical tasks related to examination paper representation: paper difficulty estimation examination paper retrieval, and paper clustering to evaluate our proposed model.

Paper difficulty estimation. Paper difficulty estimation is a regression task to estimate the difficulty of a given examination paper. We add an MLP head on top of PaperLM as well as other pre-trained model baselines and fine-tune the whole model. For non-pre-trained model baselines, we train it from zero. We conduct this experiment on Dataset2 where we can obtain paper-level difficulty scores. Following [13], we adopt MAE (Mean Absolute Error), RMSE (Root Mean Square Error), PCC (Pearson Correlation Coefficient), and DOA (Degree of Agreement) as our evaluation metrics.

Examination paper retrieval. The main purpose for task examination paper retrieval is to find similar examination papers in large-scale online education systems, which supports various application scenarios such as personalized paper recommendation. The similarity between two arbitrary examination papers are measured by cosine similarity function. We directly use the learned examination paper representations from pre-trained PaperLM without further fine-tuning. Following [25], we adopt NDCG@N (Normalized Discounted Cumulative Gain), the most widely-used evaluation metric for ranking tasks, to evaluate our model. We conduct this experiment on Dataset3.

Paper clustering. To further evaluate our model in the scenario where there is no annotation available, we perform paper clustering on Dataset4. We conduct experiments with two different number of clusters in Dataset4 to evaluate model's ability to distinguish knowledge concepts at different levels. We apply K-means [27] clustering over all the paper representations and use the NMI (Normalized Mutual Information) metric for evaluation.

5.6 Results and Discussion

The results of three tasks are presented in Table 3. Note that HiAttention and EOE are not pre-trained models, which limits their applicability to unsupervised tasks such as examination paper retrieval and paper clustering. Consequently, we don't evaluate their performances on these two down-stream tasks. We discuss our observations from the experiments as follows.

PaperLM outperforms baselines. As Table 3 shows, our proposed model PaperLM consistently achieves better performance than all baselines on paper difficulty estimation, examination paper retrieval, and paper clustering, which demonstrates that PaperLM can effectively capture the domain-specific textual and hierarchical structural features of examination papers, benefiting learning an informative and distinctive representation of examination papers. Notably, PaperLM largely surpasses other methods on tasks examination paper retrieval and paper clustering, which highlights the model's ability to learn distinct representations for different examination papers, making it highly valuable for practical applications. The observation also confirms our model's discriminability in features without fine-tuning.

Structural information is beneficial. The experimental results obviously show that pre-trained models (e.g. PaperLM, Longformer) largely outperform traditional methods (e.g. HiAttention, EOE), confirming the effectiveness of pre-trained language models. Additionally, we obtain the observation that pre-trained models with both textual and structural information (e.g. PaperLM, MarkupLM) outperform pre-trained models utilizing only textual information(e.g. BigBird, BERT), which demonstrates that the structural information of examination papers is essential for learning an informative and distinct representation.

A test item is richer than a single word. We notice that models operating at the test item granularity level (e.g. PaperLM, MarkupLM, BERT) consistently perform better than models operating at the word or character granularity level (e.g. Longformer, BigBird). Therefore, we believe that exploring information from test items can be more beneficial than collecting features from each word when representing an examination paper. In addition, well-designed modeling on feature embedding can also bring a more informative representation.

Lightweight but effective. We also observe from Table 4 that while PaperLM outperforms all baselines, it has the fewest parameters among the pre-trained models. This demonstrates that PaperLM is both lightweight and effective, which is helpful in terms of saving computation resources.

6 MODEL ANALYSIS

6.1 Ablation study

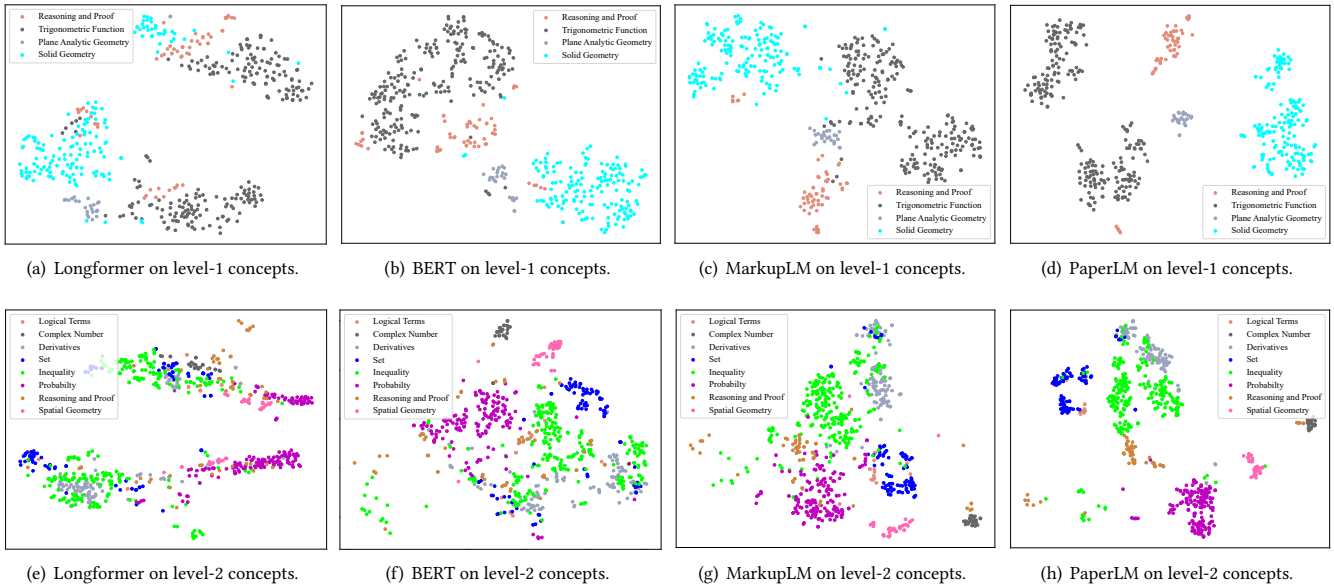
To investigate the effect of each component, we conduct ablation studies and the results are shown in Table 5. We consider two scenarios here. Firstly, we remove the EOT path embedding and two pre-training objectives related to EOT: EOT Node Relationship Prediction and Question Type Prediction. Secondly, we keep the EOT path embedding but remove the pre-training objective Paper Contrastive Learning. We observe that removing any module leads to a performance decrease on the down-stream tasks, which indicates the effectiveness of our designs. Especially, there is a relatively larger performance decrease when we remove EOT path embedding and related pre-training objectives. This highlights the crucial role of structure information in learning an informative representation of examination papers.

6.2 Visualization

As we mentioned before, our model is expected to capture the holistic examination scope of papers, which is represented by the knowledge concepts covered in the examination papers. In order to intuitively assess this capability, we project the representations of examination papers in Dataset4 into 2D space by t-SNE [35]. The projection is also conducted in two experimental scenarios, which represent knowledge concepts at the first and second levels of the knowledge hierarchy. Note that we exclude some categories as they

Table 5: Results of ablation experiments.

Tasks	Paper Difficulty Estimation				Examination Paper Retrieval		Paper Clustering	
Metrics	MAE ↓	RMSE ↓	PCC ↑	DOA ↑	NDCG@5 ↑	NDCG@10 ↑	#Cluster=4 NMI ↑	#Cluster=12 NMI ↑
PaperLM	**0.1507**	**0.1823**	**0.7003**	**0.7462**	**0.9742**	**0.9879**	**0.7062**	**0.6280**
w/o EOT path embedding	0.1589	0.1917	0.6720	0.7337	0.9679	0.9824	0.6876	0.5912
w/o PCL	0.1566	0.1897	0.6971	0.7455	0.9715	0.9868	0.6965	0.6025



(a) Longformer on level-1 concepts. (b) BERT on level-1 concepts. (c) MarkupLM on level-1 concepts. (d) PaperLM on level-1 concepts.

(e) Longformer on level-2 concepts. (f) BERT on level-2 concepts. (g) MarkupLM on level-2 concepts. (h) PaperLM on level-2 concepts.

Figure 3: Different models' visualization results of examination papers on Dataset4.

contain too few sample examination papers. We conduct the visualization on PaperLM along with three pre-trained model baselines and the results are shown in Figure 3, where examination papers in different categories are marked with different colors. From the projection results at both levels, we observe that compared with other models, PaperLM has the capability to cluster samples within the same category more closely together, while effectively separating samples belonging to different categories. This intuitively demonstrates that PaperLM is able to capture the examination scope of examination papers.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a pre-trained model, namely PaperLM, to learn representations of examination papers by integrating both the text and hierarchical structure information of papers. Specifically, we first extracted the Examination Organized Tree from a hierarchical examination paper, and converted its path into embedding as the model input. Then, we proposed three pre-training objectives: EOT Node Relationship Prediction, Question Type Prediction, and Paper Contrastive Learning. These objectives aim to capture the hierarchical structure of examination papers, bridge the gap between text and structure, and model the holistic features of papers respectively. We evaluated our proposed model on real-world datasets with three down-stream tasks: paper difficulty estimation, examination paper retrieval, and paper clustering. The experimental results demonstrated the effectiveness of our proposed model.

For future research, we will investigate the fusion of information from additional modalities like examination paper images, to further enhance the representation learning capabilities of our model. Meanwhile, we hope to explore other meaningful applications of paper representation, such as intelligent paper generation.

ACKNOWLEDGMENTS

This research was partial supported by grants from the National Key Research and Development Program of China (Grant No. 2021YFF0901003) and the National Natural Science Foundation of China (Grant No. U20A20229), the University Synergy Innovation Program of Anhui Province (No. GXXT-2022-042), and the Laboratory of Cognitive Intelligence (iED2022-002).

REFERENCES

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[2] Lidong Bing, Bai Sun, Shan Jiang, Yan Zhang, and Wai Lam. 2010. Learning ontology resolution for document representation and its applications in text mining. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1713–1716.

[3] Liqun Chen, Dong Wang, Zhe Gan, Jingjing Liu, Ricardo Henao, and Lawrence Carin. 2021. Wasserstein contrastive representation distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16296–16305.

[4] Xiang Deng, Prashant Shiralkar, Colin Lockard, Binxuan Huang, and Huan Sun. 2022. DOM-LM: Learning Generalizable Representations for HTML Documents. *arXiv preprint arXiv:2201.10608* (2022).

[5] Ronald L Flaugher, Richard S Melton, and Charles T Myers. 1968. Item rearrangement under typical test conditions. *Educational and Psychological Measurement* 28, 3 (1968), 813–824.

[6] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 6894–6910.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems* 33 (2020), 21271–21284.

[8] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. 2021. Contrastive embedding for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2371–2381.

[9] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[10] Ye Huang, Wei Huang, Shiwei Tong, Zhenya Huang, Qi Liu, Enhong Chen, Jianhui Ma, Liang Wan, and Shijin Wang. 2021. STAN: Adversarial Network for Cross-domain Question Difficulty Prediction. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 220–229.

[11] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4083–4091.

[12] Zhenya Huang, Xin Lin, Hao Wang, Qi Liu, Enhong Chen, Jianhui Ma, Yu Su, and Wei Tong. 2021. Disenqnet: Disentangled representation learning for educational questions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 696–704.

[13] Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question Difficulty Prediction for READING Problems in Standard Tests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[14] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE, 1–6.

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems* 33 (2020), 18661–18673.

[17] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. PMLR, 1188–1196.

[18] Junlong Li, Yiheng Xu, Lei Cui, and Furu Wei. 2022. MarkupLM: Pre-training of Text and Markup Language for Visually Rich Document Understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 6078–6087.

[19] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang, and Furu Wei. 2022. Dit: Self-supervised pre-training for document image transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3530–3539.

[20] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021. Selfdoc: Self-supervised document representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5652–5660.

[21] Qi Liu, Zai Huang, Zhenya Huang, Chuanren Liu, Enhong Chen, Yu Su, and Guoping Hu. 2018. Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1821–1830.

[22] Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Hui Xiong, Yu Su, and Guoping Hu. 2021. EKT: Exercise-Aware Knowledge Tracing for Student Performance Prediction. *IEEE Transactions on Knowledge and Data Engineering* 33, 1 (2021), 100–115.

[23] Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019. Graph Convolution for Multimodal Information Extraction from Visually Rich Documents. In *Proceedings of NAACL-HLT*. 32–39.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[25] Yixiao Ma, Shiwei Tong, Ye Liu, Likang Wu, Qi Liu, Enhong Chen, Wei Tong, and Zi Yan. 2021. Enhanced Representation Learning for Examination Papers with Hierarchical Document Structure. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2156–2160.

[26] Katherine MacNicol. 1956. Effects of varying order of item difficulty in an unspeeded verbal test. *Unpublished manuscript, Educational Testing Service, Princeton, NJ* (1956).

[27] J MacQueen. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*. University of California Los Angeles LA USA, 281–297.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[29] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[30] Seongsik Park, Dongkeun Yoon, and Harksoo Kim. 2022. Improving Graph-based Document-Level Relation Extraction Model with Novel Graph Structure. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4379–4383.

[31] Shuai Peng, Ke Yuan, Liangcai Gao, and Zhi Tang. 2021. MathBERT: A Pre-Trained Model for Mathematical Formula Understanding. *arXiv e-prints* (2021), arXiv–2105.

[32] Yuanyuan Qi, Jiayue Zhang, Yansong Liu, Weiran Xu, and Jun Guo. 2020. CGTR: Convolution Graph Topology Representation for Document Ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2173–2176.

[33] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[34] Fu Sun, Feng-Lin Li, Ruize Wang, Qianglong Chen, Xingyi Cheng, and Ji Zhang. 2021. K-AID: Enhancing Pre-trained Language Models with Domain Knowledge for Question Answering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4125–4134.

[35] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[36] Feng Wang and Huaping Liu. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2495–2504.

[37] Suhang Wang, Charu Aggarwal, and Huan Liu. 2019. Beyond word2vec: Distance-graph tensor factorization for word and document embeddings. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1041–1050.

[38] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3733–3742.

[39] Chenhao Xie, Wenhao Huang, Jiaqing Liang, Chengsong Huang, and Yanghua Xiao. 2021. Webke: Knowledge extraction from semi-structured web with pre-trained markup language model. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2211–2220.

[40] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1192–1200.

[41] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740* (2020).

[42] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5065–5075.

[43] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 1480–1489.

[44] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 7370–7377.

[45] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems* 33 (2020), 17283–17297.