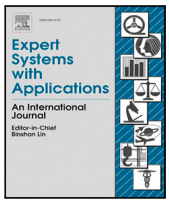
Expert Systems With Applications 278 (2025) 127236

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Adaptive ensemble learning for efficient keyphrase extraction: Diagnosis, aggregation, and distillation

Kai Zhang [a,c], Hongbo Gang [b,c], Feng Hu [b,c], Runlong Yu [d], Qi Liu [b,c,*]

[a] School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
[b] School of Data Science, University of Science and Technology of China, Hefei, China
[c] State Key Laboratory of Cognitive Intelligence, Hefei, China
[d] Department of Computer Science, University of Pittsburgh, Pittsburgh, USA

ARTICLE INFO

Keywords:
Natural language processing
Cognitive diagnosis
Adaptive learning
Keyphrase extraction
Knowledge distillation

ABSTRACT

Keyphrase extraction (KE) refers to the process of identifying words or phrases that signify the primary themes of a document. Although keyphrase extraction is important in many downstream applications, including scientific document indexing, search, and question answering, the challenge lies in executing this extraction both adaptively and effectively. To this end, we propose a novel ***D**istillation-based **A**daptive **E**nsemble **L**earning **(DAEL)*** method specifically designed for efficient keyphrase extraction, encompassing diagnosis, aggregation, and distillation processes. Specifically, we initiate with a *Cognitive Diagnosis Module (CDM)* to evaluate the diverse capabilities of individual KE models. Following this, an *Adaptive Aggregation Module (AAM)* is employed to create a weight distribution uniquely suited to each data instance. The process concludes with a *Knowledge Distillation Module (KDM)* to distill the superior performance of the ensemble model into a single model, thereby refining its efficiency and reducing computational cost. Extensive testing on real-world datasets highlights the superior performance of the proposed model. In comparison with leading-edge methods, our approach notably excels in processing text with complex structures or significant noise, marking a substantial advancement in KE effectiveness.

1. Introduction

Keyphrase Extraction (KE) refers to the process of identifying the keyphrases (e.g., *"graph neural networks"*, *"intelligent analysis"*, and *"sentiment analysis"* as exemplified in Fig. 1) from a source document. This task is a crucial component of Natural Language Processing (NLP) and boasts a broad spectrum of applications, such as condensing the documents into summaries (Papagiannopoulou & Tsoumakas, 2020), enhancing retrieval of information (Hasan & Ng, 2014), and creating the models for various topics (Hasan & Ng, 2014; Song, Feng, & Jing, 2023; Zhao et al., 2017).

In the task of keyphrase extraction, a central challenge lies in identifying the crucial information, largely because potential keyphrases often display multiple and hidden relationships (Song, Feng, & Jing, 2022a; Sun, Qiu, Zheng, Wang, & Zhang, 2020). Most current keyphrase extraction models, including those highlighted in Bennani-Smires, Musat, Hossmann, Baeriswyl, and Jaggi (2018), Liang, Wu, Li, and Li (2021), Song, Liu, Feng and Jing (2023), mainly use the pre-trained word embeddings. These models generally follow a two-step process: generating candidate keyphrases and estimating their significance (Song, Jing, & Xiao, 2021). The first step involves using heuristic methods to select a set of words or phrases from the source document as potential keyphrases (Liang et al., 2021). The second step encompasses two main elements: representing semantics of the text and calculating the importance of these keyphrases.

Recently, the landscape of natural language processing has been transformed with the introduction and widespread adoption of advanced, pre-trained transformer models such as BERT (Kenton & Toutanova, 2019) and RoBERTa (Liu et al., 2019). These Pre-trained Language Models (PLMs) have not only revolutionized the modeling method to various NLP tasks but have also proven to be exceptionally effective. As a key part of NLP workflows, they serve as sophisticated embedding layers, generating contextualized representations for a multitude of applications. This evolution in text representation technology has notably enhanced the research field of keyphrase extraction. The emergence of embedding-based keyphrase extraction methods (Song, Feng, & Jing, 2022b; Song, Jiang, Liu, Shi & Jing, 2023; Zhang et al., 2022), capitalizing on the strengths of

* Corresponding author at: School of Data Science, University of Science and Technology of China, Hefei, China.
E-mail addresses: kkzhang08@ustc.edu.cn (K. Zhang), elysiumghb@mail.ustc.edu.cn (H. Gang), fenghufh3@mail.ustc.edu.cn (F. Hu), ruy59@pitt.edu (R. Yu), qiliuql@ustc.edu.cn (Q. Liu).

https://doi.org/10.1016/j.eswa.2025.127236
Received 27 October 2024; Received in revised form 5 March 2025; Accepted 9 March 2025
Available online 25 March 2025

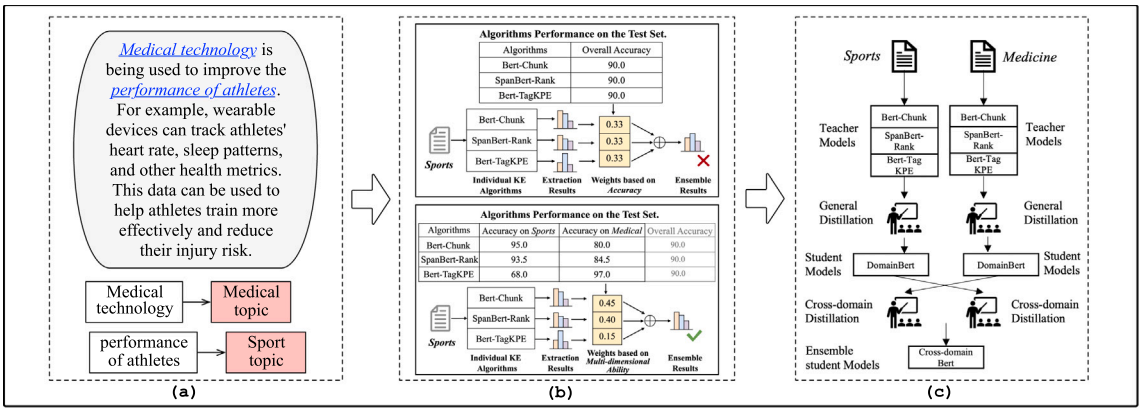


Fig. 1. Part (a) shows the traditional ensemble strategy based on Accuracy. Since the three methods perform consistently across the whole dataset, they are aggregated equally when encountering new data. However, part (b) shows that there are differences among the methods from a fine-grained perspective. When dealing with Sports news, more emphasis should be placed on the two methods that are more capable in Sports. Part (c) is the knowledge distillation process, which can reduce model parameters while maintaining the effect of the integrated model.

these PLMs, represents a notable breakthrough. Additionally, their ability to efficiently and accurately identify critical keyphrases in large volumes of text highlights the remarkable progress of current PLMs technologies. In summary, these methods have demonstrated outstanding results, surpassing the previous approaches, and setting new benchmarks for the state-of-the-art performance in keyphrase extraction. With the rise of Large Language Models (LLMs), there has been a notable exhibition of their strong zero-shot learning capabilities (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022). Recent work (Song, Geng et al., 2023) and our experiments on several open source LLMs have demonstrated the effectiveness of prompt-based approaches using ChatGPT and ChatGLM (GLM et al., 2024), highlighting the potential of LLMs for Keyphrase Extraction.

Nevertheless, despite the significant advancements made by these methods, keyphrase extraction still faces many challenges that necessitate further exploration and resolution. Specifically, the previous methods primarily concentrate on aligning candidate keyphrases and their corresponding document representations within a unified semantic space. However, they often overlook the key aspects such as the extractability, adaptability, and efficiency of keyphrase extraction (KE) models. This narrow focus results in certain shortcomings. For instance, as illustrated in Fig. 1(a), the text document titled *"Medical technology is being used to improve the performance of athletes. For example,...,reduce their injury risk"*. includes some keyphrases: *"medical technology"* and *"performance of athletes"*. However, due to the text's complexity, such as long sentences, or the presence of high noise levels, like text from diverse domains, previous models struggle to effectively capture the various features in the text, leading to limited performance.

To address these limitations, a practical solution involves leveraging ensemble learning methods. These methods aggregate various keyphrase extraction models, aiming to enhance the overall effectiveness of keyphrase extraction in more complex scenarios. Regrettably, traditional methods in keyphrase extraction often fail to yield satisfactory results in practice, sometimes even adversely affecting the process. Specifically, in the KE task, there are two notable challenges in ensemble learning that demand attention and solutions.

- First, the traditional ensemble approach tends to overlook the nuanced, multi-dimensional capabilities of individual Knowledge Extraction (KE) models, focusing instead solely on their overall performance as measured by a single metric like *"Accuracy"*. This limited perspective can reduce the effectiveness of the ensemble. For instance, as shown in Fig. 1(b), Bert-Chunk and SpanBert-Rank excel in the *"sports"* domain, yet fall behind Bert-TagKPE in the *"medical"* domain. Consequently, when dealing with new *"sports"* data, it is advisable to prioritize Bert-Chunk and SpanBert-Rank, rather than treating all three methods equally during the model aggregation process.
- Traditional ensemble methods, while effective in improving performance, often result in increasing model size and the computational complexity. This strategy of seeking performance gains at the cost of heightened model intricacy may compromise research fairness. Besides, the increased complexity may not always justify the performance improvements, especially when considering resource constraints and practical applicability. Hence, it is crucial to strike a balance between resource utilization and performance during model integration, maintaining efficiency and effectiveness.

The observation indicates that traditional ensemble methods lack in two key areas: first, in the adaptive ensemble of the multi-dimensional strengths of individual models, and second, in effectively managing the efficiency of the overall ensemble model. Thus, we argue that the previous "*ensemble pattern*" can be further explored to improve the keyphrase extraction without significantly consuming computing resources. In this work, we introduce a pioneering approach known as ***Distillation-based Adaptive Ensemble Learning (DAEL)***, specifically designed for efficient keyphrase extraction. This approach integrates diagnosis, aggregation, and distillation processes. Initially, a *Cognitive Diagnosis Module (CDM)* assesses the diverse capabilities of individual KE models. Subsequently, an *Adaptive Aggregation Module (AAM)* is employed to create a weight distribution uniquely suited to each data instance. Finally, a *Knowledge Distillation Module (KDM)* distills the enhanced performance of the ensemble model into a singular model, thereby streamlining efficiency and minimizing computational demands. Our extensive testing on real-world datasets demonstrates the model's superiority, particularly in handling texts with complex structures or substantial noise, thereby significantly advancing KE effectiveness compared to leading methods.

The remainder of this paper is organized as follows. Section 2 reviews related work in keyphrase extraction, cognitive diagnosis and knowledge distillation. In Section 3, we provide foundational concepts and methodologies that underpin our research. Section 4 introduces the proposed DAEL model and its theoretical foundations. In Section 5, we describe our implementation details, experimental setup, experimental results and detailed analysis. Finally, Section 6 concludes the paper, while Section 7 discusses the advantages of our diagnostic approach and future research directions.

2. Related works

Keyphrase Extraction. With the rise of information extraction, keyphrase extraction algorithms have become a hot research topic.

Keyphrase extraction aims at selecting a set of phrases from a document that could summarize the main topics discussed in the document (Hasan & Ng, 2014). Keyphrase extraction algorithms are mainly divided into supervised and unsupervised methods. Specifically, unsupervised methods mainly use different features of the document such as topic features, phrase frequency, length features, context features, and so on. Campos et al. (2018) proposed YAKE using 5 features including case, word position, word frequency, word contextuality, and sentence difference. Besides, graph-based methods are an effective class of unsupervised keyphrase extraction methods. Among them, inspired by PageRank (Page, Brin, Motwani, & Winograd, 1999), Mihalcea and Tarau (2004) proposed TextRank, which abstracts a document as a graph, where the nodes represent phrases and edges represent the relationship between phrases

Subsequently, considerable efforts were made to enhance the document graph by incorporating additional information. Among these efforts, Singlerank (Wan & Xiao, 2008) incorporated phrase co-occurrence information as weights on edges based on TextRank model. Bougouin, Boudin, and Daille (2013) introduced TopicRank, a model that replaced keyphrases with candidate words obtained from clustered topics. Multipartite (Boudin, 2018) combined topical information with a multipartite graph structure. PositionRank (Florescu & Caragea, 2017) incorporated positional information of phrase into a biased weighted PageRank. Liu, Huang, Zheng, and Sun (2010) proposed Topical PageRank (TPR) which used LDA to obtain the topic distribution of each word. More recently, pre-trained language models have been used for keyphrase extraction. SIFRank (Sun et al., 2020) combined the autoregressive pre-trained ELMo to compute phrase and document embeddings. It then computed cosine similarity to select the keyphrase. AttentionRank (Ding & Luo, 2021) extracted attention weights from BERT and computes self and cross attention to determine the relevance between phrases and documents. Compared with these unsupervised methods, Xiong, Hu, Xiong, Campos, and Overwijk (2019) presented BLING-KPE and took the task as an n-gram level keyphrase chunking task. Sun, Liu, Xiong, Liu, and Bao (2021) considered both of the phraseness and informativeness when extracting keyphrases in a multi-task training architecture. Kong et al. (2023) proposed the PromptRank, an encoder–decoder architecture that feeds documents into the encoder and calculates the probability of generating candidates using a designed prompt in the decoder. To take this further, Large language models, including ChatGPT and numerous open-source LLMs, have demonstrated strong performance across various downstream natural language processing tasks, attributed to their zero-shot learning capabilities. Song, Geng et al. (2023) investigated a prompt-based method that directly combines instructions and documentation as inputs, prompting ChatGPT and ChatGLM (GLM et al., 2024) for Keyphrase Extraction.

Cognitive Diagnosis. Cognitive Diagnosis (CD) is a fundamental task in many real-world scenarios (e.g., business (Liu, Yang, Gao, Li, & Liu, 2023) and education (Gao et al., 2021, 2023; Wang et al., 2020)). The main goal of Cognitive Diagnosis is to measure learners' proficiency profiles of abilities to finish specific tasks from their observed behaviors (Wang et al., 2020). For instance in education, it can be used to infer student (as *learner*) knowledge proficiency (as *ability*) by fully exploiting their responses of answering each exercise (as *task*). Most of the existing cognitive diagnosis models (CDMs) (De La Torre, 2009; Gao et al., 2021; Lord, 1952) are well designed from psychometric theories of human measurement. Among them, item response theory (IRT) (Lord, 1952) is the most established CDMs which assumes the probability of the learner s_i correctly finishing a task e_j, i.e., $r_{ij} = 1$, increases with learner ability θ_i while decreasing with task difficulty β_j. Among them, the user ability and task difficulty are trainable unidimensional parameters (Liu, 2021). A typical formulation of IRT is $P(r_{ij} = 1) = sigmoid((\theta_i - \beta_j) \cdot a_j)$, where a_j is an optional task discrimination item. Recently, some works extended the previous basic models to capture the more complex relationships among users, tasks, and abilities. The typical method is NeuralCD (Wang et al., 2020) which introduced neural networks $F(\cdot)$ to model high-level interaction between learners/abilities and tasks, i.e., $P(r_{ij} = 1) = F(\theta_i - \beta_j)$. Inspired by the psychometric theories from human measurement, the multi-dimensional evaluation of KE models can also benefit from the fine-grained assessment of human learning performance.

Ensemble Learning. Ensemble Learning (EL) can fuse the knowledge of the individual models together to achieve competitive performance via voting schemes based on some learned features, which is widely used in machine learning tasks (Dong, Yu, Cao, Shi, & Ma, 2020; Papagiannopoulou & Tsoumakas, 2020). Traditional voting schemes include unweighted averaging and the weighted voting. Among them, unweighted averaging of the outputs of the base learners in an ensemble is the most followed approach for fusing the outputs (Ganaie, Hu, Malik, Tanveer, & Suganthan, 2022). It considers the output results of each learner equally but ignores the differences between learners. On the other hand, weighted voting (Ganaie et al., 2022) and Adaptive Ensemble strategy for KE (AEKE) (Jin et al., 2023) methods tend to assign different weights to different learners based on their unidimensional ability. Such ability is often assessed by a single traditional metric on the history datasets. But the weights are constant during model aggregation. In ensemble strategies of keyphrase extraction, mainstream methods employed unweighted averaging and weighted voting methods to aggregate individual KE models.

However, the above methods still suffered from relying on the unidimensional ability (e.g., *Accuracy*, *Precision*) of individual KE models to achieve aggregation, resulting in limited performance in the ensemble. To solve that, we develop an adaptive ensemble strategy for keyphrase extraction from the perspective of multi-dimensional abilities.

Knowledge Distillation. Knowledge Distillation (KD), originally conceptualized by Hinton, Vinyals, and Dean (2015), is a process aimed at transferring knowledge from a larger, well-trained "teacher" model to a smaller, yet comparably effective "student" model. The standard approach in KD involves aligning the output distributions of both teacher and student models by minimizing the Kullback–Leibler divergence loss, which often involves a constant temperature hyperparameter.

To improve the efficacy of distillation, diverse knowledge transfer strategies have been explored. These strategies generally fall into three main categories: logit-based methods, as demonstrated in works by Chen, Mei, Wang, Feng, and Chen (2020), Zhao, Cui, Song, Qiu, and Liang (2022); representation-based methods, exemplified by Chen, Liu, Zhao, and Jia (2021); and relationship-based methods, as seen in studies by Park, Kim, Lu, and Cho (2019), Yim, Joo, Bae, and Kim (2017). Recent innovations, such as MKD (Liu, Liu, Li, & Liu, 2022), advocate for adapting the temperature via meta-learning using an additional validation set. Although this method shows promise, especially when combined with robust data augmentation, its effectiveness may be limited with augmentation or when integrated into other knowledge distillation techniques.

However, in the field of knowledge extraction, the application of knowledge distillation technology to improve the performance of ensemble model remains underexplored. Thus, our proposed DAEL (Distillation-based Adaptive Ensemble Learning) method focuses on examining the influence and utility of knowledge distillation technology within the integration process of ensemble models.

3. Preliminary

3.1. Cognitive diagnosis for KE

Building upon the NeuralCD (Wang et al., 2020) which is a cognitive diagnostic model, we introduce the definition of the cognitive diagnosis problem for keyphrase extraction algorithms. First, we denote the algorithms to be evaluated as learners and the NeuralCD as diagnosers.

Then, we evaluate the multi-dimensional abilities of learners on different skills, which are used to describe how well an algorithm performs on a particular category of samples. Besides, in our work, since the topic of documents contains the main information and represents the specific textual features of keyphrase (Meng, Wang, Yuan, Zhou, & He, 2022), we take the topics of documents as skills. For instance, topics on *Sports* and *Medical* convey a totally different message. Thus, we designate each document topic as a unique skill, aligning one topic with one specific skill.

In designing our diagnoser, we consider a well-trained learner set $S = \{s_1, \ldots, s_N\}$, a sample set $E = \{e_1, \ldots, e_M\}$ which is the dataset in our task, and a skill (topic) set $C = \{c_1, \ldots, c_P\}$. N and M denote the number of learners to be aggregated and samples in the dataset. P denotes the number of skills as a hyper-parameter in our task. Then the learner's output results on each sample as response logs R, which are denoted as a set of triplet (s, e, r_{ij}), where $s \in S$, $e \in E$ and r_{ij} is the score that learner i got on sample j. The top 5 results of keyphrase extraction are transferred to a score (0 or 1). We denote $r_{ij} = 1$ if learner i predicts more than one keyphrase correctly and $r_{ij} = 0$ otherwise. Meanwhile, an explicitly pre-defined sample-skill relevancy matrix Q should also be given. $Q = \{Q_{ij}\}_{M \times P}$, where $Q_{ij} = 1$ if sample e_i is related to skill p_j and $Q_{ij} = 0$ otherwise. Given the learner-sample response matrix R and the sample-skill matrix relevancy Q, we could estimate the multi-dimensional abilities of different learners on different skills through the diagnoser.

3.2. Adaptive aggregation for KE

Fig. 1(a) shows a critical limitation in traditional ensemble strategies for keyphrase extraction algorithms: their singular focus on a single metric (e.g., *Accuracy*), overlooking the nuances of model's multi-dimensional abilities. To address this gap, we utilizes the cognitive diagnosis results to inform the development of adaptive ensemble strategies. The method starts with the cognitive diagnosis analysis, yielding insights into each model's multi-dimensional capabilities and the specific characteristics (e.g., the difficulty, discrimination, and topic relevance) of the data set. Armed with this information, we then confront a new document, denoted as n, with a bespoke aggregation strategy. This strategy dynamically adjusts weights based on the diagnostic outcomes, taking into account the multi-dimensional abilities of the algorithms and the specific attributes of the data, including difficulty, discrimination, and topic.

In this work, multi-dimensional abilities represent the distinct competencies of the keyphrase extraction model, while the difficulty, discrimination, and topic characteristics pertain to the data samples. Our overarching objective is to forge a coherent link between the diagnostic results and the voting weights w assigned to each algorithm. This adaptive weighting strategy is designed to enhance ensemble performance on each new document by tailoring the algorithmic response to the requirements and nuances of the data at hand.

3.3. Knowledge distillation for KE

In the realm of knowledge distillation (KD), as established by Hinton et al. (2015), the foundational concept involves training a "student" model to mimic the output distribution of a "teacher" model. This approach is based on the premise that the "teacher" model's output distribution offers a more informative and effective learning signal to the "student" model than the gold standard labels alone. Building upon this concept, (Clark, Luong, Khandelwal, Manning, & Le, 2019) expanded the utility of KD by integrating it with Multi-Task Learning (MTL). Their research demonstrated that applying knowledge distillation within an ensemble framework, particularly when learning from multiple related tasks, can significantly enhance the performance of models. In this work, we try to use the integrated KE model as a teacher model, and use the knowledge distillation method to distill a single KE model, and keep its effect at a high level.

3.4. Problem definition

Given the multi-dimensional abilities of KE algorithms and features of the new document, our goal is to design an adaptive ensemble strategy to adjust the aggregation weights to improve the keyphrase extraction. Specifically, given the learner-sample response matrix R and the sample-skill matrix relevancy Q, the Cognitive Diagnosis Module utilizes NeuralCD to get multi-dimensional abilities of the learner h_a, the difficulty h_d and discrimination h_d^{disc} of the sample. It defines a probability $y = f(h_a, h_d, h_d^{disc})$ and uses cross entropy loss L as the objective of the diagnosis process. The adaptive weighting strategy takes L and (h_a, h_d, h_d^{disc}) as input to determine a new sample's distribution across various implicit topics c_n. It then embeds the token sequence of a new sample, denoted as D^w into e_d to compare with the representations of the samples entering in the diagnosis. Subsequently, it calculates the average difficulty and discrimination of the K closest samples to define the new sample's value d_n and $disc_n$. Finally, the weight w is determined based on $(h_a, c_n, d_n, disc_n)$.

4. Methodology

4.1. Overview

Since previous research either ignore the adaptability of the ensemble model or ignore the efficiency of the ensemble model. The goal of our model is to design a better adaptivity and efficient ensemble model based on the Cognitive Diagnosis and knowledge distillation technique. The overall model architecture is illustrated in Fig. 2. Specifically, the proposed DAEL model mainly contains three components: 1) Cognitive Diagnosis Module (CDM): aiming to achieve fine-grained evaluation for keyphrase extraction models on multi-dimensional ability; 2) Adaptive Aggregation Module (AAM): aggregating KE models through adaptive strategies after diagnosing capabilities of them; 3) Knowledge Distillation Module (KDM): distilling the integrated model into one model, so that the distilled KE model can reduce the number of parameters while maintaining the performance.

4.2. Cognitive diagnosis module

Following previous cognitive diagnosis research, three important elements need to be considered: learner factors, sample factors, and the interaction function between them (Wang et al., 2020). Inspired by the NeuralCD (Wang et al., 2020) model, the architecture of our diagnostic framework is shown in Fig. 2. Specifically, we use one-hot vectors of the corresponding learner and text representation for the sample as input. Then, we learn the interaction among the factors and characterize the learner and the sample. Finally, the goal of our model is to predict the performance of the algorithms on samples in the dataset.

Learner Factors. In the task, we only focus on the ability on the different skills. Therefore, each learner is represented with a one-hot vector $s_z \in \{0, 1\}^{1 \times N}$ as input, where N denotes the number of learners to be evaluated.

Sample Factors. Previous diagnostic works choose an independent one-hot vector to represent the items (e.g., data items). They represent sample e_d input as one-hot vector $e_d \in \{0, 1\}^{1 \times M}$, where M denotes the number of samples.

Skill Factors. We want to make the topics as skills, as topic information is valuable in keyphrase extraction tasks. However, the published datasets do not contain topic labels for documents. To this end, in this paper, we employ the LDA (Blei, Ng, & Jordan, 2003) to obtain the topic labels by unsupervised clustering of the documents. Especially, LDA has better interpretability and the topical tokens for the clusters can be used as the explicit description for skills, which is great of importance for Cognitive Diagnosis. After clustering the documents into P topics by LDA, we can obtain the sample-skill matrix $Q \in \{0, 1\}^{M \times P}$. By this method, the topic label of each sample will be obtained. With

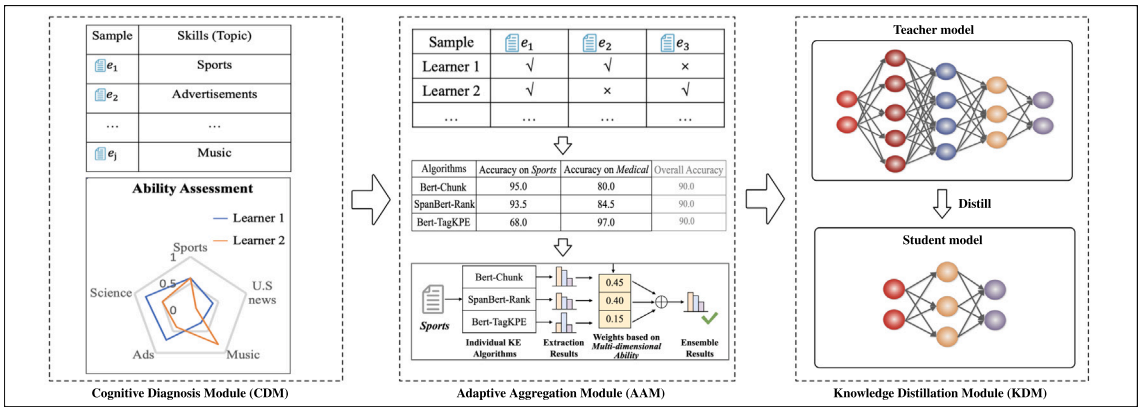


Fig. 2. The model's overall framework, DAEL, is divided into three key components: 1) the Cognitive Diagnosis Module (CDM), which conducts detailed assessments of keyphrase extraction models across various dimensions; 2) the Adaptive Aggregation Module (AAM), which combines KE models using tailored strategies based on their assessed capabilities; and 3) the Knowledge Distillation Module (KDM), which condenses the collective insights from the ensemble into a single, more efficient model.

Table 1
Several important mathematical notations.

Notation	Description
$\mathcal{S}$	The set of learners.
$\mathcal{E}$	The set of data samples.
$\mathcal{C}$	The set of skills.
$\mathcal{Q}$	The sample-skill mappings of samples.
$\mathcal{R}$	The set of response from learners to samples.
h_a	The ability of learners.
h_d	The difficulty of samples.
h_d^{disc}	The discrimination of samples.

the representation of learners, samples, and skills (i.e., s_z, e_d, and c_k respectively), we will eventually predict the performance of the learners on the samples based on the existing response logs. Previous works have achieved the generality of diagnosis frameworks in education. The core of the cognitive diagnostic layer lies in the representation and interaction of model ability and sample factors. Following these works, we have designed a simple and universal diagnostic layer.

Following previous NeuralCD method, with the model we can get multi-dimensional abilities of the learner h_a, the difficulty h_d and discrimination h_d^{disc} of the sample. Note that, detailed explanations of all symbols are provided in Table 1. Among them, h_a indicates the ability of the learner to process samples on different topics. The h_d represents the degree of difficulty the learner to solving the problem. Besides, the h_d^{disc} indicates the capability of samples to differentiate the proficiencies of learners. Samples with low discrimination mean that of low quality: they tend to have annotation errors or do not make sense.

Prediction. With the representation of learners, samples, and skills, we will eventually predict the performance of learners on the samples based on the existing response logs. Specifically, we exploit neural networks to model the relationship between learner ability factor h_a and skill difficulty factor h_d. The probability Y is defined as the ability compared with the sample in the covered topic as:

$$Y = (h_a - h_d) \times h_d^{disc}. \tag{1}$$

Then, we use the full connection layers F to predict the score of learner z on the sample d (i.e., the score y):

$$y = \sigma(F(Y)). \tag{2}$$

Finally, the whole objective of the diagnoser is defined with the cross entropy loss function:

$$\mathcal{L} = -\sum_i (r_i \log y_i + (1 - r_i) \log(1 - y_i)), \tag{3}$$

where r is the true score. Based on Eq. (3), we can get the multi-dimensional abilities of the KE models.

4.3. Adaptive aggregation module

Utilizing cognitive diagnosis module, we are able to discern the multi-dimensional capabilities of various keyphrase extraction models. Drawing on these diagnostic insights, we have developed an adaptive aggregation module (AAM) to more effectively amalgamate the outcomes of each extraction algorithm when confronted with new test samples.

The Input of AAM. The AAM receives inputs comprising both the capabilities of single KE models and the features of the new sample. The abilities of KE models, ascertained from the prior diagnostic module, reflect multi-dimensional proficiency across diverse KE topics. The features of the new sample encompass details about the topic, difficulty, and distinctiveness. Specifically, the topic information correlates with the evaluated abilities of the algorithms, ensuring a tailored approach. The difficulty and distinctiveness of the sample provide insights into the models' implicit adeptness in addressing such challenges. Together, these three attributes (i.e., topic, difficulty, and distinctiveness) comprehensively represent the information of new sample, equipping AAM with a nuanced understanding for effective aggregation.

In summary, drawing from the topic model obtained in Section 4.2, when a new sample is introduced, the distribution across various implicit topics is determined, represented as c_n. Each dimension of c_n signifies the probability of the sample's association with a specific implicit topic. However, it is vital to note that since unseen samples are not utilized as inputs in the diagnostic module, direct assessments of their difficulty and discrimination are not readily available. To this end, we design a non-parametric module to predict the difficulty and discrimination. Specifically, as there is a close relationship between original texts and the factors of samples including the difficulty and discrimination, we choose to predict difficulty and discrimination based on semantic K-nearest neighbor (Peterson, 2009) methods. Here, given the token sequence of original texts of keyphrase extraction samples $D^w = \{d_1^w, d_2^w, \ldots, d_{n_w}^w\}$, we map each word of D^w into word embedding by BERT (Devlin, Chang, Lee, & Toutanova, 2018), and get the document embedding by applying mean-pooling, where n_w is the length of the word sequence. We use the document embedding e_d as input representation for the new sample:

$$e_d = MeanPool(BERT([d_1^w, d_2^w, \ldots, d_{n_w}^w])). \tag{4}$$

Then, we match and retrieve the textual representations of the new samples with the representations of the samples entering in the diagnosis and find the K closest samples. These samples are able to get the difficulty $\{d_1, \ldots, d_k\}$ and discrimination $\{disc_1, \ldots, disc_k\}$ by diagnosis. Finally, we average the difficulty and discrimination retrieved as the difficulty d_n and discrimination $disc_n$ of the new sample.

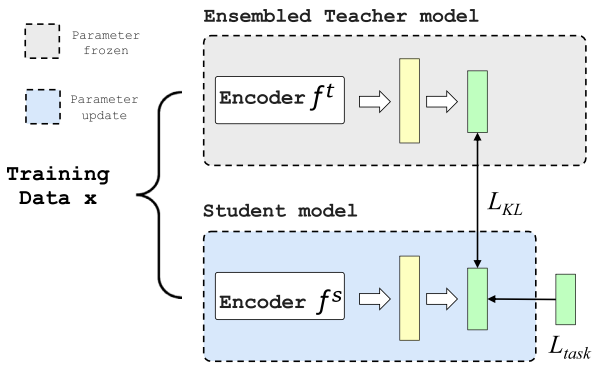


Fig. 3. Knowledge Distillation (KD) framework. Blue means the parameter needs to be updated, while gray means not. This method makes knowledge transfer more friendly and effective.

Weight Prediction. Upon acquiring the aforementioned inputs, the next step involves determining the most suitable ensemble weights for each new sample. To maintain the interpretability of these weights, we have devised an ensemble strategy for the new samples based on:

$$w = SoftMax(h_a \cdot c_n \times d_n \times disc_n), \tag{5}$$

where $w \in \mathbb{R}^{N\times 1}$, $h_a \in \mathbb{R}^{N\times P}$, $c_n \in \mathbb{R}^{1\times P}$, d_n and $disc_n$ are single numbers.

4.4. Knowledge distillation module

In neural networks, "knowledge" primarily encompasses the acquired weights and biases. However, large deep neural networks exhibit a vast array of knowledge sources. Traditional knowledge distillation techniques leverage the teacher model's logits to transfer knowledge, whereas alternative approaches concentrate on weights or activations of the intermediate layers. Additionally, pertinent knowledge forms include the dynamics between various activation types and neurons, as well as teacher model's parameters themselves.

Previous discussions have highlighted numerous studies on the enhanced performance of ensembles from theoretical standpoints. Yet, these investigations predominantly focus on methods like Boosting, Bagging, and ensembles of models and features. Building on this foundation, we propose a novel Knowledge Distillation Module (KDM) to distill the essence of the ensemble model into a single model, aiming to preserve the ensemble model's superior performance. Fig. 3 illustrates that the traditional knowledge distillation centers on the teacher model's final output layer. The underlying assumption is that the student model is trained to replicate the teacher model's predictions. This replication process is facilitated through a specialized loss function, known as distillation loss (i.e., L_{KL}), which quantifies the discrepancy between the student's and the teacher's logits. By minimizing this loss during training, the student model progressively improves its ability to generate predictions that align with those of the teacher model.

The foundational setup for the teacher and student models can be described as follows: the ensemble teacher model, denoted as f^t, comprises three distinct models: Bert-Chunk, SpanBert-Rank, and Bert-TagKPE; the student model, denoted as f^s, is Bert-Chunk. It is worth to note that, the θ is the model parameters (e.g., θ^s is the parameters of the student model. θ^t is the parameters of the teacher model).

Note that, in the teacher model, indicated by the gray (gray) section, we freeze the parameters of the ensemble model to prevent them from updating during the learning process. On the other hand, in the student model, highlighted in blue, we allow the parameters to be updated through backpropagation. This approach enables the student model to assimilate knowledge from the teacher model efficiently, without incurring significant computational overhead. Specifically, the vanilla knowledge distillation loss measuring the KL-Divergence of teachers and students can be formulated as:

$$\mathcal{L}_{KL}\left(\boldsymbol{f}\left(\tau \mid \theta^s\right), \boldsymbol{f}\left(\tau \mid \theta^t\right)\right) = \tau^2 \sum_j \boldsymbol{f}_j\left(\tau \mid \theta^t\right) \cdot \log \frac{\boldsymbol{f}_j\left(\tau \mid \theta^t\right)}{\boldsymbol{f}_j\left(\tau \mid \theta^s\right)} \tag{6}$$

where τ is the temperature used in knowledge distillation process, which controls how much to rely on the teacher's soft predictions. $j \in \{1, K\}$ is the number of classes. In sum, ensemble in deep learning may be very different from the ensemble in random features. It may be more accurate to study ensemble/knowledge distillation in deep learning as feature learning process, instead of feature selection process.

5. Experiments

In the section, we aim to showcase the efficacy of our DAEL model. We begin by comparing DAEL against several cognitive diagnosis baselines. This comparison is based on response logs obtained from algorithms tested on two distinct keyphrase extraction datasets. Subsequently, we conduct a series of comprehensive experiments including diagnostic analysis, hyper-parameter sensitivity study, model aggregation and present some cases. Collectively, these experiments are designed to illustrate the effectiveness of our *"diagnostic-aggregation-distillation"* framework from multiple, consistent perspectives.

5.1. Baselines

Keyphrase Extraction Baselines. To better test the effectiveness of the proposed framework, we select 24 representative keyphrase extraction algorithms including unsupervised methods (e.g. TextRank, TopicRank, YAKE, and SIFRank) and supervised methods (e.g. BERT-JointKPE and BERT-SpanKPE). All models are shown in Table 2 with a general description. Supervised methods are trained on the OpenKP training set (134k documents). We obtain the response logs of learners on all samples on the datasets. Following the past research (Wang et al., 2020), we split the datasets into the training, validation and test set as 7:1:2.

Diagnoser Baselines. We evaluate the performance of our proposed framework with other well-known CDMs. The details are illustrated as follows:

- *IRT* (Lord, 1952) is the most popular cognitive diagnosis method, it models students' latent traits and the parameters of exercises like difficulty and discrimination with a logistic-like function.
- *DINA* (De La Torre, 2009) is the first method to design the Q-matrix and it uses binary variables to represent mastery of skills.
- *MIRT* (Reckase & Reckase, 2009) is a multidimensional extension of IRT, modeling multiple knowledge proficiency of students and exercises.
- *MCD* (Toscher & Jahrer, 2010) uses matrix factorization for modeling the deep interaction.
- *NeuralCD* (Wang et al., 2020) is a neural cognitive diagnostic model, which leverages multi-layers for modeling complex interactions of students and exercises, aiming to diagnose students' cognition by predicting the probability of the answering exercise correctly.

Dataset Description. To illustrate the generality of our proposed framework DAEL, we conduct experiments on two common keyphrase extraction datasets, i.e., OpenKP (Xiong et al., 2019) and Inspec (Hulth, 2003). OpenKP is an open-domain keyphrase extraction dataset with various domains and topics. Each document contains the most relevant keyphrases generated by expert annotations. Furthermore, it is often used to evaluate supervised algorithms. The cognitive diagnosis model of DAEL is based on NeuralCD. In the datasets used by NeuralCD, specifically MATH and ASSIST, the average number of response logs per

Table 2
Keyphrase extraction models.

Type	Methods	Description
Unsupervised	Firstphrase	Choose the first phrase in sentences.
	YAKE (Campos et al., 2018)	Based on 5 statistical features including case, word position, frequency, contextuality, and sentence difference.
	TextRank (Mihalcea & Tarau, 2004)	Based on phrase graph and PageRank algorithm.
	SingleRank (Wan & Xiao, 2008)	Incorporate phrase co-occurrence information on TextRank.
	TopicRank (Bougouin et al., 2013)	Choose topics obtained by clustering candidate.
	TopicalPageRank (Liu et al., 2010)	Combine with LDA.
	PositionRank (Florescu & Caragea, 2017)	Incorporate position into PageRank.
	MultipartiteRank (Boudin, 2018)	Combine topical information with a multipartite graph.
	SIFRank (Sun et al., 2020)	Combine ELMo with sentence embeddings.
Supervised	BERT-JointKPE (Sun et al., 2021)	A multi-task model is trained jointly on the chunking task and the ranking task.
	BERT-RankKPE (Sun et al., 2021)	Learn the salience phrases in the documents using a ranking network.
	BERT-ChunkKPE (Sun et al., 2021)	Classify high-quality keyphrases using a chunking network.
	BERT-TagKPE (Sun et al., 2021)	Modify the sequence tagging model.
	BERT-SpanKPE (Sun et al., 2021)	Modify the span extraction model.
	RoBERTa-Variants*5 (Sun et al., 2021)	Five methods based on RoBERTa.
	SpanBERT-Variants*5 (Sun et al., 2021)	Five methods based on SpanBERT.

Table 3
Statistics of datasets.

Statistics	OpenKP	Inspec
Document Number	6616	1500
Document Len Average	900	128
Keyphrase Average	2.2	9.8
Keyphrase Len Average	2.0	2.5

Table 4
Evaluation of all diagnosers through predicting learner performance on unknown samples.

Methods	OpenKP			Inspec		
	AUC	Accuracy	RMSE	AUC	Accuracy	RMSE
DINA	0.5629	0.5447	0.5594	0.5377	0.5115	0.5782
IRT	0.5760	0.5403	0.5424	0.5595	0.5450	0.5437
MIRT	0.5691	0.5556	0.5645	0.5233	0.5175	0.5886
MCD	0.8543	0.7786	0.3765	0.6643	0.6212	0.4790
NeuralCD	**0.9140**	**0.8692**	**0.3395**	**0.8832**	**0.7615**	**0.3792**

student is 84 and 78, respectively. The experimental results from NeuralCD indicate that this volume of response logs is sufficient for training a model to diagnose student abilities. Accordingly, the 6616 samples in the OpenKP valid set provide sufficient data for each keyphrase extraction method to generate an adequate number of response logs for training DAEL's cognitive diagnostic model. This sufficiency obviates the need to employ the entire training set. Therefore, we choose the valid set for our tasks. The Inspec dataset consists of short documents selected from scientific journal abstracts which are labeled by the authors. We choose the valid (500 documents) and train (1,000 documents) sets in this paper. Table 3 shows the statistics of the two datasets. From the table, it is obvious that documents in Inspec tend to have more ground truth than those in OpenKP. In previous studies, the top N selected by the keyphrase extraction algorithm was set to $\{5, 10, 15\}$ for Inspec and $\{1, 3, 5\}$ for OpenKP considering this difference between the two datasets. For the result of the method, the top 5 keyphrases are compared with the ground truth in our work.

5.2. Implementation

In our experiment we use the pre-trained uncased BERT-based model with a 768 dimensions hidden representation as our tool. As the number of skills P is the most important hyper-parameter, we conduct sensitivity experiments on it in section 5.3.3. To set up the training process, we initialize all network parameters with Xavier initialization. The Adam optimizer (Kingma & Ba, 2014) is used in the experiment while the learning rate is set to 0.0002. We train all diagnosers for 20 epochs and select the best model on validation set for testing. All diagnosers are implemented by Pytorch and are run on a Linux server with two Intel(R) Xeon(R) E5-2650 v4 CPUs and an NVIDIA A100 GPU.

5.3. Evaluation metrics

Our experimental metrics consist of three components: metrics for learner performance prediction, and metrics for model aggregation. The first is commonly used in cognitive diagnostic to test the validity of diagnostic methods (Li et al., 2022; Wang et al., 2020, 2022). Meanwhile, the model aggregation method is employed to demonstrate the improvement of diagnostic results for the KE task.

Learner Performance Prediction. Generally, obtaining the ground truth regarding learners' abilities is challenging, making the evaluation of cognitive diagnosis models difficult. Most studies use the indirect method of assessing learners' performance prediction as a means to evaluate these models. Common evaluation metrics include Accuracy, Root Mean Square Error (RMSE), and Area Under the Curve (AUC). In this context, better predictions are indicated by higher values in Accuracy and AUC, whereas a lower value in RMSE signifies a more accurate prediction.

Model Aggregation. In order to further illustrate the usefulness of the diagnostic results for the extraction task, we realize model aggregation based on each learner's proficiency on the topics. The aggregation

Table 5
Model aggregation results of popular keyphrase extraction models. The top part lists some unsupervised methods, the middle part lists the supervised methods, and the bottom part lists the ensemble methods.

Methods	OpenKP			Inspec		
	P@5	R@5	F_1@5	P@5	R@5	F_1@5
Firstphrase	19.5	36.7	23.6	24.0	15.0	17.3
YAKE (Campos et al., 2018)	12.1	29.1	16.7	21.0	13.6	15.5
TextRank (Mihalcea & Tarau, 2004)	5.5	14.2	7.9	31.7	19.2	22.6
SingleRank (Wan & Xiao, 2008)	14.4	34.5	19.7	33.0	20.2	23.6
TopicRank (Bougouin et al., 2013)	14.4	30.3	19.6	28.2	16.9	20.0
TopicalPageRank (Liu et al., 2010)	13.0	34.4	19.6	32.9	20.0	23.5
PositionRank (Florescu & Caragea, 2017)	15.1	36.0	20.6	32.8	19.9	23.3
MultipartiteRank (Boudin, 2018)	14.8	34.9	20.1	28.2	17.3	20.2
SIFRank (Sun et al., 2020)	12.3	29.2	16.7	33.2	21.9	29.1
Aggregation(Unsupervised)	14.3	30.2	19.6	37.6	22.7	25.1
BERT-JointKPE (Sun et al., 2021)	22.7	57.1	30.3	37.9	24.3	27.9
SpanBERT-RankKPE (Sun et al., 2021)	23.2	61.8	33.9	38.7	24.9	28.6
RoBERTa-TagKPE (Sun et al., 2021)	23.0	58.9	31.8	36.9	23.7	27.2
Averaging	23.7	61.0	33.5	39.1	25.0	28.9
Weighted Voting(*Precision*)	24.0	61.4	33.7	39.7	25.2	29.4
AEKE	24.5	62.0	34.1	40.3	25.8	29.8
DAEL (Diagnosis-Aggregation-Distillation)	**24.8**	**67.8**	**35.6**	**40.5**	**26.2**	**31.3**

is tested on both OpenKP and Inspec datasets with several traditional keyphrase extraction metrics including Precision, Recall and F1-score.

5.4. Main results

Learner Performance Prediction. Table 4 presents the experimental results of five baseline diagnoser methods for predicting learner performance on unknown samples. The NeuralCD method, which serves as the diagnostic component of DAEL, exhibits the best performance on both the OpenKP and Inspec datasets, underscoring its effectiveness in evaluating the capabilities of keyphrase extraction algorithms. However, traditional models, such as IRT, MIRT, and DINA, show subpar performance. In contrast, models like MCD and NeuralCD, both implemented using neural networks, demonstrate relatively better results. This contrast highlights the complexity in capturing the relationship between the model' abilities and the features of samples, and concurrently underscores the efficiency of neural networks in this context. Notably, our method leverages NeuralCD, further emphasizing the superiority of DAEL in KE task.

Model Aggregation. Building on the precise diagnostic outcomes, we can implement a model aggregation strategy among all the models we have evaluated. The diagnostic results clearly delineate each model's strengths, thus we can choose the best models for each topic based on our new multidimensional metric, rather than traditional evaluation methods. In our experiment, we select the top 3 models for each topic and compile their responses to the samples. A keyphrase that frequently appears in these responses is selected as the final output of our integrated model. This is because such phrases, consistently chosen by top three models in the given topic, are highly likely to be essential. Besides, we make a distinction between supervised and unsupervised methods. We separately integrate all unsupervised models and supervised models to better assess their impact. The results are detailed in Table 5, which presents the model aggregation results of popular keyphrase extraction methods. It is observed that unsupervised methods, lacking sufficient training, perform poorly on the news dataset OpenKP, leading to less effective aggregation results for unsupervised methods on this dataset. However, when considering all methods collectively, the aggregation achieves relatively good results on both datasets.

Regarding the results, there are some key points to note. Firstly, in the domain of keyphrase extraction, two datasets, OpenKP and Inspec, are utilized for different methodologies: OpenKP is typically used in supervised method experiments, while Inspec is favored for unsupervised approaches. As previously mentioned, the OpenKP dataset features fewer ground truth keyphrases, leading us to adopt different top N metrics for evaluation: *@1, @3, @5* for OpenKP and *@5, @10, @15* for Inspec. This variance in evaluation metrics partly explains why models tend to score higher in recall on OpenKP compared to Inspec. Further analysis of unsupervised methods on OpenKP reveals that their outputs often exceed the length of the ground truth, resulting in a mismatch. This can be attributed to two main factors. On the one hand, these methods demonstrate a limited capability in processing long and complex texts. On the other hand, this mismatch is a significant reason for the failure of aggregating all unsupervised methods. In conclusion, these observations serve as a simple yet effective demonstration of the value of diagnostic results in enhancing the understanding and application of keyphrase extraction models.

Knowledge Distillation. Regarding the efficacy of knowledge distillation, our developed model sets a new standard, not just in terms of its effectiveness but also regarding its operational efficiency. The distilled student model, which is a central component of the Diagnosis-Aggregation-Distillation framework (DAEL), remarkably contains only about one-third of the parameter volume found in the comprehensive ensemble model, yet it manages to match, if not surpass, the teacher model's performance levels. To put this into perspective, the DAEL framework is engineered with approximately 130 million parameters, leading to a video memory requirement of around 17 GB during its training phase. This is in stark contrast to the AEKE model, which is significantly more resource-intensive, comprising 350 million parameters and demanding about 50 GB of video memory throughout the training process. Moreover, the DAEL model not only achieves parity with the teacher model in terms of raw performance but also demonstrates superior results on specific benchmarks. For instance, on the OpenKP benchmark, the DAEL model records impressive scores of 24.8 and 67.8, thereby clearly outperforming the AEKE model. This not only highlights the DAEL model's efficiency in utilizing resources but also underscores its capability to distill and leverage knowledge effectively, thereby setting a new benchmark in the field of knowledge distillation.

5.5. Visualize results

Visualization of Distribution. In our analysis, we present a visualization of the data distribution in Fig. 4, focusing on sample factors. The vertical axis, dis, represents discrimination, reflecting the dataset's ability to effectively distinguish between learners of different abilities. The horizontal axis, $diff$, represents difficulty, indicating the challenge of predicting the correct labels for the dataset samples. Based

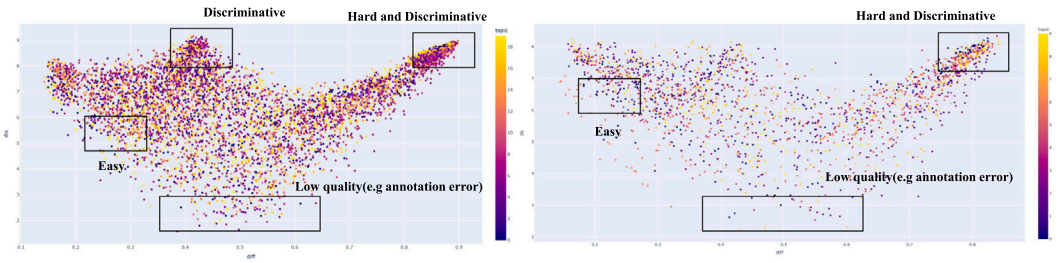


Fig. 4. Visualization of data distribution based on sample factors including skill difficulty and discrimination factors.

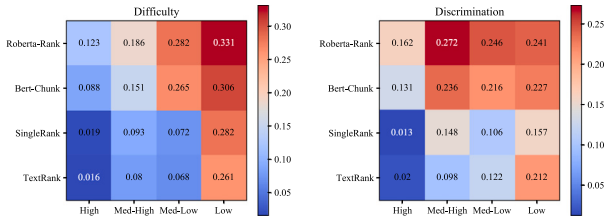


Fig. 5. A heat map showing the relationship between the difficulty factor and the discrimination factor of the sample and the traditional indicator *Precision@5*. Four keyphrase extraction models and all data within a topic in the OpenKP dataset are used as an example.

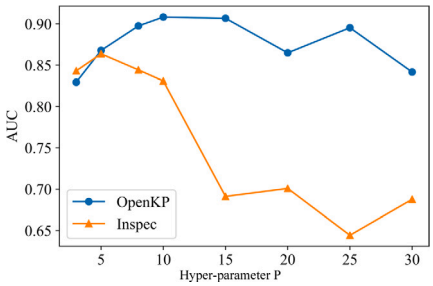


Fig. 6. Hyper-parameter Sensitivity Study.

on these two dimensions, we define four labels: *Easy*, *Discriminative*, *Low Quality*, and *Hard and Discriminative*. Samples characterized by low difficulty and moderate discrimination are classified as *Easy*. Those with moderate difficulty and high discrimination are labeled as *Discriminative*. Samples exhibiting low discrimination, regardless of their difficulty level, are categorized as *Low Quality* because they do not effectively differentiate between learners' abilities. Finally, samples that possess both high difficulty and high discrimination are classified as *Hard and Discriminative*. This visualization reveals a distinct half-moon shape pattern in both datasets, signifying that the majority of discriminative examples are classified as either straightforward or particularly challenging. Notably, across various topics, the distribution of difficulty and discrimination shows minimal variation, suggesting a uniform challenge level across the board. However, it is important to highlight the presence of some samples with low discrimination, which likely indicates issues related to the quality of the text documents or inaccuracies in the ground truth. Furthermore, our analysis shows that the Inspec dataset contains relatively fewer low-quality data instances compared to others. This is attributed to Inspec's composition, which mainly consists of abstracts from scientific and technical articles, enriched with authors' tags. This implies that the source and nature of the dataset significantly influence the quality of data and, consequently, the performance and reliability of KE models.

Diagnostic Outcome Heat Map. In our concluding analysis, we showcase specific example of the diagnostic outcome from the DAEL applied to OpenKP dataset, demonstrating the validity of sample factors. In Fig. 5, we examine four keyphrase extraction algorithms evaluated across all data within a single topic of OpenKP. The data are segmented into four subsets based on a spectrum of difficulty and discrimination, ranging from high to low. For the purpose of this analysis, we employ the traditional metric *Precision@5* to gauge the performance of each model across these subsets. The results indicate a clear trend: as the difficulty decreases, the average performance of the algorithms improves. This trend showcases how the discrimination factor is instrumental in distinguishing between more and less capable methods, using traditional performance metrics as a benchmark.

Interestingly, in subsets of low-quality data, traditional models exhibit some commendable performances. This observation suggests that despite the general advancement in keyphrase extraction methodologies, traditional models may still hold value, particularly in handling specific types of data. This nuanced finding underscores the importance of considering both novel and established approaches within the dynamic landscape of KE research.

5.6. Hyper-parameter analysis

In our work, the number of skills (i.e., P) plays a pivotal role as a hyper-parameter. It is instrumental in defining the effectiveness of topic clustering and influences the formulation of assessment skills. To explore the impact of P on our analysis, this section is dedicated to examining its sensitivity. The performance of DAEL, measured by AUC with varying numbers of topics denoted by P, is illustrated in Fig. 6. For OpenKP dataset, the experimental data reveals an initial increase in the effectiveness of ensemble result with the growth of P, which is then followed by a decrease. Notably, the optimal number of topic clusters for the OpenKP is identified as 10. Specifically, when P is set to a low value, the quality of document topic clustering is compromised, which in turn adversely affects the cognitive diagnosis of multi-dimensional abilities and the ensuing ensemble process. Conversely, when P exceeds 10, the ensemble results begin to stabilize and show less variation in effectiveness. In the case of Inspec, the effectiveness of ensemble result peaks when P is set to 5. However, when P reaches 10, it begins to decline sharply. This indicates that a large P does not align with the characteristics of the Inspec dataset. To balance the ensemble performance between the two datasets, we have chosen to set P to 10 for the experiments conducted in our study.

5.7. Case study

In this section, we present a representative sample from OpenKP, followed by a brief analysis of certain cases in which our model fails. As shown in Fig. 7(a), the text "Evelyn Rodriguez was hit around 4 pm..". clearly identifies "Evelyn Rodriguez" as a victim, which is one of the ground truth labels. In the context of legal domain keyphrase extraction, it is crucial to prioritize the victim's name. Therefore, the ability to extract "Evelyn Rodriguez" serves as a key indicator of the ensemble model's effectiveness. This rationale guided our selection of this particular case. To compare the ensemble results of both DAEL and weighted voting, we aggregate the results of three Keyphrase Extraction

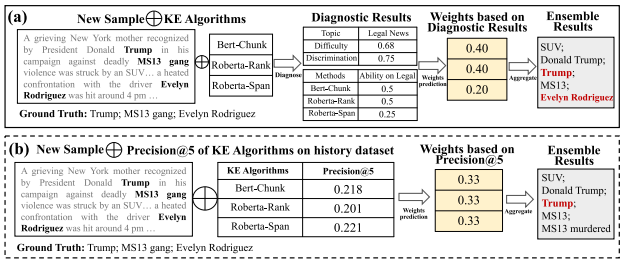


Fig. 7. Visualized keyphrases extracted by DAEL (a) and traditional strategy (b).

(KE) methods — BERT-Chunk (Sun et al., 2021), RoBERTa-Rank (Sun et al., 2021), RoBERTa-Span (Sun et al., 2021) — using both our strategy and traditional weighted voting method.

Fig. 7(a) details the DAEL process. Upon inputting a new sample into the diagnostic module, we obtain its corresponding diagnostic results. The sample is identified as a legal news report, with its difficulty and discrimination metrics indicating high textual quality. The abilities of the three KE methods in handling legal news topics are also revealed. Leveraging these diagnostic outcomes, DAEL adaptively adjusts the weights of different methods during aggregation, placing greater emphasis on approaches like BERT-Chunk and RoBERTa-Span that are better suited for extracting legal keyphrases. Consequently, the ensemble model prioritizes the victim's name, "Evelyn Rodriguez", elevating its ranking and leading to an optimized ensemble result.

Conversely, Fig. 7(b) demonstrates the traditional method, which bases its aggregation on the historical performance of the three methods, evaluated solely on the metric *Precision@5*. Since their overall performance on this metric is similar, uniform weights are applied to all new samples. As shown in the ensemble results in Fig. 7(b), the traditional weighted voting method extracts the phrase "SUV", ranking it at the top, while overlooking the phase "Evelyn Rodriguez". However, this phrase is irrelevant to the legal domain and does not serve as a valid ground truth for the case. As a result, overall performance metrics, which primarily reflect average outcomes, overlook the unique characteristics of individual samples and the application of uniform weights fails to leverage the specific strengths of each method within the legal domain. In contrast, DAEL's final results exclude the phrase "SUV", indicating that DAEL's ensemble strategy is better able to adapt to specific topics by prioritizing relevant phrases and downgrading irrelevant ones. This comparison highlights the adaptability and effectiveness of our proposed model. Unlike traditional methods that apply fixed weights during aggregation, DAEL's weights are flexible and context-sensitive.

Despite its advantages, the DAEL model faces several challenges, primarily including ambiguous phrases, rare topics, and domain-specific terminology. With respect to ambiguous phrases, the model requires a sophisticated semantic understanding of the entire document to distinguish between different interpretations of a given phrase. However, the DAEL model aggregates results from individual keyphrase extraction methods, without performing a holistic analysis of the document. This limitation in context processing is the reason DAEL struggles with ambiguous phrases. In the case of rare topics and domain-specific terminology, the supervised model employed by DAEL has limited coverage of topics. For example, the training dataset KP20k (Meng et al., 2017) used in BERT-JointKPE is confined to the computer science domain. As a result, models trained on such datasets struggle to capture patterns that are relevant to rare topics or other domains. DAEL encounters a similar challenge when integrating these supervised methods, as effective performance on rare topics and specific domains requires the construction of domain-specific datasets for additional fine-tuning.

5.8. Analysis of efficiency

To convincingly demonstrate that our DAEL model not only matches state-of-the-art performance but also offers significant advantages in terms of efficiency, we conducted a comparative analysis emphasizing the reduction in parameter count between the ensemble model and the distillation model. This comparison underscores the core benefits of computational efficiency and resource optimization in modern machine learning applications.

Ensemble models, by their design, aggregate outputs from multiple individual models, which results in a significant increase in total parameter count. For instance, the "teacher" model in our framework (i.e., Chunk-BERT, SpanBert-Rank and Bert-TagKPE), an ensemble model, comprises 300+ million parameters, indicative of the substantial computational overhead typical of ensemble approaches. However, through our innovative distillation approach, we reduced the complexity by transferring knowledge from the ensemble model to a single "student" model, labeled Chunk-BERT. Post-distillation, the student model's parameter count is reduced to 100 million, approximately 33% of the original size.

Despite this dramatic reduction in model size, the distilled model (i.e., DAEL) outperforms existing SOTA (state-of-the-art) models in keyphrase extraction (KE) tasks. Specifically, on the OpenKP dataset, the ensemble model (i.e., AEKE) achieves an F1 score of 34.1, while the distilled model achieves an F1 score of 35.6. Besides, on the Inspec dataset, the ensemble model scores 29.8 (F1), while the student model achieves 31.3. The above results demonstrate that the reduction in complexity does not come at the expense of significant performance loss. Instead, our DAEL framework strikes an optimal balance between efficiency and accuracy, making it a highly viable alternative to traditional ensemble methods, especially in resource-constrained environments. This advancement is critical for applications requiring scalable, efficient computing solutions while maintaining competitive task performance.

Finally, it is important to clarify that, unlike previous methods, DAEL requires diagnosing each new sample to determine its topic, evaluating the performance of each method on that topic, and ultimately adjusting the weights of predictions from different methods accordingly. However, this process is both automated and efficient, as the diagnostic model (NeuralCD) is implemented using a Multi-Layer Perceptron (MLP) layer, which does not require additional training. Moreover, since the sample sizes in our datasets are all within 6,616, the computational overhead remains manageable.

5.9. Leveraging LLMs as keyphrase extractor

To study the keyphrase extraction ability of large language models (LLMs) under zero-shot learning, we adopted a cue-based strategy. The experimental subjects include several open-source LLMs, such as ChatGLM3-6B (GLM et al., 2024), Baichuan2-7B (Yang et al., 2023), Qwen2-7B (Yang et al., 2024), and Llama3-8B (Dubey et al., 2024). For ChatGPT, we referenced the original test results provided by Song, Geng et al. (2023). The experimental results (see Table 6) show that the performance of these LLMs is comparable to some unsupervised learning methods. Notably, these models were not specifically trained or fine-tuned on the two datasets, yet the keyphrase extraction method based on direct cues still yielded good results. This suggests that LLMs have potential research value in keyphrase extraction tasks, particularly in cue design. More complex and effective cue strategies can be further explored in the future.

Compared to traditional methods, LLMs offer a new approach to keyphrase extraction, not only reducing reliance on labeled data but also providing more room for innovation through the flexibility of prompt design. This presents broad research prospects for improving keyphrase extraction tasks. However, compared to the method proposed in this paper, the cue-based LLM approach is not ideal. This is

Table 6
Keyphrase extraction results of LLMs.

LLMs	OpenKP			Inspec		
	P@5	R@5	F_1@5	P@5	R@5	F_1@5
ChatGLM3-6B (GLM et al., 2024)	10.6	30.1	15.6	30.7	17.9	22.6
Baichuan2-7B (Yang et al., 2023)	7.7	19.2	11.0	21.0	12.2	15.4
Qwen2-7B (Yang et al., 2024)	13.4	37.5	19.8	27.8	16.8	20.9
Llama3-8B (Dubey et al., 2024)	6.0	17.1	8.9	30.9	17.7	22.5
ChatGPT (gpt-3.5-turbo)	11.4	32.8	16.9	26.8	15.3	19.5
DAEL (Diagnosis-Aggregation-Distillation)	**24.8**	**67.8**	**35.6**	**40.5**	**26.2**	**31.3**

because the broad generalization of keyphrase understanding by large models means that many potential keyphrases could be considered appropriate, making it difficult to determine which is the most suitable one.

Thus, future research can focus on optimizing cue design and selection strategies to improve LLM performance in keyphrase extraction tasks. Leveraging the powerful language generation capabilities of LLMs, more targeted and diversified cue strategies might enhance the precision and reliability of keyphrase extraction.

6. Conclusion

In this study, we have introduced and examined a novel methodology, termed ***Distillation-based Adaptive Ensemble Learning (DAEL)***, specifically designed to improve the efficiency of keyphrase extraction. Our approach uniquely combines the processes of diagnosis, aggregation, and distillation. The first stage involves a Cognitive Diagnosis module, which evaluates the distinct strengths of various keyphrase extraction models. This is followed by the deployment of an adaptive aggregation module, responsible for generating a customized weight distribution for each data instance. The final stage encompasses the knowledge distillation module, which concentrates the collective capabilities of the ensemble into a single, more efficient model, significantly reducing computational requirements. Through rigorous testing on diverse real-world datasets, we have demonstrated the superiority of our model. This marks a substantial leap forward in the field of keyphrase extraction, setting a new benchmark when compared to existing leading techniques.

7. Discussion

This section discusses the advantages of our diagnostic approach, its ability to generalize across different domains, and potential future research directions. First, keyphrase extraction is inherently a challenging task, and even state-of-the-art methods struggle to achieve optimal performance, limiting the application of keyphrase extraction techniques in real-world scenarios. While neural networks and pre-trained language models are often employed to improve performance, they lack explainability, particularly in their understanding of semantic information in text tasks. As previously mentioned, our framework introduces fine-grained metrics to assess the suitability of algorithms for specific contexts or datasets. All of our diagnostic results, including ability and sample factors, are highly interpretable, providing valuable insights for other research in this field. Model aggregation, for instance, is a simple and straightforward application. For example, we can use the SpanBert-Chunk for medical-themed data and the Bert-TagKPE for sports-themed data, allowing the models' strengths to complement one another. Our diagnostic results consider the interaction between algorithms and samples, resulting in more accurate and appropriate outcomes.

Our experiments demonstrate the performance of DAEL on the Inspec and OpenKP datasets. The Inspec dataset primarily encompasses domains such as Computers and Control, and Information Technology. These fields are characterized by data-driven decision-making and domain-specific conventions, traits that are also prevalent in legal and financial sectors. As a result, DAEL is naturally capable of generalizing to these domains. The OpenKP dataset, on the other hand, consists of approximately seventy thousand web pages sampled from the Bing search engine index. It includes content such as news articles, multimedia pages from video sites, and index pages with numerous hyperlinks. Unlike the Inspec dataset, the content in OpenKP is not restricted to any specific domain and often features informal text, including non-academic writings. Therefore, DAEL's performance on the OpenKP dataset serves as a strong indication of its ability to generalize to informal text.

Looking ahead, several directions warrant further exploration. First, it is essential to enhance the adaptability of diagnostic techniques to a wide range of NLP tasks. Given the significant differences among various NLP tasks, applying Cognitive Diagnosis techniques often presents challenges. Second, as NLP generative tasks resemble real-life subjective questions, we aim to explore new techniques for handling such response logs. Lastly, many opportunities remain for investigating how to apply diagnostic results. For example, understanding how to leverage diagnostic results to inform course learning and assist in model training is an area that deserves further study.

CRediT authorship contribution statement

Kai Zhang: Conceptualization, Methodology, Writing – original draft. **Hongbo Gang:** Methodology, Validation, Writing – review & editing. **Feng Hu:** Methodology. **Runlong Yu:** Methodology, Visualization. **Qi Liu:** Supervision, Project administration.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Kai Zhang reports financial support was provided by National Natural Science Foundation of China. Kai Zhang reports a relationship with National Natural Science Foundation of China that includes: funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was partially supported by the National Natural Science Foundation of China (Grants No. 62406303), Anhui Provincial Natural Science Foundation (No. 2308085QF229), Anhui Province Science and Technology Innovation Project (202423k09020010) and the Fundamental Research Funds for the Central Universities (No. WK2150110034).

References

Bennani-Smires, K., Musat, C.-C., Hossmann, A., Baeriswyl, M., & Jaggi, M. (2018). Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd conference on computational natural language learning* (pp. 221–229).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Boudin, F. (2018). Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 2 (short papers)* (pp. 667–672).

Bougouin, A., Boudin, F., & Daille, B. (2013). Topicrank: Graph-based topic ranking for keyphrase extraction. In *International joint conference on natural language processing* (pp. 543–551).

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A. M., Nunes, C., & Jatowt, A. (2018). A text feature based automatic keyword extraction method for single documents. In *European conference on information retrieval* (pp. 684–691). Springer.

Chen, P., Liu, S., Zhao, H., & Jia, J. (2021). Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5008–5017).

Chen, D., Mei, J.-P., Wang, C., Feng, Y., & Chen, C. (2020). Online knowledge distillation with diverse peers. In *Proceedings of the AAAI conference on artificial intelligence, vol. 34* (pp. 3430–3437).

Clark, K., Luong, M.-T., Khandelwal, U., Manning, C. D., & Le, Q. V. (2019). Bam! born-again multi-task networks for natural language understanding. arXiv preprint arXiv:1907.04829.

De La Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ding, H., & Luo, X. (2021). AttentionRank: Unsupervised keyphrase extraction using self and cross attentions. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 1919–1928).

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Florescu, C., & Caragea, C. (2017). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1105–1115).

Ganaie, M. A., Hu, M., Malik, A., Tanveer, M., & Suganthan, P. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, Article 105151.

Gao, W., Liu, Q., Huang, Z., Yin, Y., Bi, H., Wang, M.-C., Ma, J., Wang, S., & Su, Y. (2021). Rcd: Relation map driven cognitive diagnosis for intelligent education systems. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 501–510).

Gao, W., Wang, H., Liu, Q., Wang, F., Lin, X., Yue, L., Zhang, Z., Lv, R., & Wang, S. (2023). Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval* (pp. 983–992).

GLM, T., Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., et al. (2024). ChatGLM: A family of large language models from GLM-130b to GLM-4 all tools. arXiv preprint arXiv:2406.12793.

Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 1262–1273).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on empirical methods in natural language processing* (pp. 216–223).

Jin, X., Liu, Q., Yue, L., Liu, Y., Zhao, L., Gao, W., Gong, Z., Zhang, K., & Bi, H. (2023). Diagnosis then aggregation: An adaptive ensemble strategy for keyphrase extraction. In *CAAI international conference on artificial intelligence* (pp. 566–578). Springer.

Kenton, J. D. M.-W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Kong, A., Zhao, S., Chen, H., Li, Q., Qin, Y., Sun, R., & Bai, X. (2023). PromptRank: Unsupervised keyphrase extraction using prompt. arXiv preprint arXiv:2305.04490.

Li, J., Wang, F., Liu, Q., Zhu, M., Huang, W., Huang, Z., Chen, E., Su, Y., & Wang, S. (2022). HierCDF: A Bayesian network-based hierarchical cognitive diagnosis framework. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 904–913).

Liang, X., Wu, S., Li, M., & Li, Z. (2021). Unsupervised keyphrase extraction by jointly modeling local and global context. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 155–164).

Liu, Q. (2021). Towards a new generation of cognitive diagnosis. In *IJCAI* (pp. 4961–4964).

Liu, Z., Huang, W., Zheng, Y., & Sun, M. (2010). Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 366–376).

Liu, J., Liu, B., Li, H., & Liu, Y. (2022). Meta knowledge distillation. arXiv preprint arXiv:2202.07940.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, C., Yang, L., Gao, W., Li, Y., & Liu, Y. (2023). MuST: An interpretable multi-dimensional strain theory model for corporate misreporting prediction. *Electronic Commerce Research and Applications*, *57*, Article 101225.

Lord, F. (1952). A theory of test scores. *Psychometric Monographs*.

Meng, R., Wang, T., Yuan, X., Zhou, Y., & He, D. (2022). General-to-specific transfer labeling for domain adaptable keyphrase generation. arXiv preprint arXiv:2208.09606.

Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P., & Chi, Y. (2017). Deep keyphrase generation. arXiv preprint arXiv:1704.06879.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404–411).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*: *Technical report*, Stanford InfoLab.

Papagiannopoulou, E., & Tsoumakas, G. (2020). A review of keyphrase extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*, Article e1339.

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3967–3976).

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*, 1883.

Reckase, M. D., & Reckase, M. D. (2009). *Multidimensional item response theory models*. Springer.

Song, M., Feng, Y., & Jing, L. (2022a). Hyperbolic relevance matching for neural keyphrase extraction. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 5710–5720).

Song, M., Feng, Y., & Jing, L. (2022b). Utilizing BERT intermediate layers for unsupervised keyphrase extraction. In *Proceedings of the 5th international conference on natural language and speech processing* (pp. 277–281).

Song, M., Feng, Y., & Jing, L. (2023). A survey on recent advances in keyphrase extraction from pre-trained language models. *Findings of the Association for Computational Linguistics: EACL 2023*, 2108–2119.

Song, M., Geng, X., Yao, S., Lu, S., Feng, Y., & Jing, L. (2023). Large language models as zero-shot keyphrase extractor: A preliminary empirical study. arXiv preprint arXiv:2312.15156.

Song, M., Jiang, H., Liu, L., Shi, S., & Jing, L. (2023). Unsupervised keyphrase extraction by learning neural keyphrase set function. In *Findings of the association for computational linguistics: ACL 2023* (pp. 2482–2494).

Song, M., Jing, L., & Xiao, L. (2021). Importance estimation from multiple perspectives for keyphrase extraction. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2726–2736).

Song, M., Liu, H., Feng, Y., & Jing, L. (2023). Improving embedding-based unsupervised keyphrase extraction by incorporating structural information. In *Findings of the association for computational linguistics: ACL 2023* (pp. 1041–1048).

Sun, S., Liu, Z., Xiong, C., Liu, Z., & Bao, J. (2021). Capturing global informativeness in open domain keyphrase extraction. In *CCF international conference on natural language processing and Chinese computing* (pp. 275–287). Springer.

Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. (2020). SIFRank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access*, *8*, 10896–10906.

Toscher, A., & Jahrer, M. (2010). Collaborative filtering applied to educational data mining. *KDD Cup*.

Wan, X., & Xiao, J. (2008). Single document keyphrase extraction using neighborhood knowledge. In *AAAI, vol. 8* (pp. 855–860).

Wang, F., Liu, Q., Chen, E., Huang, Z., Chen, Y., Yin, Y., Huang, Z., & Wang, S. (2020). Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence, vol. 34* (pp. 6153–6161).

Wang, F., Liu, Q., Chen, E., Huang, Z., Yin, Y., Wang, S., & Su, Y. (2022). NeuralCD: A general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering*.

Xiong, L., Hu, C., Xiong, C., Campos, D., & Overwijk, A. (2019). Open domain web keyphrase extraction beyond language modeling. In *Proceedings of the EMNLP-IJCNLP 2019* (pp. 5175–5184).

Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al. (2023). Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

Yim, J., Joo, D., Bae, J., & Kim, J. (2017). A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4133–4141).

Zhang, L., Chen, Q., Wang, W., Deng, C., Zhang, S., Li, B., Wang, W., & Cao, X. (2022). MDERank: A masked document embedding rank approach for unsupervised keyphrase extraction. In *Findings of the association for computational linguistics: ACL 2022* (pp. 396–409).

Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11953–11962).

Zhao, H., Lu, M., Yao, A., Guo, Y., Chen, Y., & Zhang, L. (2017). Physics inspired optimization on semantic transfer features: An alternative method for room layout estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10–18).