Multi-Agent Debate based Concept Augmentation for Enhanced Cognitive Diagnosis

Pengyang Shao
National University of Singapore
Singapore

Lei Chen*
University of Science and Technology of China
Hefei, Anhui, China
chenlei.hfut@gmail.com

Fei Liu
Hefei University of Technology
Hefei, Anhui, China

Yonghui Yang
National University of Singapore
Singapore

Xun Yang
University of Science and Technology of China
Hefei, Anhui, China

Meng Wang*
Hefei University of Technology
Hefei, Anhui, China
eric.mengwang@gmail.com

Abstract

Cognitive Diagnosis (CD) models are constrained by the data quality of students' response logs. Recent advancements in Large Language Model (LLM) based data augmentation show promise for enhancing CD. However, ensuring the reliability and accuracy of LLM-generated annotations remains a significant challenge. In this paper, we propose **M**ulti-**A**gent based **C**oncept **A**ugmentation for **C**ognitive **D**iagnosis (**MACA-CD**), a novel approach that enhances CD by generating and fusing reliable concept descriptions and relations based solely on concept names. MACA-CD consists of two main components: (1) a Multi-Agent Debate (MAD) based concept augmentation process that generates diverse and reliable concept descriptions and relations, reducing reliance on behavioral data. For concept descriptions, two agents generate outputs that include definitions, core features, and real-world applications, and continue debating until a judge agent determines that consensus has been reached. Concept relations are then identified using a Breadth-First Search approach to efficiently and progressively uncover relationships based on concept descriptions, with each step carried out by MAD. (2) a concept augmentation-enhanced CD model that refines concept embeddings using a graph self-supervised learning fusion layer and a pairwise comparator-based Description Fusion Layer, leading to more reliable and accurate concept embeddings. Experimental results on three real-world datasets show that MACA-CD consistently outperforms existing methods under various real-world scenarios.

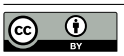
Keywords

Cognitive Diagnosis, Large Language Models, Multi-Agent Debate

Resource Availability:

The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.18051128.

*Corresponding Author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3780196

1 Introduction

Among AI applications in education [38, 39, 54, 55, 58], Cognitive Diagnosis (CD) [12] has been widely studied for its capacity to support fine-grained and personalized requirements [42, 45]. Recent advances in neural network-based CD (NCD) models have demonstrated superior capability in capturing fine-grained mastery levels over various knowledge concepts [49, 56], which serves as the foundation for various downstream tasks, such as course recommendation [35, 59, 66], computerized adaptive testing [70], and education level comparison across regions [22].

Although NCD models have shown strong performance across various tasks, recent studies highlight their heavy reliance on data quality [23, 29]. On the one hand, a key advantage of NCD models is their ability to provide fine-grained estimates of each student's mastery over individual knowledge concepts. However, achieving such high-precision diagnostics often requires sufficient interactions between students and all concepts [44]. On the other hand, NCD models typically assume that users have adequate interaction records during the training phase. This assumption limits their robustness in practical scenarios. For instance, in cold-start settings, some students may exhibit very few interactive behaviors during training, making accurate diagnosis difficult [30]. Similarly, in inductive scenarios—where users unseen during training appear at test time, possibly with abundant behavior data—existing NCD models still require retraining to accommodate them, revealing a lack of flexibility and generalization capability [29].

One promising direction for enhancing CD is leveraging auxiliary information, such as descriptive texts [30, 50] or dependencies among knowledge concepts [13, 25]. However, collecting such data often requires time-consuming expert annotation. Motivated by recent advances in Large Language Models (LLMs), researchers have proposed using LLMs to generate detailed descriptions that enhance CD performance [15, 30, 33]. These approaches typically utilize entity names, corresponding behavioral data, and carefully crafted prompts to guide LLMs in the automatic annotation process. Despite their effectiveness, these methods face two major limitations: (1) Existing approaches rely on feeding behavioral data into LLMs, thereby failing to eliminate the dependency on high-quality behavioral data. In other words, they remain ineffective for students who lack behavioral data during the training process. (2) Existing approaches rely solely on single-pass LLM generation, making it

difficult to ensure the reliability and consistency of the outputs. The generated descriptions are often sensitive to prompt design and may vary across runs, lacking mechanisms for verification or refinement.

To this end, we propose **M**ulti-**A**gent based **C**oncept **A**ugmentation for **C**ognitive **D**iagnosis (**MACA-CD**), which focuses on how to generate reliable concept descriptions and relations solely based on concept names. Specifically, MACA-CD can be divided into two main parts. First, to improve annotation stability and accuracy, inspired by the recent success of LLM based agents [27, 31, 52, 61], we propose a Multi-Agent Debate (MAD) based concept augmentation process, which focuses on generating diverse and reliable concept descriptions and relations by leveraging multiple debating agents and a judge agent. Notably, to improve the efficiency of generating concept relations while ensuring quality, we design a three stage relation generation method based on the Breadth-First Search (BFS) approach. This process relies solely on concept names, ensuring it is not affected by sparse behavioral data. Second, we propose a novel concept augmentation-enhanced CD model, which integrates the generated concept relations and descriptions to refine concept embeddings in CD. For concept relations, we introduce a graph self-supervised learning concept fusion layer to improve representation quality. For concept descriptions, we propose a pairwise Comparator-based Description Fusion Layer, which innovatively generates fusion weights based on the reliability of free embeddings. Extensive experiments have been conducted on three real-world datasets, demonstrating stable improvements of MACA-CD under scenarios of cold-start students, zero-shot concepts, and standard random splits. Our main contributions can be summarized as:

- To address the limitations of dependency on behavioral data and reliability, we propose MACA-CD , which utilizes a Multi-Agent Debate-based concept augmentation process.
- To leverage the concept augmentation, MACA-CD further introduces a concept augmentation-enhanced CD model to refine concept embeddings.
- Extensive experiments on three real-world datasets demonstrate the stable effectiveness of MACA-CD under scenarios of cold-start students, zero-shot concepts, and standard random splits.

2 Related Work

2.1 Cognitive Diagnosis

Typically, educational CD models involve with the student set S ($|S| = M$), the exercise set E ($|E| = N$), and the concept set K ($|K| = T$) [12, 14, 24]. These models take students' response logs $R = \{(m, n, r_{mn})\}$ and expert-labeled exercise-concept relations $\mathbf{Q} = [q_{nt}]_{N \times T}$ as input, and output students' comprehension levels on all concepts $\mathbf{C}$ [41, 49]. $r_{mn} = 1$ indicates that student m's correctly answer exercise n; while $r_{mn} = 0$ indicates a wrong answer. $q_{nt} = 1$ indicates that exercise n is related to concept t.

To further boost CD performance, researchers have studied how to incorporate supplementary information to enhance CD performance, e.g., integrating concept relations via graph neural network and bayesian network [13, 25, 28]. Other studies have also emphasized the value of textual data [33, 50]. However, this approach incurs additional costs and may introduce errors in the annotations—for example, the presence of cyclic concept dependencies observed in Junyi [8]. Inspired by recent success of LLMs, researchers have proposed leveraging LLMs for data or representation augmentation [11, 15, 33]. For example, Liu et al. introduce exercise and concept refiners based on entity descriptions and entity behaviors to generate more coherent and reasonable detailed descriptions [33]. Gao et al. utilize LLM-based agents to simulate diverse students based on their profiles and historical behaviors, thereby generating additional behavior data [15]. Although effective to some extent, these methods face two key limitations. First, the quality of LLM-generated content is not always reliable. Second, they still rely on entity behavior data, making them less effective for cold-start scenarios.

2.2 Multi-Agent Debate

Recent studies suggest that enforcing structured or unified conditional frameworks can effectively help improve generation consistency under complex constraints [46, 47]. Among these generation methods, Multi-Agent Debate (MAD) has emerged as a promising approach [7, 27, 40, 67]. MAD systems leverage multiple agents to express arguments and introduce a judge agent for final responses, particularly excelling in tasks requiring deep contemplation [27]. Efficiency studies demonstrate that sparse communication topologies achieve comparable performance with reduced computation [26], while group debate frameworks facilitate knowledge sharing through inter-group interactions [31].

Given MAD's capacity for reliable generation, researchers now deploy it for high-quality data augmentation [17, 19, 68]. For example, ALGPT employs dynamic multi-agent teams for autonomous multimodal annotation [69], while FMAD generates labeled reasoning datasets by evaluating stepwise contributions via confidence metrics from multi-agent systems [68]. Despite growing adoption of MAD for data augmentation, few work has explored its application in educational contexts. To address this gap, we pioneer MAD for educational knowledge structuring, leveraging multi-agent debate to generate two types of reliable outputs: (1) comprehensive concept descriptions ($\mathcal{T}^{des}$) and (2) a directed concept dependency graph (**P**). This approach harnesses MAD's deliberative strengths to construct accurate augmentation for the CD task.

3 The Proposed MACA-CD

Our proposed MACA-CD comprises two main components. One is a Multi-Agent Debate (MAD) based concept augmentation module. After obtaining the augmentations, the other component is a concept augmentation-enhanced CD model, which focuses on efficiently fusing the augmentations to improve CD performance.

3.1 MAD based Concept Augmentation

Figure 1 illustrates the process of the MAD based Concept Augmentation. Existing LLM-based approaches for augmenting CD, which often assume access to rich context (e.g., student interaction logs, linked exercises). We focus on addressing a more challenging cold-start scenario where only the concept name is available.

3.1.1 The components of MAD. As shown in Figure 1, we follow the standard MAD setup [27, 60], which consists of two debating

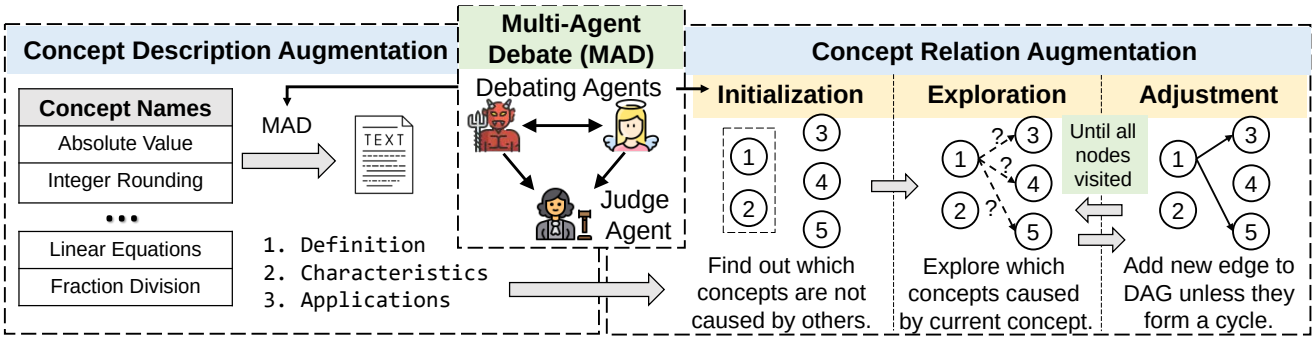


Figure 1: The process of Multi-Agent Debate based Concept Augmentation.

agents ($\mathcal{D}_0$, $\mathcal{D}_1$) and a judge agent ($\mathcal{J}$). We encourage two debating agents ($\mathcal{D}_0$ and $\mathcal{D}_1$) to generate diverse responses by setting a high temperature. In the first round, $\mathcal{D}_0$ and $\mathcal{D}_1$ receive the same task query. In subsequent rounds, each agent must reference the other's output and engage in a structured debate. The judge agent $\mathcal{J}$ evaluates the outputs:

$$\mathcal{J}(\mathcal{D}_0(q_0), \mathcal{D}_1(q_1)) = \begin{cases} \text{True}, & \text{consensus is reached} \\ \text{False}, & \text{otherwise,} \end{cases} \tag{1}$$

where q_0, q_1 denotes the input for two debating agents, respectively. If consensus is reached, $\mathcal{J}$ will end the debate and we can take one of $(\mathcal{D}_0(q_0), \mathcal{D}_1(q_1))$ as final output; otherwise another round of debate will happen. In the following sections, we will introduce the process of augmentation from both side of concept descriptions and relations.

3.1.2 Concept Description Augmentation. This process can be formulated as $\mathcal{T}^{des} = MAD(\mathcal{T}^{name})$. The initial query to MAD is designed to generate rich concept descriptions from multiple perspectives: (1) providing a **clear definition** grounded in academic or industry consensus, along with a delineation of concept boundaries; (2) listing two **core features** with concise explanations; and (3) highlighting one or two real-world examples to demonstrate **practical applications**. The process is repeated until all concepts have been successfully processed.

3.1.3 Concept Relation Augmentation. After generating the concept descriptions, we turn our focus to enhancing the relationships between concepts. While concept descriptions provide valuable contextual information, understanding how these concepts relate to one another is equally important. Following HierCDF [25], we assume that concept relations inherently involve dependencies, with certain concepts relying on others for their definition or understanding. Consequently, we model these relations as a Directed Acyclic Graph (DAG) $\mathbf{P}$, where the directionality of the edges represents the dependency relationships between concepts. A straightforward approach to discovering such structure is to employ LLM-based causal inference. However, existing methods typically require exhaustive pairwise reasoning [1, 48] or are heavily dependent on prior knowledge [3, 4].

To overcome these limitations about efficiency [51, 53], we draw inspiration from Breadth-First Search (BFS) [6, 20] and decompose the concept relation augmentation process into three sequential stages, each implemented using MAD to ensure the reliability of the generated outputs. Below, we provide a detailed introduction to the three stages:

(1) **Graph initialization stage**. We begin by identifying root concepts—those that are not causally dependent on any other concept. Formally, a concept $t \in K$ is considered a root if:

$$\forall \bar{t} \in K \setminus t, \quad \neg\text{causes}(\bar{t}, t) \tag{2}$$

All agreed-upon root concepts are inserted into a BFS queue $\mathcal{B}$ to initialize the construction of the DAG. Throughout the following two stages, we maintain two dynamic sets: $\mathcal{N}_{\text{visited}}$, which tracks the concepts that have already been explored, and $\mathcal{B}$, a queue of candidate concepts for BFS traversal.

(2) **Graph exploration stage**. In this stage, we aim to guide the process using BFS to explore child nodes for each unexplored concept in $\mathcal{B}$. First, we dequeue a concept t from the front of $\mathcal{B}$ and mark it as visited by adding it to $\mathcal{N}_{\text{visited}}$ to avoid revisiting the same concept. Next, we identify child nodes for the current concept t. The set of unexplored nodes can be represented as:

$$K \setminus (\mathcal{N}_{\text{visited}} \cup \mathcal{B}) \tag{3}$$

After determining the set of child nodes using MAD, we cannot immediately decide whether adding an edge for each child concept is appropriate. Instead, we temporarily store the identified children in a candidate set $\mathcal{N}_{\text{tmp}}$, leaving the final decision for the next stage.

(3) **Graph adjustment stage**: For each concept in $\mathcal{N}_{\text{tmp}}$, we tentatively add an edge from t to the concept $\bar{t}$ and verify whether this addition maintains the DAG property. If no cycle is introduced, the edge is retained ($p_{t,\bar{t}} = 1$); otherwise, it is discarded ($p_{t,\bar{t}} = 0$). Next, we append any newly discovered child concepts (i.e., those not already in $\mathcal{B}$) to the end of $\mathcal{B}$ for further exploration.

Note that, the exploration and adjustment stages alternate until all concepts have been visited in a BFS-consistent manner (all concepts have been added to $\mathcal{N}$_visited). This procedure progressively expands the graph while preserving its acyclic structure, forming the final DAG $\mathbf{P}$. By leveraging BFS, we avoid the inefficiency of exhaustive pairwise exploration; meanwhile, the use of MAD helps mitigate the uncertainty inherent in LLM-generated outputs.

3.2 Concept Augmentation-enhanced CD model

The next step is to integrate these enhanced concept representations into the CD model. Note that, concept embeddings $\mathbf{O}$ =

$[\mathbf{o}_1, ..., \mathbf{o}_t, ..., \mathbf{o}_T]^\top$ play a central role in modeling both student comprehension and exercise difficulty in current CD models [33, 50]. Therefore, we focus on how to incorporate external information to enhance concept embeddings $\mathbf{O}$.

3.2.1 Graph Self-Supervised Concept Fusion Layer. To effectively integrate hierarchical relations among concepts, we aim to propagate and fuse semantic information over the concept graph. To this end, we employ a LightGCN-style encoder that refines concept embeddings by aggregating signals from their graph neighbors [18]. The propagation process at the k-th layer can be defined as:

$$\mathbf{O}^{(k)} = \tilde{\mathbf{P}}\mathbf{O}^{(k-1)}, \tag{4}$$

where $\mathbf{O}^{(k)}$ and $\mathbf{O}^{(k-1)}$ denote the concept embedding matrix at k-th and $k-1$-th layers, respectively. $\tilde{\mathbf{P}}$ denotes the normalized version of the concept relation graph $\mathbf{P}$.

However, since the concept graph is constructed based on LLM-generated content, it inevitably contains noisy or spurious edges. In addition, not all relations in DAG are equally reliable or beneficial for the downstream CD task. To address this, inspired by recent advances in graph self-supervised learning [32, 57, 64], we introduce a representation-level augmentation strategy. Specifically, for each concept embedding $\mathbf{o}_t$, we generate two perturbed views at the k-th as follows:

$$\mathbf{o}_t'^{(k)} = \mathbf{o}_t^{(k)} + \Delta_t', \ \mathbf{o}_t''^{(k)} = \mathbf{o}_t^{(k)} + \Delta_t'', \tag{5}$$

where Δ_t' and Δ_t'' denote the added noise, which are subject to $\|\Delta\|_2 = \epsilon$ and $\Delta = \bar{\Delta} \odot \text{sign}(\mathbf{o}_t^{(k)})$, $\bar{\Delta} \in \mathbb{R}^d \sim U(0,1)$. This design ensures that perturbations lie on a hypersphere and remain in the same hyperoctant as the original embedding, preserving semantic consistency. The overall graph representations can be formulated as $\mathbf{o}_t = \sum_k \mathbf{o}_t^{(k)}$. To encourage robustness and invariance under such augmentations, we apply a contrastive learning objective that pulls positive pairs $(\mathbf{o}_t', \mathbf{o}_t'')$ closer while pushing apart other embeddings in the batch. The final self-supervised loss is formulated as:

$$L_{ssl} = \sum_{t=1}^{T} -log \frac{exp((\mathbf{o}_t' \cdot \mathbf{o}_t'')/\tau)}{\sum_{i=1}^{T} exp((\mathbf{o}_t' \cdot \mathbf{o}_i'')/\tau)} \tag{6}$$

where τ denotes the temperature parameter that controls the concentration level of the distribution in the contrastive softmax function. A lower τ sharpens the distribution, encouraging stronger distinctions between positive and negative pairs.

3.2.2 Pairwise Comparator based Description Fusion Layer. Although the semantic embedding $\mathbf{o}_t^{text}$ contains rich textual information, a key challenge remains: how to effectively combine it with the original ID embedding $\mathbf{o}_t$. As shown in Figure 2, to address the challenge, we adopt a simple MLP to project the textual embedding $\mathbf{o}_t^{text}$ into the same dimensional space as the ID embedding. Then, we propose a dynamically weighted fusion strategy defined as:

$$\mathbf{o}_t^{final} = (1 - g_t) \cdot MLP(\mathbf{o}_t^{text}) + g_t \cdot \mathbf{o}_t, \tag{7}$$

The key lies in determining the concept-wise weight g_t, which reflects the degree to which the ID embedding should be trusted. Our intuition is that if a concept has been sufficiently trained through student behavior data, its ID embedding $\mathbf{o}_t$ already captures meaningful information. In such cases, g_t should be higher to favor the ID embedding. Conversely, for concepts that lack sufficient optimization—indicated by their ID embedding remaining close to its initialization—we should down-weight $\mathbf{o}_t$ and rely more on the semantic embedding. To realize this adaptive fusion, we design a gating mechanism that learns the concept-wise confidence score g_t based on the relationship between the current ID embedding $\mathbf{o}_t$ and its initialization $\mathbf{o}_t^{\text{init}}$. Rather than relying on predefined metrics such as L_2 or KL divergence, we adopt a pairwise comparator network that captures richer interaction signals [9, 63], as follows:

$$g_t = max(\zeta, \sigma(\text{MLP}(\sqrt{(\mathbf{o}_t - \mathbf{o}_t^{\text{init}})^2}; \mathbf{o}_t \odot \mathbf{o}_t^{\text{init}}]))). \tag{8}$$

This design allows the model to learn flexible, non-linear patterns to identify which concept embeddings are under-trained and should rely more on textual semantics. However, one issue with this design is during initialization: at the start, the distance between $\mathbf{o}_t^{\text{init}}$ and $\mathbf{o}_t$ is too close, leading to an input that is all zeros. In this scenario, the MLP does not receive sufficient training. To address this, we introduce a truncation coefficient ζ, where if g_t is less than ζ, it is set to ζ. This ensures the model avoids the problem of receiving zero inputs during the early stages of training.

3.2.3 Diagnosis Layer. To obtain comprehension matrix C and difficulty D, we follow KaNCD [50] by combining latent embeddings of students, exercises and concepts via a simple yet effective probabilistic matrix factorization [36], formulated as:

$$c_{mt} = \sigma(< \mathbf{u}_m, \mathbf{o}_t^{final} >), \ d_{nt} = \sigma(< \mathbf{v}_n, \mathbf{o}_t^{final} >), \tag{9}$$

where σ represents the sigmoid activation function, $\mathbf{u}_m$ denotes the student m's embedding, $\mathbf{v}_n$ denotes the exercise n's embedding, and $\mathbf{o}_t$ denotes the concept t's embedding. $<,>$ denotes the inner dot. The remaining question is how to establish the relationship between representations and predicted outcomes [33, 49, 57]. In this paper, we adopt a simple yet effective diagnostic function, SimpleCD [33], which is formulated as follows:

$$\mathbf{x}_{mn} = \mathbf{Q}_n \odot (\sigma(\mathbf{c}_m) - \sigma(\mathbf{d}_n)), \tag{10}$$

where $\mathbf{x}_{mn}$ denotes the hidden vector for student m and exercise n. MLPs are used to map $\mathbf{x}_{mn}$ to final prediction $\hat{r}_{mn}$, formulated as $\hat{r}_{mn} = \text{MLPs}(\mathbf{x}_{mn})$. The comprehension matrix $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_m, \ldots, \mathbf{c}_M]^\top \in \mathbb{R}^{M \times T}$ denotes students' comprehension levels across all concepts. $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_n, \ldots, \mathbf{d}_N]^\top \in \mathbb{R}^{N \times C}$ denotes exercise difficulty. Finally, we adopt binary cross entropy loss for model optimization:

$$L_{cd} = \sum_{(m,n,r_{mn} \in R_{train})} -[r_{mn} log(\hat{r}_{mn}) + (1 - r_{mn}) log(1 - \hat{r}_{mn})], \tag{11}$$

where R_{train} denotes the training set. By combing Eq.(11) and Eq.(6), we can obtain the overall loss as follows:

$$L_{all} = L_{cd} + \beta L_{ssl}, \tag{12}$$

where β denotes the balancing parameter between CD task and self-supervised learning.

3.3 Model Discussion

We provide analyses about concept augmentation and the proposed augmentation-enhanced CD model, respectively.

Concept augmentation Process (Section 3.1): For concept description augmentation, our method requires $O(T)$ MAD calls, where T denotes the total number of concepts. In other words, although each MAD call may involve multiple rounds of agent interaction, the number of outer-loop calls scales linearly with T. As for relation augmentation, our method adopts a batch reasoning strategy to reduce complexity. During the graph initialization stage, for each concept, we prompt to the MAD whether it is causally dependent on any other concept in the set. This requires $O(T)$ MAD calls. During the graph exploration stage, we dequeue one concept

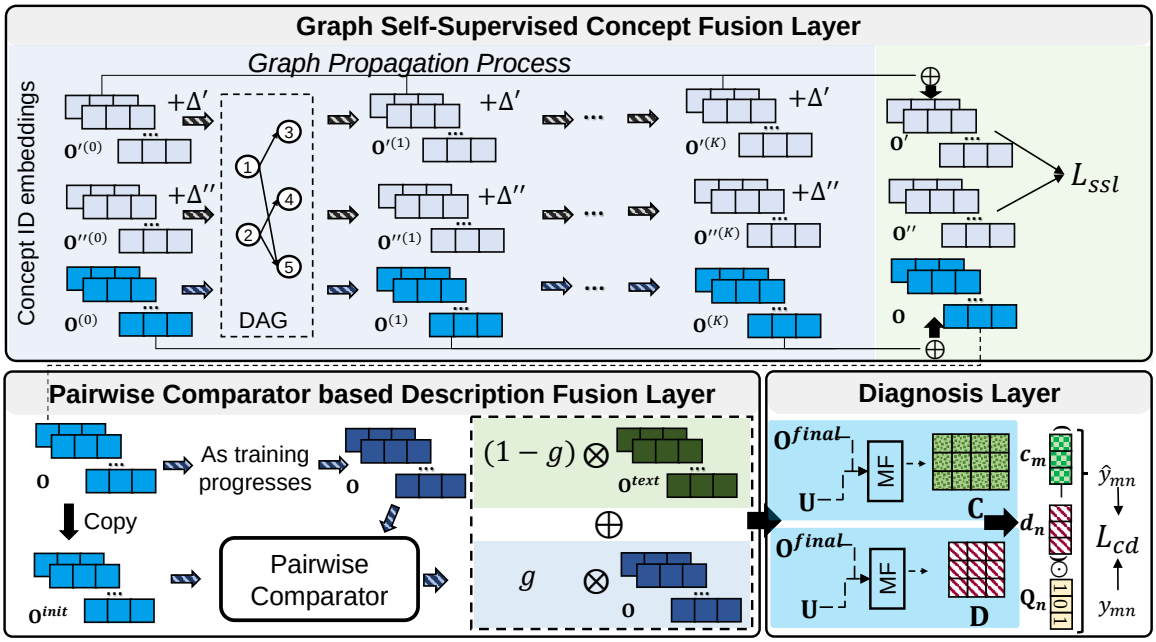


Figure 2: The overall structure of Concept Augmentation-enhanced CD model.

and query its causal links to the remaining unexplored concepts in batch at each step, leading to $O(T)$ total MAD calls. During the graph adjustment stage, we validate whether adding a new edge would introduce a cycle. This is also done in batch per parent node, maintaining $O(T)$ MAD calls overall.

Augmentation-enhanced CD model (Section 3.2): Compared to KaNCD [50], Our model introduces two augmentation-driven modules to enhance concept representations. For the graph self-supervised concept fusion, at each layer, we perform sparse matrix multiplication, and the additional time complexity is $O(|\mathcal{E}| \cdot d)$, where $|\mathcal{E}|$ denotes the number of edges in concept DAG $\mathbf{P}$ and d is the dimension of $\mathbf{O}$. The self-supervised learning loss involves comparing perturbed views of concept embeddings, which requires $O(T^2 \cdot d)$ operations. For the description fusion, the gating mechanism is computed per concept, yielding a lightweight complexity of $O(T \cdot d)$. Due to page limitations, we have placed other models discussions in Appendix, e.g., pseudo codes for the concept relation augmentation process.

4 Experiments

In the section, we try to answer these Research Questions (RQ):

- **RQ1:** Does MACA-CD outperform baselines across all datasets? (Section 4.2)
- **RQ2:** Are both components in MACA-CD essential for cognitive diagnosis? (Section 4.3)
- **RQ3:** How does the choice of LLM backbones influence CD performance? (Section 4.4)
- **RQ4:** How does MACA-CD perform under different hyperparameter configurations? (Section 4.5)
- **RQ5:** Does MACA-CD maintain its competitiveness when evaluated with alternative DOA metrics? (Section 4.6)

Table 1: The detailed statistics of three datasets.

Dataset	Eedi-2020	SLP	XES3G5M
#Students	4,918	1,760	2,000
#Exercises	948	1,346	1,624
#Concepts	53	60	2,41
#Response Logs	1,382,727	135,788	207,204
Q Density	1	1	1
#Sparsity	98.922%	94.268%	93.621%

4.1 Experimental Settings

4.1.1 Datasets. We select three publicly available datasets for the experiments, i.e., Eedi-2020[1], SLP [65], and XES3G5M [34]. All these three datasets provide three essential elements: students' response logs on exercises, exercise-concept relations Q, and concept names. The detailed statistics of these datasets are in Table 1. Sparsity denotes the sparsity of the dataset, calculated as $\frac{|R|}{M \cdot N}$. Q Density denotes the average number of concepts per exercise. To establish well-defined experimental settings, we performed careful data curation: for the Eedi-2020 dataset, we specifically selected response logs from Tasks 3 and 4, excluding students with fewer than 10 response records to ensure data reliability. The SLP dataset was constructed by combining response logs from three core disciplines (Biology, Mathematics, and Physics), similarly applying the 10-response minimum threshold for student inclusion. Following the established practice in [33], we randomly sampled 2,000 students from the XES3G5M dataset - a substantial sample size that adequately supports both model training and evaluation in cognitive diagnosis

[1]https://eedi.com/projects/neurips-education-challenge

research. To maintain conceptual clarity, our analysis exclusively focused on leaf-node knowledge concepts for the XES3G5M dataset.

4.1.2 Evaluation and Settings. Following previous works [13, 57], we adopt two widely-used evaluation metrics: Area Under the Curve (AUC) [37], and Degree of Agreement (DOA) [30, 50]. While AUC assess predictive performance, DOA specifically quantifies the alignment between predicted concept comprehension levels and actual student response patterns. Consistent with the methodology in [43, 44], we primarily report DOA results on the testing set in our main findings, reserving discussion of alternative DOA configurations for subsequent experimental analyses. This multi-metric approach enables comprehensive evaluation of both overall performance and fine-grained concept mastery prediction.

To systematically evaluate performance of MACA-CD under different settings, we adopt the following three distinct data partitioning strategies: (1) A **standard** random split where all response logs are divided into training and test sets at the ratio of 8:2; (2) A **cold**-start scenario where students are first split equally into warm and cold groups - all response logs from warm students are included in the training set, while for each cold student, a maximum of 2 response logs are allocated to training with the remainder assigned to the testing set; (3) A **zero-shot** concept scenario where 20% of concepts are randomly selected and all associated response logs are exclusively placed in the test set, ensuring the training set contains no information about these withheld concepts. These three experimental settings allows us to rigorously assess MACA-CD 's capabilities across standard, cold-start, and zero-shot scenarios.

4.1.3 Baselines. We select the following representative CD models as baselines:

- **NCDM** [49]. It applies neural networks into CD, and uses high-dimensional representations to represent students' abilities.
- **KaNCD** [50]. It adopts matrix factorization techniques on representations of student abilities and exercise difficulties.
- **ISG-CD** [43]. It focuses on how to better utilize the student-exercise bipartite graph based on subgraph construction and information bottleneck principles.
- **IDCD** [23]. It introduces an encoder-decoder based diagnostic module with inductive learning to guarantee identifiability.
- **DFCD** [33]. It leverages LLMs to refine exercise/concept coherence, then fuses these semantic features with response-specific patterns to enhance CD.

It should be noted that MIRT lacks the capability for concept-specific diagnosis, as it represents student ability through undifferentiated latent embeddings and thus cannot infer comprehension levels on individual knowledge concepts (nor measure DOA). For fair comparison with DFCD, which enhances both concept and exercise texts through textual and behavioral data—we retain only its concept-side augmentation in our experiments, as our method does not utilize exercise texts.

4.1.4 Hyperparameter Settings. For the MAD component, we employ Qwen3-Turbo [62] with its temperature parameter set to 0.5 to promote diversity across all datasets. Regarding the CD model configuration: (1) we maintain a consistent batch size of 8,192 student-exercise response logs through random sampling; (2) the learning rate is tuned from {0.0001, 0.0005, 0.001, 0.005, 0.01} for all models, with the Adam optimizer [21] universally applied; (3) the free embeddings (i.e., $\mathbf{u}_m$, $\mathbf{v}_n$, and $\mathbf{o}_t$) are fixed at 16 dimensions. For graph self-supervised learning, we perform hyperparameter searches on: (i) the graph propagation depth {1,2,3,4} in Eq.(4), (ii) the balancing coefficient β in Eq.12 across {0.1, 0.01, 0.001, 0.0001, 0}, while fixing the contrastive temperature τ in Eq.6 at 0.2. The threshold in Eq. (8) is searched from {0, 0.2, 0.4, 0.6, 0.8}.

4.2 Overall Performance

We report overall performance in Table 2. There are several observations from these three tables.

- First, our proposed MACA-CD consistently outperforms baselines across all three scenarios. It yields about 1% AUC improvement for cold-start students, over 0.6% gains in standard settings, and achieves the largest improvements—at least 5% AUC—in the challenging zero-shot concept scenario.
- Second, while ISG-CD shows relatively strong performance in standard random scenarios, this advantage disappears in cold-start and zero-shot conditions. For instance, its AUC drops below KaNCD's by nearly 5% for cold-start students on the Eedi-2020 dataset. This limitation occurs because ISG-CD requires existing edges - a condition unmet when handling new students or exercises related to zero-shot concepts in the testing set.
- Third, our analysis reveals DFCD also demonstrates strong performance across all scenarios. This outcome serves dual significance: (1) it validates the necessity of using LLMs for educational data augmentation, and (2) confirms the superior effectiveness of the MAD based concept augmentation process compared to direct prompting.
- Finally, MACA-CD shows substantially larger gains in cold-start and zero-shot settings than in standard scenarios. This supports the principle in Eq. 8: when sufficient behavioral data are available, cognitive diagnosis models already perform well, whereas augmentation becomes crucial in data-sparse cases. These results confirm that our data augmentation pipeline is essential for practical educational scenarios with limited behavioral data.

4.3 Ablation Study

The core idea of MACA-CD lies in how to utilize Concept Description and Relation Augmentation to enhance Cognitive Diagnosis (CD). In this section, we conducted ablation studies to analyze whether each of these two modules is effective individually. Specifically, we considered the following experimental setups: when neither Section 3.1.1 nor Section 3.1.2 is applied, the model is equivalent to SimpleCD (a simplified model proposed in [33]); when Section 3.1.1 and Section 3.1.2 are applied individually, and when both modules are used together (i.e., MACA-CD). The results of these ablation studies on the SLP dataset are shown in Figure 3.

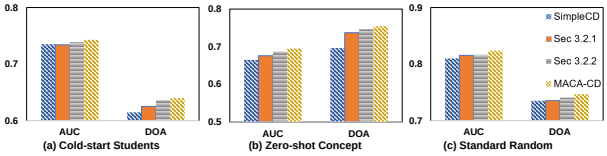


Figure 3: Ablation Study on the SLP dataset.

Table 2: Overall performance under various scenarios across three datasets. The optimal results are marked in bold, while the second-best are underlined. Asterisks (*) denote performance improvements that are statistically significant.

Model \ Metric	Eedi-2020		SLP		XES3G5M	
	AUC	DOA	AUC	DOA	AUC	DOA
	Cold-start Students Scenario					
NCDM	0.6699	0.5237	0.7224	0.5564	0.6881	0.4944
KaNCD	0.6903	0.5755	0.7348	0.5662	0.6916	0.5122
ISG-CD	0.6562	0.5365	0.7092	0.5952	0.6931	0.5658
IDCD	0.6718	0.5616	0.7263	0.5702	0.6942	0.5495
DFCD	0.6922	0.5942	0.7359	0.6341	0.6987	0.5762
MACA-CD	**0.7024***	**0.6131***	**0.7428***	**0.6396***	**0.7072***	**0.6047***
	Zero-shot Concept Scenario					
NCDM	0.4961	0.4897	0.5012	0.5155	0.5134	0.5089
KaNCD	0.4993	0.4535	0.5127	0.5216	0.5038	0.4825
ISG-CD	0.6575	0.6849	0.5477	0.5505	0.5641	0.5643
IDCD	0.5727	0.7033	0.5448	0.5401	0.5120	0.6614
DFCD	0.6623	0.7244	0.6389	0.7054	0.6085	0.6842
MACA-CD	**0.6991***	**0.7263***	**0.6952***	**0.7544***	**0.7068***	**0.7024***
	Standard Random Scenario					
NCDM	0.7632	0.6716	0.8110	0.7406	0.7351	0.5477
KaNCD	0.7724	0.7164	0.8153	0.7413	0.7681	0.6360
ISG-CD	0.7788	0.7202	0.8187	0.7429	0.7782	0.6706
IDCD	0.7762	0.7218	0.8171	0.7402	0.7655	0.6602
DFCD	0.7780	0.7211	0.8168	0.7425	0.7704	0.6688
MACA-CD	**0.7844***	**0.7235***	**0.8236***	**0.7466***	**0.7930***	**0.7056***

From the ablation studies, we observe the following key findings: First, both modules prove to be effective across different scenarios, indicating that both are necessary. Second, we found that the effects of the two modules are most significant in the zero-shot concept scenario, while the improvements are least noticeable in the random split scenario. This empirically validates the core principle of Eq. (8): when sufficient behavioral data exists (standard scenario), cognitive diagnosis models can achieve adequate accuracy without augmentation. However, in data-sparse real-world cases (cold-start/zero-shot), the augmentation becomes crucial, confirming that our data augmentation pipeline is essential for handling practical educational complexities where behavioral data is insufficient.

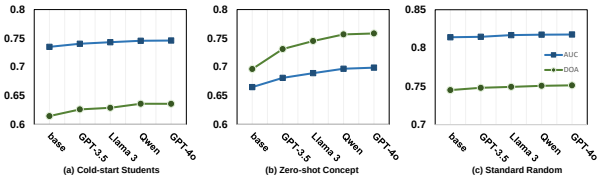


Figure 4: Varying LLM backbones on the SLP dataset.

4.4 Performance with varying LLM backbones

We systematically evaluate the performance of the MACA framework with various LLMs as backbones to assess its practical utility. This includes SimpleCD (base), which operates without external information, as well as mainstream LLM backbones like GPT-3.5 [5], Llama 3 [16], Qwen3-Turbo [62], and GPT-4o [2].

Key findings from Figure 4 are as follows: First, the quality of cognitive annotations generated by different LLMs has a significant impact on the downstream CD task. Performance improvements are most notable in zero-shot scenarios, where the system can leverage the cognitive annotations in the absence of task-specific training data. In contrast, in standard random splits, the improvements are minimal, with less than a 0.5% AUC improvement between GPT-3.5, Llama 3, and SimpleCD, indicating that for well-established tasks, cognitive annotations provide limited value beyond the baseline models. Second, while GPT-4o theoretically holds the potential for optimal performance, Qwen3-Turbo demonstrates superior efficiency, delivering comparable efficacy with lower cost. This advantage, coupled with its ability to handle larger-scale tasks without significant loss in performance, makes Qwen3-Turbo the preferred backbone for the final backbone.

4.5 Hyperparameter Analyses

During the training of the model, two crucial parameters play a significant role: ζ in Eq. (8) and the balancing parameter β in Eq. (12). These two parameters are key to the fusion process of the description and relation. We analyze these parameters through experiments to determine the conditions under which our model performs best.

Several important findings can be observed from Figure 5. First, we find that when the balancing parameter β takes values between $\{0.0001, 0.001, 0.01, 0.1\}$, the model performance remains relatively

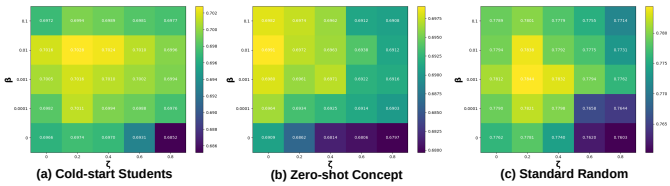


Figure 5: AUC performance when varying β and ζ on the Eedi-2020 dataset.

consistent. However, when $\beta = 0$, the model performance rapidly declines. This highlights the positive role of the self-supervised loss. Based on our experimental observations, we conclude that introducing the self-supervised loss helps ensure the uniformity of representations, preventing the model from quickly overfitting. Instead, the model gradually explores better parameters within a relatively reasonable search space, ultimately improving performance. Second, we found that the best performance occurs when ζ is set to a relatively low value, such as 0.2. This is because the introduction of ζ is meant to prevent $\mathbf{o}_t$ from being unoptimized, which would result in the MLP input in Eq. (8) being all zeros. A slight adjustment to the lower bound suffices to achieve this goal. Increasing ζ forces the free embedding to contribute more during training, but this may not be a practical requirement. Third, when comparing different scenarios, we find that the zero-shot concept scenario behaves differently from the other two. In this scenario, the optimal result is obtained when $\zeta = 0$. The reason for this is that, in the training set, the concepts corresponding to zero-shot concepts do not have any corresponding training response logs. Consequently, these parameters cannot be optimized. Therefore, not forcing them to participate in the final embeddings leads to better performance.

4.6 DOA Analyses

In Table 2, we adopt the classic DOA calculation formula [50] and record its consistency on the testing set. We note that some researchers emphasize different approaches to DOA calculation. For instance, Chen et al. highlight that DOA should be validated for consistency across all response logs, including both training and testing sets [10]. Further, researchers introduced DOA@K, an improvement over the classic DOA [29, 33]. The main difference is that DOA@K focuses only on the top K concepts with the highest number of response logs in the experiment. This improvement enhances operational efficiency and addresses issues where certain knowledge points have insufficient response logs, leading to less accurate results.

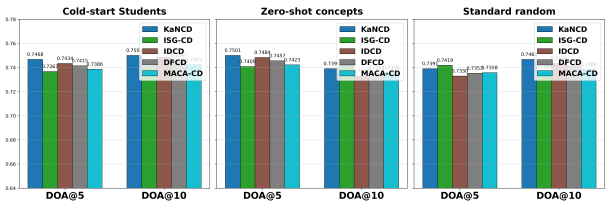


Figure 6: DOA analyses on the Eedi-2020 dataset.

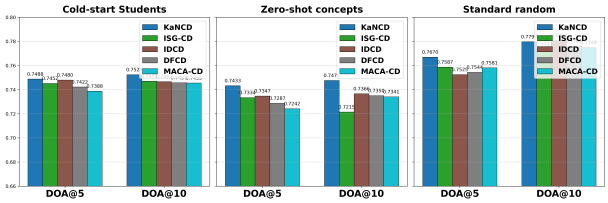


Figure 7: DOA analyses on the SLP dataset.

Figures 6 and 7 display the results of DOA@5 and DOA@10 on the Eedi-2020 and SLP datasets, respectively. Analysis shows that, while our results are not the best, they maintain a very strong performance. This finding aligns with the trend observed in historical works [33, 50], which suggests that initially, DOA and AUC move in the same direction, but as overfitting on the training set increases, a discrepancy between the two metrics starts to emerge. We believe this phenomenon is due to class imbalance. When one paradigm dominates, its influence grows stronger as training progresses, causing certain logs to have less impact, and ultimately leading the model to produce diagnostic results that contradict these samples. Such contradictions are reflected directly in the DOA@K values. It is also important to highlight that, across various scenarios on different datasets, our method consistently demonstrates superior performance on the DOA@K metric. Combined with the overall performance data in Table 2, our proposed MACA-CD not only achieves significant improvement in accuracy but also yields good results in DOA, reflecting the effectiveness of our approach.

5 Conclusion and Future Work

Current Cognitive Diagnosis models are often limited by the availability and quality of students' behavioral data, which can significantly hinder their ability to provide accurate assessments. Existing methods tend to rely heavily on such data, making them less effective in scenarios like cold-start situations. While recent advancements in LLM based augmentation have shown promise in enhancing CD, the reliability and accuracy of LLM-generated annotations remain challenging. In this paper, we proposed MACA-CD, a novel approach that generates and fuses reliable concept descriptions and relations based solely on concept names. By leveraging Multi-Agent Debate, MACA-CD reduced the reliance on behavioral data and ensures more accurate and stable concept representations. Experimental results on three real-world datasets demonstrated that MACA-CD consistently outperforms existing methods, providing robust and reliable diagnosis across various practical scenarios. Future work will explore the application of LLM-based data augmentation in other educational domains where data may be limited or difficult to obtain.

Acknowledgments

This work has been supported by grants from the National Natural Science Foundation of China under Grant 72188101 and Grant U22A2094.

References

[1] Ahmed Abdulaal, Adamos Hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C Castro, and Daniel C Alexander. 2024. Causal modelling agents: Causal graph discovery through synergising metadata-and data-driven reasoning. In *12th International Conference on Learning Representations, ICLR 2024*, Vol. 2024. International Conference on Learning Representations (ICLR).

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. 2025. Integrating large language model for improved causal discovery. *IEEE Transactions on Artificial Intelligence* (2025).

[4] Taiyu Ban, Lyvzhou Chen, Xiangyu Wang, and Huanhuan Chen. 2023. From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. *arXiv preprint arXiv:2306.16902* (2023).

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Alan Bundy and Lincoln Wallen. 1984. Breadth-first search. In *Catalogue of artificial intelligence tools*. Springer, 13–13.

[7] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. [n. d.]. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. In *The Twelfth International Conference on Learning Representations*.

[8] Haw-Shiuan Chang, Hwai-Jung Hsu, and Kuan-Ta Chen. 2015. Modeling Exercise Relationships in E-Learning: A Unified Approach.. In *EDM*. 532–535.

[9] Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced LSTM for natural language inference. *arXiv preprint arXiv:1609.06038* (2016).

[10] Xiangzhi Chen, Le Wu, Fei Liu, Lei Chen, Kun Zhang, Richang Hong, and Meng Wang. 2023. Disentangling Cognitive Diagnosis with Limited Exercise Labels. In *Thirty-seventh Conference on Neural Information Processing Systems*.

[11] Zhiang Dong, Jingyuan Chen, and Fei Wu. 2025. Knowledge is power: Harnessing large language models for enhanced cognitive diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 164–172.

[12] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

[13] Weibo Gao, Qi Liu, Zhenya Huang, Yu Yin, Haoyang Bi, Mu-Chun Wang, Jianhui Ma, Shijin Wang, and Yu Su. 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*. 501–510.

[14] Weibo Gao, Qi Liu, Hao Wang, Linan Yue, Haoyang Bi, Yin Gu, Fangzhou Yao, Zheng Zhang, Xin Li, and Yuanjing He. 2024. Zero-1-to-3: Domain-Level Zero-Shot Cognitive Diagnosis via One Batch of Early-Bird Students towards Three Diagnostic Objectives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8417–8426.

[15] Weibo Gao, Qi Liu, Linan Yue, Fangzhou Yao, Rui Lv, Zheng Zhang, Hao Wang, and Zhenya Huang. 2025. Agent4edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 23923–23932.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[17] Yong Guan, Hao Peng, Lei Hou, and Juanzi Li. 2025. Mmd-ere: multi-agent multi-sided debate for event relation extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*. 6889–6896.

[18] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 639–648.

[19] Mahmood Hegazy, Aaron Rodrigues, and Azzam Naeem. 2025. MAFA: A multi-agent framework for annotation. *arXiv preprint arXiv:2505.13668* (2025).

[20] Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. [n. d.]. Efficient Causal Graph Discovery Using Large Language Models. In *ICLR 2024 Workshop: How Far Are We From AGI*.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Luc T Le. 2009. Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing* 9, 2 (2009), 122–133.

[23] Jiatong Li, Qi Liu, Fei Wang, Jiayu Liu, Zhenya Huang, Fangzhou Yao, Linbo Zhu, and Yu Su. 2024. Towards the identifiability and explainability for personalized learner modeling: an inductive paradigm. In *Proceedings of the ACM Web Conference 2024*. 3420–3431.

[24] Jiatong Li, Qi Liu, and Mengxiao Zhu. 2025. Generative Cognitive Diagnosis. *arXiv preprint arXiv:2507.09831* (2025).

[25] Jiatong Li, Fei Wang, Qi Liu, Mengxiao Zhu, Wei Huang, Zhenya Huang, Enhong Chen, Yu Su, and Shijin Wang. 2022. HierCDF: A Bayesian Network-based Hierarchical Cognitive Diagnosis Framework. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 904–913.

[26] Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving Multi-Agent Debate with Sparse Communication Topology. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 7281–7294.

[27] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17889–17904.

[28] Mengfan Liu, Pengyang Shao, and Kun Zhang. 2021. Graph-based exercise- and knowledge-aware learning network for student performance prediction. In *Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1*. Springer, 27–38.

[29] Shuo Liu, Junhao Shen, Hong Qian, and Aimin Zhou. 2024. Inductive cognitive diagnosis for fast student learning in web-based intelligent education systems. In *Proceedings of the ACM Web Conference 2024*. 4260–4271.

[30] Shuo Liu, Zihan Zhou, Yuanhao Liu, Jing Zhang, and Hong Qian. 2025. Language Representation Favored Zero-Shot Cross-Domain Cognitive Diagnosis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 836–847.

[31] Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051* (2024).

[32] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip S Yu. 2022. Graph self-supervised learning: A survey. *IEEE transactions on knowledge and data engineering* 35, 6 (2022), 5879–5900.

[33] Yuanhao Liu, Shuo Liu, Yimeng Liu, Chanjin Zheng, Wei Zhang, and Hong Qian. 2025. A dual-fusion cognitive diagnosis framework for open student learning environments. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2* (2025).

[34] Zitao Liu, Qiongqiong Liu, Teng Guo, Jiahao Chen, Shuyan Huang, Xiangyu Zhao, Jiliang Tang, Weiqi Luo, and Jian Weng. 2023. Xes3g5m: A knowledge tracing benchmark dataset with auxiliary information. *Advances in Neural Information Processing Systems* 36 (2023), 32958–32970.

[35] Haokai Ma, Yimeng Yang, Lei Meng, Ruobing Xie, and Xiangxu Meng. 2024. Multimodal conditioned diffusion model for recommendation. In *Companion Proceedings of the ACM Web Conference 2024*. 1733–1740.

[36] Andriy Mnih and Russ R Salakhutdinov. 2007. Probabilistic matrix factorization. *Advances in neural information processing systems* 20 (2007).

[37] Joel Myerson, Leonard Green, and Missaka Warusawitharana. 2001. Area under the curve as a measure of discounting. *Journal of the experimental analysis of behavior* 76, 2 (2001), 235–243.

[38] Yilmazcan Ozyurt, Tunaberk Almaci, Stefan Feuerriegel, and Mrinmaya Sachan. 2025. Personalized Exercise Recommendation with Semantically-Grounded Knowledge Tracing. *arXiv preprint arXiv:2507.11060* (2025).

[39] Yilmazcan Ozyurt, Stefan Feuerriegel, and Mrinmaya Sachan. 2024. Automated knowledge concept annotation and question representation learning for knowledge tracing. *arXiv preprint arXiv:2410.01727* (2024).

[40] Chen Qian, Zihao Xie, YiFei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. [n. d.]. Scaling Large Language Model-based Multi-Agent Collaboration. In *The Thirteenth International Conference on Learning Representations*.

[41] Hong Qian, Shuo Liu, Mingjia Li, Bingdong Li, Zhi Liu, and Aimin Zhou. 2024. ORCDF: An Oversmoothing-Resistant Cognitive Diagnosis Framework for Student Learning in Online Education Systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2455–2466.

[42] Pengyang Shao, Le Wu, Kun Zhang, Defu Lian, Richang Hong, Yong Li, and Meng Wang. 2024. Average user-side counterfactual fairness for collaborative filtering. *ACM Transactions on Information Systems* 42, 5 (2024), 1–26.

[43] Pengyang Shao, Yonghui Yang, Chen Gao, Lei Chen, Kun Zhang, Chenyi Zhuang, Le Wu, Yong Li, and Meng Wang. 2025. Exploring Heterogeneity and Uncertainty for Graph-based Cognitive Diagnosis Models in Intelligent Education. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 1233–1243.

[44] Pengyang Shao, Kun Zhang, Chen Gao, Lei Chen, Miaomiao Cai, Le Wu, Yong Li, and Meng Wang. 2025. Breaking student-concept sparsity barrier for cognitive diagnosis. *Frontiers of Computer Science* 19, 11 (2025), 1911363.

[45] Fei Shen, Xin Jiang, Xin He, Hu Ye, Cong Wang, Xiaoyu Du, Zechao Li, and Jinhui Tang. 2025. Imagdressing-v1: Customizable virtual dressing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6795–6804.

[46] Fei Shen and Jinhui Tang. 2024. Imagpose: A unified conditional framework for pose-guided person generation. *Advances in neural information processing systems* 37 (2024), 6246–6266.
[47] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. 2024. Advancing Pose-Guided Image Synthesis with Progressive Conditional Diffusion Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=rHzapPnCgT
[48] Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. 2024. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454* (2024).
[49] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yuying Chen, Yu Yin, Zai Huang, and Shijin Wang. 2020. Neural cognitive diagnosis for intelligent education systems. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6153–6161.
[50] Fei Wang, Qi Liu, Enhong Chen, Zhenya Huang, Yu Yin, Shijin Wang, and Yu Su. 2022. NeuralCD: a general framework for cognitive diagnosis. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[51] Maolin Wang, Jun Chu, Sicong Xie, Xiaoling Zang, Yao Zhao, Wenliang Zhong, and Xiangyu Zhao. 2025. Put Teacher in Student's Shoes: Cross-Distillation for Ultra-compact Model Compression Framework. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*. 4975–4985.
[52] MAOLIN WANG, YINGYI ZHANG, CUNYIN PENG, YICHENG CHEN, WEI ZHOU, JINJIE GU, CHENYI ZHUANG, RUOCHENG GUO, BOWEN YU, WANYU WANG, et al. 2025. Function Calling in Large Language Models: Industrial Practices, Challenges, and Future Directions. (2025).
[53] Maolin Wang, Yao Zhao, Jiajia Liu, Jingdong Chen, Chenyi Zhuang, Jinjie Gu, Ruocheng Guo, and Xiangyu Zhao. 2023. Large multimodal model compression via efficient pruning and distillation at AntGroup. *arXiv preprint arXiv:2312.05795* (2023).
[54] Shanshan Wang, Ying Hu, Xun Yang, Zhongzhou Zhang, Keyang Wang, and Xingyi Zhang. 2024. Personalized forgetting mechanism with concept-driven knowledge tracing. *arXiv preprint arXiv:2404.12127* (2024).
[55] Shanshan Wang, Fangzheng Yuan, Keyang Wang, Xun Yang, Xingyi Zhang, and Meng Wang. 2025. Dual-state personalized knowledge tracing with emotional incorporation. *IEEE Transactions on Knowledge and Data Engineering* (2025).
[56] Shanshan Wang, Zhen Zeng, Xun Yang, Ke Xu, and Xingyi Zhang. 2024. Boosting neural cognitive diagnosis with student's affective state modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 620–627.
[57] Shanshan Wang, Zhen Zeng, Xun Yang, and Xingyi Zhang. 2023. Self-supervised graph learning for long-tailed cognitive diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 110–118.
[58] Shanshan Wang, Xueying Zhang, Xun Yang, Xingyi Zhang, and Keyang Wang. 2025. Learning states enhanced Knowledge Tracing: Simulating the diversity in real-world learning process. *Expert Systems with Applications* 274 (2025), 126838.
[59] Kaifang Wu, Yonghui Yang, Kun Zhang, Le Wu, Jing Liu, and Xin Li. 2022. Multi-Relational Cognitive Diagnosis for Intelligent Education. In *CAAI International Conference on Artificial Intelligence*. Springer, 425–437.
[60] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.
[61] Xiao Xia, Dan Zhang, Zibo Liao, Zhenyu Hou, Tianrui Sun, Jing Li, Ling Fu, and Yuxiao Dong. 2025. Scenegenagent: Precise industrial scene generation with coding agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 17847–17875.
[62] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).
[63] Junhao Ye, Jiahui Li, Lu Chen, Yuren Mao, Yunjun Gao, and Tianyi Li. 2024. LEAP: A Low-cost Spark SQL Query Optimizer using Pairwise Comparison. *Proceedings of the VLDB Endowment* 18, 3 (2024), 675–687.
[64] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1294–1303.
[65] LU Yu, PIAN Yang, and SHEN Ziding. 2021. SLP: A multi-dimensional and consecutive dataset from k-12 education. In *International Conference on Computers in Education*.
[66] Yongfu Zha, Xinxin Dong, Haokai Ma, Yonghui Yang, and Xiaodong Wang. 2025. Align-for-Fusion: Harmonizing Triple Preferences via Dual-oriented Diffusion for Cross-domain Sequential Recommendation. *arXiv preprint arXiv:2508.05074* (2025).
[67] Dan Zhang, Sining Zhoubian, Min Cai, Fengzu Li, Lekang Yang, Wei Wang, Tianjiao Dong, Ziniu Hu, Jie Tang, and Yisong Yue. 2025. Datascibench: An llm agent benchmark for data science. *arXiv preprint arXiv:2502.13897* (2025).
[68] Shaowei Zhang and Deyi Xiong. 2025. Debate4MATH: Multi-Agent Debate for Fine-Grained Reasoning in Math. In *Findings of the Association for Computational Linguistics: ACL 2025*. 16810–16824.
[69] Yijie Zhou, Xianhui Cheng, Qiming Zhang, Lei Wang, Wenchao Ding, Xiangyang Xue, Chunbo Luo, and Jian Pu. 2024. Algpt: Multi-agent cooperative framework for open-vocabulary multi-modal auto-annotating in autonomous driving. *IEEE Transactions on Intelligent Vehicles* (2024).
[70] Yan Zhuang, Qi Liu, GuanHao Zhao, Zhenya Huang, Weizhe Huang, Zachary Pardos, Enhong Chen, Jinze Wu, and Xin Li. 2023. A Bounded Ability Estimation for Computerized Adaptive Testing. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=tAwjG5bM7H

A Model Discussion

A.1 Detailed Procedures of Concept Relation Augmentation

We present detailed pseudo codes for the concept relation augmentation process in Algorithm 1. The proposed algorithm efficiently constructs a reliable concept DAG by combining the strengths of MAD and BFS. It avoids exhaustive pairwise comparisons by adopting a structured, batched reasoning strategy that incrementally expands the graph. This significantly improves computational efficiency without compromising reasoning depth. Meanwhile, the MAD framework ensures semantic agreement between agents before committing any relation, enhancing the quality and consistency of the generated structure.

Algorithm 1: Concept Relation Augmentation via BFS

Input: Concept set $K = \{t_1, t_2, \ldots, t_T\}$
Output: Concept relation DAG $\mathbf{P}$
$\mathcal{N}_{\text{visited}} \leftarrow \emptyset, \mathcal{B} \leftarrow \emptyset, \mathbf{P} \leftarrow \emptyset$; `// Initialization`
Stage 1: Graph Initialization;
foreach $t \in K$ **do**
  Query MAD: Is t causally dependent on any $\bar{t} \in K \setminus \{t\}$?;
  if *MAD returns False* **then**
    $\mathcal{B}$.enqueue(t);
Stage 2-3: Graph Exploration and Adjustment;
while $\mathcal{B} \neq \emptyset$ **do**
  $t \leftarrow \mathcal{B}$.dequeue();
  $\mathcal{N}_{\text{visited}}$.add($t$);
  $\mathcal{N}_{\text{tmp}} \leftarrow \emptyset$;
  foreach $\bar{t} \in K \setminus (\mathcal{N}_{visited} \cup \mathcal{B})$ **do**
    Query MAD: Is t a prerequisite of $\bar{t}$?;
    if *MAD returns True* **then**
      $\mathcal{N}_{\text{tmp}}$.add($\bar{t}$);
  foreach $\bar{t} \in \mathcal{N}_{tmp}$ **do**
    Temporarily add edge $(t \rightarrow \bar{t})$ to $\mathbf{P}$;
    if $\mathbf{P}$ *remains acyclic* **then**
      Confirm $p_{t,\bar{t}} = 1$;
      if $\bar{t} \notin \mathcal{B}$ **then**
        $\mathcal{B}$.enqueue($\bar{t}$);
    else
      Remove edge $(t \rightarrow \bar{t})$, set $p_{t,\bar{t}} = 0$;
return concept DAG $\mathbf{P}$