Front. Digit. Educ., 2025, 2(2): 17
https://doi.org/10.1007/s44366-025-0054-y

RESEARCH ARTICLE

Large Language Models Are Zero-Shot Cross-Domain Diagnosticians in Cognitive Diagnosis

Haiping Ma[a,b], **Changqian Wang**[a], **Siyu Song**[a], **Shangshang Yang**[b], **Limiao Zhang**[a], **Xingyi Zhang**[a]

[a] National Key Laboratory of Oto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei 230601, China

[b] School of Computer Science and Technology, Anhui University, Hefei 230601, China

Abstract With the rapid development of online education, cognitive diagnosis has become a key task in intelligent education, particularly for student ability assessments and resource recommendations. However, existing cognitive diagnosis models face the diagnostic system cold-start problem, whereby there are no response logs in new domains, making accurate student diagnosis challenging. This research defines this task as zero-shot cross-domain cognitive diagnosis (ZCCD), which aims to diagnose students' cognitive abilities in the target domain using only the response logs from the source domain without prior interaction data. To address this, a novel paradigm, large language model (LLM)-guided cognitive state transfer (LCST) is proposed, which leverages the powerful capabilities of LLMs to bridge the gap between the source and target domains. By modelling cognitive states as natural language tasks, LLMs act as intermediaries to transfer students' cognitive states across domains. The research uses advanced LLMs to analyze the relationships between knowledge concepts and diagnose students' mastery of the target domain. The experimental results on real-world datasets shows that the LCST significantly improves cognitive diagnostic performance, which highlights the potential of LLMs as educational experts in this context. This approach provides a promising direction for solving the ZCCD challenge and advancing the application of LLMs in intelligent education.

Keywords cognitive diagnosis, large language model, prompt engineering, AI in education, intelligent tutoring systems

1 Introduction

In recent years, with the development of online education, cognitive diagnosis has become an important research task in the application of intelligent education (Liu et al., 2023; Yang et al., 2024b; Yang et al., 2024c). Cognitive diagnosis is widely used in student ability assessments and resource recommendations (Yu et al., 2024a; Yu et al., 2024b), with its goal being to diagnose students' mastery of knowledge concepts associated with exercises through a large number of students–exercise response logs, which facilitates personalized learning (Ma et al., 2024a; Zhang et al., 2024). The existing work is primarily based on research within well-established knowledge domains with the available response logs of students and exercises. However, for knowledge concepts in a new subject (hereinafter referred to as a domain) without response logs, existing methods struggle to provide accurate diagnoses of students. Nevertheless, research on the cold-start problem of diagnostic systems in the domain with no interaction logs is relatively scarce. As shown in Figure 1, the mathematics domain encompasses a vast number of knowledge concepts. Existing cognitive diagnosis models (CDMs) analyze exercises associated with these knowledge concepts and a large set of student responses to generate detailed diagnostic feedback for each student. For instance, the value of the diagnostic feedback indicates the student's level of mastery of these concepts, including the system of linear equations in two variables presenting 0.4, quadratic radical equation presenting 0.6, and inverse proportionality function presenting 0.3. Mastery is represented as a value between 0 and 1, where 0 implies no understanding of the concept and 1 signifies complete mastery of the ability to solve related

Received January 7, 2025; revised February 17, 2025; accepted March 3, 2025

Xingyi Zhang (✉)
E-mail: xyzhanghust@gmail.com

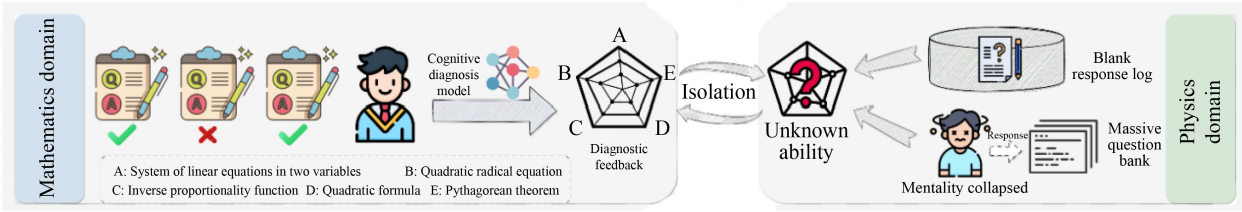


Figure 1 Illustration of the zero-shot cross-domain task.

problems with minimal errors. However, when faced with a lack of response records for knowledge concepts in the physics domain, it is difficult to accurately diagnose students' cognitive states in that domain. This is because although students can be asked to perform exercises from the target domain with no prior interaction records, the monotonous task of performing these exercises may lead to a subjective decline in their cognitive abilities. Moreover, when the set of knowledge concepts in the target domain is large or there are no corresponding exercises, diagnosing through the performance of exercises is not a feasible strategy. Furthermore, from a model perspective, CDMs rely on analyzing students' response records within a specific domain to generate diagnostic feedback, which implies that in the absence of response data for a given subject, the CDM would no longer function effectively. This task is defined in this research as zero-shot cross-domain cognitive diagnosis (ZCCD), which refers to diagnosing students' cognitive abilities in a target domain with completely blank response records, using only response logs from a source domain. ZCCD is a practical and meaningful task (Bai et al., 2025; Ma et al., 2024a). A promising approach to address this task is to build a bridge between sources and target domains and to leverage the relationships between knowledge concepts across similarity and precedence domains. This method allows students' cognitive states in the source domain to be transferred in a zero-shot manner to the target domain, therefore enabling the diagnosis of their mastery of knowledge concepts in the target domain. To address this issue, TechCD propagates students' cognitive states by pre-constructing knowledge concept graphs (KCGs) (Gao et al., 2023). However, the construction of KCGs depends heavily on information-rich interaction datasets and significant human efforts. Similarly, Zero-1-to-3 requires a substantial amount of response data from a subset of students in the target domain to propagate those students' cognitive states to cold-start students (Gao et al., 2024a). Both approaches rely on extensive prior information, which constrains their applicability in real-world scenarios.

Recently, the pretrained large language models (LLMs) have demonstrated remarkable performance in general tasks across both natural language processing (NLP) and recommendation systems (Agostinelli et al., 2023; Brown et al., 2020; Hou et al., 2024; Kojima et al., 2022; Li et al., 2024; OpenAI et al., 2023; Wang et al., 2024a; Wang et al., 2024b; Yang et al., 2023a). Trained on vast and diverse textual corpora, LLMs not only exhibit exceptional language comprehension but also possess strong reasoning and generative capabilities which enable them to capture intricate semantic structures and underlie logical relationships within text. One of the intuitive advantages of LLMs lies in their ability to emulate human thought processes. The extensive knowledge acquired during pretraining allows LLMs to generate intermediate reasoning steps, and these steps resemble the analytical processes employed by educational experts, thereby providing an additional interpretative layer for diagnostic outcomes. For instance, when evaluating the relationship between force and equilibrium force or between reflection and refraction in physics, LLMs offer detailed explanations that reflect the underlying physical principles, intuitively illustrating cross-domain knowledge dependencies. A key challenge in addressing ZCCD is identifying an intermediary that bridges the source and target domains, which necessitates an understanding of both interdisciplinary relationships and the characteristics of students' cognitive states. Leveraging their inherent world knowledge, LLMs offer a promising solution for mitigating the semantic gap between these domains. LLMs have shown exceptional summarization and reasoning capabilities in recommendation systems (Jiang et al., 2024; Ren et al., 2024). This is because after modelling user behavioural preferences as natural language descriptions, the language model understands the instructions and further completes the recommendation task. Inspired by this, this research views the cognitive modelling of students in the source domain as a language modelling task. After acquiring the student's prior cognitive state, the diagnostic task is formatted into a natural language text, which is then treated as a language-processing task. Therefore, LLMs cannot only serve as intermediaries between the source and target domains but also replace the CDM in the target domain. To this end, a novel diagnostic paradigm for the ZCCD task is proposed to integrate the cognitive diagnosis task with LLMs for the first time. This paradigm is called LLM-guided cognitive state transfer (LCST), which reexamines students' cognitive states from a natural language perspective. Llama 3.2,

Gemma, and other advanced LLMs are chosen in this research, as educational experts for analyzing the internal relationships among knowledge concepts and diagnosing students' cognitive states in the target domain due to their accessibility and high performance. Specifically, to obtain students' prior abilities, the abundant response data from the source domain are leveraged. Using mainstream CDMs, this research diagnoses the students' mastery of knowledge concepts in the source domain, which serves as the primary input for the target domain diagnostic feedback. Moreover, this research integrates students' cognitive states from the source domain, formats their prior abilities into natural language descriptions, and designs appropriate prompt templates to enable LLMs to function as intermediaries between the source and target domains, which unlocks students' potential in cognitive diagnosis tasks. To verify LLMs' performance as educational experts in diagnostic tasks, the research formats the diagnostic feedback as proficiency parameters and reintegrates them into the CDM's parameter set for diagnosing prior cognitive states. Extensive experiments are conducted on real-world datasets. The experimental results demonstrate a significant performance improvement achieved by the LCST and reveal the enormous potential of LLMs in cognitive diagnosis tasks, which offers a promising research direction.

Based on the previous research, this research propels further contributions from three aspects. First, this research identifies the real-world ZCCD task and combines cognitive diagnosis tasks with LLMs to propose the novel LCST diagnostic paradigm for the first time, which addresses the cold-start challenge in zero-shot cross-domain diagnosis scenarios. Second, given the outstanding summarization and recommendation performance of LLMs in other domains, the LCST utilizes LLMs as the cognitive diagnosis module for the target domain. Based on prior knowledge, it not only bridges the source and target domains but also enables the functionality of mainstream CDMs. Third, the research conducts extensive experiments on real-world datasets to validate the effectiveness of the proposed LCST paradigm in solving diagnostic tasks for cold-start domains.

2 Related Work

2.1 | Cognitive Diagnosis

Cognitive diagnosis is a fundamental and crucial task in the field of education that aims to infer students' mastery of knowledge concepts (Gao et al., 2024b; Ma et al., 2024a; Ma et al., 2024b; Yang et al., 2024a; Yang et al., 2023b; Yang et al., 2023c). Existing cognitive diagnostic methods are based on available student response logs and focus on modelling interaction behaviours. For instance, disentangling cognitive diagnosis improves diagnostic performance in scenarios with limited exercise labels by utilizing student response records to model students' abilities, exercise difficulties, and label distributions (Chen et al., 2024). The oversmoothing-resistant cognitive diagnosis framework designs a novel perceptual graph convolutional network to capture key response signals in interaction behaviours, which alleviates the over-smoothing problem and enhances the diagnostic capability of the model (Qian et al., 2024). Agent4Edu introduces a personalized learning simulator that leverages LLM-driven generative agents to simulate real learner behaviour (Gao et al., 2025), which aims to bridge the gap between offline evaluation metrics and real-world online performance in personalized learning. LLM-based question generation utilizes LLMs to generate questions aligned with Bloom's taxonomy learning objectives (Elkins et al., 2024), which ensures a more pedagogically grounded approach to question creation. While these works have made significant progress, they are all based on rich student response records. When faced with a blank domain lacking student response logs, these methods struggle to function effectively. As a solution, TechCD uses preconstructed KCGs to transfer students' cognitive states from existing domains to the cold-start target domain (Gao et al., 2023). Zero-1-to-3 generates simulated practice logs for cold-start students by analyzing the behaviour patterns of early-bird students and refining their cognitive states using virtual data to mitigate cold-start challenges (Gao et al., 2024a). However, these approaches rely on the availability of student response records in the target domain, which are difficult to obtain and costly in real-world scenarios, to bridge the source and target domains. Therefore, this research proposes an alternative approach that leverages the advantages of LLMs to build a bridge between source and target domains and extend further to replace the cognitive diagnostic models for the target domain.

2.2 | Large Language Models

Recently, LLMs have demonstrated remarkable capabilities in solving various tasks in the field of NLP (Mesnard et al., 2024; Touvron et al., 2023a; Touvron et al., 2023b). Thanks to self-supervised training on vast datasets and the extensive absorption of knowledge from various domains (Grattafiori et al., 2024), LLMs are capable of solving specific tasks through tailored prompts. As a result, the growing interest in leveraging these advanced models addresses traditional tasks. For instance, the personalized prompt for sequential recommendation effectively enhances cold-start

recommendations by constructing personalized soft prompts based on user profiles (Wu et al., 2024). The spatiotemporal graph transfer learning unifies different tasks into a single template and employs a two-stage prompt pipeline with learnable prompts to achieve domain and task transfer (Hu et al., 2024), enabling the prompts to effectively capture domain knowledge and task-specific attributes. Moreover, in the medical field, the Prompt-Eng designed precise prompts that included both positive and negative aspects and highlighted that designing paired prompts helped the model generalize effectively (Ahmed et al., 2024). Despite the impressive results achieved by LLMs across many domains, no work explored the integration of LLMs with cognitive diagnosis tasks.

3 Problem Statements

The formatting for the ZCCD task is conducted as follows. The available M mature source domains are designated as $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, ..., $\boldsymbol{S}_M$, and the target domain, which requires cold-start initialization, is designated as $\boldsymbol{T}$. The overlapping set of students shared between the m-th source domain and target domain is denoted as $\boldsymbol{U}_m$. In the m-th source domain, the sets of exercises and knowledge concepts are represented as $\boldsymbol{E}_{\boldsymbol{S}_m}$ and $\boldsymbol{K}_{\boldsymbol{S}_m}$, respectively, while in the target domain, they are represented as $\boldsymbol{E}_{\boldsymbol{T}}$ and $\boldsymbol{K}_{\boldsymbol{T}}$. Notably, $\boldsymbol{E}_{\boldsymbol{S}_m} \cap \boldsymbol{E}_{\boldsymbol{T}} = \varnothing$ and $\boldsymbol{K}_{\boldsymbol{S}_m} \cap \boldsymbol{K}_{\boldsymbol{T}} = \varnothing$. The student–exercise response records in the m-th source domain $\boldsymbol{S}_m$ are recorded as $\boldsymbol{R}_{\boldsymbol{S}_m} = \{(u_i, e_j, r_{ij}) \mid r_{ij} \in \{0, 1\}, u_i \in \boldsymbol{U}_m, e_j \in \boldsymbol{E}_{\boldsymbol{S}_m}\}$, where $r_{ij} = 1$ indicates that student u_i answered exercise e_j correctly, otherwise $r_{ij} = 0$. Problem definition can be recorded $\boldsymbol{R}_{\boldsymbol{S}} = \{\boldsymbol{R}_1, \boldsymbol{R}_2, ..., \boldsymbol{R}_M\}$ given the student's response in the source domains and the set of knowledge concepts $\boldsymbol{K}_{\boldsymbol{T}}$ in the target domain. The goal of the research is to diagnose the potential mastery of knowledge concepts in the target domain for students with blank interaction records.

4 Proposed LCST Diagnostic Paradigm

4.1 | Overview

To transfer students' cognitive states from the source domain to the blank target domain, as illustrated in Figure 2, the LCST diagnostic paradigm is proposed, which combines traditional CDMs with LLMs. The proposed LCST comprises four modules, including the pre-established cognitive state, bridge source and target domain, cognitive diagnosis of the target domain, and feedback constraint. In the pre-established cognitive state module, given rich response records in the source domain, the LCST utilizes a traditional CDM to pre-diagnose students' mastery of knowledge concepts in the source domain. This information is then transformed into a natural language form. The zero-shot, few-shot, and chain-of-thought prompts for the bridge source and target domain module are proposed to explore the internal relationships between knowledge concepts. This module takes input from both the source and target domains, including all knowledge concepts and instruction prompts. It is used to activate LLMs' potential to infer knowledge concept relationships. In the cognitive diagnosis of the target domain module, the information from the previous two modules is integrated, and the prompts are designed to utilize LLMs' internal knowledge and pre-acquired prior knowledge to complete the diagnosis task in the target domain. Finally, to ensure that the feedback format is easily assessable, this research incorporates the feedback constraint module into the LCST. Notably, while the CDM in the pre-established cognitive states module requires training, LLMs do not require training or fine-tuning. They only need to be utilized with pretrained models and well-designed prompts for downstream tasks.

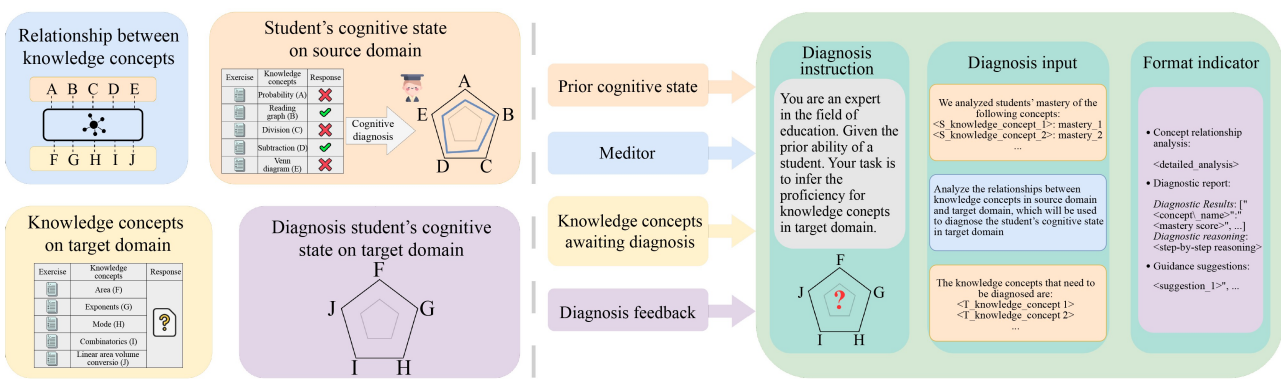


Figure 2 Overview of the LCST. LCST: large language model-guided cognitive state transfer.

4.2 | Pre-Established Cognitive States

The cognitive diagnosis task aims to analyze students' response records and generate diagnostic feedback. Traditional CDMs are only applied in source domains with rich response logs. Therefore, when faced with a zero-shot target domain, the diagnostic process becomes challenging. LLMs have great potential for propagating cognitive states from the source domain to the target domain. However, due to input constraints and the sensitivity of educational contexts, it is not feasible to use all response logs from the source domain as input directly. The focus of this phase is on obtaining the cognitive states of students in the source domain, which lays the foundation for subsequent state transfer. To achieve this goal, this research leverages the available response data from the source domain and combines it with cognitive diagnostic models such as neural cognitive diagnosis (NCD) and knowledge-association based extension of NCD, to pre-diagnose students' cognitive states in the source domain.

Specifically, for student $u_i \in \boldsymbol{U}_m$ in the m-th source domain $\boldsymbol{S}_m$, given their response record set $\boldsymbol{R}^i_{S_m}$ within that domain, the domain-specific CDM $\boldsymbol{M}_m$ will establish their cognitive state, which is expressed as follows:

$$\theta^*_{i_m} = \arg\min_{\theta_{i_m} \in \theta_{i_m}} \sum_{\left(e_j, r_{ij}\right) \in \boldsymbol{R}^i_{S_m}} \mathcal{L}\left(r_{ij}, \boldsymbol{M}_m\left(e_j \mid \theta_{i_m}\right)\right), \quad (1)$$

where the proficiency vector $\theta^*_{i_m} \in \mathbb{R}^{1 \times \|\boldsymbol{K}_{S_m}\|}$ represents the student's mastery of all $\|\boldsymbol{K}_{S_m}\|$ knowledge concepts in the source domain $\boldsymbol{S}_m$, where $\theta^*_{i_m}[n] \in [0, 1]$ indicates the student's proficiency in the n-th knowledge concept, with values ranging from 0 to 1. $\mathcal{L}$ represents the loss function used to pretrain the model $\boldsymbol{M}_m$, typically, the cross-entropy loss.

Next, the parameterized proficiency values are stored in a set, which serves as a precursor for subsequent natural language formatting, as follows:

$$\boldsymbol{C}^i_m = \left\{\left(k_n, \theta^*_{i_m}[n]\right) \mid n \in \{1, 2, \ldots, \|\boldsymbol{K}_{S_m}\|\}\right\}, \quad (2)$$

where $\boldsymbol{C}^i_m$ denotes the set of proficiency values for student u_i in the m-th source domain, and k_n represents the name of the n-th knowledge concept in source domain $\boldsymbol{S}_m$.

4.3 | Bridge Source and Target Domain

To bridge the gap between the source and target domains, previous works have utilized preconstructed knowledge concept maps. However, this approach relies on both source and target domain response logs and incurs high labour costs. Therefore, this research proposes directly leveraging LLMs as educational experts to analyze two types of relationships between knowledge concepts, including hierarchical relationships and similarity relationships. Specifically, a specialized prompt template is designed to investigate the reasoning capabilities of LLMs regarding the internal relationships among knowledge concepts. The template consists of three components, including task description, knowledge concept injection, and format indicator. The task description is used to adapt the concept reasoning task to a natural language-processing task. The knowledge concept injection aims to provide a range of knowledge concepts from both domains to be bridged. The format indicator is employed to constrain the output format, which makes the reasoning results easier to integrate into cognitive diagnostic models based on knowledge concept relationships. This task is defined as concept reasoning, and an example of such a prompt for the smart learning partner (SLP) physics dataset is shown in Figure 3.

To further stimulate the internal knowledge of LLMs, a few-shot chain-of-thought prompt is injected into the prompt template to enhance their reasoning ability (Wei et al., 2022). By adding expert-analyzed knowledge concept relationships and reasoning justifications, the above template is adjusted as shown in Figure 4.

Compared to zero-shot prompting, injecting a small number of prompts guides the LLMs to perform the reasoning task better, thereby reducing the uncertainty of the generated text and increasing confidence in the results.

4.4 | Cognitive Diagnosis of the Target Domain

After obtaining the students' cognitive abilities in the source domain and establishing a bridge between the source and target domains, the proposed LCST is capable of utilizing this prior knowledge to complete the diagnostic process in the target domain. Given the

- Task description: You are an expert in the field of education. Your task is to analyze the relationships between knowledge concepts in the physics domain, specifically focusing on similar relationships and hierarchical relationships.
- Knowledge concept injection: "Application and harm of sound", "three elements of sound", "balanced force", …, "reflection of light."
- Format indicator: JSON object.

Figure 3 Example of a prompt for the smart learning partner.

Step l:

- Task description (+ few-shot chain of thought): You are...

Here is an example:

"Similar relationships": Both concepts are related to the propagation properties of light. Reflection is the phenomenon of light returning when it encounters an interface, while refraction is the change in the propagation path of light when it enters another medium from one medium. Both describe the propagation behaviour of light and although the physical processes that occur are different, they are closely related to the properties of light.

"Hierarchical relationships": Force is the basic concept of the interaction between objects, and balanced force is a specific state or subset of force. When multiple forces act in opposite directions and are equal in magnitude, they form balanced forces. Therefore, force is a prerequisite knowledge concept for mastering balanced force.

Step 2:

- Task description: You are…
- Knowledge concept injection: "Electric energy and electric work", "kinetic energy and potential energy", …, "liquid pressure."

Figure 4 Adjusted prompt for the smart learning partner.

availability of source domain response logs, the diagnostic task in the target domain is divided into two sub-tasks, including single-domain diagnosis using diagnostic feedback from a single domain as prior knowledge and multi-domain diagnosis using diagnostic feedback from multiple domains as prior knowledge.

Similar prompt templates for these two diagnostic tasks are designed to integrate the prior knowledge from the previous modules, which instructs LLMs to analyze and diagnose the students' potential mastery of specific knowledge concepts. Each prompt includes a task description, a textual representation of the student's prior cognitive state, a concept reasoning task, and a format indicator. The task description is used to adapt the diagnostic task to a natural-language-processing task. The format indicator is designed to assess the performance of the LLMs in generating results.

For the single-domain diagnosis task, student's prior cognitive state is injected $\boldsymbol{C}_m$ from the m-th source domain, the set of knowledge concepts in that domain $\boldsymbol{K}$ and the knowledge concepts of the target domain $\boldsymbol{K}_S$, and the knowledge concepts of the target domain $\boldsymbol{K}_T$ into the prompt template, as shown in Figure 5.

For the multi-domain diagnosis task, minor adjustments are made to the above template by replacing the prior cognitive state $\boldsymbol{C}^i = \{\boldsymbol{C}^i_1, \boldsymbol{C}^i_2, ..., \boldsymbol{C}^i_{M'}\}$ where $\boldsymbol{M}'$ represents the number of available source domains and the knowledge concept set for the source domains in the template, as shown in Figure 6.

These two tasks are adjusted based on the availability of interaction logs from the source domain. To enhance the credibility and persuasiveness of the diagnostic tasks, the explanation generation commands are injected into the prompt templates of the two tasks. These explanations are intended to clarify why a particular diagnostic conclusion is made for the student's abilities, and they also contribute to providing personalized learning support for the students. Specifically, LLMs are requested to generate "why" explanation texts and "how" guidance texts for each diagnostic result, thereby ensuring the reasoning and transparency of the diagnostic process.

- Task description: You are an expert in the field of education. Given the prior ability of a student in $\boldsymbol{S}_m$ domain. Your task is to infer the proficiency for knowledge concepts $\boldsymbol{K}_T$ in $\boldsymbol{T}$ domain.
- Prior cognitive state: Students' mastery of the following concepts is analyzed in this research.
  1. Knowledge concept $\boldsymbol{k}_1$: $\boldsymbol{\theta}^*_{i_m}[1]$
  2. Knowledge concept $\boldsymbol{k}_2$: $\boldsymbol{\theta}^*_{i_m}[2]$
  …
- Concept reasoning: Analyze the relationships between knowledge concepts in $\boldsymbol{S}_m$ domain and $\boldsymbol{T}$ domain, which will be used to diagnose the student's cognitive state $\boldsymbol{T}$ domain.
- Format indicator: JSON object.

Figure 5 Example prompt for the single-domain diagnosis.

- Task description: You are an expert in the field of education. Given the prior ability of a student in $\{S_1, S_2, ..., S_M\}$ domains. Your task is to infer the proficiency for knowledge concepts K_T in T domain.
- Prior cognitive state: Students' ability to master knowledge concepts in multiple domains. In S_1 domain: … In S_2 domain: …
- Concept reasoning: Analyze the relationships between knowledge concepts in $\{S_1, S_2, ..., S_M\}$ domains and T domain, which will be used to diagnose the student's cognitive state in T domain.
- Format indicator: JSON object.

Figure 6 Adjusted prompt for the single-domain diagnosis.

4.5 | Feedback Constraints

To facilitate the evaluation and extraction of the results, format indicators are designed for each of the three tasks to constrain the output format. For the concept reasoning task, the output needs to be constrained into two parts, including pairs of knowledge concepts that exhibit similarity relationships and pairs that exhibit hierarchical relationships. The instructions were added at the end of the concept reasoning task template, as shown in Figure 7.

For the single-domain diagnosis and multi-domain diagnosis tasks, to enhance the transparency and interpretability of the diagnostic process, the format indicator in this section constrains the output into three parts, including the concept relationship analysis, the diagnostic report, and the guidance suggestion. Instructions are added to the end of the single-domain diagnosis and multi-domain diagnosis task templates, respectively, as shown in Figure 8.

Finally, the mastery score is extracted for each knowledge concept from the output, which is guided by carefully designed prompt templates. These scores are then vectorized to form a student ability vector representing the mastery levels of different knowledge concepts in the target domain. The length of this vector remains consistent with the number of knowledge concepts in the target domain.

5 Experiments

Because the key contribution of this work is to diagnose the potential mastery levels of knowledge concepts in the target domain in which interaction records are blank, comprehensive experiments on real-world datasets are conducted to address the following four research questions:

(1) Can the proposed LCST effectively handle the ZCCD task?

(2) Can the LCST effectively utilize information from diverse source domains for the ZCCD task?

(3) How does the LCST perform on the concept reasoning task?

(4) How do different LLMs perform in cross-domain diagnosis?

5.1 | Dataset Description

The experiments are conducted on five real-world datasets, including mathematics, physics, chemistry, history, and geography, which are collected from the

- Format indicator: Your response should be structured into two distinct parts, including similarity relationships and hierarchical relationships.
1. Similarity relationships: [{"*concept_1*": "< *concept_name* >", "*concept_2*": "<*concept_name* >", "*reasoning*": "< *explanation* >"}, …]
2. Hierarchical relationships: [{"*parent_concept*": "< *concept_name* >", "*child_concept*": "<*concept_name* >", "*reasoning*": "< *explanation* >"}, …]

Figure 7 Added concept reasoning task template.

- Format indicator: Your response should be structured into three parts, including concept relationship analysis, diagnostic report, and guidance suggestion.
1. Concept relationship analysis: "< *detailed_analysis* >."
2. Diagnostic report: The inferred cognitive state of the student in the T domain.
   (1) Diagnostic results: ["< *concept_name* >": "< *mastery score* >", …]
   (2) Diagnostic reasoning: "< *step-by-step reasoning* >"
3. Guidance suggestion: "< *suggestion_1* >", …

Figure 8 Added instructions in the single-domain diagnosis and multi-domain diagnosis task templates.

SLP online education platform (Lu et al., 2021). These datasets summarize students' academic performance data across different subjects over three-year-study, including their response records, knowledge concept texts, and the relationships between exercises and knowledge concepts. Each subject is treated as a domain, alternating between being the cold-start target domain and other domains that serve as source domains. For each dataset, each student's response records are divided into training and testing sets, with a 4:1 split. The test set for the target domain is used not only to evaluate the pretraining performance of the cognitive diagnostic model but also to test the performance of baselines in solving the ZCCD task. The statistics of the processed datasets are shown in Table 1.

Table 1 Statistics of the datasets

Subject	Statistics			
	Student	Exercise	Concept	Log
Mathematics	224	847	32	14,227
Physics	224	1,536	48	11,049
History	114	636	19	6,458
Geography	114	371	19	3,651
Chemistry	31	315	15	2,164

5.2 | Experimental Setup

5.2.1 Baselines

The Llama 3.21 is the backbone of the LCST. The CDM is first pretrained using the dataset from the target domain, obtaining the model's network parameters specific to that domain. The student ability component in these network parameters is presented with the vectorized student abilities, after which the cross-domain diagnostic performance is tested on the target domain's test set. To demonstrate the effectiveness of the proposed LCST in solving the ZCCD task, it is applied to four widely used cognitive diagnostic models, including NCD (Wang et al., 2020), knowledge-sensed cognitive diagnosis (Ma et al., 2022), relation map-driven cognitive diagnosis (RCD) (Gao et al., 2021), and knowledge-association based extension of the NCD (Wang et al., 2022). The random oracle model as a baseline for comparison is selected, where the random oracle method represents the lower and upper bounds of the performance of CDMs in solving the ZCCD task. The random method means that predicts the initial ability of students in the target domain from a uniform (0, 1) distribution randomly, which is the most common method in existing CDMs. The oracle method uses CDM to train the target domain ability from the response records of students in the target domain directly.

5.2.2 Evaluation Metrics

Given that the true knowledge mastery of students is unknown, mainstream methods have been used in previous research to evaluate the effectiveness of CDMs by predicting students' exercise performance based on acquired knowledge mastery vectors. Since cognitive diagnosis is a binary classification task that predicts whether a student will answer a given question correctly, three well-known metrics are selected to assess prediction performance. First, accuracy serves as an intuitive metric for evaluating overall prediction correctness. This research sets a threshold of 0.5, meaning that if the predicted probability exceeds 0.5, the student is expected to answer correctly. Otherwise, an incorrect response is predicted:

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\hat{y}_i^b = y_i\right), \tag{3}$$

where ACC indicates accuracy, $\hat{y}_i^b$ is the predicted binary response, y_i is the ground truth response, N denotes the size of the dataset, and $\mathbb{1}(\cdot)$ is the indicator function. However, in scenarios with imbalanced class distributions, accuracy alone may not provide a comprehensive assessment of model performance. Second, this research incorporates the area under curve (AUC) as an additional evaluation metric (Bradley, 1997). AUC offers a holistic measure of the model's discriminative ability across different probability thresholds. Third, root mean square error (RMSE) is introduced as an evaluation metric to quantify the deviation between predicted probabilities and actual binary labels (Pei et al., 2017):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i - y_i\right)^2}, \tag{4}$$

where $\hat{y}_i$ represents the predicted probability of a correct response and N denotes the size of the dataset.

5.2.3 Parameter Settings

In the pretraining phase, this research initializes all parameters in the network using Xavier (Glorot & Bengio, 2010), and uses the Adam optimizer during training (Kingma & Ba, 2014). The batch size is fixed at 32, while the learning rate is fixed at 0.01. Then, the dimensions of the latent features for the students and exercises are set to be equal to the number of knowledge concepts. Finally, this research sets the temperature for all LLMs to 0.2 and runs each task three times, taking the average as the diagnostic feedback. All models are implemented in PyTorch, while all experiments are

conducted on a Linux server equipped with a Tesla V100 GPU.

5.3 | Overall Performance

To validate the effectiveness of the proposed diagnostic paradigm in solving the ZCCD task, the LCST is compared with other baselines by using single-domain and multi-domain information as prior knowledge.

5.3.1 Performance on the Single-Domain Diagnosis Task

First, this research explores the possibility of using a single domain as the source domain to solve the ZCCD task, which includes four cross-domain scenarios. This research alternates each dataset to serve as the target domain, with other datasets acting as the source domains. The experimental results shown in Table 2 demonstrate that the proposed LCST outperforms the widely applied random method in most scenarios and even surpasses oracle method in some cases, achieving the performance upper bound. This illustrates the immense potential of LLMs, powered by their strong reasoning capabilities, in cognitive diagnosis tasks. Furthermore, four aspects of performance are shown. First, in most scenarios, the performance of the LCST is consistently close to the oracle, especially in domains such as mathematics and physics, where there is a clear hierarchical relationship between knowledge concepts. This confirms that by implementing the concept reasoning task, LLMs can enhance their diagnostic capability as educational experts effectively. Second, given that different CDMs model students and exercises, the parameterized diagnostic feedback shows performance fluctuations across various CDMs in the same scenario. Among them, NeuralCD exhibits the least fluctuation, since it does not interfere much with the student's cognitive state parameters. In contrast, CDMs that employ more complex modelling approaches somewhat distort the parameterized diagnostic feedback but still maintain high performance. Third, at optimal performance, the proposed LCST consistently outperforms TechCD. TechCD addresses the ZCCD challenge by pre-constructing KCGs, which rely on both information-rich interaction datasets and extensive human annotations. In contrast, the LCST leverages the strengths of LLMs, which requires only a set of knowledge concepts within the domain to infer the relationships between them. By eliminating the need for complex model designs, such as those in TechCD, which may reduce interpretability, the LCST offers a more efficient and cost-effective solution while maintaining clarity and ease of implementation. Fourth, compared to mainstream CDMs pretrained on the target domain, the LCST paradigm, which integrates LLMs, is able to better capture students' mastery of complex knowledge concepts by relying solely on prediagnosed cognitive states without directly receiving students' response records. This highlights the promising research direction for using LLMs in cognitive diagnosis tasks.

Table 2 Performance on the single-domain diagnosis task in four scenarios

LLMs	Method	Source: mathematics, Target: physics			Source: physics, Target: mathematics			Source: geography, Target: history			Source: history, Target: geography		
		ACC	RMSE	AUC	ACC	RMSE	AUC	ACC	RMSE	AUC	ACC	RMSE	AUC
TechCD	None	0.6734	0.4691	0.7137	0.6508	0.4861	0.6504	0.6349	0.48300	0.6860	0.6255	0.4626	0.6768
NeuralCD	Random	0.6021	0.5070	0.6232	0.5747	0.5674	0.6173	0.5917	0.5377	0.5541	0.4667	0.6049	0.5306
	LCST	**0.6902**	**0.4464**	**0.7204**	**0.6568**	**0.4882**	**0.6833**	**0.6572**	**0.4920**	**0.6340**	**0.6536**	**0.4857**	**0.7296**
	Oracle*	0.6979	0.4379	0.7510	0.6593	0.4855	0.6931	0.7249	0.4424	0.6531	0.6546	0.4858	0.7410
KSCD	Random	0.6721	0.4577	0.7196	0.5982	0.4927	0.6246	**0.7369**	0.4373	0.6667	0.6689	0.4636	0.7328
	LCST	**0.7230**	**0.4345**	**0.7459**	**0.6687**	**0.4590**	**0.7186**	0.7314	**0.4291**	**0.6955**	**0.6973**	**0.4609**	**0.7441**
	Oracle*	0.7192	0.4333	0.7495	0.6626	0.4600	0.7161	0.7489	0.4270	0.6885	0.6787	0.4622	0.7395
R	Random	0.6289	0.4718	0.6637	**0.6268**	**0.4733**	**0.7228**	0.7358	**0.4337**	0.6832	0.5115	0.5132	0.5225
	LCST	**0.6667**	**0.4629**	**0.6752**	0.6098	0.4791	0.7181	**0.7391**	0.4350	**0.6950**	**0.5869**	**0.4934**	**0.6035**
	Oracle*	0.7011	0.4454	0.7171	0.6702	0.4584	0.7304	0.7402	0.4316	0.6984	0.6470	0.4742	0.6847
KaNCD	Random	0.4187	0.6479	0.5985	0.5935	0.5049	0.6267	0.7052	0.4533	0.6424	0.6022	0.5250	0.6564
	LCST	**0.6979**	**0.4442**	**0.7215**	**0.6170**	**0.4896**	**0.6599**	**0.7456**	**0.4255**	**0.7212**	**0.6219**	**0.5130**	**0.6740**
	Oracle*	0.7022	0.4398	0.7403	0.6687	0.4620	0.7194	0.7544	0.4320	0.6658	0.6601	0.4855	0.7089

Notes. CD: cognitive diagnosis, KSCD: knowledge-sensed cognitive diagnosis, RCD: relation map-driven cognitive diagnosis, KaNCD: knowledge-association-based extension of neural cognitive diagnosis, LCST: large language model-guided cognitive state transfer, ACC: accuracy, RMSE: root mean square error, AUC: area under curve. The best results are highlighted in bold, while * denotes the upper bounds.

5.3.2 Performance on the Multi-Domain Diagnosis Task

This research further explores the potential of integrating multisource domains as prior information to solve the ZCCD task. This includes three scenarios in which physics, mathematics, and chemistry are used as target domains, while the remaining subjects provided prior information as source domains. From Table 3, three results can be extracted. First, information from multiple domains provides significantly richer prior knowledge than single-domain information. By offering higher-quality input, LLMs can provide more accurate and reliable student ability predictions. Moreover, having more domains means richer relationships between knowledge concepts, which also helps LLMs deliver more comprehensive diagnostics for students. Second, the knowledge concepts in the three domains exhibited stronger hierarchical and similarity relationships. As a result, the LCST demonstrated performance that not only far exceeded the random method in most scenarios but also slightly surpassed the oracle method. This suggests that domains with closer relationships can more effectively address the cold-start challenge in the blank target domain. Third, LLMs show excellent generalization across different scenarios, indicating that they can effectively adapt to various subjects and knowledge points. Compared to traditional CDM diagnostic paradigms, LLMs are capable of processing more diverse input features, further enhancing their performance in multitasking environments.

5.4 | Performance on the Concept Reasoning Task

To explore the performance of the LCST in handling the concept reasoning task, RCD is selected, which involves knowledge concept relationship modelling, as the validation model and tests are conducted in the physics domain. The standard RCD method uses an approach to obtain knowledge concept relationships from student response logs that is referred to as statistics-based. This research utilizes the LCST-enhanced method to model the relationships among knowledge concepts. The experimental results shown in Figure 9 demonstrate that the LCST-enhanced method outperforms the statistics-based method in RCD, surpassing it on both metrics. This proves that the knowledge concept relationships inferred through the LCST are more beneficial for RCD in modelling knowledge concepts and align better with real-world scenarios.

Furthermore, as shown in Figure 10, four reasoning results of both methods regarding knowledge concept relationships can be extracted. First, the LCST-enhanced method infers a broader range of knowledge concept relationships, while the statistics-based method relies on the statistical analysis of existing response logs, leading to the loss of many knowledge concept relationships, which hinders CDM from performing fine-grained modelling. Second, the statistical results of the statistics-based method regarding the similarity relationships between knowledge concepts are biased, particularly in the upper half of the knowledge concepts shown in Figure 10(a). This is because the statistical results are constrained by the distribution of student responses and abilities in the corresponding logs. In contrast, as shown in Figure 10(b), the LCST-enhanced method infers knowledge concept relationships with a more uniform distribution and provides more comprehensive reasoning of hierarchical relationships, proving that the LCST has great potential for handling the concept reasoning task.

Table 3 Performance on the multi-domain diagnosis task in three scenarios

LLMs	Method	Source: mathematics and chemistry Target: physics			Source: physics and chemistry Target: mathematics			Source: physics and mathematics Target: chemistry		
		ACC	RMSE	AUC	ACC	RMSE	AUC	ACC	RMSE	AUC
NeuralCD	Random	0.5285	0.6146	0.4549	0.5500	0.5765	0.5486	0.5893	0.5401	0.5986
	LCST	**0.7276**	**0.4723**	**0.6561**	**0.6842**	**0.4681**	**0.7411**	**0.6786**	**0.4458**	**0.7285**
	Oracle*	0.7195	0.4612	0.6921	0.6895	0.4722	0.7293	0.6768	0.4549	0.7116
KSCD	Random	0.7073	0.4644	0.6467	0.7026	0.4429	0.7579	0.6893	0.4454	0.7093
	LCST	**0.7358**	**0.4465**	**0.7017**	**0.7158**	**0.4340**	**0.7746**	**0.6964**	**0.4443**	**0.7207**
	Oracle*	0.7236	0.4481	0.7006	0.7026	0.4379	0.7660	0.7107	0.4381	0.7456
KaNCD	Random	0.7358	0.4549	0.6809	0.6868	0.4644	0.7465	0.6857	0.4629	0.7189
	LCST	**0.7439**	**0.4513**	**0.6871**	**0.7026**	**0.4519**	**0.7532**	**0.6982**	**0.4588**	**0.7254**
	Oracle*	0.7520	0.4426	0.6801	0.7132	0.4423	0.7516	0.6946	0.4483	0.77408

Notes. CD: cognitive diagnosis, KSCD: knowledge-sensed cognitive diagnosis, RCD: relation map-driven cognitive diagnosis, KaNCD: knowledge-association based extension of neural cognitive diagnosis, LCST: large language model-guided cognitive state transfer, ACC: accuracy, RMSE: root mean square error, AUC: area under curve. The best results are highlighted in bold, while * denotes the upper bounds.

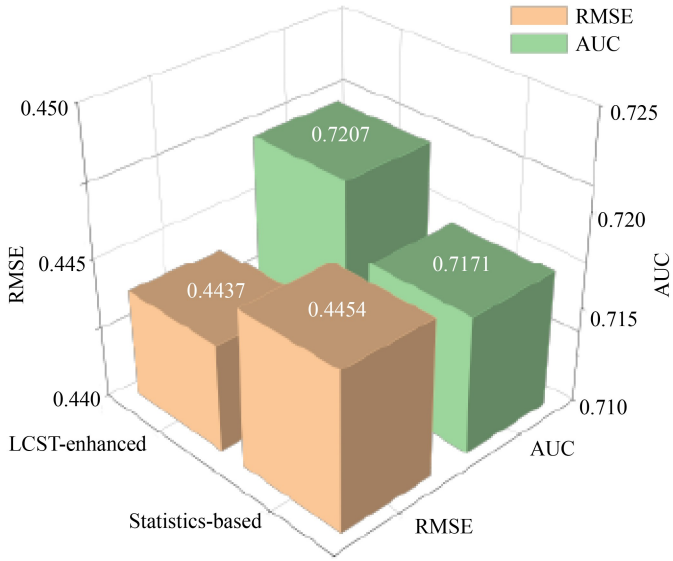
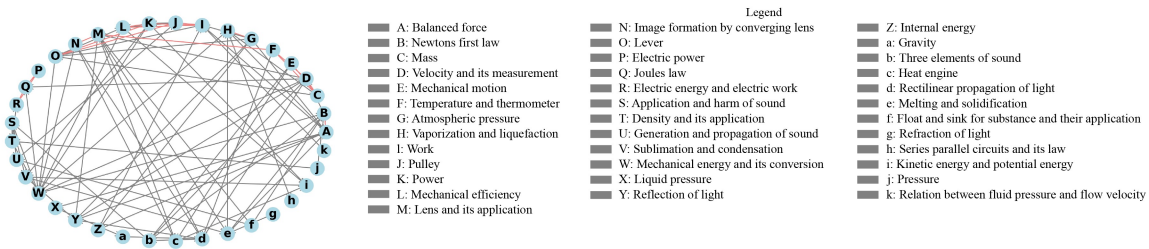


Figure 9 Visualization results of concept reasoning task in the physics domain with statistics-based and LCST-enhanced methods. LCST: large language model-guided cognitive state transfer, ACC: accuracy, RMSE: root mean square error, AUC: area under curve.

(a) Statistics-based knowledge concept relationship



(b) LCST-enhanced knowledge concept relationship

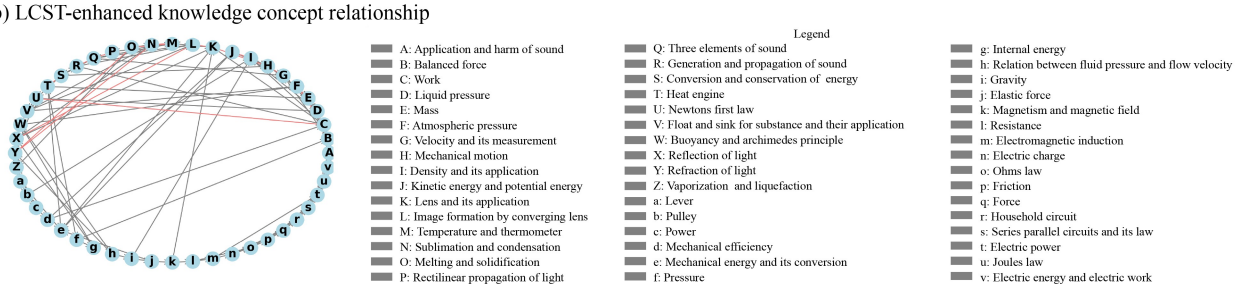


Figure 10 Performance comparison of statistics-based and LCST-enhanced methods on the concept reasoning task. (a) Statistics-based knowledge concept relationship, (b) LCST-enhanced knowledge concept relationship. LCST: large language model-guided cognitive state transfer.

5.5 | Performance of Different LLMs on Diagnostic Tasks

To explore the performance of different LLMs in cognitive diagnosis tasks, several mainstream open-source LLMs with varying parameter sizes are selected, including Llama3.2:3b, Gemma:7b, Qwen2.5:3b, Hermes3:3b, and Qwen2.5:7b, as the backbone of the LCST. Mathematics as the source domain and physics as the target domain are set, using NeuralCD as the pretrainer to diagnose students' abilities in the source domain. Two results can be extracted from Table 4. First, Gemma-7b exhibits the best diagnostic performance, albeit with the longest generation time. In contrast, Hermes3:3b shows worse diagnostic performance than the random method, which may be due to Amnesia Mode leading to unstable generated feedback (Teknium et al., 2024). Second, a larger parameter size does not necessarily indicate better performance for diagnostic tasks. Smaller LLMs may perform better than larger ones in diagnostic tasks. For example, Llama3.2:3b, despite its smaller parameter

size, shows near-optimal performance and takes the least amount of time. In terms of balancing performance and efficiency, Llama3.2:3b may be the most suitable backbone for the LCST to address the ZCCD task. This is attributed to the training set and quality of the LLMs, since Qwen2.5-based LLMs, primarily pretrained on Chinese corpora, may not perform optimally under English prompt templates.

Table 4 Performance comparison of LLMs on the ZCCD task

LLMs	ACC	RMSE	AUC	Time (s)
Random	0.6092	0.5225	0.6472	None
Llama3.2:3b	0.7072	0.4505	0.7314	**6.0100**
Gemma-7b	**0.7105**	**0.4451**	**0.7363**	9.9900
Qwen2.5:3b	0.6831	0.4601	0.7120	7.6700
Hermes3:3b	0.5758	0.5410	0.6317	6.0500
Qwen2.5:7b	0.6990	0.4531	0.7249	8.7400
Oracle* (NeuralCD)	0.7050	0.4453	0.7397	None

Notes. LLMs: large language models, ACC: accuracy, RMSE: root mean square error, AUC: area under curve. The best results are highlighted in bold and the runner-up is underlined. The asterisk * here indicates the Oracle result.

6 Discussion

In this work, the LCST method is proposed to address data scarcity in the target domain encountered in ZCCD. The experimental results indicate that the LCST not only significantly enhances diagnostic accuracy across multiple datasets but also, in certain cases, achieves performance on par with oracle methods that rely on target domain data, thereby validating the effectiveness and applicability of LLMs in cross-domain cognitive diagnosis. Nonetheless, the proposed method has several limitations, including high sensitivity to prompt template designs, inadequate coverage of domain-specific knowledge, and a lack of clarity in the model's internal reasoning processes. Future research will focus on exploring more robust prompt designs, adaptive fine-tuning techniques, and the integration of external knowledge graphs to further improve the model's ability to capture cross-domain knowledge and enhance diagnostic performance.

7 Conclusions

In this research, the application of LLMs in cognitive diagnosis tasks is explored to diagnose students' potential mastery of knowledge concepts in a zero-shot cross-domain scenario. A novel diagnostic paradigm, LCST, not only establishes a bridge between the source and target domains but also facilitates the diagnostic task in the target domain. The experimental results on real-world datasets demonstrate the effectiveness of the proposed method. This research is intended to further investigate the application of LLMs tailored to the characteristics of cognitive diagnosis tasks.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Grant Nos. 62107001, U21A20512, 62302010, and 62202006), the Chinese Postdoctoral Science Foundation (Grant No. 2023M740015), the Postdoctoral Fellowship Program (Grade B) of the China Postdoctoral Science Foundation (Grant No. GZB20240002), and the Anhui Province Key Laboratory of Intelligent Computing and Applications, China (Grant No. AFZNJS2024KF01).

Conflict of Interest The authors declare that they have no conflict of interest.

Ethics Statements The authors declare that their Institutional Ethics Committee confirmed that no ethical review was required for this study. Written informed consent for participation was not required because all participants' data was anonymized before the statistical analyses were conducted.

Data Availability Statements The authors confirm that all data generated or analyzed during this study are included in this published article.

Authors Contributions Haiping Ma was responsible for the scheme design, technical guidance, and the writing and revision of the thesis. Changqian Wang was responsible for the scheme improvement, experimental design, and implementation, as well as the writing of the thesis. Siyu Song assisted in the implementation of some experiments. Shangshang Yang was involved in the writing and revision of the thesis. Limiao Zhang was involved in the writing and revision of the thesis. Xingyi Zhang provided guidance on the research route of the thesis. All authors approved the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

Agostinelli, V., Wild, M., Raffel, M., Fuad, K. A. A., & Chen, L. (2023). Simul-LLM: A framework for exploring high-quality simultaneous translation with large language models. *arXiv Preprint*, arXiv:2312.04691.

Ahmed, A., Hou, M., Xi, R., Zeng, X., & Shah, S. A. (2024). Prompt-Eng: Health-care prompt engineering: Revolutionizing healthcare applications with precision prompts. In: *Proceedings of the ACM on Web Conference 2024*. New York: ACM, 1329–1337.

Bai, Y., Li, X., Liu, Z., Huang, Y., Guo, T., Hou, M., Xia, F., & Luo, W. (2025). csKT: Addressing cold-start problem in

knowledge tracing via kernel bias and cone attention. *Expert Systems with Applications*, *266*(25), 125988.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., & et al. (2020). Language models are few-shot learners. *arXiv Preprint*, arXiv:2005.14165.

Chen, X., Wu, L., Liu, F., Chen, L., Zhang, K., Hong, R., & Wang, M. (2024). Disentangling cognitive diagnosis with limited exercise labels. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. New York: ACM, 18028–18045.

Elkins, S., Kochmar, E., Cheung, J. C., & Serban, I. (2024). How teachers can use large language models and Bloom's taxonomy to create educational quizzes. *arXiv Preprint*, arXiv:2401.05914.

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*. New York: ACM, 501–510.

Gao, W., Liu, Q., Wang, H., Yue, L., Bi, H., Gu, Y., Yao, F., Zhang, Z., Li, X., & He, Y. (2024a). Zero-1-to-3: Domain-level zero-shot cognitive diagnosis via one batch of early-bird students towards three diagnostic objectives. *arXiv Preprint*, arXiv:2312.13434.

Gao, W., Liu, Q., Yue, L., Yao, F., Lv, R., Zhang, Z., Wang, H., & Huang, Z. (2025). Agent4Edu: Generating learner response data by generative agents for intelligent 20 education systems. *arXiv Preprint*, arXiv:2501.10332.

Gao, W., Liu, Q., Yue, L., Yao, F., Wang, H., Gu, Y., & Zhang, Z. (2024b). Collaborative cognitive diagnosis with disentangled representation learning for learner modeling. *arXiv Preprint*, arXiv:2411.02066.

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 coldstart cognitive diagnosis. In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 983–992.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*, *9*, 249–256.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., & et al. (2024). The Llama 3 herd of models. *arXiv Preprint*, arXiv:2407.21783.

Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., & Zhao, W. X. (2024). Large language models are zero-shot rankers for recommender systems. *arXiv Preprint*, arXiv:2305.08845.

Hu, J., Liu, X., Fan, Z., Yin, Y., Xiang, S., Ramasamy, S., & Zimmermann, R. (2024). Prompt-based spatio-temporal graph transfer learning. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. New York: ACM, 890–899.

Jiang, M., Bao, K., Zhang, J., Wang, W., Yang, Z., Feng, F., & He, X. (2024). Item-side fairness of large language model-based recommendation system. *arXiv Preprint*, arXiv:2402.15215.

Kingma, D. &, 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. *arXiv Preprint*, arXiv:2205.11916.

Li, Z., Chen, Z. Z., Ross, M., Huber, P., Moon, S., Lin, Z., Dong, X. L., Sagar, A., Yan, X., & Crook, P. A. (2024). Large language models as zero-shot dialogue state tracker through function calling. *arXiv Preprint*, arXiv:2402.10466.

Liu, S., Yu, X., Ma, H., Wang, Z., Qin, C., & Zhang, X. (2023). Homogeneous cohort-aware group cognitive diagnosis: A multi-grained modeling perspective. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. New York: ACM, 4094–4098.

Lu, Y., Pian, Y., Shen, Z., Chen, P., & Li, X. (2021). SLP: A multi-dimensional and consecutive dataset from K-12 education. In: *Proceedings of the 29th International Conference on Computers in Education*. 261–266.

Ma, H., Li, M., Wu, L., Zhang, H., Cao, Y., Zhang, X., & Zhao, X. (2022). Knowledge-sensed cognitive diagnosis for intelligent education platforms. In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. New York: ACM, 1451–1460.

Ma, H., Wang, C., Zhu, H., Yang, S., Zhang, X., & Zhang, X. (2024a). Enhancing cognitive diagnosis using un-interacted exercises: A collaboration-aware mixed sampling approach. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(8), 8877–8885.

Ma, H., Xia, A., Wang, C., Wang, H., & Zhang, X. (2024b). Diffusion-inspired cold start with sufficient prior in computerized adaptive testing. *arXiv Preprint*, arXiv:2411.12182.

Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., & et al. (2024). Gemma: Open models based on Gemini research and technology. *arXiv Preprint*, arXiv:2403.08295.

Open AI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & et al. (2023). GPT-4 technical report. *arXiv Preprint*, arXiv:2303.08774.

Pei, H., Yang, B., Liu, J., & Dong, L. (2017). Group sparse bayesian learning for active surveillance on epidemic dynamics. In: *Proceedings of the 32nd AAAI Conference on AI and 13th Innovative Applications of AI Conference and 8th AAAI Symposium on Educational Advances in AI*. New York: ACM, 800–807.

Qian, H., Liu, S., Li, M., Li, B., Liu, Z., & Zhou, A. (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*. New York: ACM, 2455–2466.

Ren, X., Wei, W., Xia, L., Su, L., Cheng, S., Wang, J., Yin, D., & Huang, C. (2024). Representation learning with large language models for recommendation. In: *Proceedings of the ACM on Web Conference 2024*. New York: ACM, 3464–3475.

Teknium, R., Quesnelle, J., & Guang, C. (2024). Hermes 3 technical report. *arXiv Preprint*, arXiv:2408.11857.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & et al. (2023a). Llama: Open and efficient foundation language models. *arXiv Preprint*, arXiv:2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., & et al. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv Preprint*, arXiv:2307.09288.

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*, *34*(4), 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*, *35*(8), 8312–8327.

Wang, W., Fang, T., Li, C., Shi, H., Ding, W., Xu, B., Wang, Z., Bai, J., Liu, X., Cheng, J., & et al. (2024a). CANDLE: Iterative conceptualization and instantiation distillation from large language models for commonsense reasoning. *arXiv Preprint*, arXiv:2401.07286.

Wang, W., Jiao, W., Huang, J., Dai, R., Huang, J.-t., Tu, Z., & Lyu, M. R. (2024b). Not all countries celebrate Thanksgiving: On the cultural dominance in large language models. In: *Proceedings of the 62ndAnnual Meeting of the Association for Computation Linguistics*. Bangkok: ACL, 6349–6384.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *arXiv Preprint*, arXiv:2201.11903.

Wu, Y., Xie, R., Zhu, Y., Zhuang, F., Zhang, X., Lin, L., & He, Q. (2024). Personalized prompt for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3376–3389.

Yang, F., Chen, Z., Jiang, Z., Cho, E., Huang, X., & Lu, Y. (2023a). PALR: Personalization aware LLMs for recommendation. *arXiv Preprint*, arXiv:2305.07622.

Yang, S., Chen, M., Wang, Z., Yu, X., Zhang, P., Ma, H., & Zhang, X. (2024a). DisenGCD: A meta multigraph-assisted disentangled graph learning framework for cognitive diagnosis. *arXiv Preprint*, arXiv:2410.17564.

Yang, S., Qin, L., & Yu, X. (2024b). Endowing interpretability for neural cognitive diagnosis by efficient Kolmogorov–Arnold networks. *arXiv Preprint*, arXiv:2405.14399.

Yang, S., Wei, H., Ma, H., Tian, Y., Zhang, X., Cao, Y., & Jin, Y. (2023b). Cognitive diagnosis-based personalized exercise group assembly via a multi-objective evolutionary algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *7*(3), 829–844.

Yang, S., Yu, X., Tian, Y., Yan, X., Ma, H., & Zhang, X. (2024c). Evolutionary neural architecture search for transformer in knowledge tracing. *arXiv Preprint*, arXiv:2310.01180.

Yang, S., Zhen, C., Tian, Y., Ma, H., Liu, Y., Zhang, P., & Zhang, X. (2023c). Evolutionary multi-objective neural architecture search for generalized cognitive diagnosis models. In: *Proceedings of 2023 the 5th International Conference on Data-Driven Optimization of Complex Systems*. New York: IEEE, 1–10.

Yu, X., Qin, C., Shen, D., Ma, H., Zhang, L., Zhang, X., Zhu, H., & Xiong, H. (2024a). RDGT: Enhancing group cognitive diagnosis with relation-guided dual-side graph transformer. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3429–3442.

Yu, X., Qin, C., Zhang, Q., Zhu, C., Ma, H., Zhang, X., & Zhu, H. (2024b). DISCO: A hierarchical disentangled cognitive diagnosis framework for interpretable job recommendation. *arXiv Preprint*, arXiv:2410.07671.

Zhang, D., Zhang, K., Wu, L., Tian, M., Hong, R., & Wang, M. (2024). Path-specific causal reasoning for fairness-aware cognitive diagnosis. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. New York: ACM, 4143–4154.