Front. Comput. Sci., 2025, 19(1): 191318
https://doi.org/10.1007/s11704-024-40029-1

LETTER

Foundation model enhanced derivative-free cognitive diagnosis

Mingjia LI, Hong QIAN (✉), Jinglan LV, Mengliang HE, Wei ZHANG, Aimin ZHOU

Shanghai Institute of AI for Education and School of Computer Science and Technology,
East China Normal University, Shanghai 200062, China

1 Introduction

Cognitive diagnosis, aiming at inferring students' underlying cognitive states according to their response logs, is a fundamental approach to understanding and improving individual learning in intelligent education systems [1]. In recent decades, many cognitive diagnosis models have been proposed. Item response theory (IRT) [2] proposes to utilize mathematical models to model the cognitive process of students responding to questions, where the students' proficiency (cognitive state) is simply represented as a scalar and the interaction between students and questions is modeled by a logistic function. Furthermore, deterministic input, noisy and gate model (DINA) [3] proposes to apply the Q-matrix to capture the relationship between questions and knowledge attributes. This finer-grained modeling enables cognitive diagnosis to place the scale of describing cognitive states at the level of knowledge attributes, and thus has become the paradigm of modern cognitive diagnosis. Subsequent studies are mostly based on this Q-matrix grounded paradigm [4–7]. The recent emergence of deep learning-based cognitive diagnosis, for instance, neural cognitive diagnosis models (NCDM) [4] and extending NeuralCD with the knowledge associations consideration (KANCD) [5], aims to utilize the powerful fitting capability of neural networks to capture the complex interactive relations between students and questions.

However, current cognitive diagnosis models still have limitations due to oversimplified problem formulations, such as treating responses as merely right or wrong and using Q-matrix. This leads to a loss of critical information and subsequent defects. (1) The diagnostic granularity is coarse, which diminishes the practical significance of the diagnostic results. Most of the existing diagnostic approaches describe the cognitive states of students at the level of knowledge attributes, i.e., the output of these methods is a score of the mastery level of students on each attribute. However, what exactly do these scores mean? What is the model's confidence in these scores? How can we help students improve their mastery level of knowledge points with low scores? The answers to these questions are all unclear. (2) Lack of individualization. For example, due to the binary treatment of responses, students with the same right or wrong answers may be diagnosed as having the same cognitive states, even though their mistakes may be different. (3) In need of a large amount of data. Simplification often leads to loss of details, such as the loss of question difficulties when reducing questions to a Q-matrix, which can only be indirectly estimated from students' correct rates. This is why existing methods require a large number of training samples. In the era of foundational models, these limitations could be mitigated through large language models (LLMs) enhanced cognitive diagnosis.

To address these limitations, this paper proposes a fine-grained, individualized and sample-efficient cognitive diagnosis (FineCD) method. FineCD stands out by utilizing LLMs (specifically, we utilize OpenAI GPT-4) to deal with side information such as question statements. FineCD enables us to analyze cognitive states at a more detailed level, going beyond knowledge attributes to identifying specific mistakes. This can help teachers understand students' cognitive states more specifically. Besides, individualization and sample efficiency are also improved since more details are taken into consideration. To the best of our knowledge, FineCD is the first attempt to apply LLMs to enhance cognitive diagnosis. To study the performance of FineCD, we conduct experiments on real-world datasets with online education background, where a granular, individualized and small sample diagnosis is highly in demand. The results show that, by leveraging the power of LLMs to incorporate with rich question-specific side information, FineCD can provide a more fine-grained and customized understanding of learners' cognitive profiles via a more sample efficient manner.

2 The proposed FineCD

This paper introduces FineCD, a method that utilizes LLMs to aid in cognitive diagnosis by incorporating side information. The process consists of three steps: (1) analyzing potential reasons for a student's responses, (2) identifying the most likely combination of potential reasons, and (3) summarizing the student's cognitive state using natural language. Throughout these steps, LLMs play distinct essential roles.

Analysis of potential reasons In the first step, LLMs

Received January 7, 2024; accepted June 5, 2024

E-mail: hqian@cs.ecnu.edu.cn

Special Issue—Excellent Young Computer Scientists Vision on Foundation Models

analyze students' responses to identify potential reasons for the answers. For correct answers, LLMs determine the skills involved, while for incorrect answers, they identify potential reasons for mistakes, considering factors such as lack of proficiency in certain skills. Here is an example of how exactly LLMs are utilized to discover potential reasons. Prompt: You are an elementary school math teacher. Given the response logs of a student, please list 1 to 3 technical mistakes the student may make leading to the incorrect answer using brief description (e.g., the student confuses triangle with circle). Question: 9^2 =?; A. 11; B. 18; C. 81; D. 92. Student's answer: A. GPT-4's response: (1) Misunderstanding of exponent notation, i.e., the student may interpret the exponent 9^2 as simply meaning adding 2 to the base number 9, leading to the incorrect answer of 11. (2) Confusion between addition and exponentiation, i.e., the student may confuse the operation of exponentiation with addition, thinking that 9^2 means $9+2$. (3) Misremembering multiplication facts, i.e., the student might not remember that 9^2 is the same as 9×9 and so does not compute the multiplication to arrive at the correct answer of 81. From this example, it can be seen that the potential reasons given by the LLM are quite reliable.

Formally, for N questions in a quiz, LLMs are utilized to construct a potential reason list $P_i=\{p_1,p_2,\ldots,p_{K(i)}\}$ for each question q_i, where $K(i)$ is the number of possibilities for the student to give such a response for question q_i and p_j is the j-th possibility. Besides, the series of the potential reason lists of N questions in a quiz is defined as the student's candidate cognitive state, denoted as $C=(P_1,P_2,\ldots,P_N)$.

Identification among possibilities In the previous step, the LLM is asked to analyze students' response logs and list potential reasons for each response given by the students. With only analyzing individual questions, it cannot be further determined which possibility actually matches the student's cognitive state when dealing with the question in the quiz. At this point, it is necessary to consider the overall performance of the student in the entire quiz in order to make a judgment. For instance, in the aforementioned example in the previous subsection, if Alice gives the correct answer to another question about isosceles triangles, then it is highly probable that Alice's mistake was not due to the second possibility. Formally, given the student's candidate cognitive state $C=(P_1,P_2,\ldots,P_N)$, the next step is to determine which possibility is most likely to reflect the student's real cognitive state, i.e., to find a combination of possibilities, denoted as a vector $c=(c_1,c_2,\ldots,c_N)$ where $1\leqslant c_i\leqslant K(i)$ is an index indicator of the c_i-th possibility for question q_i.

To find the most likely combination of possibilities which can best reflect the student's cognitive state, this paper proposes to leverage LLMs as the simulator of students, i.e., LLMs are required to play a role of a student with a specific cognitive state and respond to the questions. If the answers provided by LLMs are close to the student's answers reported in the response logs, then the presumed cognitive state likely reflects the true cognitive state of the student. Formally, the LLM with a proper prompt asking them to play a role of such a student to deal with questions is regarded as a probability model, denoted as $F(\cdot)$, whose parameter is the cognitive state described by the combination of possibilities c, and the answers generated by the LLM under a given prompt are considered as samples, denoted as $s=(s_1,s_2,\ldots,s_N)$, where s_i is the output answer for question q_i. Given the observed dataset (i.e., the student's answers reported in the response logs) $r=(r_1,r_2,\ldots,r_N)$, the log-likelihood function of the estimated parameter c is $L(c;r)=\sum_{i=1}^{N}\log\Pr(s_i=r_i)$, where $s\sim F(c)$. Notably, if the questions are all objective questions with unique certain answers, the probability $\Pr(s_i=r_i)$ can be explicitly estimated, and then the whole problem just boils down to an optimization problem that minimizes the negative log-likelihood function, i.e., $c^*=\arg\min_{c\in C}(-L(c;r))$, where c^* is the optimal combination of possibilities. To address the above optimization problem, the most direct approach is to iterate through all possible combinations of size $\prod_{i=1}^{N}K(i)$. However, as the number of problems N increases, the search space for such combinatorial optimization problem grows exponentially. Thus when N is large, the cost of this approach is unacceptable. Therefore, this paper proposes to solve this optimization problem in a derivative-free greedy search manner. The detailed procedure is shown in Algorithm 1.

Algorithm 1 Derivative-free search for cognitive state

Input: Response logs of a student: $r=(r_1,r_2,\ldots,r_N)$
Output: The optimal combination of possibilities: c^*
Randomly initialize $c=(c_1,c_2,\ldots,c_N)$, where $c_i\in\mathbb{N}^+$, $1\leqslant c_i\leqslant K(i)$;
$L^*\leftarrow -L(c;r)$, $c^*\leftarrow c$;
while *not converge* **do**
  $(c_1,c_2,\ldots,c_N)\leftarrow c^*$;
  for $j=1$ **to** N **do**
    $c'\leftarrow(c_1,c_2,\ldots,c_{j-1},t,c_{j+1},\ldots,c_N)$, where t, c_j, and $1\leqslant t\leqslant K(i)$ is a positive integer that minimizes $-L(c';r)$;
    if $-L(c';r)<L^*$ **then**
      $L^*\leftarrow -L(c';r)$, $c_{\text{temp}}\leftarrow c'$;
    end
  end
  $c^*\leftarrow c_{\text{temp}}$.
end

Summarizing the cognitive state Via Algorithm 1, we can determine the optimal combination of possibilities c^*, representing the inferred cognitive state of the student. In contrast to traditional cognitive diagnosis, where a radar chart displays a student's mastery levels on different attributes, the inferred combination of possibilities c^* identifies the specific technical mistakes and thus offers a more fine-grained description of the student's cognitive state, although it may be a bit verbose, less intuitive, and contain redundant information. Therefore, in order to provide a more concise, clear, and intuitive description of the cognitive state, LLMs are required to give a brief summary according to c^*, generating a more readable conclusion about the student's cognitive state.

3 Experiments

This section shows experiments on real-world cognitive diagnosis tasks under the online education background. The dataset used is called *NeurIPS20* [8], which is derived from students aged from 8 to 10 who practice math on Eedi. Eedi is

a prominent online educational platform. The *NeurIPS20* dataset was used in a cognitive diagnosis competition at conference NeurIPS 2020, and the questions in *NeurIPS20* are all multiple-choice questions. We select practice logs from *NeurIPS20* that contain the side information of the question statement. The students' response logs are grouped according to individual quizzes. Students usually complete these quizzes within 1 hour, so the cognitive state during a quiz is assumed to be constant. On average, students answer about 11 questions in each quiz, which typically involve several related math attributes. As a result, a sub-dataset with 6147 students' response logs on 458 questions is obtained.

In experiments, one question is randomly extracted from the question set of each quiz as the test set, and the rest are used as the training set. The algorithms need to predict whether the student can answer the question correctly in the test set according to the student's responses in the training set. To study the performance of FineCD, we compare FineCD with DINA [3], IRT [2], KANCD [5], NCDM [4], LLM-Naive, and Human. DINA and IRT are traditional cognitive diagnosis algorithms, and NCDM as well as KANCD are representative algorithms of the deep learning-based cognitive diagnostic methods emerging in recent years. These four compared methods take the setting of traditional cognitive diagnosis, in which the input questions are featured by a Q-matrix rather than the question statements in FineCD. Furthermore, there is currently no mature approach to utilizing the side information of question statements in the literature of cognitive diagnosis. Therefore, this paper introduces the following two compared methods: LLM-Naive and Human. LLM-Naive directly provides the question statements and the student's responses to the LLM, which is then required to predict the student's response on the testing set. In the Human baseline, several graduate students majoring in computer science and technology are employed as human experts to carry out this task. Both LLM-Naive and Human take exactly the same setting as FineCD and thus are fair baselines. The prediction accuracy (ACC), area under the curve (AUC), and root mean square error (RMSE) are chosen as metrics, and the results of 10 independent repetitions are shown in Table 1. Besides, by conducting t-test with significance level = 0.05, FineCD significantly outperforms all compared methods on all metrics. Both DINA and IRT show effectiveness similar to random guessing, and the results of NCDM and KANCD are also unsatisfied. This highlights the challenge of traditional methods in small-sample individualized scenarios within online education. LLM-Naive, which can be viewed as an ablation version of FineCD, shows significantly inferior performance. The results of LLM-Naive emphasize the importance of the proposed three-step strategy, which is further supported by the post-experiment interviews in the Human baseline, which reveals that human experts also rely on strategy with a similar reasoning logic for making predictions.

Table 1 The performance of FineCD and compared methods on *NeurIPS20*. The significantly best results are marked in bold

	FineCD	IRT	DINA	NCDM	KANCD	LLM-Naive	Human
ACC/% ↑	**76.88**	49.74	49.85	55.70	65.31	64.37	71.50
AUC/% ↑	**78.29**	49.93	49.81	52.97	70.43	68.56	71.23
RMSE/% ↓	**38.73**	70.89	70.81	66.55	58.90	57.59	52.98

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62106076).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

References

1. Liu Y J, Zhang T C, Wang X C, Yu G, Li T. New development of cognitive diagnosis models. Frontiers of Computer Science, 2023, 17(1): 171604
2. Haberman S J. Identifiability of parameters in item response models with unconstrained ability distributions. ETS Research Report Series, 2005, 2005(2): i−22
3. De La Torre J. . Dina model and parameter estimation: a didactic. Journal of Educational and Behavioral Statistics, 2009, 34(1): 115−130
4. Wang F, Liu Q, Chen E H, Huang Z Y, Chen Y Y, Yin Y, Huang Z, Wang S J. Neural cognitive diagnosis for intelligent education systems. In: Proceedings of the 34th AAAI Conference on Artificial Intelligence. 2020, 6153–6161
5. Wang F, Liu Q, Chen E H, Huang Z Y, Yin Y, Wang S J, Su Y. NeuralCD: a general framework for cognitive diagnosis. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(8): 8312−8327
6. Chen X Z, Wu L, Liu F, Chen L, Zhang K, Hong R C, Wang M. Disentangling cognitive diagnosis with limited exercise labels. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. 2023, 792
7. Shen J H, Qian H, Zhang W, Zhou A M. Symbolic cognitive diagnosis via hybrid optimization for intelligent education systems. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. 2024, 14928–14936
8. Wang Z, Lamb A, Saveliev E, Cameron P, Zaykov Y, Hernández-Lobato J M, Turner R E, Baraniuk R G, Barton C, Jones S P, Woodhead S, Zhang C. Diagnostic questions: the neurips 2020 education challenge. 2020, arXiv preprint arXiv: 2007.12061