WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 4, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v4n1.11
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Python Language Teaching Evaluation System Based on LLM large model

Haotian Yang

Department of Engineering, Huihua College, Hebei Normal University, Shijiazhuang, Hebei Province, 050091, China

ABSTRACT

With the rapid development of information technology and artificial intelligence, programming education is very important. However, the traditional Python language teaching evaluation method has some problems, such as strong subjectivity and low efficiency, which is difficult to meet the current situation of contemporary programming education. In order to overcome the limitations of traditional teaching methods and comprehensively evaluate students 'programming ability, the system uses LLM model to analyze code, and combines with the Streamlit framework to build a user interface to achieve convenient interaction and feedback. The experimental results show that the system has high accuracy and reliability, which can effectively evaluate students 'programming level and provide support for personalized teaching. Then it can promote the development of programming education

KEYWORDS

Computational thinking; Large language model; Evaluate; Python language

1. INTRODUCTION

With the rapid development of the information technology industry, more and more people are starting to learn programming languages. Among them, Python language has become the first choice of most people in a simple and easy to learn way. Since its release by Guido van Rossum in 1991, Python has become one of the most popular programming languages in the world after more than 20 years of change. And Python language has an unconquerable position in the field of artificial intelligence. With the explosive growth of machine learning and big data analysis, there is a large talent gap in the artificial intelligence industry, so learning Python is a compulsory course for job seekers and researchers.

Although there are many evaluation methods in the teaching of Python programming language. Although traditional methods such as examinations, assignments and lab reports can reflect learning outcomes, they have their limitations, such as the difficulty of comprehensively assessing programming ability and the inconsistency of evaluation standards. The process assessment method focuses on the learning process, including classroom performance, group cooperation and project completion, and can understand students 'learning situation more comprehensively. However, the evaluation workload is large and the standard is difficult to unify. In order to evaluate students' ability comprehensively and objectively, the comprehensive assessment method combines a variety of evaluation methods, such as examinations, assignments, and classroom performance. This can improve the accuracy of assessment, but may increase the complexity. In the article of Ji Jin-jie [1], it has been pointed out that there is less need to digitally empower the evaluation of computational thinking, and further attention should be paid to the data collection, data analysis, data visualization

and other aspects of the computational thinking evaluation of primary and secondary school students to effectively use data to carry out teaching activities. The Python language teaching evaluation system developed based on LLM big model can better solve this problem. LLM model stands for Large Language Model, which has powerful natural language processing ability. It can understand students' code and problems and give corresponding suggestions and solutions. Compared with the traditional teaching evaluation method, based on the teaching evaluation has an automatic LLM, personalized, the characteristics of comprehensive, can automatically analyze students code, to reduce the workload of teachers, and LLM model from multiple dimensions to analyze the user's programming ability, like code integrity, and code efficiency, innovative thinking, etc, to assess.

At present, LLM model has been applied in different fields. Liu Changhong [2] proposed a new model of intelligent postgraduate teaching quality evaluation based on artificial intelligence technology, explored multi-dimensional and diversified evaluation indicators, and expounded the idea of a system architecture with artificial intelligence and big data analysis. Yang Yingxiu [3] verified the education evaluation ability from the three dimensions of evaluation environment, evaluation system and evaluation system, and optimized the education evaluation system, faced up to the network relationship of education evaluation, and Dongtai recognized the action process of education evaluation objects. Shen Cheng [4] discussed the application of ChatGPT in the evaluation of elementary school students' computational thinking, and proposed that the development trend of LLM evaluation application was the combination of man-machine, complementary advantages and comprehensive evaluation. Li Yi [5] analyzed the realistic premise, action mechanism and practice path of ChatGPT to enable the reform of education evaluation, and improved the ethical standard machine supervision and guarantee system of ChatGPT education evaluation. Xue Xin [7] pointed out that the AI big model promoted the personalization and efficiency of education model, and improved the equity and quality of education through the application of automated scoring, personalized learning, virtual teaching assistant and so on.

In order to achieve this goal, this paper will build a data-driven Web interface by using the Streamlit framework to accept user input for Python scoring and simple user interaction, and will call ZhiPuAI's LLM model to generate a model based on the user evaluation system, and improve the accuracy through caching.

2. METHODOLOGY

2.1. Preparation of the Data Set

In order to better train and fine-tune the large model, this experiment collects a large number of Python programming questions and answers data, as well as comments and annotations of related questions. The channels of collecting data sets include online programming platforms such as CSDN platform, Baidu Library, and PythonTip platform, crawling and downloading questions and annotations, totaling 2500 questions. The data set is divided into training set, validation set and test set, which are used for model training, adjusting model parameters and evaluating the model respectively. There are about 1500 questions in the training set, about 700 questions in the validation set and 300 questions in the test set. All the questions in the dataset are divided into four categories, including grammar and basic knowledge, data structure and algorithm, input and output file operation, application development and project practice.

2.2. Model Fine-tuning and Training

Will finish cleaning the Data set, which is used in the training set about 1500 questions, converted to JSON format for transmission, transfer to ZhiPuAI the model above, using the API interface to create a training request, at the same time set the API URL, Headers, and Data. Use the requests library to send the training request, and ensure that the Headers contains the correct Content-Type and

Authorization information, and the data contains the correct message list, temperature, max_tokens and other parameters. After sending the request, the server will send back a response, at this point, we need to parse the corresponding content to obtain the training results and model performance. Based on the returned data, we can adjust the parameters of the model to improve performance and accuracy. The following figure1 shows how this works:

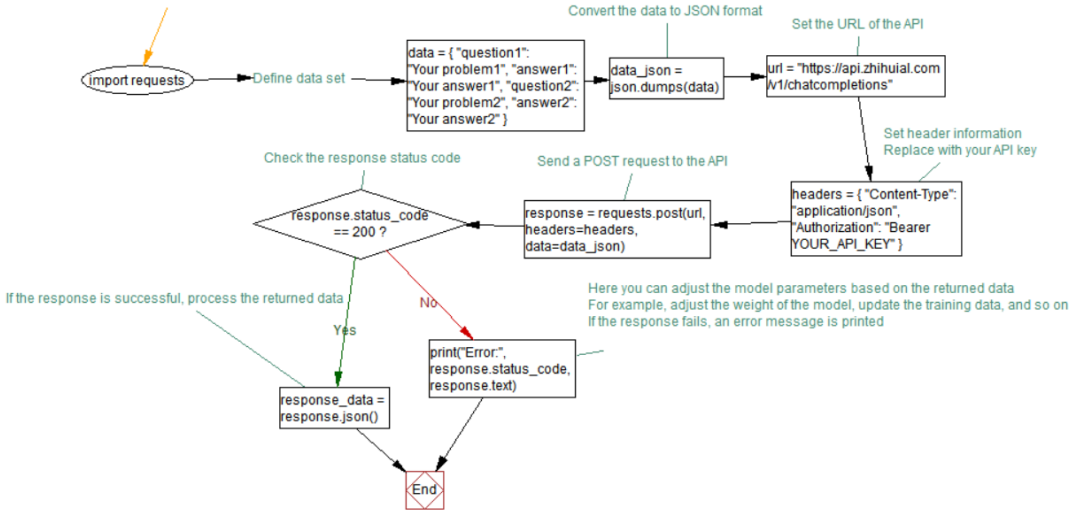


Figure 1. Process of Model fine-tuning and training

2.3. Key Module Implementation

Score input module: The score input module is the first time that the user interacts with the system. The user enters his/her name and scores in four fields. The module is implemented using Streamlit's slider and text input. The slider is used to collect the user's score in each field. Each slider corresponds to a score item, and the user can select the score by dragging the slider. Text input is used to collect the user's name, and this information can be applied to generate personalized ratings. The process is shown in the following figure2:

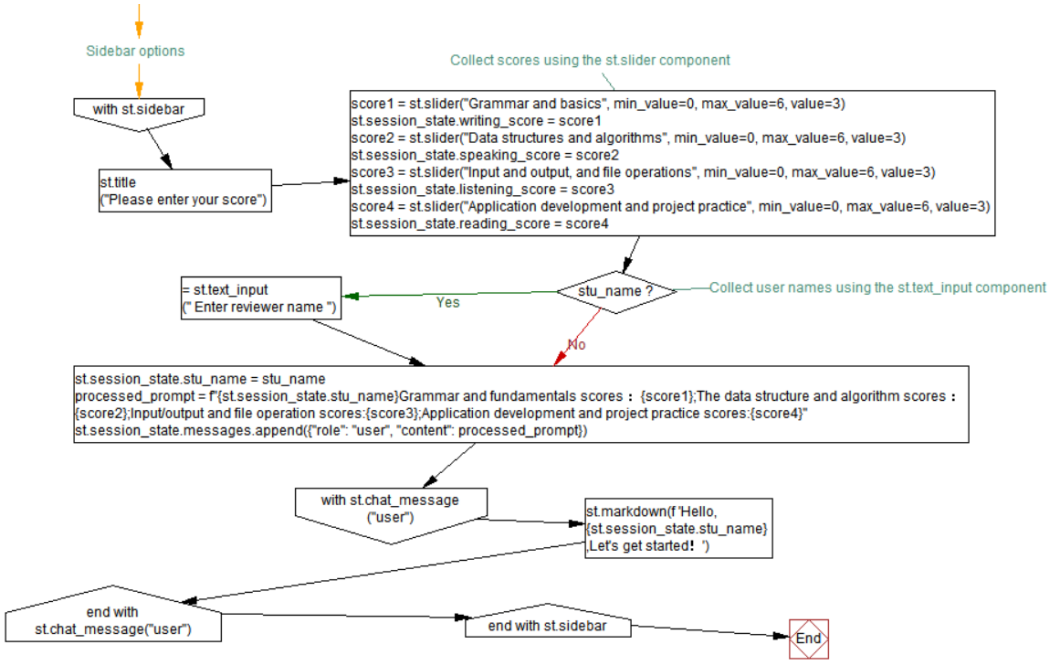


Figure 2. Process of score input module

Evaluation generation module: The evaluation generation module is the core of the system, and it is responsible for sending the user's rating information to the large model. This module builds a JSON format request and sets the necessary HTTP header information. Then, an HTTP client is used to send the request to the API endpoint of the AI service. Once the API request is complete, the AI starts generating the rating content. Finally, the system stores the rating content in the chat history to be displayed to the user in the chat interface. The process is shown in the following figure3:

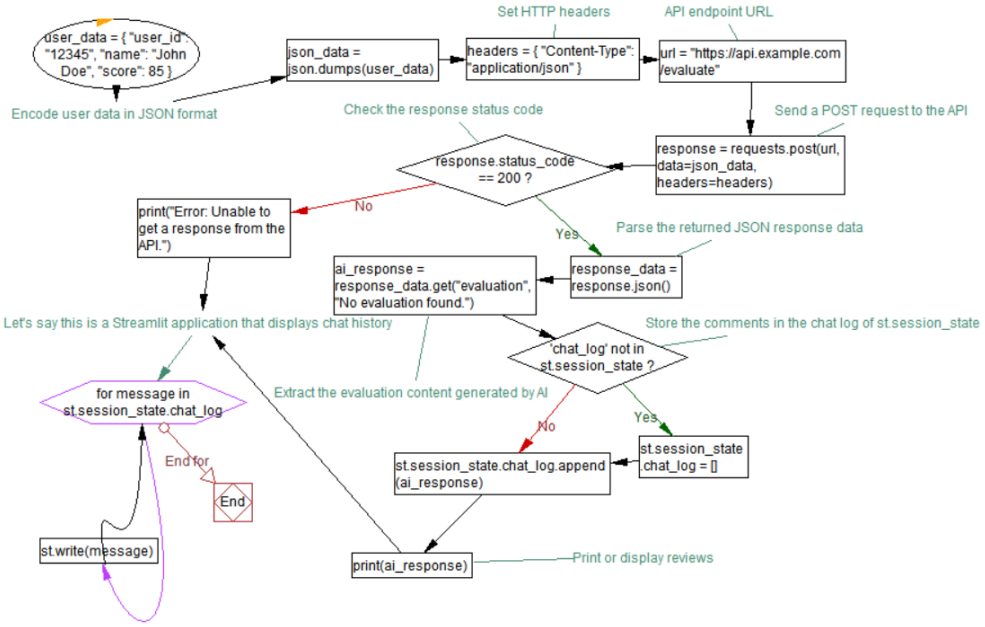


Figure 3. Process of evaluation generation module

Chat module: The chat module is an interactive component in the system. Users can chat with the AI assistant through the chat box, and the AI generated answers are displayed in Markdown format, making the answers easier to read and understand. In addition, the chat module provides a button that allows the user to clear the current chat history. When the user clicks this button, the chat history in st.session_state will be cleared and the chat interface reset to its initial state so that the user can start a new conversation. The process is shown in the following figure4:

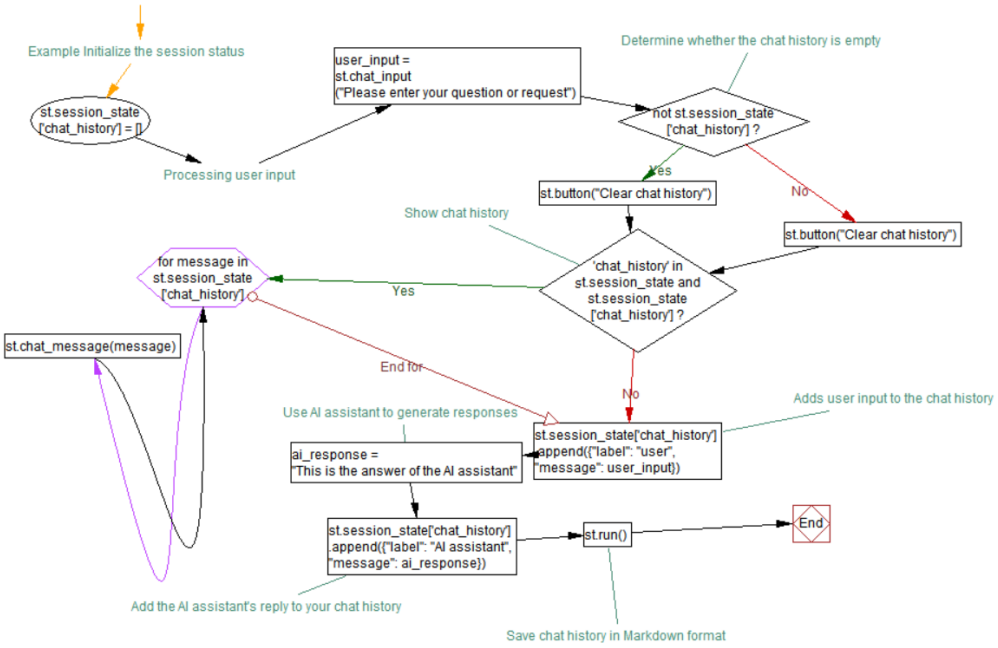


Figure 4. Process of chat module

3. EXPERIMENT DESIGN

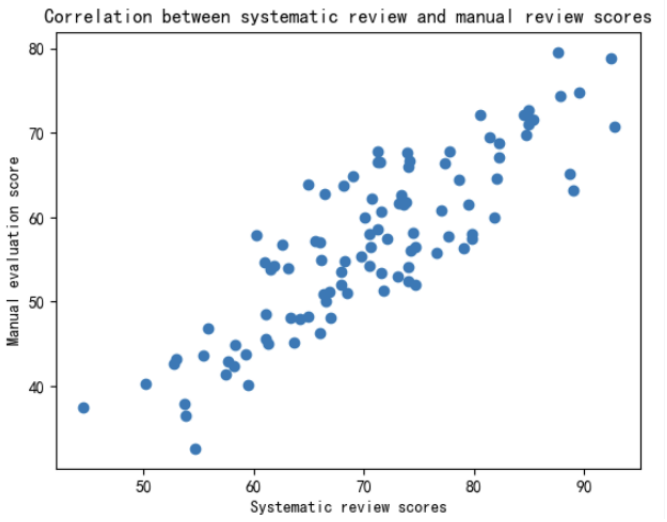
The experimental group was based on the LLM large model Python language teaching evaluation system, and the control group was the workers who participated in Python language teaching and scored the papers. The data set was 200 questions found from the CSDN platform, which were divided into four categories, covering grammar and basic knowledge, data structure and algorithm. Input and output file operation, application development and project practice, each category has a total of 15 single-choice questions, multiple choice questions, judgment questions, and 5 programming questions. 500 different people are found to complete the questions, and workers and put them into the big model for scoring and testing according to their scores.

3.1. Experimental Process

Firstly, 50 different Python language learners were invited to complete these questions, and their scores were manually scored according to their scores. Then these scores were put into the Python language teaching evaluation system, and the scoring analysis was also carried out. The evaluation standard was the examination standard of "National Youth Software Programming Grade Examination (Python)" [6]. In the process of scoring, we care about one aspect: the accuracy of assessment. For the accuracy of evaluation, correlation analysis and consistency analysis were used to evaluate the accuracy.

3.2. Analysis of Correlation Results

In this experiment, to assess the system evaluation and artificial correlation analysis, using the Pearson coefficient as a measure of relative standard and calculating the correlation coefficient between two variables, in particular, the correlation coefficient is close to 1, shows two coefficient is positive correlation, the more close to zero, shows two related coefficient is in the form of opposite, the more close to 1, The correlation coefficient is close to 1, indicating that the two coefficients are positively correlated.



T

Figure 5. Pearson correlation

The Pearson correlation coefficient is 0.859, indicating that there is a positive correlation between the two evaluation modes, and the P value is much less than 0.05, indicating that the correlation is statistically significant. In addition, through the heat map can show that most of the data are concentrated near the line, proved that there are two variables linear trend,

This result shows that there is a high degree of unity between the systematic evaluation and the manual evaluation, which can be used as an effective auxiliary tool in the evaluation of students.

3.3. Analysis of Consistency Results

In this study, consistency was used to analyze the relationship between the evaluation system and the manual evaluation, and Kappa statistic was used as the measure of the degree of consistency, with a calculated number of 500.

Firstly, a joint table is constructed to show the evaluation results of two evaluators for each object, such as:

Table 1. A joint table is constructed to show the evaluation results of two raters for each object

	Grammar and basics	Data structures and algorithms	Input and output file operations	Apply development and project practices
Systematic review	10	15	20	25
Human evaluation	12	18	22	28

Next, we set out to calculate the observational agreement probability Po, which represents the fraction of raters who actually agree with each other. $p_o = \frac{\sum f_{jj}}{n}$ The observation agreement probability is calculated as:, where is the sum of the elements on the diagonal, that is, the number of times the systematic and human reviews are the same on each category, and n is the total number of reviews. f_{ii} Then, we need to calculate the expected consistency probability Pe, which represents the proportion of reviewers who agree among themselves by chance. The expected agreement probability is calculated by the formula, and finally, we can calculate the Kappa value, which is given by the formula $p_e = \frac{\sum f_{i\times} f_i}{n2}$. $K = \frac{(Po - Pe)}{(1 - Pe)}$ By calculating the above formula, the following graph is obtained:

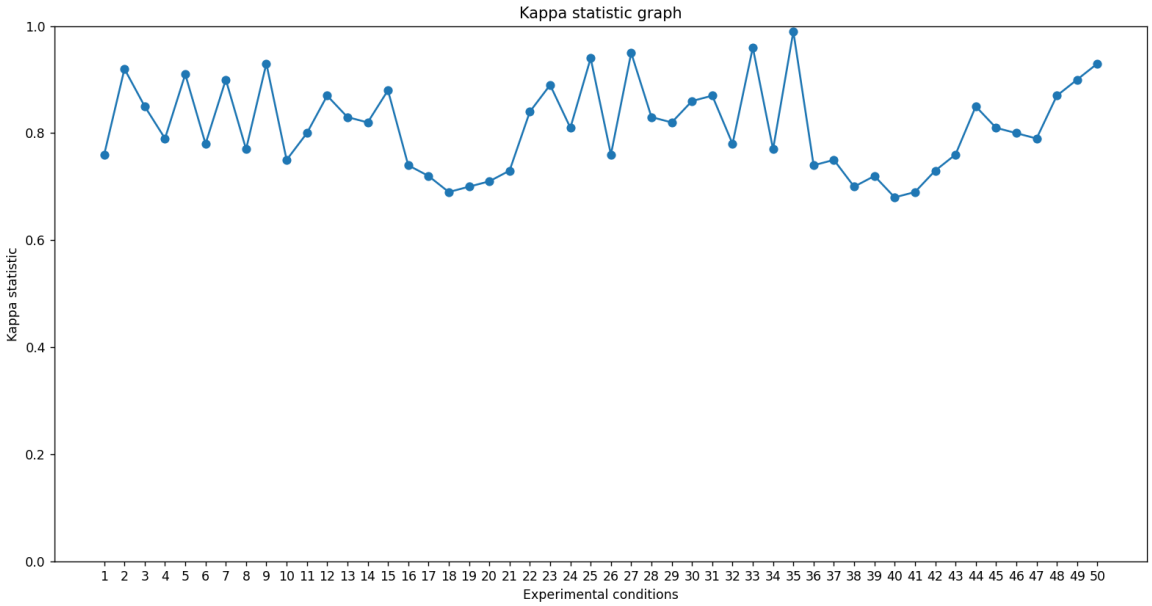


Figure 6. Kappa coefficient curve

It can be seen from the graph that the consistency level between the system evaluation and the manual evaluation is generally high, and most KAPPA values exceed 0.7, indicating that there is good consistency between the evaluation system and the manual evaluation. In detail, about 80% of the values are above 0.7, indicating that the system has good reliability, and some data are between 0.6 and 0.7. Some data are between 0.6 and 0.7. Although it is caused by some special circumstances and

the judgment of the supervisor, the overall consistency is still acceptable. According to the above data, it is concluded that systematic evaluation has high reliability. Compared with manual evaluation, systematic evaluation can be used as a tool to assist teachers in evaluation, thus improving the efficiency and accuracy of evaluation.

4. CONCLUSIONS AND PROSPECTS

4.1. Summary of the Paper

In this study, the effectiveness and accuracy of the Python language teaching system based on LLM large model are deeply explored through carefully designed experiments. In the experiment, combined with the official standards of the National Youth Software Programming Grade Examination (Python), 500 learners of different levels were tested, and educators related to Python language teaching were invited to manually score. Through the use of Pearson's related technology and consistency analysis, The relationship between the results of the system evaluation and the results of the manual evaluation is compared. And the following conclusions are drawn: 1. There is a significant positive correlation between the LLM large model in the Python language teaching evaluation system and the human evaluation, which indicates that the model has a strong ability of automatic scoring. 2. The consistency analysis results using Kappa statistic show that the system evaluation results are stable and reliable. 3. Compared with manual evaluation, the evaluation time based on the large model is shorter and faster.

The experimental results show that there is a positive correlation between the evaluation results of the large model scoring system and the human evaluation results, and the consistency analysis also determines the reliability of the system, which just shows that the Python language teaching evaluation system based on LLM large model has high accuracy and authenticity, and can provide an effective evaluation tool for programming education.

4.2. Looking to the Future

Although the Python language teaching evaluation system based on LLM large model is found to have high accuracy and reliability in this study, it is still necessary to be aware that factors such as model algorithm and sample statistics size may have an impact on the results. Therefore, future research should be committed to further optimize model algorithm, improve the evaluation accuracy and comprehensiveness, and through the analysis of data provide personalized learning advice and path, along with the continuous improvement of the model to create automated grading, easing the burden on teachers' work, improve the efficiency of teaching. Furthermore, the application of the technology of large LLM model in other programming languages, and disciplines of teaching evaluation, will provide a broader field of education with effective evaluation tool.

REFERENCES

[1] Ji Jinjie: Review of Computational Thinking Evaluation Research of Chinese Primary and Secondary School Students - Educational Communication and Technology - 2023-06-15

[2] Li Y, Zheng P Y, ZHANG T (2024) The realistic premise, mechanism and practice path of ChatGPT Empowering the reform of Education evaluation. Modern Distance Education, (online first edition).

[3] Shen, C., Bai, Y. (2023) The Application of Large Language Model in Computational Thinking Assessment of Primary School Students: A Case Study of ChatGPT. Digital Teaching in Primary and Secondary Schools, 25-28. (in Chinese)

[4] Liu Changhong, Jie Ping 'an, Hu Zhenxin, Jiang Aiwen. A new model of artificial intelligence Empowering graduate classroom teaching quality evaluation [Z]. Software Guide.

[5] Yang Yingxiu (2024) Research on the fitting of Educational evaluation ability and its ability structure. Teaching and Management, 16(16): 1-6. (in Chinese)

[6] Peng H L. Research on spatial relation recognition based on attention mechanism and graph neural network [D]. East China Normal University, 2023(02).

[7] Xue Xin, Application of AI Big Model in Education Industry [J]. Telecom Express, 2024, (03):33-38. (in Chinese)