Education and
Lifelong Development Research

Vol. 1, No. 4, p. 156-169, 2024

Original article

Capturing fine-grained teacher performance from student evaluation of teaching via ChatGPT

Boxuan Zhang[1], Xuetao Tian[2]*

[1]*The High School Affiliated To Renmin University of China, Tongzhou Campus, Beijing, China*

[2]*Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education (Beijing Normal University), Faculty of Psychology, Beijing Normal University, Beijing, China*

Keywords:
Student evaluation of teaching
teacher performance
automated labeling
ChatGPT

Cited as:
Zhang, B., & Tian, X. (2024). Capturing fine-grained teacher performance from student evaluation of teaching via ChatGPT. Education andLifelong Development Research, 1(4), 156-169.
https://doi.org/10.46690/elder.2024.04.01

Abstract:
Student evaluation of teaching (SET) is a vital component of educational enhancement, yet conventional assessment tools face inherent limitations. While open-ended questions provide a platform for students to convey authentic sentiments, the absence of automated labeling tools poses a challenge in the case of large-scale applications. In response, this study undertakes a comprehensive exploration, centered on the utilization of ChatGPT for capturing fine-grained teacher performance from SET. Based on a collected dataset and manual coding, the performance of ChatGPT with various strategies including zero-shot and few-shot, and some supervised models, including CNN, LSTM and BERT, are evaluated and compared. As a result, ChatGPT exhibits the promise of achieving commendable performance with a small number of labeled samples. This approach reduces the dependency on extensive labeled data, offering an effective solution. However, in terms of performance, a discernible margin persists in comparison to advanced supervised models, BERT. Our study also acknowledges there are various factors, such as task complexity and prompt clarity, influencing ChatGPT's performance and consistency. In summation, while the integration of ChatGPT into practical SET applications holds significant promise, further explorations are imperative to ensure the alignment of its capabilities with the intricate demands.

1. Introduction

The student evaluation of teaching (SET) is an important aspect of the education system. In addition to link to administrative decisions about teachers' promotions and merit pay raises (Annan et al., 2013; Beran & Rokosh, 2009; Emerson & Records, 2007; Toni & Sudin, 2024), it also provides valuable feedback that can help improve teaching practices and enhance the overall learning experience (Alwaely et al., 2023; Hoon et al., 2015; Smith, 2008). Students' perspectives and opinions can offer unique insights into the effectiveness of teaching methods, communication styles, classroom management and so on. The main benefit of student evaluations is that they give teachers an opportunity to understand how their teaching strategies impact students directly. By collecting feedback, teachers can identify their strengths and areas for improvement. This feedback loop fosters a culture of continuous growth and professional development among educators (Kulik, 2001; Nasser-Abu Alhija & Fresko, 2002; Pineda & Steinhardt, 2023; Wright, 2011).

Although student evaluation of teaching is very popular, it still mainly adopts the form of a rating scale, which has lots of disadvantages like biased rating and low reliability (Denson et al., 2010). Open-ended questions provide a new solution, where students have the opportunity to freely express themselves in their own words (Nasser-Abu Alhija & Fresko, 2009). However, reviewing these response data is an extremely time-consuming process, making it difficult to apply on a large scale. Previous research has explored the application of sentiment analysis techniques to solve automated labeling based on open-ended questions (Ren et al., 2023). Despite the

Yandy
Scientific
Press

*Corresponding author.
E-mail address: xttian@bnu.edu.cn (X. Tian); fanozhang@163.com (B. Zhang)

Received October 12, 2024; revised November 1, 2024; accepted November 18, 2024; available online December 12, 2024.

initial success, the methods always rely on a large amount of labeled data, therefore requiring lots of manual labor.

Recently, with the emergence of large-scale language models (LLMs) like ChatGPT (OpenAI, 2022), there are lots of researches focusing on the applications of ChatGPT in education (Alnaqbi & Fouda, 2023;Sedaghat, 2023; Situmorang et al., 2023; Zhu et al., 2023). It may become a new solution for the SET based on open-ended questions. The process only needs to define a task-specific prompt, combine it with the text to be analyzed, and input it to ChatGPT. ChatGPT will output the corresponding result via a dialogue. In this case, this paper aims to explore ChatGPT's potential for capturing teacher performance from SET.

2. Literature review

2.1 The forms of student evaluation of teaching

The student evaluation of teaching have been proved to have many benefits, but it is acknowledged that the evaluation instruments used for assessing teachers by students often rely heavily on rating scales and rarely adopted open-ended items (Denson et al., 2010; Onwuegbuzie et al., 2009;Ren et al., 2023). This reliance highlights a significant debate in the field: how to balance quantitative assessment with qualitative feedback, and which approach better captures teaching effectiveness. As is well known, rating scales have lots of disadvantages: (1) *Biased rating scores*. It is recognized that rating scores obtained from Likert-format scales can be subject to biases (Emerson & Records, 2007). Various biases can influence students' ratings, such as leniency or severity biases, central tendency bias, or contrast effects. These biases can impact the accuracy and reliability of the evaluation results, potentially leading to skewed or inaccurate understanding on teacher performance. Also, these biases challenge the validity of using standardized scales across diverse educational settings, suggesting that context-specific factors may not be adequately captured. (2) *Variability in evaluation instruments*. The instruments can vary significantly in terms of item quality, operationalization of the teaching-effectiveness construct, and specific dimensions included (Annan et al., 2013; Donnon et al., 2010). This variability can make it challenging to compare results across different instruments or institutions. Additionally, the development of evaluation instruments is often influenced by practical considerations, and psychometric evaluation of these instruments is not always systematically conducted. This can raise concerns about the validity and reliability of the evaluation process. (3) $Haloeffect$. The halo effect refers to the tendency for students to evaluate a teacher's overall performance based on their general impression or a single positive/negative aspect. This can lead to inflated or deflated ratings, where students' perceptions of one dimension heavily influence their ratings across other dimensions (Clayson & Haley, 2011). The halo effect can distort the accuracy of the evaluation results, as it does not provide a comprehensive and nuanced assessment of teaching effectiveness (Beran et al, 2007). Compared to highly-structured rating scales, evaluation text obtained through open-ended questions can offer valuable insights and a more nuanced understanding of students' evaluations on their teachers (Stupans et al., 2015).

Open-ended questions provide students with the opportunity to freely express themselves in their own words. This can lead to a more comprehensive and detailed feedback that focuses on what students perceive as most important. Also, such an approach allows students to provide qualitative feedback, share specific examples, and highlight aspects of teaching that may not be captured by rating scales alone (Hammond et al., 2003; Hodges & Stanton, 2007). It enables students to provide richer descriptions of their experiences, offer suggestions for improvement, and express their thoughts in a more personalized manner. This qualitative feedback can provide context, clarity, and deeper insights into teaching practices, allowing teachers to better understand their strengths and areas for growth. Moreover, such open comments can help identify variables in teaching that may not be covered by pre-determined rating scale items (Nasser-Abu Alhija & Fresko, 2009). They can shed light on specific teaching strategies, classroom dynamics, communication styles, or other factors that students consider crucial to their learning experience. By giving students the freedom to express themselves, open-ended questions can uncover a broader range of variables that influence teaching effectiveness. However, analyzing and interpreting these comments can be time-consuming and subject to subjective interpretations. Managing a large volume of comments can also pose challenges (Brockx et al., 2012; Rajput et al., 2016). Therefore, it is necessary to explore an automated labeling method based on open-ended questions in student evaluation of teaching. By combining quantitative and qualitative methods, SET can offer a more balanced and insightful understanding of teaching practices, contributing to a more comprehensive theoretical framework for evaluating education.

2.2 Automated labeling method for open-ended questions in student evaluation of teaching

With the development of natural language processing (NLP), extracting valuable opinions from writing data has become possible (Chong et al., 2020). Especially, sentiment analysis techniques (Medhat et al., 2014) can be helpful for analyzing text data in student evaluation of teaching, when dealing with a large volume of open-ended responses (Okoye et al., 2023; Rajput et al., 2016). Traditional sentiment analysis tools, like VADER and TextBlob, offer robust solutions for basic sentiment classification but may struggle with the complexity of educational text (Hutto & Gilbert, 2014; Hazarika et al., 2020). Nowadays, sentiment analysis based on machine learning can be performed at two levels: document-level analysis and aspect-level analysis (Jin et al., 2023;Srinivas & Hanumanthappa, 2017). Document-level sentiment analysis, also known as overall sentiment analysis, aims to determine the overall polarity of a piece of text as either positive, negative, or neutral. This approach provides a high-level summary of the sentiment expressed in the entire document. While document-level sentiment analysis is relatively straightforward and easy to implement, it lacks fine-grained information about specific aspects of teachers. Aspect-level sentiment analysis

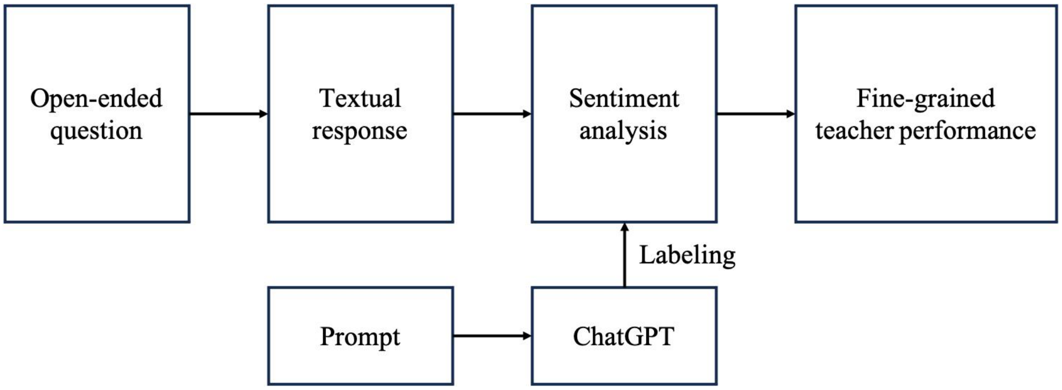

Fig. 1. The procedure of capturing fine-grained teacher performance from student evaluation of teaching via ChatGPT.

aims to identify the sentiment expressed toward specific aspects or dimensions within the text. In the SET context, this approach can analyze comments to determine the sentiment related to various aspects of teaching, such as teacher quality, communication, classroom management, course content, and more (Chantamuang et al., 2022). Although this approach is more complex and requires more advanced NLP techniques, such as aspect extraction and sentiment classification for each aspect, it provides more detailed and actionable insights on teaching practices. Previous study has utilized aspect-level sentiment analysis into student evaluation of teaching (Ren et al., 2023; Sindhu et al., 2019). Despite the initial success, aspect-level sentiment analysis is always implemented by supervised learning (Khanam, 2023; Su & Peng, 2023), where output aspects must be predefined and thus the trained models are difficult to transfer and use across populations and scenarios. In other words, for a specific task, it is necessary to establish lots of labeled data and further train a task-specific model. For a complex task, the cost of manually labeling data is also enormous. To advance this field, future research should explore the unsupervised methods that reduce dependency on pre-labeled data, thereby enhancing the generalizability and applicability of automated labeling techniques in SET.

3. The present research

For the application of student evaluation of teaching based on open-ended questions, following the paradigm of aspect-level sentiment analysis, we employ ChatGPT as the automated labeling model. The procedure is shown in Fig. 1.

Our methodology involves the formulation of prompts to interact with ChatGPT and directly extract aspect-level sentiment analysis results from the dialogue. To assess its efficacy, we annotated a dataset pertaining to student evaluation of teaching in Chinese junior schools. Subsequently, we conducted a comparative analysis between ChatGPT, human, and supervised models to investigate the potential of ChatGPT for capturing fine-grained teacher performance from student evaluation of teaching. We considered two prompt scenarios: one wherein we provided task descriptions without any labeled data (named as zero-shot prompt) and another where we included a limited amount of labeled data (named as few-shot prompt). Our research centers around three primary questions:

- Q1: What is the performance of ChatGPT for capturing fine-grained teacher performance from student evaluation of teaching when no labeled data is provided?
- Q2: How does the performance of ChatGPT, without labeled data, compare to that of supervised models for analyzing student evaluation of teaching?
- Q3: Can the performance of ChatGPT be enhanced by the inclusion of labeled samples, and if so, what strategies contribute to this improvement? This following sections presents our data collection, research design, and experimental results to answer these questions and shed light on the potential of ChatGPT in the context of analyzing student evaluation of teaching.

4. Materials and methods

4.1 Participants

The dataset is collected from Chinese junior school. A total of 99 teachers for grade 7 to grade 9 participated in the study, consisting of 12 males and 87 females. Among them, 33 people teach mathematics, 34 people teach English, and the other 32 people teach Chinese. Each of them were evaluated by 10 students from their classes.

4.2 Tool

For each teacher, his/her students were asked to answer an open-ended question, that is, “Some students think he/she is a good teacher, but some students don’t think so. What do you think of him/her? Please elaborate on his/her performance in various aspects.” In Chinese cultural background, the students may feel pressure when evaluating their teachers. Therefore, to eliminate the sensitivity of the question, the question was presented after two opposite sentiment orientations. The answers should be more acceptable. An example is presented in Table 1.

4.3 Data cleaning

For the collected data, some short responses were deleted due to the content being incomplete or obviously meaningless like repeated punctuation marks. Then, 897 valid responses on student evaluation of teaching were retained.

Table 1. An example of the collected responses on student evaluation of teaching.

Question:	Some students think he/she is a good teacher, but some students don't think so. What do you think of him/her? Please elaborate on his/her performance in various aspects.
Response:	Teacher, we appreciate your lively and interesting classes, but there is a recurring issue with procrastination. When the bell rings after class, our attention starts to wane as we feel the focus is lost. Some students find you to be a good teacher, while others have concerns about this procrastination tendency. Additionally, there have been instances of short temper, which affects the overall classroom atmosphere. It would be beneficial if you could address these issues to further improve the learning experience for all students. Overall, your teaching is good.

Table 2. The aspects and their descriptions within the open-ended responses on student evaluation of teaching.

ID	Aspect	Description
1	Teacher quality	Stable personal characteristics, including personality, temper and morality.
2	Teacher image	Physical characteristics, including dress and posture.
3	Teaching method	Instruction methods during the teaching.
4	Teaching content	Subject knowledge or life experience passed on to students.
5	Teaching ability	Knowledge level and classroom management.
6	Teaching attitude	Attitudes towards both students and teaching process.
7	Teaching effectiveness	Students' achievement, interest and self-confidence during the teaching process.
8	Teacher-student relationship	Dynamic and interpersonal connection that exists between a teacher and their students.
9	Classroom atmosphere	Overall attitudes and emotions within the teaching environment.

4.4 Manual coding on open-ended responses

The responses on open-ended question may involve several aspects of teaching performance. Referring to previous research (Wang, 2018), they were annotated from the following nine aspects: teacher quality, teacher image, teaching method, teaching content, teaching ability, teaching attitude, teaching effectiveness, teacher-student relationship and classroom atmosphere. Their descriptions are presented in Table 2.

During the coding process, three annotators separately assigned each sentence of the responses into one specific aspect via Nvivo-11 and identified its sentiment tendency as positive, negative or neural. To assess the inter-rater reliability of the annotation process, Cohen's Kappa coefficient was employed and the result indicated an initial Kappa value of 0.86, which is generally interpreted as almost perfect agreement according to the guidelines proposed by Landis & Koch (1977). Any inconsistent results will be discussed until they are consistent. For an open-ended response, its annotation results are obtained by combining the results on all its sentences. Specifically, for a specific aspect, if there are only some sentences with positive sentiment, it is labeled as "1"; If there are only some sentences with negative sentiment, it is labeled as "0"; If there are both some sentences with positive sentiment and some sentences with negative sentiment, it is labeled as "2"; If the aspect is not mentioned, it is labeled as "-1". Finally, the label distribution on the nine aspects are shown in Table 3.

4.5 Labeling methods

To support the following research, we leverage three kinds of methods to establish the automated labeling on the open-ended responses provided by students: zero-shot prompt on ChatGPT, supervised modeling, and few-shot prompt on ChatGPT.

4.5.1 Zero-shot prompt on ChatGPT

In the case of zero-shot prompt, ChatGPT will not be provided with any labeled sample, but only a task description. There are three types of prompts attempted as shown in Table 4: (1) Simultaneity: Simultaneously output sentiment tendencies across all dimensions with one prompt; (2) Parallel: Use different prompts for each dimension and output sentiment tendencies one dimension at a time; (3) Pipeline: Use a pipeline strategy, where which dimensions have sentiment tendencies is first determined before making specific judgments about tendencies.

4.5.2 Supervised modeling

In the case of supervised modeling, the labeled dataset will be divided into training set and testing set, where the training set is used to train the labeling model and the testing set is used to evaluate the trained model. In this study, we adopt a cross-validation strategy to ensure that each sample is scored once. Concretely, the dataset is divided into three equal parts and each part serves as the testing set once. We utilize three supervised models to establish the labeling model, including convolutional neural network (CNN) (Kim, 2014), long short-term memory (LSTM) (Nowak et al., 2017) and bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2019). The choice of these models is rooted in their representation of different stages in the evolution of deep

Table 3. Label distribution of the collected data.

Label	Aspect ID								
	1	2	3	4	5	6	7	8	9
2	51	3	37	71	139	61	24	12	8
1	466	78	469	257	436	600	336	222	376
0	39	4	38	78	26	20	63	35	13
-1	341	812	353	491	296	216	474	628	500

Table 4. Zero-shot prompts on ChatGPT.

Type	Prompt
Simultaneously	For this response <textual response> , are the sentiments positive or negative for the aspects of “teacher quality”, “teacher image”, “teaching method”, “teaching content”, “teaching ability”, “teaching attitude”, “teaching effectiveness”, “teacher-student relationship” and “classroom atmosphere”? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”.
Parallel	For this response <textual response>, is the sentiment positive or negative for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”.
Pipeline	(1) For this response <textual response>, which aspects of “teacher quality”, “teacher image”, “teaching method”, “teaching content”, “teaching ability”, “teaching attitude”, “teaching effectiveness”, “teacher-student relationship” and “classroom atmosphere” are mentioned? (2) For this response <textual response>, is the sentiment positive or negative for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”.

learning applications within the field of natural language processing (NLP). Each model captures unique aspects of language processing, reflecting advancements in handling text data from capturing local features to understanding global context and finally leveraging pre-trained language models for comprehensive semantic modeling.

4.5.3 Few-shot prompt on ChatGPT

In the case of few-shot prompt, ChatGPT will be provided with ten labeled samples for each aspect and sentiment tendency, which are randomly selected from the dataset. Two distinct strategies have been employed: static and dynamic, as shown in Table 5. The static strategy involves directly placing samples in the prompt and inputting them to ChatGPT. Conversely, the dynamic strategy entails evaluating ChatGPT’s response and providing feedback based on manual coding after making an assessment.

4.6 Performance evaluation

To evaluate the performance of above-mentioned automated labeling methods, we adopt commonly-used indicators of multi-class classification including accuracy (Acc), weighted F1-score (W-F1), and macro F1-score (M-F1). Especially, the label “0” and “2”, indicating that negative sentiments of some aspects are expressed, should be given special attention and are more beneficial for improving the teaching process. Thus, we also report the recall (Rec), precision (Prec) and F1-score on the both labels. Additionally, considering that ChatGPT is an online application and its output will may fluctuate over time, the methods based on ChatGPT will perform two times and report their consistency (Cons) as well as the average performance.

5. Results

In this section, we provide a concise and precise description of the experimental results.

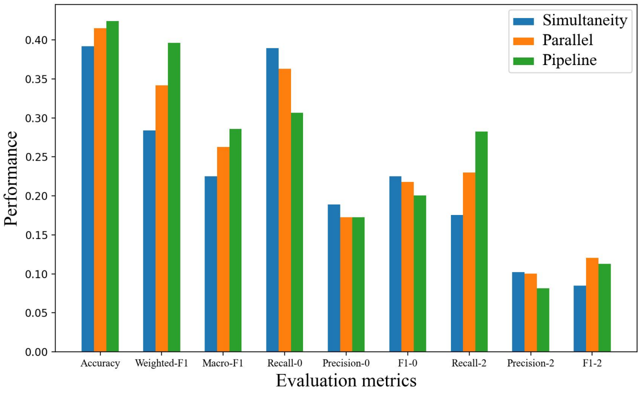

Fig. 2. Performance comparison among the three strategies of zero-shot prompt on ChatGPT.

Table 5. Few-shot prompts on ChatGPT.

Type	Prompt
Static	for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. There are some examples for you: (1) <example of textual response>: <true label>; (2) <example of textual response>: <true label>; ... (10) <example of textual response>: <true label¿>; Please only output “-1” or “0” or “1” or “2”.
Dynamic	For this response <example of textual response>, is the sentiment positive or negative for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”. ChatGPT: 1 Yes, you are right. Next one: For this response ¡example of textual response¿, is the sentiment positive or negative for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”. ChatGPT: 2 No, you are wrong. The true label is “-1”. Next one: For this response <example of textual response>, is the sentiment positive or negative for the aspect of <aspect name>? If there are only positive sentences, output “1”; If there are only negative sentences, output “0”; If there are both positive and negative sentences, output “2”; And if the aspect is not mentioned, output “-1”. Please only output “-1” or “0” or “1” or “2”. (The above process will be repeated ten times, before zero-shot prompt is utilized on ChatGPT.)

5.1 Performance of zero-shot prompt on ChatGPT (for Q1)

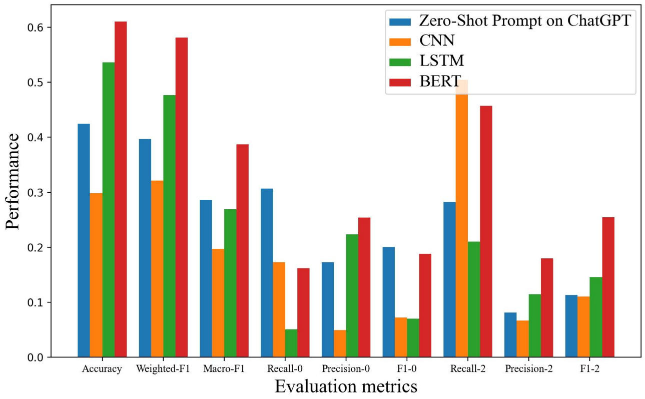

Fig. 3. Performance comparison between supervised models and zero-shot prompt on ChatGPT.

Table 6, 7, 8, and Fig. 2 present the performance of three strategies of zero-shot prompt on ChatGPT. First, the consistencies of the three strategies are 0.922, 0.844, and 0.736 respectively. Generally speaking, they are all acceptable, but implying that the consistency of ChatGPT labelings will decrease as the steps of task increases. Also, Table 9 shows the consistencies among the three strategies, demonstrating that ChatGPT has a relatively consistent semantic understanding of different prompts. Then, the overall performance on automated labeling is improved as the task is dismantled. Although the improvement is may seem limited, it should indicate that ChatGPT is better at handling a simple task than a complex task. Moreover, as shown in the above dataset, there are fewer negative evaluations from the students in actual situations, but ChatGPT will output a large number of negative evaluations, resulting in low precision of negative evaluation scores. In this case, ChatGPT is able to generate stable labeling, but not enough to replace manual labeling obviously.

5.2 Performance of supervised modeling (for Q2)

ChatGPT, lacking labeled data, cannot attain a performance level comparable to manual labeling. However, it’s worth noting that even supervised models, despite access to labeled data, are unable to achieve perfect human-like consistency. Therefore, it is necessary to quantity the performance of supervised models in order to further evaluate the potential of ChatGPT. Table 10, 11, 12 present the performance of CNN, LSTM, and BERT, and Fig. 3 shows the comparison between them and zero-shot prompt on ChatGPT. Intuitively, the comparative performance analysis of CNN, LSTM, and BERT models reveals a clear order of effectiveness: BERT exhibits superior performance over LSTM, which, in turn, outperforms CNN. Such a performance hierarchy is the same as existing studies (Ren et al., 2023;Tian et al., 2022). For ChatGPT, the ability to surpass the performance of the CNN model emphasizes its potential efficacy for student evaluation of teaching. Moreover, from the performance of each aspect,

Table 6. Performance of simultaneity strategy via zero-shot prompt on ChatGPT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Cons	.977	.966	.944	.942	.944	.941	.776	.917	.887	.922
Acc	.519	.094	.526	.310	.493	.627	.347	.239	.376	.392
W-F1	.372	.028	.406	.177	.373	.560	.252	.126	.265	.284
M-F1	.201	.084	.299	.207	.271	.327	.257	.197	.187	.225
Rec-0	.039	.376	.527	.257	.308	.700	.397	.600	.308	.390
Prec-0	.096	.046	.253	.400	.147	.167	.323	.201	.067	.189
F1-0	.055	.082	.342	.312	.198	.269	.356	.301	.110	.225
Rec-2	.020	.333	.122	.014	.115	.205	.354	.292	.125	.176
Prec-2	.100	.042	.092	.037	.368	.191	.045	.034	.009	.102
F1-2	.033	.074	.105	.020	.175	.198	.080	.062	.017	.085

Table 7. Performance of parallel strategy via zero-shot prompt on ChatGPT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Cons	.865	.870	.807	.777	.863	.871	.822	.847	.875	.844
Acc	.505	.100	.528	.325	.497	.659	.407	.295	.419	.415
W-F1	.418	.059	.477	.267	.411	.608	.314	.220	.304	.342
M-F1	.297	.061	.364	.253	.285	.409	.252	.211	.233	.263
Rec-0	.192	.250	.526	.237	.289	.725	.206	.343	.500	.363
Prec-0	.127	.013	.281	.269	.131	.257	.204	.155	.117	.173
F1-0	.153	.024	.367	.252	.180	.379	.205	.213	.190	.218
Rec-2	.353	.333	.338	.141	.065	.361	.125	.167	.188	.230
Prec-2	.183	.015	.110	.079	.180	.233	.047	.028	.030	.101
F1-2	.241	.029	.165	.101	.095	.283	.068	.048	.052	.120

Table 8. Performance of pipeline strategy via zero-shot prompt on ChatGPT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Cons	.778	.788	.687	.656	.738	.744	.741	.719	.770	.736
Acc	.483	.128	.530	.372	.482	.627	.420	.343	.435	.424
W-F1	.436	.119	.525	.351	.439	.622	.374	.319	.385	.397
M-F1	.299	.081	.378	.284	.313	.400	.280	.261	.279	.286
Rec-0	.192	.250	.408	.160	.212	.525	.198	.314	.500	.307
Prec-0	.165	.020	.254	.283	.129	.195	.214	.164	.131	.173
F1-0	.178	.036	.313	.205	.160	.284	.206	.216	.208	.200
Rec-2	.333	0	.378	.155	.151	.336	.188	.500	.500	.282
Prec-2	.109	0	.092	.078	.167	.158	.039	.049	.040	.081
F1-2	.164	0	.147	.104	.159	.215	.065	.089	.075	.113

supervised models can be significantly affected by the data distribution, e.g. BERT achieves better performance of label “2” on the aspect teaching content (ID = 4) and teaching ability (ID = 5); and LSTM does not output label “0” or “2” on the aspect of teacher image (ID = 2). As a generative model, ChatGPT may be easier to avoid this situation.

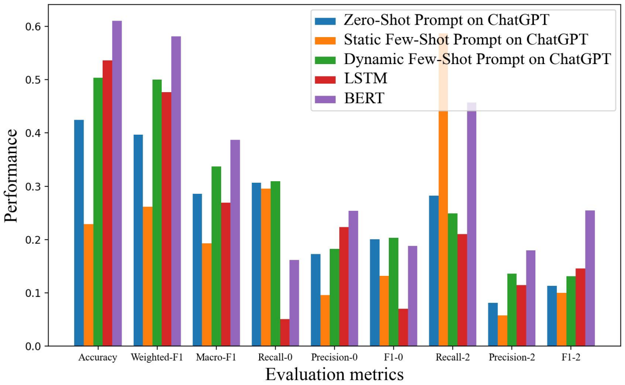

Fig. 4. Performance comparison between few-shot prompt on ChatGPT and supervised models as well as zero-shot prompt on ChatGPT.

Table 9. Consistencies among the three strategies of zero-shot prompt on ChatGPT.

	Simultaneously	Parallel	Pipeline
Simultaneously	1.00	–	–
Parallel	0.78	1.00	–
Pipeline	0.70	0.75	1.00

5.3 Performance of few-shot prompt on ChatGPT (for Q3)

ChatGPT's advantage in automated labeling lies in its potential to reduce the manual effort required for labeling data, which can be a time-consuming and resource-intensive process. However, as above-mentioned, the performance of ChatGPT without labeled data cannot be satisfactory. Thus, the idea of using a small amount of labeled data to improve ChatGPT's performance is indeed intriguing. Table 13 and 14 present the performance of two strategies of few-shot prompt on ChatGPT. Based on the obtained results, the integration of a limited number of samples within a prompt yields an outcome of notably low consistency when employed with ChatGPT. This outcome signifies that such a strategy actually leads to a deviation in their understanding of the task. While acknowledging the possible influence of data sampling in this context, the findings underscore that the static few-shot prompt on ChatGPT is insufficient for achieving proficiency in automated labeling for student evaluation of teaching. In contrast, dynamic strategy provides feedback after ChatGPT's initial labeling, which may trigger better reflection. Table 14 demonstrates that dynamic few-shot prompt on ChatGPT can significantly improve performance the maintain consistency. Also, Fig. 4 shows the comparison between few-shot prompt on ChatGPT and supervised models. Based on the empirical findings, the few-shot strategy demonstrates the potential to serve as a bridge connecting ChatGPT's inherent language understanding capabilities with the task-specific demands in analyzing student evaluation of teaching. This approach, involving the integration of a limited number of labeled samples, has led to ChatGPT surpassing the performance of LSTM, notably in terms of weighted F1-score and macro F1-score. However, while ChatGPT's performance has showcased promising advancement, a discernible performance gap remains between ChatGPT and BERT, particularly for the label "2". It is evident that content encompassing both positive and negative sentiments poses a formidable challenge for ChatGPT in achieving precise recognition, which emerges as a significant hurdle in its pursuit of automated labeling for student evaluation of teaching.

6. Discussions

The ensuing section draws attention to a pivotal aspect of our study, where we illuminate several prominent issues that demand thorough discussion. These issues, arising from the intricacies of our research methodology and the nuances of the data under examination, merit in-depth exploration to provide a comprehensive understanding of the research outcomes. By addressing these issues candidly, we aim to enhance the transparency and robustness of our research, ultimately contributing to a more holistic comprehension of the broader research landscape.

6.1 Usability of ChatGPT for analyzing student evaluation of teaching

Previous studies have explored the scoring ability of LLMs like ChatGPT extensively (Hommel, 2023; Wang et al., 2024; Zhao et al., 2024). However, there are also conflicting conclusions between them, mainly focusing on whether they are effective raters and their stability. In our study, the performance of ChatGPT, surpassing that of LSTM while falling short of BERT's level, underscores its intriguing potential and usability for analyzing student evaluation of teaching. In this case, ChatGPT shows the capacity to implement automated SET. In terms of usability, ChatGPT presents itself as a valuable tool that balances performance with the accessibility of labeled data. While BERT showcases superior accuracy, ChatGPT exhibits commendable results that might be suitable for certain contexts within education evaluation. Nonetheless, the observed performance gap between ChatGPT and BERT, particularly in scenarios with mixed sentiments (both positive and negative ones), hints at the challenges of capturing intricate linguistic nuances. This raises considerations of granularity and precision in automated labeling, especially when evaluating the teaching performance in education. As such, the usability of ChatGPT for analyzing student evaluation of teaching lies in its potential to complement existing solutions. ChatGPT may offer valuable insights, particularly when rapid assessment and feedback are required. Yet, its dynamic nature and occasional deviation from optimal accuracy necessitate vigilant integration to achieve a stable evaluation. Overall, in this study, ChatGPT with few labeled samples has demonstrated its usability but does not achieve a state-of-the-art performance. Its specific role and integration strategies should be thoughtfully considered in light of its strengths and limitations within the context of automated labeling and the broader educational landscape. Meanwhile, with the development of

Table 10. Performance of supervised modeling via CNN.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Acc	.462	.390	.231	.389	.219	.488	.095	.152	.260	.298
W-F1	.371	.491	.278	.440	.229	.502	.094	.182	.306	.321
M-F1	.231	.193	.162	.317	.183	.262	.107	.127	.194	.197
Rec-0	0	0	.053	.180	.077	.100	.492	.343	.308	.172
Prec-0	0	0	.036	.094	.025	.059	.141	.036	.051	.049
F1-0	0	0	.043	.123	.037	.074	.219	.065	.087	.072
Rec-2	.628	.333	.460	.479	.453	.312	.708	.417	.750	.504
Prec-2	.142	.014	.033	.128	.136	.078	.027	.029	.016	.067
F1-2	.231	.027	.062	.202	.209	.124	.052	.054	.030	.110

Table 11. Performance of supervised modeling via LSTM.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Acc	.483	.826	.512	.377	.510	.641	.408	.593	.477	.536
W-F1	.371	.838	.402	.382	.400	.547	.341	.607	.404	.477
M-F1	.228	.280	.242	.308	.291	.259	.254	.310	.254	.270
Rec-0	.026	0	.053	.115	0	0	.159	.029	.077	.051
Prec-0	.250	0	.333	.150	0	0	.222	.056	1	.224
F1-0	.047	0	.091	.130	0	0	.185	.038	.143	.070
Rec-2	.275	0	.189	.437	.597	.312	.083	0	0	.210
Prec-2	.119	0	.073	.186	.421	.202	.029	0	0	.114
F1-2	.166	0	.105	.261	.494	.245	.044	0	0	.146

Table 12. Performance of supervised modeling via BERT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Acc	.564	.924	.564	.506	.537	.641	.452	.701	.605	.611
W-F1	.516	.930	.514	.520	.466	.557	.433	.717	.584	.582
M-F1	.386	.456	.375	.442	.353	.320	.349	.433	.368	.387
Rec-0	.128	0	.237	.231	.039	.100	.365	.200	.154	.162
Prec-0	.263	0	.360	.367	.125	.400	.307	.212	.250	.254
F1-0	.172	0	.286	.284	.059	.160	.333	.206	.191	.188
Rec-2	.765	.333	.405	.732	.763	.492	.333	.167	.125	.457
Prec-2	.229	.143	.134	.267	.387	.242	.076	.087	.050	.179
F1-2	.353	.200	.201	.391	.513	.324	.124	.114	.071	.255

large language models, the performance could be expected to further improve.

6.2 Manually-designed prompt on ChatGPT

Obviously, the manually-designed prompt significantly influences the performance of ChatGPT. A well-constructed prompt can guide ChatGPT's responses towards desired outcomes, enhance its understanding of the task, and yield more relevant and accurate responses. Conversely, an inadequately designed prompt may lead to ambiguous or off-topic responses, affecting the overall performance. In this study, we artificially tried many versions of the prompts on the web version of ChatGPT. For each strategy, we chose the one that performed best during our trial process. However, we cannot exhaust all possibilities, so we have summarized some valuable experiences. First, the prompt should include a clear task

Table 13. Performance of static few-shot prompt on ChatGPT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Cons	.411	.473	.427	.428	.462	.458	.409	.446	.382	.433
Acc	.274	.108	.253	.202	.263	.259	.255	.204	.245	.229
W-F1	.307	.151	.302	.210	.271	.313	.273	.220	.304	.261
M-F1	.217	.086	.219	.205	.211	.221	.224	.167	.186	.193
Rec-0	.231	.500	.329	.218	.077	.500	.286	.214	.308	.296
Prec-0	.089	.024	.138	.184	.028	.110	.181	.067	.044	.096
F1-0	.128	.046	.194	.200	.041	.180	.222	.102	.076	.132
Rec-2	.549	.833	.541	.648	.648	.648	.521	.333	.563	.587
Prec-2	.071	.005	.045	.093	.169	.077	.033	.012	.011	.058
F1-2	.126	.011	.083	.163	.268	.138	.063	.023	.022	.100

Table 14. Performance of dynamic few-shot prompt on ChatGPT.

Aspect	1	2	3	4	5	6	7	8	9	Ave
Cons	.832	.747	.716	.599	.666	.672	.600	.641	.647	.680
Acc	.515	.227	.579	.511	.523	.622	.534	.512	.512	.504
W-F1	.428	.261	.570	.508	.498	.632	.543	.544	.517	.500
M-F1	.303	.143	.407	.378	.359	.416	.364	.342	.321	.337
Rec-0	.090	.500	.500	.301	.154	.475	.222	.271	.269	.309
Prec-0	.144	.035	.262	.367	.124	.189	.289	.130	.103	.183
F1-0	.110	.066	.344	.331	.135	.270	.251	.175	.149	.203
Rec-2	.441	.333	.176	.078	.126	.402	.083	.292	.313	.249
Prec-2	.212	.019	.090	.154	.452	.163	.033	.064	.035	.136
F1-2	.285	.036	.119	.102	.195	.232	.047	.105	.063	.131

definition as far as possible. A clear task definition, including background, dimension explanation, etc., in the prompt can establish a strong foundation for effective communication and interaction with ChatGPT. It enhances the ability to understand, process, and respond to the task, resulting in more accurate, relevant, and contextually appropriate responses. Second, fine-grained logic of the task is related to good performance. We find that ChatGPT seems to be easier to understand small task requirements. When asked to conduct a one-dimensional analysis, even if there is a certain emotion present, ChatGPT always performs better. Finally, making it self-reflect is more effectively than directly providing standard response. If there is some labeled data, it may be better to let ChatGPT make a judgment first and then provide the correct label to stimulate its reflection than directly inputting them to ChatGPT. These experiences may be helpful for future research on automated labeling based on ChatGPT.

6.3 "Halo effect" of ChatGPT for student evaluation of teaching

The halo effect is a cognitive bias that influences the way people perceive and judge others based on a single positive trait, characteristic, or impression. This bias occurs when a person's overall judgment of someone is disproportionately influenced by a specific positive quality or feature, leading them to assume that the person possesses other positive qualities as well, regardless of whether those qualities are actually present. In this study, we find that ChatGPT can exhibit a form of the halo effect as well. ChatGPT will infer other dimensions based on the sentiment tendencies of some dimensions. For example, if the evaluation of teaching attitude is positive, then that of teaching effectiveness is often positive. There is also an extreme example. The written comments is "The teacher is very good in all aspects", and ChatGPT just outputs the sentiment tendency of each aspect as positive, while manual coding tends to believe that it lacks any fine-grained description. In our opinion, unlike human halo effects, which stem from unconscious cognitive biases Ambady & Rosenthal, 1993; Nisbett & Wilson, 1977), ChatGPT's pattern reflects its algorithmic design aimed at providing contextually relevant responses. This behaviour highlights the need for cautious interpretation and collaboration with LLMs. Drawing on social cognition theory (Fiske & Taylor, 2013), we may better understand how these patterns differ from human evaluative processes and form a good cooperation with AI technology.

6.4 Ethical considerations including algorithmic bias and privacy concerns

In leveraging advanced LLMs like ChatGPT for analyzing student evaluation of teaching, it is crucial to address significant ethical challenges, particularly algorithmic bias and privacy concerns. Algorithmic bias can manifest in several forms: data representation bias may skew outcomes, leading to unfair evaluations for certain demographic groups; confirmation bias might arise from ChatGPT's tendency to infer other dimensions based on initial sentiment tendencies, reinforcing stereotypes; and feedback loop bias could amplify existing biases over time without adequate oversight. Privacy issues are equally important, as handling sensitive student evaluations requires robust data security measures. Transparency and consent are essential. Students should be informed about AI processing of their feedback, and explicit consent must be obtained. Future research should focus on exploring privacy-preserving technologies, and establishing comprehensive ethical guidelines tailored to AI applications in educational evaluations.

7. Conclusions

In this study, we embarked on an exploration into the realm of student evaluation of teaching, leveraging ChatGPT. Through a rigorous analysis of a collected dataset based on open-ended question, we sought to uncover insights into the performance, usability, and potential challenges associated with employing ChatGPT for capturing fine-grained teacher performance from student evaluation of teaching. Our findings are summarized as follows:

- Q1: What is the performance of ChatGPT for capturing fine-grained teacher performance from student evaluation of teaching when no labeled data is provided?

Finding: We observed that ChatGPT demonstrates a capacity to analyze open-ended responses and extract valuable insights. However, it is not enough to replace manual labeling.

- Q2: How does the performance of ChatGPT, without labeled data, compare to that of supervised models for analyzing student evaluation of teaching?

Finding: The performance of ChatGPT surpasses that of the CNN model, which emphasizes its potential efficacy for analyzing student evaluation of teaching. Meanwhile, supervised models may be affected by class-imbalanced data, but ChatGPT can clearly avoid this situation.

- Q3: Can the performance of ChatGPT be enhanced by the inclusion of labeled samples, and if so, what strategies contribute to this improvement?

Finding: Introducing labeled samples by a dynamic strategy enhances ChatGPT's performance, pointing to opportunities for further fine-tuning and optimization. However, there is still a discernible performance gap remains between ChatGPT and BERT.

In general, utilizing ChatGPT as the automated labeling model reduces the dependence on extensive labeled data, offering a more efficient solution. From this perspective, the contribution of this research lies in its exploration of ChatGPT's potential as a tool for analyzing student evaluation of teaching, shedding light on its underlying benefits and limitations. As above-discussed, ChatGPT's cross-time consistency and performance are influenced by various factors, such as task complexity and prompt clarity. Understanding these dynamics is crucial for effective deployment. Additionally, it is essential to acknowledge the following study limitations:

- Our study employed a specific paradigm for capturing fine-grained teacher performance from student evaluation of teaching, where sentiment analysis was applied individually to each student's comment. This approach provided valuable insights into individual student feedback. However, it is important to acknowledge that an alternative paradigm, which involves aggregating all comments for the same teacher and then scoring, presents an intriguing avenue for future research. It is worth noting that such a holistic evaluation method would likely rely on a large amount of labeled data for effective evaluation.
- While our study has shed light on various factors influencing ChatGPT's performance, such as task complexity and prompt clarity, it remains essential to recognize that our analysis may not encompass the full spectrum of elements affecting ChatGPT's performance. Moreover, the relationships between these factors and ChatGPT's performance may not be entirely transparent. A more comprehensive, systematic study is needed to identify any additional variables that could influence ChatGPT's performance.
- An important concern that warrants further investigation is the "halo effect" associated with ChatGPT's performance. To ensure the reliability and fairness, it is crucial to explore strategies and techniques aimed at mitigating biases that may occur during the labeling process. Future research should focus on refining ChatGPT's capabilities to provide objective and unbiased labeling during analyzing student evaluation of teaching.

Future research in this domain may focus on addressing these limitations, refining the ChatGPT model, and extending the study to different educational contexts and levels. Also, the integration of educational psychology theories could provide a richer framework for analyzing ChatGPT's performance in student evaluations of teaching. For examples, according to Social Judgment Theory (Sherif & Hovland, 2023), initial positive impressions can disproportionately influence overall judgments. Whether ChatGPT can uncover these potential contents remains to be explored. Expectancy-Value Theory (Eccles et al., 1983) posits that students' expectations of success and the value they place on a subject significantly impact their evaluation of teaching. This may explain the emergence of the ChatGPT's halo effect. Moreover, the integration of ChatGPT into practical education evaluation applications presents a tantalizing prospect. More explorations are needed to seamlessly align its capabilities with the demands of education evaluation.

Acknowledgements

This work was supported by the Beijing Education Science Planning Project under Grant CCFA24122.

The data collection was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Beijing Normal University (protocol code: 202112300091, and date of approval: December 30, 2021). Data available on request from the authors.

Conflict of interest

The authors declare no competing interest.

References

Alnaqbi, N.M., & Fouda, W. (2023). Exploring the role of chatgpt and social media in enhancing student evaluation of teaching styles in higher education using neutrosophic sets. International Journal of Neutrosophic Science, 20(4), 181-190.

Alwaely, S.A., El-Zeiny, M.E., Alqudah, H., Alamarnih, E.F.M., Salman, O.K.I., Halim, M., Khasawneh, M.A.S. (2023). The impact of teacher evaluation on professional development and student achievement. Revista de Gestao Social e Ambiental, 17(7), e03484.

Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology, 64(3), 431.

Annan, S., Tratnack, S., Rubenstein, C., Sawin, E., & Hulton, L. (2013). An integrative review of student evaluations of teaching: Implications for evaluation of nursing faculty. Journal of Professional Nursing, 29(5), e10-e24.

Beran, T., & Rokosh, J. (2009). Instructors' perspectives on the utility of student ratings of instruction. Instructional Science, 37, 171-184.

Beran, T., Violato, C.,& Kline, D. (2007). What's the "use" of student ratings of instruction for administrators? one university's experience. Canadian Journal of Higher Education, 37, 27-43.

Brockx, B., Van Roy, K., & Mortelmans, D. (2012). The student as a commentator: Students' comments in student evaluations of teaching. Procedia - Social and Behavioral Sciences, 69, 1122-1133.

Chantamuang, O., Polpinij, J., Vorakitphan, V., & Luaphol, B. (2022). Sentence-level sentiment analysis for student feedback relevant to teaching process assessment. In Multi-disciplinary Trends in Artificial Intelligence (p. 156-168). Berlin Heidelberg: Springer-Verlag.

Chong, C., Sheikh, U.U., Samah, N., & Sha'ameri, A. (2020). Analysis on reflective writing using natural language processing and sentiment analysis. IOP Conference Series: Materials Science and Engineering, 884, 012069.

Clayson, D., & Haley, D. (2011). Are students telling us the truth? a critical look at the student evaluation of teaching. Marketing Education Review, 21, 101-112.

Denson, N., Loveday, T.,& Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction? Higher Education Research & Development, 29, 339-356.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding, J. Burstein, C. Doran,& T. Solorio (Eds.), 2019 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics.

Donnon, T., Delver, H., & Beran, T. (2010). Student and teaching characteristics related to ratings of instruction in medical sciences graduate programs. Medical Teacher, 32, 327-332.

Eccles, J. S., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, Values, and Academic Behaviors. In J. T. Spence (Ed.), Achievement and Achievement Motivation (pp. 75-146). San Francisco, CA: W. H. Freeman.

Emerson, R., & Records, K. (2007). Design and testing of classroom and clinical teaching evaluation tools for nursing education. International Journal of Nursing Education Scholarship, 4, Article12.

Fiske, S. T., & Taylor, S. E. (2013). Social cognition: From brains to culture (2nd ed.). Sage Publications.

Hammond, I., Taylor, J., & Mcmenamin, P. (2003). Value of a structured participant evaluation questionnaire in the development of a surgical education program. The Australian & New Zealand Journal of Obstetrics & Gynaecology, 43, 115-118.

Hazarika, D., Konwar, G., Deb, S., & Bora, D. J. (2020). Sentiment Analysis on Twitter by Using TextBlob for Natural Language Processing. Proceedings of the International Conference on Research in Management & Technovation (pp. 63-67).

Hodges, L., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. Innovative Higher Education, 31, 279-286.

Hommel, B. E. (2023). Expanding the methodological toolbox: Machine-based item desirability ratings as an alternative to human-based ratings. Personality and Individual Differences, 213, 112307.

Hoon, A., Oliver, E., Szpakowska, K., & Newton, P. (2015). Use of the "stop, start, continue" method is associated with the production of constructive qualitative feedback by students in higher education. Assessment & Evaluation in Higher Education, 40(5), 755-767.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the Eighth International Conference on Weblogs and Social Media. The AAAI Press.

Jin, Y., Cheng, K., Wang, X., & Cai, L. (2023, 07). A review of text sentiment analysis methods and applications. Frontiers in Business, Economics and Management, 10, 58-64.

Khanam, Z. (2023). Sentiment analysis of user reviews in an online learning environment: Analyzing the methods and future prospects. European Journal of Education and Pedagogy, 4, 209-217.

Kim, Y. (2014). Convolutional neural networks for sentence classification. A. Moschitti B. Pang, & W. Daelemans (Eds.), 2014 Conference on Empirical Methods in Natural Language Processing (pp. 1746–1751). ACL.

Kulik, J. (2001). Student ratings: Validity, utility, and controversy. New Directions for Institutional Research, 2001.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174.

Latif, E., Mai, G., Nyaaba, M., Wu, X., Liu, N., Lu, G., & Zhai, X. (2023). Artificial general intelligence (AGI) for education. arXiv:2304.12479.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9), 195:1-195:35.

Medhat, W., Hassan, A.H., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5, 1093-1113.

Nasser-Abu Alhija, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC , 27, 187-198.

Nasser-Abu Alhija, F., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from studentsv written comments? Studies In Educational Evaluation, 35, 37-44.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of personality and social psychology, 35(4), 250.

Nowak, J., Taspinar, A., & Scherer, R. (2017). LSTM recurrent neural networks for short text and sentiment classification. L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L.A. Zadeh, & J.M. Zurada (Eds.), 16th International Conference Artificial Intelligence and Soft Computing (Vol. 10246, pp. 553–562). Springer.

Okoye, K., Nunez-Daruich, S., de la O, J.F.E., Castano-Gonzalez, R., Escamilla, J., & Hosseini, S. (2023). A text mining and statistical approach for assessment of pedagogical impact of students' evaluation of teaching and learning outcome in education. IEEE Access, 11, 9577-9596.

Onwuegbuzie, A., Daniel, L., & Collins, K.M. (2009). A meta-validation model for assessing the score-validity of student teacher evaluations. Quality and Quantity, 43, 197-209.

British Council. (2015). OpenAI. (2022, November). Introducing ChatGPT.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Annual Conference on Neural Information Processing Systems.

Pineda, P., & Steinhardt, I. (2023). The debate on student evaluations of teaching: Global convergence confronts higher education traditions. Teaching in Higher Education, 28(4), 859–879.

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-based sentiment analysis of teachers' evaluation. Applied Computational Intelligence and Soft Computing, 2016, 1-12.

Ren, P., Yang, L., & Luo, F. (2023). Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis. Education and Information Technologies, 28, 797-814.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347.

Sedaghat, S. (2023). Early applications of chatgpt in medical practice, education and research. Clinical Medicine, 23, clinmed.2023-0078.

Sherif, M., & Hovland, C. I. (1961). Social judgment: Assimilation and contrast effects in communication and attitude change. Yale Univer. Press.

Sindhu, I., Daudpota, S., Badar, K., Bakhtyar, M., Baber, J., & Nurunnabi, M. (2019). Aspect based opinion mining on student's feedback for faculty teaching performance evaluation. IEEE Access, 7, 108729-108741.

Situmorang, D., Mini, R., Ifdil, I., Liza, L., Rusandi, M.A., Hayati, I., & Fitriani, A. (2023). The current existence of chatgpt in education: a double-edged sword? Journal of public health, Online.

Smith, C. (2008). Building effectiveness in teaching through targeted evaluation and response: Connecting evaluation to teaching improvement in higher education. Assessment & Evaluation in Higher Education, 33.

Srinivas, A., & Hanumanthappa, M. (2017). Viale modern approaches for sentiment analysis: A survey. International Journal of Advanced Research in Computer Science, 8, 115-120.

Stupans, I., Mcguren, T., & Babey, A.M. (2015). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. Innovative Higher Education, 41, 33-42.

Su, B., & Peng, J. (2023). Sentiment analysis of comment texts on online courses based on hierarchical attention mechanism. Applied Sciences, 13, 4204.

Tian, X., Jing, L., Luo, F., & Liu, F. (2022). Inference during reading: multi-label classification for text with continuous semantic units. Applied Intelligence, 52(6), 6292–6305.

Toni, M., & Sudin, M. (2024). Research and development (R&D) interactive media that is effective, efficient and fun for students. Asian Journal of Social and Humanities, 2(6), 1239–1252.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, T., Roziere, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, H. (2018). Multi-grain sentiment analysis of teaching reviews based on topic (Unpublished master's thesis). South China University of Technology.

Wang, P., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., & Sui, Z. (2024). Large language models are not fair evaluators. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp.

9440-9450). Association for Computational Linguistics.

Wright, S. (2011). Student evaluations of teaching: Combining the meta-analyses and demonstrating further evidence for effective use. Assessment & Evaluation in Higher Education, 37, 1-17.

Zeng, A., Xu, B., Wang, B., Zhang, C., Yin, D., Rojas, D., Feng, G., Zhao, H., Lai, H., Yu, H., Wang, H., Sun, J., Zhang, J., Cheng, J., Gui, J., Tang, J., Zhang, J., Li, J., Zhao, L., ... Wang, Z. (2024). ChatGLM: A family of large language models from GLM-130B to GLM-4 all tools. arXiv:2406.12793.

Zhao, Y., Yan, L., Sun, W., Xing, G., Wang, S., Meng, C., Cheng, Z., Ren, Z., & Yin, D. (2024). Improving the robustness of large language models via consistency alignment. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (pp. 8931-8941). ELRA and ICCL.

Zhu, I.C., Sun, M., Luo, J., Li, T., & Wang, M. (2023). How to harness the potential of chatgpt in education? Knowledge Management & E-Learning, 15, 133-152.