Towards Learning Analytics for Interdisciplinary Learning: Leveraging Knowledge-empowered Fine-Tuned GPT Models*

Tianlong Zhong[1], Gaoxia Zhu[2,*], Swee Chiat Low[3], and Siyuan Liu[3]

[1] *Energy Research Institute @ NTU, Graduate College, Nanyang Technological University, 50 Nanyang Ave, Singapore*
[2] *National Institute of Education, Nanyang Technological University, 1 Nanyang Walk, Singapore*
[3] *College of Computing & Data Science, Nanyang Technological University, 50 Nanyang Ave, Singapore*

Abstract

GPT models' ability to automatically score students' writing makes them promising to assess students' interdisciplinary learning quality, a significant but unaddressed gap. While standard GPT models have challenges in understanding contextual knowledge, previous research suggests that knowledge-empowered fine-tuned (KEFT) GPT models can overcome the limitations. This study examined 1) whether KEFT GPT models can accurately label interdisciplinary learning quality based on learning process and outcome data, and 2) how to implement these models within a learning analytics (LA) platform, including three major steps. First, to establish a ground truth dataset, two pairs of researchers independently coded and discussed the interdisciplinary learning quality of 400 online posts and 190 sections from 16 essays based on an interdisciplinary learning quality codebook. Second, we employed KEFT GPT models to evaluate interdisciplinary learning quality. Results indicated that the models achieved accuracy comparable to human researchers. Third, the models were integrated into an LA platform, TopicWise, which automates evaluation and provides tailored feedback. This study showcased the feasibility of applying KEFT GPT models in LA to analyse student learning processes and outcomes. Next, we will conduct user studies to examine TopicWise's impact on students' interdisciplinary learning and identify areas for improvement.

Keywords

GPT, Prompt engineering, Fine-tuning, Interdisciplinary learning, Learning analytics

1. Introduction

Interdisciplinary learning combines perspectives, methods, and strategies from various disciplines to address complex issues that cannot be fully understood within a single field [1], [2], [3]. This approach can engage students with real-world challenges, foster critical thinking, creativity, and critical problem-solving skills, and enhance their career readiness [4], [5]. A significant challenge in this domain is assessing the quality of interdisciplinary learning based on both the learning process and outcome data, as itorigi often requires posthoc labour-intensive qualitative analysis of textual data from multiple perspectives, such as diversity, cognitive advancement, disciplinary grounding, and integration [6], [7]. This challenge limits the possibility of effectively providing students with timely feedback.

ChatGPT, a chatbot powered by foundation large language models (LLMs) like GPT-3.5 and GPT-4o, developed by OpenAI [8], has shown promise in addressing the issue of effectively analysing student text and providing feedback. For instance, Lee et al. [9] applied chain-of-thought in automatic essay scoring with accuracy above 60%. Latif and Zhai [10]used fine-tuned GPT models for auto-scoring in science education and achieved an average accuracy of 83.8%. [11] utilised GPT-3 and

Second International Workshop on Generative AI for Learning Analytics, 2025.
TIANLONG001@e.ntu.edu.sg (T. Zhong); gaoxia.zhu@nie.edu.sg (G. Zhu); c210139@e.ntu.edu.sg (S. C. Low); syliu@ntu.edu.sg (S. Liu)
0000-0002-9467-2881 (T. Zhong); 0000-0003-4589-0775 (G. Zhu); 0009-0002-6231-6433 (S. C. Low); 0000-0002-5367-4709 (S. Liu)

GPT-4 to provide feedback to student essays and found the GPT models can provide more readable and consistent feedback than human teachers in data science courses. These studies show promising results in applying LLMs to automatically evaluate students' learning processes and outcomes and provide feedback, an important research topic of LA [12], [13].

LA is a research area that focuses on gathering, analysing, and reporting data about learners and their environments to gain insights and improve both the learning experience and the conditions that support it [14], [15]. However, developing effective LA to provide tailored feedback to users requires the backend models to “acquire” task-specific knowledge, which standard GPT models lack. Zhong et al. [16] adopted "knowledge-empowered approaches" to integrate domain knowledge and codebook rules into prompts to enhance LLM performance. They found that such approaches could enhance the GPT-3.5 model's performance in evaluating students' interdisciplinary learning quality on short posts (learning process data), but the effects of knowledge-empowered GPT models in essay (learning outcome data) evaluation remain unclear. Furthermore, even though the importance of interdisciplinary learning is well recognised, there is a lack of interdisciplinary LA that can provide auto-assessment and real-time feedback. To address the research gaps, this ongoing work takes initiative steps to develop interdisciplinary LA and explore the following two questions (RQs):

RQ1: Can knowledge-empowered approaches increase GPT models' accuracy in analysing interdisciplinary learning quality?

RQ2: How can a prototype LA platform be designed to leverage GPT models for providing automated feedback on students' interdisciplinary learning quality?

2. Literature review

2.1. Interdisciplinary Learning and LA

Interdisciplinary learning refers to the process of incorporating knowledge and perspectives from multiple disciplines to solve problems or explain phenomena beyond the boundary of a singular discipline [7], [17]. However, this domain faces challenges in analysing and reporting students' interdisciplinary learning quality, which calls for more rigorous and robust methods [6]. Qualitative analysis, such as essay evaluation, is commonly used for interdisciplinary learning assessments. For instance, Boix-Mansilla et al. [19] introduced the rubric for interdisciplinary writing, encompassing four key dimensions: purposefulness, disciplinary grounding, integration, and critical awareness. Kidron and Kali [17] expanded the integration dimension into the following sub-dimensions: integrative lens, idea connection, disciplinary analysis through an integrative lens, and synthesis, and use the updated rubric to assess students' essays. However, these assessments are post-hoc and occur after data collection is done. There remains a gap in analysing learning process data in real time and providing just-in-time feedback to guide and enhance students' interdisciplinary learning.

LA can effectively analyse real-time learning processes by collecting process data from digital tools like learning management systems (LMS), automatically analysing data with algorithms, and providing visualised dashboards and personalised feedback [20]. Various applications of LA have been utilised in interdisciplinary learning. For instance, Lee et al. [18] used machine learning methods to analyse STEM learning behaviours, categorising them as passive, active, constructive, or interactive. Iku-Silan et al. [21] created a chatbot powered by natural language processing (NLP) technology to support interdisciplinary learning. This chatbot provided students with personalised advice and resources sourced from an interdisciplinary knowledge database. Tang et al. [22] designed a platform aimed at enhancing K-12 STEM education by integrating machine learning into scientific lesson plans. For instance, their platform used machine learning to analyse data related to heart disease risk factors, enabling students to engage in scientific discovery more interactively. Yet, LA tools that can analyse the interdisciplinary learning quality of students’ generated data are lacking.

2.2. Fine-tuning and Knowledge-empowered Approaches in GPT

A few techniques for enhancing GPT performances have been explored in the literature. Prompt engineering is an important strategy for improving a model's performance by designing and optimising model input [23]. Studies have shown that designing prompts can enhance the performance of GPT for various tasks, including classification and reasoning [23], [24]. Moreover, for some complex tasks, chain-of-thought (CoT) prompting is regarded as a useful technique of prompt engineering [25]. CoT induces the model to solve a problem step-by-step, thus mimicking a chain of thought and improving the model's reasoning ability [26].

However, GPT's expertise in specific domains may be limited, which can result in nonsensical or inappropriate responses to specific prompts [27]. Fine-tuning is a technique that can help mitigate this limitation and improve GPT performance on specific tasks. Fine-tuning refers to the additional training of pre-trained models to customise them for specific tasks or datasets [28]. One benefit of this approach is the ability to tailor models to enhance their performance in specific tasks, which requires only 50 to 100 examples for training [29]. The fine-tuned GPT models have been shown to be effective in several studies. Chae and Davidson [30] suggested that fine-tuning is an optimal solution for researchers due to its relatively high accuracy and low cost.

However, fine-tuning methods rely heavily on the pre-training data [29], which may limit their ability to handle tasks requiring knowledge not included in their initial training set. The knowledge-empowered method, which incorporates external knowledge into the model, may further improve GPT performance on specific tasks by expanding the model's understanding beyond the pre-training dataset [31]. The basic premise of this technique is that by integrating additional information, models can enhance their comprehension of content and generate better output [32]. Hu et al. [33] combined domain knowledge (geo-knowledge) with GPT and showed that external knowledge is indispensable for guiding the behaviour of GPT models. Similarly, Yang et al. [34] conducted a study to use external knowledge bases to enhance pre-trained language models for machine reading comprehension. They found that incorporating structured knowledge from knowledge bases significantly improved models' accuracy on benchmarks like ReCoRD and SQuAD1.1. Overall, these studies have shown that by integrating external knowledge into models, the performance of models in specific tasks significantly improved [16].

The promising results highlight the potential of using knowledge-empowered strategies to analyse the interdisciplinary learning quality of students' work. In a recent study, we employed knowledge-empowered approaches—such as dictionary-based knowledge to address terminology that GPT models struggle to understand, and rule-based knowledge to capture implicit mechanisms outlined in codebooks—to fine-tune GPT models. The findings revealed that these strategies significantly enhanced the performance of GPT-3.5 in evaluating interdisciplinary learning process data (e.g., online posts). However, this approach has yet to be applied to GPT-4 models or to learning outcome data such as final essays. Building on these strategies, this study seeks to develop an interdisciplinary LA platform capable of processing real-time data. The platform will provide students with timely feedback on the quality of their interdisciplinary learning and offer actionable suggestions to foster improvement.

3. The prototype learning analytics platform

The following sections will present a prototype interdisciplinary LA platform, TopicWise, by giving an overview of the platform, detailing how Knowledge-empowered fine-tuned GPT models have been trained and their performance, and showing the user interface design.

3.1. Overview of the LA platform

This platform is designed to evaluate students' interdisciplinary learning and provide actionable feedback for improvement. As Figure 1 shows, students can upload files (e.g., essays) or short texts (e.g., posts, discussions) through the user interface. After that, the text will be delivered to

knowledge-empowered fine-tuned models for processing to generate relevant feedback, including comments on the text and suggestions for improvement from four dimensions of interdisciplinary learning quality: diversity, cognitive advancement, disciplinary grounding, and integration [7]. The feedback will be shown on the user interface, and the text data and feedback will be saved in the Mongo database. Students can access their feedback in real-time and review them anytime, which can potentially help them better understand the strengths and weaknesses of their writing and help them improve.

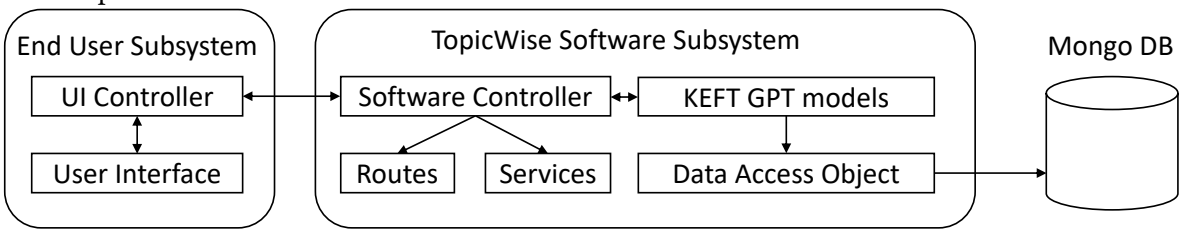

Figure 1: Architecture of TopicWise

3.2. Knowledge-empowered fine-tuned GPT models

3.2.1. The Dataset

To prepare ground truth data for model training and testing, we manually analysed 400 posts collected from the Miro platform and 16 essays. The existing literature on manual content analysis of essays indicates that smaller units, rather than entire texts, are more appropriate for studying the general standard of essays [35]. Therefore, to retain the consistency and integrity of ideas, we analysed students' essays at their most granular level, focusing on the smallest sections, which typically consist of several paragraphs and represent the deepest layer of content organisation. This approach divided the 16 essays into 190 data points.

Thereafter, two human coders independently labelled students' posts. Another two coders labelled essays, all using the codebook of interdisciplinary learning quality, which consists of diversity (the number of disciplines represented in the text), cognitive advancement (the depth and clarity of the articulated viewpoints), disciplinary grounding (the extent to which the text applies disciplinary knowledge), and integration (the degree to which perspectives from multiple disciplines are synthesised). We used Cohen's Kappa score, as shown in Table 1, to evaluate the inter-rater reliability between human raters on each dimension of interdisciplinary learning quality. Human coders subsequently discussed and settled their differences, reaching an agreement on each item, which was regarded as the ground truth of interdisciplinary learning quality.

For both the post and the essay dataset, 80% of the data were randomly selected as training data, which were applied to fine-tune the GPT-3.5 and GPT-4o-mini models, while the remaining 20% of the data were tested, which is explained in detail in Section 3.2.2. The frequency of each code for each dimension in the training dataset and the testing dataset is displayed in Table 2. The dimensions refer to the elements of the interdisciplinary learning quality [7], each of these dimensions is further divided into three levels.

Table 1
Inter-rater reliability (Cohen's Kappa) between human raters on notes and essays

Data	Diversity	Cognitive advancement	Disciplinary grounding	Integration	Overall
Post	0.83	0.82	0.83	0.73	0.83
Essay	0.84	0.71	0.67	0.57	0.72

Table 2
The frequency of each code in each dimension

	Training data	Testing data

	Dimension	Level 0	Level 1	Level 2	Level 0	Level 1	Level 2
Post	Diversity	54	89	59	10	31	10
	Cognitive advancement	79	63	60	27	15	9
	Disciplinary grounding	67	134	1	10	39	2
	Integration	163	32	7	41	8	2
Essay	Diversity	0	35	116	0	8	31
	Cognitive advancement	5	97	49	3	15	21
	Disciplinary grounding	5	94	52	1	21	17
	Integration	49	74	28	9	16	14

3.2.2. Knowledge-empowered Fine-tuning Strategy

We systematically crafted the prompts for GPT models following the stages in Figure 2. We first adopted the interdisciplinary learning codebook, drawing upon educational theories. Following that, we created a template to translate the natural language of the codebook into a structure that GPT could process. For instance, we used a standardised format like a conditional statement (i.e., if-else) to represent the rules in the codebook. Ultimately, each tailored prompt was generated using the template and contained the following components (see Table 4): (1) A system message defining GPT's persona; (2) A tailored task instruction outlining the task and its needs; and (3) A rule derived from the codebook, offering guidelines and examples relevant to different levels of a specific dimension.

We also utilised CoT to effectively instruct GPT models with step-by-step tasks. Guided by CoT, there are three main steps (see Table 3) in the prompts: Firstly, task clarification provides essential details such as the requirements and desired output. Secondly, in the task breakdown, the tasks are divided into smaller, manageable parts. Lastly, the logical sequence instruction guides GPT in understanding the relations and mechanisms among these breakdown tasks. Through these three steps, we created a structured framework designed to address complex tasks with CoT methods.

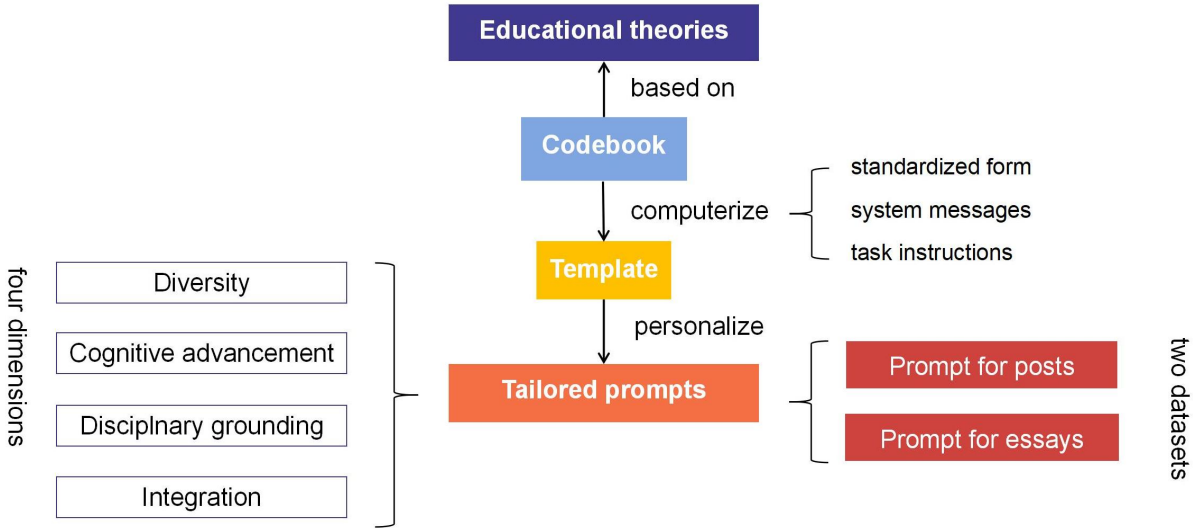

Figure 2: Stages to build a tailored prompt

Table 3
Elements of prompts

Prompt	Element	Example

Tailored Prompt	System messages	"You are an encyclopaedia that can precisely evaluate the disciplines reflected in the following notes."
	Rules sourced from the Codebook	"Please see all the information as a single paragraph and evaluate the cognitive advancement level of students essays. Return only numerical values 0, 1, and 2."
	Tailored task instructions	"Return 2 if the content provides detailed reasoning and specific examples to demonstrate a deep understanding of the topic."
Chain-of-Thought Prompting	Task clarification	"Please see all the information as a single paragraph and answer the below two questions about the cognitive advancement of essays. Please return yes or no."
	Task breakdown	"Question 1: Does the paragraph have basic explanations or causalities or examples or mechanisms or elaborations of phenomena."; "Question 2: Does the paragraph provide detailed reasoning and specific examples to demonstrate a deep understanding of the topic?"
	Logical sequence instructions	"Question 2 is an extended one based on Question 1."

In this study, two types of knowledge, dictionary-based and rule-based knowledge, were integrated into prompts to enhance the models' performance.

Dictionary-based knowledge includes specific words with predefined categories [35]. For example, in analysing the diversity dimension, a discipline dictionary was provided. The dictionary included eleven disciplines: 'Arts and humanities'; 'Business and economics'; 'Clinical, pre-clinical and health'; 'Computer science'; 'Education'; 'Engineering and technology'; 'Law'; 'Life sciences'; 'Physical sciences'; 'Psychology'; and 'Social sciences'. If a student mentions terms like "copyright," GPT might struggle to classify them correctly. To address this, prompts were structured as: "If students mention terms such as WORD (copyright), it reflects content related to the LABEL 'Law.'" This method was applied across other dimensions of interdisciplinary learning quality to improve accuracy. This study applied Dictionary-based knowledge on Diversity and Cognitive advancement dimensions because GPT models do not have context knowledge about these two dimensions and thereby need specific examples. The dictionary-based knowledge can also be applied to other circumstances when LLMs cannot understand specific instances.

Rule-based knowledge, on the other hand, uses task-specific logic derived from relationships outlined in codebooks. For instance, if no disciplines are mentioned in a student's text (Diversity = 0), disciplinary grounding should also be 0, as no disciplinary knowledge is present. Similarly, if fewer than two disciplines are mentioned (Diversity < 2), Integration is likely 0, as interdisciplinary synthesis is absent. These rules were embedded into prompts using structures like: "IF DIMENSION A (Diversity) is 0, THEN DIMENSION B (Disciplinary Grounding) is likely to be 0." By encoding such logic, rule-based knowledge ensures the model considers the interplay between dimensions, enhancing its ability to perform deductive coding effectively. This study applied rule-based knowledge on Disciplinary grounding and Integration dimensions because these two dimensions rely on the outcomes of Diversity. The rule-based knowledge can also be applied to other circumstances when LLMs cannot understand the implicit rules in the task, especially in tasks with interdependence.

3.2.3. Model performance

Experiments were operated on GPT-3.5 and GPT-4o-mini models. To answer RQ1, we tested the models in four modes: prompts (directly use prompts), fine-tuning (apply fine-tuning), knowledge-empowered prompts (embed knowledge in prompts) and knowledge-empowered fine-tuning (apply knowledge-empowered prompts and fine-tuning). Cohen's Kappa scores, presented in Tables 4 and 5, were used to measure the agreement between GPT-generated labels and human-coded ground truth for both posts and essays, with human inter-rater reliability serving as the benchmark.

The results indicated that knowledge-empowered approaches enhance both prompt-based and fine-tuning methods. Knowledge-empowered methods demonstrated clear improvements for posts (learning process data), validating their effectiveness. For essays (learning outcome data), these methods enhanced performance on Diversity and Disciplinary grounding for knowledge-empowered prompts and augmented accuracy on most dimensions, except for Diversity (0.91 vs. 0.78) in fine-tuned models.

Overall, integrating knowledge-empowered strategies with fine-tuning significantly increased GPT models' agreement with human coders, achieving or surpassing expert-level proficiency in analysing interdisciplinary learning quality

Table 4

Cohen's Kappa scores in student posts

		Diversity	Cognitive advancement	Disciplinary grounding	Integration
GPT-3.5	Prompts	0.39	0.42	0.13	0.15
	Knowledge-empowered prompts	0.48	0.42	0.14	0.41
	Fine-tuning	0.55	0.84	0.46	0.50
	Knowledge-empowered fine-tuning	0.72	0.85	0.54	0.66
GPT-4o-mini	Prompts	0.39	0.42	0.13	0.18
	Knowledge-empowered prompts	0.60	0.42	0.14	0.41
	Fine-tuning	0.71	0.90	0.51	0.64
	Knowledge-empowered fine-tuning	0.87	0.90	0.77	0.78
Human coders		0.83	0.82	0.83	0.73

Table 5

Cohen's Kappa scores in student essays

		Diversity	Cognitive advancement	Disciplinary grounding	Integration
GPT-3.5	Prompts	0.12	0.19	0.25	0.07
	Knowledge-empowered prompts	0.13	0.33	0.27	0.07
	Fine-tuning	0.75	0.68	0.51	0.40
	Knowledge-empowered fine-tuning	0.84	0.79	0.62	0.51
	Prompts	0.03	0.52	0.38	0.29

GPT-4o-mini	Knowledge-empowered prompts	0.19	0.25	0.44	0.15
	Fine-tuning	0.91	0.73	0.42	0.65
	Knowledge-empowered fine-tuning	0.78	0.79	0.56	0.68
Human coders		0.84	0.71	0.67	0.57

3.3. User interface

After getting the models, we implemented them into an LA platform we are developing: TopicWise (https://a-ori-topic-wise.vercel.app/).

Figure 3 presents a screenshot of TopicWise for providing scores and feedback for essays. In the "Scores Comparison", students can view their scores across the four interdisciplinary learning dimensions—diversity, cognitive advancement, disciplinary grounding, and integration—and compare them with the average scores from the database. This comparative feature intends to help students understand their performance in the context of their peers and highlight areas for improvement. The "Paragraph Annotations" section provides a more granular analysis, offering scores for each specific paragraph in the essay. Additionally, the platform explains the reasoning behind each score and provides targeted feedback for improvement. This detailed breakdown identifies strengths and weaknesses in students' writing, aiming to foster a deeper understanding of interdisciplinary learning principles and guiding their revisions. The platform also supports real-time feedback for online discussion posts, as shown in Figure 4. When students upload their posts to TopicWise, the system quickly analyses the content, assigns scores for interdisciplinary dimensions, and delivers immediate feedback. This instant evaluation enables students to refine their posts during discussions, promoting more effective engagement with interdisciplinary concepts and improving learning outcomes over time.

TopicWise's ability to deliver timely feedback gives it the potential to support students in interdisciplinary learning through reflective essay writing and interactive online discussion. It has the potential to provide students with accessible, actionable insights into their performance, empowering them to make meaningful progress in their interdisciplinary learning.

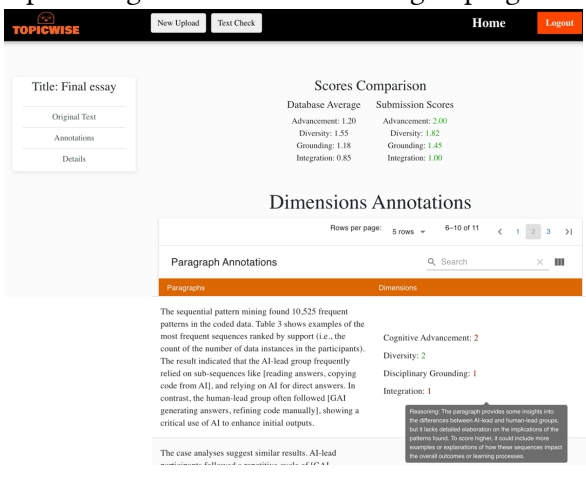

Figure 3: A screenshot of TopicWise for essay scoring and feedback

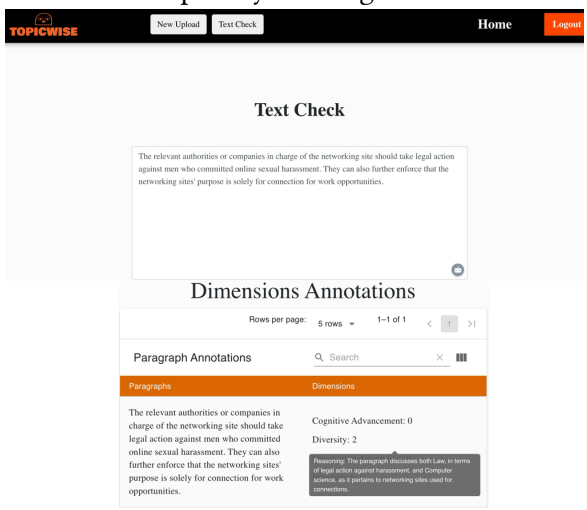

Figure 4: A screenshot of TopicWise for post scoring and feedback

4. Discussion and Conclusion

The study found that knowledge-empowered approaches can enhance the performance of GPT-3.5 and GPT-4o-mini models, achieving accuracy comparable to human experts. The trained models were subsequently integrated into a prototype LA platform and were able to offer automated scoring and feedback on students' interdisciplinary learning quality. By leveraging these advancements, students can gain insights into their performance in real-time. The use of knowledge-based fine-tuning highlights its potential as a robust method for enhancing LA, making it a promising approach in educational contexts.

Knowledge-empowered approaches amplified both prompt-based and fine-tuned model performance. By incorporating small-scale domain-specific dictionaries and rule-based logic, the study extended findings on external knowledge integration in prompt engineering in automated essay analysis [35]. This method also addresses the issue of relying on large knowledge graphs [36] or extensive knowledge bases [34], emphasising the efficacy of tailored knowledge enhancements during fine-tuning. Interestingly, this study found that knowledge-empowered fine-tuning was more effective for evaluating posts than essays. Posts typically require less complex reasoning than essays; reasoning demands deeper critical thinking and subject-specific expertise, making them more challenging for GPT models to analyse [37]. Although CoT prompting was used to assist with reasoning tasks, essays' intricate structure and nuanced content posed greater difficulties for the models. This highlights a gap in the capabilities of GPT models when handling more cognitively demanding tasks, suggesting the need for further refinement to support evaluations requiring advanced reasoning skills.

The study also introduces TopicWise, a prototype interdisciplinary LA platform that automates the evaluation process and provides tailored feedback to students. This platform not only aims to provide automated scoring but also aims to deliver tailored feedback to students, helping them understand and improve their performance. By enabling dynamic feedback and continuous monitoring, the platform has the potential to enhance students' interdisciplinary learning practices. User studies need to be conducted next to evaluate and refine the tool.

However, this study acknowledges several limitations. First, the models are trained based on a relatively small dataset of online posts and essays from a specific cohort of undergraduate students. Whether the methods can be generalised to other datasets needs further research, raising questions about generalizability. Further research is needed to determine whether these methods are applicable to broader and more diverse datasets. Second, the study only tested GPT-3.5 and GPT-4o-mini, leaving unexplored the potential of other language models, such as LLaMA and Gemini, which could offer different perspectives or improved capabilities. Third, the LA platform has not been tested by users like instructors and students. We plan to conduct user studies after further refining the tool.

Despite these limitations, the study highlights the potential of combining fine-tuning with knowledge-empowered strategies for evaluating both learning process data (e.g., online posts) and outcome data (e.g., essays). The integration of these trained models into an LA platform further enhances the approach by providing immediate, data-driven feedback. The platform has the potential to support educators in fostering interdisciplinary skills while optimising the assessment process. Future work will focus on expanding the dataset, testing additional models, and conducting user studies to ensure that the platform meets the needs of educators and students.

Acknowledgements

The research is conducted with the support of the Energy Research Institute @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University, Singapore.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools in essay writing.

References

[1] V. Boix-Mansilla, "Learning to Synthesize: The Development of Interdisciplinary Understanding," in *The Oxford Handbook of Interdisciplinarity*, R. Frodeman, J. T. Klein, C. Mitcham, and J. B. Holbrook, Eds., Oxford University Press, 2010, pp. 288–306.

[2] R. W. Bybee, *The case for STEM education: challenges and opportunities*. Arlington, VA: National Science Teachers Association, 2013.

[3] L. Ivanitskaya, D. Clark, G. Montgomery, and R. Primeau, "Interdisciplinary Learning: Process and Outcomes," *Innovative Higher Education*, vol. 27, no. 2, pp. 95–111, Dec. 2002, doi: 10.1023/A:1021105309984.

[4] M. Brassler and J. Dettmers, "How to Enhance Interdisciplinary Competence—Interdisciplinary Problem-Based Learning versus Interdisciplinary Project-Based Learning," *Interdisciplinary Journal of Problem-Based Learning*, vol. 11, no. 2, Art. no. 2, Jul. 2017, doi: 10.7771/1541-5015.1686.

[5] M. E. Madden *et al.*, "Rethinking STEM Education: An Interdisciplinary STEAM Curriculum," *Procedia Computer Science*, vol. 20, pp. 541–546, Jan. 2013, doi: 10.1016/j.procs.2013.09.316.

[6] I. E. F. Gvili, M. Weissburg, J. Yen, M. E. Helms, and C. Tovey, "Development of scoring rubric for evaluating integrated understanding in an undergraduate biologically-inspired design course," *International Journal of Engineering Education*, 2016, Accessed: Jul. 25, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Development-of-scoring-rubric-for-evaluating-in-an-Gvili-Weissburg/53fb00b8bf56209192de2da3528aa31adafc5f66

[7] T. Zhong, G. Zhu, C. Hou, Y. Wang, and X. Fan, "The influences of ChatGPT on undergraduate students' demonstrated and perceived interdisciplinary learning," *Educ Inf Technol*, May 2024, doi: 10.1007/s10639-024-12787-9.

[8] OpenAI, "GPT-4 Technical Report," Mar. 27, 2023, *arXiv*: arXiv:2303.08774. Accessed: Apr. 03, 2023. [Online]. Available: http://arxiv.org/abs/2303.08774

[9] G.-G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, "Applying large language models and chain-of-thought for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100213, Jun. 2024, doi: 10.1016/j.caeai.2024.100213.

[10] E. Latif and X. Zhai, "Fine-tuning ChatGPT for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100210, Jun. 2024, doi: 10.1016/j.caeai.2024.100210.

[11] W. Dai *et al.*, "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study," *Computers and Education: Artificial Intelligence*, vol. 7, p. 100299, Dec. 2024, doi: 10.1016/j.caeai.2024.100299.

[12] K. Alalawi, R. Athauda, R. Chiong, and I. Renner, "Evaluating the student performance prediction and action framework through a learning analytics intervention study," *Educ Inf Technol*, Aug. 2024, doi: 10.1007/s10639-024-12923-5.

[13] F. Ouyang, M. Wu, L. Zheng, L. Zhang, and P. Jiao, "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course," *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, p. 4, Jan. 2023, doi: 10.1186/s41239-022-00372-4.

[14] C. Lang *et al.*, Eds., *Handbook of Learning Analytics*, First. Society for Learning Analytics Research (SoLAR), 2017. doi: 10.18608/hla17.

[15] G. Siemens, "Learning Analytics: The Emergence of a Discipline," *American Behavioral Scientist*, vol. 57, no. 10, pp. 1380–1400, Oct. 2013, doi: 10.1177/0002764213498851.

[16] T. Zhong, C. Cai, G. Zhu, and M. Ma, "Enhancing the Analysis of Interdisciplinary Learning Quality with GPT Models: Fine-Tuning and Knowledge-Empowered Approaches," in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky*, A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt, Eds., Cham: Springer Nature Switzerland, 2024, pp. 157–165. doi: 10.1007/978-3-031-64312-5_19.

[17] A. Kidron and Y. Kali, "Promoting interdisciplinary understanding in asynchronous online higher education courses: a learning communities approach," *Instr Sci*, pp. 1–31, Jun. 2023, doi: 10.1007/s11251-023-09635-7.

[18] H.-Y. Lee, Y.-P. Cheng, W.-S. Wang, C.-J. Lin, and Y.-M. Huang, "Exploring the Learning Process and Effectiveness of STEM Education via Learning Behavior Analysis and the

Interactive-Constructive- Active-Passive Framework," *Journal of Educational Computing Research*, vol. 61, no. 5, pp. 951–976, Sep. 2023, doi: 10.1177/07356331221136888.
[19] V. Boix-Mansilla, E. D. Duraisingh, C. R. Wolfe, and C. Haynes, "Targeted Assessment Rubric: An Empirically Grounded Rubric for Interdisciplinary Writing," *The Journal of Higher Education*, vol. 80, no. 3, pp. 334–353, May 2009, doi: 10.1080/00221546.2009.11779016.
[20] D. Gašević, V. Kovanović, and S. Joksimović, "Piecing the learning analytics puzzle: a consolidated model of a field of research and practice," *Learning: Research and Practice*, vol. 3, no. 1, pp. 63–78, Jan. 2017, doi: 10.1080/23735082.2017.1286142.
[21] A. Iku-Silan, G.-J. Hwang, and C.-H. Chen, "Decision-guided chatbots and cognitive styles in interdisciplinary learning," *Computers & Education*, vol. 201, p. 104812, Aug. 2023, doi: 10.1016/j.compedu.2023.104812.
[22] J. Tang, X. Zhou, X. Wan, M. Daley, and Z. Bai, "ML4STEM Professional Development Program: Enriching K-12 STEM Teaching with Machine Learning," *Int J Artif Intell Educ*, vol. 33, no. 1, pp. 185–224, Mar. 2023, doi: 10.1007/s40593-022-00292-4.
[23] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," 2021, *arXiv*. doi: 10.48550/ARXIV.2102.07350.
[24] T. Sorensen *et al.*, "An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 819–862. doi: 10.18653/v1/2022.acl-long.60.
[25] X. Wang *et al.*, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," 2022, *arXiv*. doi: 10.48550/ARXIV.2203.11171.
[26] J. Wei *et al.*, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," 2022, *arXiv*. doi: 10.48550/ARXIV.2201.11903.
[27] L. Hamilton, D. Elliott, A. Quick, S. Smith, and V. Choplin, "Exploring the Use of AI in Qualitative Analysis: A Comparative Study of Guaranteed Income Data," *International Journal of Qualitative Methods*, vol. 22, p. 16094069231201504, Jan. 2023, doi: 10.1177/16094069231201504.
[28] K. W. Church, Z. Chen, and Y. Ma, "Emerging trends: A gentle introduction to fine-tuning," *Nat. Lang. Eng.*, vol. 27, no. 6, pp. 763–778, Nov. 2021, doi: 10.1017/S1351324921000322.
[29] OpenAI, "Fine-tuning." Accessed: Jul. 02, 2024. [Online]. Available: https://platform.openai.com
[30] Y. Chae and T. Davidson, "Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning," Aug. 24, 2023. doi: 10.31235/osf.io/sthwk.
[31] K. Zupanc and Z. Bosnić, "Automated essay evaluation with semantic analysis," *Knowledge-Based Systems*, vol. 120, pp. 118–132, Mar. 2017, doi: 10.1016/j.knosys.2017.01.006.
[32] F.-Y. Min, M. Yang, and Z.-C. Wang, "Knowledge-based method for the validation of complex simulation models," *Simulation Modelling Practice and Theory*, vol. 18, no. 5, pp. 500–515, May 2010, doi: 10.1016/j.simpat.2009.12.006.
[33] Y. Hu *et al.*, "Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages," *International Journal of Geographical Information Science*, vol. 37, no. 11, pp. 2289–2318, Nov. 2023, doi: 10.1080/13658816.2023.2266495.
[34] A. Yang *et al.*, "Enhancing Pre-Trained Language Representations with Rich Knowledge for Machine Reading Comprehension," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 2346–2357. doi: 10.18653/v1/P19-1226.
[35] T. D. Ullmann, "Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches," *Int J Artif Intell Educ*, vol. 29, no. 2, pp. 217–257, May 2019, doi: 10.1007/s40593-019-00174-2.
[36] W. Liu *et al.*, "K-BERT: Enabling Language Representation with Knowledge Graph," *AAAI*, vol. 34, no. 03, pp. 2901–2908, Apr. 2020, doi: 10.1609/aaai.v34i03.5681.
[37] Ž. Bašić, A. Banovac, I. Kružić, and I. Jerković, "ChatGPT-3.5 as writing assistance in students' essays," *Humanit Soc Sci Commun*, vol. 10, no. 1, p. 750, Oct. 2023, doi: 10.1057/s41599-023-02269-7.