# Enhancing the Analysis of Interdisciplinary Learning Quality with GPT Models: Fine-Tuning and Knowledge-Empowered Approaches

**Conference Paper** · July 2024

DOI: 10.1007/978-3-031-64312-5_19

CITATIONS
0

READS
139

**4 authors**, including:

Tianlong Zhong
Nanyang Technological University
**13** PUBLICATIONS **48** CITATIONS
SEE PROFILE

Gaoxia Zhu
Nanyang Technological University
**77** PUBLICATIONS **813** CITATIONS
SEE PROFILE

Min Ma
Nanyang Technological University
**2** PUBLICATIONS **0** CITATIONS
SEE PROFILE

# Enhancing the Analysis of Interdisciplinary Learning Quality with GPT Models: Fine-Tuning and Knowledge-Empowered Approaches

Tianlong Zhong(✉), Chang Cai, Gaoxia Zhu, and Min Ma

Nanyang Technological University, 50 Nanyang Ave, Singapore 639798, Singapore
tianlong001@e.ntu.edu.sg

**Abstract.** Assessing the interdisciplinary learning quality of student learning processes is significant but complex. While some research has experimented with ChatGPT for qualitative analysis of text data through crafting prompts for tasks, the in-depth consideration of task-specific knowledge, like context and rules, is still limited. The study examined whether considering such knowledge can improve ChatGPT's labeling accuracy for interdisciplinary learning quality. The data for this research consists of 252 online posts collected during class discussions. This study utilized prompt engineering, fine-tuning, and knowledge-empowered approaches to evaluate student interdisciplinary learning and compare their accuracy. The results indicated that unmodified GPT-3.5 lacks the capability for analyzing interdisciplinary learning. Fine-tuning significantly improved the models, doubling the accuracy compared to using GPT-3.5 with prompts alone. Knowledge-empowered approaches enhanced both the prompt-based and fine-tuned models, surpassing the researchers' inter-rater reliability in assessing all dimensions of student posts. This study showcased the effectiveness of combining fine-tuning and knowledge-empowered approaches with advanced language models in assessing interdisciplinary learning, indicating the potential of applying this method for qualitative analysis in educational settings.

**Keywords:** GPT · Prompt engineering · Fine-tuning · Interdisciplinary learning · Qualitative coding

## 1 Introduction

Interdisciplinary learning is an educational approach that synthesizes perspectives, strategies, and methodologies from multiple disciplines to comprehensively address complex issues beyond the scope of a single field [1]. In interdisciplinary learning, students gain experience tackling complex real-world problems, thereby enhancing their problem-solving skills and employability [2]. One challenge in this research area is evaluating interdisciplinary learning quality based on the


A. M. Olney et al. (Eds.): AIED 2024 Workshops, CCIS 2151, pp. 157–165, 2024.
https://doi.org/10.1007/978-3-031-64312-5_19

learning process data, which involves much qualitative coding of text and artifacts [6]. Qualitative coding refers to the process of making sense of qualitative data such as online discussions, interviews, observations, and essays by tagging them with particular labels related to research aims [4]. Analyzing interdisciplinary learning quality is a qualitative coding task involving coding process data from multiple dimensions such as diversity, cognitive advancement, disciplinary grounding, and integration [15]. However, automated qualitative coding remains a thorny topic. While some studies employ machine learning methods to accomplish this, they still encounter challenges like the need for large datasets with ground truth and issues with low transferability [3].

ChatGPT (Chat Generative Pre-trained Transformer), a chatbot based on GPT foundation models (i.e., GPT-3.5 and GPT-4) developed by OpenAI [10], shows its potential to support qualitative coding. There are two types of coding: inductive coding (labels emerging from context) and deductive coding (labels adapted from predefined frameworks or codebooks) [5]. This study focuses on the latter one, which inherently functions as a classification task, with researchers categorizing text according to predefined criteria in codebooks. Large language models (LLMs) like GPT can classify text into predefined labels [8]. For instance, Xiao and colleagues [14] combined LLMs with an expert-drafted codebook to label children's curiosity-driven questions, which demonstrated fair to substantial agreements with expert-coded results. Qualitative analysis requires a deep understanding of task-specific knowledge (e.g., context and rules)—capabilities unmodified GPT models lack. However, few studies provide a clear definition of knowledge within the context of LLMs. In our research, we use "knowledge-empowered approaches" to describe the integration of external information, such as domain knowledge and codebook rules, into prompts to improve LLMs' performance. Furthermore, while strategies like prompt engineering and fine-tuning are recognized as beneficial for improving LLMs' performance [7,11], their application in the context of qualitative coding with GPT is still rare. This research seeks to experiment with whether incorporating prompt engineering, fine-tuning, and task-specific knowledge into GPT can help automatically analyze students' interdisciplinary learning quality.

We adopted the codebook and employed prompt engineering, fine-tuning, and knowledge-empowered approaches to compare the agreement between GPT models' outputs and ground truth (human labels). The study revealed that generic LLMs like GPT-3.5, trained on broad, non-specialized texts, are not immediately equipped for assessing interdisciplinary learning. For effective evaluation, a combination of fine-tuning and knowledge-empowered approaches is essential. Fine-tuning methods doubled GPT-3.5's performance compared to using prompts alone. Moreover, incorporating knowledge-empowered approaches enabled fine-tuned GPT models to surpass researchers' inter-rater reliability across all dimensions of student posts. The contribution of this research is twofold. First, this study introduces a novel combination of fine-tuning and knowledge-empowered approaches to enhance the capabilities of GPT models in qualitative analysis of interdisciplinary learning—a challenge inadequately tackled by unmodified GPT

models. This approach offers a valuable methodology for advancing qualitative analysis. Second, this study provides a practical framework for the automated analysis of interdisciplinary learning quality, which serves as a beneficial resource for both practitioners and researchers in the interdisciplinary learning area.

## 2 Methods

### 2.1 The Dataset

This research collected data after getting approval from the institutional review board at the researchers' institutions. The study was conducted in an interdisciplinary digital literacy class at a Southeast Asian university in 2023. The participants were 130 first- and second-year undergraduate students. This study collected learning process data: student posts on the Miro platform, an online cooperation platform (https://miro.com), during weekly learning.

In the end, 252 posts on the Miro platform were included in this study. The posts were written during the activity in the Artificial Intelligence module about preparing a debate on "Artificial Intelligence or the Internet, which has more profound implications for our society?" The posts were about examples, analysis frameworks, and arguments about the debate topic.

The dataset was divided into a training dataset and a testing dataset so that we could do fine-tuning, which is explained in detail in Sect. 2.3. The frequency of each code for each dimension in the training and the testing dataset is displayed in Table 1. The dimensions refer to the elements of interdisciplinary learning quality, which are elaborated in Sect. 2.2.

**Table 1.** The frequency of each code in each dimension

| Dimension | Training data | | | Test data | | |
|---|---|---|---|---|---|---|
| | Level 0 | Level 1 | Level 2 | Level 0 | Level 1 | Level 2 |
| 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 |

### 2.2 Human-Labeled Interdisciplinary Learning Quality

In this study, we adopted a codebook [15], which consists of four dimensions of interdisciplinary learning: diversity, cognitive advancement, disciplinary grounding, and integration. Based on interdisciplinarity, each dimension is divided into three levels. For instance, level 0 of diversity refers to the text containing no disciplinary perspective, as opposed to level 1 and level 2, in which the text

contains one or more disciplinary perspectives respectively. Subsequently, two coders independently labeled students' posts, with each post as an analysis unit. The coders, who have over a year of interdisciplinary learning and research experience, are familiar with qualitative analysis. The inter-rater reliability between human raters on each dimension is evaluated by Cohen's Kappa score (Diversity: 0.68, Cognitive advancement: 0.75, Disciplinary grounding: 0.67, Integration 0.54, overall: 0.75). They then discussed and resolved their disagreements to reach a consensus on every item, which was considered the ground truth for interdisciplinary learning quality.

### 2.3 Strategies for Enhancing GPT Model Performance

#### Prompt Engineering

*Tailored Prompt.* To optimize the use of GPT for the online posts and essay datasets, we systematically crafted the prompts by integrating the latest prompt engineering methods and tailoring them to fit the specific context in the following stages. To begin with, we adopted a codebook drawing upon educational theories [15]. Subsequently, we built a template to convert natural language in the codebook into a format that GPT can interpret. For example, we used a standardized form like the "if... then..." structure to express rules in the codebook. Additionally, system messages and task instructions were incorporated to enhance GPT's comprehension of the task. In the end, each tailored prompt was formed based on the template and consists of the following elements (see Table 2): (1) A system message that refers to a persona description for GPT; (2) A customized task instruction that provides information and requirements about the task; (3) A rule sourced from the codebook containing a set of guidelines that includes concise explanations of various examples that are relevant to different levels of a particular dimension.

*Chain-of-Thought Prompting.* In this work, we used CoT prompting to direct GPT by providing step-by-step tasks. Drawing from the CoT, the prompt consists of three primary components (see Table 2): First, the task clarification provides information about the tasks, including background, requirements, and expected output, assisting GPT in understanding the core of the task. Second, the main task is split into several smaller ones in the task breakdown section. By combining these elements, we developed a systematic, structured prompting framework that should be able to handle complex or multi-layered questions with acceptable precision and depth in responses.

**Fine-Tuning Methods.** A GPT model can be fine-tuned by adjusting it to make it more effective for specific purposes [7]. In our study, we fine-tuned the GPT-3.5 baseline model and created four fine-tuned models on four dimensions of the dataset. Then, we applied these fine-tuned models to the validation dataset and calculated Cohen's Kappa scores between GPT labels and ground truth (human labels) on these dimensions to compare the performance.

**Table 2.** Elements of prompts

| Prompt | Element | Example |
|---|---|---|
| Tailored Prompt | System messages | "You are an advanced researcher that can precisely follow the user's instructions" |
| | Customized task instructions | "Please evaluate the cognitive advancement level of students' posts, and then return ONLY numerical values 0, 1 and 2" |
| | Rules sourced from the Codebook | "Return 1 if the content has explanations, reasons, relationships, or mechanisms mentioned without explanation in detail; or elaborations of terms, phenomena" |
| Chain-of-Thought Prompting | Task clarification | "Please read the post and check the following levels about the cognitive advancement" |
| | Task breakdown | "First, please check if the content has basic explanations, reasons, relationships, or mechanisms. If not, please return 0. Second, Return 2 if the content provides extended reasoning with details, mechanisms, and examples. The explanations tend to be longer than average and contain logical words such as "thus", "because", "however", and "but". Return "1" on other conditions" |

To deal with imbalanced data (see Table 1), we used ChatGPT to generate additional samples, thereby augmenting the imbalanced dataset. The specific method is to allow ChatGPT to alternatively substitute the original samples with synonyms while maintaining the sentence's structure and meaning. Through oversampling, we ensure that each data category is represented equally in the training dataset.

**Knowledge-Empowered Strategies.** The knowledge in our study refers to external information that can help GPT better understand task-specific expertise. We incorporated two types of knowledge into the prompt: dictionary-based and rule-based knowledge.

*Dictionary-Based Knowledge.* In automated labeling tasks, a dictionary refers to a collection of words linked to a specific category [13]. To improve the performance of LLMs like GPT, we associate words with labels in the prompt.

For instance, in coding the diversity dimension, the objective is to identify the range of disciplines represented in student work. However, when students mention words such as "COVID", GPT struggles to categorize them under a specific discipline. Therefore, we integrated dictionary-based knowledge in the following format: "When students discuss topics like WORD (COVID), it indicates coverage of the LABEL (Clinical, Pre-Clinical, and Health)." Similarly, dictionaries were also applied to other dimensions of interdisciplinary learning quality.

*Rule-Based Knowledge.* While dictionary-based knowledge primarily depends on matching text with dictionary labels, rule-based knowledge draws mechanisms from the rules derived from tasks. In qualitative analysis, codebooks and inherent links between the dimensions can serve as a source for these rules. For instance, if the text does not specifically fall into any discipline (diversity = 0), then the disciplinary grounding should also be 0 (no disciplinary knowledge nor methods). However, we found that the GPT models could not easily find these inherent links in the codebook without explicit prompts. Thus, we added rules to prompts based on knowledge from the codebook in the following format: "IF DIMENSION A (Diversity) is 0, then DIMENSION B (Disciplinary Grounding) is likely to be 0; "IF DIMENSION A (Diversity) is smaller than 2 (no or just one discipline mentioned), then DIMENSION C (Integration) is very likely to be 0 (no connecting nor comparing ideas across disciplines)". This rule-based knowledge, which considers the interactions between different dimensions, has the potential to enhance GPT model performance beyond using text input alone.

## 3 Results

We conducted experiments with GPT-3.5 to assess the interdisciplinary learning quality of student posts and essays in four conditions: prompts, fine-tuning, knowledge-empowered prompts, knowledge-empowered fine-tuning. In the prompts condition, we directly utilized prompts in the GPT-3.5 model. In the fine-tuning condition, we applied the same prompts but with an additional step of fine-tuning the GPT models. The knowledge-empowered prompts condition integrated knowledge into the prompts without fine-tuning. Lastly, in the knowledge-empowered fine-tuning condition, we employed these knowledge-empowered prompts and also fine-tuned the GPT models. Figure 1 illustrates Cohen's Kappa scores for students' interdisciplinary learning quality regarding posts, reflecting the agreement between GPT-generated labels and the ground truth (human labels) in each condition. The inter-rater reliability among human raters for each dimension serves as the benchmark.

The results indicated that, by using only the prompt, GPT-3.5 shows minimal agreement with human labels, as reflected in Cohen's Kappa scores (all lower than 0.35). After using fine-tuning approaches, Cohen's Kappa scores show a twofold increase (ranging from 0.39 to 0.74), achieving moderate agreement with the ground truth. However, Cohen's Kappa scores on the disciplinary grounding and integration dimensions are still fairly lower than human raters (for disciplinary grounding: 0.61 versus 0.67, and for integration, 0.39 versus 0.54).

Additionally, knowledge-empowered approaches enhance both prompt-based and fine-tuning methods. By incorporating knowledge into prompts, the performance on each dimension increases. Similarly, after combining fine-tuning and knowledge-empowered approaches, GPT models' performance shows a notable increase. The efficacy of integrating fine-tuning with knowledge-enhanced methods has been validated across all dimensions in student posts: diversity (0.80 vs. 0.68), cognitive advancement (0.79 vs. 0.75), disciplinary grounding (0.71 vs. 0.67), and integration (0.70 vs. 0.54). The results suggest that knowledge-empowered fine-tuning approaches are outperforming the proficiency of human judges. We additionally compared the models between imbalanced data and balanced data with oversampling in the knowledge-empowered fine-tuning condition. The result showed that models trained by GPT-augmented balanced data exceed the models trained by imbalanced data on all the four dimensions: diversity (0.80 vs. 0.70), cognitive advancement (0.79 vs. 0.73), disciplinary grounding (0.71 vs. 0.61), and integration (0.70 vs. 0.57).

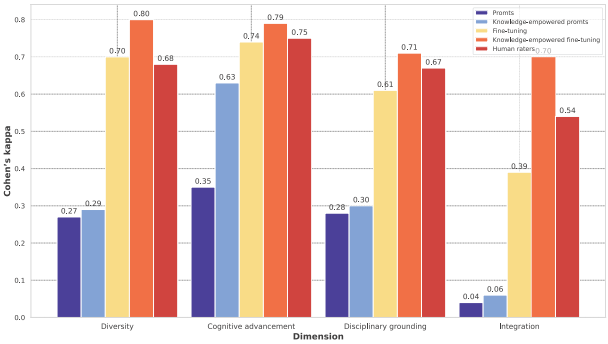


**Fig. 1.** Cohen's Kappa scores derived from prompt-based, fine-tuned, and knowledge-powered GPT models versus human raters in student posts

## 4 Discussion and Conclusion

This study developed a method for evaluating interdisciplinary learning with GPT models. Starting with human experts evaluating students' work to establish a benchmark, we then tested the GPT 3.5 model's performance on four conditions: prompts, fine-tuning, knowledge-empowered prompts, and knowledge-empowered fine-tuning. The findings indicate that knowledge-empowered approaches in prompts and fine-tuning can significantly enhance the effectiveness of GPT 3.5 and achieve accuracy surpassing that of human experts.

Our results align with previous research [7] on fine-tuning's benefits for natural language tasks like open-domain question answering and table-to-text generation. Moreover, we found oversampling beneficial for handling imbalanced

data, resonating with previous studies [12]. Our findings indicate that knowledge-empowered strategies can improve both prompt-based and fine-tuned GPT models. Echoing previous work on the value of external knowledge for prompt engineering in automatic essay analysis [13], our research shows that even small-scale domain-specific dictionaries or rules can significantly improve LLM effectiveness. This contrasts with studies relying on large knowledge graphs to solve knowledge-driven problems [9]. Furthermore, our findings highlight the importance of incorporating knowledge in the fine-tuning process, which hasn't received much attention.

This study still has some limitations. First, the results of this study are based on a limited number of online notes generated by a particular cohort of undergraduate students. Whether the methods can be generalized to other datasets needs further research. Secondly, this study focuses solely on measuring agreement levels; a more thorough error analysis of disagreement data points is essential for a deeper comprehension of the model's capabilities. Third, our study exclusively evaluated the GPT-3.5 model, leaving out numerous other LLMs like LLaMA and Gemini from future investigation.

Our research offers a practical framework for the automated analysis of interdisciplinary learning quality, paving the way for large-scale studies that were previously unfeasible due to the intensive labor required for qualitative analysis. The precision approach is beneficial for understanding the complex dynamics of interdisciplinary education and sets a new benchmark for the use of artificial intelligence in educational research.

**Acknowledgement.** This study was supported by the NTU Edex Teaching and Learning Grants (Grant No. NTU EdeX 1/22 ZG). We are indebted to the students and instructor who participated in this study.

## References

1. Boix-Mansilla, V.: Learning to Synthesize: The Development of Interdisciplinary Understanding, pp. 288–306. Oxford University Press, Oxford (2010)
2. Brassler, M., Dettmers, J.: How to enhance interdisciplinary competence-interdisciplinary problem-based learning versus interdisciplinary project-based learning. Interdiscip. J. Probl.-Based Learn. **11**(22) (2017)
3. Chejara, P., et al.: EFAR-MMLA: an evaluation framework to assess and report generalizability of machine learning models in MMLA. Sensors **21**(8), 2863 (2021)
4. Elliott, V.: Thinking about the coding process in qualitative data analysis. Qual. Rep. **23**(11) (2018)
5. Fereday, J., Muir-Cochrane, E.: Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development. Int. J. Qual. Methods **5**(1), 80–92 (2006)
6. Gvili, I.E.F., et al.: Development of scoring rubric for evaluating integrated understanding in an undergraduate biologically-inspired design course. Int. J. Eng. Educ. (2016)
7. Liu, J., et al.: What makes good in-context examples for gpt-3? arXiv preprint arXiv:2101.06804 (2021)

8. Liu, P., et al.: Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput. Surv. **55**(9), 195:1–195:35 (2023)
9. Liu, W., et al.: K-BERT: enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, pp. 2901–2908 (2020)
10. OpenAI: Openai. https://openai.com/
11. Reynolds, L., McDonell, K.: Prompt programming for large language models: beyond the few-shot paradigm. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–7 (2021)
12. Shelke, M.S., et al.: A review on imbalanced data handling using undersampling and oversampling technique. Int. J. Recent Trends Eng. Res. **3**(4), 444–449 (2017)
13. Ullmann, T.: Automated Analysis of reflection in writing: validating Machine learning approaches. Int. J. Artif. Intell. Educ. **29**(2), 217–257 (2019)
14. Xiao, Z., et al.: Supporting qualitative analysis with large language models: combining codebook with GPT-3 for deductive coding. In: Companion Proceedings of the 28th International Conference on Intelligent User Interfaces, pp. 75–78 (2023)
15. Zhong, T., et al.: The influences of chatgpt on undergraduate students' perceived and demonstrated interdisciplinary learning. OSF preprint (2023)