

## **Work-in-Progress: Fine-Tuning Large Language Models for Automated Feedback in Complex Engineering Problem-Solving**

**Mrs. Paula Francisca Larrondo, Queen's University**

Paula is a Ph.D. Candidate at Queen's University. She is working towards a neural network model to provide automated feedback to open-ended student work in the context of complex problem-solving in Engineering Education (Co-supervised by Dr. Brian Frank from the Department of Electrical and Computing Engineering and Dr. Julian Ortiz from the Robert M. Buchan Department of Mining). She is a Geologist with an M.Sc. from Universidad de Chile (Chile) and an M.Sc. in Mining Engineering (Geostatistics) from the University of Alberta (Canada).

**Prof. Brian M Frank P.Eng., Queen's University**

Brian Frank is the DuPont Canada Chair in Engineering Education Research and Development, and the Director of Program Development in the Faculty of Engineering and Applied Science at Queen's University where he works on engineering curriculum development,

**Julian Ortiz, Queen's University**

Dr. Ortiz is a Mining Engineer from Universidad de Chile and Ph.D. from University of Alberta. Currently, he is Professor and Mark Cutifani / Anglo American Chair in Mining Innovation at University of Exeter - Camborne School of Mines, in the United Kingdom, where he conducts research related to geostatistical ore body estimation and simulation, and geometallurgical modeling using statistical learning. Dr. Ortiz's previous roles include Head of Department at Queen's University and Universidad de Chile.

# Work-in-Progress: Fine-Tuning Large Language Models for Automated Feedback in Complex Engineering Problem-Solving

## Abstract

*This paper presents **work in progress** (WIP) toward using artificial intelligence (AI), specifically through Large Language Models (LLM), to support rapid quality feedback mechanisms within engineering educational settings. It describes applying to LLMs to improve the feedback processes by providing information directly to students, graders, or course instructors teaching courses focused on complex engineering problem-solving. We detail how fine-tuning an LLM with a small dataset from diverse problem scenarios achieves classification accuracies close to approximately 80%, even in new problems not included in the fine-tuning process. Traditionally, open-source LLMs, like BERT, have been fine-tuned in large datasets for specific domain tasks. Our results suggest this may not be as critical in achieving good performances as previously thought. Our findings demonstrated the potential for applying AI-supported personalized feedback through high-level prompts incentivizing students to critically self-assess their problem-solving process and communication. However, this study also highlights the need for further research into how semantic diversity and synthetic data augmentation can optimize training datasets and impact model performance.*

**Keywords:** Automated formative feedback, Complex problem-solving, Engineering Design, Large Language Models

## Introduction

Complex problem-solving skills (CPS) are key to meeting the demands of engineering graduates' future roles [1]. Developing these skills requires frequent practice in a variety of authentic and open-ended complex problems [2, 3]. Despite its critical role in education, assessing CPS effectively remains a significant challenge. Jonassen highlights the assessment of problem-solving skills as a notably weak area in instruction, emphasizing that evaluations should go beyond the student's ability to recall conceptual or procedural knowledge and focus on problem understanding, students' problem-solving performance, cognitive abilities and domain-specific knowledge [2]. Furthermore, providing timely, personalized feedback on the student's understanding or problem-solving performance in complex, open-ended and ill-structured problems in large classes poses significant challenges. Given that student work in these courses often includes the evaluation of extended reports, large teams of teaching assistants are usually recruited to minimize the response time. However, achieving consistency in the quality and quantity of feedback is difficult, even in the presence of multiple calibration strategies.

Recent advances in natural language processing (NLP) technologies, through the use of artificial intelligence (AI) agents and Large Language Models (LLM), have already provided significant advantages in the holistic assessment of high-order features such as argumentation, use of evidence or scientific thinking [4-6]. With the evolution of Automated Feedback Systems (AFS) [7-9] and, more recently, the release of Open AI's ChatGPT, LLMs have become commonplace in higher education among students and instructors [10, 11]. The emergence of LLMs in higher and secondary education has triggered an influx of publications on the opportunities and

challenges of incorporating these technologies in instruction and evaluation [10, 12, 13]. However, the unique nature of engineering design problems, characterized by their complexity and the validity of multiple solutions, presents distinct challenges to generating targeted elaborated feedback. In addition, to maximize student engagement and learning from the exposure to this complex problem, novelty is introduced by changing the scenario and problem to be solved each term. The novelty introduced by crafting a new problem every term can reduce pattern detection accuracy, thereby impacting the pertinence of the automated generated feedback.

This contribution is part of a larger study on the impact of AI-generated feedback on open-ended student work. The study explored a fine-tuned LLM classification method for generating on-demand automated feedback on students' written drafts before their final submission. The generated feedback prompts are designed to provide students insights into the engineering design process' clarity and completeness of their discourse on the engineering design process, prioritizing critical self-evaluation over direct corrections. This study explicitly examines how models fine-tuned with sentences derived from problem scenarios from previous terms can be used to classify students' responses to a novel scenario in the upcoming term, especially when dealing with limited datasets for fine-tuning.

## **Related Work and Research Questions**

Complexity in engineering workplace problems commonly originates from having unclear or multiple solution paths, high uncertainty, multidisciplinary requirements, conflicting goals, non-engineering constraints, and numerous criteria evaluations of solutions. Complex problems are often associated with open-ended problems since they can have multiple solutions or with ill-structured problems because they could have been vaguely defined or have unclear goals and unstated constraints [14-18]. However, complexity in problem-solving extends beyond the open-endedness or ill-structured nature of problems. Jonassen and Hung [15] argue that complexity depends on the breadth of knowledge required to solve the problem, the attainment level of domain knowledge, the intricacy of problem-solution procedures, and the relational complexity, which involves processing multiple relationships simultaneously.

In this context, the breadth of knowledge required to solve the problem and attain domain acquaintance will depend on each student's particular cognitive skills and domain knowledge. So, in addition to contemplating multiple valid solutions, feedback had to be constructed on the student's individual progress, as recommended by Hattie and Timperley [19]. Furthermore, recent studies on formative feedback indicate that the most effective way to improve learning is through elaborated feedback that incorporates details on how to improve on the task and the process and develop self-regulated learning strategies [20-26]. Effective feedback has different effects depending on the operating level. According to Hattie and Timperley's [19] feedback framework, "These include the level of task performance, the level of process of understanding how to do a task, the regulatory or metacognitive process level, and/or the self or personal level (unrelated to the specifics of the task)" [19 p.86]. Another critical criterion of feedback effectiveness is the student's action upon the received feedback [26, 27]. Computer-based learning environments offer an advantage by ensuring students have enough time to act on the recommendation [9, 21, 22]. The latter is particularly critical in large enrollment courses where

instances for personalized feedback from instructors or graders are scarce, and developing self-regulating skills on CPS and written communication is essential to designing an effective feedback process [25].

Advancements in AI, mainly through large pre-trained language models, have revolutionized NLP tasks [28, 29]. Based on a “transformer” architecture, Bidirectional Encoder Representations from Transformers (BERT) and OpenAI’s Generative Pre-Trained Transformer (GPT) are prominent examples of these models [30, 31]. A fundamental mechanism behind the success of LLM is self-attention, which allows the models to recognize relationships between words regardless of their order in the textual sequence, thereby improving the ability of the models to deal with long-term dependencies [32]. Pre-trained LLMs are originally trained in massive text corpus prior to being fine-tuned on a specific task. This approach has been proven effective for improving performance on various NLP tasks, such as sentence classification, question answering and named entity recognition [33]. While early automated feedback systems relied on domain-expert rules and were limited in addressing the diversity of open-ended assignments [34-36], data-driven approaches, though promising in highly semantically diverse responses, often face challenges due to the lack of extensive training datasets [4, 37, 38].

AFS based on LLMs holds the potential for a more effective and efficient solution. Applications range from personalized hints for programming assignments [39] to reflective writing [40], including feedback on the appropriateness of the topic of a data science project proposal and the description clarity of goals, benefits, novelty and overall clarity of the report [41]. Despite the promising results from studies like Dai, et al. [41], the study also highlights potential limitations in the reliability of LLM-generated feedback compared to instructors’ assessments of student performance. Among the evaluation metrics used in their study, the authors assess the alignment between ChatGPT-generated feedback and instructor feedback regarding student performance. As outlined by Hattie and Timperley [19], feedback should aim to reduce the gap between current and desired understanding. This goal can be attained in the feedback process when the instructors affirm the student’s effort or indicate if there are areas for further improvement. If the automated feedback generator is unable to give feedback that accurately indicates how students perform, the generated feedback can inadvertently mislead the student and could negatively affect the student’s learning.

According to Dai, et al. [41], affirmative automated feedback provided by ChatGPT was present in most reports (between 85 and 95%, depending on the assessed dimension). In contrast, for the same reports, a minority received positive instructor feedback (4 and 20%, respectively). Among several possible explanations, the authors highlight the absence of specific ChatGPT training on the measurement of assignment quality or the use of an indication of golden feedback. As discussed by the authors, this pattern of predominantly affirmative feedback from automated systems indicates that feedback generated directly from a language model API like ChatGPT may not align perfectly with human instructors’ assessments. This discrepancy could stem from various factors, including the model’s training and the instructors’ unintended focus on areas for student improvement over providing positive reinforcement. Without human intervention or specific fine-tuning, the feedback might focus on the process or task levels, missing more relevant feedback at a self-regulation level. Such feedback is essential as it reinforces the

application of generic problem-solving processes used in engineering design, a key component to developing complex problem-solving skills [17, 18].

Given these challenges, the project aims to assess the efficacy of LLM-based automated quality feedback on the student's individual progress in complex problem-solving tasks at a self-regulated level. The feedback mechanism, based on a set of measurements obtained from the LLM classification of the student sentences, aims to identify missing elements of the engineering design process. The classes and classification probability of the different sentences in the student's work are translated into a 5-point completeness/conciseness and clarity scoring matrix. Feedback prompts are selected for the student based on the scoring of all sentences classified to each dimension.

In this context, the present contribution focuses on the impact of a small fine-tuning dataset on the accuracy of identifying the presence and absence of dimensions within *problem definition* in a design process from a new scenario in the upcoming term. Specifically, the research questions this contribution aims to answer are the following:

**RQ1:** To what extent do the limited annotated datasets impact the classification accuracy for abstract dimensions of problem definition in the context of complex engineering design problem-solving?

**RQ2:** How does fine-tuning using different problem scenarios impact the classification accuracy underlying the feedback mechanism?

**RQ3:** Could automatically generated sentences compensate for the scarcity of annotated data?

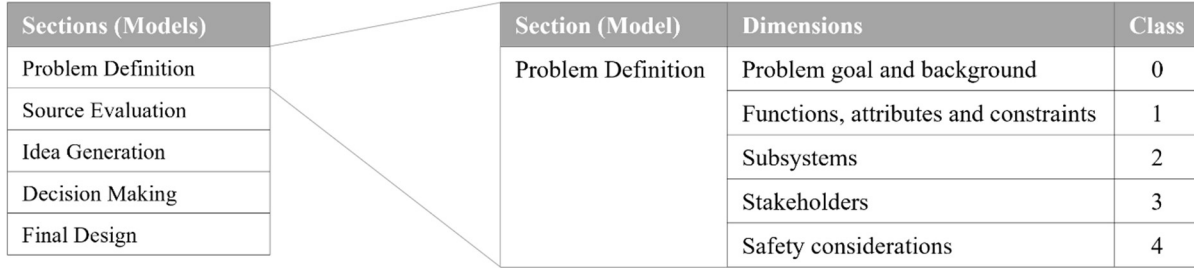
## Methods

### *Context and Course Description:*

The study was set in the context of a first-year engineering design course at a medium-sized, research-intensive Canadian university, part of the faculty-wide Engineering Design and Practice Sequence. This sequence prepares students for real-world, open-ended design problems [42]. Part of the student summative evaluations includes team reports at different stages of the design process, such as *Scoping and Preliminary Design Concept*, *Full Design Proposal*, or *Final Design*, following Dym et al.[17] framework. Our analysis focuses on the *Problem Definition* sections of these reports, which include identifying the problem goal, stakeholders, safety considerations, functions, attributes, and constraints, among other possibilities that can be added as the students progress in their project design.

### *Feedback Mechanism and Model Fine-tuning:*

A multi-class classification approach using a fine-tuned LLM is the base for the feedback-triggering mechanism. This method involves classifying student sentences into pre-defined classes [43, 44] that reflect specific dimensions of the engineering design generic problem-solving process [17, 18]. An example of the dimensions included in the testing done for the *Problem Definition* stage, presented in this contribution, is included in Figure 1.



Sections (Models)	Section (Model)	Dimensions	Class
Problem Definition	Problem Definition	Problem goal and background	0
Source Evaluation		Functions, attributes and constraints	1
Idea Generation		Subsystems	2
Decision Making		Stakeholders	3
Final Design		Safety considerations	4

Figure 1 illustrates the models and dimensions adopted for the problem-definition stage of a generic problem-solving process.

The fine-tuning process is based on the “further in-domain pre-training” strategy described by Sun, et al. [45], using sentences from 32 reports taken randomly from a pool of approximately 140 reports each term. The selected reports are from two terms equivalent to two different problem scenarios: model and construction of a scale prototype of a Hyperloop vehicle [46] run in the 2019 academic year and create a working prototype of an assistive robotic arm, done in the 2022 academic year. After ethics approval, text was extracted from PDF and Word documents, followed by minor preprocessing and splitting using the spaCy library<sup>1</sup>.

#### *Annotation and Dataset Preparation:*

Manual annotation was initially performed by one of the authors, and ambiguous cases were discussed with the course instructor. Human annotation is a time and resource-consuming task that can limit the size of the fine-tuning dataset. To expedite annotation validation, we employed GPT version 4.0 for comparison, addressing discrepancies between human and GPT-4 classifications. The use of GPT for intercoder agreement validation is supported by recent research showing that generative language models can outperform trained annotators [47].

Given the labour-intensive nature of manual annotation, we explored alternative strategies to expand the dataset generation and mitigate the instabilities associated with small datasets for fine-tuning in text classification [45]. One such strategy was to generate a synthetic dataset: 36 instructor-devised sentences were augmented to 450 sentences using ChatGPT. Additionally, we experimented with fine-tuning the model directly using the instructor’s exemplar to further assess the fine-tuning capabilities without manual annotation or AI-generated augmentation. This approach also allowed us to explore the effectiveness of utilizing well-structured, expert-created content as a standalone fine-tuning source.

In addition to sentence annotation for fine-tuning, 169 and 109 sentences from the *Problem Definition* sections of the 2022 and 2019 reports were manually annotated for validation purposes. The purpose of these datasets is to evaluate the model’s classification accuracy performance across different fine-tuning exercises. Details regarding the number of sentences and the sources for each fine-tuning dataset used in this study are summarized in Table 1. The 2022 dataset was in the context of an assistive robotic arm design, and 2019 was a scale model hyperloop vehicle design. In each dataset, the name “2019” or “2022” corresponds to the academic year from which the training data was drawn, and the final number corresponds to the number of sentences used to fine-tune the model.

<sup>1</sup> <https://spacy.io/>

Table 1 details the fine-tuning and validation datasets used in this work.

Dataset	Number of sentences per class						Description
	0	1	2	3	4	Total	
2022_745	169	277	73	123	103	745	Fine-tuning dataset. Include all annotated sentences from the 2022 term.
2022_397	92	147	40	63	55	397	Randomly selected subsets of the training dataset 2022_745. Classes' relative proportions were kept similar.
2022_200	46	74	20	32	28	200	
2022_100	23	37	10	16	14	100	
2022_36	6	11	9	5	5	36	
2022_Exemplar	6	11	9	5	5	36	Fine-tuning dataset comprising 36 sentences corresponding to the instructor's exemplar.
2022_GPT	90	120	120	60	60	450	Synthetic dataset generated through using ChatGPT to augment the course instructor's exemplar of 36 sentences to a total of 450 sentences.
2022_2019_866	182	316	122	143	103	866	2022_745 fine-tuning dataset plus 121 sentences from the 2019 term.
2022_Validation	34	51	32	26	26	169	Validation dataset, comprising sentences from the 2022 term.
2019_Validation	19	43	24	23	0	109	Validations dataset comprising sentences from the 2019 term.

### Model Selection and Fine-tuning:

Despite the advancements in GPT models, we opted for the distilBERT base uncased model due to its open-source nature, domain specificity, and cost-effectiveness [48-51]. This choice was supported by the model performance in highly specific domains where in-domain further pre-training is advised for LLM. Bosley, et al. [48] showed that although not explicitly, commercial models like GPT might still need a form of outsourced fine-tuning by including examples in the prompt, known as in-context learning, to achieve performances observed in fine-tuned BERT-type models. Furthermore, to achieve the specific level of targeted feedback, according to Hattie and Timperley's [19] feedback framework, using a GPT model would require in-context learning to ensure that the feedback generated encourages student self-regulation and does not focus on the task as described by Dai, et al. [41]. In addition, the need for explainability in the feedback process and concerns about student data privacy also play a role in the decision. Finally, although the cost of text processing is relatively low in unitary terms, reports consist of various thousands of tokens. Moreover, when using data drawn from courses with approximately 1000 students and several submissions each term, the cost becomes significant, especially in the context of testing the applicability of such a tool.

The pretrained distilBERT base uncased model was adopted based on their performance and accuracy after testing BERT, distilBERT, RoBERTa, BART and distilBART. The adopted model is a general-purpose distilled pre-trained version of the BERT model [52]. A first fully connected pre-classifier layer streamlines the output from 768 to 32 features [53], followed by a 30% dropout layer for optimization. The model then classifies these features into the target number of classes using a fully connected layer and a SoftMax layer for standardized output. We use the Adam optimizer with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ , a learning rate of  $1e-05$ , a maximum sentence length of 256, and batch sizes of 32. The model was trained over 50 epochs to check for overall and within each class overfitting. These relatively large number of epochs are also



recommended for small fine-tuning datasets [50]. A performance analysis was conducted to determine the optimum fine-tuning dataset size and impact of sentence sources. Although performance in multi-class classification tasks can be reported using different measures, for comparison purposes, we reported the accuracy calculated as the number of correct predictions divided by the total number of predictions turned into percentages. This measurement is suitable in balanced datasets with relatively homogenous class proportions, as is the case in the dataset details in Table 1.

## Results

### *Effect of Fine-Tuning Dataset:*

Our initial analysis examines classification accuracies of distilBERT models fine-tuned with datasets ranging from 745 to 36 sentences (Table 1). We compared these to a zero-shot classification baseline, where class names were included in the classification prompt without prior fine-tuning. As illustrated in Figure 2, all fine-tuned models outperformed the baseline, demonstrating the value of fine-tuning across all dataset sizes. Notably, it is observed that models fine-tuned on relatively smaller datasets can achieve comparable accuracy to larger ones after additional training epochs, aligning with findings by Zhao, et al. [50] that suggest extended fine-tuning on small datasets can partially offset the need for larger annotation volumes. Indeed, when fine-tuned with half the size of the existing dataset, similar accuracy values are obtained after 21 epochs. For smaller datasets, it is unclear at this time if fine-tuning for more than 50 epochs would compensate for the loss of accuracy or if this factor has a ceiling.

### *Class-specific Performance:*

The average improvement after additional training epochs is not consistent among the different classes. Classes with higher misclassification rates, such as *Safety Considerations* (Figure 3), are more affected when fewer annotated sentences are available during fine-tuning. For this particular class, when fine-tuned with half the sentences (*2022\_397 dataset*), the resultant classification accuracy is systematically 10 points below that when all sentences are considered. Overall, the *Safety Considerations* class shows the lowest accuracies across all epochs compared to the other classes despite this class not being the least represented. Possible explanations for this result are included in the discussion section. The previous behaviour is accentuated when comparing the accuracy obtained with smaller fine-tuned datasets of 200 sentences or less. As the number of sentences available for fine-tuning decreases, not only does the overall accuracies decrease, but the number of classes that consistently show lower accuracies increases, and the number of epochs needed to be stabilized increases as well.

### *Synthetic Dataset Comparison:*

For the purpose of comparing the models fine-tuned with the synthetic dataset (*2022\_GPT*), the model fine-tuned with the manually annotated dataset, *2022\_397*, was chosen as they have a similar number of sentences per class. The comparison shows a systematic decrease in accuracy across all classes, with the exception of the *Safety Considerations* class, which shows an increase that ranges between 10 and 20 points (Figure 4). Among the classes for which the model fine-tuned with synthetic sentences derived from an exemplar resulted in lower classification accuracies, the performance of the *Subsystems* and *Functions, Attributes, and Constraints* classes



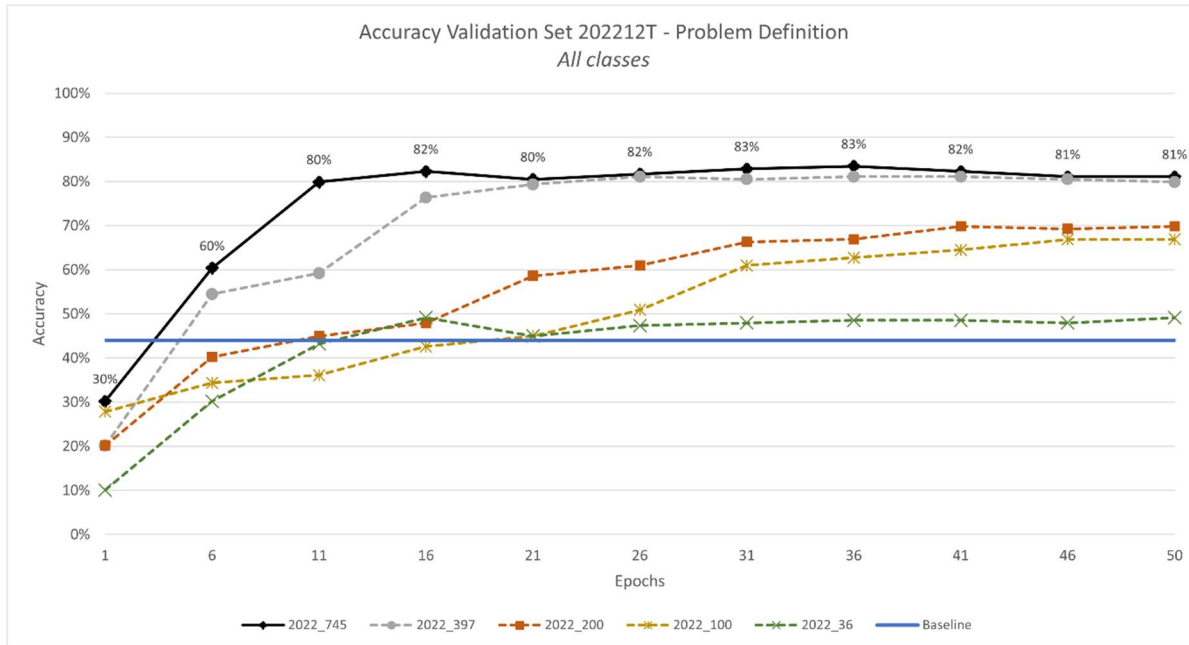


Figure 2 shows the overall validation accuracy across epochs for various fine-tuned models, illustrating the benefits of fine-tuning even with limited data.

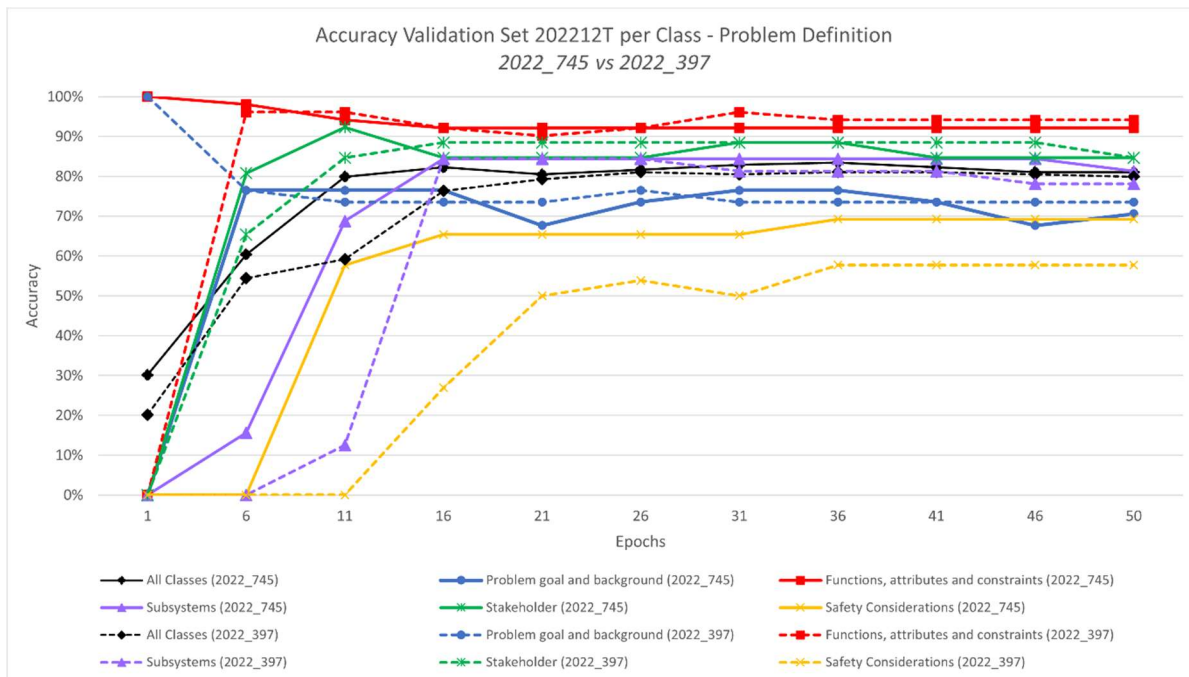


Figure 3 contrasts class-specific accuracies between models fine-tuned on the full dataset (2022\_745 full lines) versus a reduced dataset (2022\_397 dash lines), highlighting disparities in misclassification rates.

stands out. These two classes have relatively more sentences than the rest of the classes. Particularly, the *Subsystems* classes have significantly more synthetic sentences for fine-tuning than the manually annotated dataset. The latter is due to the exemplar's class proportion differing from those in the students' reports. Further analysis of the possible relation between the number

of sentences and observed accuracy in the presence of synthetic data is included in the next section.

#### *Cross-term Predictive Accuracy to Novel Scenarios:*

To assess the model's performance when applied to a new scenario, we tested a fine-tuned model using sentences from the *2022\_745* dataset. These sentences were extracted from students' reports on the creation of a working prototype of an assistive robotic arm. We applied this model to classify sentences from the *2019\_Validation* dataset, which includes reports on the modelling and constructing a scale prototype of a Hyperloop vehicle. Although accuracy varied among classes and over epochs, the overall accuracy decreased by an average of 5%, as shown in Figure 5. The difference in accuracy was notably smaller than anticipated. These discrepancies were slightly higher for the *Subsystems* and *Stakeholder* classes.

We then tested whether these asymmetric decreases in accuracy persisted after incorporating additional sentences from the 2019 problem scenario, attempting to simulate the early availability of an instructor's exemplar or guided solution. Consequently, we generated an expanded dataset (*2022\_2019\_866*), adding 121 sentences from the 2019 scenario into the *2022\_745* dataset. These sentences are different from those used in the validation process, serving to develop a new fine-tuned model. When validated against the 2022 sentences, accuracies decreased by an average of 1% across all classes and epochs; notably, despite the differing backgrounds of the two scenarios, the accuracy of the *Problem Goal and Background* categories increased. However, these results require further analysis, as the absence of the *Safety Considerations* class in the 2019 instructions, which led to no sentences attributed to this class, could have influenced the outcomes. When this expanded model, which incorporates both the 2022 and 2019 scenarios, was validated against the *2019\_Validation* dataset—as would be done with student reports in the upcoming term—overall accuracy further decreased to an average of 76% across all classes and sufficient training time, as shown in Figure 5.

## **Discussion**

Our results on the influence of fine-tuning dataset size on classification accuracy suggest that AI educational tools based on LLM have the potential to achieve high efficiency even with limited data. The performance of fine-tuned distilBERT models across a range of dataset sizes shows that accuracies of approximately 80% can be achieved with datasets of as little as 400 sentences, which is equivalent, in this case, to approximately 14 reports, in contrast to the order of thousands usually reported in the literature for fine-tuning LLM models [45, 53, 54]. This observation aligns with the research conducted by Zhao, et al. [50], which suggests that with sufficient training, smaller datasets can nearly match the classification accuracy achieved by larger datasets.

The retention of relatively high accuracies in classifying sentences from novel scenarios using models fine-tuned with data from previous terms represents a significant advancement in the development of automated feedback systems for engineering design education. This ability to transfer problem-solving knowledge and skills to new situations is crucial, as highlighted by Jonassen [2]. Our results, presented in Figure 5, demonstrate the model's robustness across different training scenarios, whether they are based on more diverse fine-tuning datasets or solely

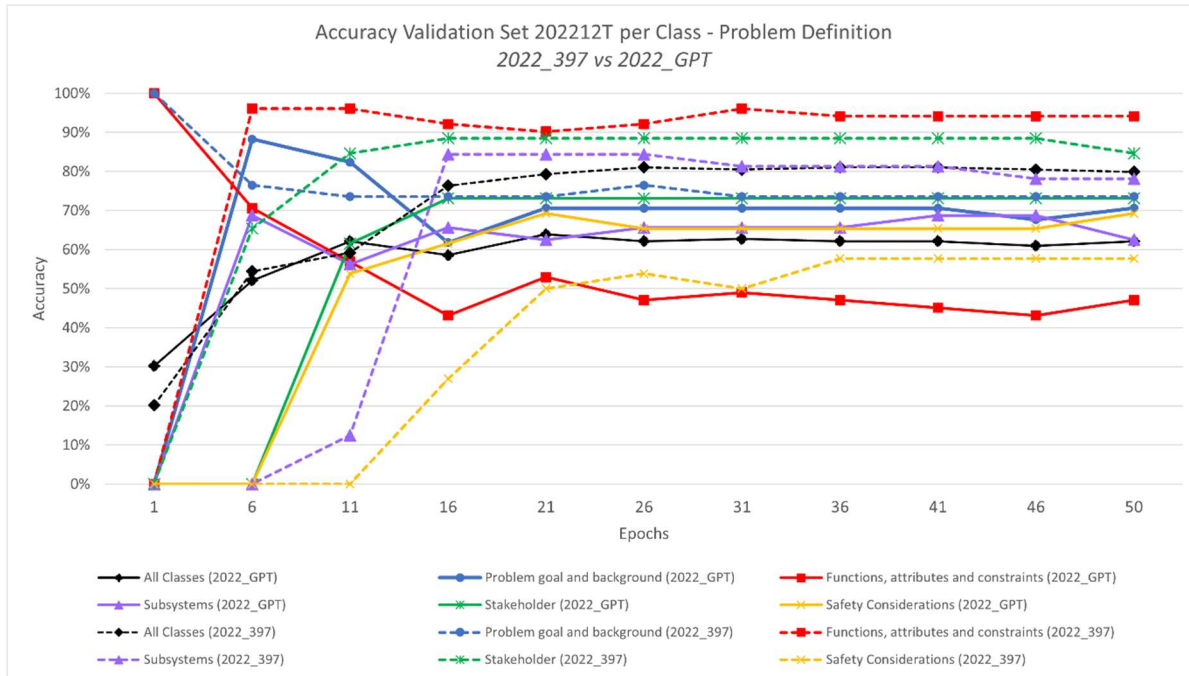


Figure 4 compares class accuracy between models trained on synthetic (full lines) and a manually annotated dataset (2022\_397 dash lines), highlighting the limitations of synthetic data.

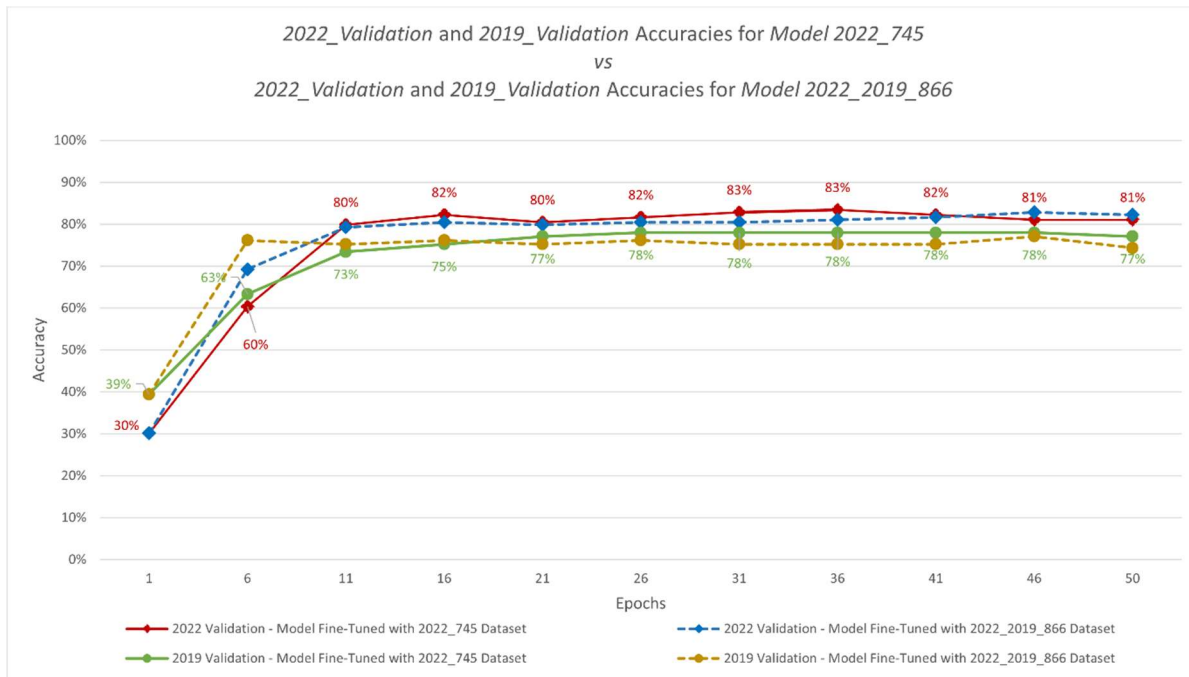


Figure 5 illustrates the impact of training diversity on model accuracy when applied to novel scenarios. It displays the accuracies for models fine-tuned with sentences solely from the 2022\_745 dataset (shown with solid lines) compared to models trained with an expanded dataset that includes sentences from both 2022 and 2019 (dashed lines). This comparison demonstrates that using models fine-tuned with sentences from previous term scenarios (indicated by diamond markers) to classify sentences from novel scenarios (indicated by circle markers) does not significantly reduce accuracy.

on historical data. This finding is particularly important for our study, as it is expected that student-generated sentences, which could extend beyond those derived from an instructor's exemplar, will not be available for annotation before the start of the term. Further testing in the effect of broadening the problem scenarios with cases from additional terms is underway.

The analysis of our model's performance showed that the classification accuracy across different classes can vary as much as 20% for the larger dataset. The range of variation for models fine-tuned with smaller datasets can increase to 30%, suggesting the variation, in some cases, could be linked directly to a smaller number of annotated sentences for fine-tuning. However, this possible explanation does not apply to all classes. The *Safety Considerations* class exhibits systematically lower accuracies across the different test cases despite containing nearly 40% more sentences than the class with the smallest number of sentences for fine-tuning. Further analysis is needed to understand the impact of linguistic and semantic features of the different classes in the fine-tuning process.

Furthermore, it is possible that the linguist and semantic diversity, or the lack of it, may have played a role in the lower accuracies observed in the model trained with ChatGPT augmented sentences that we derived solely from the instructor's exemplar. Students' responses from different reports exhibit a diverse linguist and semantic diversity compared to the instructor's exemplar. Nevertheless, using synthetic data generated through techniques such as augmentation with tools like ChatGPT offers a promising strategy for enriching underrepresented classes in existing annotated fine-tuning datasets. Further testing on the utility of using ChatGPT augmentation capabilities on an instructor's exemplar of novel scenarios to be added to existing annotated sentences from previous terms is underway and will be presented in future publications.

## **Conclusion**

Our study provides valuable insights into the application of LLM models to engineering design education. We show that open-source distilBERT models, when fine-tuned with small datasets of 400 sentences on the problem definition dimensions of a generic problem-solving process used in engineering design, can achieve an accuracy of approximately 80%. This performance suggests that, as shown by Zhao et al. [50], the need for large datasets to fine-tune LLMs effectively may not be as restrictive as previously thought. These findings are particularly promising for instructors or educational settings with limited resources for extensive dataset development who would prefer to develop in-house fine-tuned models.

Our results confirm that the fine-tuning process retains relatively high accuracy even when incorporating data from various terms or using it across terms, indicating the model's robustness in handling diverse problem scenarios without significant performance reduction. This feature is particularly relevant in engineering education, where learning to transfer knowledge and skills to new problems is essential. Our analysis suggests that AI tools based on LLM have the potential to support a wide range of educational content beyond engineering design.

However, further research is needed on how linguistic and semantic features, among other factors, influence the model's performance and how synthetic data might be used effectively to supplement training datasets. While synthetically generated annotations can improve underrepresented classes in training datasets, more detailed exploration is needed to understand the underlying requirement to achieve the potential fully.

In summary, our findings provide positive preliminary results on the use of LLMs to achieve meaningful accuracies in triggering automated feedback prompts at a self-regulation level anchored to a generic problem-solving process widely used to develop complex problem-solving skills in engineering design.

## **Limitations**

In this section, we explore some of the limitations associated with the use of large language models (LLMs) in educational contexts. In addition to the well-documented issues of bias, privacy, and copyright concerns that have been detailed by several authors [7, 10-13]. AI-based tools have specific limitations beyond the scope of this study when used in assessment contexts that differ from providing formative feedback. The accuracy achieved in this study is suitable for triggering responses when classifying several sentences as feedback prompts based on average behaviour rather than precise individual statements. Furthermore, the targeted feedback prompts generated aim to incentivize students' critical self-evaluation of their problem-solving process and communication, and they are not intended for specific corrections at a sentence level.

From a technical perspective, the results presented in this study are also constrained by the specificities of the dataset used, the selected LLM, and the parameters utilized in the fine-tuning process. These factors limit generalizability and applicability across different educational settings or disciplines.

## **Acknowledgements**

The DuPont Canada Chair in Engineering Education Research and Development supported this work. The secondary use of student reports for fine-tuning purposes reported here was approved by the Queen's University General Research Ethics Board.

ChatGPT assistance was used to identify areas where clarity and conciseness would improve the first author's drafts. AI-generated suggestions were examined and incorporated as deemed appropriate by the first author. All authors reviewed and revised the final submission.

## References

- [1] International Engineering Alliance, "Graduate Attributes and Professional Competencies," Jun. 21, 2021. Accessed: Oct. 14, 2021. [Online]. Available: <https://www.ieagrements.org/assets/Uploads/IEA-Graduate-Attributesand-Professional-Competencies-2021.1-Sept-2021.pdf>
- [2] D. H. Jonassen, *Learning to solve problems: A handbook for designing problem-solving learning environments*. Routledge, 2010.
- [3] S. Sheppard, K. Macatangay, A. Colby, W. M. Sullivan, and L. S. Shulman, *Educating engineers: Designing for the future of the field*, 1st ed. San Francisco, CA, USA: Jossey-Bass, 2009.
- [4] H. S. Lee, A. Pallant, S. Pryputniewicz, T. Lord, M. Mulholland, and O. L. Liu, "Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty," *Science Education*, vol. 103, no. 3, pp. 590-622, 2019.
- [5] H. Zhang *et al.*, "eRevise: Using natural language processing to provide formative feedback on text evidence usage in student writing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Jul. 2019, vol. 33, no. 01, pp. 9619-9625, doi: <https://doi.org/10.1609/aaai.v33i01.33019619>.
- [6] M. Zhu, O. L. Liu, and H.-S. Lee, "The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing," *Computers & Education*, vol. 143, p. 103668, 2020.
- [7] U. Maier and C. Klotz, "Personalized feedback in digital learning environments: Classification framework and literature review," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100080, 2022/01/01/ 2022, doi: <https://doi.org/10.1016/j.caeai.2022.100080>.
- [8] A. P. Cavalcanti *et al.*, "Automatic feedback in online learning environments: A systematic literature review," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100027, 2021, doi: 10.1016/j.caeai.2021.100027.
- [9] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerd, "A review of automated feedback systems for learners: Classification framework, challenges and opportunities," *Computers & Education*, vol. 162, p. 104094, 2021, doi: 10.1016/j.compedu.2020.104094.
- [10] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023/04/01/ 2023, doi: <https://doi.org/10.1016/j.lindif.2023.102274>.
- [11] K. A. Gamage, S. C. Dehideniya, Z. Xu, and X. Tang, "ChatGPT and higher education assessments: more opportunities than concerns?," *Journal of Applied Learning and Teaching*, vol. 6, no. 2, 2023.
- [12] J. Rudolph, S. Tan, and S. Tan, "ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?," *Journal of Applied Learning and Teaching*, vol. 6, no. 1, 2023.
- [13] J. Finnie-Ansley, P. Denny, B. A. Becker, A. Luxton-Reilly, and J. Prather, "The robots are coming: Exploring the implications of openai codex on introductory programming," in *Proceedings of the 24th Australasian Computing Education Conference*, 2022, pp. 10-19.
- [14] D. Jonassen, J. Strobel, and C. B. Lee, "Everyday problem solving in engineering: Lessons for engineering educators," *Journal of engineering education*, vol. 95, no. 2, pp. 139-151, 2006.
- [15] D. H. Jonassen and W. Hung, "All Problems are Not Equal: Implications for Problem-Based Learning," *Interdisciplinary Journal of Problem-based Learning*, vol. 2, no. 2, p. 4, 2008.
- [16] N. Shin, D. H. Jonassen, and S. McGee, "Predictors of well-structured and ill-structured problem solving in an astronomy simulation," *Journal of Research in Science Teaching*, vol. 40, no. 1, pp. 6-33, 2003, doi: <https://doi.org/10.1002/tea.10058>.
- [17] C. L. Dym, P. Little, E. J. Orwin, and E. Spjut, "Engineering Design: A Project-Based Introduction," 2009.

- [18] S. McCahan, P. Anderson, M. Kortschot, P. E. Weiss, and K. A. Woodhouse, *Designing engineers: an introductory text*. John Wiley & Sons, 2015.
- [19] J. Hattie and H. Timperley, "The power of feedback," *Review of educational research*, vol. 77, no. 1, pp. 81-112, 2007, doi: <https://doi.org/10.3102/003465430298487>.
- [20] B. Wisniewski, K. Zierer, and J. Hattie, "The power of feedback revisited: A meta-analysis of educational feedback research," *Frontiers in psychology*, vol. 10, p. 3087, 2020.
- [21] M. Theobald and H. Bellhäuser, "How am I going and where to next? Elaborated online feedback improves university students' self-regulated learning and performance," *The Internet and Higher Education*, vol. 55, p. 100872, 2022.
- [22] F. M. Van der Kleij, R. C. Feskens, and T. J. Eggen, "Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis," *Review of educational research*, vol. 85, no. 4, pp. 475-511, 2015.
- [23] E. Panadero, A. Lipnevich, and J. Broadbent, "Turning self-assessment into self-feedback," *The impact of feedback in higher education: Improving assessment outcomes for learners*, pp. 147-163, 2019.
- [24] D. Nicol, "The power of internal feedback: exploiting natural comparison processes," *Assessment & Evaluation in Higher Education*, vol. 46, no. 5, pp. 756-778, 2021/07/19 2021, doi: 10.1080/02602938.2020.1823314.
- [25] N. Winstone and D. Carless, "Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach," 2019.
- [26] D. Carless and D. Boud, "The development of student feedback literacy: enabling uptake of feedback," *Assessment & Evaluation in Higher Education*, vol. 43, no. 8, pp. 1315-1325, 2018.
- [27] N. E. Winstone, R. Ajjawi, K. Dirkx, and D. Boud, "Measuring what matters: the positioning of students in feedback processes within national student satisfaction surveys," *Studies in Higher Education*, vol. 47, no. 7, pp. 1524-1536, 2022.
- [28] T. Brown *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877-1901, 2020.
- [29] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [30] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38-45.
- [31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [32] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [33] B. Min *et al.*, "Recent advances in natural language processing via large pre-trained language models: A survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1-40, 2023.
- [34] S. Narciss *et al.*, "Exploring feedback and student characteristics relevant for personalizing feedback strategies," *Computers & Education*, vol. 71, pp. 56-76, 2014/02/01/ 2014, doi: <https://doi.org/10.1016/j.compedu.2013.09.011>.
- [35] Y. Qian and J. D. Lehman, "Using Targeted Feedback to Address Common Student Misconceptions in Introductory Programming: A Data-Driven Approach," *SAGE Open*, vol. 9, no. 4, p. 2158244019885136, 2019, doi: 10.1177/2158244019885136.
- [36] S. Basu, G. Biswas, and J. S. Kinnebrew, "Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment," *User Modeling and User-Adapted Interaction*, vol. 27, pp. 5-53, 2017.
- [37] L. Zhang, Y. Huang, X. Yang, S. Yu, and F. Zhuang, "An automatic short-answer grading model for semi-open-ended questions," *Interactive Learning Environments*, vol. 30, no. 1, pp. 177-190, 2022, doi: 10.1080/10494820.2019.1648300.



- [38] O. L. Liu, C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M. C. Linn, "Automated scoring of constructed-response science items: Prospects and obstacles," *Educational Measurement: Issues and Practice*, vol. 33, no. 2, pp. 19-28, 2014.
- [39] M. Pankiewicz and R. S. Baker, "Large Language Models (GPT) for automating feedback on programming assignments," in *Proceedings of the 31st International Conference on Computers in Education*, Matsue, Shimane, Japan, J.-L. Shih, A. Kashiara, W. Chen, and H. Ogata, Eds., December, 2023 2023, vol. 1: Asia-Pacific Society for Computers in Education (APSCE), pp. 68-77.
- [40] J. Nehyba and M. Štefánik, "Applications of deep language models for reflective writings," *Education and Information Technologies*, vol. 28, no. 3, pp. 2961-2999, 2023.
- [41] W. Dai *et al.*, "Can large language models provide feedback to students? A case study on ChatGPT," in *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 2023: IEEE, pp. 323-325.
- [42] B. Frank, N. Simper, and J. Kaupp, "Formative feedback and scaffolding for developing complex problem solving and modelling outcomes," *European Journal of Engineering Education*, vol. 43, no. 4, pp. 552-568, 2018.
- [43] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872-1897, 2020.
- [44] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, p. Article 62, 2021, doi: 10.1145/3439726.
- [45] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune bert for text classification?," in *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, 2019: Springer, pp. 194-206.
- [46] E. Musk, "Hyperloop alpha," *SpaceX: Hawthorne, CA, USA*, 2013.
- [47] F. Gilardi, M. Alizadeh, and M. Kubli, "Chatgpt outperforms crowd-workers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.
- [48] M. Bosley, M. Jacobs-Harukawa, H. Licht, and A. Hoyle, "Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research," 2023.
- [49] H. Liu *et al.*, "Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 1950-1965, 2022.
- [50] Z. Zhao, Z. Zhang, and F. Hopfgartner, "A Comparative Study of Using Pre-trained Language Models for Toxic Comment Classification," presented at the Companion Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 2021. [Online]. Available: <https://doi.org/10.1145/3442442.3452313>.
- [51] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, pp. 1-85, 2022.
- [52] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [53] S. Mohammadi and M. Chapon, "Investigating the Performance of Fine-tuned Text Classification Models Based-on Bert," in *2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 14-16 Dec. 2020 2020, pp. 1252-1257, doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00162.
- [54] K.-E. Chang, Y.-T. Sung, R.-B. Chang, and S.-C. Lin, "A new assessment for computer-based concept mapping," *Journal of Educational Technology & Society*, vol. 8, no. 3, pp. 138-148, 2005.