Leveraging Large Language Models to Automatically Investigate Core Tasks Within Undergraduate Engineering Work-Integrated Learning Experiences

Pauline Aguinalde
College of Education
University of Florida
Gainesville, USA
aaguinalde@ufl.edu

Jinnie Shin, Ph.D.
College of Education
University of Florida
Gainesville, USA
jinnie.shin@coe.ufl.edu

Bruce F. Carroll, Ph.D.
Department of Mechanical & Aerospace Engineering
University of Florida
Gainesville, USA
bfc@ufl.edu

Kent J. Crippen, Ph.D.
College of Education
University of Florida
Gainesville, USA
kcrippen@coe.ufl.edu

***Abstract*—This full research paper aims to investigate methods for systematically identifying core tasks within undergraduate engineering work-integrated learning (WIL) opportunities, such as internships and co-ops. It achieves this by automatically analyzing WIL opportunities using transformer models. A dataset of 4,833 engineering internship postings from the last ten years was obtained through a partnership with the University's Career Connections Center. From this, a subset of 374 aerospace engineering internships, yielding 1,913 unique job tasks, was extracted for human labeling. We applied the Llama 2 architecture, a sophisticated pre-trained LLM, to extract a list of specific responsibilities and tasks from the internship postings. The job tasks were used to train an automated classification system to map each task to the established seven ABET student outcomes. Each job task was human-labeled by three subject matter experts, achieving a high level of inter-rater reliability of 0.998, according to Krippendorff's alpha. RoBERTa resulted in the optimal model indicating a label ranking average precision of 0.892 on the validation set and 0.857 on the testing set. Our findings provide novel insights into understanding the evolving skill expectations of undergraduate interns, offering a basis for tailoring engineering education to address these demands. Furthermore, the automated analysis of internship tasks demonstrates the potential for a scalable way to address the gap in understanding the core responsibilities within WIL experiences.**

Keywords—work-integrated learning, internships, natural language processing, aerospace engineering, ABET

I. INTRODUCTION

Work-integrated learning (WIL) experiences, such as internships or co-ops, provide educational opportunities that equip students for their future careers. These opportunities offer numerous benefits for undergraduate engineering students, as demonstrated in previous research. Students who participate in internships often achieve higher grades and enhanced academic competencies, including writing, research, and practical application of theory to practice [1,2,3]. Internships have been shown to increase the likelihood of engineering students graduating in engineering by 10% [3]. Furthermore, engaging in internship experiences benefits the student professionally, positively influencing their readiness to work, as well as the acquisition of technical, and non-technical competencies in engineering [1,4,5]. Thus, participation in WIL experiences is widely recognized by both institutions and students as a beneficial endeavor during undergraduate engineering programs [6].

Despite the importance of WIL experiences in improving academic and practical outcomes in engineering education, the understanding of tasks conducted by student interns remains ambiguous [7]. The lack of clarity regarding specific tasks in internships poses several challenges, particularly in effectively aligning students' learning experiences across industry and academic domains. This may indicate that WIL experiences are poorly defined and misaligned with the curriculum [7]. Furthermore, studies [7,8] have shown that engineering students often lack professional skills, such as communication, teamwork, and leadership. Even more worrisome, these skills are frequently perceived by students as less important than hard skills, such as technical knowledge. This perception highlights the need to define the exact skill expectations within industry settings. Given the ambiguity surrounding the tasks assigned to engineering interns and the challenges this poses for aligning learning experiences with academic and industry demands, it is essential to systematically analyze these tasks.

Investigating the core tasks within WIL experiences presents significant challenges. Mainly, there is a substantial amount of these opportunities available ubiquitously, posing difficulties in comprehensively analyzing each of them manually. Additionally, the rapidly evolving nature of industry expectations over time can pose challenges in maintaining a relevant and current understanding of the core tasks. Thus, the process of understanding these tasks must be systematized to extract meaningful insights. Recent studies have introduced natural language processing (NLP) approaches to understand job postings that include details such as descriptions, skills, or requirements [9,10]. Techniques such as text classification where job postings are

sorted based on their attributes and job tasks, have been utilized in the postdoctoral engineering field [9] as well as in a general job search context [10]. The latter utilized deep learning approaches, particularly transformer models to accomplish this aim, resulting in an advanced framework to investigate jobs on a larger scale with a mean reciprocal rank of 0.905.

Hence, the purpose of this study is to investigate methods for systematically identifying core tasks within undergraduate engineering WIL opportunities using transformer models. Identifying core tasks can aid in aligning students' perceptions with industry expectations. By utilizing transformer models to support this aim, we can efficiently process and analyze WIL data at a large scale. Furthermore, this study is a component of a broader project aimed at optimizing an undergraduate engineering student's involvement in WIL experiences. Two research questions are addressed to guide the study as below.

RQ1. Can standardized WIL tasks, based on the ABET criteria [11], effectively represent a wide variety of employer-defined job descriptions?

RQ2. Which transformer models among BERT, RoBERTa, and DeBERTa, acquire the highest classification performance for core tasks within undergraduate engineering WIL opportunities?

II. Related Work

A. Work Integrated Learning

WIL experiences have been an integral component of education for many undergraduate students. In engineering education, these experiences are often completed in the form of internships and co-ops, extending learning beyond conventional academic coursework by allowing students to integrate their foundational engineering knowledge from their courses into the workplace [8]. Numerous studies have found that students particularly benefit from engineering WIL experiences due to their improved transition to full-time employment [7,8]. Involvement in WIL has been found to result in several positive learning outcomes, such as higher grades, increased perceived job readiness [2,4], and gains in both technical and non-technical skills in the workplace [1].

Recognizing the significant advantages of WIL experiences, it is important to further explore how these experiences translate into employable skills in the engineering field. Research in the WIL context indicates that communication, technical analysis, technical writing, and teamwork are highly valued attributes according to industry and faculty perceptions [12]. Conversely, expectations that WIL supervisors have for their interns tend to be less clear. Many supervisors have not received explicit training in managing interns, leading to varied expectations for these students [7]. Moreover, [13] conducted a deductive and inductive analysis of job postings aimed at undergraduate engineering students and found that employers prioritize different information literacy proficiencies from potential hires in contrast to academia. In particular, they found that 70% of the expected information was on standards and codes, whereas 21% were related to laws and regulations. Since many companies view WIL programs as a means to secure potential full-time hires, understanding the core tasks in undergraduate engineering WIL experiences would help both institutions and potential interns align with industry expectations for these roles.

B. Transformer Models

Text classification methods are widely adopted to categorize various types of text into distinct groups [14]. In recent years, numerous sophisticated models have been developed and utilized for this task, evolving from shallow learning approaches to deep learning algorithms [15]. In the context of understanding job postings and job tasks, text classification methods have been implemented for various purposes. Recent studies [9] evaluated the job expectations from postdoctoral engineering and computer science job postings, by mapping them to the Knowledge, Skills, and Attributes (KSA) framework. To do this, they utilized NLP preprocessing techniques such as NLTK to extract textual features. Results from their analysis indicated that the top three commonly identified KSAs are communication (69.9% of the data), academic writing (61.4% of the data), and leadership skills (34.5% of the data). Similar work [10] employed a multilabel classification approach in their job analysis and used various deep learning models such as neural networks and transformers. This enabled them to identify required skills from job descriptions in a generalized context, with their results indicating a recall of 92.24 based on the top 100 skills in their prediction.

There is a gap in the literature regarding specifically mapping job tasks to their associated categories using transformer models. A literature review by [16] found that the use of deep learning models was mainly for matching descriptions in job ads to their corresponding job titles. Additionally, they found that studies investigating skills within job ads utilize unsupervised techniques (e.g., latent Dirichlet allocation, latent semantic indexing, etc.) for their analyses. [17] evaluated the capacities of transformer-based models to match job descriptions and candidates' resumes, achieving close to 80% accuracy. Similarly, [18] conducted a case study to investigate the task of job classification using traditional machine-learning models, such as support vector machines (SVMs), and transformer models like DeBERTa. Their findings showed a precision close to 0.89 at a recall of 0.95, outperforming many other large language models, such as davinci-002, 003, and GPT-3.5.

Employing transformer models for categorizing job tasks expands their current use of matching job ads to job titles. Adapting them in the context of mapping skills within job postings to predefined categories contributes to the understanding of expected skills for undergraduate engineering WIL experiences. Furthermore, pre-trained large language models are widely used as they have been trained on large corpora. They can then be fine-tuned to adapt them to various tasks such as text classification. Specifically, transformer models are used to model the structure of text based on contextual information input [10]. For this study, encoder-based transformer models were selected—particularly, variations of BERT—BERT, RoBERTa, and

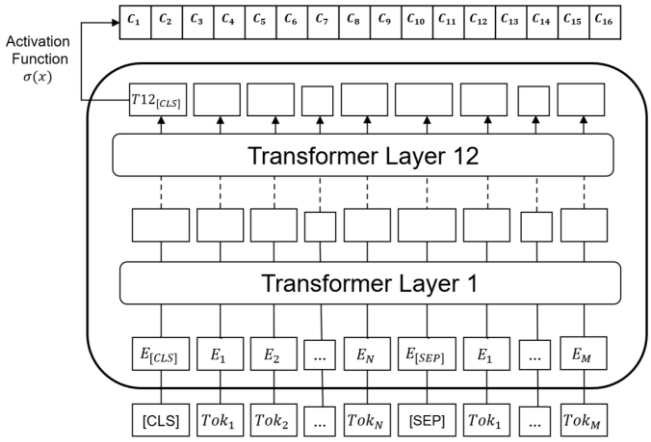

Fig. 1. Transformer Model Architecture.

DeBERTa. To visualize the overarching structure of the models, Figure 1 presents a generalized representation of their architecture. The selection of each model in the study was based on their varying pretraining procedures, allowing us to evaluate which performs best in the context of job task classification.

To actualize the job task category classification process, each response is treated as an input, undergoing a bidirectional self-attention layer, which outputs a corresponding likelihood value of belonging to a given label [19]. The bidirectionality self-attention layer considers the semantic context of a given word from both the left and right sides. BERT was initially trained for two tasks: masked-word prediction and next-sentence prediction. Although RoBERTa and DeBERTa are variations of BERT, they all employ some form of masked-language-modeling, typically on 15% of all the tokens. Distinctively, RoBERTa [20] uses dynamic masking where the masked tokens are randomized at each epoch, allowing it to handle textual variations. In contrast, DeBERTa uses an enhanced mask decoder to take absolute positions into account, in addition to the relative position, in order to have a more nuanced understanding of the relationships within the text. In next-sentence prediction, the model is given a pair of sentences, and it attempts to predict whether the second came following the first. These approaches capture textual patterns and relationships to enable multilabel classification of job tasks. Each text input (e.g., job task) initially undergoes a vectorization and tokenization process. Segment and position embeddings are then added. DeBERTa slightly differs in that content and position vectors for each token are separated, resulting in each vector having separate corresponding weights. The job task will thus begin with a [CLS] token (to be used for classification), followed by embedded tokens, and include a separating token [SEP] to indicate each sentence segment. The [CLS] token is utilized to represent the text input and fed into the hidden layers to output the corresponding categories it belongs to.

Following the embeddings, the resulting vectors pass through each model's 12 transformer layers. BERT, RoBERTa, and DeBERTa [21] have been pre-trained on substantial amounts of data, with RoBERTa being trained with the largest corpus. Each of the models then attempts to predict the masked tokens as it learns the contextual dependencies of each response. These approaches yield diverse methodologies in capturing textual patterns and relationships within the job tasks to classify them according to their category.

III. METHODS

To understand the core tasks within undergraduate engineering WIL experiences, a job posting dataset was collected in partnership with the University's Career Center. A list of unique tasks from these postings was extracted and manually labeled to correspond to categories based on the seven student outcomes defined as Criterion Three by the Accreditation Board for Engineering and Technology (ABET), which is a non-profit accreditation organization focused on college and university programs. This extracted and labeled dataset was then used to train and fine-tune BERT, RoBERTa, and DeBERTa to automatically align engineering WIL tasks with the established outcomes. The models were evaluated using a selection of metrics appropriate for the multilabel context. Figure 2 provides an overview of the methodology implemented in this study.

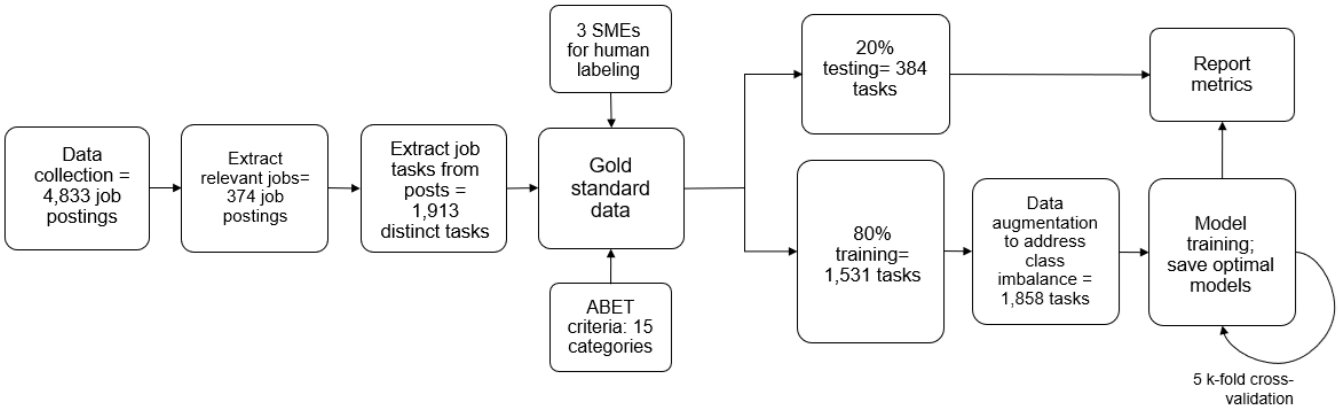

Fig. 2. A Conceptual Representation of the Analysis Framework.

A. Data Collection

A total of 374 internship postings for engineering students, particularly those targeting aerospace engineering students, were analyzed in this study. This dataset was identified from a total of 4,833 job postings collected since 2014 and was obtained in collaboration with the University's Career Connections Center. It was derived from the online employment social network platform exclusive to university members, where recruiters and companies connect with potential hires. Each job posting consisted of the job title, position type (i.e., co-op or internship), the posting date, desired majors, and a description– each provided by their respective companies. The content of the description showed considerable variation across postings due to the unstructured format of the subsection. Namely, descriptions included desired qualifications, expected responsibilities/tasks, company details, compensation, or links to access additional information. From this, a subset of 374 aerospace engineering internship postings were extracted for processing.

B. Data Preprocessing

We extracted a total of 1,913 specific job tasks from the internship postings. These tasks were labeled into 15 categories. We first extracted the description section of the aerospace internship postings. The lengthy descriptions included a wide range of information irrelevant to the tasks expected for the internship. For the purpose of this study, only the specific responsibilities associated with the role are to be utilized. We used Llama 2, a sophisticated pre-trained LLM, which was applied to first summarize the description into a list of specific responsibilities and tasks from the internship. This yielded a total of 1,913 unique entities of tasks for human labeling. An example of the output from Llama and its processed output is shown in Table I.

TABLE I. PREPROCESSING OUTPUTS

Llama 2 output	Based on the job posting, the following are the job responsibilities: 1. Production of CADD drawings. 2. Perform field work (evaluation of existing sites) as directed by senior engineer/designer. 3. Participation in electrical design, including power distribution, lighting, fire alarm, and low-voltage building systems. 4. Assists with engineering duties on projects of various complexity. 5. Performs other duties as assigned by the senior engineering staff.
Preprocessed output	['production of cadd drawings', 'perform field work (evaluation of existing sites) as directed by senior engineer/designer', 'participation in electrical design, including power distribution, lighting, fire alarm, and low-voltage building systems', 'assists with engineering duties on projects of various complexity', 'performs other duties as assigned by the senior engineering staff']

A total of 15 distinct categories were identified for assigning tasks (see Table I). To prepare a general list of labels that captures the job tasks, we adopted the ABET outcomes. The Engineering Accreditation Commission of ABET[1] has established criteria emphasizing high standards in engineering education. These standards ensure that students are well-prepared for the profession [22]. By utilizing the ABET outcomes to investigate core tasks, WIL tasks are systematically mapped to existing high-quality standards that reflect the ever-evolving skill expectations in the industry.

Three subject matter experts (SMEs) were then tasked with labeling each item according to its corresponding category or categories. We also added a label of 'other' for tasks that do not align with the predefined categories, as well as "not a task" for handling irrelevant outputs. To ensure reliability among the raters, Krippendorff's alpha was calculated [23], achieving a high level of 0.998. After this process, a gold standard dataset of the labeled data was derived by taking the majority vote among the SMEs.

Table II presents the fifteen categories to which job tasks belong. The most common tasks across the dataset involve conducting engineering design and development tasks, collaborating with an engineering or multidisciplinary team, creating written documents, developing or modifying code, and conducting experiments. The completed gold standard dataset was then divided into an 80/20 training and test set split. The training set was used for cross-validation. The validation set performance is derived from this cross-validation set. Alternatively, the test set was set aside as unseen data throughout the entire training process to ensure its independence from the model's learning process during training.

TABLE II. LABEL INSTANCES PER CATEGORY

Category	N	%
Conduct engineering design and development Tasks (ABET Outcome 2 and 4)	219	11.448
Analyze the operation or functional performance of a component or system (ABET Outcome 1 and 2)	74	3.868
Perform thermal science or fluid dynamic Analysis (ABET Outcome 1 and 2)	35	1.830
Perform solid mechanics analysis (ABET Outcome 1 and 2)	31	1.620
Perform dynamics or vibrational analysis	22	1.150

[1] https://www.abet.org/about-abet/

(ABET Outcome 1 and 2)		
Perform control analysis (ABET Outcome 1 and 2)	47	2.457
Create or revise technical drawings (ABET Outcome 3)	37	1.934
Develop or modify computer codes and/or public software including CAD, CAM, FEA, CFD, Matlab, C++, Python, or related computational tools (ABET Outcome 1)	169	8.834
Conduct experimental programs, including the testing of prototypes, components, hardware, or Products (ABET Outcome 6)	166	8.677
Conduct manufacturing activities or processes (ABET Outcome 2)	83	4.339
Conduct quality control activities or troubleshooting a failure of a component (ABET Outcome 1 and 2)	83	4.339
Create written documentation of procedures, processes, or results (ABET Outcome 3)	171	8.939

Furthermore, to address a mean imbalance ratio per label (IRLb) of 3.628 and a coefficient of variation imbalance ratio of 0.837—identifying five of the fifteen categories as belonging to the minority, using IRLb > Mean IRLb as the threshold—a strategy was developed [24]. To address this imbalance, a data augmentation approach was conducted as outlined by [25] using the python library *nlpaug 1.1.11*. In particular, the back translation function was utilized. This method translates the input text from its original language to a second language, and then translates it back to the original. In this case, each job task was translated from English to German and back to English, creating a slightly varied version of the original text while preserving contextual meaning. Augmented versions of tasks were generated using this technique for the minority labels until the threshold IRLb for all minority instances reached a value lower than the original mean of 3.628. The more balanced dataset was then used for further model development and fine-tuning.

C. Model Development and Fine-Tuning

The pre-trained models BERT, RoBERTa, and DeBERTa were fine-tuned using the *AutoModel* class from the Hugging Face Transformers library to classify the engineering tasks into the predefined 15 tasks. To train the models, a system with a NVIDIA GeForce RTX 3060 GPU was utilized. To begin the training process, only the training subset was used. From the dataset, the labels not associated with the tasks (i.e., "other" and "not a task") were joined into one label to streamline the categories and focus on the predefined outcomes. Each task was tokenized using the *AutoTokenizer* from Hugging face to match the corresponding model architecture. The trainer was developed employing binary cross-entropy (BCE) loss for the multilabel classification context. The 5-fold cross-validation technique was set up to promote a more robust model across subsets of the data.

To identify optimal parameters for the models, a grid search was utilized. The search iterated through variations in the learning rate (1e-5, 3e-5, 5e-5, 7e-5), batch size (32, 64), and weight decay (0.01, 0.001). The epoch was set to 20 throughout the process. The optimal models were then saved for evaluation using the previously reserved validation set.

D. Model Evaluation

Metrics used in our multilabel classification task involved the Label Ranking Average Precision (LRAP), Matthew Correlation Coefficient (MCC), and the Binary Cross-Entropy (BCE) loss.

LRAP. is a metric utilized in multilabel classification contexts, which calculates the average precision score for each sample based on ranking of relevant labels, as shown in (1), where L_{ij} represents the number of correct predictions with a score greater or equal to the label being evaluated (2), and the $rank_{ij}$ indicates the number of labels with a predicted score greater than the current label's score (3).

$$LRAP(y,\hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{samples}-1} \frac{1}{||y_i||_0} \sum_{j:y_{ij}=1} \frac{|L_{ij}|}{rank_{ij}} \quad (1)$$

$$L_{ij} = \{k: y_{ik} = 1, \widehat{y_{ik}} \geq \widehat{y_{ij}}\} \quad (2)$$

$$rank_{ij} = |\{k: \widehat{y_{ik}} \geq \widehat{y_{ij}}\}|, \quad (3)$$

LRAP assesses how well the model ranks the correct labels for a given instance—investigating whether higher-ranked labels were true labels, ranging from 0-1—with 1 indicating optimal performance.

MCC. The MCC is a method of calculating the correlation to evaluate the quality of a prediction to an observation, summarized by equation (4). Its value ranges from [-1,1], with 1 indicating a perfect agreement among the predicted label and the true label, 0 comparable to random predictions, and -1 indicating total disagreement.

$$MCC = \frac{TPxTN - FPxFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (4)$$

There are two ways to report MCC in a multi-label classification context, proposed by [26]. The micro-average MCC flattens the labels and predictions into a single binary classification problem. This reflects the model performance across all job task categories combined, providing a single MCC value. On the other hand, macro-average MCC treats each label as one binary classification task, taking into account the classification performance for each category. The average across categories is then calculated to reflect the overall performance. Results from the macro-average MCC weights each category equally in the calculation, addressing imbalance within the distribution of the labels.

BCE. has been recognized to be effective in the multi-label classification context. The objective function is a minimization problem defined in equations (5-6). BCE loss function, combined with a sigmoid activation function, calculates the error by measuring the difference between the predicted probabilities of belonging to a category with the true labels.

$$min_\theta \; -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{L}[y_{ij} log(\sigma(f_{ij})) + (1-y_{ij}) log(1-\sigma(f_{ij})) \quad (5)$$

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (6)$$

IV. RESULTS

Tables III and IV showcase the results of the training and the test set, respectively. The best RoBERTa model was trained using a learning rate of 710-5, a batch size of 32, and a weight decay of 0.01. RoBERTa outperformed on every metric, with a LRAP of 0.892, micro-MCC of 0.781, macro-MCC of 0.806, and a BCE loss of 0.107 for the validation set. Performances of the BERT and DeBERTa models follow closely behind, with BERT producing better results for the micro-MCC (0.773 vs. 0.768 for DeBERTa), macro-MCC (0.791 vs. 0.788 for DeBERTa), and BCE loss (0.111 vs. 0.130 for DeBERTa). Conversely, DeBERTa achieved a higher LRAP (0.886 vs. 0.883 for BERT).

RoBERTa's LRAP of 0.892 reflects a strong predictive performance, indicating that the model ranks the correct job categories higher for a given job task proficiently. That is, it ranks a job task's true categories higher than irrelevant tasks. Moreover, a micro-MCC of 0.781 indicates a strong overall model performance. Similarly, the macro-MCC of 0.806 reflects strong model performance as an aggregation of each category. Compared to micro-MCC which treats the multi-label classification task as a single binary problem (i.e., whether it classified a given task to the correct categories or not), the macro-MCC indicates that the model prediction for belonging to a category consistently correlates highly across each of the 15 categories since they are weighted equally for the average calculation. The low error calculated from the BCE loss further supports the preceding metrics indicating that the model's probability outputs for each category correspond to the task's true category associations.

Likewise, RoBERTa outperformed the other models on every metric on the test set. It maintained a strong ability to rank relevant categories higher than irrelevant ones with an LRAP of 0.857. Performance on the micro and macro-MCCs indicate a moderate classification ability overall and across categories, with 0.605 and 0.559, respectively. The BCE loss indicates that the model predicts categories for the job tasks fairly close to their true categories. Conversely, DeBERTa's performance lags behind with lower LRAP, micro and macro-MCCs (0.850, 0.590, and 0.539, respectively). Additionally, its BCE indicates a relatively worse performance with 0.183. BERT's performance on the test set falls slightly above DeBERTa on the first metric, with an LRAP of 0.852, and below it with a micro-MCC of 0.588, macro-MCC of 0.532, while its BCE loss matched the RoBERTa model with 0.149.

TABLE III. MODEL PERFORMANCE ON VALIDATION SET

Model	**Metrics**			
	LRAP	***Micro-MCC***	***Macro-MCC***	***BCE loss***
BERT	0.883	0.773	0.791	0.111
RoBERTa	0.892	0.781	0.806	0.107
DeBERTa	0.886	0.768	0.788	0.130

TABLE IV. MODEL PERFORMANCE ON TEST SET

Model	**Metrics**			
	LRAP	***Micro-MCC***	***Macro-MCC***	***BCE loss***
BERT	0.852	0.588	0.532	0.149
RoBERTa	0.857	0.605	0.559	0.149
DeBERTa	0.850	0.590	0.539	0.183

V. CONCLUSION

The purpose of this study was to investigate the use of transformer models as a way to systematically identify core tasks within undergraduate engineering WIL experiences. To address RQ1, three SMEs evaluated the development of the general list of 15 job tasks for aerospace engineering students based on the ABET outcomes. The results indicated that close to 12% of the tasks, which are 232 out of 1,913 unique tasks, were labeled as the "other" category. In other words, this indicates that close to 88% of the tasks were successfully classified into the fifteen general categories informed by the ABET outcomes (see Table II). Specific examples of other tasks include "ensuring compliance with safety and environmental policies" and "setting targets and leading projects from product planning to production launch." The first example may directly relate to ABET outcome 5, highlighting the curricular importance of industry- or discipline-specific content, such as attention to engineering standards and codes, public safety and health, and the implications for the organization and society. Still, our 15 tasks covered the majority, close to 88% of the engineering tasks. We found that specific tasks, such as conducting engineering design and development tasks (ABET Outcomes 2 and 4), and creating written documentation of procedures, processes, or results (ABET Outcome 3), were frequently included in internship postings. This aligns with the findings of [13], which indicated that one of the primary job categories involved engineers "communicating their information through different sources, such as written documents and oral communication."

To address RQ2, our results indicated that the best-performing model was RoBERTa—which exceeded the others on all evaluation metrics used in the study. Results from the model training process indicated that the use of transformer models for classifying job tasks into their corresponding categories is effective. Additionally, the results from the model indicated that it is capable of ranking relevant categories for each task higher than irrelevant ones based on its LRAP score. Additionally, its moderate to high micro and macro-MCC scores and the low BCE loss demonstrate its ability to predict a given job task's true categories. RoBERTa has been identified to showcase comparable performance accuracies in complex multilabel text classification research. Oftentimes, DeBERTa showcased improved accuracy compared to RoBERTa. [27] showcased that in six out of the seven large-scale multi-label text classification datasets, DeBERTa outperformed RoBERTa by a small margin. The inconsistency between these findings and our experiments, where RoBERTa

performed the best, may largely be due to issues related to overfitting with relatively smaller sample sizes, as indicated by [28]. Further analyses to elucidate the different conditions where RoBERTa performs better than DeBERTa may be critical.

Results from this study show that standardized WIL tasks can represent employer-defined tasks within their indicated job descriptions, as well as indicate the frequency at which they are prevalent in job postings. Moreover, our model shows potential in automatically classifying these core tasks within the context of aerospace engineering. These findings are beneficial to support engineering institutions in the further enhancement of the curriculum. By reflecting the evolving nature of task expectations, students can be prepared for the workplace.

While this study was conducted thoroughly to eliminate limitations, we acknowledge that future research could help improve the generalizability and interpretability of our findings. This study focused on job postings specifically aimed at undergraduate aerospace engineering students. Future research can include a wider array of engineering disciplines to provide a more comprehensive understanding of core tasks within engineering WIL experiences. Furthermore, our dataset spanning the last ten years may not precisely reflect the most recent/current expectations of tasks within aerospace engineering. Utilizing a large dataset with a scope that is within a more recent time frame as opposed to ten years may be beneficial for a more modern understanding.

Additionally, conducting an error analysis by investigating confusion matrices and qualitatively examining misclassified job tasks in future research may provide more insight into optimizing the model further and identifying areas of improvement. Findings from this study benefit both students and institutions in matching their perceptions with expectations within industry. Moreover, this study is a component of a larger project aimed at optimizing WIL experiences. With about 600,000 undergraduate students enrolled in an engineering program every year [29], gaining a clear understanding of the competencies expected within the industry is essential. Aligning perceptions with expectations ensures that students can be better prepared for WIL experiences that contribute to their professional development.

ACKNOWLEDGMENT

We thank Adam Sardouk, Brandon Bulnes, and Dylan Garrison, for contributing as subject matter experts in data labeling. This work was conducted with support from the University of Florida's Research Opportunity Seed Fund (ROSF) to support interdisciplinary research in the field of AI and engineering education.

REFERENCES

[1] L. Y. Y. Luk and C. K. Y. Chan, "Students' learning outcomes from engineering internship: a provisional framework," *Studies in Continuing Education*, vol. 44, no. 3, pp. 526–545, Sep. 2022, doi: 10.1080/0158037X.2021.1917536.

[2] S. A. Rolland, J. W. Jones, and G. Bunting, "The impact of a year in industry on academic outcomes in higher education (engineering)," *European Journal of Engineering Education*, vol. 48, no. 4, pp. 747–760, Jul. 2023, doi: 10.1080/03043797.2023.2194244.

[3] J. Main, B. Johnson, N. Ramirez, H. Ebrahiminejad, M. Ohl, and E. Groll, "A Case for Disaggregating Engineering Majors in Engineering Education Research: The Relationship between Co-Op Participation and Student Academic Outcomes," *International Journal of Engineering Education*, vol. 36, no. 1, pp. 170–185, 2020.

[4] M. E. Al-Atroush and Y. E. Ibrahim, "Role of Cooperative Programs in the University-to-Career Transition: A Case Study in Construction Management Engineering Education.," *The international journal of engineering education*, vol. 1, 2022.

[5] Rahdiyanta, D., Nurhadiyanto, D., & Munadi, S. (2019). The Effects of Situational Factors in the Implementation of Work-Based Learning Model on Vocational Education in Indonesia. International Journal of Instruction, 12(3), 307-324.

[6] S. Chopra and L. Golab, "Undergraduate Engineering Applicants' Perceptions of Cooperative Education: A Text Mining Approach.," *International Journal of Work-Integrated Learning*, vol. 23, no. 1, pp. 95–112, 2022.

[7] S. M. Zehr and R. Korte, "Student internship experiences: learning about the workplace," *ET*, vol. 62, no. 3, pp. 311–324, Feb. 2020, doi: 10.1108/ET-11-2018-0236.

[8] P. Ackerman and K. Arcieri, "Co-ops are Great! but What are the Numbers Telling Us?," presented at the 2022 ASEE Annual Conference & Exposition, Aug. 2022, doi: 10.18260/1-2--40896.

[9] J. Zhu, E. Zerbe, M. Ross, and C. Berdanier, "The stated and hidden expectations: applying natural language processing techniques to understand postdoctoral job postings," presented at the 2021 ASEE Virtual Annual Conference Content Access, Jul. 2021, doi: 10.18260/1-2--37896.

[10] A. Bhola, K. Halder, A. Prasad, and M.-Y. Kan, "Retrieving Skills from Job Descriptions: A Language Model Based Extreme Multi-label Classification Framework," in *Proceedings of the 28th International Conference on Computational Linguistics*, Stroudsburg, PA, USA, 2020, pp. 5832–5842, doi: 10.18653/v1/2020.coling-main.513.

[11] ABET, "2023-2024 Criteria for Accrediting Engineering Programs," 2022. https://www.abet.org/accreditation/accreditation-criteria/criteria-for-accrediting-engineering-programs-2023-2024/ (accessed May 08, 2024).

[12] A. Eggleston, R. Rabb, and R. Welch, "Employer and Student Mismatch in Early-Career Skill Development," presented at the 2022 ASEE Annual Conference & Exposition, Aug. 2022, doi: 10.18260/1-2--40895.

[13] M. Phillips, D. Zwicky, and J. Lu, "Initial study of information literacy content in engineering and technology job postings," in *2020 IEEE Frontiers in Education Conference (FIE)*, Oct. 2020, pp. 1–3, doi: 10.1109/FIE44824.2020.9274195.

[14] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning--based Text Classification," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, Jun. 2021, doi: 10.1145/3439726.

[15] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: from text to predictions," *Information*, vol. 13, no. 2, p. 83, Feb. 2022, doi: 10.3390/info13020083.

[16] I. Rahhal, I. Kassou, and M. Ghogho, "Data science for job market analysis: A survey on applications and techniques," *Expert Syst. Appl.*, vol. 251, p. 124101, Oct. 2024, doi: 10.1016/j.eswa.2024.124101.

[17] C. Li, E. Fisher, R. Thomas, S. Pittard, V. Hertzberg, and J. D. Choi, "[2011.02998] Competence-Level Prediction and Resume & Job Description Matching Using Context-Aware Transformer Models," *arXiv*, Nov. 2020.

[18] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, "Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification," in *Natural language processing and information systems: 28th international conference on applications of natural language to information systems, NLDB 2023, derby, UK, june 21–23, 2023, proceedings*, vol. 13913, E. Métais, F. Meziane, V. Sugumaran, W. Manning, and S. Reiff-Marganiec, Eds. Cham: Springer Nature Switzerland, 2023, pp. 3–17.

[19] 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.

[20] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach.," *arXiv preprint arXiv:1907.11692*, 2019.

[21] P. He, X. Liu, J. Gao, and W. Chen, "Deberta: Decoding-enhanced bert with disentangled attention.," *arXiv preprint arXiv:2006.03654*, 2020.

[22] R. M. Felder and R. Brent, "Designing and teaching courses to satisfy the ABET engineering criteria," *J. Eng. Educ.*, vol. 92, no. 1, pp. 7–25, Jan. 2003, doi: 10.1002/j.2168-9830.2003.tb00734.x.

[23] K. Krippendorff, "Computing Krippendorff's Alpha -—Reliability,"

2011.

[24] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, “MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation,” *Knowledge-Based Systems*, vol. 89, pp. 385–397, Nov. 2015, doi: 10.1016/j.knosys.2015.07.019.

[25] Y. Shi *et al.*, “Improving Imbalanced Learning by Pre-finetuning with Data Augmentation,” in *Proceedings of Machine Learning Research*, Oct. 2022, pp. 68–82.

[26] C.-X. Li, “Exploiting label correlations for multi-label classification,” Master thesis, 2011.

[27] L. Galke *et al.*, “Are We Really Making Much Progress in Text Classification? A Comparative Review,” 2023.

[28] J. Carreras Timoneda and S. Vallejo Vera, “BERT, roberta or deberta? comparing performance across transformer models in political science text,” *J. Polit.*, Apr. 2024, doi: 10.1086/730737.

[29] American Society for Engineering Education, “Profiles of Engineering and Engineering Technology,” 2022.