Fine-Tuning Language Models for Pedagogy-Aligned Lesson Plans in Cybersecurity Education

Samar Althagafi[1], Miada Almasre[2], Wafaa Alsaggaf[3], Lana Alshawwa[4]
Information Technology Department, King Abdulaziz University, Jeddah, Saudi Arabia[1,2,3]
Medical Education Department, King Abdulaziz University, Jeddah, Saudi Arabia[4]

***Abstract*—Lesson planning in cybersecurity is time-consuming and cognitively demanding, especially for less experienced instructors, and manual approaches often lack flexibility across courses and contexts. We present a framework for generating pedagogy-aligned lesson plans using a large language model, integrating measurable objectives (Revised Bloom's Taxonomy), explicit learning theories, and evidence-based teaching strategies. We constructed a domain-specific knowledge base for cybersecurity topics and organized it with sentence-level embeddings and KMeans clustering. A pretrained large language model (GPT-3.5) was then fine-tuned to produce lesson plans that follow this structure. On a held-out test set, the model achieved BLEU 73.5, ROUGE-1 82.2, ROUGE-L 78.2, and BERTScore F1 97.4, reflecting strong lexical and semantic fidelity to reference plans. Although the study is limited to a single academic program and relies primarily on automated metrics, the framework offers practical support for instructors by reducing preparation time, enhancing consistency, and ensuring alignment with pedagogical standards. Future work will expand the curricular scope and involve expert review and classroom validation to assess educational impact.**

***Keywords*—Fine-Tuning; large language models; lesson planning; cybersecurity**

I. Introduction

Lesson plans are essential instruments in education, providing teachers with a clear framework for delivering effective and organized instruction [1]. They outline learning objectives, teaching methods, classroom activities, instructional materials, and assessment strategies required to achieve educational goals [2], [3]. By establishing a structured approach, lesson plans enhance teaching effectiveness and ensure instructional consistency and clarity [3], [4]. Although their components may vary depending on curriculum, educational context, or teacher preference, the primary purpose of lesson plans remains the improvement of instructional quality and student outcomes [5], [1]. Aligning lesson plans with established learning theories, such as constructivism, cognitivism, and constructionism, further strengthens their effectiveness by grounding instruction in well-established principles of how learners acquire, process, and apply knowledge [6], [7]. In addition, the integration of contemporary teaching strategies, including project-based learning, problem-based learning, and flipped classrooms, promotes active engagement, collaboration, and deeper understanding [8], [9].

Within this broader educational context, cybersecurity presents distinctive challenges. As a discipline, it is concerned with safeguarding the confidentiality, integrity, and availability of digital systems [10]. Given its rapidly evolving nature, cybersecurity education must combine strong theoretical foundations with practical, hands-on learning experiences [7], [11]. Effective lesson plans in this domain must balance theory with application, incorporate up-to-date content, and prepare learners to respond to emerging threats [12]. However, lesson planning is often time-intensive and cognitively demanding, particularly for novice teachers with limited pedagogical experience [3], [5]. Traditional lesson planning methods, which rely heavily on manual preparation, can be inefficient and may lack the flexibility to meet diverse instructional needs [2].

Recent advances in large language models (LLMs) such as BERT [13], GPT-3/4 [14], [15], and LLaMA [16] have enabled high-quality, controllable text generation for educational applications [17]. Although pre-training on extensive text corpora provides broad language competence [18], [19], these models often show limited effectiveness in specialized domains. Fine-tuning addresses this limitation by adapting pre-trained models to domain-specific tasks, thereby improving contextual accuracy and alignment with specialized requirements [20], [21]. Such approaches have already been applied in areas such as legal rule classification [22], transportation safety [23], and language-specific text summarization [24].

The effectiveness of fine-tuned models is typically assessed using automatic evaluation metrics that capture different aspects of text quality. In this study, such metrics are employed to assess lexical overlap, structural similarity, and semantic alignment between generated and reference lesson plans, providing a reproducible basis for evaluating instructional content.

Despite recent progress, existing studies have not fully explored the use of fine-tuned large language models for generating structured lesson plans in technically specialized domains such as cybersecurity. This gap is particularly relevant given the growing demand for scalable, pedagogy-aligned, and domain-specific instructional resources. Rather than relying solely on prompt engineering or post-generation filtering, the present study adapts a pre-trained language model to generate cybersecurity lesson plans by embedding pedagogical structure within the data representation and model fine-tuning process. Accordingly, the study addresses the following research questions:

- RQ1. How accurately can a fine-tuned large language model generate structured lesson plans for cybersecurity education?
- RQ2. To what extent do the generated lesson plans align with established pedagogical standards and instructional design principles?

The contributions of this research are as follows:

1) This work designs and constructs a domain-specific lesson planning dataset for cybersecurity to support pedagogy-aware language model fine-tuning.
2) It proposes a structured framework that incorporates explicit pedagogical elements into the fine-tuning process of a large language model for lesson plan generation.
3) It provides an empirical evaluation demonstrating that pedagogy-aligned fine-tuning enables the generation of structurally consistent and instructionally valid lesson plans.

The remainder of this paper is organized as follows: Section II reviews related work, Section III describes the methodology, Section IV presents the results and discussion, and Section V concludes the paper.

II. Related Work

The rapid progress of large language models has generated significant interest in their role within teaching and learning. Recent surveys highlight both their potential to transform instructional practices and the challenges they introduce in terms of reliability, bias, and alignment with educational objectives [17], [20]. Kasneci et al. [17] emphasize the ability of these models to reduce teacher workload and broaden access to instructional resources, while also noting the importance of producing outputs that educators can trust and adapt. Similarly, Zhang et al. [20] report that instruction-tuned models show promising capabilities but remain highly sensitive to the design of adaptation strategies and evaluation protocols. Together, these studies provide valuable context for educational use of LLMs, while also indicating that effective deployment depends on careful consideration of pedagogical alignment rather than generic application alone.

Recent work has explored how large language models can support assessment and instructional content generation. Sridhar et al. [25] investigated the use of GPT-4 to generate learning objectives, demonstrating that model outputs can follow established instructional design frameworks such as Bloom's taxonomy when guided by structured prompts. Similarly, Scaria et al. [26] examined automated question generation across Bloom's taxonomy levels and validated the quality of exam-style questions through expert review. Latif and Zhai [27] applied a fine-tuned ChatGPT model to automated scoring, reporting performance comparable to human graders on large collections of student responses. Collectively, these studies highlight the effectiveness of LLMs in generating and evaluating individual instructional components. At the same time, they primarily focus on isolated elements such as objectives, questions, or assessment outcomes, and therefore provide limited insight into how these components can be coherently integrated into complete, pedagogy-aligned lesson plans.

Evaluating the quality of generated content requires reliable and reproducible measures, leading to the widespread adoption of automatic evaluation metrics in natural language processing. BLEU [28] remains a common standard for measuring precision-oriented n-gram overlap, while ROUGE [29] is frequently employed to capture recall-oriented coverage in summarization. More recently, BERTScore [30] has been introduced to provide a semantic perspective by comparing contextual embeddings, offering closer alignment with human judgments of similarity. Together, these metrics provide complementary views of lexical fidelity, content coverage, and semantic similarity, and their primary strength lies in enabling consistent and scalable evaluation of generated text. However, they implicitly assume that textual similarity is a sufficient proxy for instructional quality, which limits their ability to fully capture pedagogical effectiveness when used in isolation.

Beyond assessment and evaluation, prior research has examined lesson planning systems as a means of providing more direct support for teachers. Zheng et al. [1] proposed an automatic lesson plan generator that improves plan quality through iterative self-critique, while Hu et al. [2] evaluated GPT-4 generated teaching plans and reported benefits such as efficiency and creativity alongside limitations related to inconsistency and insufficient pedagogical grounding. Fan et al. [3] introduced *LessonPlanner*, an interactive system designed to help novice teachers structure lesson components more effectively. Earlier systems, including Strickroth's *Platon* [5] and retrieval-augmented planning approaches explored by Kloker et al. [31], further demonstrate the feasibility of digital tools for lesson organization. Collectively, these approaches highlight the potential of automated and interactive systems to support lesson planning; however, they are largely designed for general instructional contexts and provide limited support for domain-specific requirements, such as those found in cybersecurity education.

A growing body of research has examined model adaptation through fine-tuning as a means of producing more consistent and domain-sensitive outputs. Within educational contexts, Tsai et al. [32] fine-tuned BERT and GPT-2 for short-answer question generation and reported improvements in quality and diversity over template-based approaches. Jia et al. [33] embedded Gagné's nine events of instruction into a fine-tuned model to generate lesson plans grounded in established pedagogical theory, while Gharbi et al. [34] demonstrated the potential of fine-tuned models to support personalized learning at scale in MOOCs. Beyond education, fine-tuning has been successfully applied to legal rule classification [22], transportation safety [23], and domain-specific summarization [24], highlighting its versatility across application areas. While these studies confirm the effectiveness of fine-tuning for improving domain alignment, they largely address either isolated instructional tasks or general learning contexts, leaving limited exploration of fine-tuned systems that generate fully structured lesson plans for specialized domains such as cybersecurity, where alignment between theoretical concepts and practical application is essential.

III. Methodology

The study employed a four-phase methodology integrating pedagogical theory with computational modeling. The methodology was designed to incorporate pedagogical alignment through both dataset construction and model fine-tuning, ensuring that instructional structure and pedagogical constraints are embedded in the generation process rather than applied during subsequent evaluation. Phase 1 identified relevant learning theories and teaching strategies through a literature review

and expert validation. Phase 2 constructed the knowledge base through schema design, sample generation, and expert evaluation. Phase 3 refined the model through dataset clustering, balancing, and fine-tuning. Finally, Phase 4 evaluated model performance using established automatic metrics. The overall process is illustrated in Fig. 1.

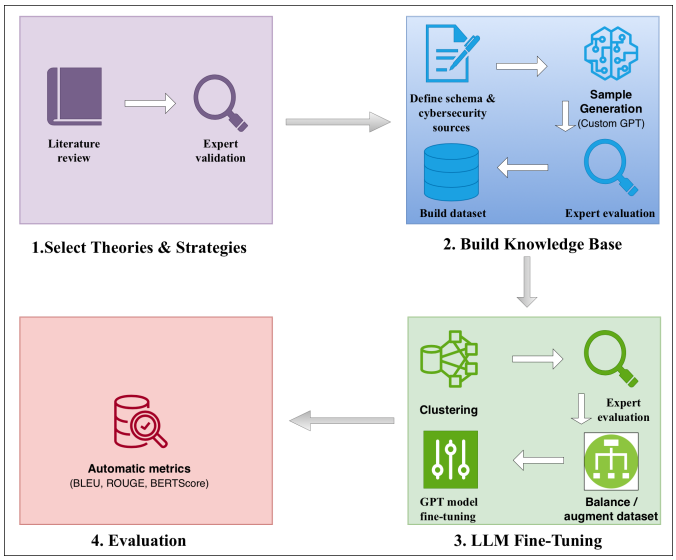


Fig. 1. Four-phase methodology pipeline.

A. Selection of Learning Theories and Teaching Strategies

This phase guided the design of lesson plans by identifying pedagogical foundations most relevant to cybersecurity education. A targeted review of pedagogical literature was conducted, complemented by consultations with an instructional design specialist. This process ensured that the selected learning theories and teaching strategies were both theoretically grounded and practically applicable to the demands of cybersecurity instruction, where adaptability and professional relevance are critical [7].

A range of learning theories were considered in shaping the instructional approach, with the following perspectives identified as particularly relevant.

Differential theories of learning transfer highlight that knowledge can only be successfully applied across contexts under specific conditions, such as having sufficient prior knowledge, possessing cognitive flexibility, and learning in environments similar to those of application [7]. Constructionist theories of learning emphasize creating tangible outcomes to build knowledge, promote creativity, encourage hands-on experimentation, and support iterative problem-solving [7]. The general cognitive science perspective emphasizes meta-cognitive control, analogical reasoning, and structured problem-solving as fundamental processes in learning; it supports instructional strategies that guide learners from basic comprehension to expert application through reflection and goal-oriented practice [7].

Cognitive-associationist theories focus on strengthening mental links between declarative, procedural, and conditional knowledge, with learning taking place incrementally through sequenced tasks that develop complex competencies [7]. Similarly, Cognitive-Constructivist theory emphasizes that learners actively build mental models through problem-solving, exploration, and reflection, integrating new knowledge with existing experience to produce higher-level cognitive schemas [7]. Coordinative transfer theory seeks to reduce negative transfer by improving learners' ability to determine when and how to apply existing knowledge [7].

More broadly, cognitive learning theories view learning as an active process of acquiring, organizing, and integrating knowledge, with a focus on internal mechanisms such as attention, memory encoding, and schema activation [35]. Experiential learning theory defines learning as a cyclical process involving concrete experience, reflective observation, abstract conceptualization, and active exploration [35]. Transformative learning theory emphasizes the importance of learners critically analyzing their preexisting assumptions and ideas in response to disorienting situations; through introspection and discussion, learners reframe their perspectives and develop more expansive and informed worldviews [35]. Finally, social learning theory affirms that people acquire skills and behaviors through social interaction, modeling, and observation, encompassing cooperation, peer review, and role modeling in education [35].

Building on these theories, a variety of teaching strategies were selected to translate abstract principles into practical methods that effectively support cybersecurity education, such as Problem-Based Learning (PBL), in which students collaborate to explore complex, real-world issues without predetermined answers. In this approach, teachers serve as facilitators, guiding students as they investigate, evaluate, and overcome obstacles [11], [9]. Similarly, Project-Based Learning (PrBL) involves students in long-term projects that apply cybersecurity concepts to real-world challenges. This strategy requires learners to create a tangible product or solution and promotes sustained engagement, teamwork, and iterative problem-solving [9].

Another approach is the use of practical exercises, which immerse students in real-world cybersecurity scenarios. Conducted in controlled physical or virtual environments, these sessions enable learners to develop technical competence, apply theoretical knowledge, and gain experience with industry-standard tools and protocols [36]. The Flipped Classroom model further complements this by requiring students to study core materials (e.g., videos, texts, or mini-lectures) before class, allowing in-class time to be dedicated to higher-order activities such as collaborative analysis, simulations, and practical problem-solving under instructor guidance. This method promotes self-paced study outside of class and deeper participation during class [9], [4].

Case-Based Learning extends this applied approach by using real-world cybersecurity cases to engage students in analyzing complex problems, comparing approaches, and proposing well-justified solutions. It encourages critical thinking and helps learners apply academic knowledge to practice [9]. Lecture-Based Learning also remains valuable for presenting fundamental cybersecurity concepts, where instructors can arrange content, provide expert framing, and establish conceptual baselines. When enriched with examples, questions,

and discussions, lectures can enhance engagement and comprehension, though they are most effective when combined with active learning techniques [9], [37].

To further enhance engagement, Role-Playing activities assign learners specific roles, immersing them in authentic cybersecurity scenarios. This approach improves empathy, decision-making, and communication, making it especially beneficial for developing soft skills and ethical reasoning in collaborative, high-pressure environments [38]. Finally, Simulation activities replicate real-world cybersecurity operations in secure, controlled environments, enabling students to practice technical skills, incident response, and decision-making under pressure [39].

Together, these learning theories and teaching strategies provide a solid foundation for designing cybersecurity instruction that is both grounded in established principles and responsive to real-world practice.

B. Build Knowledge Base

To fine-tune a large language model for generating cybersecurity lesson plans, a standardized structure was developed based on instructional design concepts. Each plan included five components: course title, topic, learning objectives (based on Revised Bloom's Taxonomy [40]), learning theory, and teaching strategy. Theories and strategies were taken from the previously presented curated lists, ensuring that they related to known educational frameworks. Because no existing dataset combined these features with cybersecurity content, a customized knowledge base was created using courses from the Diploma in Cybersecurity program at King Abdulaziz University (KAU), thereby ensuring both disciplinary relevance and educational validity.

1) Custom GPT for Knowledge Base Generation: To construct a pedagogically grounded knowledge base, OpenAI's Custom GPT was configured as a dedicated instance of ChatGPT. Custom GPT leverages prompt engineering and metadata settings to control tone, structure, and output format [41]. Within this setup, the model was instructed to generate structured instructional entries that could later serve as training data for fine-tuning a larger model specialized in lesson plan creation. For each course title and topic, the system consistently generated three elements: learning objectives, a learning theory, and a teaching strategy all defined according to established instructional design parameters. In particular, the Custom GPT created three measurable objectives for each topic based on Revised Bloom's Taxonomy [40]. Best practices for 60-90 minute sessions were followed to maintain focus and manage cognitive load [42].

2) Sample Generation and Expert Evaluation: To evaluate the instructional quality of the Custom GPT, a pilot dataset was created using one course from the cybersecurity diploma program. The course included nine topics, and for each, the model generated a structured lesson plan with three components: measurable learning objectives, an aligned learning theory, and a related teaching strategy. Eight professionals evaluated this dataset, including three curriculum and pedagogy specialists and five classroom-experienced cybersecurity instructors.

Each group received a tailored evaluation form with shared core criteria and role-specific emphasis. This ensured that the lesson plans were assessed for both educational quality and practical application. Educational experts examined eight criteria, including the clarity, measurability, alignment, and cognitive complexity of objectives; the appropriateness of the selected learning theory; the suitability of the teaching strategy; and the overall coherence of the topic, objectives, theory, and method. Cybersecurity instructors employed the same criteria, excluding the theoretical dimension.

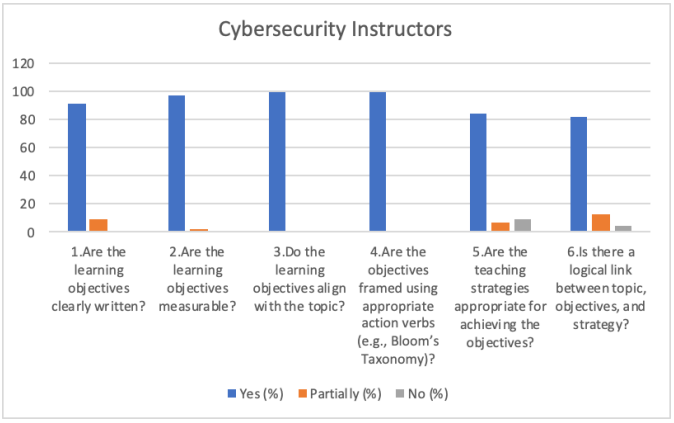


Fig. 2. Evaluation results from cybersecurity instructors across nine topics: percentage distribution of yes, partially, and no responses for six criteria.

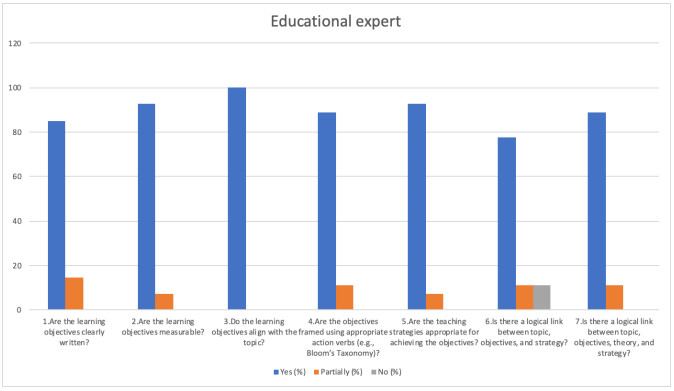


Fig. 3. Evaluation results from educational experts across nine topics: percentage distribution of yes, partially, and no responses for seven criteria.

Cybersecurity instructors reported high agreement on all rating criteria. Learning objectives were almost consistently rated as clear, measurable, well-aligned with the topic, and correctly framed with Bloom's verbs. Teaching strategies (84.4%) and the logical relationship between components (82.2%) also received strong ratings, indicating instructional coherence (see Fig. 2).

Educational experts similarly reported strong agreement across all criteria. Learning objectives were rated as clear (85.2%), measurable (92.6%), well-aligned with the topic (100%), and appropriately framed using Bloom's action verbs (88.9%). Teaching strategies (92.6%) and the logical relationship between topic, objectives, and strategy (77.8%) also scored highly, with the extended link to theory rated at 88.9%. These results highlight the instructional validity of the framework (see Fig. 3).

TABLE I. ELBOW METHOD AND SILHOUETTE SCORES FOR TF-IDF-BASED CLUSTERING

k	SSE	Silhouette Score
2	113.46	0.0123
3	111.95	0.0104
4	109.82	0.0126
5	108.32	0.0137
6	106.07	0.0193
7	104.25	0.0180
8	103.06	0.0179
9	101.17	0.0220
10	99.63	0.0241

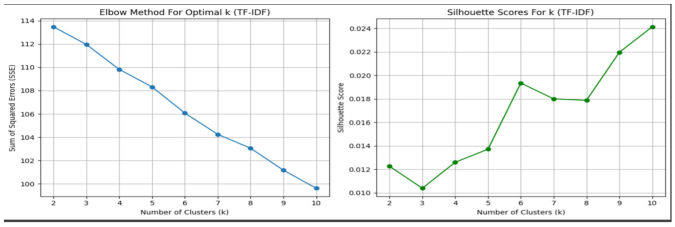


Fig. 4. Evaluation of TF-IDF clustering using the elbow method and silhouette scores.

TABLE II. NUMBER OF TOPICS BEFORE AND AFTER BALANCING USING TF-IDF-BASED CLUSTERING

Cluster	Original Samples	After Balancing
0	8	65
1	23	65
2	16	65
3	65	65[a]
4	7	65
Total Rows	119	325

[a] Unchanged during balancing.

TABLE III. ELBOW METHOD AND SILHOUETTE SCORES FOR SBERT-BASED CLUSTERING

k	SSE	Silhouette Score
2	42742.39	0.0635
3	40998.04	0.0470
4	39532.42	0.0599
5	38017.05	0.0649
6	36452.24	0.0651
7	35318.29	0.0726
8	35471.88	0.0617
9	34536.95	0.0649
10	33433.46	0.0646

After expert evaluation of the pilot course, the dataset was extended to cover the full diploma program, resulting in 119 structured entries across 14 courses. Topics were drawn directly from the official curriculum, while Custom GPT generated the corresponding objectives, theories, and strategies using the validated schema. The additional entries were generated automatically and were not subjected to further expert evaluation.

C. LLM Fine-Tuning

1) Dataset Preparation: To prepare the dataset for supervised fine-tuning, consistent and repeatable labels were required, since the Course Title and Topic fields alone did not provide enough overlap to enable generalization. To address this, an unsupervised clustering approach was used to introduce broader instructional categories. Specifically, the KMeans algorithm was used to group semantically related subjects, with each cluster assigned a pseudolabel representing a provisional instructional category [43]. For effective clustering, two feature extraction methods were tested on the Topic field: Term Frequency-Inverse Document Frequency (TF-IDF), which captures surface-level word relevance, and Sentence-BERT (SBERT), which generates dense semantic embeddings [44], [45]. Using both approaches enabled a comparison analysis to identify the representation most effective for structuring the dataset.

- TF-IDF: The first feature extraction approach converted topic phrases into numerical vectors using TfidfVectorizer with English stop-word removal. KMeans clustering was then applied to group related topics. The optimal number of clusters was determined using the Elbow Method [46] and Silhouette Score [47], with k=5 providing the best balance of compactness and interpretability (Table I, Fig. 4). The resulting clusters represented distinct instructional themes, which were assigned pseudo-labels for supervised fine-tuning. To address class imbalance (Table II), oversampling was applied, and lightweight data augmentation introduced paraphrased learning objectives (e.g., describe → explain) to enhance linguistic diversity. After balancing, the Silhouette Score for k=5 increased from 0.0137 to 0.1059, indicating stronger cohesion (Fig. 6b).
- SBERT: is a neural embedding approach built on a fine-tuned BERT architecture that captures sentence-level semantics using siamese and triplet networks [45]. Unlike TF-IDF, which relies on surface word frequencies, SBERT generates dense embeddings that encode contextual meaning, making it more effective for clustering short instructional topics. The pre-trained `all-MiniLM-L6-v2` model from the `sentence-transformers` library was used to encode topic entries, which were then clustered with KMeans. As with TF-IDF, the Elbow Method and Silhouette Score identified k=5 as the most interpretable solution (Table III, Fig. 5). The clusters represented coherent instructional themes and were assigned pseudo-labels for supervised fine-tuning.To address class imbalance (Table IV), oversampling and paraphrasing-based augmentation were applied. After balancing, the Silhouette Score for k=5 increased to 0.1150, reflecting stronger semantic cohesion (Fig. 6a).

To compare TF-IDF and SBERT, a mixed evaluation was performed. Internal metrics (Elbow, Silhouette) provided numerical validation, but six cybersecurity education experts reviewed representative clusters to assess semantic coherence, topic distinctiveness, and pedagogical relevance. While TF-IDF achieved slightly higher quantitative scores (Fig. 6b),

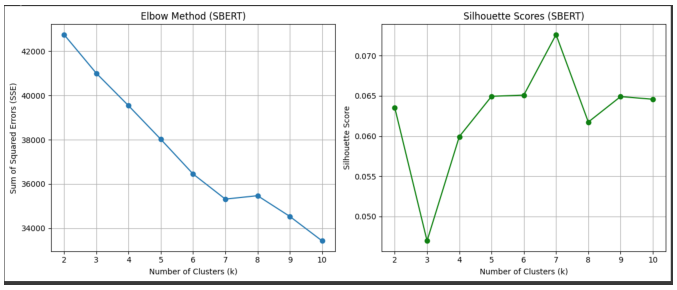
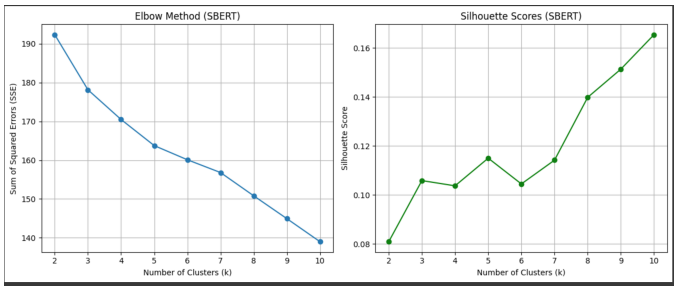


Fig. 5. Evaluation of SBERT clustering using the elbow method and silhouette scores.

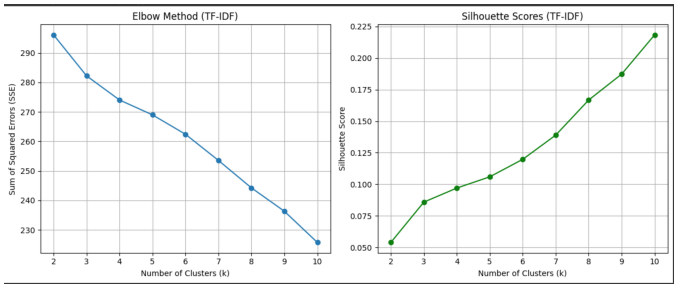
TABLE IV. NUMBER OF TOPICS BEFORE AND AFTER BALANCING USING SBERT-BASED CLUSTERING.

Cluster	Original Samples	After Balancing
0	30	53
1	53	53[a]
2	12	53
3	11	53
4	13	53
Total Rows	119	265

[a] Unchanged during balancing.



(a) SBERT-based clustering.



(b) TF-IDF-based clustering.

Fig. 6. Comparison of clustering evaluation between SBERT and TF-IDF (post-balancing).

experts consistently preferred SBERT-based groupings, citing clearer semantics and greater instructional utility. Consequently, SBERT was adopted for subsequent model development.

2) Fine-Tuning Process: The fine-tuning process was carried out using the OpenAI API, which provides access to state-of-the-art instruction-following language models.

- Base Model: The experiments used `gpt-3.5-turbo-1106`, recommended by OpenAI for instruction-oriented tasks [48]. GPT-3.5 Turbo is an efficient variant of the GPT-3.5 family optimized for conversational and structured text generation, making it well suited for producing lesson plans.
- Inputs: Fine-tuning data was formatted in `.jsonl`, with each entry structured as a dictionary. A chat-style prompt format was adopted with three roles:
  - System: standardized as "You assist instructors with designing effective lesson plans."
  - User: structured inputs (domain, course title, topic).
  - Assistant: instructional outputs (learning objectives, theoretical context, teaching strategy).

  The dataset was partitioned using stratified sampling to preserve instructional category proportions: 70% training, 15% validation, and 15% testing.
- Training: OpenAI's fine-tuning framework allows adjustment of three hyperparameters: `n_epochs`, `batch_size`, and `learning_rate_multiplier`. This study used the platform defaults: `n_epochs = 3`, `learning_rate_multiplier = 2`, and `batch_size = 1`.

D. Evaluation

We evaluated the fine-tuned model on a held-out test split comprising 15% of the dataset, which was excluded from both training and validation. This setup ensured that the assessment reflected the model's ability to generalize to unseen lesson plan inputs rather than relying on memorized exemplars. The use of a reserved test partition also provides a more reliable indication of performance when the model is applied to new instructional material. Performance was assessed using BLEU, ROUGE, and BERTScore, which together provide complementary perspectives on lexical fidelity, structural coverage, and semantic alignment with the reference dataset. Within this evaluation, the metrics serve as consistent indicators of the extent to which the generated lesson plans align with the intended instructional material.

IV. RESULTS AND DISCUSSION

This section presents the experimental results to assess the generated lesson plans in terms of lexical similarity, semantic alignment, and consistency with the intended instructional structure.

The evaluation of the fine-tuned model on the held-out test set (15%, $n = 40$) indicates strong lexical and semantic similarity between the generated lesson plans and the reference texts. The results are summarized in Table V. The BLEU score (73.54) reflects high n-gram precision, suggesting close surface-level correspondence in instructional vocabulary and phrasing. This result indicates that the model consistently reproduces domain-relevant instructional language, contributing to clarity and standardization in the generated pedagogical content.

TABLE V. AUTOMATED EVALUATION METRICS FOR THE FINE-TUNED MODEL

Metric	Score
BLEU	73.54
ROUGE-1	82.19
ROUGE-2	68.91
ROUGE-L	78.16
ROUGE-Lsum	81.28
BERTScore P	97.40
BERTScore R	97.46
BERTScore F1	97.43

TABLE VI. COMPARATIVE EVALUATION OF FINE-TUNED LLMS (SCORES ON A 0–100 SCALE)

Study	Model	BLEU	R-1	R-2	R-L	BERT-F1
This study	GPT-3.5	73.54	82.19	68.91	78.16	97.43
Russian Summ. [24]	GPT-3	23.10	11.40	1.40	10.00	89.00
LLMs Cyber Q&A [49]	LLaMA-3	96.70	76.20	–	76.20	–

ROUGE scores provide further evidence of lexical and structural correspondence between the generated lesson plans and the reference texts. ROUGE-1 (82.19) demonstrates strong unigram recall, ROUGE-2 (68.91) reflects solid bigram coverage despite the stricter matching requirement, and ROUGE-L (78.16) indicates that generated outputs preserve sequential coherence with reference lesson plans. Together, these values show that the model effectively captures key instructional concepts and maintains structural organization, which are essential for high-quality instructional content.

BERTScore results (Precision = 97.40; Recall = 97.46; F1 = 97.43) indicate very high semantic similarity between the generated lesson plans and the reference texts. These values suggest that, even when lexical variation occurs, the generated content preserves meaning and reflects the intended pedagogical intent. Moreover, the consistently high semantic scores indicate that the fine-tuned model reliably adheres to the required instructional structure (three learning objectives, one theoretical context, and one instructional strategy), supporting pedagogical consistency across generated outputs.

Overall, these findings indicate that the fine-tuned model performs reliably in generating structured educational content. Its ability to balance lexical fidelity with semantic adequacy, while consistently adhering to the intended instructional structure, underscores its potential for practical use in lesson plan generation where pedagogical relevance is essential. To contextualize these results, Table VI compares the proposed model against recent fine-tuning efforts in other domains, including Russian text summarization [24] and cybersecurity question answering [49]. Although these tasks differ in format lesson plan generation, abstractive summarization, and factual question answering they share the common challenge of producing outputs that are lexically precise, semantically coherent, and domain-specific.

As shown, the Russian summarization study reported lower n-gram and ROUGE values, with BLEU = 23.10, ROUGE-1 = 11.40, and ROUGE-L = 10.00, but achieved comparatively stronger semantic alignment (BERTScore F1 = 89.00) [24]. These outcomes are consistent with prior findings in abstractive summarization, where lexical overlap is often limited due to paraphrasing. In contrast, the cybersecurity Q&A study attained very high BLEU (96.70) and ROUGE-1/ROUGE-L (76.20) scores [49], reflecting the fact-oriented and terminologically constrained nature of short-answer tasks. Within this spectrum, the present study demonstrates a balanced profile, combining strong lexical correspondence (BLEU = 73.54; ROUGE-1 = 82.19; ROUGE-L = 78.16) with near-perfect semantic similarity (BERTScore F1 = 97.43). This suggests that structured instructional generation tasks benefit from maintaining both consistency in phrasing and preservation of pedagogical meaning.

V. CONCLUSION

This study addressed the challenge of generating structured, pedagogy-aligned lesson plans for cybersecurity, a discipline that requires both strong theoretical grounding and practical proficiency. By constructing a domain-specific dataset, integrating explicit learning objectives, learning theories, and evidence-based teaching strategies, and embedding these elements into the fine-tuning process, the study proposed a structured framework for lesson plan generation. The experimental results demonstrated strong lexical and semantic alignment between the generated and reference lesson plans, indicating that the framework effectively preserves instructional structure and pedagogical intent.

Despite these contributions, the study has limitations. The dataset was derived from a single academic program, which may not fully represent the diversity of cybersecurity curricula, and the evaluation relied primarily on automated similarity metrics that cannot capture all aspects of instructional effectiveness in real classroom settings. Future work should therefore expand the dataset to cover broader curricular contexts and incorporate expert review and classroom-based evaluations to assess instructional quality, adaptability, and learning impact more comprehensively.

More broadly, this work demonstrates that integrating pedagogical representations into the fine-tuning of large language models enables AI systems to function as pedagogy-aware instructional assistants rather than generic content generators. The proposed framework offers a practical and transferable approach to AI-assisted lesson planning, in which pedagogical design is embedded directly into the generation process. While this study focuses on cybersecurity education, the framework can be adapted to other domains by redefining the domain knowledge base and corresponding pedagogical mappings.

Within the scope of this study, the results indicate that pedagogy-aligned fine-tuning can effectively support the generation of structurally consistent and instructionally valid lesson plans in a specialized domain such as cybersecurity.

ACKNOWLEDGMENT

The project was funded by KAU Endowment (WAQF) at king Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore, acknowledge with thanks WAQF and the Deanship of Scientific Research (DSR) for technical and financial support.

REFERENCES

[1] Y. Zheng, X. Li, Y. Huang, Q. Liang, T. Guo, M. Hou, B. Gao, M. Tian, Z. Liu, and W. Luo, "Automatic lesson plan generation via large language models with self-critique prompting," in *Proceedings of the International Conference on Artificial Intelligence in Education*. Springer, 2024, pp. 163–178.

[2] B. Hu, L. Zheng, J. Zhu, L. Ding, Y. Wang, and X. Gu, "Teaching plan generation and evaluation with gpt-4: Unleashing the potential of llm in instructional design," *IEEE Transactions on Learning Technologies*, 2024.

[3] H. Fan, G. Chen, X. Wang, and Z. Peng, "Lessonplanner: Assisting novice teachers to prepare pedagogy-driven lesson plans with large language models," in *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 2024, pp. 1–20.

[4] S. Puranik, "Innovative teaching methods in higher education," *BSSS Journal of Education*, vol. 9, no. 1, pp. 67–75, 2020.

[5] S. Strickroth, "Platon: Developing a graphical lesson planning system for prospective teachers," *Education Sciences*, vol. 9, no. 4, p. 254, 2019.

[6] N. D. A. Lubis *et al.*, "Learning theory," OSF, 2021, [Online]. Available: https://osf.io. Accessed: Oct. 7, 2025.

[7] D. T. Hickey and R. J. Kantor, "Contrasting cognitive theories of learning transfer to advance cybersecurity instruction, assessment, and testing," *Journal of Cybersecurity Education, Research and Practice*, vol. 2024, no. 1, 2024.

[8] T. E. Dorgu, "Different teaching methods: A panacea for effective curriculum implementation in the classroom," *International Journal of Secondary Education*, vol. 3, no. 6, pp. 77–87, 2015.

[9] A. M. Alnajim, S. Habib, M. Islam, H. S. AlRawashdeh, and M. Wasim, "Exploring cybersecurity education and training techniques: a comprehensive review of traditional, virtual reality, and augmented reality approaches," *Symmetry*, vol. 15, no. 12, p. 2175, 2023.

[10] W. S. Admass, Y. Y. Munaye, and A. A. Diro, "Cyber security: State of the art, challenges and future directions," *Cyber Security and Applications*, vol. 2, p. 100031, 2024.

[11] M. Shivapurkar, S. Bhatia, and I. Ahmed, "Problem-based learning for cybersecurity education," *Journal of The Colloquium for Information Systems Security Education*, vol. 7, no. 1, pp. 6–6, 2020.

[12] M. Mukherjee, N. T. Le, Y.-W. Chow, and W. Susilo, "Strategic approaches to cybersecurity learning: A study of educational models and outcomes," *Information*, vol. 15, no. 2, p. 117, 2024.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[17] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[18] M. Moradi, K. Yan, D. Colwell, M. Samwald, and R. Asgari, "Exploring the landscape of large language models: Foundations, techniques, and challenges," *arXiv preprint arXiv:2404.11973*, 2024.

[19] R. Patil and V. Gudivada, "A review of current trends, techniques, and challenges in large language models (llms)," *Applied Sciences*, vol. 14, no. 5, p. 2074, 2024.

[20] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, G. Wang *et al.*, "Instruction tuning for large language models: A survey," *ACM Computing Surveys*, 2023.

[21] B. Weng, "Navigating the landscape of large language models: A comprehensive review and analysis of paradigms and fine-tuning strategies," *arXiv preprint arXiv:2404.09022*, 2024.

[22] D. Liga and L. Robaldo, "Fine-tuning gpt-3 for legal rule classification," *Computer Law & Security Review*, vol. 51, p. 105864, 2023.

[23] O. Zheng, M. Abdel-Aty, D. Wang, C. Wang, and S. Ding, "Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety," *arXiv preprint arXiv:2307.15311*, 2023.

[24] A. Nikolich, I. Osliakova, T. Kudinova, I. Kappusheva, and A. Puchkova, "Fine-tuning gpt-3 for russian text summarization," in *Computational Methods in Systems and Software*. Springer, 2021, pp. 748–757.

[25] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr, "Harnessing llms in curricular design: Using gpt-4 to support authoring of learning objectives," *arXiv preprint arXiv:2306.17459*, 2023.

[26] N. Scaria, S. Dharani Chenna, and D. Subramani, "Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation," in *International Conference on Artificial Intelligence in Education*. Springer, 2024, pp. 165–179.

[27] E. Latif and X. Zhai, "Fine-tuning chatgpt for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100210, 2024.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[29] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.

[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[31] S. Kloker, H. Bukoli, and T. Kateete, "New curriculum, new chance—retrieval augmented generation for lesson planning in ugandan secondary schools: Prototype quality evaluation," *arXiv preprint arXiv:2408.07542*, 2024.

[32] D. C. L. Tsai, W. J. W. Chang, and S. J. H. Yang, "Short answer questions generation by fine-tuning bert and gpt-2," in *Proceedings of the 29th International Conference on Computers in Education (ICCE)*, 2021, pp. 123–132.

[33] L. Jia, C. Qi, Y. Wei, H. Sun, and X. Yang, "Fine-tuning large language models for educational support: Leveraging gagne's nine events of instruction for lesson planning," *arXiv preprint arXiv:2503.09276*, 2025.

[34] M. Gharbi, M. T. Mohtadi, and A. F. Merkazi, "Revolutionizing moocs with fine-tuned chatgpt: Personalized learning at scale," *International Journal of Computing and Digital Systems*, vol. 16, no. 1, pp. 1–11, 2024.

[35] D. C. M. Taylor and H. Hamdy, "Adult learning theories: implications for learning and teaching in medical education: Amee guide no. 83," *Medical Teacher*, vol. 35, no. 11, pp. e1561–e1572, 2013.

[36] J. Vykopal, P. Seda, V. Švábenský, and P. Čeleda, "Smart environment for adaptive learning of cybersecurity skills," *IEEE Transactions on Learning Technologies*, vol. 16, no. 3, pp. 443–456, 2022.

[37] Z. Abulhul, "Teaching strategies for enhancing student's learning," *Journal of Practical Studies in Education*, vol. 2, no. 3, pp. 1–4, 2021.

[38] H. Rathore and H. Griffith, "Enhancing cybersecurity education through research-based role-playing," in *Proceedings of the IEEE Integrated STEM Education Conference (ISEC)*, 2025, pp. 1–6.

[39] H. Kavak, J. J. Padilla, D. Vernon-Bido, S. Y. Diallo, R. Gore, and S. Shetty, "Simulation for cybersecurity: state of the art and future directions," *Journal of Cybersecurity*, vol. 7, no. 1, p. tyab005, 2021.

[40] D. R. Krathwohl, "A revision of bloom's taxonomy: An overview," *Theory Into Practice*, vol. 41, no. 4, pp. 212–218, 2002.

[41] OpenAI, "Introducing gpts," https://openai.com/blog/introducing-gpts, 2023, accessed: Oct. 7, 2025.

[42] American College Health Association, "Guidance for designing successful learning objectives," https://www.acha.org/education-events/guidance-for-designing-successful-learning-objectives/, 2025, accessed: May 20, 2025.

[43] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[44] J. Ramos, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the 1st Instructional Conference on Machine Learning*, vol. 242, no. 1, New Jersey, USA, 2003, pp. 29–48.

[45] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[46] D. J. Ketchen and C. L. Shook, "The application of cluster analysis in strategic management research: An analysis and critique," *Strategic Management Journal*, vol. 17, no. 6, pp. 441–458, 1996.

[47] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[48] OpenAI, "Gpt-3.5 turbo," https://platform.openai.com/docs/models/gpt-3.5-turbo, 2023, accessed: Feb. 9, 2025.

[49] Y. Huang, "Fine-tuning of large language models for domain-specific cybersecurity knowledge," *arXiv preprint arXiv:2509.25241*, 2025.