Toward the Use of LLMs to Support Curriculum Mapping to Established Frameworks

Mr. Eric L Brown, Tennessee Technological University

Eric L. Brown is an education leader with over 28 years of experience in higher education, currently serving as the Associate Director of Workforce Development for the Cybersecurity, Education, Research, and Outreach Center at Tennessee Tech University. As a senior lecturer in the Computer Science department, Eric teaches various cybersecurity courses and agile-focused software engineering.

His prior experiences include serving as a District Solutions Advocate for the Tennessee Department of Education, where he played a key role in the Chief Information Officer's leadership team. In addition, Eric served for eight years as a school board member in Putnam County, TN, with four years in leadership positions, giving him valuable insights into K12 education.

Today, Eric's work focuses on cybersecurity education and workforce development in the K-16 sector, building pathways for students and professionals in emerging cyber fields.

Douglas A. Talbert, Tennessee Technological University

Dr. Doug Talbert is a Professor of Computer Science at Tennessee Tech University, where he has worked since 2002. He work focuses on trustworthy human-AI interatction, especially in the area of clinical informatics.

Dr. Jesse Roberts, Tennessee Technological University

I am an Assistant Professor at Tennessee Technological University in the computer science department. My research focuses on the development, evaluation, and application of empirical methods for NLP with emphasis on transformer-based LLMs. I prefer applications that contribute to digital humanities, robotics, law, cognitive science, and the preservation of endangered languages - primarily Cherokee (Tsalagi) and Irish (Gaeilge). But I'm also interested in game theory, computational physics, and any idea that can positively impact people.

Toward the Use of LLMs to Support Curriculum Mapping to Established Frameworks

Eric L. Brown, Douglas A. Talbert, Jesse Roberts
Department of Computer Science
Tennessee Tech University
Email: elbrown@tntech.edu, dtalbert@tntech.edu, jtroberts@tntech.edu

Abstract

Large language models (LLMs) have revolutionized content creation across various domains, yet their application in highly specialized fields such as curriculum gap analysis remains challenging. This study (full paper format), to be presented at the 2025 ASEE Annual Conference's Computers in Education Division (COED), investigates the efficacy of LLMs in developing and evaluating curriculum pathways based upon established frameworks.

Traditional curriculum plans, such as those from institutions designated by the NCAE-C in cybersecurity defense education (CAE-CD) and cyber operations (CAE-CO) designation tracks, involve a manual process of identifying knowledge units for specific courses. Throughout the mapping process, gaps are identified in curriculum plans based on the knowledge of the subject matter expert(s) (SME). Performing this task for one framework is challenging enough; consider the increased complexity and risk of error when multiple frameworks are cross-referenced into the plan. Improvement opportunities exist in the curriculum mapping and gap analysis process. This leads to the question of whether an LLM can speed up the curriculum mapping process compared to a manual process conducted by an SME, which will be evaluated through a set of "human in the loop" experiments.

To evaluate this question, the paper details the results of the following experiments involving a computer science/cybersecurity curriculum being mapped to the CAE-CD knowledge units (KU):

1. A single SME will create a manual KU curriculum mapping.
2. Provide an LLM with full curriculum details, including catalog descriptions and syllabi, and create a mapping for a single CAE-CD knowledge unit.
3. Provide an LLM with details of all CAE-CD knowledge units and information for one course (catalog description and syllabus) and create a knowledge unit mapping for that course.
4. Provide an LLM with details of all CAE-CD knowledge units, course descriptions, and syllabi, and create a knowledge unit mapping for the full curriculum.

We will evaluate the results using the following metrics (where applicable):

1. Accuracy - precision, recall, F1, and error rate of LLM vs SME
2. Efficiency - time to complete and reduction of human effort
3. Qualitative - expert review and comparison of output quality

Our analysis will show that the manual curriculum mapping and gap analysis process can be accelerated and improved by suggesting possible outcomes from the LLM, allowing the SME to

assume the more effective role of critical reviewer over content developer. We will also show that this assistance comes with constraints. The paper also introduces the CIRCLE methodology—a structured, human-in-the-loop process for aligning curriculum to frameworks using LLMs—which synthesizes best practices derived from the experiments.

Introduction

Academic standards management requires multiple retrospective efforts to ensure that the curriculum and programs align well with established standards. These manual processes demand extensive work hours from subject matter experts (SMEs) for tasks such as document acquisition, classification, review, revision, and curriculum (re)alignment/mapping. While each task is demanding in isolation, their cumulative effect can be overwhelming, particularly when SMEs must meticulously map internal curriculum documents to expansive, multifaceted frameworks. This process mirrors a means-ends analysis problem-solving strategy, where SMEs must continuously compare internal documents to external standards, identify gaps, and determine appropriate modifications. Research in cognitive load theory suggests that such processes impose a heavy cognitive burden, as problem solvers must simultaneously track multiple states, constraints, and potential solutions [1]. The high cognitive demand of this alignment task may hinder efficiency and accuracy, leading to inconsistencies and increased effort over time. Given these challenges, there is a need for more structured, automated, or AI-assisted approaches to alleviate the cognitive load associated with curriculum mapping and improve the overall effectiveness of academic standards management [2].

The advent of advanced machine learning mechanisms—evolving from early neural networks to modern transformer architectures—has ushered in a new renaissance in artificial intelligence and its practical applications. The rapid development of large language models (LLMs), capable of processing substantial volumes of unstructured text and generating structured outputs, now empowers framework mapping projects at a quality level that was inconceivable less than a decade ago. SMEs now have access to AI tools that facilitate comprehensive reviews of local guidance documents and alignment exercises with strategic frameworks. In practice, instructional design teams have used tools like ChatGPT and Copilot to accelerate the development of course maps, learning outcome alignments, and instructional components, demonstrating how LLMs can meaningfully support expert-led curriculum alignment [3]. Complementing these real-world use cases, a growing body of scholarly literature also highlights the potential of LLMs to automate repetitive instructional tasks and enhance personalized learning systems [4]. These efforts improve the quality and consistency of local guidance and mitigate time constraints that would otherwise hinder thorough manual review. Consequently, the goal of this paper is not to advocate for full automation of framework mapping, but rather to critically evaluate emerging tools and techniques that assist SMEs in producing higher-quality outcomes in less time—a "human-in-the-loop" model for process improvement. We propose the CIRCLE methodology, a six-phase framework for prompt-driven curriculum review to support this human-in-the-loop model. Developed as a synthesis of the most effective practices observed across experiments, CIRCLE guides SMEs through a repeatable process of content collection, chunking, prompting, and evaluation using LLMs. While not tested as a standalone experiment, the CIRCLE framework emerged organically from this study and provides a reusable scaffold for broader application.

For this exercise, our evaluation will look at mapping a set of college courses to the National Centers of Academic Excellence in Cybersecurity (NCAE-C) Cybersecurity Defense Education (CAE-CD) knowledge unit framework. This will be a limited-scope exercise to evaluate the usefulness of LLMs at a high level of review. We will address further extensions of the exercise in the Discussion section of this paper.

In this paper, we provide a process by which educators and evaluators may leverage the power of LLMs as collaborative reasoning assistants (CRAs) to support curriculum mapping. Critically, the developed process does not cede control to the model and is instead part of the growing body of expert-in-the-loop work that seeks to augment rather than automate the capability of human experts. Additionally, we document important failure modes specific to curriculum mapping (hallucinated criteria, etc.) that will likely lead to critical errors in the developed mapping without appropriate expert oversight.

Background

All organizational structures have some level of guiding framework that dictates best practices. While some frameworks may be suggestive, others may be regulatory (and consequently mandatory). A topic of current popularity, given the current geopolitical climate, being heavily reviewed is the establishment and maintenance of evaluation frameworks for critical infrastructure [5] [6]. Consider the exhaustive list of frameworks, many overlapping and possibly conflicting, for such a complicated sector with interconnecting layers such as power, water, and transportation [7].

In the healthcare sector, numerous accreditation frameworks exist to ensure quality and safety, including the Joint Commission (TJC), Det Norske Veritas (DNV), the Center for Improvement in Healthcare Quality (CIHQ), and the Centers for Medicare and Medicaid Services (CMS). Each organization provides distinct standards and survey processes, resulting in overlapping and sometimes conflicting requirements. Healthcare providers must navigate these frameworks to maintain compliance and ensure high-quality care. Questions about the use of ML/AI in healthcare include [8]

1. How can it be used to safely assist patient care?
2. How can it be used and not violate patient privacy?
3. How can it be verified?

Considerable work has been conducted in the education sector to incorporate LLMs into curriculum development and review processes amid frequently changing frameworks [9] [10]. Like the other sectors, issues of accuracy, bias, and privacy are issues where “human in the loop” in the form of effective SMEs are essential. Within education, the use of LLMs to create content and framework review mechanisms creates a need for education of use as well as education of proper material consumption. How can we train SMEs to effectively use these tools, setting the appropriate expectations for outcomes and review requirements?

While we would love to be able to create our own version of J.A.R.V.I.S. (see Tony Stark - Marvel Universe), this is not our real world. ML tools supporting guiding frameworks’ development, implementation, and management will always be a “human in the loop” system. The human element is essential to deal with issues of accuracy review, human and/or model bias,

data privacy, and general safety [11] [12]. SME education regarding the benefits and limitations of the tools is of equal importance [13].

Methodology

As addressed in the abstract, the process was evaluated through four experiments. Each experiment examined the positive and negative aspects of using an LLM as a CRA. The key distinction across experiments is how the LLM was tasked with processing curriculum data: Experiment 1 used no LLMs; Experiments 2 and 3 limited the evaluation to either a single knowledge unit or a single course, respectively; and Experiment 4 extended the scope to a complete mapping of all knowledge units across all courses.

General Background for All Experiments

For each LLM-enhanced experiment, the following protocols were followed:

1. Each experiment was conducted on ChatGPT 4o [14], Claude 3.5 Sonnet [15], and Gemini [16]. Prompts were submitted via a chat window. No programming or API calls were required.
2. Each experiment was conducted in a project environment (e.g., a "Gem" in Gemini) pre-loaded with two core reference documents: the CAE-CD Knowledge Units Documentation (framework) and the CAE-NICE-ABET Curriculum Mapping (SME-reviewed outcomes). Additional documents—such as syllabi or batch course PDFs—were uploaded at the time of prompting based on the specific experiment design.
3. The SME-reviewed document was considered the sole source of truth for evaluation purposes.
4. 16 documents were included in the experiments. KU mapping information was redacted.
5. The same prompts were used for each LLM (background and action) to provide consistency.

Note on Token Window Awareness. When using LLMs for curriculum mapping or similar document-rich tasks, practitioners must be mindful of each model's context window limitations. Although modern models like Gemini and Claude 3.5 advertise token capacities ranging from 128k to over 1M, practical limits may be substantially lower, especially when reference documents contain complex structures (e.g., PDFs with multi-column layouts, headers, and embedded formatting) [17] [18]. For this study, combined inputs of the KU framework and 16 course syllabi exceeded 60,000 tokens, which strained or exceeded Gemini's tolerance and reduced output fidelity in other models. To avoid hallucinations or dropped content, we recommend chunking large documents into smaller, thematically grouped sections and supplying background frameworks via persistent project memory where possible. SMEs should additionally design prompts with token-awareness, ideally measuring the average syllabi + framework input lengths and staying well below model-specific limits (e.g., 8k or 32k tokens) to avoid truncation or hallucination.

Table 1 outlines the structure and focus of the four experiments. The prompting strategies, evaluation patterns, and feedback loops captured in these experiments later informed the development of the CIRCLE methodology—a structured framework for repeatable LLM-supported curriculum mapping, described in the Discussion section.

Table 1: Summary of Experimental Conditions and Evaluation Criteria

Exp	Title	Task Description	Evaluation Focus
1	Manual SME Mapping	SME maps all knowledge units to all courses manually using institutional documents and past mappings.	Time to complete, used as ground truth for comparison with LLMs.
2	One KU to All Courses	LLM maps a single CAE-CD knowledge unit (e.g., PLE) to a full set of syllabi. Two variations were run using separate and combined inputs.	Precision, recall, and efficiency. Course-level KU coverage with feedback.
3	All KUs to One Course	LLM maps all knowledge units to a single syllabus. Feedback includes justification, comparison to SME mapping, and statistical scoring.	F1 score, precision, recall, quality of justification, and SME refinement opportunity.
4	All KUs to All Courses	LLM maps all knowledge units to all syllabi in a single or chunked batch. Tested with full curriculum PDF and grouped syllabi.	LLM accuracy across all syllabi. Basic precision and recall.

Experiment 1 - Manual SME review / No LLM Mapping

Experiment 1 was conducted in early 2024. Its results serve as the controlled reference point for the outcomes of the subsequent experiments. The manual review was conducted during a re-designation process in 2020 and revisited in 2024. The KU mapping used historical data from a similar process completed in 2016, data from course update documents for the past four years, and newly revised knowledge unit documentation. In addition to the CAE KUs, further mapping was completed to align with the NICE framework and ABET goals established by the Computer Science department. This additional mapping is mentioned as it will become a point of interest in Experiments 3 and 4. The process took one reviewer 30 hours to complete over a two-week period. While this timing was not precision clocked, the timing results from prior notes with general time indicators are acceptable for overall effort comparison. For each mapped course, the reviewer had to complete the following steps:

1. Collect a recent copy of the syllabus for each course in the designated curriculum.
2. Conduct follow-up interviews with the recent instructor, should the syllabus not provide sufficient topic coverage.
3. Review each KU and its related topics to determine if sufficient coverage exists to map the KU to the course.
4. Document mapping gaps to address in the course.

In addition to this effort, a review of the course KU mappings occurred in 2024. This review started with the last published document and re-evaluated the accuracy based on the current course material offering. This review required ten hours to complete over a three-day period. The process and review required approximately 40 person-hours to complete the effort. The work product from the effort, later called the CAE-NICE-ABET Curriculum Mapping document, was used as the "sole source of truth" for the mapping comparisons in the remaining experiments. This document also contained information regarding NICE Framework and ABET mappings.

Experiment 2 - LLM Mapping One KU to All Courses

This experiment focused on answering the question, "For a collection of syllabi, how many address knowledge unit XXX?" This experiment was conducted by providing up to six individual syllabi simultaneously (limited by the platforms and their associated context windows). A second version of this experiment was conducted by providing a single combined PDF document, which included all 16 syllabi. This document was optimized and text-recognized using Adobe Acrobat to assist with readability by the LLM. The authors used the Policy, Legal, Ethics, and Compliance (PLE) knowledge unit, which was known to be unique to one specific syllabus, where many of the others could have been generalized. This selection was made to help assess the accuracy of the evaluation. For ease of identification, the single combined document experiment will be called Experiment 2a. The experiment where groups of six documents were evaluated over the course of three runs will be called Experiment 2b.

A project/gem was created in each LLM. The project contained a general prompt (see below) and the CAE-CD KU Documentation document. The individual prompts within the projects included a copy of the syllabi and the instance prompt.

General prompt for project/gem setup

> You are an expert in the National Centers of Academic Excellence in Cybersecurity's Cyber Defense Education designation program. You can provide assistance with knowledge unit mapping suggestions based on the knowledge units listed in the CAE-CD KU Documentation document. You are polite, friendly, and helpful. With your understanding of the knowledge units provided by the CAE-CD KU Documentation.pdf file, you will look at a group of course syllabi and evaluate which courses either directly or indirectly relate to a specific KU provided by the user. Your response will be in the form of:
>
> ```
> Direct Match
>     CSC XXXX COURSENAME
>     CSC XXXX COURSENAME
> Inferred Match
>     CSC XXXX COURSENAME
>     CSC XXXX COURSENAME
>     In the template, XXXX will represent the course number
>     and COURSENAME will represent the name of the course.
> ```

Instance prompt

> ```
>     For these course syllabi, which implement the PLE
>     knowledge unit?
> ```

Experiment 3 - LLM Mapping ALL KUs to One Course – Time Taken + Recall/Precision of KUs

This was the most enlightening of the experiments, as discussed in the Results section. This experiment focuses on answering the question, "For a specific syllabus, which knowledge units should be mapped, either via direct or inferred match?" Like the other LLM experiments, a background prompt was provided to focus the project. An instance prompt was submitted to evaluate the specific syllabus. An analysis prompt was then entered to complete the outcome analysis and calculate the accuracy scores. These prompts were specifically crafted to allow some "dreaming" room for the LLM to allow the human reviewer to see other possibilities that had not been discovered. While the SME-reviewed document was considered the sole source of truth, the practicality is that an SME can always learn and reconsider decisions based on new observations. This experiment would show the value of an LLM as a CRA.

General prompt for project/gem setup

> You are an expert in the National Centers of Academic Excellence in Cybersecurity's Cyber Defense Education designation program. You can assist with knowledge unit mapping suggestions using the knowledge units listed in the CAE-CD KU Documentation document. You are polite, friendly, and helpful. You look for direct matches of the syllabus to knowledge units as well as inferred matches to find possible knowledge unit alignments. Identify which knowledge units directly match and which are inferred.

Instance prompt

> Create a KU mapping for the attached syllabus using the CAE-CD KU Documentation document only. Do not use a table format. Use the format of:
>
> ```
> Direct Match
> KU NAME (XXX) – JUSTIFICATION
> Inferred Match
> KU NAME (XXX) – JUSTIFICATION
> ```
>
> Where KU NAME is the full name of the knowledge unit, XXX is the three-letter abbreviation, and JUSTIFICATION is an explanation of why you made the match.

Analysis prompt

I am evaluating course mappings to CAE-CD Knowledge Units for the CSC XXXX course. Previously, you provided an evaluation that included direct and inferred KU matches. I have also conducted my own evaluation, found in the document named **CAE-NIST-ABET Curriculum Mapping**. Should you find multiple knowledge units in the human evaluation in the form of YYY-1, you should identify this as a single match to YYY and not multiple matches. Ignore the NICE Framework KSA section. Do not use tables in any part of your response.

Here's what I need in one conversation:

1. **Your Evaluation**: Provide the direct and inferred Knowledge Units (KUs) you mapped to CSC XXXX and their rationale.
2. **Comparison**: Compare your evaluation with the human evaluation found in the **CAE-NIST-ABET Curriculum Mapping Document** and categorize KUs into:
    - **True Positives (TP)**: KUs identified by both evaluations.
    - **False Positives (FP)**: KUs you identified but not in the human evaluation.
    - **False Negatives (FN)**: KUs in the human evaluation but not in yours.
3. **Metrics Calculation**: Calculate the precision, recall, and F1 score using my evaluation as the standard and present the results in this format:
    - Precision = TP / (TP + FP) = [Calculation] = [Value]
    - Recall = TP / (TP + FN) = [Calculation] = [Value]
    - F1 Score = 2 × (Precision × Recall) / (Precision + Recall) = [Calculation] = [Value]
4. **Analysis of Discrepancies**: For False Positives and False Negatives, provide a deeper evaluation and discuss possible reasons for differences in mapping.
5. **Recommendations**: Suggest how to refine course descriptions and mapping methodology for better alignment with CAE-CD standards.

I am attaching my evaluation document and the CAE-CD KU documentation for reference.

Experiment 4 - Mapping All KUs to All Courses

This experiment addressed the question, "If I give an LLM everything simultaneously, can it do all of the mapping in one prompt?" Like Experiment 2, this experiment was conducted in two different versions. The first version of the experiment was conducted by providing up to nine individual syllabi simultaneously (limit of the platforms). A second version of this experiment was conducted by providing a single combined PDF document, which included all 16 syllabi. This document was optimized and text-recognized using Adobe Acrobat to assist with readability by the LLM.

Instance prompt

> Create a knowledge unit mapping for each course for which a syllabus is provided. I will provide placeholders for information in all capital letters. Please preserve the formatting and overall template that I provide. Match knowledge units (KU) to each course where most of the topics are addressed. Make note of which KUs are direct matches and which KUs are suggestions based upon inference. Produce a PDF named Experiment4.pdf, which contains your results. This is the template:
>
> ```
> CSC XXXX – COURSE NAME
> KU NAME
> KU TOPIC X NAME – DESCRIPTION
> KU NAME
> KU TOPIC X NAME – DESCRIPTION
> ```

Results

General analysis element considerations

- **True Positive (TP)**: Both the LLM and the manual SME review (experiment 1) agree that the KU outcomes belong to this course syllabus, meaning the KU is correctly identified as relevant.
- **False Positive (FP)**: The LLM identifies the KU as relevant, but the manual SME review disagrees, determining that the KU outcomes do not belong to this course syllabus.
- **True Negative (TN)**: Both the LLM and the manual SME review agree that the KU outcomes do not belong to this course syllabus, meaning the KU is correctly excluded.
- **False Negative (FN)**: The manual SME review identifies the KU as relevant, but the LLM fails to recognize this, leading to the omission of a KU that should belong to this course syllabus.

Experiment 1 - Manual SME review / No LLM Mapping – Time Taken

This experiment served as the control for the other work. As stated earlier, this manual process took approximately 40 hours of work to complete using a completely manual process. This effort's outcomes were considered the sole source of truth for the remaining experiments. However, it is important to acknowledge that human assessments are not infallible. Factors such as fatigue, cognitive overload, and inherent biases can lead to errors during manual evaluations. For instance, fatigue and repetition have been shown to impair decision-making and increase the likelihood of mistakes, particularly in tasks requiring sustained attention. [19] One of the positives of using LLMs as a CRA is the ability to see other considerations that may have been overlooked during an extended, possibly rushed, mapping process. Secondly, we must also consider that humans and LLMs both have limited "context windows." Miller suggests that humans can hold seven discrete things in short-term memory before information is lost. [20] Others have discussed techniques that can be used to extend the size of the context window in LLMS. [21]. Gemini - 1 million tokens, Claude - 200,000 (170,000), ChatGPT (4,096 with chunking extensions)

Experiment 2 - LLM Mapping One KU to All Courses – Time Taken + Recall/Precision of Courses

As stated in the methodology section, this experiment was conducted using two different methods - 2a and 2b. Experiment 2a (single combined syllabus document) did not yield useful results. Two of the three LLMs provided a listing of all the KUs and failed the mapping process. Gemini failed to conduct the analysis. Secondly, six syllabi were evaluated per run (Experiment 2b). The project had access to the CAE-CD KU Documentation. The batches were conducted in the same conversation, allowing the evaluations to improve with each run. Each LLM evaluated a total of 16 syllabi.

The PLE KU was selected as it was only mapped by the SME to CSC 3570; however, the outcome goal language in the CSC 2570 syllabus would strongly indicate a match with PLE as well. Ethics outcomes listed in the KU and CSC 3040 would also be indicated. This would allow us to see the accuracy of the models compared to the human-reviewed document.

This experiment provided a view of how the LLM could be used to make suggestions for the SME. Two LLMs (ChatGPT and Gemini) identified CSC 3570 as a possible mapping. All three models strongly indicated a mapping to CSC 2570 based on outcomes described in the syllabus. Upon further review, PLE could be mapped to this class. Two LLMs (ChatGPT and Claude) identified CSC 3040 due to the ethics goals outlined in the course syllabus.

Table 2: Experiment 2b - Matches Identified for Policy, Legal, Ethics, and Compliance (PLE) Knowledge Unit

LLM	Run	Direct Matches	Inferred Matches
ChatGPT	1	CSC 2570	CSC 2310, CSC 2510
	2	CSC 2570, CSC 3570, CSC 3040	CSC 3300, CSC 3410
	3	CSC 2570, CSC 3570, CSC 3040	CSC 3300, CSC 4610
Claude	1	CSC 2570	CSC 1300
	2	CSC 2570, CSC 3040, CSC 3300	CSC 1300
	3	None	CSC 4610, CSC 4615/20-001
Gemini	1	CSC 2570	CSC 2310, CSC 2400, CSC 2510, CSC 1300, CSC 1310
	2	CSC 3570	CSC 3300, CSC 3410, CSC 3710, CSC 2700, CSC 3040
	3	CSC 3300	CSC 4320, CSC 4610, CSC 4615, CSC 4100

This experiment also provides a point of interest in how the LLMs addressed the request. ChatGPT created a cumulative mapping, retaining learned items from previous runs. This would be consistent with its sliding window attention mechanism, where each 4,096 token chunk is processed, summarized, and overlapped with the next chunk, allowing the past information to be carried over into the next chunk. Claude AI could complete the task to a point but overran its context window by Run 3. Gemini provided warnings about context window overruns in each

run, largely due to the size of the KU reference document. This impacted its direct matching abilities and contributed to hallucinating tendencies in its inferred KU matches.

Experiment 3 - LLM Mapping ALL KUs to One Course – Time Taken + Recall/Precision of KUs

As mentioned earlier, Experiment 3 proved to be the most valuable of the experiments in terms of positive outcomes assisting the SME in the mapping process. The experiment allowed the LLM to focus on a single syllabus and the KU documentation. Details of the reasoning process and suggestions for syllabus improvement were better in this experiment. This can be attributed to the working datasets being more compatible with the model's context windows. It should be noted that while Gemini did better in this experiment than any of the other LLM-enhanced experiments, it still issued context window errors for every prompt. The summary of the evaluations can be found in Table 3.

Table 3: Comparison of ChatGPT, Claude AI, and Gemini Experiments Results

Course	ChatGPT			Claude AI			Gemini		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
CSC1300	0.50	0.67	0.57	0.20	1.00	0.33	0.25	1.00	0.40
CSC1310	0.50	1.00	0.67	0.50	1.00	0.67	0.50	1.00	0.67
CSC2310	0.43	1.00	0.60	0.29	0.67	0.40	0.29	1.00	0.44
CSC2400	0.25	1.00	0.40	0.17	1.00	0.29	1.00	0.67	0.80
CSC2510	0.43	0.75	0.55	0.33	1.00	0.50	0.44	0.40	0.42
CSC2570	0.56	0.63	0.59	0.75	0.75	0.75	0.47	0.44	0.46
CSC2700	0.13	1.00	0.22	0.20	1.00	0.33	0.25	1.00	0.40
***CSC3040**	0.75	0.75	0.75	NaN	NaN	NaN	0.40	1.00	0.57
CSC3300	0.40	1.00	0.57	0.50	1.00	0.67	0.43	0.67	0.52
CSC3410	0.50	1.00	0.67	0.20	1.00	0.33	0.13	1.00	0.22
CSC3570	0.65	0.92	0.67	0.70	0.70	0.70	0.64	0.88	0.74
CSC3710	0.25	1.00	0.40	0.67	1.00	0.80	0.20	0.50	0.29
CSC4100	0.50	1.00	0.67	0.33	1.00	0.50	0.50	1.00	0.67
***CSC4320**	0.43	0.50	0.46	NaN	NaN	NaN	0.33	0.33	0.33
CSC4610	0.43	0.75	0.55	0.50	1.00	0.67	0.20	0.25	0.22
CSC4615	0.71	1.00	0.83	0.67	1.00	0.80	0.50	1.00	0.67

Major takeaways

1. The models generally performed better (i.e., higher precision and fewer false positives) when the syllabus mapped fewer KUs (1 to 3 KUs).
2. When syllabi had more KUs mapped (6 or more KUs), the models tended to produce higher recall but suffered from lower precision due to increased false positives. This could be considered good or bad. While many KUs were detrimental to precision, the number of false positives allowed the SME to review other possibilities for the course. Sometimes,

dreaming can lead to innovative changes.

3. Feedback regarding key elements affecting mapping decisions was better in this method, allowing the SME to review their previous decisions. The number of mapped KUs also influenced the quality of this feedback.
4. Claude proved to be the most conservative in its mapping processes. ChatGPT provided the most feedback, providing a greater balance between precision and recall. Gemini suffered from accuracy issues due to continued context window challenges. However, it should be noted that its evaluations in Experiment 3 were competitive with the other models, given less stress on its context window.
5. It should be noted that CSC 3040 and CSC 4320 did not have any KU mappings in the human-generated document. Claude, in its conservative position, acknowledged this and did not provide any accuracy calculation information. ChatGPT and Gemini used inferred KUs and incorrectly forced a mapping in their accuracy reviews.

This discussion must also consider syllabus structure. As stated earlier, these documents were not edited before the experiment and were not identical in structure. Document complexity can contribute to the depletion of valuable context window space. A review of the syllabi yielded the following structural considerations to improve review:

Experiment 4 - Mapping All KUs to All Courses – Time Taken + Recall/Precision per KU for All Courses and Per Course for All KUs

Experiment 4 focused on the optimistic goal of "throwing everything in all at once, and beautiful things will come out." As we have seen the results from the other experiments, current context windowing schemes in the generally available commercial models will prevent this from being true. Only highly specialized models could begin to address this goal using techniques such as semantic compression [22], positional interpolation [18], or segment-based adjustment for rotary position embeddings [23]

Initially, the projects were provided all 16 syllabi—compiled into a single, optimized PDF totaling approximately 45 pages—and the KU documentation. Based on estimated token counts, this exceeded 60,000 tokens, which approached or surpassed the effective processing limits of the models. While ChatGPT and Claude attempted to produce output (with extensive hallucination), Gemini refused to process the prompt entirely, likely due to context window exhaustion.

Using a similar approach as Experiment 2b, a group of six syllabi and the KU documentation were included in the conversation. The results of each run were collected into a single document. This document was introduced into a second chat conversation, along with the SME mapping document, to mitigate further context window issues. The outcomes from this method were reasonable, with F1 scores ranging from 0.7692 to 0.8659. The limitation of this model was the quality of feedback regarding the decision process. Unlike Experiment 3, where rich detail was provided to the SME for review, this process yielded very basic mapping information that would not significantly increase SME insight or productivity.

Table 4: Summary of Structural Elements

Structural Element	**Impact on Mapping**	**Recommendation**
Detailed Descriptions	Improves precision and recall when detailed; hurts when vague	Ensure course descriptions include specific topics and technologies related to KUs
Explicit KU-Related Terms	Reduces false positives and false negatives	Use KU-related terminology explicitly in syllabi (e.g., secure coding, network defense)
Learning Outcomes	Helps models map outcomes directly to KUs	Clearly define learning outcomes and align them with KU terminology
Weekly/Topical Breakdown	Helps models assess depth of coverage	Include a detailed weekly or topical breakdown of the syllabus
Hands-On Activities	Improves recall for practical KUs	Specify hands-on activities with clear reference to skills and technologies related to KUs

Table 5: Experiment 4 - Precision, Recall, and F1 Scores

LLM	**Precision**	**Recall**	**F1 Score**
ChatGPT	0.8554	0.8765	0.8659
Claude	0.8214	0.8118	0.8166
Gemini	0.7831	0.7558	0.7692

Discussion

Discussion

The primary objective of these experiments was to evaluate the practical viability of LLM-assisted curriculum mapping in real-world academic settings, where document structures and quality vary significantly. Moving beyond idealized prompting scenarios, this study demonstrates how LLMs can effectively augment, rather than replace, human expertise. When used in tandem with subject matter experts (SMEs), LLMs can act as Collaborative Reasoning Assistants (CRAs), expediting curriculum alignment with minimal data preparation [24].

To guide practical adoption, this study introduces the CIRCLE methodology—a structured, iterative framework for integrating LLMs into curriculum mapping workflows. While CIRCLE provides a repeatable process for SME-guided alignment, institutions must still choose how to apply LLM assistance based on their goals, timeline, and resource constraints.

Each of the four experiments conducted in this study reflects a different balance between time required, mapping accuracy, feedback quality, and scalability. Table 6 summarizes these tradeoffs as a decision-support matrix to help curriculum directors and accreditation teams select an appropriate strategy.

Table 6: Decision-Support Matrix for Curriculum Mapping Strategy Selection

Experiment	**Time**	**Precision / Recall**	**Feedback**	**Use Case**
1 – Manual SME Mapping	High	High	High	Formal review or accreditation prep
2 – One KU to All Courses	Low–Moderate	Moderate	Moderate	Targeted check for KU coverage across curriculum
3 – All KUs to One Course	Moderate	High	Moderate	Balanced SME review at course level
4 – All KUs to All Courses	Low	Moderate	Low	Rapid triage or baseline scan

The matrix shows that Experiment 1 (manual SME mapping) yields the most comprehensive results but requires a substantial time investment. Experiment 2 offered a scalable but coarse-grained approach, suitable for identifying which courses may touch on specific KUs, though with higher variance in accuracy. Experiment 4 enables faster triage but provides limited interpretability. Experiment 3 offers a practical middle ground—balancing quality, feedback depth, and efficiency, making it a strong candidate for institutions adopting LLM-based tools for the first time.

Limitations and considerations

The experiments suggest that even lower-quality curriculum documents can yield useful feedback when modest improvements are made. These improvements enable more accurate comparisons to established frameworks and help surface potential new mappings. High recall values were observed in several cases, highlighting the LLMs' ability to identify the most relevant KUs. However, higher recall often comes at the cost of increased false positives, which in turn require additional SME review and filtering. The optimal balance between precision and recall will depend on factors such as the SME's time constraints and the complexity of the target framework.

Another key consideration is the potential for bias in existing course documents, which LLMs may inadvertently amplify. This risk can be mitigated by involving multiple reviewers or by integrating bias-review protocols into the alignment workflow. Finally, the structural design of syllabi plays a critical role in alignment accuracy. Institutions that adopt accessible, template-based formats—such as those with ADA-compliant headers—can not only improve machine readability for students but also enhance document ingestion and analysis by LLMs.

Precision vs. Recall Considerations.

- *For exploratory review:* Favor higher recall to surface potential gaps.
- *For formal accreditation prep:* Emphasize precision and documented justifications.

CIRCLE methodology

While not tested as an independent experiment, the **CIRCLE methodology** is a synthesis of best practices and patterns derived from the four experiments conducted in this study. It reflects the authors' applied experience in Agile software development and the iterative interaction observed between human experts and LLMs across varying mapping scopes. In particular, **Experiment 3** demonstrated the most effective balance of time, precision, recall, and actionable LLM output, serving as a practical model for this framework.

CIRCLE provides a human-in-the-loop methodology for aligning curriculum content with external frameworks through iterative LLM prompting. Its six phases are:

- **Collection** – Aggregate syllabi, outcomes, and supporting documentation. Quantity can partially compensate for quality in this phase.
- **Identification of Chunks** – Segment the collected content into manageable units. This phase is especially critical given LLM context window limitations.
- **Review of Background** – SME reviews or uploads the relevant framework(s) to ground the prompt.
- **Crafting Focused Prompts** – Refine prompts to target specific KU-course relationships. This phase was heavily exercised in Experiments 2 and 3.
- **Looking Back with Retrospective Prompting** – Use results from previous prompts to iterate on mappings. This improves SME reflection and was central to the analysis phase of Experiment 3.
- **Evaluating Results** – Use both SME insight and statistical comparisons (e.g., F1 score) to evaluate output. Precision vs. recall tradeoffs were most evident in Experiments 3 and 4.

As such, CIRCLE provides a scalable, reusable structure for SMEs seeking to incorporate LLMs into their curriculum review processes. While this framework has not yet been evaluated as a complete pipeline, the experiments exercised its components individually. Future work will seek to validate CIRCLE holistically as a standalone approach.

The results of this study affirm that LLMs can significantly reduce both the time commitment and cognitive burden associated with SME-led framework mapping. In our evaluation, the traditional manual process required approximately 40 hours, drawing from a previously developed five-year-old mapping report as a foundation. In contrast, the LLM-enhanced approach—including collecting data, configuring three LLM environments, iterating on prompts, and analyzing outputs—required just 17 hours.

Table 7: Time Investment Comparison Across Mapping Approaches

Approach	**Estimated Time Requirement**
Manual SME Mapping	40 hours (baseline, includes follow-up and historical review)
LLM-Enhanced (multi-model)	17 hours (includes all three models, prompt development, and output analysis)
Streamlined Using CIRCLE (Experiment 3)	<10 hours (expected for single-model KU-to-course mapping with refined prompts)

These reductions streamline curriculum review workflows and lay the groundwork for more frequent, data-informed, and iterative quality assurance cycles. By reducing SME time requirements without sacrificing mapping fidelity, institutions can shift from static, multi-year reviews to lighter, continuous updates driven by evolving standards.

It is worth noting that this timeline included overhead from exploratory prompt engineering and parallel execution of multiple experimental conditions. Based on these findings, we anticipate that future implementations of the CIRCLE methodology—centered around the streamlined practices observed in Experiment 3—could further reduce the required time while preserving or enhancing mapping accuracy.

While these results are promising, they should be interpreted as preliminary benchmarks. Time savings and output quality will vary based on curriculum complexity, document structure, and SME familiarity with LLM-assisted workflows. Nonetheless, the experiments provide strong evidence that CIRCLE offers a viable and scalable pathway for integrating LLMs into real-world curriculum alignment efforts.

Broader Impacts

Although this study focuses on cybersecurity education, the proposed framework can be adapted to other disciplines. The proposed CIRCLE method could benefit from fields such as healthcare and engineering, which also require strict compliance with established frameworks. The process leans on decades of experience with Agile software development models.

This application supports a broader framing of prompt engineering and project structuring as a

systems development activity, akin to Agile software practices [25]. This study advances the concept of "**prompting as coding**," framing prompt engineering and project scaffolding as systems development tasks. Drawing from Agile practices, this model encourages continuous iteration, structured documentation, and human oversight, paralleling how software teams manage modular complexity. Institutions can build reusable prompt structures that scale across curriculum domains by viewing LLM interactions as modular and repeatable processes. By adopting a standardized approach to prompt design and encouraging the use of accessible, templated syllabus structures (e.g., ADA-compliant formatting), institutions can scale this process to evaluate and align hundreds of courses efficiently. Future research could explore fine-tuning LLMs on institution-specific corpora of syllabi and KU mappings to improve precision and reduce false positives. Looking ahead, CIRCLE-inspired tooling could be embedded within Learning Management Systems (LMS) to provide real-time KU alignment feedback during course design, transforming LLMs from post-hoc reviewers into proactive curriculum development assistants.

Regarding institutional impacts, the ability to conduct LLM-enhanced reviews can introduce efficiencies in multiple areas. Curriculum designers can use the feedback to inform the development of new courses and the revision of existing courses. Institutional leadership can use the feedback to enhance high-level accreditation reviews and develop model continuous improvement processes across all units. As seen in the experiments, taking the summary of a review process and introducing that summary into a new conversation can yield new insights while remaining within the constraints of context windows. While these experiments were done with commercially available LLMs, consider the power of specially trained models created by institutions using highly curated document sets.

Furthermore, this study offers an implicit decision-support model for selecting curriculum mapping strategies based on competing priorities. By examining tradeoffs across *time*, *recall*, and *precision*, institutions can better tailor their LLM usage:

- **Experiment 1** maximizes precision and recall but requires the highest SME time investment.
- **Experiment 4** minimizes time but at the cost of feedback depth and interpretability.
- **Experiments 2 and 3** offer hybrid strategies; notably, **Experiment 3**'s KU-to-course mapping emerged as the best balance between insight, accuracy, and SME utility.

This insight enables educational institutions—especially those navigating evolving frameworks like CAE-CD or ABET—to select appropriate methods aligned with their internal constraints. Beyond education, similar tradeoffs exist in fields like *healthcare compliance*, *engineering accreditation*, and *critical infrastructure resilience*, where experts must map local policies or training programs to broad external standards. The combination of **CIRCLE** and this empirical tradeoff model offers a versatile decision-making aid.

Conclusion

In this paper, we have experimentally reviewed various techniques using LLMs as collaborative reasoning assistants to provide reasoning support to SMEs involved in framework mapping exercises. Our results have demonstrated that while LLMs are not a "data savior," they are effective tools to expedite the mapping process for an SME. While precise timing comparisons

are difficult due to the experimental conditions, results suggest a greater than 50% reduction in time to completion.

We have also proposed an Agile-influenced CIRCLE evaluation framework to further increase the productivity of the process. Aside from the inherent reduction in direct labor, part of the improvement of this model should be considered in terms of scalability. With proper chunking identification, the process could address much larger datasets or curricula with minimal increases in time and effort, enabling institutions to rapidly map new or updated courses to evolving frameworks without a proportional increase in human workload. Future iterations of this work could leverage models with expanded context windows or chunking strategies designed to optimize the use of context while preserving essential relationships between framework knowledge units.

Beyond education, the CIRCLE methodology can be used in other sectors, such as critical infrastructure and healthcare. The iterative, Agile approach facilitates initial discovery and refinement of application. As with education, the Collection and Identification will be the most involved as document repositories are initially developed.

A key insight emerging from this study is the multi-objective optimization inherent to curriculum mapping. Institutions must often balance time constraints, alignment fidelity, and review depth. This work contributes a structured understanding of how to navigate those tradeoffs, offering specific guidance for when to use manual review, targeted LLM evaluation (as in Experiments 2 and 3), or fully automated suggestions. The CIRCLE framework provides a generalizable, iterative methodology rooted in these findings—one that can guide practitioners through the nuanced process of aligning curriculum content to evolving standards using AI-augmented workflows.

References

[1] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257–285, 1988.

[2] R. M. Harden, "Curriculum mapping: A tool for transparent and authentic teaching and learning," *Medical Teacher*, vol. 23, no. 2, pp. 123–137, 2001.

[3] J. Simmons, "Integrating large language models in online course design." University of San Diego News Center, Feb. 2025. Available: `https://www.sandiego.edu/news/detail.php?_focus=95467`, Accessed: Apr. 26, 2025.

[4] S. Guizani, T. Mazhar, T. Shahzad, *et al.*, "A systematic literature review to implement large language model in higher education: issues and solutions," *Discover Education*, vol. 4, Feb. 2025. Published: Feb. 15, 2025, Accessed: Apr. 26, 2025.

[5] A. Ardebili, A. . Lezzi, M. . Pourmadadkar, M. Risk, A. A. Ardebili, M. Lezzi, and M. Pourmadadkar, "Risk assessment for cyber resilience of critical infrastructures: Methods, governance, and standards," *Applied Sciences 2024, Vol. 14, Page 11807*, vol. 14, p. 11807, 12 2024.

[6] S. M. Ali, A. Razzaque, M. Yousaf, and R. U. Shan, "An automated compliance framework for critical infrastructure security through artificial intelligence," *IEEE Access*, 2024.

[7] D. Guo, M. Shan, and E. K. Owusu, "Resilience assessment frameworks of critical infrastructures: State-of-the-art review," *Buildings 2021, Vol. 11, Page 464*, vol. 11, p. 464, 10 2021.

[8] K. Palaniappan, E. Y. T. Lin, and S. Vogel, "Global regulatory frameworks for the use of artificial intelligence (ai) in the healthcare services sector," *Healthcare 2024, Vol. 12, Page 562*, vol. 12, p. 562, 2 2024.

[9] R. Ejjami, "The future of learning: Ai-based curriculum development," *Int J Multidiscip Res*, vol. 6, 2024.

[10] A. R. Vargas-Murillo, I. N. M. D. L. A. Pari-Bedoya, and F. D. J. Guevara-Soto, "The ethics of ai assisted learning: A systematic literature review on the impacts of chatgpt usage in education," *ACM International Conference Proceeding Series*, pp. 8–13, 6 2023.

[11] P. Liu, "Trustworthy civil infrastructure inspection through human-machine intelligence integration," 10 2024.

[12] O.-M. C. Osazuwa and M. O. Musa, "The expanding attack surface: Securing ai and machine learning systems in security operations," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 2498–2505, 5 2024.

[13] A. Perrella, F. F. Bernardi, M. Bisogno, and U. Trama, "Bridging the gap in ai integration: enhancing clinician education and establishing pharmaceutical-level regulation for ethical healthcare," *Frontiers in Medicine*, vol. 11, 2024.

[14] OpenAI, "Chatgpt: Language model by openai." `https://openai.com/chatgpt`, 2023. Accessed: 2025-01-14.

[15] Anthropic, "Claude: An ai assistant by anthropic." `https://www.anthropic.com/index/claude`, 2023. Accessed: 2025-01-14.

[16] G. DeepMind, "Gemini: Multimodal ai model by deepmind." `https://www.deepmind.com/research/gemini`, 2023. Accessed: 2025-01-14.

[17] C. An, J. Zhang, M. Zhong, L. Li, S. Gong, Y. Luo, J. Xu, and L. Kong, "Why does the effective context length of llms fall short?," *arXiv preprint arXiv:2410.18745*, 2024.

[18] S. Chen, S. Wong, L. Chen, and Y. Tian, "Extending context window of large language models via positional interpolation," 2023.

[19] J. A. Yeow, P. K. Ng, K. S. Tan, T. S. Chin, and W. Y. Lim, "Effects of stress, repetition, fatigue and work environment on human error in manufacturing industries," *Journal of Applied Sciences*, vol. 14, pp. 3464–3471, 12 2014.

[20] G. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, pp. 81–97, 3 1956.

[21] H. Jin, X. Han, J. Yang, Z. Jiang, Z. Liu, C.-Y. Chang, H. Chen, and X. Hu, "Llm maybe longlm: Self-extend llm context window without tuning," 2024.

[22] W. Fei, X. Niu, P. Zhou, L. Hou, B. Bai, L. Deng, and W. Han, "Extending context window of large language models via semantic compression," 2023.

[23] R. Li, J. Xu, Z. Cao, H.-T. Zheng, and H.-G. Kim, "Extending context window in large language models with segmented base adjustment for rotary position embeddings," *Applied Sciences 2024, Vol. 14, Page 3076*, vol. 14, p. 3076, 4 2024.

[24] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," in *AI Magazine*, vol. 35, pp. 105–120, Association for the Advancement of Artificial Intelligence, 2014.

[25] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," *arXiv preprint arXiv:2102.07350*, 2021.