

# Evaluating 21st-Century Competencies in Postsecondary Curricula with Large Language Models: Performance Benchmarking and Reasoning-Based Prompting Strategies

Zhen Xu<sup>1</sup>, Xin Guan<sup>2†</sup>, Chenxi Shi<sup>3†</sup>, Qin hao Chen<sup>4</sup>, Renzhe Yu<sup>5\*</sup>

## Abstract

The growing emphasis on 21st-century competencies in postsecondary education, intensified by the transformative impact of generative artificial intelligence (GenAI) on the economy and society, underscores the urgent need to evaluate how they are embedded in curricula and how effectively academic programs align with evolving workforce and societal demands. Curricular analytics, particularly recent advancements powered by GenAI, offer a promising data-driven approach to this challenge. However, the analysis of 21st-century competencies requires pedagogical reasoning beyond surface-level information retrieval, and the capabilities of large language models (LLMs) in this context remain underexplored. In this study, we extend prior research on curricular analytics of 21st-century competencies across a broader range of curriculum documents, competency frameworks, and models. Using 7,600 manually annotated curriculum-competency alignment scores (38 competencies and 200 courses across five curriculum document types), we evaluate the informativeness of different curriculum document sources, benchmark the performance of general-purpose LLMs on mapping curricula to competencies, and analyze error patterns. We further introduce a reasoning-based prompting strategy, curricular chain-of-thought (CoT), to strengthen LLMs' pedagogical reasoning. Our results show that detailed instructional activity descriptions are the most informative type of curriculum document for competency analytics. Open-weight LLMs achieve accuracy comparable to proprietary models on coarse-grained tasks, demonstrating their scalability and cost-effectiveness for institutional use. However, no model reaches human-level precision in fine-grained pedagogical reasoning. Our proposed curricular CoT yields modest improvements by reducing bias in instructional keyword inference and improving the detection of nuanced pedagogical evidence in long text. Together, these findings highlight the untapped potential of institutional curriculum documents and provide an empirical foundation for advancing AI-driven curricular analytics.

## Notes for Practice

- Existing curriculum documents in digital formats are a promising yet underutilized resource for postsecondary institutions to assess the integration of 21st-century competencies in course design. Institutions may benefit the most from combining multiple data sources to ensure adequate coverage of key pedagogical elements.
- Large language models (LLMs) perform reasonably well in coarse-grained competency analytics, with open-weight models achieving performance comparable to proprietary ones. However, none of the models reach human-level precision, with two common error patterns: biased inference from instructional keywords in short texts and difficulty identifying nuanced pedagogical evidence in long, unstructured documents.
- The proposed curricular chain-of-thought (CoT) prompting yields modest performance gains and reduces hallucination by guiding LLMs to extract key instructional elements prior to reasoning, improving the reliability of pedagogical inference, particularly for more advanced models.

## Keywords

Curricular analytics, 21st-century competency, large language models (LLMs), prompt engineering, chain of thought (CoT).

**Submitted:** 09/02/2025 — **Accepted:** 22/12/2025 — **Published:** 25/02/2026

<sup>1</sup> Email: [zx2393@tc.columbia.edu](mailto:zx2393@tc.columbia.edu) Address: Teachers College, Columbia University, New York, New York, United States. ORCID iD: <https://orcid.org/0009-0004-3131-910X>

<sup>2</sup> Email: [xg2413@tc.columbia.edu](mailto:xg2413@tc.columbia.edu) Address: Teachers College, Columbia University, New York, New York, United States. ORCID iD: <https://orcid.org/0009-0001-6055-1555>

<sup>3</sup> Email: [cs4477@tc.columbia.edu](mailto:cs4477@tc.columbia.edu) Address: Teachers College, Columbia University, New York, New York, United States. ORCID ID: <https://orcid.org/0009-0009-1630-7078>

<sup>4</sup> Email: [qc2354@columbia.edu](mailto:qc2354@columbia.edu) Address: Graduate School of Arts and Sciences, Columbia University, New York, New York, United States. ORCID ID: <https://orcid.org/0009-0001-9105-6899>

<sup>5</sup> \*Corresponding author. Email: [renzheyu@tc.columbia.edu](mailto:renzheyu@tc.columbia.edu) Address: Teachers College and Data Science Institute, Columbia University, New York, United States. ORCID ID: <https://orcid.org/0000-0002-2375-3537>.

<sup>†</sup> Equal contribution.

## 1. Introduction

Globalization and rapid technological advancements, particularly the recent emergence of generative artificial intelligence (GenAI), are profoundly transforming the labour market and the broader economic landscape. As AI automates routine and repetitive tasks, human labour is expected to shift toward roles that require higher-order cognitive skills, creativity, and social-emotional intelligence (OECD, 2023c). In this context, foundational knowledge and technical skills are no longer sufficient for future workforce readiness. Employers now place growing importance on a broad set of transferable, domain-agnostic competencies, commonly referred to as 21st-century competencies, such as critical thinking, complex problem-solving, collaboration, communication, adaptability, and digital literacy (OECD, 2023a). Recognizing these global shifts, policymakers and organizations worldwide are underscoring the urgency of equipping future generations with 21st-century competencies to meet the demands of an increasingly digital and AI-driven economy (OECD, 2023c; McKinsey Global Institute, 2023; World Economic Forum, 2025).

Correspondingly, these shifts carry profound implications for teaching and learning, especially in postsecondary education, where academic preparation directly connects to workforce readiness. Postsecondary education institutions are increasingly expected to integrate and foster higher-order, interdisciplinary 21st-century competencies within their traditional subject-based curricula (UNESCO, 2023; OECD, 2023b; National Education Association, 2024). In response, many countries have introduced new teaching and learning frameworks to guide systematic reforms and embed these competencies into national learning outcomes, encouraging institutions to align with evolving societal and workforce needs (Office of Educational Technology, 2023; Department for Education, 2023). However, institutions still face challenges in translating these policy expectations into practice (OECD, 2023a; Buckingham Shum & Crick, 2016), particularly in redesigning curricula and instruction. Effective reform requires faculty training and pedagogical support, digital and technical infrastructure, and flexible institutional policies—resources that many institutions lack or possess unevenly. Therefore, examining curriculum design and implementation at scale is critical to understanding how institutions are responding to the evolving demands of 21st-century education and adapting to future workforce preparation.

Traditional curricular analytics and evaluation have long relied on experts' manual analysis, which is labour-intensive and difficult to scale. With ongoing advancements in postsecondary education data infrastructure and analytical techniques, more institutions and stakeholders are seeking automated, data-driven approaches. In this context, curricular analytics is a promising approach that leverages computational methods to support data-driven curriculum decision-making, program improvement, and institutional evaluation (Hilliger et al., 2020). By leveraging digital records, curricular analytics enables scalable and systematic analysis of various dimensions of curriculum and instruction, such as learning objectives, pedagogical design and practices, learning outcomes, and their alignment with workforce competencies (Pistilli & Heileman, 2017; Chou et al., 2015; Hilliger et al., 2024).

However, the analysis of 21st-century competencies in postsecondary curricula remains at an early stage of exploration, primarily due to technical and data limitations. Technically, it requires pedagogically informed reasoning to interpret learning activities and instructional design, shifting the task beyond traditional natural language processing (NLP) approaches (e.g., rule-based matching or embedding-based alignment) toward pedagogical reasoning. Although supervised machine learning offers partial solutions, it depends on large, manually annotated datasets that are scarce in postsecondary education, limiting scalability. Moreover, from a data quality perspective, such analysis also requires detailed, high-quality curriculum documents, as overly general materials often lack sufficient information. As a result, despite institutions maintaining extensive collections of curriculum documents, their potential to generate actionable insights into competency development remains largely untapped.

Recent advancements in GenAI, particularly large language models (LLMs), show promise for addressing these challenges. Some recent studies evaluate LLMs' ability to analyze higher-order interdisciplinary competencies in domains such as business and information science (Jayalath et al., 2025; Jovanović et al., 2025) and suggest that LLMs show initial promise. However, other work highlights persistent limitations, especially for tasks requiring pedagogically grounded reasoning and interpretation of instructional design (Zamecnik et al., 2024; Xu et al., 2025). Without careful prompt design, fine-tuning, or contextual adaptation, LLMs struggle to reliably perform complex pedagogical reasoning tasks. Therefore, fully realizing the potential of LLMs for assessing 21st-century competencies in postsecondary curricula requires systematic cross-contextual evaluation to identify when LLMs are effective and how their pedagogical reasoning and contextual inference can be improved.

This study aims to advance AI-assisted curricular analytics by systematically examining the capabilities of LLMs for analyzing 21st-century competencies in postsecondary curricula. We extend prior work to a broader setting that spans multiple types of curriculum documents, academic disciplines, and competency frameworks. Our goal is to assess LLMs' current capabilities and limitations and to identify strategies for improving LLM-assisted pedagogical reasoning in curricular analytics. Specifically, we address the following research questions:

- RQ1:** What types of curriculum documents are more informative for conducting 21st-century competency analytics?
- RQ2:** What is the performance of zero-shot LLMs as a baseline for 21st-century competency analytics tasks, and what are the challenges?
- RQ3:** Do reasoning-based prompting strategies address these challenges and improve the reliability of LLM-based 21st-century competency analytics?

The contribution of this study is threefold. First, we evaluate the feasibility of analyzing 21st-century competencies using different types of digital curriculum documents in postsecondary institutions, providing empirical evidence and practical guidance for data-driven competency analytics. Second, we establish baseline performance benchmarks for off-the-shelf LLMs across diverse curriculum document types and varying levels of conceptual specificity in competency frameworks, highlighting systematic limitations in current models. Third, we introduce a reasoning-based prompt strategy that enhances LLM performance in curricular analytics. Together, these contributions advance the empirical and methodological foundations for using GenAI to support large-scale, evidence-based curricular analytics in postsecondary education.

## 2. Related Work

### 2.1 Mapping 21st-Century Competencies in Postsecondary Curricula

Postsecondary institutions are increasingly expected to prepare graduates not only with disciplinary knowledge but also with a broad set of competencies necessary to navigate an increasingly complex and dynamic workforce (OECD, 2023a). These competencies, often referred to as *21st-century competencies*, *21st-century skills*, or *future skills*, are not singular task-specific skills but a diverse, interdisciplinary set of transferable capabilities (Griffin et al., 2012). Over the past decade, the 21st-century competencies have been a dynamic and rapidly evolving concept, as various governmental, academic, non-profit, and corporate entities have accelerated the revision of regional frameworks to better respond to evolving economic development, societal priorities, and technological advancements. A systematic review from Cambridge (Kotsiou et al., 2022) identified 99 21st-century competency frameworks developed since 2010, encompassing 341 distinct terms and constructs. Although there is no single shared definition, several competencies commonly found across frameworks include creativity and innovation, critical thinking, problem-solving, communication and collaboration, and information and communication technology (ICT) literacy, among others.

Despite the growing institutional commitment, the systematic integration of 21st-century competencies into curricula remains relatively limited and uneven (OECD, 2023a, 2023b; Buckingham Shum & Crick, 2016). The absence of a shared definition and understanding of how such competencies develop across disciplines and contexts impedes curriculum reform. Practical challenges further constrain progress, such as difficulties in assessing 21st-century competencies, limited faculty training and pedagogical support, resource constraints that restrict the development of active learning environments, and rigid program structures that limit curricular innovation and skill-oriented instructional approaches. Therefore, while the value is widely recognized, more systematic and scalable approaches are needed to assess how effectively these 21st-century competencies are embedded within and across curricula and to generate evidence to inform continuous curriculum improvement and institutional decision-making.

In this context, curricular analytics is a promising approach to addressing these challenges. As a subfield of learning analytics, curricular analytics leverages computational approaches to provide evidence-based insights to drive curriculum decision-making and program improvement (Dawson & Hubball, 2014; Greer et al., 2016). Over the past few decades, a growing body of curricular analytics research has leveraged curriculum documents such as textbooks (Li et al., 2024; Yang et al., 2023); general catalogues (Jovanović et al., 2025; Ohland & Collins, 2002; Irwin, 2002; Walker, 2024; Light, 2024); syllabi (Zamecnik et al., 2024; Fiesler et al., 2020; Gorski, 2009; Hong & Hodge, 2009; Homa et al., 2013); learning management system records (Shorman, Khder, et al., 2024) to explore various aspects of curriculum design and delivery topics, such as learning objectives and workforce alignment (Doyle et al., 2025; Tan & Lim, 2023; Lohr et al., 2025; Nguyen et al., 2024; M. Zhang et al., 2022; Nguyen et al., 2024; Zamecnik et al., 2024; Javadian Sabet et al., 2024, 2024; Zamecnik et al., 2024; Walker, 2024); pedagogical strategies (Liu et al., 2024; Kozov et al., 2024; Lyu et al., 2024); course delivery (Tang & Sae-Lim, 2016); and assessment methods (Jovanović et al., 2025).

Although skill coverage has long been a well-established topic in curricular analytics, prior research has primarily focused on domain- and task-specific skills. For example, Kawintiranon and colleagues (2016) analyze the coverage of 13 computer science-specific knowledge areas in the Computer Engineering Curricular Guideline (CE2016) (Durant et al., 2015) from the course descriptions of five universities. Similarly, studies such as Light (2024), Walker (2024), and Kitto and colleagues (2020) map course descriptions to job skill profiles derived from Lightcast (formerly Burning Glass Technologies) using open-access general catalogs. Other representative studies, such as Javadian Sabet and colleagues (2024), align curriculum texts with detailed work activities from O\*NET, a U.S. Department of Labor database containing thousands of real-world task descriptions across occupations. Recent studies have explored interdisciplinary and higher-order competencies in curricula using LLMs, as these competencies require contextual inference beyond surface-level semantic matching. Jovanovic and colleagues (2025), Jayalath and colleagues (2025), and Zamecnik and colleagues (2024) analyze enterprise skills and graduate qualities such as self-management, problem-solving, ethical awareness, oral communication, international perspective, and so on. These early efforts highlight both the potential and key challenges in assessing 21st-century competencies: results depend heavily on the quality and informativeness of curriculum documents, model performance varies across skills (Zamecnik et al., 2024), and LLMs often struggle with competencies that are implicit or require deep pedagogical interpretation. Building on the gaps identified in previous single-framework and single-domain studies, these findings underscore the need for broader, cross-domain investigations that integrate multiple data types, disciplines, and competency models.

## 2.2 GenAI for Curricular Analytics

The analysis of 21st-century competencies is closely tied to advances in text mining techniques. Early approaches relied on rule-based methods or keyword matching, which struggled to capture complex constructs. The introduction of transformer-based models marked a significant shift, enabling semantic analysis through contextual embeddings and improving performance for well-defined, domain-specific skills. However, these methods remain limited when addressing abstract competencies in 21st-century frameworks, which require not only semantic alignment but also pedagogical inference from course materials, such as learning activities and assessments.

The advent of GenAI has expanded these possibilities. Among various technical strategies, zero-shot prompting with carefully designed task guidance prompts has been examined to evaluate the baseline capability of LLMs in performing competency analytics tasks. For instance, Sridhar and colleagues (2023) and Jayalath and colleagues (2025) evaluated GPT-4 in zero-shot settings for tasks such as generating learning objectives and mapping seven enterprise skills to assessments. They found that while performance varied across different skills, GenAI achieved an overall alignment comparable to expert annotations. However, some other studies report more limited performance. For example, Zamecnik and colleagues (2024) compared zero-shot and retrieval-augmented generation (RAG) methods for extracting seven graduate attributes from curriculum documents, while Xu and colleagues (2025) extended this by evaluating multiple LLMs across four types of curriculum documents and comparing the performance against traditional NLP approaches. Their findings show that although RAG improves accuracy for well-defined tasks, LLMs in zero-shot mode alone remain unreliable for consistent skill-level analysis. Parallel efforts in the labour market domain, using GenAI to extract skills from job postings and resumés (Nguyen et al., 2024), illustrate faster progress due to the availability of standardized taxonomies and large-scale annotated corpora (Decorte et al., 2022). This has enabled experimentation beyond prompt-based methods (Nguyen et al., 2024; Senger et al., 2024), including supervised learning and fine-tuning approaches (Thakrar & Young, 2025; Herandi et al., 2024). Recent studies have also leveraged synthetic data generation and data augmentation to improve the recognition of rare or implicit skills (Decorte et al., 2022; Senger et al., 2024).

In summary, while GenAI-based approaches are expanding the analytical capacity of curricular analytics, current practices remain largely prompt-based due to the scarcity of benchmark datasets for postsecondary education. The key technical gap lies in the current LLMs' limited ability to perform in-context reasoning about how competencies are developed and demonstrated within curricular documents. The challenge stems from both the abstract nature of 21st-century competencies and the heterogeneous structure and quality of curriculum documents (Nguyen et al., 2024; Herandi et al., 2024). So technically, this makes higher-order competency analytics fundamentally different from prior task-specific skill mapping, shifting the problem space from information retrieval and classification toward contextual inference and pedagogical reasoning.

## 2.3 Reasoning-Based Prompt Engineering

Among various prompting strategies, a line of research focuses on enhancing LLM reasoning through prompt engineering, which is directly relevant to our analytical scenario. This line of work began with the chain-of-thought (CoT) approach, inspired by the “scratchpad” idea in Nye and colleagues (2021), in which models improve on multi-step tasks by showing intermediate steps. Building on this idea, Wei and colleagues (2022) introduced standard CoT prompting, in which few-shot examples guide the model to output step-by-step reasoning, followed by the final answer. This significantly improved performance on math problems, such as those in the GSM8K benchmark. Later, Kojima and colleagues (2022) showed that simply adding “Let’s think step by step” enabled zero-shot CoT, achieving performance close to few-shot versions. To reduce reliance on

human-written examples, Zhang and colleagues (2022) proposed Auto-CoT, using an LLM to generate its own reasoning traces. Least-to-most prompting (Zhou et al., 2022) breaks down complex problems into simpler sub-problems, improving results over standard CoT. Tree-of-thought prompting (Yao et al., 2023) explores multiple reasoning paths like a search tree, helping with planning and strategy tasks.

More recently, prompting approaches that explicitly decompose inputs into key semantic components have emerged. For example, summary CoT (SumCoT) (Wang et al., 2023) encourages models to summarize key contextual elements (e.g., who, what, when, why) before generating final responses, while rephrase-and-respond (RaR) (Deng et al., 2024) guides models to restate questions in semantically distinct yet complementary ways before answering, thereby enhancing reasoning diversity and robustness. These methods have a common intuition: structuring model reasoning around core semantic components can support a more systematic interpretation of complex inputs. This intuition is particularly relevant in curricular analytics. From a learning sciences and pedagogical perspective, 21st-century competencies are developed through instructional strategies and learning activities that are intentionally designed and context dependent (Retnawati et al., 2018; Ghanizadeh et al., 2020). As a result, meaningful pedagogical inference depends on identifying and interpreting the instructional components that encode these learning intentions. This perspective aligns with curricular analytics tasks, in which valid reasoning requires recognizing and synthesizing key pedagogical information embedded in course materials.

### 3. Study Context

Figure 1 shows the overall design of this study. For **RQ1**, we identify representative types of curriculum documents and conduct a rigorous human annotation process to assess their value for 21st-century competency analytics. This process yields a benchmark dataset of 7,600 pairs of curriculum-competency alignment scores. For **RQ2**, we evaluate the performance of zero-shot LLMs as a baseline on this benchmark through quantitative and qualitative analyses, focusing on limitations in analytical granularity and sensitivity to curriculum and competency characteristics. Finally, for **RQ3**, we assess the effectiveness of proposed reasoning-based prompting strategies in improving LLMs’ competency analytics performance by addressing the identified limitations.

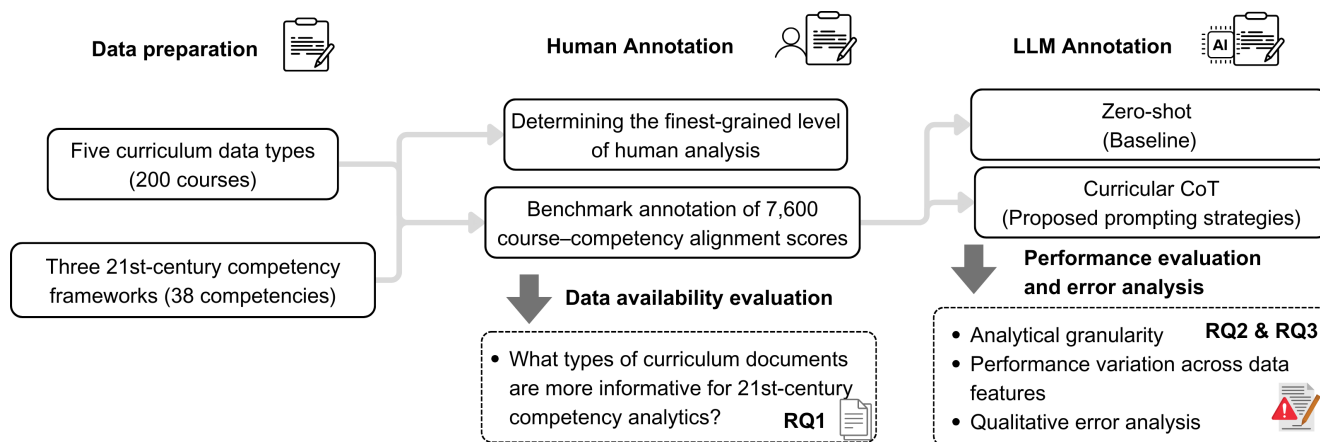


Figure 1. Overall study design.

### 3.1 Data

#### 3.1.1 Curriculum Documents

We draw on De Silva and colleagues (2024), which provides a systematic review of data used in curricular analytics, to identify representative types of digital curriculum documents commonly available at postsecondary institutions. Table 1 summarizes five types of curriculum documents, categorized by pedagogical element coverage and data source.

**Table 1.** Overview of curriculum document types for curricular analytics in postsecondary education.

| Data type                   | Data source                       | Potential pedagogical element coverage   |
|-----------------------------|-----------------------------------|--|
| Concise course description  | General catalogue                 | High-level course themes, intended learning goals, overall instructional approach, general assessment plan   |
| Detailed course description | Course syllabus                   | Course topics, articulated learning goals, planned learning activities, instructional or delivery format, assessment plan                          |
| Learning objective          | Course syllabus                   | Explicit learning goals, such as knowledge, skills, and competencies   |
| Instructional schedule      | Course syllabus                   | Temporal organization of course content, sequencing of instructional activities, types of learning activities, assessment plan                     |
| Learning activity content   | Learning management system record | Specific learning tasks and prompts, instructional materials, formats of learning activities, detailed assessment components, and grading criteria |

We then create samples of these curriculum document types from three distinct sources:

- *course syllabi* from Open Syllabus<sup>1</sup>, a nonprofit organization maintaining an archive of over 20.9 million course syllabi from postsecondary education institutions worldwide;
- *general catalogue* from publicly accessible websites of two institutions in the United States: a large, urban, public two-year college and a public four-year university;
- *learning management system* data from a public four-year university in the United States, including instructor-generated prompts and instructions for various learning activities as logged within the Canvas learning management system.

From the general catalog, we extract short-form descriptions and basic metadata of individual courses, including the course title and department. From course syllabi, we extract overall descriptions, learning goals, and activity schedules for each course. From learning management system data, we retrieve full instructional prompts and content for learning activities (e.g., assignments, discussions, and quizzes) and concatenate them at the course level. For comparability across data types, we apply stratified sampling by subject area and document length for each data type, followed by a manual review to remove low-quality or incomplete samples. This process yields five datasets, each containing 40 course-level samples, for a total of 200 curriculum documents.

### 3.1.2 Competency Frameworks

We select 21st-century competency frameworks that were (1) published after 2010, (2) issued by authoritative organizations (e.g., government agencies or international institutions), (3) applicable to postsecondary education, and (4) characterizing general, cross-disciplinary competencies rather than domain-specific skills. Table 2 summarizes the three frameworks selected for this study, which vary in definitional granularity. More detailed information on these frameworks can be found in the Appendix Tables 8, 10, 11, and 12.

**Table 2.** Overview of 21st-century competencies frameworks included in this study.

| Competency framework                        | Country        | Year | Education/labour market | Count of competencies |
|---|----------------|------|-------------------------|-----------------------|
| O*NET (Workforce Competencies) <sup>2</sup> | US             | 2013 | labour market           | 21                    |
| EU (Key Competences) <sup>3</sup>           | European Union | 2018 | education               | 8                     |
| ESDC (Success Model) <sup>4</sup>           | Canada         | 2024 | labour market           | 9                     |

## 3.2 Human Annotation

### 3.2.1 Rubrics for Curriculum-Competency Mapping

We recruited two graduate students in education with expertise in curriculum and instructional design. During the tutorial session, the annotators were introduced to the task, and each received three sample curriculum documents from each of the five curriculum document types (15 in total), which is a holdout set to avoid circularity between rubric development and model evaluation. They independently annotated the samples and refined the author’s draft rubric, adjusting its levels and descriptions to enhance the alignment between competencies and course documents. Finally, the author and two annotators jointly finalized

<sup>1</sup><https://www.opensyllabus.org/>

<sup>2</sup><https://files.eric.ed.gov/fulltext/EJ1109948.pdf>

<sup>3</sup><https://op.europa.eu/en/publication-detail/-/publication/297a33c8-a1f3-11e9-9d01-01aa75ed71a1/language-en>

<sup>4</sup><https://www.srdc.org/project/Research-report-to-support-the-launch-of-Skills-for-Success-Structure-evidence-and-recommendations-Final-report/>

a version that defines the finest level of analytical granularity achievable by human raters across all data and framework types, as shown in Table 3.

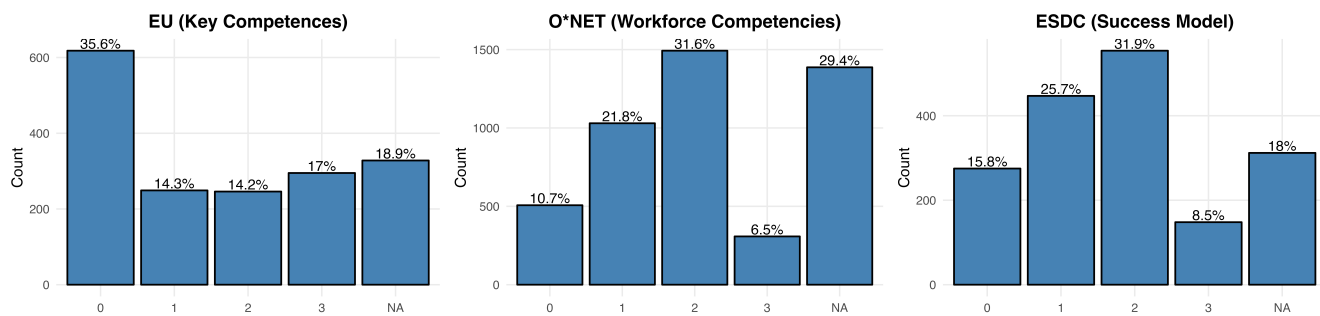
**Table 3.** Curriculum-competency alignment rubric.

| Score | Description  |
|-------|--|
| 3     | The competency is explicitly and clearly stated as a course objective.                           |
| 2     | The competency can be reasonably inferred from the course document.                              |
| 1     | The competency is vaguely implied but not clearly emphasized.                                    |
| 0     | The competency is clearly unrelated to the course.   |
| NA    | The competency may be relevant, but there is insufficient information to determine its presence. |

### 3.2.2 Annotation and Inter-rater Reliability

Before the formal annotation, we conducted a calibration round in which two annotators annotated 15 curriculum documents from each of the five document types (45 samples) for each of the three competency frameworks (38 competencies), which is 1,710 curriculum-competency pairs in total. The two annotators achieved a Cohen’s  $\kappa$  of 0.841 in the framework of EU (Key Competences), a  $\kappa$  of 0.288 in the framework of O\*NET (Workforce Competencies), and a  $\kappa$  of 0.168 in the framework of ESDC (Success Model). Follow-up discussions revealed that most disagreements stem from differing interpretations of whether course context (e.g., subject matter and course type) should inform competency inference, leading to confusion between codes 0 (clearly irrelevant) and NA (insufficient information). We therefore refined the coding guidelines to explicitly incorporate course context: competencies clearly outside the instructional scope were coded as 0, whereas potentially relevant competencies lacking sufficient textual evidence were coded as NA. After aligning this understanding, the two annotators re-annotated all samples in the calibration round. The final inter-rater reliability on these two frameworks is  $\kappa = 0.94$  in O\*NET (Workforce Competencies) and  $\kappa = 0.92$  in ESDC (Success Model).

Then, the two annotators independently worked on the remaining dataset, scoring each competency in each framework for each document. To maintain annotation reliability during this phase, random sampling and periodic cross-checks were implemented. Specifically, each annotator periodically reviewed a small subset of the other’s annotations (approximately 10%–15%), and any discrepancies were resolved through discussion. This annotation process yielded a benchmark dataset comprising 200 course documents annotated across the three frameworks (38 competencies in total), resulting in 7,600 pairs of curriculum-competency alignment scores. Figure 2 presents the distribution of human annotations, showing the count and percentage of each rubric score across the five data types. A detailed summary of the annotation distribution for each data type is provided in Table 9 in the Appendix.



**Figure 2.** Distribution of human annotation scores across competency frameworks.

## 3.3 LLM Annotation

### 3.3.1 Models

We evaluated LLM performance on the 7,600 human-annotated curriculum-competency alignment pairs using four models: GPT 3.5-turbo and GPT-4o (OpenAI, proprietary) and Llama-3-70B and Llama-3-8B (Meta AI, open-weight). In all experiments, the generation temperature was set to zero.

### 3.3.2 Prompting Strategies

We examined six prompting strategies: a zero-shot baseline, zero-shot prompting with competency definitions, and four variants of our proposed reasoning-based approach. Full prompt templates are provided in Figure 4.

1. **Zero-shot (ZERO):** The LLM is provided with the course document content and a list of competency names. It reflects the model’s default reasoning ability without any additional scaffolding or contextual support.

2. **Zero-shot prompting with competency definitions (DEF):** The model is provided with the course document, competency names, and corresponding definitions from human annotator notes. This strategy aims to reduce competency ambiguity and improve interpretive consistency by grounding model reasoning in explicit human-defined criteria.
3. **Curricular CoT:** Building on prior work (Wang et al., 2023; Deng et al., 2024), we introduce a structured reasoning-based prompting strategy for curricular analysis. The model follows a two-step reasoning process, illustrated in Figure 3:
  - *Element extraction:* Key pedagogical components are extracted using guided questions (Table 4) grounded in curriculum design theory.
  - *Standardized content representation:* The extracted elements are synthesized into a structured summary that reorganizes unstructured curriculum content into a consistent representation, which is then used in the final competency evaluation prompt.

This design mitigates traditional challenges in curricular analysis arising from heterogeneous document granularity (Arafeh, 2016), inconsistent wording and structure (Tian et al., 2024), and long or variable input length. We further distinguish the following variants of this prompting strategy.

- (a) **Curriculum + Questions + Answers (CQA):** The LLM receives the original curriculum document, competency names, guided questions, and the corresponding summarized answers and is instructed to reflect on this information before analyzing competency alignment.
- (b) **Curriculum + Questions (CQ):** The LLM receives the original curriculum document, competency names, and guided questions and reflects on the questions before analyzing competency alignment.
- (c) **Questions + Answers (QA):** The LLM receives the summarized curriculum content, competency names, and guided questions and reflects on the questions and summarized content before analyzing competency alignment.
- (d) **Answers Only (A):** The LLM receives only the summarized curriculum content and competency names to analyze competency alignment.

**Table 4.** Guided questions by curriculum document type.

| Course Description  | Learning Objectives   | Learning Activities/Instructional Schedule   |
|---|---|--|
| 1. <b>Course focus:</b> What is the primary content focus of the course?  | 1. <b>Target knowledge:</b> What domain-specific knowledge are students expected to acquire?                              | 1. <b>Activity summary:</b> What is the learning activity (one sentence)?                          |
| 2. <b>Core knowledge and skills:</b> What key knowledge and skills are emphasized in the course?                | 2. <b>Target skills:</b> What skills are students expected to develop?  | 2. <b>Activity type:</b> What type of activity is it (e.g., discussion, assignment, project, lab)? |
| 3. <b>Primary learning tasks:</b> What major tasks are students expected to complete?                           | 3. <b>Expected performance:</b> What tasks or capabilities should students be able to demonstrate upon course completion? | 3. <b>Target knowledge and skill:</b> What knowledge and skill does the activity address?          |
| 4. <b>Instructional format:</b> How is the course delivered (e.g., lecture, discussion, project-based, hybrid)? |   | 4. <b>Student deliverable:</b> What is the expected student output for this activity?              |
| 5. <b>Assessment approach:</b> How is student learning evaluated in the course?                                 |   | 5. <b>Assessment method:</b> How is performance on this activity evaluated?                        |

### 3.4 Performance Evaluation

#### 3.4.1 Evaluation Metrics

In this study, model performance is evaluated using accuracy and macro-averaged precision, recall, and F1 score. Macro-averaged metrics are reported to account for class imbalance in the ground-truth data, as shown in Figure 2 and detailed in Table 9 in the Appendix. To capture the ordinal structure of the labels and assess agreement with human annotators, we additionally report the weighted Cohen’s kappa and the intraclass correlation coefficient (ICC).



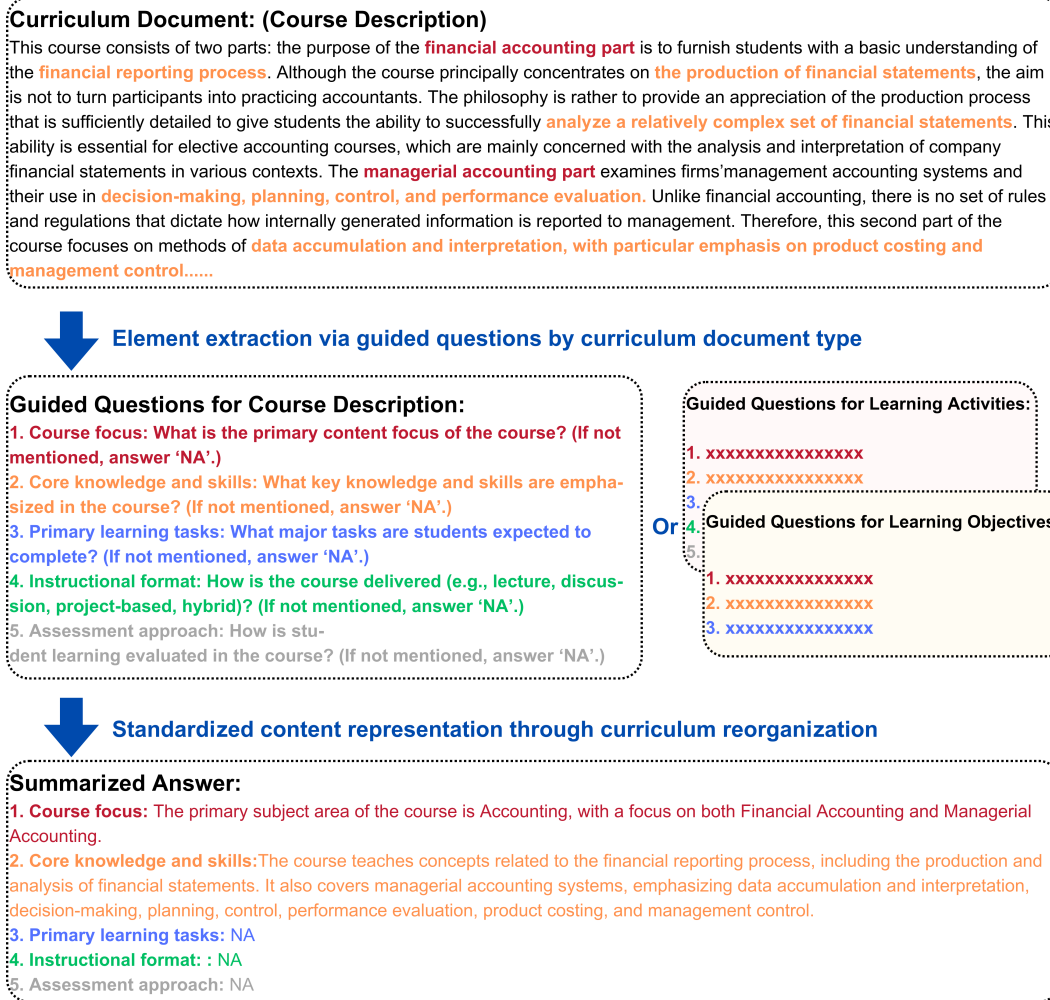


Figure 3. Full pipeline and illustrative example of curricular CoT.

### 3.4.2 Analytical Granularity

We further examine LLMs' ability to distinguish fine-grained representations of 21st-century competencies in curriculum documents. Table 5 summarizes the hierarchical taxonomy used for performance evaluation, in which detailed competency categories are progressively merged into coarser levels. We first evaluated performance at the finest granularity that human annotators could reliably achieve: a five-class classification task capturing nuanced skill distinctions. We then gradually reduced the level of granularity to four-, three-, and two-class settings to assess how performance changes as task complexity decreases. We adopted an aggregation-based evaluation approach, collapsing predictions from the five-level taxonomy into coarser levels before computing performance metrics.

Table 5. Description of granularity levels used for LLM classification evaluation.

| 5-Level                      | 4-Level                  | 3-Level                   | 2-Level                    |
|------------------------------|--------------------------|---------------------------|----------------------------|
| 3: Explicitly stated         | 3: Explicitly stated     | 3: Explicitly stated      | 1 & 2 & 3: Broadly Covered |
| 2: Reasonably inferred       | 2: Reasonably inferred   | 1 & 2: Partially inferred |                            |
| 1: Possibly implied          | 1: Possibly implied      |                           |                            |
| 0: Unrelated                 | 0 & NA: Unrelated or     | 0 & NA: Unrelated or      | 0 & NA: Unrelated or       |
| NA: Insufficient information | Insufficient information | Insufficient information  |                            |

**ZERO**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

[Competency 1]  
[Competency 2]  
...  
Course title: []  
Course field: []  
Course document content: []

**DEF**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

**[Competency 1]: [Definition from human annotator notes]**  
**[Competency 2]: [Definition from human annotator notes]**  
...  
Course title: []  
Course field: []  
Course document content: []

**Curricular CoT (CQA)**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

[Competency 1]  
[Competency 2]  
...  
Course title: []  
Course field: []  
**Original course document content: []**

**Here are key questions to think about before: [Guided Questions]**

**And this is the content summary of the curriculum document: [Summarized Answer]**

**Curricular CoT (CQ)**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

[Competency 1]  
[Competency 2]  
...  
Course title: []  
Course field: []  
**Original course document content: []**

**Here are key questions to think about before: [Guided Questions]**

**Curricular CoT (QA)**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

[Competency 1]  
[Competency 2]  
...  
Course title: []  
Course field: []

**Here are key questions to think about before: [Guided Questions]**

**This is the content summary of the curriculum document: [Summarized Answer]**

**Curricular CoT (A)**

You are a curriculum analyst. For the course content provided, score each of the skills below using the following rubric: [rubric]

Your response must include all skills listed below. Respond using only the format: Skill: Score (e.g., Critical thinking: 3). Do not include explanations or any other text.

[Competency 1]  
[Competency 2]  
...  
Course title: []  
Course field: []

**This is the content summary of the curriculum document: [Summarized Answer]**

**Figure 4.** Detailed templates for each prompting strategy.

### 3.4.3 Performance Heterogeneity across Data Features

We ran a series of regression analyses to examine how LLM performance varies across curriculum document type, model version, and competency framework, and to identify data characteristics associated with greater difficulty in competency analysis. The dependent variable  $Y_i$  is specified in two forms: (1) *prediction accuracy*, a binary indicator denoting whether the LLM prediction exactly matches the human annotation, and (2) *score difference*, a continuous measure capturing the deviation between LLM-predicted and human-annotated competency scores. The regression model is specified as

$$Y_i = \alpha + \beta \cdot \mathbf{DataType}_i + \gamma \text{WordCount}_i + \delta \cdot \mathbf{Model}_i + \sigma \cdot \mathbf{Framework}_i + \eta \cdot \mathbf{SubjectMatter}_i + \epsilon_i, \quad (1)$$

where  $\mathbf{DataType}_i$ ,  $\mathbf{Model}_i$ , and  $\mathbf{Framework}_i$  are categorical indicators. Learning objectives (data type), GPT-4o (LLM model), and the EU Key Competences framework serve as reference categories, as they exhibit the highest overall performance.  $\text{WordCount}_i$  controls for variation in text length. Subject matter is included as a set of categorical controls—STEM, Humanities and Social Sciences, Applied Disciplines, and Other/Interdisciplinary—to account for disciplinary context. We estimated separate regressions for each classification granularity (from two-class to five-class) and for each dependent variable (accuracy and score difference).

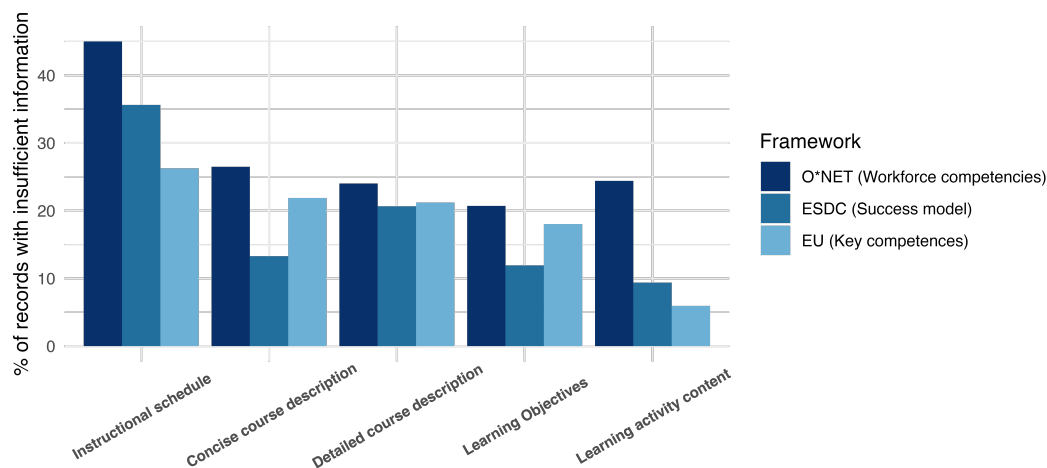
## 4. Results

### 4.1 RQ1: What Types of Curriculum Documents Are More Informative for Conducting 21st-Century Competency Analytics?

During human annotation, annotators could assign an NA label when curriculum documents lacked sufficient information to determine whether a competency was covered. We use the proportion of NA annotations as an indicator of each document type’s usability for 21st-century competency analytics. Figure 5 shows the percentage of documents labelled as NA based on finalized annotations after inter-rater reliability was established.

Across all three competency frameworks, instructional schedules exhibit the highest rates of insufficient information for competency inference—up to 45% under ONET (Workforce Competencies), 36% under ESDC (Success Model), and 27% under the EU (Key Competences). In contrast, learning objectives and learning activity content provide clearer pedagogical cues and show the lowest NA rates. Usability also varies by framework granularity: finer-grained frameworks, such as ONET’s distinction between “written expression” and “written comprehension,” are associated with higher NA rates, while broader competency definitions (e.g., ESDC’s general “writing” category) are more readily inferred from limited information.

Overall, detailed descriptions of specific learning activities provide the richest information for evaluating 21st-century competencies. Learning objectives and course descriptions can also provide helpful cues, whereas instructional schedules generally lack the depth needed to assess 21st-century competencies. However, we note that these findings are specific to our sample and may not reflect general patterns. Information sufficiency ultimately depends on both the data type and the targeted competency, and it should be determined in the context of the specific datasets.



**Figure 5.** Percentage of curriculum documents with insufficient information (NA: competency may be relevant, but available content is insufficient to determine its presence) across competency frameworks and curriculum document types.

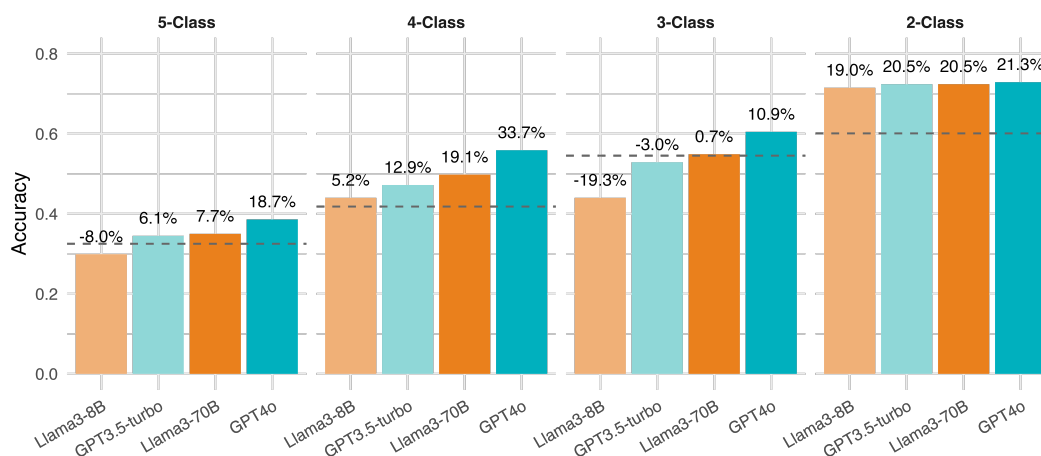
## 4.2 RQ2: What Is the Performance of Zero-Shot LLMs as a Baseline for 21st-Century Competency Analytics Tasks, and What Are the Challenges?

After assessing the usability of the five curriculum document types for 21st-century competency analytics, we evaluated LLMs’ baseline performance in a zero-shot setting, reflecting their default reasoning ability without additional scaffolding or contextual support.

### 4.2.1 Performance Variation across Analytical Granularity

We first examined how LLM baseline performance varies with the granularity of competency classification. As shown in Figure 6, more advanced models such as GPT-4o and Llama3-70B consistently achieve higher prediction accuracy across all levels of classification granularity, particularly in more complex, finer-grained tasks. In the three- to five-class settings, smaller models such as Llama3-8B often perform at or near the random-guessing baseline, whereas GPT-4o exceeds the baseline by 10.9%–33.7%. This gap indicates that greater model scale and reasoning capacity are critical for handling nuanced competency distinctions in zero-shot settings.

Despite these differences, LLMs overall exhibit limited effectiveness in fine-grained competency analysis compared to humans. In the five-class setting, all models perform only marginally above random, suggesting a shared difficulty in distinguishing among different levels of competency coverage. Performance improves substantially as the task becomes easier. When reduced to a binary classification task, as shown in Table 5, all models achieve accuracies above 70% (GPT-3.5-turbo: 72.4%; GPT-4o: 72.9%; Llama3-70B: 71.5%; Llama3-8B: 72.4%). Full results for additional evaluation metrics are reported in Table 18 in the Appendix.



**Figure 6.** Zero-shot accuracy across levels of classification granularity and comparison with random guessing.

**Note:** Accuracy here represents averages across data types within each competency framework. Instructional schedules are excluded due to insufficient information in many records. The grey dashed line marks the majority class guessing (MCG) baseline, which accounts for class imbalance, defined as Expected accuracy =  $\max_i P(i)$ . MCG was computed for each data type within each framework, and the final score reflects their average. Numbers on the bars indicate the deviation from the corresponding MCG accuracy.

### 4.2.2 Performance Heterogeneity across Data Features

We next examined which data characteristics pose greater challenges for LLMs in producing accurate and reliable predictions. Table 6 reports the regression results. Across all analytic settings, both curriculum document type and competency framework are significantly associated with variation in LLM prediction accuracy.

Among curriculum document types, learning objectives (the reference category) consistently show the highest accuracy after controlling for task context. In contrast, instructional schedules are associated with the largest performance decline, with accuracy reductions ranging from 0.114 in the five-class setting to 0.216 in the binary classification task. This is followed by detailed course descriptions, concise course descriptions, and learning activity content from the learning management system. Importantly, this ordering of performance degradation is stable across all levels of classification granularity.

Framework characteristics also play a substantial role. The EU Key Competences framework, representing the most general level of competency definition, is associated with significantly higher prediction accuracy than the two more fine-grained frameworks across all classification settings. Text length, as measured by word count, is not significantly associated with prediction accuracy. Finally, model capacity matters primarily for fine-grained tasks: more advanced models (e.g., GPT-4o and

Llama3-70B) significantly outperform smaller models in multi-class settings, whereas in the binary classification task, accuracy differences across models are no longer statistically significant.

**Table 6.** Regression coefficients for LLM alignment with human annotations across classification granularity.

|           |                             | Prediction accuracy |           |           |           | Score difference |           |           |           |
|-----------|-----------------------------|---------------------|-----------|-----------|-----------|------------------|-----------|-----------|-----------|
|           |                             | 5 Class             | 4 Class   | 3 Class   | 2 Class   | 5 Class          | 4 Class   | 3 Class   | 2 Class   |
| Data type | Concise course description  | -0.043***           | -0.034*** | -0.029*** | -0.069*** | 0.225***         | 0.171***  | 0.089***  | 0.084***  |
|           | Detailed course description | -0.054***           | -0.046*** | -0.059*** | -0.088*** | 0.213***         | 0.153***  | 0.107***  | 0.126***  |
|           | Learning activity content   | -0.038**            | -0.042**  | -0.044*** | -0.074*** | -0.213***        | -0.127*** | -0.094*** | -0.089*** |
|           | Instructional schedule      | -0.114***           | -0.045*** | -0.120*** | -0.216*** | 0.645***         | 0.469***  | 0.281***  | 0.257***  |
| Length    | Word count                  | 0.008               | 0.005     | 0.001     | 0.023***  | 0.030**          | 0.031***  | 0.015**   | 0.008     |
|           | GPT 3.5-turbo               | -0.039***           | -0.014*   | -0.051*** | 0.010     | 0.052**          | 0.070***  | 0.028**   | -0.091*** |
| Model     | Llama3-70B                  | -0.044***           | -0.033*** | -0.054*** | 0.000     | 0.132***         | 0.128***  | 0.082***  | -0.029*** |
|           | Llama3-8B                   | -0.020**            | 0.038***  | 0.000     | 0.001     | -0.636***        | -0.267*** | -0.222*** | -0.225*** |
|           | O*NET                       | -0.093***           | -0.114*** | 0.012*    | 0.011*    | 0.100***         | 0.001     | 0.053***  | -0.003    |
| Framework | ESDC                        | -0.115***           | -0.163*** | -0.102*** | -0.017**  | 0.234***         | 0.214***  | 0.129***  | -0.034*** |
|           | Adjusted R <sup>2</sup>     | 0.0158              | 0.018     | 0.018     | 0.027     | 0.092            | 0.058     | 0.061     | 0.069     |

**Note:** Each column presents results from a separate regression conducted at a specific classification granularity. All regressions are run at the course competency level. The dependent variables include (1) a binary accuracy indicator, which captures whether the LLM exactly predicts the human-annotated score (1 if the LLM prediction exactly matches the human annotation, 0 otherwise), and (2) a continuous score difference variable (LLM-predicted score – human-annotated score), which measures the degree of deviation from human judgment, revealing whether it overestimates or underestimates the competency relevance. For each group tested, we use the highest performing condition as learning objectives (data type), GPT-4o (model), and EU (Key competences) (framework) as the reference group to assess performance degradation across other conditions. Statistical significance is denoted by asterisks:  $p < 0.10(\cdot)$ ,  $p < 0.05(*)$ ,  $p < 0.01(**)$ ,  $p < 0.001(***)$ .

For score differences, the regression results indicate that, after controlling for task context, LLMs generally overestimate the extent to which a course covers a given competency. This upward bias is reflected in positive and statistically significant intercepts across all classification granularities (0.128 in the five-class task, 0.019 in the four-class task, 0.051 in the three-class task, and 0.147 in the binary task). Curriculum document characteristics further shape this bias. Data type is a significant predictor of overestimation, with instructional schedules producing the largest positive score differences across all analytic settings, indicating the strongest tendency toward overprediction. Course descriptions also exhibit significant but more moderate overestimation effects. Framework characteristics contribute as well: relative to the most general framework (EU Key Competences), both O\*NET (Workforce Competencies) and ESDC (Success Model) are associated with significantly larger score differences. Finally, word count is positively associated with score overestimation, suggesting that longer or more content rich curriculum documents lead LLMs to infer higher levels of competency coverage.

### 4.2.3 Error analysis

We manually analyzed cases where LLM predictions diverged from human annotations and identified several recurring error patterns. Figure 7 provides a representative example illustrating these errors.

- Over-interpretation:** This error pattern is observed across all models and all curriculum document types we evaluated. A possible explanation is that LLMs tend to overinterpret available content, while human raters are more rigorous in seeking explicit pedagogical evidence. For example, we find that when a course document mentions an exam, human annotators are cautious about inferring writing-related competencies because exam formats can vary widely. In contrast, LLMs tend to default to the assumption that exams are writing based, leading them to confidently infer writing competencies, which reflects biased reasoning.
- Failure to detect relevant information:** This error pattern is common in lengthy curriculum document types such as learning activity content, where dense information makes it difficult for LLMs to capture and interpret key details accurately. For example, when information about skill development is embedded within subtle descriptions of teaching activities, human annotators can often detect and reason from these cues, whereas LLMs may overlook such nuanced evidence and incorrectly assign *NA*.
- Hallucination:** This error pattern is more common in concise course descriptions and instructional schedule data, which contain limited information compared to other data types. In some cases, LLMs make confident predictions even when the text offers little to no evidence to support such inferences.
- Failure to generate a response:** This error pattern is relatively rare and occurs predominantly in smaller models (e.g., Llama3-8B). In these cases, the model fails to return any prediction. It is most frequently observed in data content types used for learning activities, likely due to their unstructured format or because the input length exceeds the model’s context window.

**Error pattern 1: Over-interpretation**



**Writing expression:** Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.

Human annotation: 1 (possibly implied)



LLM prediction (zero-shot from GPT 3.5-turbo): 2 (reasonably inferred)



**Course document: syllabi excerpt**

WEEKS TOPICS 1-4

Course Orientation  
Module 1--Computers and the Internet  
Chapter 1--Introduction to Computers and the Internet.  
Chapter 5--Introducing the Internet and Email  
Chapter 6--Networks and Internet Security and Privacy  
ASSIGNMENTS/ACTIVITIES/QUIZZES  
MODULE EVALUATION

WEEKS TOPICS 5-6

Module 2--Operating System Software--Windows 7  
Chapter 7--Exploring Microsoft Windows 7  
Chapter 8--Managing Your Files  
ASSIGNMENTS/ACTIVITIES/QUIZZES  
MODULE EVALUATION  
...

WEEKS TOPICS 15

Module 7--Application Software  
Chapter 17-->Creating a Database  
Chapter 18-->Maintaining and Querying a Database  
ASSIGNMENTS/ACTIVITIES/QUIZZES  
MODULE EVALUATION  
FINAL EXAM/PROJECT

**Error pattern 2: Failing to detect relevant information**



**Writing expression:** Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.

Human annotation: 2 (reasonably inferred)



LLM Prediction (zero-shot from Llama3-8B): NA (insufficient information)



**Course document: Learning activities excerpt**

.....  
**SCAFFOLDING ASSIGNMENT 1:** Organizing, Analyzing, and Outlining for Research Writings. Please use this time to draft an outline of your research essay draft. You can use this Google doc for reference. Please MAKE A COPY first. Go to FILE in the upper left-hand corner. Click on MAKE A COPY, Save to your drive. You can also use Figma or Miro if you are a more visual learner.  
**SCAFFOLDING ASSIGNMENT 2:** Tone and Hedging for Research Writing: Pick a section of your essay assignment and start to address the language, specifically keeping tone and hedging in mind. Anywhere you see absolutist language used, edit out for balance. Submit a 300-word section of your essays with these edits.  
**SCAFFOLDING ASSIGNMENT 3:** Using Popular Sources: Please pick a couple of quotes or headlines from popular sources and tell me what you find fascinating about them. What can we learn from people's direct quotes or the way the media reports a story? 100 words in total. Please write at least THREE sentences that clearly state your hypothesis and justifications, based on the data you have collected thus far.  
.....

**Error pattern 3: Hallucination**



**Persistence:** Sustaining effort and motivation in the face of obstacles, setbacks, or difficulties, continuing to pursue goals despite challenges or repeated failures.

Human annotation: NA (insufficient information)



LLM Prediction (Zero-shot from Llama3-70b): 2 (reasonably inferred)



**Course document: course description excerpt**

This course is a two-semester course sequence that introduces principles and concepts of general organic and biological chemistry; the laboratory will provide experimental applications of these chemical topics. Che 121122 two terms are required; they are liberal arts electives and are recommended for students intending to transfer to bachelor degree allied health science curricula...

**Error pattern 4: Non-response**



**Originality:** Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.

Human annotation: 1 (Possibly implied)



LLM Prediction (Zero-Shot from GPT 3.5-turbo): BLANK (no prediction)



**Course document: learning activities excerpt**

WEEKS TOPICS 1-4

**Self Assessment:** Don't worry: This isn't graded. But completion is mandatory. (You'll get full credit for completion.) Instructions: Take your time, answer the prompt as you would any other short assignment. This is your first writing assignment for this class, and it will help me determine where you are with your writing. You may proofread and revise as you would any other short essay, but complete it in time to upload by class time on Wednesday. Prompt: Self-assessment Write a short narrative, a sketch that moves clearly from one detailed point to the next—to document and describe important moments in your development as a reader, a writer, a critical thinker, and a communicator. This "literacy narrative" should use between 2 and 5 examples from your personal history—both in school and out of school—that have shaped your perspective on reading and writing, and defined your academic interests, your ways of thinking about and strategizing your writing assignments in school, and your sense of style when you write and communicate in both school-related and non-school related contexts. As you recount key examples from your past, assesses your strengths and weaknesses as a reader, a writer, and a communicator and present two or three new things that you would like to accomplish—that is, two or three things that you would like to see become part of your personal "literacy narrative" as a college-level writer. Then, as a final thought, state what you would like to get out of this class, and why. That's all! If writing your Self-Assessment inspires any questions, feel free to sign up for a conference Week 1 to come and discuss them with me. (Conferences are the way you will receive much of your feedback on your writing, so make a habit of using them!) You can find the conferences sign-up link on the main course page, at the top of the syllabus.

**Harriot Imitation: Draft 3 (+ Reflection) Assignment & Guide:** Harriot Imitation: Draft 3 (+ Reflection) Assignment & Guide First Part: The Draft It is totally up to you how much you will change your imitation for your third draft. However, you will have the most success if these changes are based on clear goals you've set for yourself. If you have already identified specific things you would like to focus on improving in this draft, great! Focus on those. But if you are unsure, that might be a sign you need to come meet me for a conference. Sign up for a conference using the link below: [Link] Second Part: The Reflection Writing is never really done. Which is why you'll still have a chance to revise this assignment and submit a better version by the end of the quarter. But before you do, it's important to decide what you'd like to accomplish with a new draft. So, this week, it's time to return to your third draft of the Harriot Imitation with a critical eye. Directions: Compare the third draft of the Harriot Imitation to an older draft and answer the following question, supported by examples from your drafts: Based on the changes between these two drafts, WHAT do you think your biggest accomplishment was on this assignment? (A major writing challenge that you overcame.) To support this, copy and paste two different versions of one section of your Harriot Imitation—matching samples from each draft. Explain the changes you made between those two drafts. Then, explain HOW you figured out how to improve that section by applying a rhetorical concept you learned so far this quarter. Then, explain WHY you think that section is better now. Why will your later draft do a better job of persuading an audience to change their mind about something? Based on the changes between those two drafts, WHAT do you think your biggest writing struggle has been? (A major writing challenge that you haven't overcome yet, but you would still like to fix.) To support this, copy and paste an example of something you're still not satisfied with from your newest draft. Explain WHY you think that section isn't persuasive yet. Why don't you think it would help convince an audience to change their mind? Give yourself one piece of advice: If you were starting this assignment all over again, what advice would you give yourself? How would you avoid the challenges you had on your first attempt and start off stronger, if given a second chance? Remember: Writing is a solitary act. It's important to be able to identify your own mistakes. By reflecting on your work in the above ways, you will develop skills that can help you to become your own best editor, and make the rewriting process more comfortable.  
.....

**Figure 7.** Error patterns in baseline (zero-shot) LLM performance for 21st-century competency analysis.

### 4.3 RQ3: Do Reasoning-Based Prompting Strategies Address These Challenges and Improve the Reliability of LLM-Based 21st-Century Competency Analytics?

In this section, we evaluate whether our proposed reasoning-based prompting strategies improve LLMs’ pedagogically grounded reasoning for higher-order competency analysis by replicating prior experiments.

#### 4.3.1 Performance Variation across Analytical Granularity

Table 7 presents the overall prediction accuracy of five proposed prompting strategies, compared to the baseline zero-shot setting, across all levels of classification granularity. Overall, we find that curricular CoT improves prediction accuracy over the zero-shot baseline across all levels of task granularity. However, these improvements are more apparent in larger and more advanced models, such as GPT-4o and Llama3-70B. In contrast, the definition-based (DEF) strategy, which provides the model with human-annotated notes for each competency, does not improve performance and can even reduce accuracy.

**Table 7.** Accuracy of curricular CoT (DEF, CQA, CQ, QA, A) compared with the baseline (zero-shot).

| Prompting       | GPT3.5-turbo |         |         |         | GPT4o        |              |              |              |
|-----------------|--------------|---------|---------|---------|--------------|--------------|--------------|--------------|
|                 | 5 Class      | 4 Class | 3 Class | 2 Class | 5 Class      | 4 Class      | 3 Class      | 2 Class      |
| ZERO (Baseline) | 0.337        | 0.457   | 0.536   | 0.712   | 0.334        | 0.530        | 0.586        | 0.704        |
| DEF             | 0.337        | 0.452   | 0.512   | 0.671   | <b>0.337</b> | 0.524        | 0.571        | 0.690        |
| CQA             | 0.322        | 0.444   | 0.535   | 0.706   | <b>0.338</b> | <b>0.536</b> | <b>0.599</b> | <b>0.713</b> |
| CQ              | 0.321        | 0.438   | 0.517   | 0.706   | 0.334        | <b>0.531</b> | <b>0.584</b> | 0.704        |
| QA              | 0.328        | 0.446   | 0.525   | 0.707   | <b>0.341</b> | <b>0.535</b> | <b>0.594</b> | <b>0.705</b> |
| A               | 0.332        | 0.454   | 0.533   | 0.712   | <b>0.347</b> | <b>0.538</b> | <b>0.599</b> | <b>0.710</b> |

| Prompting       | Llama3-8B    |              |              |              | Llama3-70B   |              |              |              |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                 | 5 Class      | 4 Class      | 3 Class      | 2 Class      | 5 Class      | 4 Class      | 3 Class      | 2 Class      |
| ZERO (Baseline) | 0.315        | 0.442        | 0.442        | 0.710        | 0.334        | 0.473        | 0.535        | 0.701        |
| DEF             | 0.305        | <b>0.455</b> | <b>0.455</b> | 0.700        | <b>0.337</b> | 0.473        | 0.533        | 0.701        |
| CQA             | <b>0.323</b> | <b>0.452</b> | <b>0.452</b> | <b>0.715</b> | <b>0.338</b> | <b>0.484</b> | <b>0.551</b> | <b>0.710</b> |
| CQ              | <b>0.316</b> | <b>0.448</b> | <b>0.448</b> | <b>0.717</b> | 0.334        | <b>0.479</b> | <b>0.541</b> | <b>0.705</b> |
| QA              | 0.307        | 0.437        | 0.437        | 0.706        | 0.341        | <b>0.487</b> | <b>0.550</b> | <b>0.705</b> |
| A               | 0.304        | 0.424        | 0.424        | 0.696        | <b>0.347</b> | <b>0.493</b> | <b>0.557</b> | <b>0.707</b> |

**Note:** Bold numbers indicate performance improvements over each model’s zero-shot baseline. Accuracy is averaged across curriculum-competency alignment tasks within each classification granularity level.

#### 4.3.2 Performance Heterogeneity across Data Features

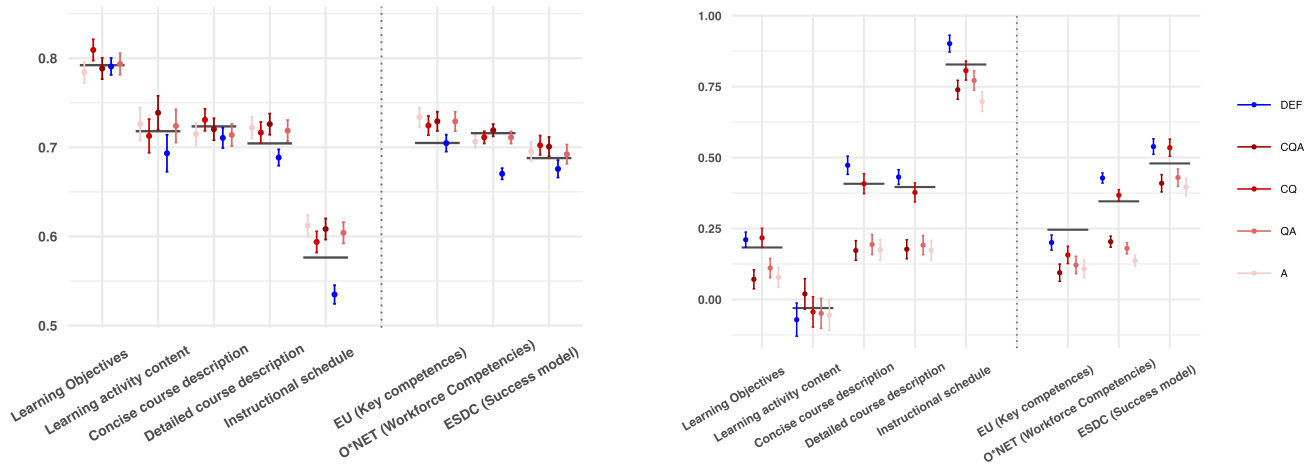
Building on the accuracy gains over the zero-shot baseline, we next examine whether our proposed prompting strategies mitigate the data-specific challenges identified in RQ2. Figure 8a compares the zero-shot baseline with the proposed strategies in binary classification tasks, which we focus on because they are most relevant to practical applications and because prior results show limited zero-shot capability for fine-grained competency analysis.

Overall, curricular CoT yields modest but consistent improvements in prediction accuracy, particularly by reducing performance gaps across curriculum document types. Notably, for the two document types that performed worst under the zero-shot condition—instructional schedules and detailed course descriptions—each curricular CoT variant improves accuracy. Even for more information rich document types, such as learning objectives and learning activity content, several curricular CoT variants outperform the zero-shot baseline, indicating that structured reasoning can further enhance performance when sufficient instructional detail is available.

In terms of competency frameworks, curricular CoT methods improve performance for more abstract frameworks, such as the EU (Key Competences) and ESDC (Success Model), but yield limited gains for O\*NET (Workforce Competencies). In contrast, the definition-based prompting strategy (DEF) does not reduce performance gaps across frameworks or document types and, in some cases, leads to declines in accuracy.

We next examine a key challenge in the zero-shot setting: LLMs’ tendency to overestimate the coverage of competencies relative to human annotations. Figure 8b presents the average differences between LLM-predicted and human rating scores. Because score differences are not meaningful in binary classification, we conduct this analysis in the five-class classification scenario. Our results indicate that CoT-based strategies can mitigate this challenge. These methods consistently reduce the systematic overestimation observed under the zero-shot condition, particularly for inputs with longer word counts and across a

range of curriculum document types and competency frameworks. In contrast, the definition-based strategy (DEF) again fails to narrow these discrepancies and, in some cases, further increases the misalignment between LLM-generated scores and human judgments.



(a) Prediction accuracy (binary classification). (b) Score differences (5-class classification).

**Figure 8.** Estimated performance of curricular CoT : (a) accuracy and (b) score differences.

**Note:** Solid grey lines indicate the estimated accuracy or score difference of the zero-shot baseline for each category. Coloured points and error bars represent the estimated means and 95% confidence intervals of other prompting strategies.

### 4.3.3 Error Analysis

We conducted a manual error analysis to examine the error patterns identified in the previous section, focusing on how the proposed curricular CoT approach mitigates these baseline errors and where it still fails. Figures 9 and 10 illustrate how the curricular CoT approach mitigates these errors through intermediate reasoning steps and the structured processing of curriculum documents.

By examining cases in which the curricular CoT approach mitigates the limitations of direct analysis of curriculum documents, we provide empirical support for our design assumptions. First, consistent with error patterns 1 and 3, when curriculum texts are highly abstract or underspecified, reasoning based on guiding questions enables the model to more accurately extract curriculum elements relevant to competency inference than direct decision making. Particularly when relevant information is missing, the model provides a more honest summary of what is available in the curriculum document and explicitly reports the lack of information instead of making overconfident assumptions, as seen in the direct inference of higher-order skills. Moreover, in error patterns 2 and 4 involving lengthy curriculum documents, the curricular CoT approach facilitates the identification of subtle yet meaningful instructional cues. By structuring the extraction process around targeted guiding questions, the model is better able to surface relevant information embedded in dense descriptions, thereby reducing the likelihood of missing evidence in long texts.

We also analyze failure cases in which the proposed methods yield limited or negative gains. A primary source of failure is variability in the quality of intermediate summaries generated by different LLMs. Because curricular CoT relies on the accurate extraction and structuring of curriculum content prior to competency inference, errors in summarization can degrade downstream performance. Comparing intermediate summaries across models reveals clear differences. GPT-4o produces the most accurate and well structured summaries, consistently capturing key instructional components. Llama3-8B generates concise but less detailed summaries. In contrast, GPT-3.5-turbo frequently hallucinates unsupported instructional elements, while Llama3-70B, although generally accurate, often introduces redundancy that adds noise to downstream analysis.

## 5. Discussion

In this study, we investigate the affordances of LLMs for analyzing the integration of 21st-century competencies in postsecondary education curricula using digital curriculum documents commonly available at postsecondary institutions. We evaluate the usability of these data sources, benchmark zero-shot LLM performance, and identify systematic reasoning errors. To address the lack of large-scale, high-quality curriculum benchmarks, we propose a reasoning-based prompting strategy that improves LLMs’ pedagogically grounded reasoning over complex and unstructured curriculum texts.



### Mitigating error pattern 1: Overinterpretation



**Collaboration:** The ability to contribute and support others to achieve a common goal.

**Human annotation:**  
0 (clearly unrelated)



**LLM prediction (Zero-Shot from GPT 4o):**  
1 (possibly implied)



**LLM prediction (CoT(CQA) from GPT 4o):**  
0 (clearly unrelated)



#### Course document: description excerpt

This is a survey course that introduces essential topics in atomic and molecular structure chemical and physical properties and chemical reactivity these concepts will form a basis for understanding biological applications and environmental issues.

#### Curricular CoT summary

Here are the rewritten course descriptions in a structured format:

\*\*1. Main subject & focus:\*\*

The primary subject area and focus of this course is Chemistry, specifically exploring essential topics in atomic and molecular structure, chemical and physical properties, and chemical reactivity.

\*\*2. Core concepts & practical skills:\*\*

This course teaches domain knowledge and practical skills in the following areas: atomic and molecular structure, chemical and physical properties, and chemical reactivity. These concepts will form a basis for understanding biological applications and environmental issues.

\*\*3. Key tasks & activities:\*\*

NA (no specific tasks or learning activities mentioned)

\*\*4. Delivery format(s):\*\*

NA (no specific delivery format mentioned)

\*\*5. Assessment methods:\*\*

NA (no information available on how student learning will be evaluated or assessed in this course)

### Mitigating error pattern 2: Failing to detect relevant information



**Deductive reasoning:** Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.

**Human annotation:**  
1 (possibly implied)



**LLM prediction (Zero-shot from GPT 4o):**  
NA (insufficient information)



**LLM prediction (CoT(A) from GPT 4o):**  
1 (possibly implied)



#### Course document: syllabi excerpt

**Muckraking Rewrite:** If you would like to rewrite the muckraking comparative analysis, please feel free to do so. There is no obligation to resubmit the essay, but you may find it helpful to respond to the comments and work on developing your writing. It is, after all, a process. I have intentionally waited to re-post this option so that you could focus on developing your topic - and the other work you have to do around midterms. This is the last paper that you will be able to rewrite - you can work on editing all of your other papers as part of writing your final paper for the class.

**Annotated Bibliography:** Write an annotated bibliography of the sources for your final paper. Your paper should be about 2-3 pages long, include 5 - 8 sources, at least two of which should be sources we have not used in class. Read Rampolla's section on annotated bibliography to prepare. Your annotated bibliography should include all sources - primary and secondary - that you have gathered for your final paper. For each source, write a short paragraph that describes the content of the source, the main point, and its relevance to your topic. This should include information about the author and a concise evaluation of the value of the source to your work. The listing should be alphabetical by author's last name (see Rampolla, beginning on p 121 for more information). Remember that this is not a book or article review; it is a concise commentary that you will use in your final paper to do a literature review.

**Comparative Analysis:** Secondary Sources: Reading secondary sources helps us to analyze how other historians have interpreted the primary sources for a period being studied. Historiography is the study of historical writing; historiographical analysis is the comparison of more than one secondary source to evaluate the historians contribution to the literature. Your essay for this week should analyze at least two secondary sources. You should have at least two secondary sources selected for your final paper; each of these sources should relate to the topic you have selected and speak to your research question. Your first step is to review Rampolla, pages 9 - 12 and 18 - 23 on secondary sources. Your second step is to select two (or more) secondary sources that you will analyze. Your review should be up to three pages, include a thesis statement that addresses your comparative analysis of the two sources, and address the following questions: what is each author's primary argument? How does each work relate to your research question? How are the works in dialogue with one another? How would you evaluate each work?

#### Curricular CoT summary

The "Muckraking Rewrite" is an optional task that allows students to rewrite their muckraking comparative analysis essay. The purpose of this task is to provide students with an opportunity to improve their writing by responding to feedback and developing their writing skills further. The activity format is a written assignment related to the topic of muckraking and historical analysis. Students are expected to revise their original essay based on comments received, focusing on enhancing their argumentation and writing clarity. The delivery format is an online submission, and the assessment method is NA, as it is not a required resubmission.

The "Annotated Bibliography" task requires students to compile an annotated bibliography for their final paper. The purpose is to help students organize and evaluate their sources, which will be used in their final research paper. The activity format is a written assignment related to research methodology and source evaluation. Students must write a 2-3 page bibliography that includes 5-8 sources, with at least two new sources not used in class. Each source should be described in a short paragraph, including its content, main point, relevance, author information, and evaluation. The delivery format is an online submission, and the assessment method is graded by rubric.

The "Comparative Analysis - Secondary Sources" task involves analyzing secondary sources to understand historiographical perspectives. The purpose is to evaluate how historians interpret primary sources and contribute to historical literature. The activity format is a written assignment related to historiography and source analysis. Students must select at least two secondary sources related to their research question, write a 3-page review with a thesis statement, and address questions about the authors' arguments and the sources' dialogue. The delivery format is an online submission, and the assessment method is graded by rubric.

Figure 9. Part 1. Illustrative example: curricular CoT mitigating off-the-shelf LLM errors.

### Mitigating error pattern 3: Hallucination

**Cooperation:** Being pleasant with others on the job and displaying a good-natured, cooperative attitude.

**Human annotation:**  
NA (insufficient information)

**LLM prediction (Zero-shot from Llama8-b):**  
2 (reasonably inferred)

**LLM prediction (CoT(A) from Llama8-b):**  
NA (insufficient information)

#### Course document: description excerpt

This course is a two-semester course sequence that introduces principles and concepts of general, organic, and biological chemistry. The laboratory will provide experimental applications of these chemical topics. CHE 121-122: Two terms are required. They are liberal arts electives and are recommended for students intending to transfer to bachelor's degree allied health science curricula. CHE 121-122 cannot be granted credit to fulfill degree requirements for science and engineering science.

#### Curricular CoT summary

Here are the rewritten course descriptions in a structured format:

**\*\*1. Main subject & focus:\*\***  
The primary subject area and focus of this course is the fundamentals of general, organic, and biological chemistry, introducing principles and concepts in a two-semester sequence.

**\*\*2. Core concepts & practical skills:\*\***  
This course teaches domain knowledge and practical skills in general, organic, and biological chemistry, with laboratory applications providing experimental applications of these chemical topics. Students will gain a solid understanding of the subject matter.

**\*\*3. Key tasks & activities:\*\***  
The course involves laboratory work, where students will apply the chemical topics learned in the course. The laboratory will provide experimental applications of these chemical topics.

**\*\*4. Delivery format(s):\*\***  
The course is a two-semester sequence, with no specific delivery format mentioned.

**\*\*5. Assessment methods:\*\***  
NA (No information available on assessment methods)

### Mitigating error pattern 4: Non-response

**Originality:** Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.

**Human annotation:**  
1 (possibly implied)

**LLM prediction (Zero-shot from GPT 3.5-turbo):**  
No prediction

**LLM prediction (CoT(A) from GPT 3.5-turbo):**  
1 (possibly implied)

#### Course document: syllabi excerpt

WEEKS TOPICS 1-4

**Self Assessment:** Don't worry: This isn't graded. But completion is mandatory. (You'll get full credit for completion.) Instructions: Take your time, answer the prompt as you would any other short assignment. This is your first writing assignment for this class, and it will help me determine where you are with your writing. You may proofread and revise as you would any other short essay, but complete it in time to upload by class time on Wednesday. Prompt: Self-assessment Write a short narrative, a sketch that moves clearly from one detailed point to the next—to document and describe important moments in your development as a reader, a writer, a critical thinker, and a communicator. This "literacy narrative" should use between 2 and 5 examples from your personal history—both in school and out of school—that have shaped your perspective on reading and writing, and defined your academic interests, your ways of thinking about and strategizing your writing assignments in school, and your sense of style when you write and communicate in both school-related and non-school related contexts. As you recount key examples from your past, assesses your strengths and weaknesses as a reader, a writer, and a communicator and present two or three new things that you would like to accomplish—that is, two or three things that you would like to see become part of your personal "literacy narrative" as a college-level writer. Then, as a final thought, state what you would like to get out of this class, and why. That's all! If writing your Self-Assessment inspires any questions, feel free to sign up for a conference Week 1 to come and discuss them with me. (Conferences are the way you will receive much of your feedback on your writing, so make a habit of using them!) You can find the conferences sign-up link on the main course page, at the top of the syllabus.

**Harriot Imitation: Draft 3 (+ Reflection) Assignment & Guide:** Harriot Imitation: Draft 3 (+ Reflection) Assignment & Guide First Part: The Draft It is totally up to you how much you will change your Imitation for your third draft. However, you will have the most success if these changes are based on clear goals you've set for yourself. If you have already identified specific things you would like to focus on improving in this draft, great! Focus on those. But if you are unsure, that might be a sign you need to come meet me for a conference. Sign up for a conference using the link below: [Link] Second Part: The Reflection Writing is never really done. Which is why you'll still have a chance to revise this assignment and submit a better version by the end of the quarter. But before you do, it's important to decide what you'd like to accomplish with a new draft. So, this week, it's time to return to your third draft of the Harriot Imitation with a critical eye. Directions: Compare the third draft of the Harriot Imitation to an older draft and answer the following question, supported by examples from your drafts: Based on the changes between those two drafts, WHAT do you think your biggest accomplishment was on this assignment? (A major writing challenge that you overcame.) To support this, copy and paste two different versions of one section of your Harriot Imitation—matching samples from each draft. Explain the changes you made between those two drafts. Then, explain HOW you figured out how to improve that section by applying a rhetorical concept you learned so far this quarter. Then, explain WHY you think that section is better now. Why will your later draft do a better job of persuading an audience to change their mind about something? Based on the changes between those two drafts, WHAT do you think your biggest writing struggle has been? (A major writing challenge that you haven't overcome yet, but you would still like to fix.) To support this, copy and paste an example of something you're still not satisfied with from your newest draft. Explain WHY you think that section isn't persuasive yet. Why don't you think it would help convince an audience to change their mind? Give yourself one piece of advice: If you were starting this assignment all over again, what advice would you give yourself? How would you avoid the challenges you had on your first attempt and start off stronger, if given a second chance? Remember: Writing is a solitary act. It's important to be able to identify your own mistakes. By reflecting on your work in the above ways, you will develop skills that can help you to become your own best editor, and make the rewriting process more comfortable.

#### Curricular CoT summary

**The Self Assessment assignment** is a written task where students are required to write a short narrative documenting and describing important moments in their development as a reader, writer, critical thinker, and communicator. The activity format is a written assignment, and the topic or concept is related to the student's personal literacy narrative. The task instructions are to recount key examples from their past, assess their strengths and weaknesses, and present two or three new things they would like to accomplish as a college-level writer. The delivery format is online submission, and the assessment method is completion-based, with full credit given for completion.

**The Harriot Imitation: Draft 3 (+ Reflection) assignment** is a written task where students are required to submit their third draft of the Harriot Imitation, along with a reflection on their progress and improvements. The activity format is a written assignment, and the topic or concept is related to imitating Harriot's writing style and structure. The task instructions are to compare the third draft to an older draft, answer specific questions about the changes made, and provide examples from the drafts to support the answers. The delivery format is online submission, and the assessment method is graded by the instructor, with the option to receive feedback through one-on-one conferences.

**Figure 10.** Part 2. Illustrative example: curricular CoT mitigating off-the-shelf LLM errors.

Regarding data feasibility, our findings suggest that digital curriculum documents in postsecondary institutions constitute a valuable yet underutilized resource for examining how 21st-century competencies are embedded within courses and programs. A key concern in prior literature is whether such data contain sufficient pedagogical detail to support meaningful analysis. From a theoretical perspective, the development of 21st-century competencies typically relies on student-centred instructional designs, such as project-based and active learning (Meyers & Nulty, 2009; Musa et al., 2012), which require curriculum documentation with adequate informational richness. Our human examination confirms that learning activity descriptions are the most informative and reliable data source for competency assessment and should be prioritized when available. Learning management systems offer a promising source for capturing such detailed records. In contrast, instructional schedules in syllabi, while also indicating activity types, often lack content-level specificity and are less feasible for analyzing 21st-century competencies. Course descriptions and learning goals, typically drawn from syllabi or institutional websites, showed moderate utility but varied widely in quality across institutions. Importantly, these findings are not intended to prescribe a fixed hierarchy among curriculum document types. Rather, we argue that the presence of explicit pedagogical elements is more critical than the source of the curriculum document in determining suitability for competency analytics. In practice, institutions may benefit from combining multiple data types to achieve sufficient pedagogical coverage, tailored to the characteristics and informational richness of their local data ecosystems. Over the long term, richer and more structured curriculum data infrastructures will be essential for enabling systematic curricular analytics to support monitoring, evaluation, and continuous improvement.

In terms of the affordances of current general-purpose LLMs for higher-order competency analytics, our results reveal both promise and clear limitations. Even the most advanced model in our study (GPT-4o) does not yet achieve the analytical precision of human experts. While human annotators can reliably distinguish subtle differences in the extent to which courses integrate specific competencies, LLMs struggle with such fine-grained distinctions, highlighting the challenge of applying general-purpose models to pedagogical judgments that require contextual interpretation. Nonetheless, LLMs perform reasonably well in coarse-grained tasks, such as identifying whether a course likely addresses a given competency. The relatively small performance gap between large proprietary models and open-weight models further suggests the cost-effectiveness and scalability of open-weight models for large-scale curricular analytics. We further identify two main sources of model error. First, when analyzing abstract course data, LLMs tend to over-infer or hallucinate based on some keywords (e.g., assuming any “final exam” implies writing comprehension), likely due to patterns learned from historical training data. Second, when processing long and unstructured course text, models often fail to locate the precise pedagogical evidence needed for accurate inference. This issue may reflect constraints in contextual processing and attention span. Together, these findings provide empirical insight into why LLMs continue to struggle with complex educational reasoning tasks, consistent with recent work (Xu et al., 2025; Siyan et al., 2025). They also motivate our element-centric prompting strategy, which aims to strengthen in-context pedagogical reasoning by explicitly directing models’ attention to instructional elements most relevant to competency integration.

As prompt engineering remains the most practical approach for LLM-supported curricular analytics, we evaluate our proposed prompting strategies across multiple curriculum document types, competency frameworks, and both open-weight and proprietary models. Our results show that the CoT-based approach yields modest performance gains, particularly for more advanced models. Qualitative analysis suggests that these gains arise because the guided extraction of structured instructional elements enables more accurate inference of competencies than direct reasoning from raw curriculum text. This intermediate step mitigates hallucination in abstract texts and improves the identification of relevant pedagogical evidence in lengthy, unstructured content. Consistent with prior CoT research, decomposing complex reasoning into explicit steps enhances performance, though the effectiveness of this approach depends on the quality of the intermediate extraction. When reasoning quality is low, this step can introduce noise and reduce accuracy, making the method most effective for larger, more capable models (Wei, Tay, et al., 2022). Interestingly, incorporating human-defined competency descriptions into prompts does not improve LLM performance. This finding supports our hypothesis that higher-order competency analytics require more than surface-level definition matching and instead depend on pedagogically grounded reasoning. Human experts, when performing this task, frequently draw on implicit knowledge when evaluating whether an activity meaningfully promotes a given skill, rather than relying solely on predefined criteria. Such pedagogical knowledge is difficult to formalize or encode in prompts for general-purpose LLMs, limiting their effectiveness. These results underscore the constraints of current LLMs and highlight the need for models or training strategies specifically designed to support pedagogical reasoning in curricular analytics.

Finally, there are still several limitations in our current study that could be addressed in future work. First, our analysis is conducted on a limited set of annotated curriculum samples. Although we conducted human checks to ensure their representativeness, future research should draw on more diverse and multiple data sources to improve generalization. Second, the LLM prompts in this study were not iteratively refined using a separate training or validation set. Future work could incorporate such a step to further optimize prompt design in the specific institutional context. Third, while our evaluation approach allows us to isolate and better understand the limitations and strengths of each data type, it does not reflect how institutions typically use curriculum documents in practice. In real-world settings, multiple document types are often used

together. Future research should therefore examine how integrating multiple data sources affects LLM performance and whether such integration can improve robustness and reliability.

Beyond addressing these limitations, our findings point to several directions for future research in LLM-assisted curricular analytics. The lack of large-scale, high-quality benchmark datasets remains a major bottleneck for curricular analytics. Unlike domains such as the labour market, where richly annotated corpora have accelerated methodological progress, postsecondary curricular analytics lacks shared benchmarks. Future work should prioritize the development and open sharing of such datasets, potentially through coordinated efforts in structured data collection and automated or semi-automated annotation of existing curriculum documents (Javadian Sabet et al., 2024). Moreover, our results indicate that the effectiveness of reasoning-based prompting depends strongly on model capacity. Hybrid pipelines that combine models of different sizes—for example, using more powerful models for content extraction and lighter models for scoring—may offer a practical trade-off between performance and computational cost.

## 6. Conclusion

In this study, we examine whether and how digital curriculum documents, commonly available at postsecondary institutions, can support the analysis of 21st-century competency integration and evaluate the extent to which LLMs can reliably perform this task. Using three competency frameworks and five types of curriculum documents, we manually annotated 7,600 curriculum-competency alignments to assess data feasibility, benchmark zero-shot LLM performance, and test a reasoning-based prompting strategy.

We first show that existing curriculum documents constitute a valuable but underutilized resource for competency-oriented analytics. Among the document types examined, detailed descriptions of learning activities provide the most informative signals for assessing competency integration. More importantly, effective analysis depends on whether curriculum documents contain explicit pedagogical elements, suggesting that combining multiple document types is often necessary to achieve adequate informational coverage in institutional settings.

We then demonstrate that current general-purpose LLMs have mixed capabilities for higher-order competency analytics. While both proprietary and open-weight models perform well in coarse-grained classifications, achieving over 70% agreement with human annotations, they fall short of human-level performance in fine-grained assessments of competency coverage. We identify two primary sources of error: systematic over-interpretation of instructional keywords and difficulty identifying relevant pedagogical evidence in long, unstructured texts.

To address these limitations, we introduce a reasoning-based prompting strategy that guides models to extract and reason over structured instructional elements before making judgments. This approach yields modest but consistent performance improvements across data types and competency frameworks, consistent with the CoT principle that decomposing complex reasoning tasks enhances model performance. However, its effectiveness depends on model capacity, as errors in intermediate extraction can introduce noise and reduce analytical accuracy.

Taken together, our work extends prior evaluations of LLMs for curricular analytics by systematically examining data feasibility, model performance, and technical interventions across multiple data types and frameworks. Our findings indicate that realizing the full potential of curricular analytics requires not only technical advances but also higher-quality curriculum data infrastructures and shared benchmark datasets. While general-purpose LLMs remain limited in their ability to perform complex pedagogical reasoning, their performance can be improved through structured reasoning designs and careful data use. Future gains are likely to come from richer and more standardized curriculum data, shared benchmark datasets, and hybrid human-AI or domain-specific modelling approaches that better align with the pedagogical nature of curricular analytics.

## Acknowledgements

We thank Joe Karaganis for providing access to Open Syllabus data.

## Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

## References

Arafeh, S. (2016). Curriculum mapping in higher education: A case study and proposed content scope and sequence mapping tool. *Journal of Further and Higher Education*, 40(5), 585–611. <https://doi.org/10.1080/0309877X.2014.1000278>

- Buckingham Shum, S., & Crick, R. D. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics*, 3(2), 6–21. <https://doi.org/10.18608/jla.2016.32.2>
- Chou, C.-Y., Tseng, S.-F., Chih, W.-C., Chen, Z.-H., Chao, P.-Y., Lai, K. R., Chan, C.-L., Yu, L.-C., & Lin, Y.-L. (2015). Open student models of core competencies at the curriculum level: Using learning analytics for student reflection. *IEEE Transactions on Emerging Topics in Computing*, 5(1), 32–44. <https://doi.org/10.1109/TETC.2015.2501805>
- Dawson, S., & Hubball, H. (2014). Curriculum analytics: Application of social network analysis for improving strategic curriculum decision-making in a research-intensive university. *Teaching and Learning Inquiry*, 2(2), 59–74. <https://doi.org/10.20343/teachlearninqu.2.2.59>
- De Silva, L. M. H., Rodríguez-Triana, M. J., Chounta, I.-A., & Pishtari, G. (2024). Curriculum analytics in higher education institutions: A systematic literature review. *Journal of Computing in Higher Education*, 1–47. <https://doi.org/10.1007/s12528-024-09410-8>
- Decorte, J.-J., Van Haute, J., Deleu, J., Develder, C., & Demeester, T. (2022). Design of negative sampling strategies for distantly supervised skill extraction. *arXiv preprint arXiv:2209.05987*. <https://doi.org/10.48550/arXiv.2209.05987>
- Deng, Y., Zhang, W., Chen, Z., & Gu, Q. (2024). Rephrase and respond: Let large language models ask better questions for themselves. *arXiv preprint arXiv:2311.04205*. <https://doi.org/10.48550/arXiv.2311.04205>
- Department for Education. (2023). *Generative artificial intelligence (AI) in education* (tech. rep.). Government of the United Kingdom. <https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education>
- Doyle, A., Sridhar, P., Agarwal, A., Savelka, J., & Sakr, M. (2025). A comparative study of AI-generated and human-crafted learning objectives in computing education. *Journal of Computer Assisted Learning*, 41(1), e13092. <https://doi.org/10.1111/jcal.13092>
- Durant, E., Impagliazzo, J., Conry, S., Reese, R., Lam, H., Nelson, V., Hughes, J., Liu, W., Lu, J., & McGettrick, A. (2015). CE2016: Updated computer engineering curriculum guidelines. In *Proceedings of the 2015 IEEE Frontiers in Education Conference (FIE 2015)*, 21–24 October 2015, El Paso, Texas, USA (pp. 1–2). IEEE. <https://doi.org/10.1109/FIE.2015.7344157>
- Fiesler, C., Garrett, N., & Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education (SIGCSE 2020)*, 11–14 March 2020, Portland, Oregon, USA (pp. 289–295). ACM. <https://doi.org/10.1145/3328778.3366825>
- Ghanizadeh, A., Al-Hoorie, A. H., & Jahedizadeh, S. (2020). Higher order thinking skills. In *Higher order thinking skills in the language classroom: A concise guide* (pp. 1–51). Springer. [https://doi.org/10.1007/978-3-030-56711-8\\_1](https://doi.org/10.1007/978-3-030-56711-8_1)
- Gorski, P. C. (2009). What we're teaching teachers: An analysis of multicultural teacher education coursework syllabi. *Teaching and Teacher Education*, 25(2), 309–318. <https://doi.org/10.1016/j.tate.2008.07.008>
- Greer, J., Molinaro, M., Ochoa, X., & McKay, T. (2016). Learning analytics for curriculum and program quality improvement (pcla 2016). In *Proceedings of the Sixth International Conference on Learning Analytics and Knowledge (LAK 2016)*, 25–29 April 2016, Edinburgh, Scotland, UK (pp. 494–495). ACM. <https://doi.org/10.1145/2883851.2883899>
- Griffin, P., McGaw, B., & Care, E. (2012). *Assessment and teaching of 21st century skills* (Vol. 10). Springer. <https://doi.org/10.1007/978-3-319-65368-6>
- Herandi, A., Li, Y., Liu, Z., Hu, X., & Cai, X. (2024). Skill-LLM: Repurposing general-purpose LLMs for skill extraction. *arXiv preprint arXiv:2410.12052*. <https://doi.org/10.48550/arXiv.2410.12052>
- Hilliger, I., Aguirre, C., Miranda, C., Celis, S., & Pérez-Sanagustín, M. (2020). Design of a curriculum analytics tool to support continuous improvement processes in higher education. In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK 2020)*, 23–27 March 2020, Frankfurt, Germany (pp. 181–186). ACM. <https://doi.org/10.1145/3375462.3375489>
- Hilliger, I., Miranda, C., Celis, S., & Pérez-Sanagustín, M. (2024). Curriculum analytics adoption in higher education: A multiple case study engaging stakeholders in different phases of design. *British Journal of Educational Technology*, 55(3), 785–801. <https://doi.org/10.1111/bjet.13374>
- Homa, N., Hackathorn, J., Brown, C. M., Garczynski, A., Solomon, E. D., Tennial, R., Sanborn, U. A., & Gurung, R. A. (2013). An analysis of learning objectives and content coverage in introductory psychology syllabi. *Teaching of Psychology*, 40(3), 169–174. <https://doi.org/10.1177/0098628313487456>
- Hong, P. Y. P., & Hodge, D. R. (2009). Understanding social justice in social work: A content analysis of course syllabi. *Families in Society*, 90(2), 212–219. <https://doi.org/10.1606/1044-3894.3874>
- Irwin, R. (2002). Characterizing the core: What catalog descriptions of mandatory courses reveal about LIS schools and librarianship. *Journal of Education for Library and Information Science*, 175–184. <https://doi.org/10.2307/40323978>
- Javadian Sabet, A., Bana, S. H., Yu, R., & Frank, M. R. (2024). Course-Skill Atlas: A national longitudinal dataset of skills taught in US higher education curricula. *Scientific Data*, 11(1), 1086. <https://doi.org/10.1038/s41597-024-03931-8>

- Jayalath, V., Barthakur, A., Dawson, S., Tingey, J., Crase, L., & Kovanović, V. (2025). Scaling curriculum mapping in higher education: Evaluating generative AI's role in curriculum analytics. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Proceedings of the 2025 International Conference on Artificial Intelligence in Education (AIED 2025)*, 22–26 July 2025, Palermo, Italy (pp. 294–308). ACM. [https://doi.org/10.1007/978-3-031-98414-3\\_21](https://doi.org/10.1007/978-3-031-98414-3_21)
- Jovanović, J., Zamecnik, A., Barthakur, A., & Dawson, S. (2025). Curriculum analytics: Exploring assessment objectives, types, and grades in a study program. *Education and Information Technologies*, 30(4), 4843–4866. <https://doi.org/10.1007/s10639-024-13015-0>
- Kawintiranon, K., Vateekul, P., Suchato, A., & Punyabukkana, P. (2016). Understanding knowledge areas in curriculum through text mining from course materials. In *Proceedings of the 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE 2016)*, 7–9 December 2016, Bangkok, Thailand (pp. 161–168). IEEE. <https://doi.org/10.1109/TALE.2016.7851788>
- Kitto, K., Sarathy, N., Gromov, A., Liu, M., Musial, K., & Buckingham Shum, S. (2020). Towards skills-based curriculum analytics: Can we automate the recognition of prior learning? In *Proceedings of the 10th International Conference on Learning Analytics and Knowledge (LAK 2020)*, 23–27 March 2020, Frankfurt, Germany (pp. 171–180). ACM. <https://doi.org/10.1145/3375462.3375526>
- Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*. <https://doi.org/10.48550/arXiv.2205.11916>
- Kotsiou, A., Fajardo-Tovar, D. D., Cowhitt, T., Major, L., & Wegerif, R. (2022). A scoping review of future skills frameworks. *Irish Educational Studies*, 41(1), 171–186. <https://doi.org/10.1080/03323315.2021.2022522>
- Kozov, V., Ivanova, G., & Atanasova, D. (2024). Practical application of AI and large language models in software engineering education. *International Journal of Advanced Computer Science and Applications*, 15(1). <https://doi.org/10.14569/IJACSA.2024.0150168>
- Li, X., Henriksson, A., Duneld, M., Nouri, J., & Wu, Y. (2024). Supporting teaching-to-the-curriculum by linking diagnostic tests to curriculum goals. *Artificial Intelligence in Education*, 14829, 118–132. [https://doi.org/10.1007/978-3-031-64302-6\\_9](https://doi.org/10.1007/978-3-031-64302-6_9)
- Light, J. (2024). Student demand and the supply of college courses. <https://doi.org/10.2139/ssrn.4856488>
- Liu, C., Hoang, L., Stolman, A., & Wu, B. (2024). HiTA: A RAG-based educational platform that centers educators in the instructional loop. In A. Olney, I. Chounta, Z. Liu, O. Santos, & I. Bittencourt (Eds.), *Artificial intelligence in education. AIED 2024. Lecture notes in computer science* (pp. 405–412, Vol. 14830). Springer. [https://doi.org/10.1007/978-3-031-64299-9\\_37](https://doi.org/10.1007/978-3-031-64299-9_37)
- Lohr, D., Berges, M., Chugh, A., Kohlhase, M., & Müller, D. (2025). Leveraging large language models to generate course-specific semantically annotated learning objects. *Journal of Computer Assisted Learning*, 41(1), e13101. <https://doi.org/10.1111/jcal.13101>
- Lyu, W., Wang, Y., Chung, T., Sun, Y., & Zhang, Y. (2024). Evaluating the effectiveness of LLMs in introductory computer science education: A semester-long field study. In *Proceedings of the 11th ACM Conference on Learning at Scale (L@S 2024)*, 18–20 July 2024, Atlanta, Georgia, USA (pp. 63–74). ACM. <https://doi.org/10.1145/3657604.3662036>
- McKinsey Global Institute. (2023). *Generative AI and the future of work in America*. <https://www.mckinsey.com/mgi/our-research/generative-ai-and-the-future-of-work-in-america>
- Meyers, N. M., & Nulty, D. D. (2009). How to use (five) curriculum design principles to align authentic learning environments, assessment, students' approaches to thinking and learning outcomes. *Assessment & Evaluation in Higher Education*, 34(5), 565–577. <https://doi.org/10.1080/02602930802226502>
- Musa, F., Mufti, N., Latiff, R. A., & Amin, M. M. (2012). Project-based learning (PjBL): Inculcating soft skills in 21st century workplace. *Procedia-Social and Behavioral Sciences*, 59, 565–573. <https://doi.org/10.1016/j.sbspro.2012.09.315>
- National Education Association. (2024). *Teaching in the age of AI: NEA members' roadmap for safe, effective, and accessible use of artificial intelligence in education* (tech. rep.). Washington, DC. <https://www.nea.org/resource-library/artificial-intelligence-education>
- Nguyen, K. C., Zhang, M., Montariol, S., & Bosselut, A. (2024). Rethinking skill extraction in the job market domain using large language models. *arXiv preprint arXiv:2402.03832*. <https://doi.org/10.48550/arXiv.2402.03832>
- Nye, M., Hewitt, J., Chen, J., Krueger, D., Duvenaud, D., Lake, B., & Zemel, R. (2021). Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*. <https://doi.org/10.48550/arXiv.2112.00114>
- OECD. (2023a). *Innovating assessments to measure and support complex skills* (N. Foster & M. Piacentini, Eds.). <https://doi.org/10.1787/e5f3e341-en>
- OECD. (2023b). *OECD digital education outlook 2023: Towards an effective digital education ecosystem*. <https://doi.org/10.1787/c74f03de-en>

- OECD. (2023c). *OECD employment outlook 2023: Artificial intelligence and the labour market*. <https://doi.org/10.1787/08785bba-en>
- Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations* (tech. rep.). U.S. Department of Education. Washington, DC. <https://www.ed.gov/sites/ed/files/documents/ai-report/ai-report.pdf>
- Ohland, M., & Collins, R. (2002). Creating a catalog and meta analysis of freshman programs for engineering students: Part 2: Learning communities. In *Proceedings of the 2002 American Society for Engineering Education Annual Conference and Exposition*, 16–19 June 2002, Montréal, Québec, Canada (pp. 7–338). ASEE PEER. <https://doi.org/10.18260/1-2--10110>
- Pistilli, M. D., & Heileman, G. L. (2017). Guiding early and often: Using curricular and learning analytics to shape teaching, learning, and student success in gateway courses. *New Directions for Higher Education*, 2017(180), 21–30. <https://doi.org/10.1002/he.20258>
- Retnawati, H., Djidu, H., Apino, E., Anazifa, R. D., et al. (2018). Teachers' knowledge about higher-order thinking skills and its learning strategy. *Problems of Education in the 21st Century*, 76(2), 215–230. <https://doi.org/10.33225/pec/18.76.215>
- Senger, E., Zhang, M., van der Goot, R., & Plank, B. (2024). Deep learning-based computational job market analysis: A survey on skill extraction and classification from job postings. In E. Hruschka, T. Lake, N. Otani, & T. Mitchell (Eds.), *Proceedings of the First Workshop on Natural Language Processing for Human Resources (NLP4HR 2024)*, 22 March 2024, St. Julian's, Malta (pp. 1–15). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.nlp4hr-1.1>
- Shorman, S., Khder, M., et al. (2024). Curriculum management system to measure the course and program outcomes. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS 2024)*, 28–29 January 2024, Manama, Bahrain (pp. 391–397). IEEE. <https://doi.org/10.1109/ICETSIS61505.2024.10459625>
- Siyam, L., Xu, Z., Raghuram, V. C., Zhang, X., Yu, R., & Yu, Z. (2025). Bringing pedagogy into focus: Evaluating virtual teaching assistants' question-answering in asynchronous learning environments. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng (Eds.), *Findings of the Association for Computational Linguistics (EMNLP 2025)*, 4–9 November 2025, Suzhou, China (pp. 9743–9774). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.findings-emnlp.518>
- Sridhar, P., Doyle, A., Agarwal, A., Bogart, C., Savelka, J., & Sakr, M. (2023). Harnessing LLMs in curricular design: Using GPT-4 to support authoring of learning objectives. *arXiv preprint arXiv:2306.17459*. <https://doi.org/10.48550/arXiv.2306.17459>
- Tan, C. W., & Lim, K. Y. (2023). Revolutionizing formative assessment in STEM fields: Leveraging AI and NLP techniques. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2023)*, 31 October–3 November 2023, Taipei, Taiwan (pp. 1357–1364). IEEE. <https://doi.org/10.1109/APSIPAASC58517.2023.10317226>
- Tang, R., & Sae-Lim, W. (2016). Data science programs in U.S. higher education: An exploratory content analysis of program description, curriculum structure, and course focus. *Education for Information*, 32(3), 269–290. <https://doi.org/10.3233/efi-160977>
- Thakrar, K., & Young, N. (2025). Enhancing talent employment insights through feature extraction with LLM finetuning. *arXiv preprint arXiv:2501.07663*. <https://doi.org/10.48550/arXiv.2501.07663>
- Tian, Z., Sun, M., Liu, A., Sarkar, S., & Liu, J. (2024). Enhancing instructional quality: Leveraging computer-assisted textual analysis to generate in-depth insights from educational artifacts. *arXiv preprint arXiv:2403.03920*. <https://doi.org/10.48550/arXiv.2403.03920>
- UNESCO. (2023). *Guidance for generative AI in education and research* (tech. rep.). UNESCO. Paris. <https://unesdoc.unesco.org/ark:/48223/pf0000386693>
- Walker, R. E. (2024). Mapping curricula to skills and occupations using course descriptions. In C. da Rocha Brito & M. M. Ciampi (Eds.), *Proceedings of the 2024 IEEE World Engineering Education Conference (EDUNINE 2024)*, 10–13 March 2024, Guatemala City, Guatemala. IEEE. <https://doi.org/10.1109/EDUNINE60625.2024.10500452>
- Wang, Y., Zhang, Z., & Wang, R. (2023). Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*. <https://doi.org/10.48550/arXiv.2305.13412>
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*. <https://doi.org/10.48550/arXiv.2206.07682>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Proceedings of the 36th International Conference on Neural Information Processing Systems*

- (NIPS 2022), 28 November 2022–9 December 2022, New Orleans, Louisiana, USA (pp. 24824–24837). ACM. <https://dl.acm.org/doi/10.5555/3600270.3602070>
- World Economic Forum. (2025). *The future of jobs report 2025*. World Economic Forum. <https://www.weforum.org/publications/the-future-of-jobs-report-2025/>
- Xu, Z., Li, X., Huan, Y., Minaya, V., & Yu, R. (2025). From course to skill: Evaluating large language model performance in curricular analytics. In A. Cristea, E. Walker, Y. Lu, O. Santos, & S. Isotani (Eds.), *Artificial intelligence in education. AIED 2025. Lecture notes in computer science* (pp. 203–211, Vol. 15882). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-98465-5\\_26](https://doi.org/10.1007/978-3-031-98465-5_26)
- Yang, H., Kim, J., & Lee, W. (2023). Analyzing the alignment between AI curriculum and AI textbooks through text mining. *Applied Sciences*, 13(18), 10011. <https://doi.org/10.3390/app131810011>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*. <https://doi.org/10.48550/arXiv.2305.10601>
- Zamecnik, A., Barthakur, A., Wang, H., & Dawson, S. (2024). Mapping employable skills in higher education curriculum using llms. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & R. Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education. EC-TEL 2024. Lecture notes in computer science* (pp. 18–32, Vol. 15160). Springer. [https://doi.org/10.1007/978-3-031-72312-4\\_2](https://doi.org/10.1007/978-3-031-72312-4_2)
- Zhang, M., Jensen, K., Sonniks, S., & Plank, B. (2022). Skillspan: Hard and soft skill extraction from english job postings. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2022)*, 10–15 July 2022, Seattle, Washington, USA (pp. 4962–4984). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2022.naacl-main.366>
- Zhang, S., Qin, L., Zhou, D., Le, Q. V., Liu, P. J., et al. (2022). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*. <https://doi.org/10.48550/arXiv.2210.03493>
- Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q., & Chi, E. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*. <https://doi.org/10.48550/arXiv.2205.10625>



## Appendix

### 1. Details of competency frameworks

**Table 8.** Summary of competency items

| Competency framework          | Competencies   |
|-------------------------------|--|
| O*NET(Workforce Competencies) | Complex problem solving; Critical thinking; Deductive reasoning; Judgment and decision making; Inductive reasoning; Category flexibility; Perceptual speed; Information ordering; Cooperation; Social Interaction; Concerns for others; Leadership; Persistence; Achievement/effort; Initiative; Originality; Innovation; Oral expression; Oral comprehension; Written expression; Written comprehension |
| EU(Key Competences)           | Literacy competence; Multilingual competence; Mathematical competence and competence in science, technology and engineering; Digital competence; Personal, social and learning to learn competence; Citizenship competence; Entrepreneurship competence; Cultural awareness and expression competence  |
| ESDC (Success Model)          | Adaptability; collaboration; Communication; Creativity and Innovation; Digital; Numeracy; Problem Solving; Reading; Writing  |

### 2. Human annotation

#### B.1 Score distribution in human annotation

**Table 9.** Score distribution across competency frameworks

| Competency frameworks | Score | Learning Objective |      | Concise course description |      | Detailed course description |      | Learning activity content |      | Instructional schedule |      |
|-----------------------|-------|--------------------|------|----------------------------|------|-----------------------------|------|---------------------------|------|------------------------|------|
|                       |       | N                  | %    | N                          | %    | N                           | %    | N                         | %    | N                      | %    |
| EU                    | 0     | 106                | 30.8 | 157                        | 40.9 | 142                         | 38.6 | 113                       | 35.3 | 100                    | 31.2 |
|                       | 1     | 57                 | 16.6 | 42                         | 10.9 | 50                          | 13.6 | 55                        | 17.2 | 45                     | 14.1 |
|                       | 2     | 57                 | 16.6 | 54                         | 14.1 | 50                          | 13.6 | 55                        | 17.2 | 30                     | 9.4  |
|                       | 3     | 62                 | 18.0 | 47                         | 12.2 | 48                          | 13.0 | 78                        | 24.4 | 60                     | 18.8 |
|                       | NA    | 62                 | 18.0 | 84                         | 21.9 | 78                          | 21.2 | 19                        | 5.9  | 85                     | 26.6 |
| O*NET                 | 0     | 50                 | 5.5  | 126                        | 12.5 | 128                         | 13.3 | 21                        | 2.5  | 182                    | 21.7 |
|                       | 1     | 201                | 22.3 | 275                        | 27.3 | 219                         | 22.7 | 213                       | 25.4 | 111                    | 13.2 |
|                       | 2     | 373                | 41.3 | 286                        | 28.4 | 320                         | 33.1 | 319                       | 38.0 | 164                    | 19.5 |
|                       | 3     | 92                 | 10.2 | 53                         | 5.3  | 67                          | 6.9  | 78                        | 9.3  | 5                      | 0.6  |
|                       | NA    | 187                | 20.7 | 268                        | 26.6 | 232                         | 24.0 | 209                       | 24.9 | 378                    | 45.0 |
| ESDC                  | 0     | 33                 | 9.6  | 47                         | 12.2 | 54                          | 14.7 | 55                        | 17.2 | 86                     | 26.9 |
|                       | 1     | 99                 | 28.8 | 132                        | 34.4 | 77                          | 20.9 | 85                        | 26.6 | 54                     | 16.9 |
|                       | 2     | 121                | 35.2 | 110                        | 28.6 | 127                         | 34.5 | 132                       | 41.2 | 64                     | 20.0 |
|                       | 3     | 50                 | 14.5 | 44                         | 11.5 | 34                          | 9.2  | 18                        | 5.6  | 2                      | 0.6  |
|                       | NA    | 41                 | 11.9 | 51                         | 13.3 | 76                          | 20.7 | 30                        | 9.4  | 114                    | 35.6 |

**B.2 human annotation notes**

**Table 10.** O\*NET(Workforce Competencies)

| Competency                              | Definition  |
|---|---|
| Complex problem solving                 | Identifying complex problems and reviewing related information to develop and evaluate options and implement solutions.   |
| Critical thinking                       | Using logic and reasoning to identify the strengths and weaknesses of alternative solutions, conclusions, or approaches to problems.  |
| Deductive reasoning                     | Applying general rules to specific situations to produce logical answers, ensuring that the reasoning process is transparent and justifiable.   |
| Judgment and decision making            | Evaluating the relative costs and benefits of various actions or solutions, and choosing the option that is most appropriate for the context.   |
| Inductive reasoning                     | Integrating multiple pieces of information to formulate general rules or conclusions, including identifying patterns or relationships among seemingly unrelated data.   |
| Category flexibility                    | Generating or applying different sets of rules to combine or group things in novel ways, and flexibly adapting categorization criteria to fit new situations or information.  |
| Perceptual speed                        | Quickly and accurately comparing similarities and differences among sets of letters, numbers, objects, pictures, or patterns, and efficiently identifying discrepancies or matches.                                 |
| Information ordering                    | Arranging actions, information, or objects in a specific sequence according to established rules, instructions, or logical patterns, ensuring that each step follows a coherent and justifiable order.              |
| Cooperation                             | Demonstrating a pleasant and cooperative attitude with others in the work environment, actively contributing to group efforts, and supporting colleagues to achieve common goals.                                   |
| Social interaction / Social orientation | Preferring to work collaboratively with others rather than independently, establishing and maintaining effective interpersonal relationships, and engaging in constructive social exchanges within a group or team. |
| Concern for others                      | Showing sensitivity to the needs and feelings of others, offering understanding and practical help, and responding appropriately to others’ emotional or situational challenges.                                    |
| Leadership                              | Willingly taking charge of situations or groups, offering direction and constructive opinions, motivating others, and assuming responsibility for group outcomes or decisions.                                      |
| Persistence                             | Sustaining effort and motivation in the face of obstacles, setbacks, or difficulties, continuing to pursue goals despite challenges or repeated failures.   |
| Achievement/effort                      | Setting personally challenging achievement goals, maintaining high standards for one’s own performance, and exerting consistent effort to master difficult tasks.   |
| Initiative                              | Proactively taking on new responsibilities and challenges, identifying and acting on opportunities without needing external prompting or supervision.   |
| Originality                             | Developing unusual or clever ideas for a given topic, problem, or situation, and producing creative solutions or approaches that differ from conventional practices.  |
| Innovation                              | Formulating and implementing novel ideas, processes, or products that improve outcomes or create added value, and actively seeking opportunities for creative change.   |
| Oral expression                         | Communicating information and ideas clearly and effectively in spoken language, ensuring that listeners can readily understand the intended meaning.  |
| Oral comprehension                      | Listening to and accurately understanding information, instructions, or ideas presented through spoken words and sentences in various contexts.   |
| Written expression                      | Conveying information and ideas effectively in writing, producing clear and understandable documents, messages, or reports tailored to the audience.  |
| Written comprehension                   | Reading and comprehending information and ideas presented in written form, accurately interpreting meaning, context, and relevant details.  |

**Table 11. EU(Key Competences)**

| Competency   | Definition   |
|--|--|
| Literacy competence  | The ability to identify, understand, express, create, and interpret concepts, feelings, facts, and opinions in both oral and written forms, using visual, audio, and digital materials across disciplines and contexts.  |
| Multilingual competence  | The ability to use different languages appropriately and effectively for communication.  |
| Mathematical competence and competence in science, technology, and engineering | The ability to develop and apply mathematical thinking and understanding to solve a range of problems in everyday situations. Competence in science, technology, and engineering involves the application of knowledge and methodology to explain the natural world and to use that knowledge to identify questions and draw evidence-based conclusions. |
| Digital competence   | The confident, critical, and responsible use of digital technologies for learning, at work, and for participation in society.  |
| Personal, social, and learning to learn competence                             | The ability to reflect upon oneself, manage time and information effectively, work with others in a constructive way, remain resilient, and manage one’s own learning and career.  |
| Citizenship competence   | The ability to act as responsible citizens and to fully participate in civic and social life, based on understanding of social, economic, legal, and political concepts and structures, as well as global developments and sustainability.   |
| Entrepreneurship competence  | The capacity to act upon opportunities and ideas, and to transform them into values for others.  |
| Cultural awareness and expression competence                                   | The understanding of and respect for how ideas and meaning are creatively expressed and communicated in different cultures and through a range of arts and other cultural forms.   |

**Table 12. ESDC (Success Model)**

| Competency                | Definition   |
|---------------------------|--|
| Adaptability              | Adjusting behavior or strategies in response to change or unexpected challenges, while maintaining focus and persistence.    |
| collaboration             | Working constructively with others to achieve shared goals, including communicating openly and sharing responsibilities.     |
| Communication             | Expressing and understanding information clearly through various modes, and adapting messages appropriately for the context. |
| Communication             | Expressing and understanding information clearly through various modes, and adapting messages appropriately for the context. |
| Creativity and Innovation | Generating and applying new ideas or approaches to improve processes, solve problems, or create value.                       |
| Digital                   | Effectively using digital tools and technologies to find, manage, and communicate information.                               |
| Numeracy                  | Interpreting and applying quantitative information and calculations to solve practical problems.                             |
| Problem Solving           | Identifying issues, analyzing relevant information, and implementing effective solutions.                                    |
| Reading                   | Understanding, interpreting, and using written information to accomplish tasks.  |
| Writing                   | Organizing and conveying information clearly and effectively in written form.  |

### 3. Comprehensive evaluation results

C.1 RQ2

**Table 13.** Zero-shot performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.248 | 0.396 | 0.299    | 0.312     | 0.274  | 0.365 |
| 5           | Llama3-70B    | 0.304 | 0.474 | 0.35     | 0.338     | 0.309  | 0.383 |
| 5           | GPT 3.5-turbo | 0.282 | 0.429 | 0.345    | 0.294     | 0.272  | 0.39  |
| 5           | GPT4o         | 0.301 | 0.456 | 0.386    | 0.362     | 0.31   | 0.416 |
| 4           | Llama3-8B     | 0.204 | 0.343 | 0.44     | 0.406     | 0.34   | 0.48  |
| 4           | Llama3-70B    | 0.246 | 0.386 | 0.498    | 0.442     | 0.376  | 0.52  |
| 4           | GPT 3.5-turbo | 0.244 | 0.369 | 0.472    | 0.401     | 0.345  | 0.49  |
| 4           | GPT4o         | 0.221 | 0.333 | 0.559    | 0.491     | 0.366  | 0.575 |
| 3           | Llama3-8B     | 0.204 | 0.343 | 0.44     | 0.406     | 0.34   | 0.48  |
| 3           | Llama3-70B    | 0.258 | 0.413 | 0.549    | 0.505     | 0.476  | 0.528 |
| 3           | GPT 3.5-turbo | 0.25  | 0.396 | 0.529    | 0.474     | 0.447  | 0.508 |
| 3           | GPT4o         | 0.224 | 0.351 | 0.605    | 0.55      | 0.467  | 0.588 |
| 2           | Llama3-8B     | 0.213 | 0.319 | 0.715    | 0.682     | 0.618  | 0.679 |
| 2           | Llama3-70B    | 0.178 | 0.264 | 0.724    | 0.701     | 0.601  | 0.721 |
| 2           | GPT 3.5-turbo | 0.192 | 0.267 | 0.724    | 0.661     | 0.602  | 0.704 |
| 2           | GPT4o         | 0.155 | 0.231 | 0.729    | 0.715     | 0.593  | 0.742 |

C.2 RQ3

**Table 14.** Zero-Shot with competency definition (DEF) performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.239 | 0.395 | 0.305    | 0.311     | 0.269  | 0.283 |
| 5           | Llama3-70B    | 0.345 | 0.521 | 0.337    | 0.316     | 0.286  | 0.311 |
| 5           | GPT 3.5-turbo | 0.254 | 0.393 | 0.337    | 0.29      | 0.275  | 0.285 |
| 5           | GPT4o         | 0.338 | 0.515 | 0.363    | 0.331     | 0.282  | 0.344 |
| 4           | Llama3-8B     | 0.239 | 0.395 | 0.305    | 0.311     | 0.269  | 0.283 |
| 4           | Llama3-70B    | 0.345 | 0.521 | 0.337    | 0.316     | 0.286  | 0.311 |
| 4           | GPT 3.5-turbo | 0.254 | 0.393 | 0.337    | 0.29      | 0.275  | 0.285 |
| 4           | GPT4o         | 0.338 | 0.515 | 0.363    | 0.331     | 0.282  | 0.344 |
| 3           | Llama3-8B     | 0.189 | 0.327 | 0.455    | 0.389     | 0.322  | 0.389 |
| 3           | Llama3-70B    | 0.298 | 0.474 | 0.533    | 0.487     | 0.467  | 0.445 |
| 3           | GPT 3.5-turbo | 0.212 | 0.333 | 0.512    | 0.436     | 0.425  | 0.417 |
| 3           | GPT4o         | 0.248 | 0.406 | 0.571    | 0.495     | 0.444  | 0.49  |
| 2           | Llama3-8B     | 0.206 | 0.304 | 0.7      | 0.637     | 0.597  | 0.597 |
| 2           | Llama3-70B    | 0.2   | 0.291 | 0.701    | 0.666     | 0.596  | 0.652 |
| 2           | GPT 3.5-turbo | 0.129 | 0.171 | 0.671    | 0.578     | 0.558  | 0.577 |
| 2           | GPT4o         | 0.147 | 0.227 | 0.69     | 0.682     | 0.579  | 0.687 |

**Table 15.** Curricular CoT (CQA) performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.304 | 0.473 | 0.323    | 0.303     | 0.286  | 0.299 |
| 5           | Llama3-70B    | 0.337 | 0.519 | 0.338    | 0.338     | 0.292  | 0.319 |
| 5           | GPT 3.5-turbo | 0.304 | 0.481 | 0.322    | 0.291     | 0.279  | 0.281 |
| 5           | GPT4o         | 0.35  | 0.505 | 0.378    | 0.375     | 0.292  | 0.351 |
| 4           | Llama3-8B     | 0.252 | 0.416 | 0.452    | 0.385     | 0.344  | 0.4   |
| 4           | Llama3-70B    | 0.279 | 0.444 | 0.484    | 0.433     | 0.357  | 0.444 |
| 4           | GPT 3.5-turbo | 0.261 | 0.415 | 0.444    | 0.367     | 0.34   | 0.366 |
| 4           | GPT4o         | 0.264 | 0.403 | 0.536    | 0.48      | 0.348  | 0.465 |
| 3           | Llama3-8B     | 0.252 | 0.416 | 0.452    | 0.385     | 0.344  | 0.4   |
| 3           | Llama3-70B    | 0.299 | 0.477 | 0.551    | 0.505     | 0.48   | 0.46  |
| 3           | GPT 3.5-turbo | 0.277 | 0.45  | 0.535    | 0.491     | 0.448  | 0.43  |
| 3           | GPT4o         | 0.273 | 0.423 | 0.599    | 0.546     | 0.464  | 0.503 |
| 2           | Llama3-8B     | 0.165 | 0.259 | 0.704    | 0.673     | 0.585  | 0.703 |
| 2           | Llama3-70B    | 0.183 | 0.281 | 0.701    | 0.655     | 0.589  | 0.645 |
| 2           | GPT 3.5-turbo | 0.209 | 0.308 | 0.712    | 0.634     | 0.595  | 0.605 |
| 2           | GPT4o         | 0.251 | 0.372 | 0.71     | 0.663     | 0.623  | 0.637 |

**Table 16.** Curricular CoT (CQ) performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.279 | 0.444 | 0.316    | 0.29      | 0.278  | 0.293 |
| 5           | Llama3-70B    | 0.332 | 0.516 | 0.334    | 0.342     | 0.289  | 0.31  |
| 5           | GPT 3.5-turbo | 0.288 | 0.445 | 0.321    | 0.273     | 0.274  | 0.293 |
| 5           | GPT4o         | 0.35  | 0.519 | 0.37     | 0.35      | 0.287  | 0.363 |
| 4           | Llama3-8B     | 0.232 | 0.389 | 0.448    | 0.37      | 0.339  | 0.385 |
| 4           | Llama3-70B    | 0.273 | 0.432 | 0.479    | 0.435     | 0.352  | 0.429 |
| 4           | GPT 3.5-turbo | 0.254 | 0.399 | 0.438    | 0.33      | 0.335  | 0.377 |
| 4           | GPT4o         | 0.252 | 0.393 | 0.531    | 0.449     | 0.338  | 0.487 |
| 3           | Llama3-8B     | 0.232 | 0.389 | 0.448    | 0.37      | 0.339  | 0.385 |
| 3           | Llama3-70B    | 0.285 | 0.463 | 0.541    | 0.498     | 0.468  | 0.455 |
| 3           | GPT 3.5-turbo | 0.265 | 0.435 | 0.517    | 0.458     | 0.439  | 0.422 |
| 3           | GPT4o         | 0.258 | 0.412 | 0.584    | 0.52      | 0.454  | 0.502 |
| 2           | Llama3-8B     | 0.262 | 0.388 | 0.717    | 0.657     | 0.626  | 0.633 |
| 2           | Llama3-70B    | 0.19  | 0.291 | 0.705    | 0.672     | 0.593  | 0.647 |
| 2           | GPT 3.5-turbo | 0.197 | 0.291 | 0.706    | 0.635     | 0.59   | 0.58  |
| 2           | GPT4o         | 0.165 | 0.256 | 0.704    | 0.688     | 0.586  | 0.695 |

**Table 17.** Curricular CoT (QA) performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.277 | 0.43  | 0.307    | 0.298     | 0.273  | 0.288 |
| 5           | Llama3-70B    | 0.318 | 0.483 | 0.341    | 0.322     | 0.291  | 0.324 |
| 5           | GPT 3.5-turbo | 0.291 | 0.454 | 0.328    | 0.261     | 0.269  | 0.3   |
| 5           | GPT4o         | 0.334 | 0.491 | 0.375    | 0.364     | 0.293  | 0.349 |
| 4           | Llama3-8B     | 0.231 | 0.381 | 0.437    | 0.373     | 0.337  | 0.396 |
| 4           | Llama3-70B    | 0.272 | 0.419 | 0.487    | 0.424     | 0.357  | 0.459 |
| 4           | GPT 3.5-turbo | 0.245 | 0.387 | 0.446    | 0.323     | 0.325  | 0.393 |
| 4           | GPT4o         | 0.256 | 0.394 | 0.535    | 0.461     | 0.347  | 0.468 |
| 3           | Llama3-8B     | 0.231 | 0.381 | 0.437    | 0.373     | 0.337  | 0.396 |
| 3           | Llama3-70B    | 0.284 | 0.445 | 0.55     | 0.501     | 0.476  | 0.48  |
| 3           | GPT 3.5-turbo | 0.253 | 0.416 | 0.525    | 0.445     | 0.426  | 0.421 |
| 3           | GPT4o         | 0.267 | 0.42  | 0.594    | 0.528     | 0.463  | 0.497 |
| 2           | Llama3-8B     | 0.287 | 0.435 | 0.706    | 0.673     | 0.645  | 0.634 |
| 2           | Llama3-70B    | 0.197 | 0.305 | 0.705    | 0.663     | 0.595  | 0.657 |
| 2           | GPT 3.5-turbo | 0.188 | 0.278 | 0.707    | 0.625     | 0.583  | 0.596 |
| 2           | GPT4o         | 0.184 | 0.299 | 0.705    | 0.661     | 0.595  | 0.643 |

**Table 18.** Curricular CoT (A) performance across different levels of classification granularity.

| $N_{class}$ | Model         | Kappa | ICC   | Accuracy | Precision | Recall | F1    |
|-------------|---------------|-------|-------|----------|-----------|--------|-------|
| 5           | Llama3-8B     | 0.261 | 0.42  | 0.304    | 0.308     | 0.271  | 0.289 |
| 5           | Llama3-70B    | 0.319 | 0.481 | 0.347    | 0.324     | 0.292  | 0.327 |
| 5           | GPT 3.5-turbo | 0.29  | 0.462 | 0.332    | 0.295     | 0.268  | 0.278 |
| 5           | GPT4o         | 0.347 | 0.503 | 0.381    | 0.361     | 0.299  | 0.347 |
| 4           | Llama3-8B     | 0.218 | 0.368 | 0.424    | 0.384     | 0.327  | 0.392 |
| 4           | Llama3-70B    | 0.269 | 0.413 | 0.493    | 0.414     | 0.352  | 0.454 |
| 4           | GPT 3.5-turbo | 0.252 | 0.406 | 0.454    | 0.374     | 0.323  | 0.356 |
| 4           | GPT4o         | 0.268 | 0.407 | 0.538    | 0.454     | 0.347  | 0.451 |
| 3           | Llama3-8B     | 0.218 | 0.368 | 0.424    | 0.384     | 0.327  | 0.392 |
| 3           | Llama3-70B    | 0.281 | 0.439 | 0.557    | 0.501     | 0.469  | 0.463 |
| 3           | GPT 3.5-turbo | 0.265 | 0.438 | 0.533    | 0.468     | 0.436  | 0.424 |
| 3           | GPT4o         | 0.279 | 0.431 | 0.599    | 0.528     | 0.465  | 0.513 |
| 2           | Llama3-8B     | 0.263 | 0.405 | 0.696    | 0.668     | 0.631  | 0.64  |
| 2           | Llama3-70B    | 0.204 | 0.317 | 0.707    | 0.655     | 0.597  | 0.641 |
| 2           | GPT 3.5-turbo | 0.213 | 0.336 | 0.712    | 0.662     | 0.596  | 0.591 |
| 2           | GPT4o         | 0.192 | 0.303 | 0.71     | 0.67      | 0.599  | 0.663 |