PRIVACY-ADAPTIVE GENERATIVE AI: FROM UNSTRUCTURED REPORTS TO COMPETENCY-BASED LEARNING ANALYTICS

Saul Garcia-Huertes, Ramon Bragós

Telecos-BCN, Universitat Politècnica de Catalunya (UPC)

ABSTRACT

Background: Challenge-Based Learning (CBL) courses generate a wealth of unstructured student artifacts (e.g., reports, prototypes) that contain rich evidence of professional competencies (CDIO Standard 11). However, extracting these process variables for longitudinal analysis has historically required prohibitive manual effort, leaving valuable curriculum insights inaccessible. As a result, assessment practices often rely on coarse indicators (e.g., grades), limiting evidence-based feedback on innovation and professional skills. **Challenge:** While Generative AI offers a scalable solution to automate this extraction, institutions face a dilemma: utilizing frontier cloud models risks violating data sovereignty regulations, while privacy-preserving local models have historically lacked the reasoning capability required for complex assessment. **Methodology:** This paper presents a Privacy-Adaptive Dual-Pipeline framework to resolve this tension. We analyze a subset of 15-year archive of engineering capstone projects ($N = 63$ validated samples) by benchmarking a local, offline Large Language Model (LLM) against a frontier Cloud model. **Results:** We identify a distinct *Inference Gap*: while local models successfully extract explicit identity metadata (Title: 78.2% match), they struggle with semantic reasoning tasks (e.g., Value Proposition: 0.0% match) compared to the cloud benchmark. **Contribution:** We propose a tiered implementation strategy that allows educators to select the appropriate privacy tier based on data sensitivity. This methodology provides the CDIO community with a validated workflow to transform unstructured repositories into structured datasets, enabling scalable, evidence-based assessment and feedback without compromising student privacy.

KEYWORDS

Generative AI, Learning Analytics, Challenge-Based Learning, Data Extraction, CDIO Standards: 2, 5, 11

INTRODUCTION

Challenge-Based Learning (CBL) is a primary vehicle for implementing CDIO Standard 5 (Design-Implement Experiences) (Kohn Rådberg et al., 2020; Malmqvist et al., 2015). By engaging students in open-ended, real-world problems, these courses foster both technical competence and professional skills. However, systematically assessing such multifaceted experiences remains challenging for engineering educators (CDIO Standard 11).

Although quantitative indicators such as grades are readily available, they fail to capture the process-level variables that shape student outcomes. Information related to team organization, prototype fidelity, or technical decision-making is typically embedded in unstructured artifacts such as project reports. In long-running courses, this results in substantial dark data.

Since the widespread availability of generative AI, about two years ago, the methodology for evaluating project-based and challenge-based subjects is moving towards observing the process rather than observing the results. However, having information on various indicators in a semi-automatic way to evaluate projects can be very useful. In addition, being able to obtain information on the characteristics of challenges that generate better learning outcomes can be very useful for the institution's Learning Analytics.

Generative AI offers a pathway to unlock this archive but introduces a tension between capability and privacy. While frontier cloud models offer the reasoning depth to analyze complex documents, transmitting student data to third parties raises sovereignty concerns. Conversely, local models ensure privacy but often lack the inference capability for high-level assessment tasks. To resolve this, we propose a Privacy-Adaptive Dual-Pipeline Framework benchmarking local extraction against a cloud baseline.

The study is situated in the Advanced Engineering Project (PAE) and Challenge-Based Innovation (CBI) courses at the Universitat Politècnica de Catalunya (UPC) (Bragós Bardia et al., 2022; Garcia-Huertes et al., 2024). Since 2011, these programs have collectively generated over 200 projects involving 1,900+ students. Using this dataset, we address the following research question: *To what extent can privacy-preserving local Large Language Models (LLMs) replicate the extraction capabilities of frontier cloud models when analyzing unstructured engineering project artifacts, and across which dimensions does their performance diverge?*

The contribution of this paper is methodological. We present a validated, replicable approach that allows institutions to quantify trade-offs between local and cloud-based extraction, enabling tiered assessment without compromising data sovereignty. Although situated within a CDIO-aligned program, the proposed methodology is designed to be transferable to similar courses across higher education institutions that rely on unstructured student artifacts for assessment.

LITERATURE REVIEW

To position the methodological contribution of this paper, this literature review is structured around three themes: assessment challenges in CDIO-aligned CBL environments; the resulting data gap in quantifying project-based competencies; and the emerging role of Generative AI in addressing this gap.

Challenge-Based Learning and the CDIO Assessment Challenge

CBL and Project-Based Learning (PBL) are foundational to the CDIO Syllabus, particularly in fostering competencies in Conceiving, Designing, Implementing, and Operating (Standards 5 and 6). The Advanced Engineering Project course aligns with these principles by requiring students to manage the full lifecycle of a solution from ideation to pitch. Research in engineering entrepreneurship education indicates that such courses enhance entrepreneurial intentions and skills, although measuring their long-term impact remains challenging (Cobb et al., 2016).

Despite their educational value, the rigorous assessment of non-technical skills such as teamwork, communication, and business model viability remains difficult (Standard 11). The CDIO initiative has long acknowledged the tension between usefulness in improving educational practice and scholarliness in generating generalizable knowledge (Edström, 2016). Addressing both requires assessment approaches that move beyond subjective faculty impressions toward robust, evidence-based data.

The Data Gap in Assessing Project-Based Competencies

Engineering Education Research (EER) on project-based courses has traditionally relied on subjective data (surveys, interviews) or aggregated metrics like final grades. Recent bibliographic analyses within the CDIO community show that while interest in outcome measurement is increasing, most studies remain survey-based and focus on constructs like entrepreneurial mindset or intention rather than observable behaviors (Garcia-Huertes & Bragós, 2023).

Deeper quantitative analyses face a data resolution problem, since key explanatory variables including challenge type, team structure, and prototype fidelity are embedded in unstructured artifacts such as reports and presentations. Extracting these process-level variables typically requires time-intensive manual content analysis, constraining study scale (Berdanier et al., 2018; Borrego et al., 2014). This methodological bottleneck limits large-scale synthesis and hinders the identification of high-impact instructional practices.

The Emerging Role of Generative AI in Educational Research Methodology

The maturation of Generative AI and Large Language Models (LLMs) represents a significant shift in educational research methodology. Systematic reviews report rapid adoption of LLMs across engineering disciplines, though primarily for instructional support rather than assessment-focused analysis (Filippi & Motyl, 2024). Emerging work, however, demonstrates that LLMs can function as *zero-shot* data annotators, offering scalable alternatives to manual coding in computational social science (Ziems et al., 2024).

Empirical studies have begun to validate this approach in educational contexts. For example, Ding et al. (2024) showed that LLMs such as LLaMA and BERT can outperform traditional machine learning in analyzing qualitative student survey responses. Similarly, Tai et al. (2024) proposed a systematic methodology for using LLMs to support deductive coding of interview transcripts, highlighting the role of recursive analysis in improving reliability.

Despite this promise, adoption is constrained by data sovereignty: while frontier cloud models offer superior reasoning, transmitting student artifacts to third parties introduces significant privacy risks (Bahrini et al., 2023; Hingle et al., 2023). Furthermore, empirical evidence on the

viability of privacy-preserving local models for complex extraction remains limited. This paper bridges this gap by benchmarking local versus cloud performance, validating a privacy-aware methodology for extracting assessment variables from unstructured artifacts at scale.

METHODOLOGICAL FRAMEWORK

To extract structured data from the unstructured historical archive of the PAE and CBI courses, we designed a privacy-first, scalable data processing pipeline. The methodology employs a Privacy-Adaptive Dual-Pipeline approach that leverages recent advances in multimodal LLMs while adhering to strict data protection standards.

Data Corpus and Execution Environment

Although the complete PAE and CBI archive spans more than 15 years (200+ projects), this study focuses on a stratified sample of 63 final project reports from recent academic periods (2019–2024). This subset was selected to validate the extraction methodology against contemporary technical vocabularies and report formats. The corpus consists of unstructured PDF files averaging 50–60 pages.

All experiments were conducted in a Google Colab environment with Institutional Account of Google Drive used for secure file storage. This setup supports reproducibility and reflects the capabilities of standard commercial laptops (e.g., NVIDIA T4 GPU), making the methodology accessible without dedicated HPC infrastructure. In addition, Google's enterprise Data Privacy Terms & Conditions (T&C) ensure compliance with GDPR requirements for educational data.

Privacy Pre-processing Layer

A core element of the framework is a common Privacy Pre-processing Layer (Figure 1). All PDF artifacts undergo automated sanitization implemented in Python. *PyMuPDF* is used for document parsing, combined with *spaCy* models and Regular Expressions to redact Personally Identifiable Information (PII) such as names and email addresses. OpenCV with Haar Cascade classifiers is additionally used to detect and obfuscate human faces in project images.

Dual-Pipeline Architecture

Sanitized documents are processed through two parallel pipelines for benchmarking purposes, enabling systematic comparison of offline local extraction against a cloud benchmark and evaluation of trade-offs between assessment depth, infrastructure, and data sovereignty:

1. **Pipeline A (Cloud Benchmark):** Redacted text is transmitted to Google's *Gemini 2.5 Flash-Lite*. This frontier model was selected for its large context window ($>$1 million tokens) and strong multimodal reasoning capabilities. Document structure analysis (e.g., table parsing, header and footer exclusion) is performed in a single inference pass, trading data locality for high-precision extraction.
2. **Pipeline B (Local/Private):** Redacted text is processed locally using *DeepSeek-R1-Distill-Llama-8B*, optimized with the *Unsloth* framework. This pipeline demonstrates fully offline execution on consumer hardware. Owing to the model's limited context window

(8k tokens), a multi-step strategy is required:

- **Context Windowing:** The first 8,000 characters and final 4,000 characters of each report are concatenated, capturing sections with high metadata density.
- **Two-Pass Extraction:** Extraction is divided into two sequential passes, with Pass 1 targeting Project Identity variables (Title, Tech Stack) and Pass 2 focusing on Business Metrics (Revenue, TRL).
- **Retry Logic:** A temperature-scaling retry mechanism is applied when malformed JSON is produced, automatically increasing temperature ($T = 0.1 \rightarrow T = 0.4$) to recover valid output.

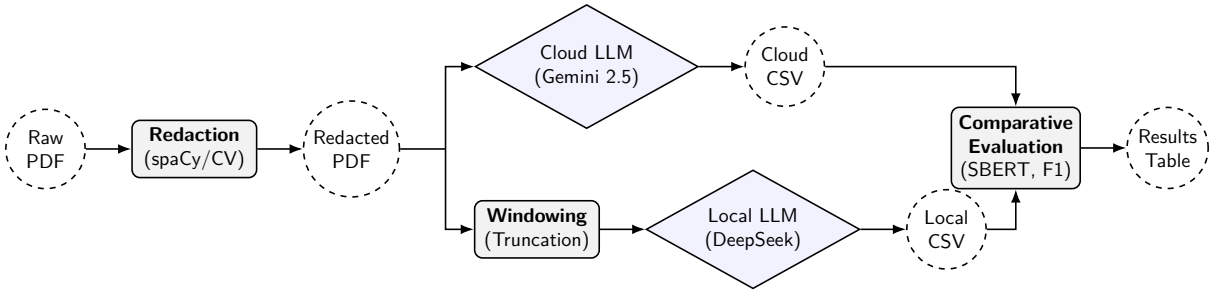

Figure 1. The methodological pipeline for assessment-oriented data extraction. Redacted reports follow two paths: a Cloud Benchmark (Gemini) or a Local Pipeline (DeepSeek) requiring windowing and truncation prior to inference.

Target Variables

To represent the multi-dimensional nature of the projects, we defined a strict schema using *Pydantic* objects to enforce structured JSON output. Variables were grouped as follows:

- **Identity Variables:** *Project Title*, *Company Name*, and *Domain*.
- **Narrative Variables:** *Problem Statement*, *Solution*, and *Value Proposition*.
- **Technical Variables:** *Tech Stack*, *Tools*, and *Data Sources*.
- **Strategic Variables:** *SDG Goals*, *TRL (Technology Readiness Level)*, and *Team Size*.

Evaluation Framework

To compare Local and Cloud extraction performance, three Natural Language Processing (NLP) metrics were employed, selected according to variable type:

1. **Semantic Similarity (SBERT):** Narrative fields (e.g., Problem Description) were evaluated using *all-MiniLM-L6-v2* embeddings and cosine similarity.
2. **F1-Score (Set Overlap):** List-based fields (e.g., Tech Stack, SDGs) were evaluated using F1-score based on entity overlap.
3. **ROUGE-L & BLEU:** Short text fields (e.g., Titles) were evaluated using standard n-gram overlap metrics to measure exact reproduction.

Table 1. Performance of privacy-preserving local extraction against cloud benchmark: variable-level assessment feasibility

Data	Variable	Description	Metric	Match
Project	Title	Official project name.	Semantic	78.2%
	Company	External partner or stakeholder.	Semantic	72.0%
	Domain	Industry sector (e.g., Bio-Tech).	Semantic	35.3%
Narrative	Problem	Summary of the user need or pain point.	Semantic	56.1%
	Solution	Summary of the proposed solution.	Semantic	56.1%
	Value Prop.	Core benefit provided to the user.	Semantic	0.0%
Technical	Tech Stack	Core technologies used (e.g., Python).	F1-Score	17.0%
	Tools	Specific hardware/software (e.g., RPi).	F1-Score	13.6%
	Data Sources	Inputs used (e.g., APIs, datasets).	F1-Score	5.6%
Strategic	SDGs	UN Sustainable Development Goals.	F1-Score	10.0%
	TRL	Technology Readiness Level (1–9).	Numeric	18.3%
	Team Size	Number of students involved.	Numeric	20.0%

METHODOLOGY VALIDATION AND RESULTS FOR ASSESSMENT

To validate the proposed Dual-Pipeline framework, we processed a stratified sample of 63 final project reports from Academic Years 2019–2024. The primary contribution is a comparative performance analysis between the Privacy-Preserving (Local) and High-Capability (Cloud) approaches, assessing the viability of offline AI for educational assessment.

Comparative Performance Analysis

Table 1 reports the quantitative agreement between the Local (DeepSeek-8B) and Cloud (Gemini 2.5) pipelines. The results reveal a distinct *Inference Gap*, characterized by increasing divergence with the abstraction level of the target variable.

Analysis of the Enriched Dataset

Given the higher semantic accuracy of the Cloud pipeline (Table 1), we synthesized extraction outputs to explore preliminary trends that were previously inaccessible in static PDF archives. Treating the Cloud outputs as a benchmark and using local data for corroboration, we examined project evolution over the five-year period.

The analysis suggests a clear technological shift in project focus. In 2019, approximately 33% of projects centered on Hardware/IoT solutions, whereas by 2024, AI-native projects had become dominant, with nearly 45% incorporating Machine Learning components. This marks a departure from earlier emphasis on web and mobile development.

In parallel, semantic extraction of strategic variables indicates a strong alignment with UN Sustainable Development Goals. SDG 9 (Industry, Innovation, and Infrastructure) emerged as the most prevalent focus ($N = 38$), followed by SDG 11 (Sustainable Cities) and SDG 3 (Good Health). This distribution suggests that the course's PDP model naturally steers student projects toward infrastructure and urban health challenges.

DISCUSSION

The quantitative divergence between the two pipelines highlights key boundary conditions for deploying LLM-based extraction in educational data mining.

First, the high agreement scores for *Project Identity* variables (Title: 78.2%, Company: 72.0%) indicate that Local LLMs function effectively as *Librarians*. Even with truncated contexts, these models accurately identify *what* a document is and *who* it relates to. This suggests that institutions can reliably build privacy-preserving, searchable indexes of historical archives using offline models, supporting low-risk assessment tasks such as project classification and curriculum mapping while minimizing exposure.

A distinct Inference Gap emerges when extraction requires semantic abstraction. Although the Local model could summarize explicit sections such as the Problem Statement (56.1%), it struggled to infer the *Value Proposition* (0.0%) or classify the *Domain* (35.3%) unless these elements were explicitly labeled. This indicates that current offline models are effective for Retrieval Augmented Generation (RAG) tasks when answers are directly stated, but lack the *Analyst* capability needed for higher-level semantic categorization.

Low F1-scores for Technical Variables (Tech Stack: 17.0%) further reveal a methodological pitfall we term the *Boilerplate Trap*. A manual audit showed that the Local model, constrained by its 8k token window, often failed to distinguish assignment instructions (e.g., "Students must use Python or Java") from actual implementation choices. In contrast, the Cloud model, leveraging its large context window ($>$1M tokens), consistently differentiated instructional boilerplate from student-generated content.

Taken together, these findings support a tiered implementation strategy. For tasks requiring strict data sovereignty, such as indexing and metadata extraction, Local models provide a viable privacy-first solution for foundational assessment activities. However, for deeper curriculum analysis that depends on disentangling instructional scaffolding from student work, the reasoning depth and context capacity of frontier Cloud models remain necessary. Importantly, the framework does not assume concurrent deployment of both pipelines in practice; rather, it provides empirical evidence to support informed selection of a single pipeline aligned with institutional regulatory, infrastructural, and assessment requirements.

LIMITATIONS AND FUTURE WORK

The primary limitation of the Local pipeline stems from hardware constraints typical of consumer environments. The model's 8k token limit required a strict windowing strategy that concatenated the introduction and conclusion sections, likely excluding technical details embedded in middle appendices and contributing to lower recall for technical variables. In addition, while the automated comparison across 63 reports provides sufficient statistical support for methodological validation, the manual human-in-the-loop verification was limited to targeted spot-checks rather than a comprehensive audit of all extracted data points.

Current work focuses on scaling the analysis from the validation sample to the complete historical archive (200+ projects), enabling robust longitudinal insights to inform course design and preparation. In parallel, we are conducting human validation of the identified project features.

The enriched dataset will also be integrated with institutional academic records, such as grades, to statistically model relationships between extracted project complexity indicators (e.g., TRL level) and academic performance. In parallel, and aligned with emerging cost-efficient AI practices, we plan to explore fine-tuning local LLMs on the structural patterns of PAE reports. This approach has the potential to reduce the observed Inference Gap, allowing local models to achieve higher accuracy while preserving data sovereignty.

CONCLUSION

This paper presents a scalable, privacy-adaptive methodology for educational data mining in engineering education. We demonstrate that unstructured student reports can be transformed into structured curriculum data using Generative AI while explicitly prioritizing data privacy through local execution.

The findings support a tiered implementation strategy. Local models are well suited for privacy-safe project indexing and discovery, whereas redacted cloud models remain necessary for deeper curriculum auditing and high-stakes competency assessment. This framework enables informed, context-sensitive decisions by combining cost-effective local extraction with targeted use of secure cloud capabilities. By adopting this tiered strategy, institutions can unlock insights from digital educational archives while maintaining strong data governance. From a CDIO perspective, the proposed methodology provides a practical decision framework for transparent, repeatable, and evidence-based learning analytics.

ACKNOWLEDGEMENTS

The authors received no financial support for this work.

Declaration of Generative AI Usage

Generative AI was used to refine data extraction code and improve linguistic clarity. All outputs were validated as described in the Methodological Framework. The authors assume full responsibility.

REFERENCES

Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R. J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). Chatgpt: Applications, opportunities, and threats. *arXiv preprint arXiv:2304.09103*.

Berdanier, C. G., Baker, E., Wang, W., & McComb, C. (2018). Opportunities for natural language processing in qualitative engineering education research: Two examples. *2018 IEEE Frontiers in Education Conference (FIE)*, 1–6.

Borrego, M., Foster, M. J., & Froyd, J. E. (2014). Systematic literature reviews in engineering education and other developing interdisciplinary fields. *Journal of Engineering Education*, *103*(1), 45–76.

Bragós Bardia, R., Aoun, L., Charosky Larrieu-Let, G., Bermejo Broto, S., Rey Micolau, F., & Pegueroles Vallés, J. R. (2022). Correlation study between the access mark and the performance in project-based and standard courses. *SEFI 50th Annual Conference of The European Society for Engineering Education*, 151–159.

Cobb, C. L., Hey, J., Agogino, A. M., Beckman, S. L., & Kim, S. (2016). What alumni value from new product development education: A longitudinal study. *Advances in Engineering Education*, *5*(1).

Ding, X., Gopannagari, M., Sun, K., Tao, A., Zhao, D. L., Varadhan, S., Hardy, B. L. B., Dalpiaz, D., Vogiatzis, C., Angrave, L., & Liu, H. (2024). Evaluation of llms and other machine learning methods in the analysis of qualitative survey responses for accessible engineering education research. *2024 ASEE Annual Conference & Exposition*.

Edström, K. (2016). Aims of engineering education research: The role of the cdio initiative. *Proceedings of the 12th International CDIO Conference*, 974–985.

Filippi, S., & Motyl, B. (2024). Large language models (llms) in engineering education: A systematic review and suggestions for practical adoption. *Information*, *15*(6).

Garcia-Huertes, S., & Bragós, R. (2023). Entrepreneurship in engineering programs: A methodology for systematic literature review. *Proceedings of the 19th International CDIO Conference*, 424–434.

Garcia-Huertes, S., Bragós, R., Vidal, E., & Bermejo, S. (2024). Leveraging generative ai for assessing student reports in engineering education. *Proceedings of the 52nd Annual Conference of the European Society for Engineering Education (SEFI)*.

Hingle, A., Katz, A., & Johri, A. (2023). Exploring nlp-based methods for generating engineering ethics assessment qualitative codebooks. *2023 IEEE Frontiers in Education Conference (FIE)*, 1–8.

Kohn Rådberg, K., Lundqvist, U., Malmqvist, J., & Hagvall Svensson, O. (2020). From CDIO to challenge-based learning experiences – expanding student learning as well as societal impact? *European Journal of Engineering Education*, *45*(1), 22–37.

Malmqvist, J., Kohn Rådberg, K., & Lundqvist, U. (2015). Comparative analysis of challenge-based learning experiences. *Proceedings of the 11th International CDIO Conference*, *8*, 87–94.

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, *23*, 1–14.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2024). Can large language models transform computational social science? *Computational Linguistics*, *50*(1), 237–291.

BIOGRAPHICAL INFORMATION

Saul Garcia-Huertes is an assistant professor at Telecos-BCN, Technical University of Catalonia (UPC), where he lectures courses on innovation and entrepreneurship topics. He is currently a PhD candidate in the Education on Engineering, Science and Technology program at UPC, and his research interests include innovation and entrepreneurship education and experiential learning. (ORCID: 0000-0002-6735-3025)

Ramon Bragós, PhD is a full professor at the Electronics Engineering Department of Technical University of Catalonia (UPC). His research focuses on electrical impedance spectroscopy applications in biomedical engineering and in Engineering Education. He is the associate dean of academic innovation at Telecos-BCN, the ICT engineering School of UPC in Barcelona. (ORCID: 0000-0002-1373-1588)

Corresponding author

Saul Garcia-Huertes
Universitat Politècnica de Catalunya (UPC)
Telecos-BCN
C. Jordi Girona 1-3
08034 Barcelona, SPAIN
saul.garcia@upc.edu