

Investigating AI Models for Digital Readiness Assessment in Higher Education Using Unstructured Quality Reports

Zisis Manouskos¹, Nikolaos Avouris¹, Sophia Daskalaki¹, Pedro M. Ruiz², Antonio Ruiz-Martínez², Ramón Sánchez-Iborra²

¹Department of Electrical and Computer Engineering, University of Patras, Greece, ²Department of Information and Communications Engineering, University of Murcia, Spain

How to cite: Manouskos, Z.; Avouris, N.; Daskalaki, S.; Ruiz, P. M.; Ruiz-Martínez, A.; Sanchez-Iborra, R. (2026). *Investigating AI Models for Digital Readiness Assessment in Higher Education Using Unstructured Quality Reports*. In: *12th International Conference on Higher Education Advances (HEAd'26)*. Valencia, 15-18 June 2026. <https://doi.org/10.4995/HEAd26.2026.21813>

Abstract

Higher Education Institutions increasingly use digital readiness assessments to plan and monitor their digital transformation. The DigiReady framework aims at assessing digital readiness across seven dimensions, which are measured through quantitative and qualitative (binary, Likert scale based) indicators based of collected evidence. This evidence is often embedded in unstructured and semi-structured institutional documents from which needs to be extracted. The institution's quality assurance documents, like assessment reports, accreditation reports and quality manuals that include descriptions of quality metrics and indicators, contain valuable information relating to its current state. This paper investigates whether large language models (LLMs) can effectively support this process of evidence extraction. We conducted two preliminary case studies in order to answer this question. The first examines evidence extraction and digital readiness dimension assignment from a single department annual report. The second extends the analysis to multiple institutional documents that contain descriptions of QA metrics and indicators from two European universities. Results show that LLMs can assist with the initial screening of documents by identifying potentially relevant excerpts and metrics, but performance remains moderate and sensitive to document structure, prompt design, and task complexity. Across both studies, fully autonomous identification of related evidence and assessment proved unreliable. The findings therefore support a hybrid human-AI workflow in which LLMs enhance scalability and prioritization, while expert review remains necessary for validation and final assessment.

Keywords: Digital Readiness; Higher Education Institution; Large Language Models, LLM-as-a-judge; Zero-shot Text Classification.

1. Introduction

Higher Education Institutions (HEIs) are under growing pressure to implement digital transformations in teaching, administration, and research. Effective action requires an evidence-based view of institutional digital readiness; otherwise, initiatives often remain fragmented or unsustainable (Alenezi, 2021; Chounta et al., 2024). The DigiReady (DR) framework addresses this need by defining digital readiness across seven dimensions, each one measured through indicators using qualitative and quantitative evidence. A computational model has been implemented, UDReady, to be used as a self-assessment platform by interested institutions (Chounta et al., 2024; Manouskos et al., 2025). The collection of the necessary evidence to measure these indicators is a tedious process, especially when this evidence is embedded in semi-structured and unstructured sources, like reports. This is especially challenging for dimensions such as Leadership and Governance (D1) and Strategy and Policies (D2), where relevant information is dispersed across institutional documents (e.g., strategic plans, QA policies, accreditation documents) (Manouskos et al., 2025). Similar difficulties arise for quantitative indicators: stakeholders must still locate and interpret institution-specific metrics, often defined in semi-structured QA documentation, and map them to DR indicators manually.

To address this gap, we aim to develop an intelligent knowledge base integrated into UDReady to support evidence identification for DR assessment. Building on previous work suggesting that LLMs can assist in document-centric tasks but should not fully automate interpretive judgments, we adopt a hybrid human–AI approach (Shrestha et al., 2019; Bharathi Mohan et al., 2024; Eigner & Händler, 2024). We investigate whether LLMs can help stakeholders identify relevant qualitative excerpts and extract institution-specific metrics that map to DR dimensions and indicators.

We evaluate this through two preliminary case studies: (1) identification of qualitative evidence in an annual report of a typical academic department, including an LLM-as-a-judge decision-support layer (Gu et al., 2025; Wei et al., 2025); and (2) attempt to extract relevant quantitative metrics from QA reports of two European institutions, compared against expert ground truth. Section 2 presents the knowledge base architecture, Section 3 the methodology and case studies, Section 4 the results, and Section 5 the discussion and future work.

2. Architectural Overview of the Knowledge Base

The proposed Knowledge Base (KB) supports institutional stakeholders in assessing digital readiness from semi-structured and unstructured sources. It operationalizes the DigiReady framework in UDReady by linking DR dimensions and topics to qualitative and quantitative evidence extracted from institutional documents. To support transparency and auditability, the KB stores extracted evidence together with provenance metadata and source context.

The KB has two main functions: evidence collection and assessment support. Stakeholders upload relevant documents, such as strategic plans, QA manuals, and reports, and the KB returns topic-aligned candidate excerpts and metrics. After stakeholder review, approved evidence is stored and reused in UDReady assessments.

The system follows an evidence pipeline of collection, processing, storage, and retrieval. This paper focuses on the processing stage, where artifacts such as PDFs, presentations, and spreadsheets are mapped to DigiReady dimensions. We prioritize high-context LLMs to generate prompt-based “evidence bundles” from long documents while preserving context (Ke et al., 2025), with human oversight to ensure accountability (Raji et al., 2020).

3. Methodology

This study includes two preliminary case studies. Case Study 1 examines whether an LLM can extract verbatim evidence from a QA report and map it to DigiReady dimensions, with a second LLM supporting evaluation. Case Study 2 analyzes QA documents from two HEIs to identify institution-specific metrics and map them to DigiReady dimensions.

3.1 Case Study 1: Evidence extraction and DigiReady mapping

The first case study used a QA report of an academic department of the University of Patras (UPAT) to examine two tasks: (a) extracting verbatim DigiReady-relevant evidence and assigning it to predefined DigiReady dimensions and topics, and (b) evaluating these assignments through an LLM-as-a-judge workflow.

For extraction and classification, ChatGPT 5.2 was used through the OpenAI chatbot interface in January 2026. It was prompted in a structured zero-shot setting to review the full report, extract only clearly DigiReady-relevant text verbatim, preserve contextual completeness, apply strict relevance filtering, and assign each segment only to predefined DigiReady dimensions and topics. Results were returned in a fixed table with the exact text, page number, dimension, and topic.

We compared two processing methods for the 116-page report: full-document ingestion and batched input in three parts (pages 1–50, 51–80, and 81–116). A human evaluator then rated how well the extracted segments matched the assigned DigiReady dimensions on a 3-point scale: 0 (No Match), 1 (Partial Match), and 2 (Full Match). To assess classification performance while preserving quality differences, we calculated weighted precision using the following formula:

$$Precision = \frac{0 \cdot No\ Match + 1 \cdot Partial + 2 \cdot Full}{3 \cdot Total}$$

This approach assigns proportional credit to partial matches while penalizing weak classifications, avoiding the misleading outcomes of binary precision metrics where all non-zero matches would be treated equally. A manual ground truth was not established for this preliminary phase due to the high resource requirements for achieving objective inter-rater reliability. Furthermore, this approach supports our goal of building an **LLM-as-a-judge** framework, designed to automatically evaluate and filter extracted content for dimensional relevance.

The same outputs were also evaluated in an LLM-as-a-judge setting, where a second LLM rated the assigned mappings on the same 0–2 scale. Two judge models, ChatGPT 5.2 and Gemini 3 Pro, were tested with two iterative prompts. **Prompt 1a** asked the model to rate alignment between each segment and its assigned DigiReady dimension and topic as 0, 1, or 2. **Prompt 1b** refined this by adding explicit scoring criteria and two annotated examples. Agreement with human ratings was measured using mean absolute score difference, with lower values indicating closer alignment.

3.2 Case Study 2: Multi-institution extraction and DigiReady mapping of QA metrics

The second case study focused on quantitative QA metrics and indicators described in institutional QA documents, like manuals. Although HEIs already defines and monitor various quality assurance metrics and indicators through their internal QA processes, identifying which are relevant to digital readiness and mapping them to the DigiReady framework remains a time-consuming manual task, especially when evidence is distributed across semi-structured and unstructured documents.

We used two models, ChatGPT 5.4 in extended thinking mode and Gemini 3 Pro, on QA document collections from two European HEIs: the University of Patras (UPAT) and the University of Murcia (UDM). Both models were accessed in April 2026. The corpus comprised five semi-structured and unstructured documents from each institution. For both institutions, expert ground truth was constructed collaboratively by institutional stakeholders and DigiReady experts, yielding 71 metrics and indicators for UPAT and 74 for UDM.

Like in case study 1, the prompt followed a structured zero-shot design, asking a model to scan a document, identify metrics relevant to digital readiness based on the descriptions of DR dimensions, and assign each to the most appropriate DR dimension in a fixed tabular output. Documents longer than 100 pages were split into 50-page batches. This applies to one document in the UPAT corpus and two in the UDM corpus. Each document or batch was processed in a separate session using the same prompt.

Model outputs were compared against expert-constructed ground truth at two levels: metric extraction, i.e., whether the model identified a QA metric relevant to the DigiReady framework,

and metric mapping, i.e., whether the extracted metric was assigned to the correct DigiReady dimension. Performance at both levels was measured using F1-score, derived from precision and recall. In addition, we assessed the accuracy of mapping for the positive extracted cases, the True Positive (TP) mapping.

4. Evaluation

4.1 Case Study 1: Evidence extraction, DigiReady classification, and LLM-as-a-judge

Case Study 1 examined whether an LLM could extract and classify digital readiness evidence from a 116-page QA report using different ingestion strategies. As described in Section 3.1, ChatGPT 5.2 was used for the extraction task. The model was prompted to identify relevant verbatim segments, assign each to the appropriate DigiReady dimension and topic, exclude irrelevant content, and provide page references. We compared two ingestion strategies: whole-document processing and page-based batching (pages 1-50, 51-80, and 81-116). Outputs were evaluated using the 3-point relevance scale described in Section 3 and weighted precision.

In the whole-document condition, ChatGPT 5.2 extracted 6 segments, all rated as Partial Match, for a weighted precision of 33%. In the batched condition, it extracted 37 segments: 13 No Match, 13 Partial Match, and 11 Full Match. Although batching increased irrelevant or weak outputs, it also produced substantially more high-quality evidence, with 11 fully aligned segments compared with none in the whole-document condition. Weighted precision was similar at 31%.

We also tested whether a second LLM could reduce manual review by reliably rating extracted mappings. ChatGPT 5.2 and Gemini 3 Pro re-scored the segments using the same 0–2 relevance scale, and agreement with human ratings was measured by mean absolute score difference. With the initial prompt, ChatGPT 5.2 aligned much more closely with human ratings than Gemini 3 Pro ($M = 0.50$, $SD = 0.84$ vs. $M = 2.33$, $SD = 0.52$). After refining the prompt with explicit criteria and examples, both models improved, and the gap narrowed (ChatGPT 5.2: $M = 0.42$, $SD = 0.71$; Gemini 3 Pro: $M = 0.65$, $SD = 0.58$).

Overall, the extraction model showed low precision, likely due to limited context, batch design, and prompt structure. Prompt refinement improved LLM-as-a-judge performance, but model variability remained. Due to the low precision of information identification and classification, the hybrid workflow is needed, with LLMs supporting initial screening and humans validating final decisions.

4.2 Results of Case Study 2: Multi-institution extraction and DigiReady mapping of QA metrics

Case Study 2 evaluated how well two LLMs extracted quantitative QA metrics relevant to digital readiness and mapped them to the appropriate DigiReady dimensions, using expert-constructed ground truth from the University of Patras (UPAT) and the Universidad de Murcia (UDM). As described in Section 3.2, performance was assessed at two levels: metric extraction, i.e., whether a model identified a ground-truth metric as relevant to digital readiness, and metric mapping, i.e., whether the extracted metric was assigned to the correct DigiReady dimension. Table 1 summarizes the main results.

Table 1. F1 scores and True Positive mapping accuracy in Case Study 2

Institution	Model	F1 (extraction)	F1 (mapping)	TP (mapping)
UPAT	ChatGPT 5.4 Thinking	0.368	0.241	65.6%
UPAT	Gemini 3 Pro	0.420	0.254	60.5%
UDM	ChatGPT 5.4 Thinking	0.462	0.321	69.4%
UDM	Gemini 3 Pro	0.295	0.188	63.6%

For UPAT, the ground-truth set contained 71 metrics and indicators. Gemini 3 Pro outperformed ChatGPT 5.4 Thinking in both extraction ($F1 = 0.420$ vs. 0.368) and mapping ($F1 = 0.254$ vs. 0.241). However, ChatGPT achieved slightly higher mapping accuracy among correctly extracted metrics (65.6% vs. 60.5%), suggesting that Gemini's advantage was driven primarily by stronger extraction recall.

For UDM, where the ground-truth set contained 74 metrics and indicators, the pattern was reversed. ChatGPT 5.4 Thinking outperformed Gemini 3 Pro in both extractions ($F1 = 0.462$ vs. 0.295) and mapping ($F1 = 0.321$ vs. 0.188) and achieved higher mapping accuracy (69.4% vs. 63.6%).

Across both institutions, mapping performance was consistently lower than extraction performance, indicating that identifying potentially relevant QA metrics is easier than assigning them reliably to the correct DigiReady dimension. The results also show clear cross-institution variation: Gemini performed better on the UPAT corpus, whereas ChatGPT performed better on the UDM corpus. This suggests that model effectiveness depends not only on model capability, but also on document structure, terminology, and local QA reporting practices. A likely contributing factor is the different nature of the two corpora: the UPAT documents were more semi-structured, consisting largely of tables of metrics and indicators, whereas the UDM documents were more unstructured, with narrative text interspersed with tables describing metrics and indicators.

Case Study 2 shows that LLMs are not yet reliable for identifying quantitative evidence in digital readiness assessments without human oversight. Performance was modest across all settings, with F1-scores below 0.5 and mapping accuracy around 60–70%, indicating frequent errors in both extraction and framework mapping. The high number of false positives at both institutions also underscores the need for expert validation, especially when metrics are only indirectly related to digital readiness.

5. Discussion, Limitations and Future Work

The two case studies show that LLM support for DigiReady evidence processing is feasible, but strongly shaped by document structure, task design, and evaluation setup.

In Case Study 1, page-based batching outperformed whole-document prompting, indicating that segmentation matters for long QA reports. The results also showed that LLM-as-a-judge can support validation, but only with explicit prompts; prompt design and judge-model choice strongly affect agreement with human ratings, making these models better suited for triage than automatic acceptance.

Case Study 2 extends these findings to multiple QA artifacts from two institutions. Across both corpora, mapping consistently underperformed extraction, showing that identifying relevant QA metrics is easier than assigning them to the correct DigiReady dimension. Performance remained modest, with all F1-scores below 0.5 and mapping accuracy around 60-70%, indicating that current LLMs are not reliable enough for unsupervised use. The relatively high number of false positives further highlights the need for expert validation. Results also varied by institution: Gemini performed better on the more semi-structured UPAT corpus, while ChatGPT performed better on the more unstructured UDM corpus, suggesting sensitivity to document organization and terminology.

Overall, LLMs can support stakeholders in identifying qualitative and quantitative evidence for digital readiness assessment, but their role is supportive rather than semi-automated or fully automated. In a knowledge-base workflow, LLMs could generate preliminary evidence extraction and mapping, which stakeholders would then validate against the original documents before inclusion in the knowledge base and use through UDReady for digital readiness assessment.

This work has important limitations: only two universities and a small document set were studied; qualitative judgments came from a single annotator; and only one prompt per model was tested, so results are preliminary. Future work will expand the corpus, include multi-rate expert annotation, compare LLM-based extraction with retrieval- and OCR-supported pipelines, and integrate the resulting components into UDReady.

Acknowledgments

The work presented here has been funded by Erasmus+ Programme of the European Union, and coordinated by IKY Hellenic Scholarships Foundation, Project number: 2025-1EL01-KA220HED000358593, Project partners are: GUNet Greece, University of Patras, Aristotle University of Thessaloniki, University of Macedonia, Arts et Métiers Institute of Technology, University of Murcia and Cyprus University of Technology

References

- Alenezi, M. (2021). Deep dive into digital transformation in higher education institutions. *Education Sciences*, 11(12), Article 770. <https://doi.org/10.3390/educsci11120770>
- Bharathi Mohan, G., Prasanna Kumar, R., Vishal Krishh, P., Keerthinathan, A., Lavanya, G., Meghana, M. K. U., Sulthana, S., & Doss, S. (2024). An analysis of large language models: Their impact and potential applications. *Knowledge and Information Systems*, 66(9), 5047–5070. <https://doi.org/10.1007/s10115-024-02120-8>
- Chounta, I.-A., Ortega-Arranz, A., Daskalaki, S., Dimitriadis, Y., & Avouris, N. (2024). Toward a data-informed framework for the assessment of digital readiness of higher education institutions. *International Journal of Educational Technology in Higher Education*, 21(1), Article 59. <https://doi.org/10.1186/s41239-024-00491-0>
- Eigner, E., & Händler, T. (2024). *Determinants of LLM-assisted decision-making* (arXiv:2402.17385). arXiv. <https://doi.org/10.48550/arXiv.2402.17385>
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., Wang, Y., Gao, W., Ni, L., & Guo, J. (2024). *A survey on LLM-as-a-judge* (arXiv:2411.15594). arXiv. <https://doi.org/10.48550/arXiv.2411.15594>
- Ke, W., Zheng, Y., Li, Y., Xu, H., Nie, D., Wang, P., & He, Y. (2025). Large language models in document intelligence: A comprehensive survey, recent advances, challenges, and future trends. *ACM Transactions on Information Systems*, 44(1), Article 18. <https://doi.org/10.1145/3768156>
- Manouskos, Z., Chounta, I.-A., Tsimpanis, K., Ortega-Arranz, A., Daskalaki, S., Dimitriadis, Y., & Avouris, N. (2025). Evaluation of UDReady: A tool for measuring digital readiness in higher education. In *Proceedings of the EUNIS 2025 annual congress in Belfast (EPIc Series in Computing*, Vol. 107, pp. 196–205). <https://doi.org/10.29007/grb9>
- Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33–44). Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372873>
- Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83. <https://doi.org/10.1177/0008125619862257>

Wei, H., He, S., Xia, T., Liu, F., Wong, A., Lin, J., & Han, M. (2025). Systematic Evaluation of LLM-as-a-Judge in LLM Alignment Tasks: Explainable Metrics and Diverse Prompt Templates (arXiv:2408.13006). arXiv. <https://doi.org/10.48550/arXiv.2408.13006>

Article in Press