EPiC Series in Computing

Volume 109, 2026, Pages 12–23

Proceedings of EUNIS 2026 Annual Congress

Digital Readiness in Higher Education: A Multi-Institutional Evidence Study and AI-Assisted Approach

Zisis Manouskos[1], Nikolaos Avouris[1], Sophia Daskalaki[1], Christos Rodosthenous[2], Vasos Papageorgiou[2], Dimitris Daskopoulos[3], Eftychios Protopapadakis[4], Nikolaos Samaras[4], Pantelis Balaouras[5,6], Konstantinos Tsimpanis[5,6]

[1] University of Patras, Department of Electrical and Computer Engineering, Patras, Greece
[2] Cyprus University of Technology, Cyprus
[3]Aristotle University of Thessaloniki, Greece.
[4] University of Macedonia, Greece
[5] Greek Universities Network, Athens, Greece
[6] National and Kapodistrian University of Athens, Greece

zisis.manouskos@ac.upatras.gr, avouris@upatras.gr, sdask@upatras.gr, c.rodosthenous@cut.ac.cy, vasos.papageorgiou@cut.ac.cy, dimitris@it.auth.gr, efprot@uom.edu.gr, samaras@uom.edu.gr, balaoura@di.uoa.gr, ktsibanis@uoa.gr

Abstract

Higher Education Institutions (HEIs) need to assess their digital readiness in order to embark on an informed digital transformation, following an evidence-based decision-making approach. This study investigates the feasibility of assessing digital readiness in Higher Education Institutions (HEIs) by analyzing whether institutional documents provide evidence aligned with the DigiReady (DR) framework. We examined 75 documents from four Greek and one Cypriot HEI, finding that evidence was concentrated in governance-related dimensions, Digital Leadership (48%), Strategy (46.7%), and Networks and Collaboration (38.7%), while operational and teaching-related areas (D3-D6) were less represented. Documents were classified into strategic/governance, implementation/operational, and evaluation/outcome-oriented types, informing systematic evidence collection and the creation of an intelligent knowledge base. To explore AI-assisted analysis, nine open-source Large Language Models (LLMs) were evaluated on paragraph-level classification. Mid-sized, instruction-tuned models (3B-7B parameters) achieved the best balance of accuracy and

L. Desnos, C. Diaz, J. Mincer-Daszkiewicz, L. Merakos, R. Vogl, S. McLellan and U. Lucke (eds.), EUNIS 2026 (EPiC Series in Computing, vol. 109), pp. 12–23

efficiency, outperforming larger models, highlighting the importance of model design and tuning. Despite limitations, including regional focus, a single benchmark document, overlapping dimensions, and dataset imbalance, AI-assisted methods show strong potential to convert fragmented institutional evidence into actionable insights for scalable, reproducible digital readiness assessment.

1 Introduction

Higher Education Institutions (HEIs) often launch digital transformation initiatives without first assessing their digital readiness, limiting evidence-based decision-making. Effective transformation requires not only technology investment but also a clear understanding of existing infrastructure, policies, teaching practices, and organizational processes.

However, institutional decision-making often relies on fragmented, narrative documentation rather than systematically analyzed evidence. Strategic plans, quality assurance reports, accreditation reviews, and policy documents constitute **authoritative sources** of institutional knowledge. Yet these unstructured data sources remain underutilized for evidence-based governance. This study argues that systematic analysis of such authoritative unstructured data is essential for enabling reproducible and data-informed digital transformation in higher education.

The DigiReady (DR) Framework addresses this need by providing a structured, data-informed assessment across seven dimensions: Digital Leadership and Governance (D1), Digital Strategy and Policies (D2), Teaching and Learning (D3), Content and Curricula (D4), Training and Support (D5), Infrastructure (D6), and Networks and Collaboration (D7) (Chounta et al., 2024; Tsimpanis et al., 2025). Each dimension includes specific topics and indicators derived from stakeholder workshops, classified as qualitative (e.g., Likert-scale assessments) or quantitative (e.g., data from institutional systems and learning management platforms) (Manouskos et al., 2025).

Assessing qualitative indicators can be subjective, while quantitative data may be difficult to obtain, often requiring analysis of unstructured documents such as strategic plans, policy reports, or online resources. To address this, we aim to develop an intelligent knowledge base that extracts and structures both qualitative and quantitative evidence for each DR dimension.

This paper addresses two research questions: (RQ1) What types of institutional evidence are available and how are they distributed across HEIs? and (RQ2) How accurately can compact language models identify and classify this evidence? Together these questions support both mapping institutional documentation and evaluating AI-based methods for scalable digital readiness assessment.

The remainder of the paper is organized as follows: Section 2 reviews related literature, Section 3 describes the evidence mapping protocol, Section 4 presents cross-institutional results, Section 5 details the language model evaluation methodology, Section 6 reports classification results, Section 7 discusses findings, limitations and future work.

2 Background and Related Work

Assessing digital readiness in Higher Education Institutions (HEIs) requires both understanding existing infrastructure and leveraging institutional knowledge effectively. HEIs generate substantial documentation, strategic plans, policy reports, and online resources that can inform digital transformation initiatives. However, the complexity and volume of such evidence make relying solely on leadership intuition insufficient (Funda, 2024).

Intelligent Decision Support Systems (IDSS) help administrators consolidate large datasets and apply predictive analytics to support strategic planning, scenario analysis, and institutional growth monitoring (Funda, 2024). Complementarily, Knowledge Management (KM) enables the creation, storage, and application of institutional knowledge, facilitating digital transformation by shifting from document-centric to researcher-centric knowledge bases enriched with AI techniques (Khilji et al.,

2024; Koperwas et al., 2017). Together, IDSS and KM provide a framework for data-driven, knowledge-informed decision-making in HEIs.

Recent advances in Large Language Models (LLMs) further enhance institutional intelligence by automating the extraction and classification of unstructured textual data. LLMs outperform traditional machine learning models in complex classification tasks by capturing contextual and semantic relationships (Alarfaj et al., 2026; Kostina et al., 2025). Their effectiveness depends on model size, prompt engineering, and reasoning strategies, with instruction-tuned medium-sized models balancing accuracy, efficiency, and stability (Niimi, 2025; Silveira et al., 2026). In the context of HEIs, LLMs can support scalable, consistent mapping of institutional authoritative documents to structured frameworks, such as DigiReady, enabling evidence-informed digital readiness assessment.

3 Multi-Institutional Evidence Mapping Methodology

HEIs generate unstructured documents that can inform multiple DigiReady (DR) dimensions. Assessing qualitative indicators and mapping evidence can be supported by intelligent systems. To this end, we are developing an intelligent knowledge base using institutional documents, strategic plans, policies, reports, and online resources.

An ongoing study is being conducted across five European Higher Education Institutions (HEIs): the University of Patras (UPAT), Aristotle University of Thessaloniki (AUTH), the University of Macedonia (UoM), the National and Kapodistrian University of Athens (NKUA), and the Cyprus University of Technology (CUT). Stakeholders from each participating institution applied a standardized protocol to identify and map **authoritative** evidence to DR dimensions. Institutional repositories, websites, quality assurance platforms, and public reports were systematically reviewed. Key artifacts were catalogued with metadata (title, type, unit, date, source) and mapped to one or more DR dimensions using consistent guidelines to ensure reliability.

The resulting evidence logs were consolidated into a shared repository, enabling cross-institutional quantitative analysis and supporting classification methods. This dataset underpins the comparative results presented in Section 4.

4 Multi-Institutional Evidence Mapping Results

In this section, we present the results of a multi-institutional evidence mapping study conducted to assess the availability and distribution of digital readiness evidence across five European Higher Education Institutions.

It should be noted that this is an ongoing study. To date, a total of **75** artifacts of evidence have been collected across the five participating institutions and classified into **15** categories. The distribution of evidence is uneven: UPAT contributed 28 artifacts, CUT 13, AUTH 11, UoM 14, and NKUA 9. Given that most participating institutions are based in Greece and operate primarily in Greek, subsequent analyses treat the dataset collectively.

The most frequently observed types of evidence were Analytical Reports (14 artifacts), Online Resources and Strategic Plans (9 artifacts each), Institutional Documents and Quality Assurance Documents (8 artifacts each). Policy/Governance Documents (7 artifacts), Certification Reports, Independent Evaluation Reports, Evaluation Reports were each represented by 4 artifacts, whereas Service Documentation, Statistical Datasets, Memoranda of Understanding, Curriculum Documents, Guidelines, and Meeting Minutes were less frequently reported. Institutional contributions varied, with UPAT providing evidence across ten categories, CUT in nine, AUTH in six, UoM in six, and NKUA in 5 categories, reflecting both the diversity and uneven distribution of documentation.

Each evidence artifact was categorized into one or more dimensions. As illustrated in Figure 1(a), dimensions D1 (Leadership and Governance), D2 (Strategy and Policies), and D7 (Networks and Collaboration) had the highest coverage, with 48%, 46.67%, and 38.67% of artifacts, respectively.

Dimension D3 (Teaching and Learning) followed with 34.67% coverage, while the remaining dimensions were represented below 32%.

Institution-specific coverage, shown in the heatmap of Figure 1(b), indicates that UPAT contributed most extensively to D1 and D2. NKUA contributed most in D1, D2 and D7, whereas coverage across the other institutions was relatively uniform.

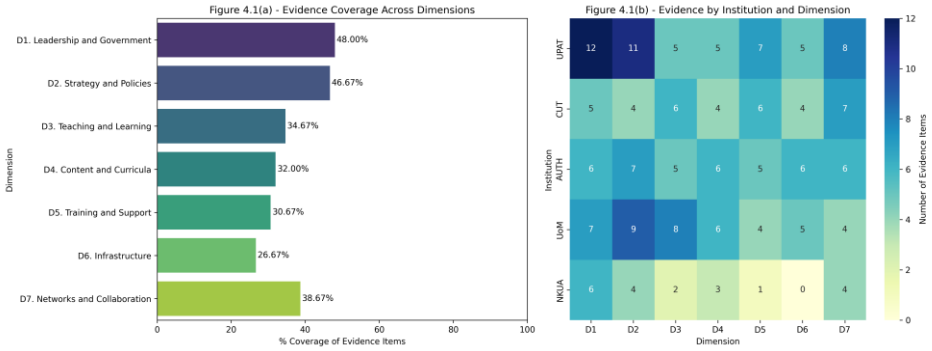

Figure 1 (a) Percentage coverage of evidence across dimensions. (b) Heatmap of evidence distribution by institution and dimension.

This preliminary analysis highlights which DigiReady dimensions are most readily assessable in Greek and Cypriot institutions participating in the study. Evidence is concentrated in D1 (Leadership and Governance), D2 (Strategy and Policies), and D7 (Networks and Collaboration), reflecting the larger number of qualitative indicators in these dimensions. Overall, authoritative unstructured materials, such as strategic plans, policies, and other institutional documents, exist across multiple dimensions and can serve as structured evidence sources for the development of institutional knowledge bases supporting digital readiness assessment.

Although most participating institutions are Greek, all belong to the European University Association (EUA) and follow the European Standards and Guidelines (ESG), indicating similar documentation exists across European HEIs. The 15 empirical evidence types can be grouped into three abstract types: **strategic and governance**, **implementation and operational**, and **evaluation and outcome oriented**. This typology aligns with quality culture research (Bendermacher et al., 2017; Uslu, 2025), reflecting institutional progression from strategy through implementation to continuous improvement. Empirically, strategic/governance documents dominate D1, D2, and D7; implementation/operational documents focus on D3-D6; and evaluation/outcome-oriented documents span multiple dimensions, supporting the development of an intelligent knowledge base for cross-institutional digital readiness assessment.

Manual extraction from multipage documents is time-consuming, underscoring the need for advanced NLP tools. Sections 5 and 6 present our methodology for human classification of content into DigiReady dimensions, along with the evaluation of LLMs for automatically identifying and categorizing relevant information.

5 Methodology for Evaluating Large Language Models for Institutional Text Classification

Higher education institutions generate evidence materials primarily in the form of document-based reports (e.g., PDF documents) that contain information relevant to assessing institutional digital readiness through the DigiReady (DR) dimensions (see Section 4). However, only specific parts of these documents, typically individual paragraphs or sentences, are relevant to dimensions. Manually

identifying such evidence requires stakeholders to read entire documents, locate relevant text segments, and map them to DR dimensions, resulting in a time-consuming process.

Recent advances in NLP, particularly LLMs, have demonstrated strong capabilities in text classification tasks (Kostina et al., 2025). These models can perform zero-shot classification and identify relevant content without task-specific training. Instruction-tuned models in the lower parameter range (≈3B–8B) offer a balance of accuracy and computational efficiency, with recent research demonstrating that ensemble strategies can further improve the stability of models in this class to match larger counterparts (Niimi, 2025). Automated prompt optimization has been shown to outperform manually designed instructions in domain-specific contexts (Silveira et al., 2026).

The objective of this study is to conduct a feasibility assessment aimed at informing the development of a protocol for constructing a DigiReady (DR) benchmark dataset. Specifically, the study evaluates how effectively open-source language models, primarily smaller-parameter models, can classify document paragraphs according to DR dimensions. The goal is not to provide a general evaluation of LLMs, but to investigate whether their categorization capabilities can be leveraged for this task. To this end, paragraphs were extracted from an external evaluation report of a Greek higher education institution and manually classified by two experts. Their annotations served as the ground truth for evaluating the performance of the language models.

This section presents the study methodology, including the development of the ground truth dataset, the experimental setup (model selection, inference configuration, and prompt design), and the evaluation of metrics used to assess model performance.

5.1 Development of a Paragraph-Level Benchmark Dataset

This study explored external evaluation reports conducted by the Hellenic Authority for Higher Education (HAHE)[1]. These reports provide structured institutional assessments in English and contain valuable information regarding institutional processes and practices. The 2015 external evaluation report of Aristotle University of Thessaloniki (AUTH) was selected as an initial benchmark source because of its comprehensive institutional coverage and relatively advanced digital maturity profile. The purpose of this benchmark was exploratory: to evaluate whether paragraph-level DigiReady classification is sufficiently learnable to justify future large-scale annotation efforts across institutional document collections.

The document comprises 28 pages of narrative text without tables or figures. A total of 118 paragraphs were manually extracted and organized into a structured dataset for systematic analysis. Each paragraph was evaluated by two experts using the DigiReady (DR) framework and classified as either “Relevant” to one or more dimensions (D1–D7) or “DR Not Relevant” for paragraphs that were not relevant to any dimension.

Inter-evaluator agreement was assessed to ensure dataset reliability. At the binary level (DR relevance vs. non-relevance), the evaluators agreed on 101 out of 118 paragraphs (86% concordance), indicating strong consistency. Evaluator 1 identified 28 relevant paragraphs, while Evaluator 2 identified 30, with agreement on 21 cases (70% positive agreement), indicating that the dataset provides a reliable foundation for binary relevance classification.

At the dimension level, agreement remained high, with concordance rates ranging from 87.3% (D3 – Teaching and Learning) to 96.6% (D5 – Training and Support), as shown in Table 1. Lower agreement in certain dimensions, particularly between D3 and D4 (Content and Curricula), reflects their conceptual overlap and interpretative complexity. Positive agreement at the dimension level was lower overall, peaking at 40% for D6 (Infrastructure), which is expected given the interrelated nature of DR dimensions (e.g., D1 with D2, and D3 with D4).

[1] https://www.ethaae.gr/en/quality-assurance/external-evaluation-reports-of-institutions

The final curated dataset includes 118 paragraphs, of which 21 are labeled as DR-related. For multi-class classification, confidence scores were computed by averaging evaluator judgments: 0 (both irrelevant), 0.5 (partial agreement), and 1 (full agreement). Across all paragraph–dimension annotations, the majority of labels (751) received a confidence score of 0, followed by 64 labels with a score of 0.5 and 11 with full agreement (1), highlighting the sparsity and complexity of dimension-level classification.

Dimension	Total Agreement	Concordance Rate (%)
D1 – Leadership and Governance	109	92.4
D2 – Strategy and Policies	111	94.1
D3 – Teaching and Learning	103	87.3
D4 – Content and Curricula	106	89.8
D5 – Training and Support	114	96.6
D6 – Infrastructure	110	93.2
D7 – Networks and Collaboration	109	92.4

Table 1: Inter-Evaluator Agreement and Concordance Rate per Dimension

5.2 Experimental Setup & Evaluation Metrics

After constructing the dataset described in the previous sections, we conducted experiments to evaluate whether open-source LLMs can (1) determine the relevance of a paragraph to the proposed framework and (2) identify the corresponding dimension(s).

We focused exclusively on open-source models due to institutional constraints on the use of proprietary large-scale cloud-based language models, particularly concerning data sovereignty, privacy, regulatory compliance, and the long-term operational costs associated with commercial AI services. This study therefore examines whether smaller, openly available models can provide a cost-effective and practically viable alternative.

5.3 Model Selection

The exclusive use of open-source LLMs was not only motivated by cost considerations but by institutional **data sovereignty requirements**. Higher education institutions manage sensitive internal documentation, including governance records, evaluation reports, and strategic plans. Processing such documents through proprietary cloud-based AI services may conflict with institutional data governance policies and European regulatory frameworks. Therefore, the adoption of deployable open-source models ensures that evidence analysis can be conducted locally, preserving data ownership, compliance, and institutional autonomy.

Model selection and performance baselines were obtained from the Open LLM Leaderboard, hosted on Hugging Face. The leaderboard provides standardized evaluation of LLMs across multiple normalized benchmarks (Fourrier et al., 2024). The leaderboard compares models across multiple benchmarks, and we focused on those achieving high performance on Instruction-Following Evaluation (IFEval), which measures accuracy in following explicit instructions and producing structured outputs; BIG-Bench Hard (BBH), which evaluates advanced logical, mathematical, common-sense, and world-knowledge reasoning through multiple-choice accuracy; and MMLU-Pro, which assesses professional-level domain knowledge and reasoning robustness across disciplines using multiple-choice accuracy.

Models were selected to represent three deployment tiers: **edge devices** (≤2B parameters), **consumer-level GPUs** (~3B parameters), and **mid-range models** (7B–32B parameters). Selection was restricted to models officially evaluated on the Open LLM Leaderboard to ensure standardized and comparable performance metrics across benchmarks.

The edge device tier includes *Granite-3.1-2b-instruct* (AVG 21.71), *Qwen2.5-1.5B-Instruct* (AVG 18.43), and *Qwen2.5:0.5b* (AVG 10.11). These models represent minimal resource deployment scenarios, where computational constraints are strict, and efficiency is prioritized over advanced reasoning performance.

The consumer-level tier (~3B parameters) comprises *Phi-4-mini-instruct* (AVG 29.41), *Falcon3-3B-Instruct* (AVG 26.60), and *Qwen2.5-3B-Instruct* (AVG 27.16). These models are suitable for deployment on consumer-grade GPUs, supporting single-instance inference while offering moderate reasoning capabilities.

The mid-range tier (7B–32B parameters) includes Internlm2.5:7b-chat (AVG 32.97), *Qwen2.5-14B-Instruct* (AVG 41.31) and *Qwen2.5-32B-Instruct* (AVG 46.60). These models provide a balance between computational cost and advanced reasoning capacity, demonstrating substantially stronger performance across instruction-following and knowledge-intensive benchmarks.

5.4 Implementation Environment, Classification Framework Design and Evaluation Metrics

All experiments were conducted in a controlled environment using **Google Colab** for execution, with the **Ollama local inference server** for model deployment and **DSPy** for structured prompting and pipeline orchestration (Khattab et al., 2023a, 2023b). This setup ensured reproducibility and consistent evaluation across all models.

Hardware allocation was tier-specific: edge models were executed on **NVIDIA T4 GPUs** to simulate constrained-resource environments, while consumer-level and mid-range models ran on **NVIDIA A100 GPUs** to accommodate larger parameter sizes and higher memory requirements.

The LLM based classification pipeline consisted of two stages. In the first stage, models extract sentences explicitly relevant to any DigiReady digital readiness dimension. In the second stage, each sentence is classified into one or more dimensions (D1–D7) with associated confidence scores. By enforcing standardized prompts, output schemas, and structured reasoning via DSPy, the framework minimizes variability from prompt interpretation and formatting differences, ensuring that observed performance reflects model capability rather than implementation artifacts.

To maintain comparability across models, all inference parameters were fixed: a **temperature** of 0.1, maximum tokens of **1024** for extraction and **512** for classification, and **3 maximum retries** in case of extraction or classification errors. No parameter tuning was performed for individual models.

This controlled configuration, combined with the tiered hardware setup, enables consistent evaluation of extraction efficiency, classification accuracy, and reasoning robustness across edge, consumer-level, and mid-range models.

The evaluation of model performance was conducted at two hierarchical levels, reflecting the dual objectives of the study: (1) identifying whether a paragraph is relevant to the DR framework, and (2) determining the confidence of its alignment to specific DR dimensions. Since models provide confidence scores at the sentence level, paragraph-level metrics were derived by averaging the confidence of all extracted sentences per dimension.

At the first level, the task was to determine whether a paragraph is relevant to the "DR" framework. Paragraphs were labeled as DR-related if the model extracted at least one sentence deemed relevant to the DigiReady framework. To evaluate performance, we calculated the overall concordance rate, as well as positive concordance rates. These metrics assess the agreement between model predictions and the ground truth.

The **second-level evaluation** examines the model's ability to align its predicted confidence scores with specific DR dimensions. While ground truth scores are defined at discrete levels (0, 0.5, 1), the model outputs continuous values in the range [0, 1]. To enable direct comparison, model

confidence scores were discretized as follows: scores below 0.5 were mapped to 0; scores from 0.5 (inclusive) to 0.75 were mapped to 0.5; and scores of 0.75 or higher were mapped to 1.

This discretization ensures compatibility with evaluator scoring. Performance was quantified for each dimension using the following metrics. **Mean Absolute Error (MAE):** The average absolute difference between predicted and ground truth scores, reflecting overall prediction accuracy. **Mean Squared Error (MSE):** The average squared difference, emphasizing larger deviations between predicted and true scores. **Agreement:** The proportion of paragraphs where the model's discretized confidence exactly matches the ground truth. **Concordance Rate:** The fraction of cases in which the model correctly identifies the presence or absence of each dimension, independent of confidence magnitude. **Precision, Recall, and F1** were **not used** because our evaluation emphasizes multi-dimensional alignment and continuous confidence scores across overlapping DR dimensions, capturing nuanced information crucial for a future human-in-the-loop methodology to handle uncategorized text, something threshold-based metrics cannot adequately reflect.

6 Large Language Model Text Classification Results

This section presents the performance of the selected LLMs on the DigiReady (DR) paragraph classification and dimension alignment tasks. Results are summarized through both numerical tables and heatmaps, highlighting overall model accuracy, per-dimension error, correlation, and concordance rate.

As described in Section 5.4, each model was first instructed to extract sentences from a given paragraph that were relevant to the DR framework. The extracted sentences were then classified by the model into one or more dimensions of the DR framework, using the provided descriptions of each dimension as context, and assigning a confidence score ranging from 0 to 1.

For each paragraph, the confidence score for a particular dimension was calculated as the mean of the confidence scores assigned to its extracted sentences for that dimension. A paragraph was considered "DigiReady-related" if the model extracted at least one relevant sentence; otherwise, it was considered "Not DigiReady-related". The overall paragraph confidence score per dimension was converted to confidence class 0, 0.5 and 1 as described in section 5.4.

6.1 Model Performance for Paragraph Level Identification

In the first-level evaluation of binary paragraph classification, 21 of 118 paragraphs in the ground truth dataset were labeled as 'DR-related.' Concordance rates varied across models. ***Qwen2.5:32b-instruct*** (overall 84.7%, positive 71.4%) and ***Qwen2.5:3b-instruct*** (72.0%, 76.2%) were the only models achieving both substantial overall and positive concordance rates (>0.7). The ***Phi-4-mini-instruct*** (63.6%, 81.0%) and ***Qwen2.5:1.5b-instruct*** (55.9%, 76.2%) showed high positive but lower overall concordance, while ***Granite3.1-dense:2b*** reached the highest positive concordance (95.2%) with low overall agreement (31.4%). Other models, including ***Qwen2.5:14b-instruct***, ***Internlm2.5:7b-chat***, and ***Falcon3:3b***, had high **overall concordance rates** but lower positive concordance (52–57%). Sentence extraction times ranged from 3.02 s to 13.34 s across models.

6.2 Model Performance for Dimension Level Identification

Table 2 summarizes the quantitative evaluation results across all models using mean absolute error (MAE), mean squared error (MSE), correlation (Corr), agreement per dimension, and concordance rate per dimension. Overall, clear performance differences were observed between smaller baseline models and larger or instruction-tuned variants.

The model ***Qwen2.5:0.5b*** showed the lowest predictive accuracy, with high errors (MAE 0.219 ± 0.221; MSE 0.190 ± 0.214) and near-zero correlation (−0.034 ± 0.063). It also had the **lowest**

agreement (85.3 ± 27.9) and concordance rate per dimension (0.723 ± 0.236), indicating unstable and inconsistent performance across dimensions, as reflected in Figure 2.

Instruction-tuned small and mid-sized models showed improved performance.

The small-sized models, ***Qwen2.5:1.5b-instruct*** and ***Granite3.1-dense:2b*** reduced errors (MAE < 0.10) and achieved high concordance (>0.85), with ***Qwen2.5:1.5b-instruct*** showing positive correlation (0.207 ± 0.100) to reference annotations.

Among mid-sized models, ***Qwen2.5:3b-instruct***, ***Phi-4-mini-instruct***, and ***Falcon3:3b*** balanced metrics, with ***Qwen2.5:3b-instruct*** attaining the lowest MAE (0.063 ± 0.028) and MSE (0.042 ± 0.024), high agreement (105.7 ± 4.6), and concordance (0.896 ± 0.039), while ***Falcon3:3b*** showed slightly lower variance, indicating robust generalization.

The mid-sized model ***Internlm2.5:7b-chat*** achieved the best combined results, with the lowest MAE (0.054 ± 0.025), lowest MSE (0.035 ± 0.019), highest correlation (0.230 ± 0.178), and the highest agreement (107.000 ± 4.472) and concordance rate per dimension (0.907 ± 0.038).

Larger models and higher-capacity chat variants showed comparable performance, ***Qwen2.5:14b-instruct*** and ***Qwen2.5:32b-instruct***, both maintaining low error levels and concordance rates exceeding 0.90, although without substantial gains over the 7B model.

Dimension-level analysis (Figure 2) shows that most models achieved consistently high concordance (≈0.8–1.0) across dimensions D3–D7, while greater variability appeared in early dimensions (D1–D2), particularly for smaller models. Correlation patterns similarly indicate that improvements in larger models primarily arise from increased stability rather than large per-dimension performance gains.

Model	MAE ± SD	MSE ± SD	Corr ± SD	Dimension Concordance
Qwen2.5:0.5b	0.219 ± 0.221	0.190 ± 0.214	-0.034 ± 0.063	0.723 ± 0.236
Qwen2.5:1.5b-instruc	0.077 ± 0.050	0.054 ± 0.042	0.207 ± 0.100	0.875 ± 0.069
Granite3.1-dense:2b	0.095 ± 0.049	0.072 ± 0.041	-0.024 ± 0.039	0.860 ± 0.069
Qwen2.5:3b-instruct	0.063 ± 0.028	0.042 ± 0.024	0.115 ± 0.178	0.896 ± 0.039
Phi-4-mini-instruct	0.072 ± 0.040	0.052 ± 0.038	0.075 ± 0.103	0.889 ± 0.049
Falcon3:3b	0.067 ± 0.026	0.048 ± 0.023	0.081 ± 0.103	0.893 ± 0.036
Internlm2.5:7b-chat	0.054 ± 0.025	0.035 ± 0.019	0.230 ± 0.178	0.907 ± 0.038
Qwen2.5:14b-instruct	0.055 ± 0.024	0.035 ± 0.018	0.162 ± 0.181	0.906 ± 0.037
Qwen2.5:32b-instruct	0.057 ± 0.024	0.038 ± 0.018	0.050 ± 0.103	0.904 ± 0.037

Table 2 Model performance metrics, showing mean ± SD for MAE, MSE, and correlation, plus per-dimension concordance rates.

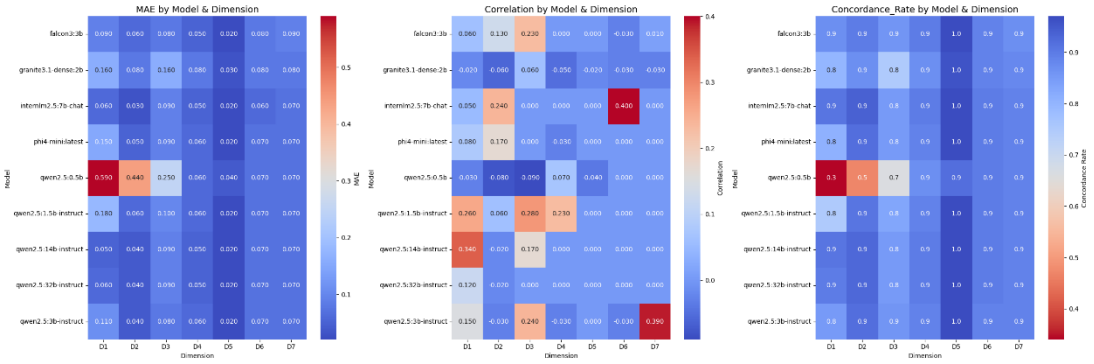

Figure 2 Heatmap distribution of metrics MAE, Correlation and Concordance rate per dimension per model

7 Discussion and Conclusions

This study forms part of a broader effort to evaluate digital readiness in Higher Education Institutions (HEIs) by exploring whether institutional documents contain identifiable evidence aligned with the DigiReady framework. This work is an initial attempt to determine whether these documents exist in institutions, checks if they can provide systematic evidence and investigates whether open-source LLMs can support their analysis in privacy-sensitive, resource-limited settings.

HEIs produce extensive documentation, including strategies, policies, and reports, but it is rarely used systematically. Analysis of 75 documents from five institutions showed strong coverage in governance-related areas: Digital Leadership (48%), Strategy (46.7%), and Collaboration (38.7%), while operational and teaching-related areas (D3–D5) were below 35%, and D6 had the lowest coverage. This reflects a strong focus on governance and highlights the need for more implementation- and evaluation-oriented evidence.

Documents can be grouped into three abstract types: strategic/governance documents, dominating D1, D2, and D7; implementation/operational documents, focused on D3–D6; and evaluation/outcome-oriented documents, spanning multiple dimensions. This typology can guide evidence identification for institutions with different structures and support the development of an intelligent knowledge base for cross-institutional digital readiness assessment.

A key limitation is the geographic scope, since four Greek and one Cypriot institution were included, which may limit generalizability. Nonetheless, the document types are common across HEIs, and the DigiReady framework has broad European relevance. Future research should include more linguistically diverse institutions.

To assess open-source language models, we used a single external evaluation report. Two experts labeled relevant paragraphs by DR dimensions, forming a ground-truth dataset. Binary relevance agreement was high (86%), but dimension-level positive agreement was lower (e.g., 40% for D6), indicating the need for clearer guidelines and expanded validation.

Evaluation revealed that mid-sized, instruction-tuned models (3B–7B parameters) provide the best balance between accuracy and computational efficiency. The top-performing model, ***Internlm2.5:7b-chat***, achieved performance comparable to larger LLMs while requiring substantially lower computational resources. These findings suggest that instruction tuning and architectural optimization may be more influential than parameter scale alone for institutional document classification tasks. Differences between binary relevance detection and multi-dimensional classification performance further indicate that hybrid or staged classification strategies may improve robustness. The relatively low correlation values, despite high concordance rates, likely reflect the strong class imbalance and sparsity of positive dimension labels, where agreement on dominant negative classes inflates concordance while reducing sensitivity to confidence variation. Dataset imbalance (17.8% DR-related paragraphs) will be addressed in future work.

Open-source deployment enables institutions to perform document analysis locally, supporting compliance with the General Data Protection Regulation (GDPR) and institutional data-governance requirements. AI-assisted methods can structure fragmented institutional documentation and support continuous, evidence-informed digital readiness assessment within institutional governance and quality assurance processes. Future work will expand the benchmark dataset and develop workflows that combine AI efficiency with expert oversight. In conclusion, mid-sized open-source AI models can support scalable, evidence-based digital transformation governance by enabling HEIs to systematically analyze institutional documents for digital readiness.

Acknowledgements

This work was funded by the Erasmus+ Programme of the European Union and coordinated by the IKY Hellenic Scholarships Foundation (Project No. 2025-1EL01-KA220HED000358593), in partnership with GUNet Greece (grant holder), 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

Alarfaj, F. K., Khan, H. U., Naz, A., & Almusallam, N. (2026). A real-time large language model framework with attention and embedding representations for misinformation detection. Engineering Applications of Artificial Intelligence, 164(Part B), 113304. https://doi.org/10.1016/j.engappai.2025.113304

Bendermacher, G. W. G., oude Egbrink, M. G. A., Wolfhagen, I. H. A. P., & Dolmans, D. H. J. M. (2017). Unravelling quality culture in higher education: A realist review. Higher Education, 73(1), 39–60. https://doi.org/10.1007/s10734-015-9979-2

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), 59. https://doi.org/10.1186/s41239-024-00491-0

Fourrier, C., Habib, N., Lozovskaya, A., Szafer, K., & Wolf, T. (2024). *Open LLM leaderboard v2*. Hugging Face. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard

Funda, V. (2024). Intelligent decision support systems in higher education institutions in developing countries: A systematic literature review. In H. Twinomurinzi, N. T. Msweli, S. Gumbo, T. Mawela, E. Mtsweni, P. Mkhize, & E. Mnkandla (Eds.), Proceedings of the NEMISA Digital Skills Summit and Colloquium 2024 (Vol. 6, pp. 189–202). EasyChair. https://doi.org/10.29007/slvr

Khattab, O., Santhanam, K., Li, X. L., Hall, D., Liang, P., Potts, C., & Zaharia, M. (2023a). Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP (arXiv:2212.14024). arXiv. https://doi.org/10.48550/arXiv.2212.14024

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., & Potts, C. (2023b). DSPy: Compiling declarative language model calls into self-improving pipelines (arXiv:2310.03714). arXiv. https://doi.org/10.48550/arXiv.2310.03714

Khilji, N., Nicolic, K., & Ikram-ur-Rehman. (2024). The influence of knowledge management on digital transformation: An overview for managing change and innovation. In K. Arai (Ed.), Advances in information and communication (Lecture Notes in Networks and Systems, Vol. 919, pp. 368–388). Springer. https://doi.org/10.1007/978-3-031-53960-2_24

Koperwas, J., Skonieczny, Ł., Kozłowski, M., Andruszkiewicz, P., Rybiński, H., & Struk, W. (2017). Intelligent information processing for building university knowledge base. Journal of Intelligent Information Systems, 48(1), 141–163. https://doi.org/10.1007/s10844-015-0393-0

Kostina, A., Dikaiakos, M. D., Stefanidis, D., & Pallis, G. (2025). Large Language Models For Text Classification: Case Study And Comprehensive Review (arXiv:2501.08457). arXiv. https://doi.org/10.48550/arXiv.2501.08457

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 L. Desnos, R. Vogl, L. Merakos, C. Diaz, J. Mincer-Daszkiewicz, & S. McLellan (Eds.), Proceedings of EUNIS 2025 Annual Congress in Belfast (EPiC Series in Computing, Vol. 107, pp. 196–205). EasyChair. https://doi.org/10.29007/grb9

Niimi, J. (2025). A Simple Ensemble Strategy for LLM Inference: Towards More Stable Text Classification (arXiv:2504.18884). arXiv. https://doi.org/10.48550/arXiv.2504.18884

Silveira, R., López Matias, J. V., Ponte, C., Donza Corrêa, I., Câmara, A., & Furtado, V. (2026). Optimizing prompts for legal document classification. In Proceedings of the Twentieth International Conference on Artificial Intelligence and Law (pp. 369–373). Association for Computing Machinery. https://doi.org/10.1145/3769126.3769217

Tsimpanis, K., Plessas, A., Balaouras, P., & Avouris, N. (2025). UDReady: A data-driven platform measuring digital readiness of higher education institutions. In R. Vogl, L. Desnos, J.-F. Desnos, S. Bolis, L. Merakos, G. Ferrell, E. Tsili, & M. Roumeliotis (Eds.), Proceedings of the EUNIS 2024 annual congress in Athens (EPiC Series in Computing, Vol. 105, pp. 254–264). EasyChair. https://doi.org/10.29007/vczq

Uslu, B. (2025). Evolving Steps of Quality Culture in Universities: Outline from European University Association's (EUA) Institutional Evaluation Reports. Yükseköğretim Dergisi, 15(Special Issue), 49-58. https://doi.org/10.53478/yuksekogretim.1553711