WSB Journal of Business and Finance
Year 2026, Vol. 60, No.2
eISSN 2657-4950

A Hybrid LLM–CBR Approach for Explainable and Stable Assessment of Academic Theses

Artur Ziółkowski[a], Paweł Tomkiewicz [a1,]

[a]*WSB Merito Gdańsk, Instytut Informatyki i Nowoczesnych Technologii, Gdańsk, Polska*

DOI 10.2478/WSBJBF-2026-0010

Abstract

The assessment of master's theses is a complex and time-consuming process, often burdened by a significant degree of subjectivity. In recent years, Large Language Models (LLMs) have demonstrated high effectiveness in analyzing academic texts; however, their application to the evaluation of degree theses raises concerns about inconsistent decisions and limited ability to objectively verify content. This article proposes a hybrid approach that integrates LLMs with the Case-Based Reasoning (CBR) methodology to support the assessment of master's theses by leveraging analogies to previously evaluated cases. The paper presents the system architecture, the formalization of a case representation, the integration of LLMs with a case base, and methods for evaluating the quality of assessments and their justifications. The analysis indicates that the proposed approach can enhance assessment consistency and improve transparency while preserving the role of the human evaluator as the final decision-maker.

Keywords: Large Language Models, Case-Based Reasoning, master's thesis assessment, AI in education, explainable AI

1. Introduction

The dynamic development of Large Language Models (LLMs) has significantly influenced text processing and analysis across a wide range of domains, including information retrieval, document analysis, and decision-support systems (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020). Trained on vast

[1]Corresponding author: Tel.: +48 600577267; E-mail address: pawel.tomkiewicz@gdansk.merito.pl

datasets, LLMs demonstrate strong capabilities for understanding context and generating coherent outputs, making them attractive for complex tasks such as assessing academic theses. However, their application in high-responsibility decision-making contexts reveals critical limitations: hallucinations (Ribeiro et al., 2016; Guidotti et al., 2018), limited stability, and insufficient explainability hamper their acceptance in academic and institutional environments. Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) has partially addressed these issues by grounding LLM outputs in externally retrieved documents, with demonstrated improvements in factual accuracy (Gao et al., 2023; Shi et al., 2023). Nevertheless, RAG approaches retrieve unstructured text fragments without imposing any formal decision structure, leaving the model without a normative anchor derived from prior expert-validated outcomes, which is insufficient for consistent and auditable academic assessment.

In parallel, Case-Based Reasoning (CBR) (Aamodt & Plaza, 1994; Kolodner, 1993) offers a well-established decision-support paradigm grounded in the reuse of historical expert-validated cases, providing transparency and explainability by explicitly linking decisions to concrete reference cases. However, classical CBR systems exhibit limited capability in analyzing rich unstructured textual data, constraining their applicability to semantically complex documents such as academic theses. Several hybrid architectures combining LLMs with knowledge bases and semantic search have been proposed (Hendri et al., 2026; Keane & Smyth, 2020), yet the integration of LLMs specifically with CBR remains underexplored, particularly in the domain of academic assessment where consistency, explainability, and alignment with expert practice are critical requirements.

This study addresses this gap by proposing and empirically evaluating a hybrid LLM-CBR decision-support model for academic thesis assessment. The study is deliberately scoped to a single academic discipline and educational level, constituting a controlled proof-of-concept aimed at isolating the effect of CBR grounding before broader generalization is attempted. The authors analyze the impact of this integration on assessment consistency, agreement with expert decisions, and the quality of generated justifications. LLM applicability to supporting assessment of legal acts and grant applications (Choi et al., 2023; Shrestha et al., 2019) further motivates the relevance of the proposed approach in broader evaluation contexts, which are however left for future work.

2. State of the Art

2.1. Large Language Models in Decision-Making Tasks

In recent years, Large Language Models have become the subject of intensive research, both in their architectures and in their potential practical applications. The literature demonstrates their high effectiveness in tasks such as document summarization, text classification, sentiment analysis, and question answering (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020). Increasingly, LLMs are also being employed for tasks that require decision-making or the formulation of qualitative judgments, including the assessment of legal documents and grant applications (Choi et al., 2023; Shrestha et al., 2019).

A significant direction aimed at improving the reliability of LLM outputs in knowledge-intensive tasks is Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), which supplements the language model with externally retrieved documents at inference time. RAG-based systems have demonstrated measurable reductions in hallucination rates and improved factual grounding across a range of tasks (Gao et al., 2023; Shi et al., 2023). However, RAG retrieves unstructured text fragments and provides them as passive context, without imposing any formal decision structure or normative grounding derived from expert-validated outcomes. This distinction is particularly relevant in academic assessment, where decisions must be consistent, justifiable, and traceable to established evaluation practice rather than to arbitrary retrieved passages.

At the same time, numerous studies indicate fundamental limitations of generative models in decision-making contexts. The problem of hallucinations, the lack of determinism in generated responses, and limited control over the reasoning process constitute significant barriers to their application in environments where high responsibility and decision auditability are required (Ribeiro et al., 2016; Guidotti et al., 2018).

Techniques such as prompt engineering and chain-of-thought prompting have been proposed in response, however their effectiveness remains strongly dependent on the application context.

2.2. Case-Based Reasoning as a Decision Support Tool

Case-Based Reasoning is one of the classical paradigms of artificial intelligence and has been widely applied in decision support systems, diagnostics, planning, and education (Aamodt & Plaza, 1994; Kolodner, 1993). A key advantage of the CBR approach is its ability to use real, historical cases as a basis for decision-making, thereby promoting consistency and increasing users' trust in the system. Unlike RAG-based retrieval, CBR does not merely surface relevant text fragments but provides structured, expert-validated decision contexts including prior outcomes and their justifications, which constitutes a fundamentally different form of grounding.

In the educational context, CBR has been used to support assessment processes, recommend educational materials, and analyze student progress (Holmes et al., 2019; Zhai et al., 2021). However, the effectiveness of such systems strongly depends on the quality of case representation and the similarity measures employed. Moreover, classical CBR systems encounter difficulties when dealing with long and complex texts, which limits their applicability to the assessment of academic theses characterized by rich semantic structures.

2.3. Hybrid LLM–CBR Approaches

The integration of generative models with classical AI methods is currently one of the most actively explored research directions. Several works have proposed combining LLMs with structured knowledge sources, including knowledge graphs, rule-based systems, and semantic search mechanisms (Hendri et al., 2026; Keane & Smyth, 2020). More specifically, within the CBR community, recent studies have explored using neural embeddings to improve case retrieval, replacing traditional similarity measures with learned representations (Mathisen et al., 2019; Pauli et al., 2023). However, these approaches treat the neural component primarily as a retrieval mechanism, without leveraging the full generative and interpretative capabilities of modern LLMs.

The integration of LLMs specifically with CBR remains relatively underexplored. Existing studies focus primarily on architectural aspects, proposing frameworks in which LLMs generate candidate solutions that are subsequently validated against a case base, or conversely, in which CBR retrieval initializes LLM generation (Hendri et al., 2026; Keane & Smyth, 2020). However, systematic empirical evaluations of such architectures within well-defined decision-making tasks are largely absent from the literature, and no study to our knowledge has examined this integration in the context of academic thesis assessment.

This gap is particularly significant given that academic assessment imposes requirements that neither standalone LLMs nor classical CBR can fully satisfy individually. The proposed hybrid model addresses this by assigning complementary roles to each component: the LLM handles semantic analysis and justification generation, while the CBR system provides normative grounding through formally structured historical cases. This division of responsibility is the key architectural distinction from both RAG-based approaches and prior hybrid LLM-CBR frameworks.

3. Research Problem and Objectives

Despite the rapidly growing popularity of Large Language Models in educational applications, their use as decision-support tools for the assessment of academic theses remains highly problematic. Although LLMs demonstrate strong capabilities in analyzing complex textual content, they do not inherently guarantee assessment stability, reproducibility, or alignment with established academic evaluation practices (Ribeiro et al., 2016; Guidotti et al., 2018). As a result, their direct application in high-stakes academic decision-making processes raises concerns regarding reliability, fairness, and institutional accountability (Hendri et al., 2026; Keane & Smyth, 2020).

At the same time, higher education institutions are increasingly facing practical pressures related to the scalability and efficiency of assessment processes. Growing student numbers, limited availability of qualified

reviewers, and the need to ensure consistency across evaluators create a demand for computational tools that can support academic staff without undermining the principles of academic integrity and expert judgment. In this context, fully automated assessment systems based solely on generative AI are generally perceived as insufficient, particularly due to issues related to explainability, auditability, and trust.

Conversely, classical Case-Based Reasoning offers a well-established decision-support paradigm grounded in the reuse of historical expert-validated cases. CBR systems provide transparency and interpretability by explicitly linking decisions to concrete reference cases, which aligns well with academic requirements for justification and traceability. However, traditional CBR approaches struggle with the semantic analysis of long unstructured textual documents such as master's theses, limiting their practical applicability in contemporary educational settings.

The research problem addressed in this study therefore concerns the question of whether the integration of Large Language Models and Case-Based Reasoning into a hybrid decision-support model can overcome the individual limitations of each approach and produce a synergistic effect. In particular, the study investigates whether anchoring LLM-generated assessments in historically evaluated cases can improve decision quality, reduce hallucinations, enhance stability, and increase the level of explainability required in academic evaluation processes. The study is deliberately scoped to a single academic discipline and educational level as a controlled proof-of-concept; potential applicability to other evaluation domains is acknowledged but left for future research.

Based on these considerations, the following research question is formulated: Does a hybrid model integrating Large Language Models and Case-Based Reasoning yield more consistent, expert-aligned assessment decisions than applying each approach independently?

To address this research question, a structured research process was conducted, comprising the following stages: first, the design of a hybrid LLM-CBR system architecture dedicated to the assessment of academic theses; second, an empirical comparison of the performance of three approaches: standalone LLM, classical CBR, and the proposed hybrid LLM-CBR model; third, an analysis of the impact of integration on assessment stability, agreement with expert decisions, and the quality of generated justifications.

4. Architecture

The proposed architecture of the hybrid decision-support system is based on the assumption of complementarity between two distinct artificial intelligence paradigms: Large Language Models and Case-Based Reasoning. The architecture has been designed to preserve the autonomy of each component while enabling close cooperation among them in the assessment of academic theses.

A fundamental design assumption is the clear separation of roles between the system components. The language model serves an interpretive and generative function, analyzing thesis content, identifying relevant qualitative features, and formulating assessment justifications. The CBR system, in turn, plays a stabilizing and referential role by providing the language model with decision-making context in the form of real, historical cases previously evaluated by experts.

An essential element of the architecture is the assumption that the final assessment decision is not generated solely by either the LLM or the CBR component, but rather emerges from a controlled interaction between the two. This approach aims to mitigate hallucinations, increase assessment consistency, and ensure decision explainability by enabling reference to specific historical cases.

The system architecture consists of five main functional layers: the input layer, the semantic analysis layer (LLM layer), the case base layer, the integration layer (CBR–LLM integration layer), and the decision and explanation layer. Data flow within the system follows a sequential–iterative pattern. The academic thesis is first processed by the semantic analysis layer, which analyzes it using the language model. Subsequently, based on the generated semantic representation, the CBR system retrieves the most similar historical cases. This information is then used to recontextualize the LLM's assessment-generation process.

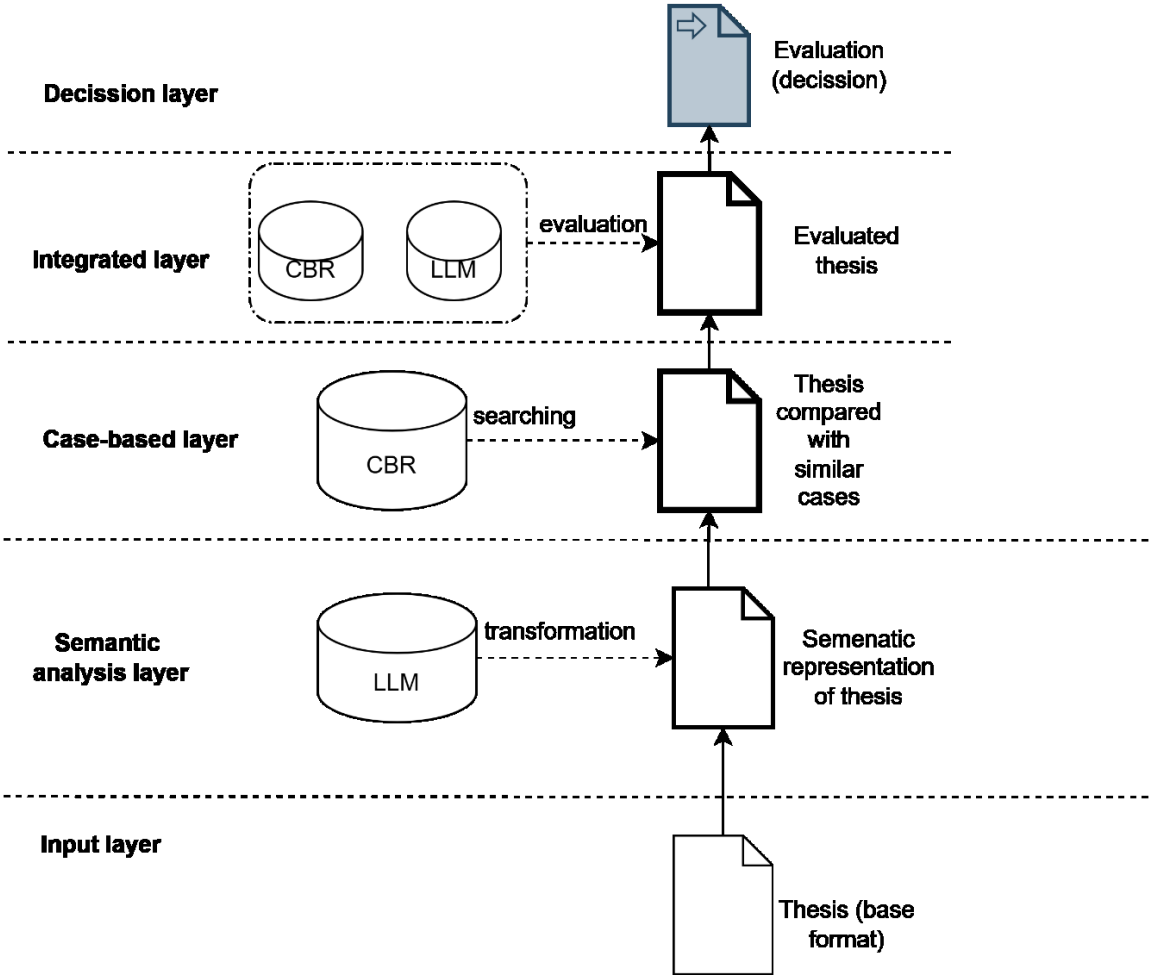

Fig. 1. Architecture of LLM-CBR integrated model for thesis evaluation

The academic thesis is first processed by the semantic analysis layer (LLM), which generates a representation of its qualitative features. Based on this representation, the CBR system retrieves similar cases from the historical case base. The selected cases are then integrated into the LLM's assessment-generation process, leading to a final decision accompanied by an explanation grounded in the reference cases.

Semantic Analysis Layer (LLM Layer)

The Large Language Model is responsible for analyzing the full content of the academic thesis, including its structure, the coherence of its argumentation, substantive correctness, and the degree to which its research objectives have been achieved. The LLM processes the document as unstructured text and generates an internal semantic representation, which is subsequently used for both the preliminary quality assessment of the thesis and the initialization of the case retrieval process within the CBR system.

In contrast to approaches based solely on LLMs, the model does not generate a final decision at this stage. Its role is limited to extracting qualitative features and preparing contextual information for subsequent decision-making. This deliberate restriction of the language model's autonomy constitutes one of the mechanisms aimed at reducing hallucinations and excessive variability in assessments.

Case Base and CBR Mechanism

The case base consists of historical academic theses evaluated by examination committees. Each case includes both a description of the thesis features (e.g., subject scope, methodology, structure) and the assessment outcome accompanied by expert justification. The case representation is designed to enable partial matching with the analyzed thesis, without requiring full structural correspondence between documents.

The CBR mechanism implements the classical case-based reasoning cycle, including case retrieval, adaptation, and use in decision-making. Similarity retrieval is based on semantic features generated by the LLM, enabling the system to overcome the limitations of traditional similarity measures commonly used in classical CBR systems.

LLM–CBR Integration Layer

A key component of the architecture is the integration layer, which enables true synergy between the LLM and CBR. The selected reference cases are provided to the language model as explicit decision-making context. The LLM receives not only the content of the evaluated thesis but also a set of similar cases together with their assessments and expert justifications. The integration process follows a defined algorithmic procedure formalized in Algorithm 1.

```
Input:  thesis T, case base C, number of retrieved cases k
Output: assessment scores S, justification J

F_T = LLM.encode(T)

similarities = []
for each case c_i in C:
    F_i = encode(c_i.features)
    sim_i = cosine_similarity(F_T, F_i)
    similarities.append((c_i, sim_i))

C* = top_k(similarities, k)

context_blocks = []
for each c_i in C*:
    block = format(c_i.description, c_i.scores, c_i.justification)
    context_blocks.append(block)

prompt = build_prompt(T, context_blocks, assessment_criteria)

S, J = LLM.generate(prompt)

return S, J
```

Algorithm 1. LLM–CBR Integration Procedure. Input: thesis T, case base C, number of retrieved cases k. Output: assessment scores S, justification J

Step 1. The LLM processes the full text of the evaluated thesis T and generates a semantic feature vector F(T) representing its qualitative characteristics across the defined assessment criteria.

Step 2. The CBR retrieval mechanism computes similarity scores between F(T) and all cases in the case base C, selecting the k most similar cases: $C^* = \{c_1, c_2, ..., c_k\}$ where similarity is measured as cosine distance between semantic feature vectors.

Step 3. Each retrieved case c_i is formatted into a structured context block containing: the thesis subject description, partial scores per criterion, and the original expert justification text.

Step 4. The formatted case blocks are injected into the LLM prompt as explicit decision context, constructing the augmented prompt P as follows:

$$P = [System\ instruction] + [Thesis\ text\ T] + [Retrieved\ cases\ C^*] + [Assessment\ instruction]$$

Step 5. The LLM generates the final assessment under constrained decision freedom, producing partial scores for each criterion and a justification that explicitly references the provided historical cases.

As shown in Algorithm 1, the LLM generates an assessment under conditions of constrained decision freedom, taking into account evaluation patterns derived from established academic practice. This layer serves a regulatory function by limiting extreme decisions and enforcing consistency with historical assessment standards.

Decision and Explainability Layer

The final assessment decision is generated by the LLM based on the thesis content and the reference cases supplied by the CBR system. A key feature of this stage is the generation of a justification that explicitly cites similar cases, highlighting both common elements and differences between the evaluated thesis and historical works.

This approach significantly increases decision explainability, enabling auditability and expert review. In contrast to classical systems based solely on LLMs, the decision justification is not purely narrative but grounded in real institutional data.

5. Case Representation and Evaluation Criteria

5.1. Case Structure and Components

In the proposed hybrid LLM–CBR model, a case is defined as a formal representation of a historical academic thesis together with its expert assessment and justification. It is assumed that a case does not merely record the final assessment outcome, but reflects the broader decision-making context in which the thesis was evaluated. This approach enables cases to be used not only as quantitative reference points, but also as carriers of qualitative knowledge.

Each case consists of three main components: (1) structural metadata of the thesis, (2) a description of qualitative features, and (3) the assessment decision together with its justification. The metadata include information such as the subject area, level of study, type of thesis (theoretical, empirical, project-based), and document length. This component serves as an initial filtering mechanism, eliminating cases that are clearly non-comparable.

The qualitative features of the thesis constitute the core element of the case representation. They include, among others, the degree of achievement of research objectives, structural coherence, methodological correctness, level of independence, and innovativeness. These features are not encoded as rigid numerical values, but are stored as semantic descriptions, which enables their further processing by the language model.

The third component of a case is the assessment decision, including the final grade and the justification prepared by the examination committee. This justification plays a particularly important role in the hybrid architecture, as it provides an argumentation pattern that the LLM can reference when generating its own assessment justifications.

5.2. Similarity Criteria and the Role of LLM in Case Retrieval

In classical CBR systems, similarity between cases is typically defined using manually selected features and distance measures. In the proposed model, a different approach is adopted, in which the language model serves as an intermediary in determining semantic similarity.

The LLM generates an internal semantic feature vector representation of the analyzed academic thesis, accounting for both its content and its argumentative structure. This representation is then used to retrieve the most similar cases from the CBR case base, as outlined in Algorithm 1. Such a solution allows the system to capture subtle qualitative aspects that are difficult to formalize using traditional numerical features.

The case retrieval process is not based on deterministic matching but rather on identifying a set of reference cases with a high degree of semantic similarity. This set subsequently serves as decision-making context for the LLM, enabling it to generate assessments grounded in real evaluation practice.

6. Experimental Setup and Procedure

6.1. Dataset

The experiment was conducted on a dataset comprising 60 academic theses completed within computer science programs at the undergraduate (engineering and bachelor's) level. All these had been previously evaluated by official examination committees in accordance with established academic procedures. Each thesis was associated with a final grade as well as written justifications prepared independently by the thesis supervisor and an external reviewer.

The dataset was deliberately scoped to a single academic discipline and educational level. This constitutes an intentional methodological choice aimed at establishing a controlled baseline for the hybrid model, allowing the effect of CBR grounding to be isolated without confounding variables introduced by cross-domain variation. Extension to broader datasets across multiple disciplines and degree levels is identified as the primary direction for future work.

The dataset was divided into two disjoint subsets: a case base consisting of 40 theses, used exclusively for constructing the CBR system, and a test set consisting of 20 theses, used solely for system evaluation. This separation ensured that information leakage between the case construction and evaluation phases was prevented, thereby enabling a fair and reliable comparison of the analyzed approaches. None of the theses included in the test set were accessible to the systems during case retrieval or decision grounding.

6.2. Thesis Assessment Criteria

The assessment of academic theses was based on a set of five qualitative criteria commonly applied in academic evaluation practice:

- substantive and methodological correctness,
- degree of achievement of the thesis objectives,
- coherence of structure and argumentation,
- level of the author's independence,
- innovativeness and practical applicability.

For the purpose of experimental clarity, each criterion was evaluated on a five-point ordinal scale (1–5). Although weighted grading schemes are often applied in institutional practice, an unweighted scale was deliberately adopted to simplify interpretation and ensure comparability across approaches. The final grade was computed as the arithmetic mean of the partial scores.

The identical set of criteria and grading scale was consistently applied across all three evaluated approaches, ensuring methodological comparability and eliminating potential bias arising from differing evaluation frameworks.

6.3. Experimental Procedure

The experimental procedure consisted of three independent evaluation scenarios, each corresponding to one of the analyzed approaches.

In the first scenario, theses were assessed using a standalone Large Language Model (LLM). The model was provided with the full text of the thesis and a formal description of the evaluation criteria. However, it had no access to historical cases, prior assessments, or expert justifications. For each thesis, the model generated partial scores for each criterion and a narrative justification of the assessment.

In the second scenario, a classical Case-Based Reasoning (CBR) system was employed. The assessment was produced by retrieving the most similar cases from the historical case base and adapting their assessment outcomes. The CBR system operated on structural metadata and a simplified feature representation, without performing a full semantic analysis of the thesis text.

In the third scenario, the proposed hybrid LLM–CBR model was applied. The language model analyzed the thesis text while being explicitly provided with a set of retrieved reference cases, including their partial scores and expert-written justifications. The final assessment was generated by the LLM, but under conditions of decision grounding in the historical case base.

Each thesis in the test set was evaluated independently in all three scenarios, using identical criteria and grading scales. This resulted in three separate assessment outcomes per thesis, enabling direct comparison across approaches.

6.4. Evaluation Metrics

System performance was evaluated using three quantitative metrics and one qualitative metric.

The first quantitative metric was agreement with expert assessment, measured as the mean absolute difference between the system-generated grade and the grade assigned by the examination committee. This metric reflects how closely the system aligns with expert judgment. Results are reported with 95% confidence intervals computed using bootstrap resampling over the test set.

The second metric was assessment stability, defined as the variance of grades generated by the system across multiple independent runs. This metric is particularly relevant for generative models, where stochasticity may lead to variability in outcomes. Effect size for variance differences between approaches was computed using Cohen's d to provide a standardized measure of practical significance beyond statistical significance.

The third quantitative metric captured internal consistency of partial scores, as measured by the standard deviation of criterion-level scores. Lower dispersion indicates greater coherence in the assessment across evaluation dimensions. Effect sizes are reported alongside all pairwise comparisons between the three approaches.

The qualitative metric concerned the quality of assessment justifications. Justifications were evaluated by three independent academic experts, each holding a doctoral degree in computer science and possessing direct experience in academic thesis supervision and review. Each expert scored all justifications on a five-point scale across three dimensions: clarity of argumentation, coherence with the thesis content, and explicitness of reference to evaluation criteria. Inter-rater reliability was assessed using Krippendorff's alpha to ensure the validity of the qualitative scores. This metric aimed to capture aspects of explainability and interpretability that are critical for practical adoption in academic environments.

6.5. Quantitative Results

Table 1 presents the mean values of the evaluation metrics obtained for the three analyzed approaches, together with 95% confidence intervals and effect sizes for pairwise comparisons.

Table 1. Quantitative evaluation results (mean values with 95% confidence intervals)

Approach	Mean difference vs. expert	Assessment variance	Consistency of partial scores
LLM	0.62 [0.51, 0.73]	0.48 [0.39, 0.57]	0.41 [0.34, 0.48]
CBR	0.35 [0.27, 0.43]	0.21 [0.15, 0.27]	0.29 [0.23, 0.35]
LLM–CBR	0.22 [0.16, 0.28]	0.15 [0.10, 0.20]	0.18 [0.13, 0.23]

Effect sizes (Cohen's d) for the comparison between LLM–CBR and standalone LLM were large across all three metrics: $d = 0.91$ for mean difference versus expert, $d = 1.14$ for assessment variance, and $d = 0.87$ for consistency of partial scores, indicating that the observed improvements are of practical as well as statistical significance.

The results indicate that the hybrid LLM–CBR model achieved the highest agreement with the assessments assigned by examination committees, as reflected by the lowest mean absolute difference relative to expert grades. At the same time, the hybrid approach exhibited the lowest assessment variance, indicating increased stability across repeated evaluations.

An analysis of the qualitative justifications revealed notable differences between the evaluated approaches. Justifications generated by the standalone LLM were characterized by high linguistic fluency and coherence; however, in a subset of cases they contained statements insufficiently grounded in the actual content of the thesis. Justifications produced by the CBR system were consistent and firmly anchored in historical data, but often exhibited a schematic and repetitive structure limiting their explanatory richness. Inter-rater reliability for the qualitative evaluation reached Krippendorff's alpha = 0.74, indicating substantial agreement among the three expert evaluators.

The highest-rated justifications were generated by the hybrid LLM–CBR model. These justifications successfully combined natural language fluency with explicit references to relevant historical cases, thereby enhancing both transparency and credibility. The ability to directly relate the evaluated thesis to concrete reference cases proved particularly valuable from the perspective of auditability and expert review.

Overall, the obtained results confirm that integrating Large Language Models with Case-Based Reasoning leads to a substantial improvement in the quality of assessment decisions. The hybrid model demonstrated not only higher agreement with expert evaluations, but also greater stability and internal consistency of partial scores. These findings are interpreted within the scope of the present controlled proof-of-concept study; broader generalization across disciplines and educational levels remains a direction for future work.

7. Conclusions

This paper proposed and empirically evaluated a hybrid decision-support model that integrates Large Language Models with the Case-Based Reasoning paradigm for the assessment of academic theses. The starting point of the study was the observation that both standalone language models and classical CBR systems exhibit significant limitations when applied to complex, high-responsibility decision-making processes that simultaneously require the analysis of unstructured textual data, assessment stability, and a high level of explainability.

The conducted controlled proof-of-concept study confirmed that integrating the two approaches leads to a synergistic effect in which the weaknesses of one paradigm are compensated by the strengths of the other. The hybrid LLM–CBR model achieved higher agreement with examination committee assessments than both the standalone language model and the classical CBR system, while also exhibiting lower assessment variance and greater internal consistency of partial scores, with large effect sizes across all three metrics. These results indicate that grounding the assessment generation process in real historical cases plays a crucial stabilizing and normative role.

An important contribution of this work is the observed improvement in the quality of assessment justifications generated by the hybrid model. In contrast to justifications produced by the standalone language model, which in some cases exhibited hallucination-like characteristics, the justifications generated by the LLM–CBR approach were explicitly anchored in evaluation practice through direct references to similar historical cases. This form of argumentation significantly enhances decision transparency and auditability, which is essential in academic and institutional contexts.

The obtained results are interpreted within the scope of the present study, which was deliberately scoped to computer science theses at the undergraduate level as a controlled proof-of-concept. The findings support the viability of the proposed hybrid architecture within this well-defined domain. Whether the approach generalizes to other disciplines, educational levels, or evaluation domains remains an open empirical question and constitutes the primary direction for future research.

Future Research Directions

The presented study opens several avenues for future research. The most immediate and primary direction is the extension of the experiments to larger and more diverse datasets, including academic theses from different fields of study and educational levels. Such an extension would allow for a comprehensive assessment of the scalability of the proposed model and its robustness across varying domain contexts, addressing the deliberate scope limitation of the present study.

Another important research direction concerns a deeper analysis of the impact of the number and selection of reference cases on the quality of decisions generated by the hybrid model. Determining the optimal number of cases provided to the language model, as well as investigating how different case selection strategies influence assessment stability and explainability, remains an open research question.

It is also worth considering the application of formal explainability evaluation methods, including studies involving end users such as thesis supervisors and reviewers. Analyzing user perceptions of the credibility and usefulness of the system-generated justifications could provide valuable insights into the model's practical applicability and acceptance.

From a technical perspective, an interesting direction for further research is the integration of the LLM–CBR model with continuous learning mechanisms that enable the dynamic expansion of the case base with new expert decisions. Such an approach could support the gradual adaptation of the system to evolving evaluation standards and institutional practices.

Acknowledgements

This work was supported by the Polish Ministry of Education and Science under the "Science for Society II" program (project No. NdS-II/SN/0194/2024/01). The project received funding of 920,000 PLN with a total value of 1,012,000 PLN.

References

Aamodt, A., & Plaza, E. 1994. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(1), 39–59.

Brown, T.B., et al. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS), 1877–1901.

Choi, J.H., Hickman, K.E., Monahan, A., & Schwarcz, D. 2023. ChatGPT Goes to Law School. Journal of Legal Education, 71(3), 387–400.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 4171–4186.

Gao, Y., et al. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. 2018. A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42.

Hendri, H.M., Arifin, F., & Andayani, S. 2026. Reviewing The Combination of Case-Based Reasoning and Machine Learning for Improving a Decision Support System. Applied Information System and Management, 9(1), 81–88. DOI: 10.15408/aism.v9i1.48904

Holmes, W., Bialik, M., & Fadel, C. 2019. Artificial Intelligence in Education: Promise and Implications for Teaching and Learning. Center for Curriculum Redesign.

Keane, M.T., & Smyth, B. 2020. Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI). arXiv preprint arXiv:2005.13997.

Kolodner, J.L. 1993. Case-Based Reasoning. Elsevier Science & Technology Books.

Lewis, P., et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), 9459–9474.

Mathisen, B.M., Aamodt, A., Bach, K., & Langseth, H. 2019. Learning similarity measures from data. Progress in Artificial Intelligence, 9(11). DOI: 10.1007/s13748-019-00201-2

Pauli, J., Hoffmann, M., & Bergmann, R. 2023. Similarity-Based Retrieval in Process-Oriented Case-Based Reasoning Using Graph Neural Networks and Transfer Learning. The International FLAIRS Conference Proceedings, 36. DOI: 10.32473/flairs.36.133040

Ribeiro, M.T., Singh, S., & Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Shi, W., et al. 2023. REPLUG: Retrieval-Augmented Black-Box Language Models. arXiv preprint arXiv:2301.12652.

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.

Vaswani, A., et al. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.

Zhai, X., et al. 2021. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. Complexity. DOI: 10.1155/2021/8812542.