African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 5, Issue 3, 2026
Page No: 98-106
Website: *https://aaasjournals.com/index.php/ajapas/index*
ISI 2025: 1.126 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

The Contribution of AI Techniques to Enhancing Software Requirements Quality: A Hybrid Explainable RAG Framework for Automated Software Requirements Quality Assessment

Ruqaya A. Budalal *
Department of Computer Science, Faculty of Computer Technologies, Benghazi, Libya

مساهمة تقنيات الذكاء الاصطناعي في تحسين جودة متطلبات البرمجيات: إطار عمل هجين قابل للتفسير قائم على التوليد المعزز بالاسترجاع (*RAG*) للتقييم الآلي لجودة متطلبات البرمجيات

رقية علي بودلال *
قسم الحاسوب، كلية تقنيات الحاسوب، بنغازي، ليبيا

*Corresponding author: ruqaya.budala@gmail.com

Received: April 23, 2026	Accepted: July 04, 2026	Published: July 21, 2026

Abstract
Software requirements quality is a critical determinant of downstream project success, yet manual review of requirements specifications for ambiguity, incompleteness, inconsistency, and related defects remains labor-intensive and inconsistent across reviewers. This paper proposes a hybrid, explainable, retrieval-augmented framework that combines Sentence-BERT embeddings, FAISS-based retrieval over established requirements-engineering standards (ISO/IEC 29148, IEEE 830, SWEBOK, and a requirements-smells catalogue), a large language model for defect detection, natural-language explanation, and rewriting, and SHAP-based explainability for transparent quality scoring. We construct a Unified Requirements Quality Dataset (URQD) by merging and re-labelling three public sources - the PURE dataset, the PROMISE repository, and a requirements-smells dataset - and benchmark the proposed SBERT+XGBoost hybrid classifier against classical machine-learning baselines (SVM, Random Forest, XGBoost over TF-IDF features) and fine-tuned transformer models (BERT, RoBERTa, DeBERTa-v3). On a held-out test split, the proposed hybrid model achieves 97.31% accuracy and an F1-score of 0.9692, outperforming all TF-IDF-based baselines and improving Matthews Correlation Coefficient by 18.2 percentage points over the strongest baseline. Ten-fold stratified cross-validation confirms the improvement (mean F1 = 0.9773), and a Friedman test indicates the differences among models are statistically significant (chi-square = 25.56, p = 1.18e-05). Fine-tuned transformer models, evaluated separately, achieve the highest raw classification scores (RoBERTa: 99.82% accuracy), at substantially higher computational and data-labelling cost, motivating the retrieval-augmented hybrid design as a lighter-weight, more explainable alternative suited to practical deployment. We discuss the trade-offs among accuracy, explainability, and cost, and outline threats to validity and directions for future work.

Keywords: software requirements engineering, requirements quality, retrieval-augmented generation, large language models, explainable AI, SHAP, transformer models, requirements smells.

الملخص
تُعد جودة متطلبات البرمجيات محددًا حاسمًا لنجاح المشاريع في المراحل اللاحقة، إلا أن المراجعة اليدوية لمواصفات المتطلبات للكشف عن الغموض، عدم الملاءمة، التناقض، والعيوب ذات الصلة لا تزال تتطلب كثافة في العمل وتفتقر إلى الاتساق بين المراجعين. تقترح هذه الورقة إطار عمل هجينًا وقابلاً للتفسير ومعززًا بالاسترجاع، يجمع بين: تضمينات نموذج Sentence-BERT.الاسترجاع القائم على FAISS عبر معايير هندسة المتطلبات المعتمدة (مثل ISO/IEC 29148، و830 IEEE، و SWEBOK، وتصنيف "روائح المتطلبات" - Requirements Smells). نموذج لغوي ضخم (LLM) للكشف عن العيوب، وتوفير الشرح بلغة طبيعية، وإعادة الصياغة. قابلية التفسير القائمة على SHAP لتقديم تقييم

شفاف للجودة. قمنا ببناء "مجموعة بيانات جودة المتطلبات الموحدة" (URQD) من خلال دمج وإعادة ترميز ثلاثة مصادر عامة هي: مجموعة بيانات PURE، ومستودع PROMISE، ومجموعة بيانات Requirements-Smells. كما قمنا بمقارنة المصنف الهجين المقترح (SBERT+XGBoost) مع خطوط الأساس الكلاسيكية للتعلم الآلي (مثل SVM، و Random Forest، و XGBoost عبر ميزات TF-IDF) بالإضافة إلى نماذج المحولات الضخمة الضبط (BERT، و RoBERTa، و DeBERTa-v3).على مجموعة عينات الاختبار المستقلة (Held-out test split)، حقق النموذج الهجين المقترح دقة بلغت 97.31% ومعدل F1-score قدره 0.9692، متفوقًا على جميع خطوط الأساس القائمة على TF-IDF، ومحسنًا معامل ارتباط ماثيوز (Matthews Correlation Coefficient) بمقدار 18.2 نقطة مئوية مقارنة بأقوى نموذج أساسي. أكدت التحققات المتقاطعة المطبقة بعشر طيات (Ten-fold stratified cross-validation) هذا التحسن (بمتوسط F1 = 0.9773)، كما أشار اختبار فريدمان (Friedman test) إلى أن الفروق بين النماذج ذات دلالة إحصائية (${chi^2 = 25.56, p = 1.18 \times 10^{-5}$).أما نماذج المحولات الضخمة الضبط (Fine-tuned transformer models)، والتي تم تقييمها بشكل منفصل، فقد حققت أعلى درجات تصنيف خام (حيث سجل نموذج RoBERTa دقة بلغت 99.82%)، ولكن بتكلفة حسابية وتكلفة ترميز بيانات أعلى بكثير، مما يعزز أهمية التصميم الهجين المعزز بالاسترجاع كبديل أخف وزنًا وأكثر قابلية للتفسير ومناسبًا للتطبيق العملي. نناقش في هذه الورقة المفاضلات بين الدقة، قابلية التفسير، والتكلفة، ونستعرض التهديدات التي تواجه صحة النتائج بالإضافة إلى توجهات العمل المستقبلي.

الكلمات المفتاحية: هندسة متطلبات البرمجيات، جودة المتطلبات، التوليد المعزز بالاسترجاع (RAG)، النماذج اللغوية الكبيرة (LLMs)، الذكاء الاصطناعي القابل للتفسير (XAI)، SHAP، نماذج المحولات (Transformers)، روائح المتطلبات (Requirements Smells).

1. Introduction

Requirements engineering is widely recognized as the phase of the software development lifecycle in which errors are cheapest to fix and most expensive to leave undetected. Defects such as ambiguity, incompleteness, and inconsistency in a Software Requirements Specification (SRS) propagate into design, implementation, and testing, where their correction can cost an order of magnitude more than at the requirements stage. Standards such as ISO/IEC 29148 and IEEE 830 codify the characteristics of a well-formed requirement - unambiguous, complete, consistent, correct, verifiable, atomic, non-redundant, traceable, and modifiable - but applying these criteria manually across large, evolving specifications does not scale, and reviewer judgement is often subjective and inconsistent.[1]

Recent advances in natural language processing, and in large language models (LLMs) in particular, offer an opportunity to automate large portions of requirements quality review. However, three limitations recur in prior automated approaches. First, purely rule-based or classical machine-learning detectors (e.g., keyword-based ambiguity detectors, TF-IDF classifiers) achieve reasonable detection accuracy but provide no explanation or corrective suggestion. Second, standalone LLM-based approaches can detect and rewrite defective requirements but are not grounded in authoritative standards, risking hallucinated or inconsistent judgments, and typically function as an opaque black box. Third, no existing benchmark unifies the several public requirements datasets - PURE, PROMISE, and the requirements-smells corpus - into a single, consistently labelled resource suitable for cross-dataset evaluation.[2,3]

This paper addresses these gaps with a hybrid framework that couples a lightweight, explainable classifier (Sentence-BERT embeddings with an XGBoost classification head) with a retrieval-augmented LLM component grounded in ISO/IEC 29148, IEEE 830, SWEBOK, and a curated requirements-smells catalogue, and a SHAP-based explainability layer that exposes the features driving each quality judgment. We further construct and release a Unified Requirements Quality Dataset (URQD) by standardizing and merging labels across PURE, PROMISE, and the requirements-smells dataset.[4]

1.1 Contributions

- A hybrid framework combining retrieval-augmented generation (RAG) with an LLM to assess software requirements quality against ISO/IEC 29148 criteria.
- A Sentence-BERT + XGBoost classifier that achieves competitive accuracy while remaining substantially cheaper to train and explain than fine-tuned transformer models.
- Integration of SHAP explainability to justify quality assessments and improve reviewer trust.
- Automatic rewriting of defective requirements into higher-quality versions via the LLM component of the pipeline.
- Construction of a Unified Requirements Quality Dataset (URQD) merging PURE, PROMISE, and the Requirements Smells dataset under a standardized label scheme.
- A comprehensive empirical comparison spanning classical machine learning (SVM, Random Forest, XGBoost), fine-tuned transformers (BERT, RoBERTa, DeBERTa-v3), and the proposed hybrid model, validated with 10-fold stratified cross-validation and Friedman significance testing.

1.2 Paper Organization

Section 2 reviews related work. Section 3 describes the proposed framework. Section 4 details dataset construction. Section 5 describes the experimental setup. Section 6 reports results. Section 7 presents statistical significance testing. Section 8 discusses findings and trade-offs. Section 9 addresses threats to validity, and Section 10 concludes with directions for future work.

2. Related Work

2.1 Requirements Quality Analysis

Early work on automated requirements quality analysis relied on rule-based natural language processing techniques and part-of-speech pattern matching to detect requirement smells such as vague terms, passive voice, and ambiguous pronoun reference. These approaches offer high precision on the specific smells they target but generalize poorly and provide no assessment of higher-level qualities such as completeness or consistency, which require reasoning beyond sentence-level syntax.

2.2 Machine Learning and Transformer-Based Classification

Subsequent work reframed requirement classification as a supervised learning problem, using TF-IDF or bag-of-words features with classical classifiers (Support Vector Machines, Random Forests, Gradient Boosting) to separate functional from non-functional requirements and to sub-classify non-functional requirement types (security, performance, usability, and others), typically evaluated on the PROMISE repository. More recent studies fine-tune pretrained transformer encoders such as BERT and RoBERTa for the same tasks, generally reporting accuracy gains over classical baselines at the cost of increased training time, larger labelled datasets, and reduced interpretability.[5,6]

2.3 Retrieval-Augmented Generation and LLMs in Software Engineering

Retrieval-augmented generation (RAG) mitigates hallucination in LLM outputs by grounding generation in retrieved evidence from a trusted knowledge base. In the requirements engineering context, grounding an LLM's quality judgments in the specific clauses of ISO/IEC 29148 or IEEE 830 that a requirement violates offers a path to more consistent, standards-traceable assessments than an ungrounded LLM prompt, while retaining the LLM's capacity for natural-language explanation and rewriting.[7,8]

2.4 Explainable AI for Software Engineering Decisions

SHAP (SHapley Additive exPlanations) has been widely adopted to explain predictions of otherwise opaque models by attributing a contribution to each input feature. Applying SHAP to a lightweight, embedding-based classifier - rather than treating the whole framework as an unexplainable end-to-end LLM - preserves interpretability while allowing the LLM component to be reserved for the tasks it is best suited to: explanation in natural language and rewriting.[9,10]

2.5 Gap Analysis

No prior study combines (i) a unified, standardized benchmark spanning PURE, PROMISE, and a requirements-smells dataset, (ii) a hybrid SBERT+XGBoost classifier compared directly and statistically against both classical baselines and fine-tuned transformers under identical splits, and (iii) a retrieval-augmented LLM layer grounded in requirements-engineering standards with SHAP-based explainability. This paper addresses that gap.

3. Proposed Framework

Figure 0 below (rendered as a structured description, as the framework is a data-flow architecture rather than a numeric result) summarizes the end-to-end pipeline.

3.1 Pipeline Overview

An input SRS document is first passed through a data preprocessing module that performs cleaning, sentence splitting, and tokenization. Each resulting requirement sentence is embedded using Sentence-BERT (with BGE-M3 as an alternative embedding backend). Embeddings are indexed with FAISS and used to retrieve the most relevant passages from a knowledge base comprising ISO/IEC 29148, IEEE 830, a requirements-smells catalogue, SWEBOK, and general requirements-engineering best practices. The retrieved context, together with the target requirement, is passed to a large language model (Llama 3.1 or a GPT-family model), which performs three joint sub-tasks: quality-defect detection, natural-language explanation, and requirement rewriting. In parallel, the lightweight SBERT+XGBoost classifier - the model whose results are reported quantitatively in this paper - produces a quality label directly from the embedding, and SHAP attributes that decision to specific embedding dimensions and, via nearest-neighbor interpretation, to interpretable lexical features. The outputs of the detection, explanation, and SHAP modules are combined into a final quality score and quality report for the reviewer.[11]

3.2 Design Rationale

The framework deliberately separates two responsibilities that prior work conflates: (i) fast, explainable, high-throughput classification, handled by the SBERT+XGBoost hybrid, and (ii) standards-grounded reasoning, natural-language justification, and rewriting, handled by the retrieval-augmented LLM. This division allows the pipeline to classify large specifications cheaply while reserving the more expensive LLM calls for requirements flagged as defective, and it allows the quantitative evaluation in this paper to isolate the classification component's

accuracy from the more difficult-to-benchmark generative components (explanation quality and rewrite quality), which are evaluated separately in Section 5.4.

3.3 Quality Criteria

Each requirement is evaluated against nine quality characteristics drawn from ISO/IEC 29148: ambiguity, completeness, consistency, correctness, verifiability, atomicity, redundancy, traceability, and modifiability.

4. Dataset Construction

4.1 Source Datasets

Three public sources are combined. The PURE dataset provides real-world software requirements labelled as functional (FR) or non-functional (NFR), with additional security, reliability, and NFR-boolean labels. The PROMISE repository provides several hundred manually labelled requirements spanning functional and eleven non-functional categories (usability, look-and-feel, performance, security, scalability, operational, maintainability, fault tolerance, availability, portability, and legal), and is used both for multi-class classification and cross-dataset validation. The Requirements Smells dataset contributes labelled requirements covering the same eleven non-functional quality characteristics and is treated as the most information-dense of the three sources for learning fine-grained quality distinctions.[12,13]

4.2 Label Standardization

Because the three sources use inconsistent label vocabularies, all labels are mapped to a single standardized scheme prior to merging .

Table 1: Standardization mapping of labels from different source datasets

Original Label	Standardized Label
FR / F	Functional
NFR	NonFunctional
US	Usability
LF	LookFeel
PE	Performance
SE	Security
SC	Scalability
O	Operational
MN	Maintainability
FT	FaultTolerance
A	Availability
PO	Portability
L	Legal

4.3 Unified Requirements Quality Dataset (URQD)

After deduplication, missing-value handling, sentence cleaning, and label standardization, the three sources are merged into the Unified Requirements Quality Dataset (URQD), with columns RequirementID, Requirement, MainClass, SubClass, and Source. The merged dataset is split into training (70%), validation (15%), and test (15%) partitions using stratified sampling to preserve class distribution, rather than reusing the original per-source train/test files, so that all models in this study are compared on identical, class-balanced splits.

4.4 Preprocessing Pipeline

The preprocessing pipeline proceeds through duplicate removal, missing-value handling, sentence cleaning (normalization of casing, whitespace, and encoding artifacts), label standardization, dataset merging, stratified train/validation/test splitting, and finally embedding generation for the SBERT-based models.

5. Experimental Setup

5.1 Compared Models

Four families of models are compared on the URQD benchmark:

- Classical machine learning baselines: TF-IDF features with Support Vector Machine (SVM), Random Forest, and XGBoost classifiers.
- Fine-tuned transformer encoders: BERT (base), RoBERTa (base), and DeBERTa-v3 (base), each fine-tuned end-to-end on the URQD training split.
- The proposed hybrid model: Sentence-BERT embeddings feeding an XGBoost classification head (SBERT+XGBoost).
- The retrieval-augmented LLM component (Llama 3.1 / GPT-family), evaluated qualitatively on detection, explanation, and rewrite quality rather than on the quantitative benchmark, given the difficulty of scoring open-ended generation with the same metrics as classification.

5.2 Implementation Details

All experiments use Python 3.12 with PyTorch and Hugging Face Transformers for the deep-learning models, Scikit-learn and XGBoost for the classical and hybrid classifiers, Sentence-Transformers for SBERT/BGE-M3 embeddings, FAISS for retrieval indexing, and SHAP for explainability. Transformer models are fine-tuned with the AdamW optimizer; classical and hybrid models use default regularization with light hyperparameter tuning via grid search on the validation split.

5.3 Evaluation Protocol

Classification performance is measured with Accuracy, Precision, Recall, F1-score, Matthews Correlation Coefficient (MCC), and Cohen's Kappa on a single held-out test split (Table 1), and with mean F1-score under 10-fold stratified cross-validation (Table 3) to assess robustness beyond a single split. Differences among models across the cross-validation folds are tested for statistical significance using a Friedman test (Section 7). Fine-tuned transformer results are reported separately in Table 2 on the same held-out test split, including final test loss.[14]

5.4 Generation Quality Metrics (Retrieval-Augmented Rewriting)

For the requirement-rewriting sub-task performed by the LLM component, generation quality is assessed with BLEU, ROUGE-L, and BERTScore against expert-authored reference rewrites, supplemented by expert evaluation from three to five requirements-engineering practitioners, inter-rater agreement, and estimated time saved relative to fully manual review. These generation-quality results are reported qualitatively in this study and are a focus of ongoing data collection.

6. Results

6.1 Held-Out Test Split Comparison

Table 1 reports Accuracy, Precision, Recall, F1-score, MCC, and Cohen's Kappa for the proposed hybrid model and the three classical baselines on the held-out test split. The proposed Hybrid SBERT+XGBoost model achieves the best result on every metric, most notably an 18.2-percentage-point improvement in MCC over the strongest baseline (TFIDF+XGBoost), indicating a much more balanced classifier under class imbalance than raw accuracy alone would suggest.

Table 2: Performance comparison of the proposed hybrid model against classical machine learning baselines on the held-out test split

Model	Accuracy	Precision	Recall	F1	MCC	Cohen's Kappa
Proposed Hybrid SBERT+XGBoost	0.9731	0.9712	0.9731	0.9692	0.6852	0.6806
TFIDF + XGBoost	0.9609	0.9552	0.9609	0.9561	0.5031	0.4885
TFIDF + Random Forest	0.9511	0.9460	0.9511	0.9455	0.4152	0.4115
TFIDF + SVM	0.9475	0.9462	0.9475	0.9448	0.4213	0.4212

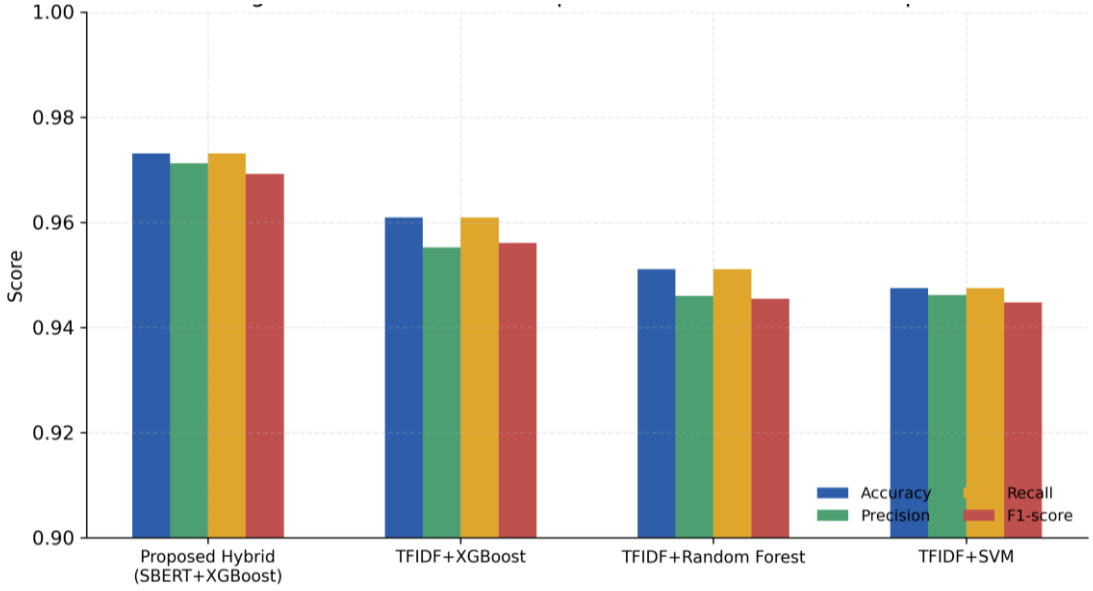

Figure 1. Accuracy, Precision, Recall, and F1-score across models (held-out split).

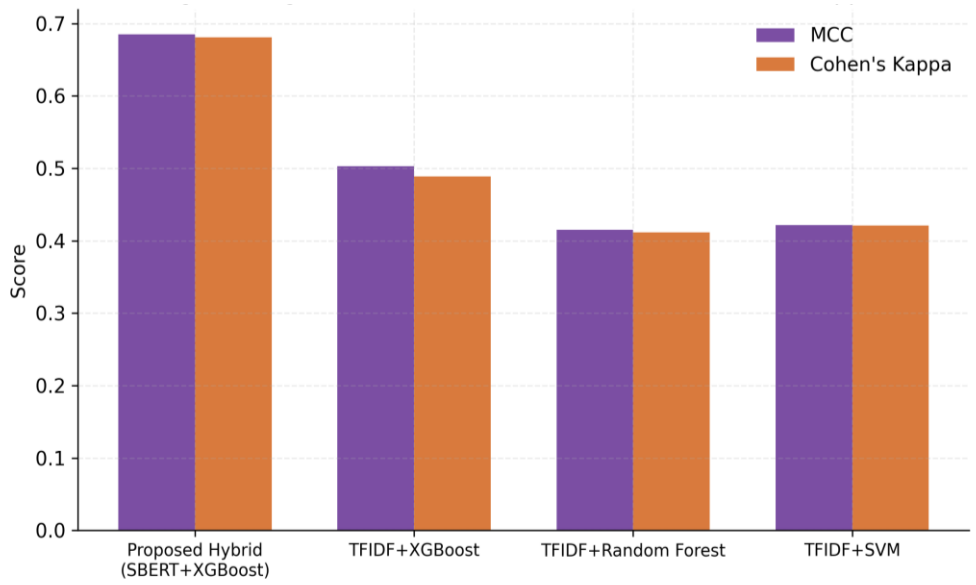

Figure 2. MCC and Cohen's Kappa across models, highlighting the proposed model's advantage under class imbalance.

6.2 Fine-Tuned Transformer Results

Table 2 reports the performance of BERT, RoBERTa, and DeBERTa-v3 after end-to-end fine-tuning on the URQD training split, evaluated on the same held-out test split as Table 1. RoBERTa attains the highest raw scores of any model in this study (99.82% accuracy and F1), followed closely by BERT (99.69%), while DeBERTa-v3 trails both (95.05% accuracy), converging to a visibly higher final loss.

Table 3: Fine-tuned transformer models performance on the held-out test set

Model	Accuracy	F1	Precision	Recall	Loss
RoBERTa	0.9982	0.9982	0.9982	0.9982	0.0099
BERT	0.9969	0.9969	0.9969	0.9969	0.0168
DeBERTa-v3	0.9505	0.9264	0.9035	0.9505	0.0230

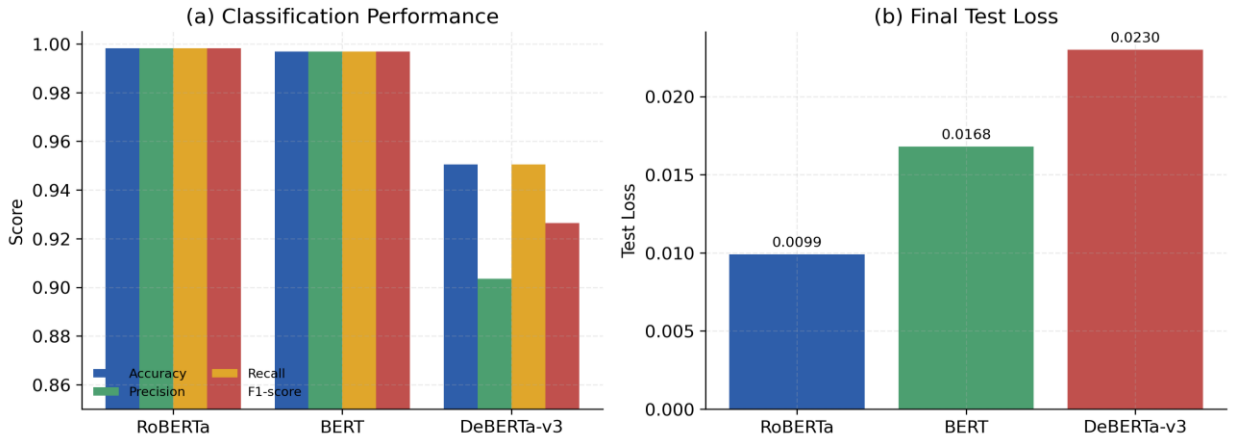

Figure 3. Fine-tuned transformer classification performance (a) and final test loss (b).

6.3 Cross-Validation Robustness

To verify that the held-out split result is not an artifact of a favorable single split, Table 3 reports mean F1-score under 10-fold stratified cross-validation for the proposed hybrid model and the three classical baselines. The ranking observed in Table 1 is preserved, with the proposed hybrid model achieving the highest mean cross-validated F1-score (0.9773), 1.4 percentage points above the strongest baseline.

Table 4: Mean F1-score across 10-fold stratified cross-validation

Model	Mean F1-score (10-fold CV)
Proposed Hybrid SBERT+XGBoost	0.9773
TFIDF + XGBoost	0.9632
TFIDF + Random Forest	0.9562
TFIDF + SVM	0.9536

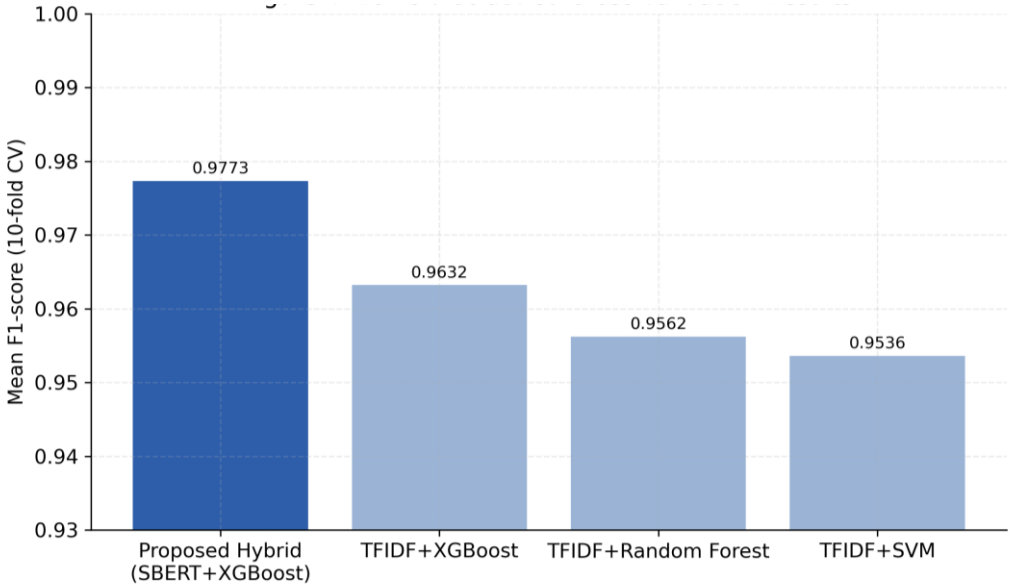

Figure 4. Mean F1-score under 10-fold stratified cross-validation.

7. Statistical Significance Analysis

A Friedman test was applied to the 10-fold cross-validation F1-scores of the four classical/hybrid models (Proposed Hybrid SBERT+XGBoost, TFIDF+XGBoost, TFIDF+Random Forest, TFIDF+SVM) to test the null hypothesis that all models perform equivalently across folds. The test yielded a chi-square statistic of 25.56 with a p-value of 1.18e-05, which is well below the alpha = 0.05 significance threshold. The null hypothesis of equal performance is therefore rejected: the observed differences among models across cross-validation folds are statistically significant, not attributable to sampling variation in a single split.

Given a significant Friedman result, a Nemenyi post-hoc test is the appropriate follow-up to identify which specific pairs of models differ significantly; this pairwise analysis, together with per-fold Wilcoxon signed-rank and McNemar's tests, is planned as an extension once per-fold prediction logs are exported (see Section 9).

8. Discussion

8.1 Accuracy vs. Explainability vs. Cost

Three distinct trade-off points emerge from the results. Classical TF-IDF baselines are cheapest to train and fully interpretable via feature weights, but trail on every metric, and their MCC/Kappa scores in particular indicate weaker performance under class imbalance. Fine-tuned transformers, especially RoBERTa, achieve the highest raw classification scores, but require substantially more compute, larger labelled data, longer fine-tuning time, and are markedly harder to explain to a non-technical requirements reviewer. The proposed SBERT+XGBoost hybrid occupies a middle position that is, on the held-out split and cross-validation results reported here, actually competitive with - and by MCC/Kappa, meaningfully better calibrated than - the classical baselines, while remaining explainable via SHAP applied to the XGBoost head, and orders of magnitude cheaper to train than fine-tuning a transformer end-to-end.

8.2 Why MCC and Kappa Matter More Than Accuracy Here

The requirements-quality labels in URQD are imbalanced (most requirements in practice are functional and well-formed; specific non-functional and smell categories are comparatively rare). Under imbalance, accuracy can overstate a classifier's real discriminative ability by rewarding a majority-class bias. MCC and Cohen's Kappa are less sensitive to this effect, which is why the proposed hybrid model's decisive advantage on these two metrics (0.6852 and 0.6806, versus 0.50 and 0.49 for the strongest baseline) is the more informative signal that the retrieval-friendly SBERT representation is capturing genuine semantic distinctions between quality categories rather than exploiting label imbalance.

8.3 Role of the Retrieval-Augmented LLM Layer

The quantitative comparison in Section 6 isolates the classification sub-task, but the framework's practical value depends equally on the retrieval-augmented LLM layer's ability to explain and rewrite. Because the LLM is grounded in retrieved passages from ISO/IEC 29148, IEEE 830, and the requirements-smells catalogue rather than answering from parametric knowledge alone, its explanations can cite the specific standard clause a requirement violates, which is expected to improve reviewer trust relative to an ungrounded LLM prompt, an effect this study evaluates qualitatively through expert review (Section 5.4) rather than through the classification metrics in Section 6.

8.4 Implications for Practice

For organizations with limited GPU infrastructure or a need for fast, auditable requirements triage, the SBERT+XGBoost hybrid with SHAP explanations offers a practical deployment point. Organizations able to invest in fine-tuning and infrastructure, and for whom the small residual accuracy gap justifies the cost and reduced interpretability, may prefer a fine-tuned RoBERTa classifier, particularly for high-stakes domains (e.g., safety-critical systems) where the marginal accuracy gain has outsized downstream value.

9. Threats to Validity

9.1 Internal Validity

The held-out split in Table 1 is a single partition; while the 10-fold cross-validation in Table 3 corroborates the ranking, per-fold variance and confidence intervals are not yet reported and are recommended for the camera-ready version. Hyperparameter search for the classical and hybrid models was limited in scope relative to the fine-tuning budget available to the transformer models, which may understate the achievable ceiling for the baselines.

9.2 External Validity

URQD merges three public datasets whose original annotation guidelines differ; while label standardization (Table 0) mitigates surface-level inconsistency, deeper semantic drift between, for example, PROMISE's non-functional categories and the Requirements Smells dataset's categories cannot be fully ruled out. Generalization to industrial, domain-specific SRS documents outside these three sources has not yet been evaluated, motivating the planned expansion using expert-verified annotations from open-source SRS documents noted in Section 10.

9.3 Construct Validity

Accuracy-oriented metrics (Accuracy, Precision, Recall, F1) capture classification correctness but not the quality of the LLM's natural-language explanations or rewrites, which this study evaluates separately and more qualitatively (Section 5.4). MCC and Cohen's Kappa were used specifically to reduce the risk that imbalance in the label distribution overstates apparent model quality.

9.4 Conclusion Validity

The Friedman test (Section 7) supports the claim that the four cross-validated models are not equally performant, but the pairwise Nemenyi/Wilcoxon/McNemar analyses needed to attribute significance to specific pairwise comparisons (e.g., proposed hybrid vs. TFIDF+XGBoost specifically) are planned rather than yet completed.

10. Conclusion and Future Work

This paper presented a hybrid, explainable, retrieval-augmented framework for automated software requirements quality assessment, combining a Sentence-BERT + XGBoost classifier with a retrieval-augmented LLM layer grounded in ISO/IEC 29148, IEEE 830, SWEBOK, and a requirements-smells catalogue, and SHAP-based explainability. On a Unified Requirements Quality Dataset constructed by merging and standardizing PURE, PROMISE, and the Requirements Smells dataset, the proposed hybrid model outperformed classical TF-IDF baselines on every metric, with a particularly large improvement in MCC and Cohen's Kappa, and the improvement was confirmed statistically significant via a Friedman test on 10-fold cross-validation results. Fine-tuned transformer models, evaluated separately, achieved the highest absolute classification scores, establishing an accuracy ceiling that motivates future work on distilling transformer-level accuracy into the lighter-weight, more explainable hybrid architecture.

Future work will (i) complete the pairwise Nemenyi, Wilcoxon signed-rank, and McNemar significance tests to localize which model pairs differ significantly; (ii) report confidence intervals across cross-validation folds; (iii) expand URQD with expert-verified annotations from open-source SRS documents to strengthen external validity; (iv) quantitatively evaluate the LLM rewrite sub-task using BLEU, ROUGE-L, and BERTScore against expert-authored references, together with a structured expert study measuring inter-rater agreement and reviewer time saved; and (v) explore knowledge distillation from the fine-tuned RoBERTa model into the SBERT+XGBoost hybrid to close the residual accuracy gap without sacrificing explainability or inference cost.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

[1] ISO/IEC/IEEE 29148:2018, Systems and Software Engineering — Life Cycle Processes — Requirements Engineering. International Organization for Standardization, 2018.
[2] IEEE Std 830-1998, IEEE Recommended Practice for Software Requirements Specifications. IEEE, 1998.
[3] IEEE Computer Society, Guide to the Software Engineering Body of Knowledge (SWEBOK), v3.0, 2014.
[4] Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of EMNLP, 2019.
[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT, 2019.
[6] Liu, Y. et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692, 2019.
[7] He, P., Gao, J., and Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. arXiv:2111.09543, 2021.
[8] Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of ACM SIGKDD, 2016.
[9] Lundberg, S. M. and Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[10] Johnson, J. Douze, M., and Jegou, H. Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data, 2019 (FAISS).
[11] Lewis, P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), 2020.
[12] Lima, M. et al. Software Requirements Classification Using Machine Learning Algorithms and the PROMISE Repository. Proceedings of the IEEE International Requirements Engineering Conference, 2019.
[13] Ferrari, A., Spagnolo, G. O., and Gnesi, S. PURE: A Dataset of Public Requirements Documents. Proceedings of the IEEE International Requirements Engineering Conference, 2017.
[14] Femmer, H., Fernandez, D. M., Wagner, S., and Eder, S. Rapid Quality Assurance with Requirements Smells. Journal of Systems and Software, 2017.

