# A Systematic Evaluation Framework for LLM-Generated Software Quality Artifacts

Priscilla de Souza Silva[1]

[1]University of Pernambuco, Street Benfica, 455 – 50720-001 – Recife – Pernambuco – Brazil

**Abstract**

Large Language Models are increasingly used to generate software quality artifacts such as test cases and requirements-related documents. However, these artifacts exhibit high variability, probabilistic behavior, and context sensitivity, challenging traditional quality assessment criteria originally designed for human-produced artifacts. Their evaluation remains largely subjective and poorly standardized, and the literature lacks a consolidated and reproducible evaluation framework tailored to LLM-generated artifacts. This PhD research investigates the design and empirical validation of a systematic evaluation framework for LLM-generated software quality artifacts. The framework aims to integrate and operationalize evaluation dimensions, criteria, and metrics into a structured and reproducible assessment protocol. The research combines a Systematic Literature Review, a practitioner survey, and controlled experimentation to ground the framework both theoretically and empirically. The study is currently at an early stage, with the SLR protocol defined and the survey instrument under development.

**Keywords**

Large Language Models, Software Quality Artifacts, Evaluation

## 1. Introduction

Despite the growing adoption of Large Language Models (LLMs) in the generation of software quality artifacts, there is currently no systematic and standardized evaluation approach tailored to their specific characteristics. Unlike traditionally produced artifacts, LLM-generated artifacts exhibit probabilistic behavior, variability across prompts, context sensitivity, and potential hallucinations [1]. These properties challenge conventional quality assessment criteria designed for human-produced artifacts and raise concerns about reliability and reproducibility.

For example, a test case generated by an LLM may appear syntactically correct while containing ambiguous preconditions, implicit assumptions, or inconsistencies with system requirements. Traditional evaluation criteria do not explicitly account for such LLM-specific risks, which makes quality assessment largely subjective and dependent on individual evaluator judgment. This gap highlights the need for a structured and empirically grounded evaluation framework specifically designed for LLM-generated software quality artifacts.

Existing studies partially address this issue. Yang et al. (2024) compare different LLMs for test case generation, while other works analyze dimensions such as clarity, completeness, and consistency [3]. Additional approaches rely on semantic similarity measures to estimate artifact quality [4]. However, these studies adopt heterogeneous criteria and procedures that are often insufficiently specified or not reproducible. As a result, evaluation practices remain fragmented, limiting comparability across studies and hindering methodological consolidation.

To address this gap, this PhD research investigates how to design and empirically validate a systematic evaluation framework for software quality artifacts generated by LLMs. The research strategy combines: (i) a Systematic Literature Review to identify reported evaluation dimensions, criteria, metrics, and procedures; (ii) a practitioner survey to capture current evaluation practices, challenges, and expectations; and (iii) the design and controlled experimental validation of a structured evaluation framework.

*Joint Proceedings of REFSQ-2026 Workshops, Doctoral Symposium, Posters Tools Track, and Education and Training Track. Co-located with REFSQ 2026. Poznan, Poland, March 23-26, 2026*

pss@ecomp.poli.br (P. d. S. Silva)

This paper presents the current status of the ongoing PhD project. The SLR protocol has been defined and preliminary searches conducted, while the survey instrument has been designed and is under refinement. The expected outcome is a validated evaluation framework that reduces evaluator subjectivity and supports more reliable and reproducible assessment of LLM-generated software quality artifacts.

## 2. Background

Many of the artifacts addressed in this research, such as test cases, acceptance criteria, and requirements-related documentation, are directly connected to Requirements Engineering practices, reinforcing the relevance of this work to the REFSQ community and to research on requirements-driven quality assurance.

Software quality artifacts encompass documents and items produced to support verification, validation, and testing activities, such as test cases, acceptance criteria, test plans, defect reports, and requirements documentation [5]. Their quality is typically assessed based on properties such as clarity, completeness, consistency, and verifiability, which are widely discussed in the software engineering literature [6].

Large Language Models are advanced artificial intelligence models designed to understand and generate human-like text based on large volumes of training data. They rely on deep learning techniques, particularly neural networks, to analyze language patterns and produce coherent and contextually relevant text. Prominent examples include OpenAI's GPT, Google's Gemini, and Meta's LLaMA, which are widely used for content creation, cognitive task support, and conversational interaction [7].

However, existing studies focus on isolated metrics, tasks, or artifact types, rather than proposing a structured and reproducible evaluation approach across software quality assurance artifacts. This gap motivates the research proposed in this PhD work.

## 3. Related Work

One of the most relevant studies in this research area is that of Molina and Gorla (2024), which investigates the use of LLMs to automate test oracles. The authors show that, although the models can produce useful assertions, they still present semantic inconsistencies and structural errors that affect the reliability of the artifacts. Another relevant work is by Dahiya et al. (2024), which examines the capability of ChatGPT to predict the testability of requirements. The study reveals that LLMs are able to identify relevant requirement characteristics, but also generate ambiguous interpretations and imprecise analyses, highlighting limitations when supporting early software quality activities.

In the field of automated test generation, Guo et al. (2025) evaluate the quality of test suites produced by generative models and hybrid approaches. The authors analyze metrics such as diversity, effectiveness, and completeness of the tests and conclude that, although promising, the generated artifacts still require refinement before being used in real development environments.

Although recent studies evaluate specific dimensions of LLM-generated artifacts, such as semantic similarity, syntactic validity, or functional coverage, these evaluations are typically task-specific and lack a unified theoretical foundation. Furthermore, most studies focus on benchmarking model performance rather than systematically structuring the evaluation of the generated artifacts themselves. There is limited work on consolidating heterogeneous criteria into a coherent, operationalized framework that supports reproducibility and cross-study comparability.

## 4. Methodology

The adopted methodology combines a Systematic Literature Review (SLR) and an exploratory survey with the goal of grounding the development and subsequent validation of the proposed evaluation framework. The SLR is used to map research gaps, evaluation criteria, and practices reported in the

literature, while the survey investigates perceptions, challenges, and expectations of software quality professionals. Together, these stages guide the definition of the following research questions:

- **RQ1:** Which evaluation dimensions, criteria, and metrics are reported in the literature for assessing software quality artifacts generated by LLMs?
- **RQ2:** How do professionals currently evaluate LLM-generated software quality artifacts, and what limitations are observed?
- **RQ3:** What requirements and design principles should guide the construction of a systematic and reproducible evaluation framework?
- **RQ4:** How can these principles be operationalized into a structured evaluation protocol?
- **RQ5:** To what extent is the proposed framework perceived as useful, applicable, and reliable by professionals dealing with LLM-generated software quality artifacts?

## 4.1. Systematic Literature Review (SLR)

The SLR, conducted in accordance with the guidelines proposed by Kitchenham and Charters (2007), aims to map evaluation criteria, metrics, quality dimensions, and challenges reported in studies that address the assessment of software quality artifacts generated with LLM support. The SLR process comprises the phases of planning, search execution, study selection, data extraction, and data synthesis.

Automatic searches will be conducted in the IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, Scopus, and Web of Science databases using a search string composed of terms related to the keywords *software quality*, *software artifact*, *evaluation*, and *LLMs*. The searches will be restricted to peer-reviewed articles written in English or Portuguese and published between 2017 and 2026. The automatic searches will be complemented by manual searches in proceedings of relevant national conferences in the field, such as SBQS, SBES, SBSI, and ENIAC.

Studies are included if they present evaluation approaches for software quality artifacts generated or supported by LLMs, and excluded if they do not evaluate artifacts, focus solely on model performance, or address LLM use outside software engineering. Study selection follows a three-stage screening process (title/abstract, introduction/conclusion, and full text), followed by methodological quality assessment and data extraction to support descriptive and comparative synthesis.

This SLR primarily addresses RQ1 by identifying evaluation dimensions, criteria, and metrics reported in the literature. Additionally, it partially informs RQ3 by extracting design principles and methodological gaps that guide the construction of the proposed framework.

## 4.2. Survey

The survey addresses RQ2 by investigating how practitioners currently evaluate LLM-generated software quality artifacts and which limitations they perceive. Furthermore, it contributes to RQ3 by identifying practitioner-driven requirements and expectations for a systematic evaluation framework.

The **questionnaire design** followed a systematic process guided by the research questions and grounded in evidence from the literature. Initially, relevant theoretical dimensions were identified based on studies on LLM-based artifact generation and evaluation [2] and on classical software quality literature [6]. Based on these dimensions, the questionnaire was structured into four thematic sections, described as follows:

- **Section A — Participant Profile:** collects information about professional role, experience in software development and quality activities, and familiarity with LLM usage, in order to contextualize responses from subsequent sections.
- **Section B — Evaluation of LLM-Generated Artifacts:** investigates perceptions of artifact quality considering criteria such as clarity, completeness, consistency, and contextual adequacy.
- **Section C — Challenges and Limitations:** addresses difficulties, limitations, and risks encountered when evaluating LLM-generated artifacts, including issues that may not be evident in superficial analyses.

- **Section D — Expectations for a Systematic Approach:** explores desired characteristics, perceived usefulness, and professional expectations regarding a structured approach for evaluating such artifacts.

The instrument will be reviewed by software quality experts and refined after a pilot study in order to ensure clarity, relevance, and alignment with the research questions.

**Data collection** is planned to reach professionals with experience in using or evaluating LLM-generated artifacts in software quality activities. A non-probabilistic convenience sampling strategy will be adopted, which is appropriate for exploratory studies, and recruitment is planned through professional networks and technical communities. The questionnaire will be administered electronically, asynchronously, and anonymously, with an estimated completion time of 8 to 12 minutes. All participants will receive an informed consent statement explaining the academic purpose of the study, data anonymization procedures, and the voluntary nature of participation, with the option to withdraw at any time.

**Data analysis** is planned to combine quantitative and qualitative techniques according to the nature of the collected responses. Closed-ended items will be analyzed using descriptive statistics, including frequency, mean, median, and standard deviation, as well as trend analysis for Likert-scale items. Open-ended responses will be examined through thematic analysis involving open coding and categorization to identify recurring themes. Finally, quantitative and qualitative results will be integrated through triangulation in order to strengthen interpretive consistency and conclusion robustness.

As required by institutional regulations, ethical approval will be obtained from the appropriate Research Ethics Committee prior to survey deployment.

### 4.3. Framework Design and Experimental Validation

The development of the proposed evaluation framework follows a Design Science Research approach, emphasizing iterative refinement grounded in evidence from the SLR and survey. This phase addresses RQ4 by operationalizing identified design principles into measurable evaluation components. It involves the formal specification of evaluation dimensions, definition of operational metrics, development of rating scales, and construction of a structured evaluation protocol. The objective is to transform abstract quality criteria into reproducible assessment procedures.

The framework will be empirically validated through controlled experimental studies, addressing RQ5. These experiments will compare LLM-generated and human-produced software quality artifacts across selected quality dimensions. Multiple evaluators will independently apply the proposed protocol to assess inter-rater reliability, discriminative capacity, reproducibility, and perceived usefulness in practical contexts.

Statistical analysis will be employed to evaluate agreement levels, variance patterns, and consistency of results, ensuring methodological robustness and empirical grounding of the framework.

## 5. Proposed Solution

This section presents the proposed evaluation framework. First, we describe its structural components and conceptual foundations. Then, we detail the evaluation protocol that operationalizes the framework in practice.

### 5.1. Framework Structure

The proposed evaluation framework is systematic and empirically grounded, designed to assess software quality artifacts generated by LLMs. Although the long-term objective is to support multiple artifact types, the initial validation focuses on test cases and acceptance criteria to ensure methodological depth and feasibility.

The framework integrates evidence from the Systematic Literature Review and practitioner insights from the survey to support structured and reproducible human-centered evaluation. It addresses three core aspects of artifact assessment, what to evaluate, how to evaluate, and how to interpret results:

- **C1 — Evaluation Criteria and Dimensions:** defines quality attributes such as clarity, completeness, consistency, semantic accuracy, and contextual adequacy, operationalized through structured descriptors and rating scales.
- **C2 — Evaluation Process:** specifies a standardized procedure for applying the criteria, including artifact preparation, criteria application, and structured documentation of results.
- **C3 — Interpretation Guidelines:** provides guidance for consolidating results, comparing artifacts, and identifying recurring limitations to support consistent decision-making.

The novelty of this research lies in consolidating heterogeneous evaluation criteria into a unified, operationalized, and empirically validated framework. The doctoral contribution includes:

1. **Formalize evaluation dimensions into a structured taxonomy**.
2. **Operationalize abstract criteria into measurable constructs**.
3. **Empirically validate reliability and discriminative capacity**.
4. **Provide reproducible evaluation procedures grounded in evidence**.

### 5.2. Evaluation Protocol

The evaluation protocol operationalizes the proposed framework by defining a structured and reproducible workflow for human-centered assessment of LLM-generated software quality artifacts. It ensures consistent application of evaluation criteria and supports subsequent empirical validation.

The protocol comprises five stages:

1. **Planning**: definition of the evaluation objective, selection of artifacts, and identification of relevant quality dimensions and criteria.
2. **Artifact Preparation**: standardization of artifacts and contextual information to ensure equivalent evaluation conditions.
3. **Criteria Application**: systematic assessment using the defined dimensions, rating scales, and qualitative descriptors, performed independently by evaluators when applicable.
4. **Result Recording**: structured documentation of scores, evidence, and justifications to enable traceability and transparency.
5. **Consolidation and Analysis**: synthesis and comparison of results to identify recurring patterns, limitations, and quality trends.

By formalizing these stages, the protocol reduces evaluator subjectivity, enhances reproducibility, and provides the methodological basis for assessing reliability and discriminative capacity in subsequent experimental validation.

## 6. Research Plan and PhD Roadmap

The PhD research is structured in sequential and iterative phases, planned to be completed by August 2029:

- **In 2025**, preliminary searches for the Systematic Literature Review (SLR) were initiated and an initial survey instrument was designed.
- **In 2026**, the SLR will be finalized and the survey will be updated and applied to software quality professionals.
- **In 2027**, a systematic evaluation framework will be designed and operationalized based on the SLR and survey findings.
- **In 2028**, the framework will be empirically validated through controlled experiments comparing LLM-generated and human-produced software quality artifacts.
- **In 2029**, the framework will be refined and consolidated in the final thesis.

## 7. Status and Risks

At the time of writing, the SLR protocol has been defined and initial search execution has been performed. The survey instrument has been designed and is undergoing refinement prior to deployment. The framework design is currently at a conceptual stage.

Potential risks include:

- Insufficient consensus in the literature regarding evaluation dimensions.
- Limited practitioner participation in the survey.
- Variability in human evaluator interpretation during experimental validation.
- Challenges in defining universally applicable criteria across diverse artifact types.

Mitigation strategies include expert review, pilot testing, triangulation of data sources, and iterative refinement of the framework based on empirical findings.

## 8. Final Considerations and Next Steps

This PhD research aims to establish methodological foundations for evaluating software quality artifacts generated by Large Language Models. By consolidating evidence from the literature and practitioner perspectives, the study seeks to reduce subjectivity and fragmentation in current evaluation practices and to provide a structured basis for reproducible assessment.

The expected doctoral contributions include: (i) a theoretically grounded taxonomy of evaluation dimensions for LLM-generated software quality artifacts; (ii) operationalized criteria and measurable constructs tailored to LLM-specific characteristics; (iii) a structured human-centered evaluation protocol; and (iv) empirical evidence regarding reliability, discriminative capacity, and practical applicability of the proposed framework.

The next steps include completing the Systematic Literature Review, conducting the practitioner survey, and integrating both evidence sources to refine and operationalize the framework prior to controlled experimental validation.

## 9. Appendices

The Survey Questionnaire is available for consultation at: https://drive.google.com/drive/folders/1f1mwCvVbnZc-Ry_GshsBI8Z0n0FMexc-?usp=drive_link

## Declaration on Generative AI

During the preparation of this work, the author used **ChatGPT** to assist with **grammar and spelling checking, paraphrasing, rewording, and text translation**. All AI-generated suggestions were carefully reviewed, edited, and validated by the author, who takes full responsibility for the final content of this publication.

## References


[1] X. Lu, M. Tufano, D. Drain, J. Humble, S. Zhang, S. K. Lahiri, C. Le Goues, S. Chandra, Generating test cases with large language models: Opportunities and pitfalls, in: Proceedings of the 45th International Conference on Software Engineering (ICSE), IEEE/ACM, 2023, pp. 1225–1236. doi:10.1109/ICSE48619.2023.00109.

[2] L. Yang, C. Yang, S. Gao, W. Wang, B. Wang, Q. Zhu, X. Chu, J. Zhou, G. Liang, Q. Wang, J. Chen, On the evaluation of large language models in unit test generation, 2024. URL: https://arxiv.org/abs/2406.18181. arXiv:2406.18181.

[3] E. Farchi, K. Nayak, P. G. Majumdar, S. Route, Technique to baseline qe artefact generation aligned to quality metrics, 2025. URL: https://arxiv.org/abs/2511.15733. arXiv:2511.15733.

[4] S. Xu, Z. Wu, H. Zhao, P. Shu, Z. Liu, W. Liao, S. Li, A. Sikora, T. Liu, X. Li, Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis, 2024. URL: https://arxiv.org/abs/2402.11398. arXiv:2402.11398.

[5] J. Estdale, E. Georgiadou, Applying the ISO/IEC 25010 Quality Models to Software Product: 25th European Conference, EuroSPI 2018, Bilbao, Spain, September 5-7, 2018, Proceedings, 2018, pp. 492–503. doi:10.1007/978-3-319-97925-0_42.

[6] D. Bowes, T. Hall, J. Petrić, T. Shippey, B. Turhan, How good are my tests?, in: Proceedings of the 8th Workshop on Emerging Trends in Software Metrics, WETSoM '17, IEEE Press, 2017, p. 9–14.

[7] S. Jana, R. Biswas, K. Pal, S. Biswas, K. Roy, The evolution and impact of large language model systems: A comprehensive analysis, Alochana J 13 (2024) 65–77.

[8] F. Molina, A. Gorla, Test oracle automation in the era of llms, 2024. URL: https://arxiv.org/abs/2405.12766. arXiv:2405.12766.

[9] M. Dahiya, R. Gill, N. Niu, H. Gudaparthi, Z. Peng, Leveraging chatgpt to predict requirements testability with differential in-context learning, in: 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI), 2024, pp. 170–175. doi:10.1109/IRI62200.2024.00044.

[10] X. Guo, H. Okamura, T. Dohi, Improving test suite generation quality through machine learning-driven boundary value analysis, Array 27 (2025) 100496. URL: https://www.sciencedirect.com/science/article/pii/S2590005625001237. doi:https://doi.org/10.1016/j.array.2025.100496.

[11] B. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE-2007-01, EBSE Technical Report, Keele University and Durham University Joint Report, 2007.