http://dx.doi.org/10.21622/ACE.2026.06.2.2265

EXPLAINABLE LARGE LANGUAGE MODELS FOR SOFTWARE REQUIREMENTS ENGINEERING: A SYSTEMATIC LITERATURE REVIEW

Hellen Wasike and John Gichuki Ndia

Murang'a University of Technology, School of Computing and Information Technology, Kenya

Emails: {waslene@gmail.com, jndia@mut.ac.ke}

Received 11 June 2026 - Accepted 28 August 2026 - Published 17 September 2026

ABSTRACT

Large language models (LLMs) have become increasingly popular in automating language-dependent software engineering activities, especially in Requirements Engineering (RE), because they are inherently language-driven. The increasing popularity of LLMs in software engineering applications has led to a lack of clarity and understanding regarding the use of Explainability within LLM-based RE solutions. There is no existing literature summarizing their techniques, assessment methodologies, issues, or future research directions. This systematic review aims to explore the state-of-the-art techniques, methodologies, and challenges related to using Explainability in LLMs for RE applications. The study followed PRISMA guidelines to identify works published between 2018 and 2026 from seven databases, including Springer, ScienceDirect, IEEE Xplore, ACM Digital Library, Wiley, arXiv, and Google Scholar. This survey considered and synthesized 57 primary sources from a total of 487 records. The most investigated areas where LLMs are applicable in RE processes include elicitation, ambiguity detection, and traceability. The approaches used for explaining the models are mainly based on prompts and post-hoc methods that do not use any formal approach. In terms of evaluation, task performance-related measures exist, but neither approach evaluates the quality of explanations provided by LLMs, and there are no benchmark sets. Challenges related to RE using LLMs include hallucinations, lack of Explainability, absence of human-centered design approaches, and unresolved ethical issues. In conclusion, Explainability appears to be treated as an afterthought rather than a core element in designing LLM-based RE.

Keywords: (Explainability, Explainable Artificial Intelligence (XAI), Large Language Models (LLMs), Requirements Engineering (RE), Software Engineering (SE))

1. INTRODUCTION

Large Language Models (LLMs) have emerged as powerful tools that can understand and generate natural language, giving rise to opportunities for automation in Software Engineering (SE) activities. Studies show that developers are widely embracing the use of LLMs in tasks like analyzing software products, documentation, and generating code [1]. Since LLMs are mainly expressed in natural language and need considerable specialized interpretation and analysis, they are well-suited for Requirements Engineering (RE) [2].

RE plays a pivotal role in the software development process because ambiguities or bugs at this stage can cascade to the subsequent phases, leading to expensive project failures. Conventional RE tasks such as requirements elicitation, analysis, traceability, and validation are often tedious and reliant on human skills. Recent studies reveal that LLMs have demonstrated promising results in automating or supporting tasks in areas like requirement

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

quality assessment, traceability analysis, and ambiguity detection [2], [3], [4]. For instance, LLM-based methodologies have been suggested to improve requirements traceability via prompt engineering approaches that associate requirements with pertinent development outputs [3]. Additionally, industrial studies have demonstrated that LLMs can detect and explain unclear requirements, thereby enabling analysts to improve requirement clarity and quality [4].

Despite these promising uses, incorporating LLMs into RE raises major concerns about transparency, reliability, and trust. Many LLM-based models work as complicated black-box models that hide their core reasoning process, making it difficult for users to interpret. Stakeholders frequently want correct reasons for automated requirement analysis and validation decisions, especially in safety-critical and industrial contexts. However, the restricted interpretability of LLM outputs makes it difficult for practitioners to comprehend how specific results or recommendations are reached [5]. This difficulty has sparked an increased interest in integrating Explainable Artificial Intelligence (XAI) methods into LLM-based software engineering tools [6].

XAI provides interpretable explanations for Machine Learning (ML) models' outputs, thereby improving their transparency. Explainable methods can be employed in RE to assist analysts in understanding how the LLMs detect ambiguities, categorize requirements, or determine traceability relationships. Studies have begun investigating how Explainability can be utilized in RE tasks [4], [7]. Despite various studies and surveys investigating the broader application of LLMs in SE and RE, most work has focused on task automation or performance improvements [8], [9], [10]. The aspect of Explainability in LLM-based RE systems has received relatively little attention. Existing literature reviews emphasize research avenues for using LLMs in RE. Still, they do not give a complete examination of how explainability techniques are integrated into these approaches, or how such techniques affect transparency and confidence in RE processes [8], [10]. As a result, there is still a lack of consolidated knowledge about the role of explainable LLMs in assisting requirements engineering processes.

The gap identified points to the need for a more focused examination of explainable LLMs within requirements engineering. Existing work has demonstrated the growing use of LLMs across RE activities. However, less attention has been given to how their outputs are explained, how those explanations are evaluated, and whether the available approaches provide sufficient evidence of transparency and trustworthiness. This study, therefore, conducts a systematic literature review (SLR) of explainable LLMs in RE. Rather than simply cataloging LLM applications in RE, it examines the explainability problem across RE tasks, techniques, and evaluation methods. It further synthesizes the findings into a taxonomy of explainability approaches and a quality-assessment framework for explainable LLM-based RE. The following research questions guide the study:

1. What are the existing applications of LLMs in RE?
2. What techniques are used in providing explanations in LLM-based RE?
3. What evaluation methods are used to assess explainable LLM-based approaches in RE?
4. What are the existing challenges in applying explainable LLMs to RE?
5. Are there research gaps and future directions that can be identified in this domain?

The remaining part of the paper is structured as follows: Section 2 reviews related work on LLMs, RE, and Explainability. Section 3 describes the research methodology of the review process. Section 4 presents an analysis of the results, while Section 5 discusses the findings, challenges, and implications of the study. Lastly, Section 6 presents the paper's conclusions and outlines future directions for research.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

2. TERMINOLOGY AND CONCEPTUAL DISTINCTIONS

The literature uses Explainability, interpretability, and transparency with some overlap, and this review therefore adopts explicit terminology to maintain consistency. Explainability is used as the overarching concept referring to mechanisms or information that help users understand or assess why an LLM produces a particular output [11], [12]. Interpretability is the extent to which humans can understand the model or its outputs. Transparency is related to the availability of information or evidence about the system and its outputs[13], [14]. XAI represents the broader field of methods concerned with making AI systems understandable. Explainable LLMs allude specifically to LLM-based systems that incorporate mechanisms for supporting the understanding, justification, or verification of their outputs [15]. Finally, explanation quality refers to the quality of the explanation itself, including properties such as faithfulness, completeness, consistency, and interpretability [12], [14]. Where primary studies use these terms differently, their original terminology is retained when reporting individual findings, while the synthesis uses Explainability as the overarching concept.

3. RELATED WORKS

Several SLRs, systematic mapping exercises, and structured surveys have been conducted to explore the relationship between LLMs, software RE, and XAI. It is essential to examine these reviews in detail to establish how the current study contributes to existing literature on this topic. Table 1 provides a comparison of 11 existing SLRs and surveys in terms of their scope, topic, coverage of LLMs within RE, Explainability, and the gap they leave behind that the current study addresses.

3.1. LLMS IN REQUIREMENTS ENGINEERING

Hou *et al.* [9] performed a detailed SLR study on the applications of LLMs, focusing on main software engineering tasks such as requirements analysis, code summarization, and program repair. Although it was thorough and rigorous, the study did not focus on software RE tasks in particular or Explainability as an issue. Similarly, Hemmat *et al.* [8] carried out a systematic review focusing explicitly on the application of LLMs in software RE in terms of elicitation, specification, analysis, and traceability. The survey pointed out instances where interpretability was considered an issue in LLM-generated outputs in RE processes. Conversely, it did not conduct a comprehensive analysis of explainability methods, evaluation measures, or how explanations could be generated, validated, or communicated within RE workflows.

Zadenoori *et al.* [10] studied the application of LLMs in RE by performing an analysis that covered various aspects, including publication trends, RE processes supported by LLMs, prompts used, and evaluation approaches. Despite identifying issues regarding the rigor of evaluation, it did not conduct a systematic analysis of XAI approaches within LLM-based RE systems. Khan *et al.* [16] performed a complementary SLR on sources that included literature on LLM techniques used for elicitation, analysis, modeling, validation, specification, prioritization, and tracing within RE processes. While they discussed interpretability-related challenges, their study did not capture a systematic review of the explainability techniques used. Additionally, Cheng *et al.* [17] identified key trends in generative AI applications in RE. They acknowledged limited empirical depth and an absence of explainability-focused analysis.

Another study provided a systematic survey on applying LLMs to improve RE [18]. It explored the abilities of LLMs in terms of opportunities, such as automated requirements elicitation, specification generation, and quality improvement, and limitations, like hallucinatory behavior, domain sensitivity, and insufficient empirical validation in actual industrial settings. Although the study briefly acknowledged transparency as a desirable property of LLM outputs, it neither systematically investigated XAI methods nor considered how XAI methods could be integrated into RE processes. Magableh [19] comes the closest to the

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

current study, having conducted an SLR on XAI techniques applied in RE. However, the scope of the study was restricted to only cross-cutting and aspect-oriented concerns. Its scope predated the widespread application of LLMs in this domain, and therefore, it did not address the integration of LLMs into XAI-driven RE processes.

3.2. NATURAL LANGUAGE PROCESSING AND AI IN REQUIREMENTS ENGINEERING

Before large language models (LLMs) emerged, Zhao *et al.* [20] systematically mapped the Natural Language Processing for Requirements Engineering (NLP4RE) research landscape. The authors surveyed the use of Natural Language Processing (NLP) techniques for requirements engineering (RE) between 1983 and 2019 from 404 articles. In doing so, the study provided a taxonomy of NLP4RE research, covering over four decades of development in linguistic analysis activities like defect detection, traceability, and requirements classification. This research preceded the LLM revolution; hence, the question of Explainability remained beyond its scope.

Another study conducted in 2024 also provided an SLR on NLP in software requirements engineering [21]. Unlike Zhao *et al.* [20], the researchers examined the literature published from 1991 to 2023 across various platforms, including Scopus, IEEE Xplore, ACM Digital Library, and Clarivate. Although the study recognized the importance of LLMs in RE, it focused on NLP in general and did not cover XAI.

3.3. EXPLAINABLE AI IN SOFTWARE ENGINEERING

There have been systematic literature reviews investigating the intersection between XAI and SE. Specifically, Mohammadkhani *et al.* [22] carried out an SLR to evaluate the prevalence of XAI adoption in SE research, known as XAI4SE. They analyzed 24 out of 869 initial studies, and their results revealed that XAI approaches in SE have mainly been utilized in software maintenance and defect prediction with a particular emphasis on traditional machine learning rather than complex language models. Significantly, they also noted a clear gap in standardized evaluation metrics for XAI approaches in SE. On a more comprehensive scale, Cao *et al.* [23] reviewed 108 studies across 23 unique SE activities to classify different XAI approaches and investigate various evaluation methods. Based on their findings, they validate that requirements engineering (RE) is one of the least studied phases in SE using XAI, and most of the literature on XAI4SE mainly revolves around software maintenance. In both studies, the role of LLMs, or their integration with XAI, within RE workflows was not specifically addressed.

Table 1: A Summary of Related SLRs

Scope	Does it cover LLMs in RE?	Does it cover Explainability?	The gap addressed by the present study	Studies
LLMs for all SE tasks	Partially	No	Broad SE scope and no RE-specific XAI treatment	[9]
LLMs in software RE and research directions	Yes	No	No systematic treatment of explainability techniques	[8]
LLMs for RE tasks	Yes	No	Evaluation gaps identified, but no XAI integration analysis	[10]
LLM applications in RE activities	Yes	No	Focuses on RE automation, and Explainability is not systematically addressed.	[16]
Generative AI in RE	Yes	No	Lacks empirical depth and has no XAI-focused analysis	[17]
Role of LLMs in advancing RE	Yes	Only acknowledged but not analyzed	Not a formal SLR, and XAI methods are not analyzed	[18]
NLP techniques across RE tasks (NLP4RE)	No (pre-LLM)	No	Pre-LLM era with no explainability or LLM focus	[20]
NLP in SRE 1991-2023 (ML, DL, LLMs)	Partially	Only acknowledged but not analyzed	Broad NLP scope. XAI and LLMs are not the primary focus	[21]

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Scope	Does it cover LLMs in RE?	Does it cover Explainability?	The gap addressed by the present study	Studies
XAI for SE tasks (ML4SE)	No	Yes	Covers classic ML/SE and does not address LLMs and RE	[22]
XAI for ML/DL in SE (23 SE tasks)	No	Yes	Covers SE broadly, where LLMs and RE-specific XAI gaps are unaddressed	[23]
XAI in RE for aspect identification	No	Yes	Narrow scope. LLM-XAI is not integrated into the RE lifecycle	[19]

4. METHODOLOGY

The study adopted a Systematic Literature Review (SLR) approach that was guided by the PRISMA framework. The goal of this study is to methodically recognize, evaluate, and synthesize the current research on explainable LLMs in RE. The PRISMA approach helped structure the review process into well-defined phases that ensured transparency, reproducibility, and rigor. It also helped minimize bias and provide a detailed overview of existing studies. Figure 1 is the PRISMA flow diagram showing a summary of the whole selection process.

4.1. SEARCH STRATEGY

Relevant studies were retrieved from major digital libraries, including Springer, ScienceDirect, IEEE Xplore, ACM Digital Library, Wiley, arXiv, and Google Scholar. The search focused on literature published between 2018 and 2026 to capture recent developments in LLMs and XAI. The study searched for a combination of keywords associated with LLMs, Explainability, and RE. Major terms and their synonyms were utilized during the search, and Boolean operators such as "OR" and "AND" were used. The adopted search string was as follows:

("Large language models" or "LLM" or "Generative AI" or "GenXAI") and ("requirements engineering" or "software requirements" or "requirements elicitation" or "requirements traceability" or "requirements verification" or "requirements validation" or "requirements specification" or "requirements analysis") and ("explainability" or "explainable AI" or "XAI" or "interpretability")

4.2. INCLUSION AND EXCLUSION CRITERIA

The review included studies that focused on SE or RE addressing Explainability, interpretability, or transparency in large language models. It also included peer-reviewed articles, conference papers, and high-quality preprints published between 2018 and 2026, as well as studies published in English.

By contrast, the review excluded studies unrelated to SE or RE and papers that did not have substantive discussions of LLMs or Explainability. Duplicate studies across databases, short papers, editorials, tutorials, and non-scholarly articles were also excluded.

4.3. STUDY SELECTION PROCESS

The review followed the PRISMA four-phase approach, which started with identification. Here, the selected databases were used for initial search results. The screening phase was followed by the review of titles and abstracts to remove irrelevant studies. Thereafter, eligibility was assessed by reviewing full-text articles against the inclusion and exclusion criteria. Lastly, the final collection of studies was selected for qualitative synthesis.

4.4. DATA EXTRACTION, SYNTHESIS, AND ANALYSIS

To ensure consistency and effective comparison across studies, the review used a structured data extraction form to collect information from each selected study. The extracted data included publication details such as author, year, and venue; the research

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

focus and objectives; RE tasks addressed; LLM techniques used; explainability methods applied; key findings and contributions; and identified challenges.

A qualitative synthesis approach was used to analyze the selected studies. The focus was on finding common themes, patterns, and trends relating to applications of LLMs in RE, explainability techniques and frameworks, challenges, and emerging research directions.

4.5. QUALITY ASSESSMENT OF INCLUDED STUDIES

To assess the strength of the evidence, all 57 included studies were subjected to a structured quality assessment tailored to the review objectives. Five criteria considered for quality assessment were research clarity, methodological transparency, evaluation rigor, Explainability of evidence, and reproducibility. Each criterion was scored as 0 (not satisfied), 1 (partially satisfied), or 2 (fully satisfied), giving a maximum score of 10. Research clarity covered the definition of objectives or research questions. Methodological transparency considered the description of models, RE tasks, data, and experimental procedures. Evaluation rigor examined the appropriateness and clarity of the evaluation methods. Explainability evidence considered whether explanation quality was explicitly assessed through properties such as faithfulness, consistency, completeness, interpretability, or usefulness, while reproducibility considered the availability of sufficient implementation and experimental details.

Based on the total scores, studies were classified as high quality (8 – 10), moderate quality (5 – 7), or low quality (0 – 4). The assessment was not used to exclude studies but to distinguish the strength of the available evidence and identify weaknesses in current research. This distinction matters because methodological rigor does not necessarily imply explanatory rigor. A study may demonstrate strong LLM performance on an RE task while providing little evidence that its explanations are faithful or useful. The resulting quality scores were therefore incorporated into the synthesis to assess the reliability of observed trends and identify areas requiring stronger empirical evidence.

4.6. VALIDITY CONSIDERATIONS

Internal validity. The main internal validity concern relates to the risk of bias in the selection of studies. Although inclusion and exclusion criteria were predefined and consistently applied, distinguishing studies that focused on Explainability and required RE from those that peripherally discussed Explainability involved a subjective judgment. Consequently, a standardized data extraction form was used to mitigate the effect of such a decision.

Construct validity. The constructs of interest, "explainability" and "requirements engineering," are multidimensional and differently defined in reviewed sources [11], [12], [13]. Reviewed studies used terms like interpretability and Explainability interchangeably, and the scope of RE tasks is varied across articles. Although the inclusion criteria attempted to account for that variation, it led to heterogeneity, which made direct comparative assessments difficult.

External validity. Some of the included studies come from conferences and preprints, which reflect an early-stage and exploratory field of research. Compared with conference proceedings, journal articles tend to report more mature findings that have undergone empirical validation. In addition, many studies utilized controlled academic datasets rather than datasets gathered from practice. Therefore, their external validity should be considered with caution.

Conclusion validity. The speed at which LLMs are evolving suggests that conclusions concerning specific model implementations like GPT-3 and GPT-4 prevailing in the reviewed literature [10], [17] may quickly become obsolete as a result of the emergence of newer architectures. The search was conducted with 2026 as a cutoff to include recent findings. However, the field is evolving faster than traditional SLR cycles can track. Moreover, potential publication bias resulting in an excessive reporting of positive rather than null or negative findings may lead to an excessively favorable conclusion about the performance of LLMs in explaining RE tasks. The inclusion of high-quality preprints was intended to mitigate this bias partially.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

5. RESULTS

This section presents the findings of the literature review based on the PRISMA framework. The analysis identified recurring patterns, research trends, and gaps in the application and Explainability of LLMs across RE activities. The research questions were used to organize the results into major themes. The relevant background literature is integrated with empirical findings for each subsection using the 57 included studies, accompanied by a summary table.

5.1. PUBLICATION TRENDS

Figure 2 presents the publication distribution of the 57 studies included in the review. Studies published between 2018 and 2022 are grouped into a single category because this period predates the widespread use of generative LLMs. This time frame also contains relatively few studies directly addressing explainable LLMs in requirements engineering. The distribution, therefore, provides a useful view of how quickly research in this area has developed.

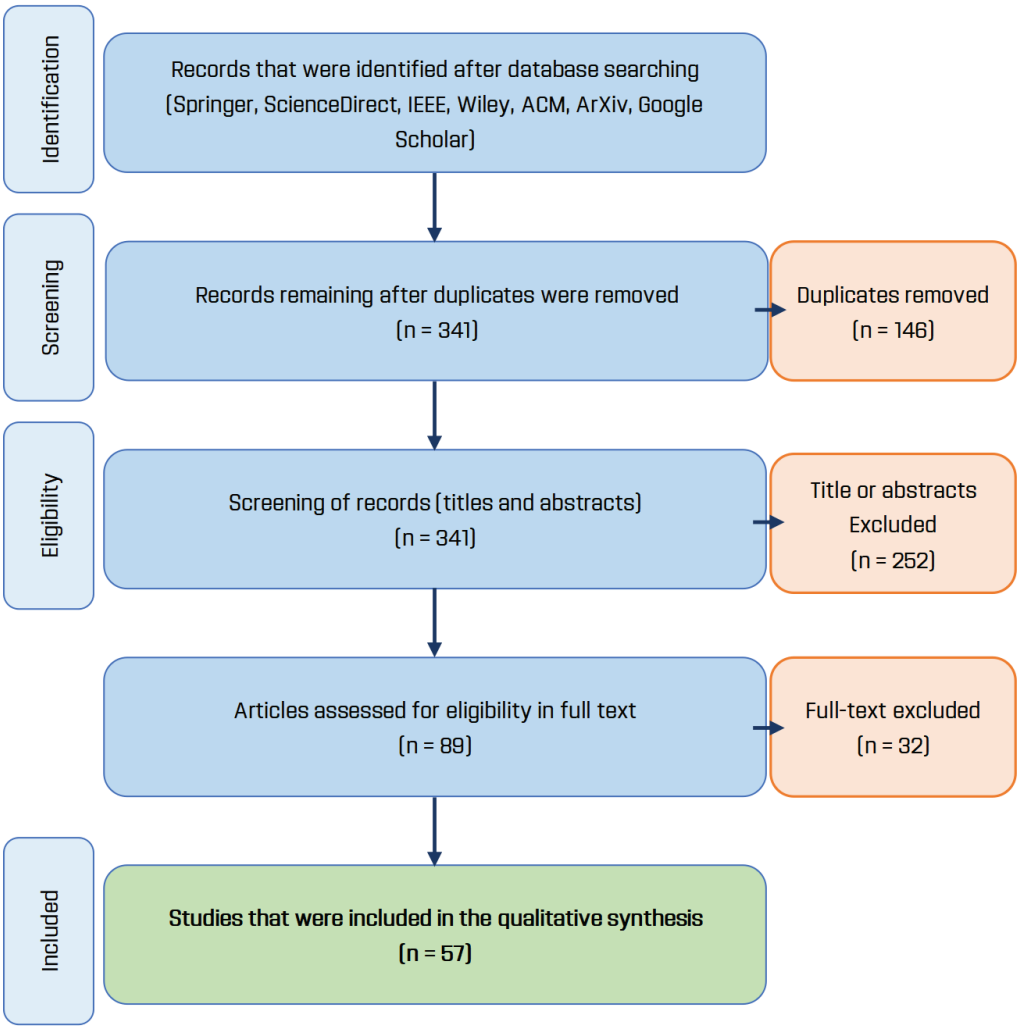


Figure 2: Distribution of the 57 included studies by publication year

The distribution is strongly concentrated in 2024 – 2026, with more than two-thirds of the included studies published during this period. This sharp increase reflects the rapid adoption of generative LLMs in RE and the more recent interest in making their outputs explainable. At the same time, the concentration of evidence in these years points to a field that is still relatively young. Many approaches remain exploratory and lack long-term or industrial validation. Meanwhile, shared benchmarks, mature evaluation protocols, and reproducible

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

practices are still developing. The publication trend therefore signals both growing research activity and the limited maturity of the evidence base.

5.2. QUALITY ASSESSMENT OF INCLUDED STUDIES

The quality assessment shows that the evidence base is predominantly moderate in quality. 38 of the 57 studies (66.7%) were classified as moderate, 13 (22.8%) as high, and 6 (10.5%) as low, as shown in Table 2. Thus, although most studies provide a reasonable methodological basis for synthesis, relatively few meet the threshold for high overall quality. The predominance of moderate-quality studies indicates that the evidence base is sufficiently developed to support thematic synthesis. However, it is not yet consistently rigorous across all dimensions. The distribution also suggests that reported advances in LLM-based RE should be interpreted with caution when evidence extends from task performance to claims about Explainability.

Table 2: Quality Distribution of the 57 Included Studies

Quality level	Score range	Number of studies	Percentage
High	8 – 10	13	22.8%
Moderate	5 – 7	38	66.7%
Low	0 – 4	6	10.5%

The criterion-level assessment in Table 3 reveals a more important pattern. Research clarity was strong, with 55 studies (96.5%) fully satisfying the criterion, while 42 (73.7%) demonstrated strong methodological transparency. In contrast, only 15 studies (26.3%) fully satisfied the evaluation-rigor criterion, and just 13 (22.8%) provided strong explainability evidence. Reproducibility was weaker, with only three studies (5.3%) fully satisfying this criterion, and 44 (77.2%) were only partially reproducible. The gap between methodological reporting and explainability evidence is particularly notable. It suggests that the field is generally capable of describing what LLMs are used for and how they are evaluated. However, it is much less consistent in establishing whether the resulting explanations are faithful, complete, or otherwise reliable.

Table 3: Criterion-Level Quality Assessment of the Included Studies

Quality criterion	Fully satisfied (2)	Partially satisfied (1)	Not satisfied (0)
Research clarity (RC)	55 (96.5%)	2 (3.5%)	0 (0.0%)
Methodological transparency (MT)	42 (73.7%)	14 (24.6%)	1 (1.8%)
Evaluation rigor (ER)	15 (26.3%)	37 (64.9%)	5 (8.8%)
Explainability evidence (EE)	13 (22.8%)	14 (24.6%)	30 (52.6%)
Reproducibility (R)	3 (5.3%)	44 (77.2%)	10 (17.5%)

RQ1: WHAT ARE THE EXISTING APPLICATIONS OF LLMS IN RE?

LLMs have become valuable tools for effecting changes in software engineering processes, especially those that require natural language processing, such as code generation, documentation, and defect detection [1], [9], [24]. Studies have confirmed that LLMs can automate complex cognitive tasks, leading to increased efficiency and reduced development costs [1], [25]. While LLMs excel at structured and semi-structured operations, their performance assessment shows inconsistencies in context and reasoning coherence [25].

RE is an ideal use case for LLMs because RE-related operations are language-intensive. Despite RE being one of the fundamental aspects of software development, its operations are plagued by consistent challenges associated with natural language ambiguity and variability [26], [27], [28], [29]. Ambiguous, inconsistent, and incomplete specifications often result

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

in budget overruns, schedule delays, and downstream defects [30], [31]. These problems become even more pressing in dynamic environments such as agile development, since requirements evolve continuously, requiring rapid and timely validation [32]. Furthermore, the growing need for automated tools that process huge volumes of requirements has given rise to the evolution of complex systems [8], [10]. As a result, AI and LLMs have become the logical choice for supporting efficient RE.

The empirical evidence from the included studies points out that LLMs have been applied to various RE tasks. The most popular application area is requirements elicitation and generation, where models can transform unstructured input data, such as stakeholder interviews, user feedback, and domain documents, into structured output data, such as user stories and requirement specifications [2], [33], [34]. In addition, domain adaptation techniques, such as instruction tuning, further improve generation performance [35]. Consequently, manual effort has been significantly reduced, and consistency has been improved in the early stages of requirements engineering.

Requirements analysis and quality assurance have also been applied in RE activities, where LLMs can be used to detect ambiguities, inconsistencies, and incompleteness in requirements specifications. Studies show that LLMs can recognize even subtle language mistakes and produce suggested refinements [4], [28]. Also, LLM-based analysis and modeling have further facilitated the generation of RE artifacts [36]. Requirements prioritization and classification are RE tasks that include activities like the identification of functional and non-functional requirements, enabling (Rejithkumar & Anish, 2025; Viet & Vinh, 2024) (Anish, 2025; Viet & Vinh, 2024).

Requirements traceability and validation are emerging as fast-growing application areas of LLMs. Studies have proposed techniques such as TraceLLM [3] and semantic API alignment [38] that improve the process of establishing trace links between requirements and downstream artifacts, including design models, code, and test cases. Additionally, Ferrari et al [39] utilized LLMs to transform requirements into structured representations such as UML sequence diagrams. In terms of validation, Xie *et al.* [40] presented promising yet inconsistent findings on the efficiency of LLMs in generating software specifications. Further, Roy *et al.* [41] expanded the scope of RE into human-centric dimensions by introducing semi-automatic techniques for requirements-to-emotion impact mapping.

Table 4: A Summary of Applications of LLMs in RE (RQ1)

RE Application	Description	Number of studies
Requirements elicitation/ generation	Converting unstructured inputs (interviews, user feedback) into structured requirements	22 (38.6%)
Ambiguity detection/quality assurance	Identifying vague, inconsistent, or incomplete requirements	3 (5.3%)
Requirements prioritization/ classification	Ranking and categorizing functional/non-functional requirements	3 (5.3%)
Requirements traceability	Linking requirements to design artifacts, code, and test cases	3 (5.3%)
Requirements modeling/ specification	Generating UML diagrams and structured specifications	6 (10.5%)
Validation/verification	Checking requirements against specifications and standards	20 (35.1%)

RQ2: WHAT TECHNIQUES ARE USED IN PROVIDING EXPLANATIONS IN LLM-BASED RE?

Explainable Artificial Intelligence (XAI) is a concept that was introduced to tackle the black-box nature of current AI models. Studies have categorized XAI methods into intrinsic and post-hoc approaches. The intrinsic approach is where the model is interpretable by design, while the post-hoc approach generates explanations after the prediction [11], [12].

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Subsequent research has expanded the notion of explainable AI to encompass other aspects of trustworthy AI, including robustness, fairness, and accountability [42], [43], [44], [45]. In SE workflows, Explainability has been beneficial in offering insights into model behavior, thus supporting activities like debugging, validation, and decision-making [15], [22], [23].

The practical relevance of XAI has been demonstrated through its domain-specific applications in SE. Petrillo et al. [46] employed transformer models that had explicit explainability mechanisms to perform classification for security requirements. Another study devised a validation framework that combined retrieval-augmented generation (RAG) with explanation capabilities [47]. These efforts contribute to the recognition of Explainability as an indispensable element required for deploying AI in dynamic SE contexts. They also set the stage for investigating how these techniques can be implemented in LLM-based RE workflows.

The reviewed literature shows that operationalization of Explainability in LLM-based requirements engineering is mainly based on prompt-based and post-hoc approaches. The most common approach is prompt-based LLMs used to generate explanations in natural language alongside their outputs. This enables users to understand the logic behind the classification, prioritization, and transformation of the requirements [4], [7]. These kinds of explanations are directly embedded into workflows, allowing technical and non-technical stakeholders to access them. Bashir *et al.* [4] carried out an industrial validation of LLM-assisted ambiguity detection research that was accompanied by natural language explanations. Similarly, Shah *et al.* [7] proposed the EQDAT-RE framework that was used for transparent qualitative data analysis in requirements engineering.

Formal XAI techniques have also been adapted in RE tasks, including classification and validation, to guarantee explanation correctness. Some of these techniques include attention visualization, feature attribution, SHAP, and example-based explanations [46]. In addition, the inherently transparent nature of retrieval-augmented generation (RAG) has increased its adoption. This improves traceability and accountability since LLM outputs are grounded in external knowledge sources [47]. Intrinsic, or built-in, Explainability, where explanation ability is included in the model, has also been especially applied in hybrid architectures. For instance, the proposed NICE [37] method was used for the non-functional classification of requirements, integrating small language models with explicit explanation generation.

Framework-level Explainability is another significant category. Here, the process of eliciting, modeling, and managing Explainability is explicitly done as a non-functional requirement [13], [14]. Chazette *et al.* [13] provided a conceptual framework for Explainability as an RE artifact. Similarly, Obaidi *et al.* [48], [49] proposed approaches to automatically generate explainability requirements from user reviews and through structured elicitation techniques. New techniques for debugging and interpretable validation of the requirements using LLMs are also emerging [50].

The cross-tabulation in Table 5 exposes a clear pattern in the current research landscape. Prompt engineering dominates Explainability across nearly every RE activity. This makes it the preferred mechanism for producing more transparent and structured outputs. Its popularity is understandable since it is simple to implement and integrates naturally with LLM workflows. However, prompt engineering explains what the model generates more effectively than why it generates it. As a result, many studies treat well-articulated responses as evidence of Explainability without establishing whether those explanations faithfully reflect the model's reasoning.

A different trend emerges in requirements traceability, where retrieval-augmented generation (RAG) is gaining traction because it grounds responses in external artifacts that stakeholders can inspect. This improves transparency and supports evidence-based decision-making. Yet, its use remains largely confined to retrieval-intensive tasks and has seen limited adoption across the broader RE lifecycle. By contrast, established XAI techniques such as SHAP and LIME appear only rarely. They are designed for predictive models and do not transfer easily to generative LLMs, thus leaving a noticeable gap between advances in XAI and their application to RE.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Perhaps the more important finding is not the choice of explainability technique but how it is evaluated. Most studies stop at demonstrating task performance or presenting plausible natural-language justifications, while objective assessment of explanation quality remains uncommon. Evidence that explanations improve requirements validation, stakeholder agreement, or software quality is still limited. This disconnect suggests that Explainability is advancing faster as a usability feature than as a measurable engineering capability, highlighting the need for more rigorous and standardized evaluation frameworks.

Table 6 combines the two most important dimensions of the review. It converts the qualitative adoption levels in Table 5 into a quantitative cross-tabulation.

Table 5: Distribution of Explainability Techniques

Explainability technique	Number of studies n (%)
Prompt engineering	18 (31.6%)
Natural-language rationales	11 (19.3%)
Retrieval-augmented generation (RAG)	2 (3.5%)
Rule/knowledge-based explanations	6 (10.5%)
SHAP	1 (1.8%)
LIME	0 (0.0%)
Other/formal XAI	19 (33.3%)

Table 6: Cross-Tabulation of Requirements Engineering Phases and Explainability Techniques used in LLM-Based RE (RQ2)

Requirements Engineering Phase	Prompt Engineering	Retrieval-Augmented Generation (RAG)	Natural Language (NL) Rationales	SHAP	LIME	Rule / Knowledge-based Explanation	Other/ formal XAI	Studies	Identified Gap
Requirements Elicitation/ generation	11	0	6	0	0	2	3	22	Heavily relies on prompt engineering. Objective evaluation of explanation quality and stakeholder trust is limited.
Requirements Specification/ modeling	3	0	0	0	0	0	3	6	Specifications are generated with minimal explainability mechanisms. Faithfulness of generated specifications to stakeholder intent is not formally evaluated. The natural language outputs are treated as sufficient evidence of transparency without verification.
Requirements Classification / Prioritization	1	0	0	1	0	1	0	3	No formal XAI method has been applied to prioritization. Comparative evaluation benchmarks and standardized explanation quality metrics remain absent.
Requirements Traceability	1	1	0	0	0	1	0	3	Explanations for why specific trace links are established are rarely generated or validated. Formal XAI techniques and the evaluation of trace explanation fidelity are absent. RAG-based grounded traceability is emerging but not yet prominent.
Ambiguity Detection	1	0	1	0	0	0	1	3	Explanations rely mainly on NL justifications with little evidence that they faithfully reflect model reasoning.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Requirements Verification / Validation	1	1	4	0	0	2	12	20	[50] Provides the most formally grounded Explainability through RAG and rule-based justification trees, but its scope is narrow. Broader adoption of RAG-based explainable validation across RE contexts and empirical evidence that explanations improve validation decisions remain limited.
Total	18	2	11	1	0	6	19	57	

RQ3: WHAT EVALUATION METHODS ARE USED TO ASSESS EXPLAINABLE LLM-BASED APPROACHES IN RE?

Evaluating XAI systems involves challenges that go beyond the limitations of evaluating traditional ML systems. SLRs on XAI4SE indicate a noticeable lack of well-established evaluation criteria for XAI techniques, which leads to inconsistencies and limits cross-study benchmarking [22], [23]. The problem is more apparent in the field of RE, where explanation quality is seen as a sociotechnical aspect that should be evaluated according to the requirements of different stakeholders and situations [51]. Previous studies on requirements for Explainability have indicated that evaluation should consider aspects that include usability, actionability, and trust. These are dimensions that cannot be assessed using standard performance measures [13], [14], [48].

Mixed use of quantitative, qualitative, and hybrid evaluation strategies is observed in the analysis of explainable LLMs across the reviewed publications. The quantitative approach prevails and is used extensively in evaluating requirements in tasks such as classification, ambiguity detection, and traceability. Recall, accuracy, precision, and F1-score have been the commonly utilized performance metrics [3], [4], [24]. Although they lay the groundwork for comparing various models, these metrics provide very little information about the quality or usefulness of the generated explanations since they focus only on task correctness and not on interpretation [52].

Qualitative evaluation techniques are becoming increasingly popular as additional measures. Explanation qualities like clarity, relevance, interpretability, and trustworthiness can be assessed using qualitative methods like expert evaluation, user studies, and case-based analysis [6], [7], [49]. It is also common practice to involve domain experts during the evaluation of model performance in RE tasks due to their ability to validate explanation relevance and applicability for decision-making processes [51]. A hybrid evaluation approach that combines automatic evaluation of model performance and human perception has emerged, bringing forth a more detailed evaluation interpretation [17], [46].

Table 7: Distribution of Evaluation Methods for Explainable LLM-Based RE (RQ3)

Evaluation Type	Metrics or Approach	Purpose	Studies n (%)	Suitability for Assessing Explanation Quality	Technical Basis for Suitability
Quantitative	Accuracy, Precision, Recall, F1-score	Measure the predictive performance of the LLM on RE tasks	14 (24.6%)	Low	Assess task correctness only. They do not evaluate whether explanations faithfully reflect the model's reasoning, provide sufficient evidence, or support stakeholder understanding.
Qualitative	Expert Review, User Studies, Case Analysis	Assess explanation clarity, usefulness, and stakeholder acceptance	38 (66.7%)	Moderate	Capture interpretability, usability, and perceived trust. They rely on subjective judgment and are difficult to reproduce consistently across evaluators and contexts.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Hybrid	Automated Metrics plus Human Judgment	Combine quantitative performance measures with qualitative assessment	4 (7.8%)	Moderate –High	Integrate objective task performance with human assessment of explanation quality. They provide a more balanced evaluation, although standardized protocols remain limited.
Intrinsic Model Evaluation	Faithfulness, Consistency, Completeness (and where applicable, Transparency)	Directly evaluate the quality of explanations generated by the model	1 (1.8%)	High	Measure whether explanations accurately reflect the model's reasoning, remain stable across similar inputs, and provide sufficient evidence for the generated outputs, which are the most appropriate indicators of explanation quality.

5.3. WHAT ARE THE EXISTING CHALLENGES IN APPLYING EXPLAINABLE LLMS TO RE?

Challenges that have been considered most pressing are reliability and transparency in both LLM4RE and XAI4SE research. Large language models have been observed to experience the hallucination problem, where outputs generated are plausible yet not true, even when they appear to be credible [5], [53]. Although comparative analyses reveal that LLMs can approach expert-level performance on benchmark tasks, they often fail to meet reliability and verifiability standards [54]. Additionally, privacy, security, and accountability limit the adoption of LLMs in safety-critical systems [5].

This review revealed five major areas of challenge across the literature reviewed. Reliability and hallucination continue to be the most frequently mentioned challenges. Outputs of large language models in requirements engineering tasks tend to be inconsistent. They can include errors that may be particular to a specific domain and require further inspection by a domain expert [40], [53], [55]. The absence of standardization of the explanation process poses structural challenges to the field. Presently, the explanations are provided in an ad hoc manner and involve prompt engineering and natural language reasoning processes without any formal grounding or consistency measures [17], [52], [56]. As a result, it becomes harder to perform inter-method or inter-model comparisons and evaluate validity.

There is a significant human-centered design gap since RE includes multiple stakeholders ranging from end users to developers, analysts, and domain experts with different levels of expertise. Scalability and domain dependence remain obstacles, as models perform poorly in context-dependent settings and for specific domains [5], [37]. Finally, researchers have inadequately addressed ethical concerns related to bias, transparency, and accountability, thus hindering adoption, especially in regulated industries [5], [19], [53].

Table 8: A Summary of Challenges in Explainable LLM-Based RE (RQ4)

Challenge Category	Description	Studies
Hallucinations & reliability	LLMs generate plausible but incorrect outputs	[53], [55], [35]
Lack of Standardized Explainability	Explanations are generated in an ad hoc and inconsistent manner without formal frameworks.	[6], [17]
Human-centered design gaps	Explanations are not tailored to diverse stakeholder backgrounds	[49], [51]
Evaluation inconsistency	No agreed metrics for measuring explanation quality in RE	[17], [19]
Scalability and domain dependency	Model performance degrades across domains, leading to poor generalization	[37], [5]
Privacy, security & ethics	Data privacy and ethical use concerns limit adoption	[5], [53]

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

RQ5: ARE THERE RESEARCH GAPS AND FUTURE DIRECTIONS IN THIS DOMAIN?

Across the reviewed studies, five recurring gaps were identified. The first one is the lack of standardized explainability frameworks [19], [52], [56]. The second class is related to the development of human-centered explainability design. Explanations do not take into account various user types and only focus on expert analysts. They do not consider domain experts and users who engage with requirement artifacts [48], [49], [51]. The third gap pertains to a need for standard evaluation measures for Explainability in the RE context [39], [56].

The fourth gap relates to existing concerns about reliability, robustness, and trust in explanation systems [54], [55]. Finally, existing approaches have a narrow scope and fail to integrate other AI paradigms such as symbolic reasoning and knowledge graphs [6], [50]. At the same time, the ethical and governance implications of using explainable models are still a critical area for future interdisciplinary research but remain largely unexplored [5], [53].

Table 9: A summary of research gaps and future directions (RQ5)

Research Gap	Description	Proposed Future Direction
Lack of Standardized Frameworks	Explainability methods are ad hoc and non-reproducible	Develop unified taxonomies and benchmarks for explainable RE
Human-centered Explainability	Explanations are not adaptive to diverse stakeholder needs	Design personalized, interactive, and on-demand explanation systems
Evaluation standards	There are agreed metrics for explanation quality in RE	Establish standardized evaluation benchmarks for explainable RE
Hallucination mitigation	LLMs produce unreliable outputs without robust verification	Integrate human-in-the-loop and formal verification mechanisms
Hybrid AI integration	LLMs lack symbolic reasoning and domain knowledge grounding	Combine LLMs with knowledge graphs and symbolic AI methods
Ethical and governance frameworks	Bias, accountability, and transparency concerns are still unresolved	Develop ethical guidelines and governance models for LLM-based RE tools

6. DISCUSSION

This section presents the interpretations and synthesis of the findings of the reviewed literature across the five research questions. It identifies overarching themes that emerge when the results are viewed together and considers the implications for scholars and practitioners. It also evaluates the validity threats of the current review. Furthermore, this discussion provides insights into how the findings point to the current status of the field and what changes need to be made to advance it.

6.1. LLMS' CAPABILITY AND INTERPRETATION

One of the most important and prevailing issues that emerges out of all five research questions is the discrepancy between the capability of LLMs and their interpretability. On one hand, RQ1 shows that LLMs possess great capability for addressing various types of tasks relevant to requirements engineering, which include requirement elicitation [34], [35], detecting ambiguities [4], maintaining traceability [3], generating specifications [40], and modeling [36], [39]. On the other hand, RQ2 demonstrates that the existing methods for explaining LLMs' decisions are mostly prompt-based and post-hoc in nature. As a result, they lack formal grounding that guarantees that the generated explanations faithfully represent the internal reasoning process [4], [7]. Such an imbalance has dire consequences: despite their high capability, LLMs cannot be easily audited, challenged, or verified for the correctness of their work due to the lack of proper explanation [5], [53].

Generally, the imbalance is characteristic of the relationship between machine learning models and Explainability. In the XAI4SE literature, Explainability is introduced as an add-on

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

rather than a core design criterion for the large language models developed [22], [23]. However, this poses a problem that cannot be overlooked, since RE documents are contractual and auditable artifacts that must withstand scrutiny from diverse stakeholders [13], [14], [26]. Therefore, in the future, LLMs should integrate Explainability from the very start.

6.2. TASK-DRIVEN ADOPTION WITHOUT PRINCIPLES-DRIVEN FRAMEWORKS

The common denominator between Research Questions 1 and 2 is that the usage of LLMs in requirements engineering has largely been task-oriented and has not leveraged principles-based approaches. The vast majority of the studies discussed above have been focused on specific RE tasks, including ambiguity detection [4], generation of traceability links [3], and requirements classification [24], [37]. They neglected implementing them into consistent RE frameworks that would cover explanation generation, validation, and communication during all stages of product development. As a result, the explanations provided by the relevant tools are still fragmented and non-interoperable. Thus, they pose challenges for evaluating the overall trustworthiness of an RE process involving several LLM-based applications [52], [56].

The absence of commonly accepted taxonomies for classifying LLM uses in RE and their corresponding explainability frameworks exacerbates the fragmentation problem. While Chazette *et al.* [13] and Obaidi *et al.* [48], [49] made considerable efforts to define and Explainability as a component of RE, these frameworks have not yet been operationalized within the scope of this paper. The field could benefit greatly from a reference architecture identifying where and at what level of consistency explainability should be provided for each particular RE task.

These findings clarify how this review differs from earlier studies. They largely examined either LLM applications in RE or Explainability in software engineering more broadly. Here, Explainability is treated as part of the problem itself, not as an additional feature. The review considers what RE tasks LLMs support, how their outputs are explained, how those explanations are evaluated, and how much confidence can be placed in the available evidence. This shifts the focus from simply asking whether LLMs perform RE tasks effectively to asking whether their outputs are sufficiently understandable and well supported to inform engineering decisions. These observations provide the foundation for a multidimensional quality framework that considers explanation quality alongside stakeholder needs and practical engineering value.

6.3. EVIDENCE QUALITY AND MATURITY OF EXPLAINABLE LLM RESEARCH IN RE

The quality assessment suggests that explainable LLM research in requirements engineering is progressing, but its maturity is uneven. The most revealing result is not the number of low-quality studies. It is the gap between methodological rigor and evidence of Explainability. Although many studies clearly define their objectives and report reasonable experimental procedures, only 13 of the 57 studies (22.8%) provide strong evidence that their explanations are faithful, complete, or otherwise meaningful. More than half (52.6%) offer no explicit evaluation of explanation quality. This reinforces a broader pattern observed throughout the review. The field has become increasingly effective at demonstrating what LLMs can do for RE, yet remains much less consistent in showing why their outputs should be trusted.

A similar pattern emerges in reproducibility. While 44 studies (77.2%) provide some implementation details, only three satisfy the reproducibility criterion in full. This is more than a reporting issue. In LLM-based RE, seemingly minor changes in model versions, prompts, retrieval configurations, or generation settings can produce different outcomes. Without sufficient experimental detail, it becomes difficult to determine whether reported improvements reflect genuine methodological advances or simply different implementation choices. The distinction between overall study quality and explainability quality is therefore important. Several studies achieved high overall quality scores despite offering limited evidence about the validity of their explanations. This suggests that Explainability has yet to

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

mature into a rigorously evaluated software quality attribute. Strengthening future research will require the same level of methodological discipline for evaluating explanations as is already expected for evaluating task performance.

6.4. THE EVALUATION CRISIS: PERFORMANCE METRICS CANNOT MEASURE EXPLANATION QUALITY

RQ3 revealed that there seems to be a clear predominance of quantitative measures of task performance in the evaluation of explainable LLM-based RE approaches. These measures include accuracy, precision, recall, and F1-score [3], [4], [24]. They tell us how well the systems perform their tasks, but not much about the quality of the explanations. This study concurs with the conclusions drawn by two systematic reviews on XAI4SE included in our study. Mohammadkhani *et al.* [22] and Cao *et al.* [23] highlight the lack of appropriate evaluation metrics as a key systemic problem in XAI for software engineering. Our review adds one additional aspect, demonstrating that this gap remains open even in the context of RE.

The evaluation crisis directly supports problems highlighted in RQ4. Due to the absence of agreed measures for evaluating faithfulness, consistency, and usefulness of explanations, researchers can hardly distinguish an explanation generated by genuine interpretable systems from those that generate plausible-sounding explanations [52], [56]. Qualitative methods of evaluation, like expert assessment, user studies, and think-aloud protocols [6], [7], [49], [51] appear to be too limited in terms of scale and reproducibility, thus preventing systematic accumulation of evidence. To solve this problem, we need to develop a comprehensive evaluation framework that would evaluate both the performance and the quality of explanation along multiple dimensions such as faithfulness, completeness, consistency, and stakeholder-perceived usefulness [13], [14], [51].

6.5. THE MISSING STAKEHOLDER PERSPECTIVE

The near-absence of a human-centered approach in existing explainable LLM approaches for RE is a consistent gap that cuts across RQ2, RQ4, and RQ5. Requirements Engineering itself is a sociotechnical process where people with diverse backgrounds and different degrees of technical knowledge not only interact with requirement artifacts but also with explanations provided alongside them [26], [27]. However, the studies reviewed do not take into account the type of stakeholders when designing the explanations. In most cases, explanation generation systems predominantly generate explanations as generic natural language output, with no adaptive mechanism to tailor depth, format, or vocabulary to the recipient's preferences [49], [51].

The notion of on-demand explanation elicitation is defined in Deters *et al.* [51] as the development of systems that generate explanations upon demand based on user queries. It is seen as a promising solution for addressing the aforementioned gaps. According to Obaidi *et al.* [48], [49], explanation requirements have to be explicitly elicited because users' needs for explanations are highly context-dependent. In general, the evidence implies that future research should pursue more active stakeholder involvement in participatory approaches for building explainability support systems based on their needs [13], [14], [48].

6.6. RELIABILITY, HALLUCINATIONS, AND THE TRUST DEFICIT

RQ4 identified hallucinations and reliability as the most significant impediments to the deployment of explainable LLMs in RE tasks. LLM hallucinations may produce requirements artifacts that look valid but contain domain-specific errors that cannot be detected without expert analysis. Systematic analysis of hallucinations and reliability issues in LLM-based RE artifacts is reported in [53], [55]. Most importantly, post hoc XAI explanations fail to reveal LLM hallucinations or reliably fix mistakes. This is because the model's explanation may be a hallucination itself and can produce an apparently sound but actually incorrect explanation for the decisions made [5], [6].

Addressing the issue of trust deficit calls for multiple measures. To start with, LLMs should

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

be integrated into architectures with other reasoning engines such as symbolic reasoning, rule-based systems, or knowledge graphs, thereby constraining possible outputs and providing formal correctness assurances for critical RE artifacts [5], [50]. Secondly, any LLM-driven artifact, such as requirements or explanations, should undergo formal approval by domain experts as part of a human-in-the-loop validation approach [54]. In addition, rigorous frameworks for testing LLM robustness are needed to determine the behavior of LLM-based RE tools in different scenarios [5], [56].

6.7. TOWARDS A QUALITY ASSESSMENT FRAMEWORK FOR EXPLAINABLE LLMS IN REQUIREMENTS ENGINEERING

One pattern that emerges consistently across the studies reviewed is that explainable LLMs are still evaluated primarily in terms of how well they perform and not how well they explain. Most authors report traditional performance measures such as accuracy, precision, recall, and F1-score to demonstrate the effectiveness of LLM-based requirements engineering solutions. Yet, comparatively little attention is paid to whether the accompanying explanations faithfully represent the model's reasoning [6], [11], [12]. This creates an important blind spot. A model can achieve excellent predictive performance while offering explanations that are superficial, unstable, or simply persuasive narratives with little connection to the underlying decision process. The broader XAI community has long argued that explanation quality cannot be reduced to predictive accuracy alone. It also depends on properties such as faithfulness, interpretability, transparency, and user trust [11], [12], [43]. Viewed from this perspective, the current evaluation practices in RE capture only one dimension of system quality and make meaningful comparisons across studies difficult.

The evidence synthesized in this review suggests that future explainable LLM frameworks should be evaluated through a broader quality assessment model built around three complementary dimensions: explanation quality, human-centered quality, and engineering utility (Fig. 2). The first dimension examines the explanation itself. It should establish whether explanations are faithful, complete, and consistent [6], [12]. Faithfulness reflects the reasoning that actually produced the model output rather than a convincing post-hoc justification. Completeness captures the evidence that influenced the recommendation or classification, while consistency is when explanations remain stable when equivalent requirements are presented under similar conditions [6], [12], [43]. These properties determine whether an explanation can be trusted as an accurate representation of the model behavior rather than an after-the-fact narrative. The second dimension recognizes that requirements engineering is inherently collaborative. Analysts, developers, customers, and domain experts interact with the same requirements but often need different forms of explanations. Consequently, evaluation should extend beyond technical correctness to include interpretability, perceived usefulness, stakeholder trust, cognitive effort, and support for decision-making [6], [39]. This reinforces earlier work that positions Explainability as a software quality attribute and a non-functional requirement rather than merely an AI capability [48]. The third dimension shifts attention to engineering outcomes. If explanations genuinely add value, they should improve activities such as ambiguity resolution, requirements validation, traceability, review efficiency, and ultimately software quality [14], [43], [57]. Without demonstrating these practical benefits, explainability risks becoming an attractive feature whose contribution to engineering practice remains largely assumed rather than demonstrated.

The framework proposed in this study addresses another weakness that became evident throughout the review. There is a lack of a common benchmark for evaluating explainable LLMs across different RE tasks. At present, studies adopt different evaluation criteria that make it difficult to compare findings, reproduce experiments, or determine which approaches genuinely advance Explainability. Establishing shared quality dimensions would provide a more consistent basis for evaluating future systems. Thus, researchers will be encouraged to look beyond predictive performance and assess explanations as software artifacts in their own right. Furthermore, it will also support the development of benchmark datasets and reproducible evaluation protocols that combine objective measures of explanation quality with human-centered assessment. Moving in this direction helps make Explainability a clear and measurable software quality attribute. This strengthens the case for trustworthy,

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

transparent, and practical LLM solutions in requirements engineering.

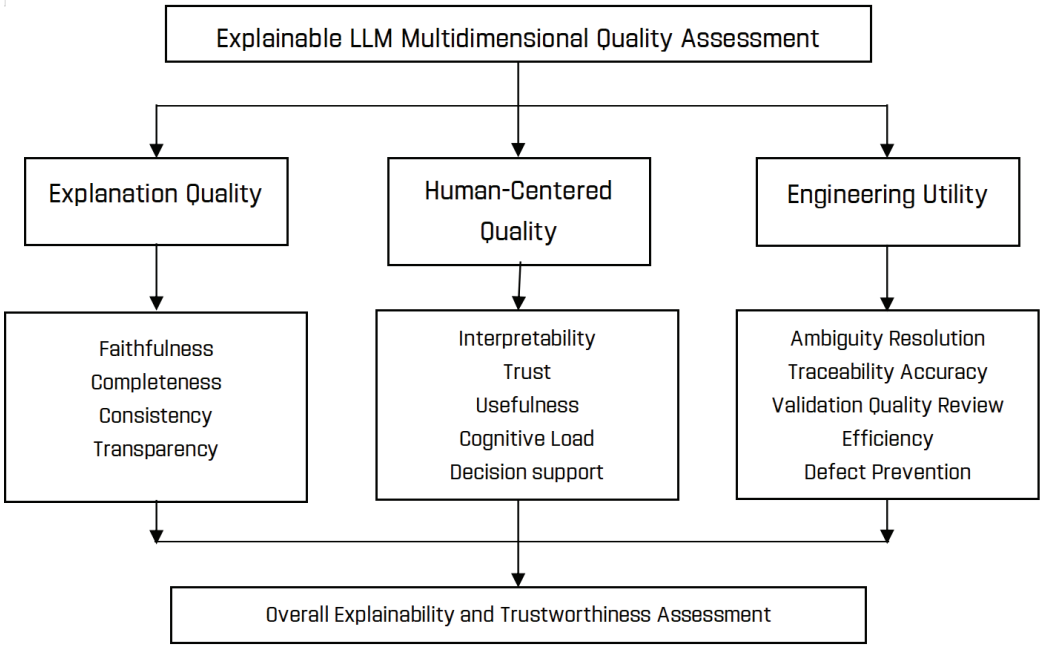


Figure 3: A conceptual quality assessment framework synthesizing the dimensions identified in the reviewed literature for evaluating explainable LLMs in requirements engineering.

The evidence summarized in Table 10 highlights a clear imbalance in current evaluation practices for explainable LLMs in RE. While most studies demonstrate strong predictive performance, far fewer examine whether the accompanying explanations faithfully represent the model's reasoning, support stakeholder understanding, or improve engineering outcomes. As a result, explanation quality is still judged largely through subjective assessments, with objective measures such as faithfulness, completeness, and consistency receiving comparatively little attention. This lack of a common evaluation approach makes it difficult to compare findings across studies or determine whether reported improvements reflect genuine advances in Explainability. Adopting standardized quality dimensions and reproducible evaluation protocols can provide a more rigorous basis for evaluating explainable LLMs. This shift allows future studies to measure explanations not only for their technical validity but also for their practical value in supporting requirements engineering activities.

Table 10: Proposed Quality Assessment Dimensions for Evaluating Explainable LLMs in Requirements Engineering

Quality Dimension	Purpose	Suggested Evaluation Metrics	Evaluation Method	Studies
Faithfulness	Determines whether the explanation generated accurately reflects the internal reasoning process of the LLM.	Sufficiency Score, Comprehensiveness Score, Deletion–Insertion Tests, Causal Fidelity	Perturbation, feature ablation, explanation consistency with model outputs	[12], [43], [6]
Completeness	Assess whether the explanation covers all relevant information and evidence that contributed to the generated requirement or recommendation.	Feature Coverage Ratio, Rationale Coverage, Information Completeness	Expert assessment, comparison with reference explanations, coverage analysis	[12], [6]
Consistency	Evaluates whether similar requirements or semantically equivalent inputs produce stable explanations under similar conditions.	Explanation, Stability Score, Similarity Index, Variance Analysis	Repeated experiments using semantically equivalent requirements or prompt variations	[43], [6]

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Quality Dimension	Purpose	Suggested Evaluation Metrics	Evaluation Method	Studies
Interpretability	Measures how easily stakeholders understand and use explanations during RE activities.	Interpretability Rating, Cognitive Load, Task Completion Time	User studies, expert reviews, questionnaires	[14], [51]
Trust	Assesses the extent to which explanations improve stakeholder confidence without encouraging over-reliance on AI recommendations.	Trust Calibration Score, Likert Trust Scale, Confidence Gain	Controlled user experiments, stakeholder surveys	[43], [48]
Usefulness	Determine whether explanations assist users in making better RE decisions.	Decision Accuracy Improvement, User Satisfaction, Error Detection Rate	Comparative human studies, case studies	[6], [39]
Engineering Utility	Evaluate the practical impact of explanations on software engineering outcomes.	Ambiguity Resolution Rate, Traceability Accuracy, Validation Accuracy, Review Time Reduction, Defect Prevention Rate	Industrial case studies, benchmark datasets, controlled experiments	[4], [57], [52], [39]

7. CONCLUSION AND FUTURE WORK

This paper conducted a systematic literature review of explainable LLMs in software RE using the PRISMA methodology, consolidating findings to answer five research questions. The paper revealed that LLMs have been employed throughout the entire range of RE activities. This shows their robust natural language processing ability, which helps eliminate manual labor while increasing consistency in the RE process. It also showed that Explainability is mainly done using prompt-based and post-hoc approaches. The study found that quantitative task-performance metrics, which are inherently inappropriate as criteria for evaluating explanations, dominated the evaluation methodology. Also, there exists no established benchmarking framework for explainable requirements elicitation. The lack of benchmarks for evaluation, the inadequacy of current explainability approaches, shortcomings in human-centered explanation design, insufficient robustness and trust management, and underdeveloped LLM integration techniques, in combination with other AI paradigms, were identified as research gaps in the review. This review consolidates the evidence and proposes a multidimensional framework for evaluating explainable LLMs in RE.

In future research, XAI should become an integral concern in LLM-based RE instead of being addressed as an afterthought. The most pressing methodological requirement in this context would be dedicated benchmarks for evaluating explanations in RE, with datasets and metrics established to assess the quality dimensions of interpretability in terms of faithfulness, completeness, consistency, and perceived usefulness by relevant stakeholders. Without these benchmarks, there would be no way to differentiate true interpretability from plausible-sounding explanations. Another priority would be to progress from prompt-based and post-hoc XAI approaches to explainability paradigms that incorporate intrinsic and hybrid Explainability. These approaches can produce faithful explanations grounded in the model's actual reasoning and thus directly address the capability-interpretability imbalance. XAI processes should also be human-centered and stakeholder-specific. The degree and modality of explanations provided by the XAI methods should vary depending on the expertise and role of each RE stakeholder. Furthermore, the validity of the XAI output should be justified through human-in-the-loop validation procedures, robustness testing, and hallucination detection capabilities. This will ensure the veracity and truthfulness of the LLM-generated explanations of requirements artifacts. Finally, there needs to be specific interdisciplinary research on ethics and governance issues around explainable AI in LLM-driven RE for the use of such tools to be justified in terms of their reliability in safety-critical and regulated industrial contexts.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

ACKNOWLEDGEMENT

The authors wish to acknowledge Murang'a University of Technology for providing access to the digital library databases used in the systematic literature search. No external funding was received for this research.

REFERENCES

[1] S. Rasnayaka, G. Wang, R. Shariffdeen, and G. N. Iyer, "An Empirical Study on Usage and Perceptions of LLMs in a Software Engineering Project," in *1st International Workshop on Large Language Models for Code (LLM4Code '24)*, New York, NY, USA: Association for Computing Machinery (ACM), Apr. 2024, pp. 111–118.

[2] C. Ellsel and R. Stark, "Advancing Requirements Engineering With Large Language Models," *Procedia CIRP*, vol. 136, pp. 701–706, 2025, doi: 10.1016/j.procir.2025.08.120.

[3] N. Alturayeif, I. Ahmad, and J. Hassine, "TraceLLM: Leveraging Large Language Models With Prompt Engineering for Enhanced Requirements Traceability," Feb. 01, 2026. doi: 10.1007/s00766-026-00460-1.

[4] S. Bashir *et al.*, "Requirements Ambiguity Detection and Explanation With LLMS: An Industrial Study," in *41st IEEE International Conference on Software Maintenance and Evolution (ICSME 2025)*, Auckland, New Zealand: IEEE, Sep. 2025, pp. 620–631.

[5] Z. Yang, Z. Sun, T. Zhuo Yue, P. Devanbu, and D. Lo, "Robustness, Security, Privacy, Explainability, Efficiency, and Usability of Large Language Models for Code," Mar. 12, 2024. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2403.07506v1

[6] A. Vitale, K.-N. Nguyen, D. Poshyvanyk, R. Oliveto, S. Scalabrino, and A. Mastropaolo, "Toward Explaining Large Language Models in Software Engineering Tasks," Dec. 23, 2025. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2512.20328v1

[7] S. T. U. Shah, A. Barcomb, and H. Leung, "EQDAT-RE: An Explainable Qualitative Data Analysis for Transparent Requirements Engineering," in *2025 IEEE Smart World Congress (SWC)*, Calgary, AB, Canada: IEEE, Aug. 2025, pp. 1328–1335.

[8] A. Hemmat, M. Sharbaf, S. Kolahdouz-Rahimi, K. Lano, and S. Y. Tehrani, "Research Directions for Using LLM in Software Requirement Engineering: A Systematic Review," *Frontiers in Computer Science*, vol. 7, Mar. 2025, doi: 10.3389/fcomp.2025.1519437.

[9] X. Hou *et al.*, "Large Language Models for Software Engineering: A Systematic Literature Review," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 8, pp. 1–79, Nov. 2024, doi: 10.1145/3695988.

[10] M. Amin Zadenoori, J. Dąbrowski, W. Alhoshan, L. Zhao, and A. Ferrari, "Large Language Models (LLMs) for Requirements Engineering (RE): A Systematic Literature Review," Sep. 14, 2025. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2509.11446v1

[11] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, Oct. 2018, doi: 10.1109/access.2018.2870052.

[12] A. B. Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI," *Information Fusion*, vol. 58, no. 1, pp. 82–115, Jun. 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

[13] L. Chazette, W. Brunotte, and T. Speith, "Explainable Software Systems: From Requirements Analysis to System Evaluation," *Requirements Engineering*, vol. 27, no. 4, pp. 457–487, Nov. 2022, doi: 10.1007/s00766-022-00393-5.

[14] L. Chazette and K. Schneider, "Explainability as a Non-Functional Requirement: Challenges and Recommendations," *Requirements Engineering*, vol. 25, no. 4, pp. 493–514, Jun. 2020, doi: 10.1007/s00766-020-00333-1.

[15] C. Tantithamthavorn, J. Jiarpakdee, and J. Grundy, "Explainable AI for Software Engineering," Dec. 03, 2020. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2012.01614v1

[16] J. A. Khan, S. Qayyum, and H. S. Dar, "Large Language Model for Requirements Engineering: A Systematic Literature Review," *Research Square (Research Square)*, vol. 4, Mar. 2025, doi: 10.21203/rs.3.rs-5589929/v1.

[17] H. Cheng *et al.*, "Generative AI for Requirements Engineering: A Systematic Literature Review," *Software Practice and Experience*, vol. 56, no. 2, pp. 141–170, Nov. 2025, doi: 10.1002/spe.70029.

[18] C. Arora, J. Grundy, and M. Abdelrazek, "Advancing Requirements Engineering Through Generative AI: Assessing the Role of LLMs," Oct. 21, 2023. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2310.13976v2

[19] A. A. Magableh, "Towards Leveraging Explainable Artificial Intelligent (XAI) in Requirements Engineering (RE) to Identify Aspect (Crosscutting Concern): A Systematic Literature Review (SLR) and Bibliometric Analysis," in *International Conference on Information Technology: Cybersecurity Challenges for Sustainable Cities (ICIT)*, IEEE, Aug. 2023, pp. 319–326.

[20] L. Zhao *et al.*, "Natural Language Processing for Requirements Engineering," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–41, Apr. 2021, doi: 10.1145/3444689.

[21] S.-C. Necula, F. Dumitriu, and V. Greavu-Şerban, "A Systematic Literature Review on Using Natural Language Processing in Software Requirements Engineering," *Electronics*, vol. 13, no. 11, p. 2055, Jan. 2024, doi: 10.3390/electronics13112055.

[22] A. H. Mohammadkhani, B. N. Sai, M. Daboussi, O. Sabnis, C. Tantithamthavorn, and H. Hemmati, "A Systematic Literature Review of Explainable AI for Software Engineering," *arXiv.org*, 2023, Accessed: Sep. 07, 2026. [Online]. Available: http://arxiv.org/abs/2302.06065

[23] S. Cao *et al.*, "A Systematic Literature Review on Explainability for Machine/Deep Learning-Based Software Engineering Research," Jan. 26, 2024. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2401.14617v2

[24] N. Van Viet and N. T. Vinh, "Large Language Models in Software Engineering," *Journal of Education For Sustainable Innovation*, vol. 2, no. 2, pp. 146–156, Dec. 2024, doi: 10.56916/jesi.v2i2.968.

[25] Z. Zheng *et al.*, "Towards an Understanding of Large Language Models in Software Engineering Tasks," *Empirical Software Engineering*, vol. 30, no. 2, Dec. 2024, doi: 10.1007/s10664-024-10602-0.

[26] I. Udousoro, "Effective Requirement Engineering Process Model in Software Engineering," *Software Engineering*, vol. 8, no. 1, pp. 1–5, 2020, doi: 10.11648/j.se.20200801.11.

[27] S. Veena, N. A. S. Vinoth, A. M. Nancy, G. S. Kumar, and R. T. Teja, "Effective System

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

for Software Requirement Management," in *AIP Conference Proceedings*, American Institute of Physics Inc, 2020, p. 240010.

[28] I. M. Siddique, "Requirements Engineering at a Turning Point: Key Research Trends Shaping the Future of Systems and Software Integration," *Journal of Geotechnical Studies*, vol. 9, no. 3, pp. 37–49, Oct. 2024, doi: 10.46610/jogs.2024.v09i03.005.

[29] S. Bin Basri, G. Kumar, F. F. Fahrurazi, P. Emieldza Balqis Azmi, A. O. Balogun, and H. Mamman, "Current Trend of Software Requirement Engineering Process in IT Small and Medium Enterprises (SMEs)-a Systematic Literature Review," in *13th International Conference on Information Technology in Asia (CITA)*, Institute of Electrical and Electronics Engineers Inc, Aug. 2023, pp. 82–87.

[30] K. Ronanki, S. Arvidsson, and J. Axell, "Prompt Engineering Guidelines for Using Large Language Models in Requirements Engineering," arXiv.org. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2507.03405

[31] M. Ebrahim, S. Guirguis, and C. Basta, "Enhancing Software Requirements Engineering With Language Models and Prompting Techniques: Insights From the Current Research and Future Directions," in *63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, Association for Computational Linguistics, Jan. 2025, pp. 486–496. Accessed: Sep. 07, 2026. [Online]. Available: https://aclanthology.org/2025.acl-srw.31/?utm_source=chatgpt.com

[32] H. Saeeda, J. Dong, Y. Wang, and M. A. Abid, "A Proposed Framework for Improved Software Requirements Elicitation Process in SCRUM: Implementation by a Real-Life Norway-Based IT Project," *Journal of Software: Evolution and Process*, no. 7, Feb. 2020.

[33] M. A. Sami, Z. Rasheed, M. Waseem, Z. Zhang, T. Herda, and P. Abrahamsson, "Prioritizing Software Requirements Using Large Language Models," *arXiv (Cornell University)*, Apr. 2024, doi: 10.48550/arxiv.2405.01564.

[34] C. Almeida *et al.*, "From Elicitation Interviews to Software Requirements: Evaluating LLM Performance in Requirement Generation," in *Proceedings of the 28th Workshop on Requirements Engineering (WER2025)*, Rio de Janeiro, Brazil: WER / Brazilian software engineering community, 2025.

[35] M. K. Habib, D. Graziotin, and S. Wagner, "ReqBrain: Task-Specific Instruction Tuning of LLMs for AI-Assisted Requirements Generation," May 23, 2025. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2505.17632v1

[36] Z. Wang, C. Feng, L. Liu, G. Jiao, and P. Ye, "The Application of LLMs in the Analysis and Modeling of Software Requirements," in *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, IEEE, Jul. 2024, pp. 1143–1153.

[37] G. Rejithkumar and P. R. Anish, "NICE: Non-Functional Requirements Identification, Classification, and Explanation Using Small Language Models," in *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, Ottawa: IEEE, Apr. 2025, pp. 284–295.

[38] R. Feldt and R. Coppola, "Semantic API Alignment: Linking High-Level User Goals to APIs," May 07, 2024. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2405.04236v1

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

[39] A. Ferrari, S. Abualhaija, and C. Arora, "Model Generation With LLMs: From Requirements to UML Sequence Diagrams," in *32nd IEEE International Requirements Engineering Conference Workshops, REW 2024, IEEE*, Jun. 2024, pp. 291–300. Accessed: Sep. 07, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10628665

[40] D. Xie *et al.*, "How Effective Are Large Language Models in Generating Software Specifications?," in *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Mar. 2025, pp. 1–12. Accessed: Sep. 07, 2026. [Online]. Available: https://doi.org/10.1109/SANER64311.2025.00014

[41] M. Roy, N. Deb, A. Cortesi, and N. Chaki, "Mapping System Requirements to Emotional Impact: A Semi-Automated Approach," *Innovations in Systems and Software Engineering*, vol. 21, no. 4, pp. 1521–1542, Aug. 2025, doi: 10.1007/s11334-025-00618-0.

[42] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable Artificial Intelligence: An Analytical Review," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 5, Jul. 2021, doi: 10.1002/widm.1424.

[43] V. Chamola, V. Hassija, A. R. Sulthana, D. Ghosh, D. Dhingra, and B. Sikdar, "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 11, pp. 78994–79015, 2023, doi: 10.1109/ACCESS.2023.3294569.

[44] A. Singhal, P. Pratap, K. K. Dixit, and K. Kathuria, "Advancements in Explainable AI: Bridging the Gap Between Model Complexity and Interpretability," in *2024 2nd International Conference on Disruptive Technologies, ICDT 2024, Institute of Electrical and Electronics Engineers Inc*, Mar. 2024, pp. 675–680. Accessed: Sep. 07, 2026. [Online]. Available: https://ieeexplore.ieee.org/document/10489277

[45] Y. V. Shankheshwaria and D. B. Patel, "Explainable AI in Machine Learning: Building Transparent Models for Business Applications," *Frontiers in Emerging Artificial Intelligence and Machine Learning*, vol. 02, no. 08, pp. 08–15, Aug. 2025, doi: 10.37547/feaiml/volume02issue08-02.

[46] L. Petrillo, F. Martinelli, A. Santone, and F. Mercaldo, "Explainable Security Requirements Classification Through Transformer Models," *Future Internet*, vol. 17, no. 1, p. 15, Jan. 2025, doi: 10.3390/fi17010015.

[47] G. C. Publio and J. E. L. Gayo, "xpSHACL: Explainable SHACL Validation Using Retrieval-Augmented Generation and Large Language Models," Jul. 11, 2025. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2507.08432v1

[48] M. Obaidi *et al.*, "How to Elicit Explainability Requirements? A Comparison of Interviews, Focus Groups, and Surveys," in *Proceedings of the IEEE International Conference on Requirements Engineering*, IEEE Computer Society, Sep. 2025, pp. 167–178. Accessed: Sep. 07, 2026. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/11190364?casa token=a6wNG5lS4NIAAAAA:SeAphJw4E7r6Gf18tNxg-bUyBNawoQKwV2clEeyhRPlmHjfxqv84Z7sAktswtcB_wje3D9_RW23s

[49] M. Obaidi *et al.*, "Automatic Generation of Explainability Requirements and Software Explanations From User Reviews," in *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)*, IEEE, Jul. 2025. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2507.07344v1

[50] A. Kleijwegt, S. G. Yaman, and R. Calinescu, "Tool for Supporting Debugging and Understanding of Normative Requirements Using LLMs," pp. 576–579, Sep. 2025, doi: 10.1109/re63999.2025.00068.

[51] H. Deters, J. Droste, M. Fechner, and J. Klünder, "Explanations on Demand - a Technique for Eliciting the Actual Need for Explanations," in *2023 IEEE 31st International*

http://dx.doi.org/10.21622/ACE.2026.06.2.2265

Requirements Engineering Conference Workshops (REW), Institute of Electrical and Electronics Engineers Inc, Sep. 2023, pp. 345–351.

[52] M. Krishna, B. Gaur, A. Verma, and P. Jalote, "Using LLMs in Software Requirements Specifications: An Empirical Evaluation," pp. 475–483, Jun. 2024, doi: 10.1109/re59067.2024.00056.

[53] J. J. Norheim, E. Rebentisch, D. Xiao, L. Draeger, A. Kerbrat, and O. L. de Weck, "Challenges in Applying Large Language Models to Requirements Engineering Tasks," *Design Science*, vol. 10, no. 16, 2024, doi: 10.1017/dsj.2024.8.

[54] C. Hymel and H. Johnson, "Analysis of LLMs Vs Human Experts in Requirements Engineering," arXiv.org. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2501.19297

[55] H. Jin and H. Chen, "Uncovering Systematic Failures of LLMs in Verifying Code Against Natural Language Specifications," in *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, Nov. 2025, pp. 3819–3823. Accessed: Sep. 07, 2026. [Online]. Available: https://www.researchgate.net/publication/394540800_Uncovering_Systematic_Failures_of_LLMs_in_Verifying_Code_Against_Natural_Language_Specifications

[56] M. Vieira, "Engineering Trustworthy Software: A Mission for LLMs," Nov. 27, 2024. Accessed: Sep. 07, 2026. [Online]. Available: https://arxiv.org/abs/2411.17981v1

[57] S. Lubos *et al.*, "Leveraging LLMs for the Quality Assurance of Software Requirements," *arXiv (Cornell University)*, pp. 389–397, Jun. 2024, doi: 10.1109/re59067.2024.00046.