

Generative AI for Requirements Engineering: A Systematic Literature Review

Haowei Cheng^a, Jati H. Husen^{a,b}, Sien Reeve Peralta^a, Bowen Jiang^a, Nobukazu Yoshioka^a, Naoyasu Ubayashi^a, Hironori Washizaki^a

^aWaseda University, Tokyo, 169-8050, Japan

^bTelkom University, Bandung, 140257, Jawa Barat, Indonesia

Abstract

Context: Generative AI (GenAI) has emerged as a transformative tool in software engineering, with requirements engineering (RE) actively exploring its potential to revolutionize processes and outcomes. The integration of GenAI into RE presents both promising opportunities and significant challenges that necessitate systematic analysis and evaluation.

Objective: This paper presents a comprehensive systematic literature review (SLR) analyzing state-of-the-art applications and innovative proposals leveraging GenAI in RE. It surveys studies focusing on the utilization of GenAI to enhance RE processes while identifying key challenges and opportunities in this rapidly evolving field.

Method: A rigorous SLR methodology was used to conduct an in-depth analysis of 27 carefully selected primary studies. The review examined research questions pertaining to the application of GenAI across various RE phases, the models and techniques used, and the challenges encountered in implementation and adoption.

Results: The most salient findings include i) a predominant focus on the early stages of RE, particularly the elicitation and analysis of requirements, indicating potential for expansion into later phases; ii) the dominance of large language models, especially the GPT series, highlighting the need for diverse AI approaches; and iii) persistent challenges in domain-specific applications and the interpretability of AI-generated outputs, underscoring areas requiring further research and development.

Conclusions: The results highlight the critical need for comprehensive evaluation frameworks, improved human-AI collaboration models, and thorough consideration of ethical implications in GenAI-assisted RE. Future research should prioritize extending GenAI applications across the entire RE lifecycle, enhancing domain-specific capabilities, and developing strategies for responsible AI integration in RE practices.

Keywords: Generative AI, Requirements Engineering, Systematic Literature

1. Introduction

Software engineering (SE) faces escalating challenges as systems grow in complexity and scale. Modern software must meet diverse functional requirements while ensuring reliability, security, and maintainability [1]. The discipline of SE encompasses the entire software lifecycle, from requirements elicitation to system maintenance, aiming to enhance development efficiency and quality through systematic methods and technical strategies. Despite significant advancements, complex projects frequently encounter issues such as delays, cost overruns, and system defects. Lederer et al. demonstrated through empirical studies that frequent change requests from users and their lack of understanding of requirements are primary contributors to cost overruns [2]. In addition, the Standish Group’s “2020 CHAOS Report” reveals that only 31% of software projects are completed on time and within budget, with a mere 46% delivering high-value returns [3]. These findings underscore the critical need to optimize the requirements engineering (RE) phase to improve the success rate of software projects and reduce costs.

The integration of AI into SE (AI for SE) has transformed traditional practices, particularly in areas such as code generation, defect prediction, and software testing, substantially improving efficiency and quality across development processes [4, 5, 6]. AI’s ability to automate and enhance these tasks has paved the way for more intelligent and adaptive SE practices. Building on these advancements, current research is increasingly focused on extending the benefits of AI to requirements engineering (AI for RE), with the goal of addressing the unique challenges in eliciting, analyzing, and validating software requirements. This shift marks an important evolution from AI’s traditional role in SE toward a more specialized focus on using generative AI (GenAI) to optimize RE practices. These AI-based approaches in SE lay the foundation for the application of more advanced GenAI techniques, such as large language models (LLMs), in various SE tasks. The success of AI in improving software development processes and quality sets the stage for exploring the potential of GenAI in the specific domain of RE.

RE holds a critical role within SE because it focuses on the systematic elicitation, analysis, specification, validation, and management of both functional and non-functional requirements [7]. The importance of RE in the software development lifecycle is widely recognized and has been codified in international standards such as ISO/IEC/IEEE 29148:2018. This standard provides a unified framework and best-

practice guidelines for the RE process, emphasizing its crucial role in project success [8].

As illustrated in Figure 1, RE encompasses five primary components:

1. **Elicitation:** This component involves gathering requirements through various techniques such as interviews, workshops, and surveys. These methods help capture stakeholders' needs and expectations.
2. **Analysis:** This stage includes modeling the gathered requirements and prioritizing them on the basis of their importance and feasibility.
3. **Specification:** Here, requirements are documented in various formats, including user stories, use cases, and formal specifications, ensuring clear communication of system expectations.
4. **Validation:** This crucial step involves reviewing the specified requirements, prototyping, and testing to ensure they accurately reflect stakeholders' needs and are feasible to implement.
5. **Management:** This overarching activity includes change control, traceability, and version control, ensuring that requirements remain consistent and up-to-date throughout the project lifecycle.

The diagram also highlights the importance of stakeholders and various tools and techniques that support the entire RE process. The primary goal of RE is to ensure that the software system aligns with stakeholders' needs, thereby reducing project risks and improving system usability and user satisfaction [9]. High-quality RE is a key determinant in the success of software projects, providing the foundation for subsequent design, implementation, and testing activities.

However, traditional RE methods often face challenges related to efficiency and accuracy, especially when addressing rapidly evolving and increasingly complex requirements. As modern software systems grow in scale and intricacy, enhancing the quality and effectiveness of RE processes remains a pressing issue in contemporary SE [10].

The evolution of RE has been marked by several distinct phases, each driven by the need to address increasingly complex software development challenges [7]. Initially, traditional RE methods relied heavily on manual processes, stakeholder interviews, and document-centric approaches. As software systems grew more intricate, these methods proved insufficient for capturing and managing the full spectrum of requirements. This led to the rise of model-driven RE, which introduced visual

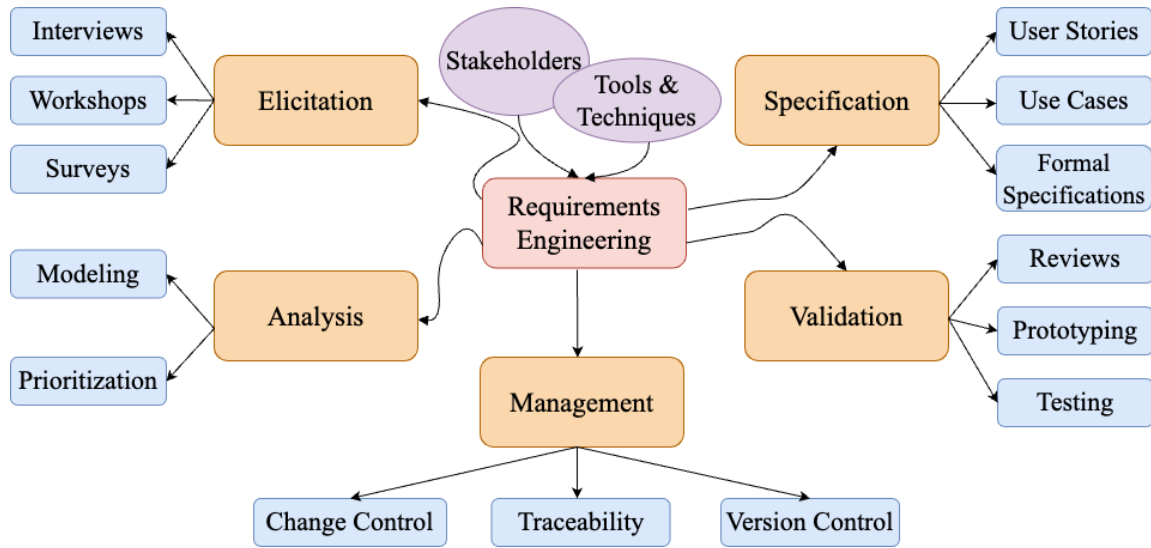


Figure 1: Concept of requirements engineering

representations and formal specifications to improve clarity and traceability [11]. In response to the growing complexity and dynamic nature of modern software systems, RE has gradually evolved into AI-driven approaches that automate many of the repetitive tasks, improve analysis, and handle larger volumes of data with greater efficiency [12]. Recently, this evolution has taken another significant step with the introduction of GenAI for RE. By leveraging advanced technologies such as LLMs, GenAI for RE represents a new paradigm that not only enhances traditional RE activities but also opens new possibilities for automating the generation, refinement, and analysis of requirements [13, 14]. This shift towards GenAI for RE marks a transformative moment in the field, offering unprecedented potential to tackle the increasing complexity of requirements and the evolving needs of stakeholders.

GenAI refers to a class of advanced AI systems capable of generating new content (e.g., text, images, music, and even code) on the basis of patterns learned from extensive training data [15]. Powered by cutting-edge deep learning techniques and neural networks, including LLMs such as GPT-3 and GPT-4, GenAI has made significant strides in fields such as natural language processing (NLP), image generation, and creative content creation.

As illustrated in Figure 2, GenAI is built upon three main pillars: deep neural networks, machine learning algorithms (especially deep learning techniques), and large-scale training data. Deep neural networks, particularly advanced architectures such as transformers and generative adversarial networks (GANs), form the backbone

of GenAI systems. A crucial component of these networks is LLMs, such as GPT-3 and GPT-4, which have revolutionized natural language processing and generation.

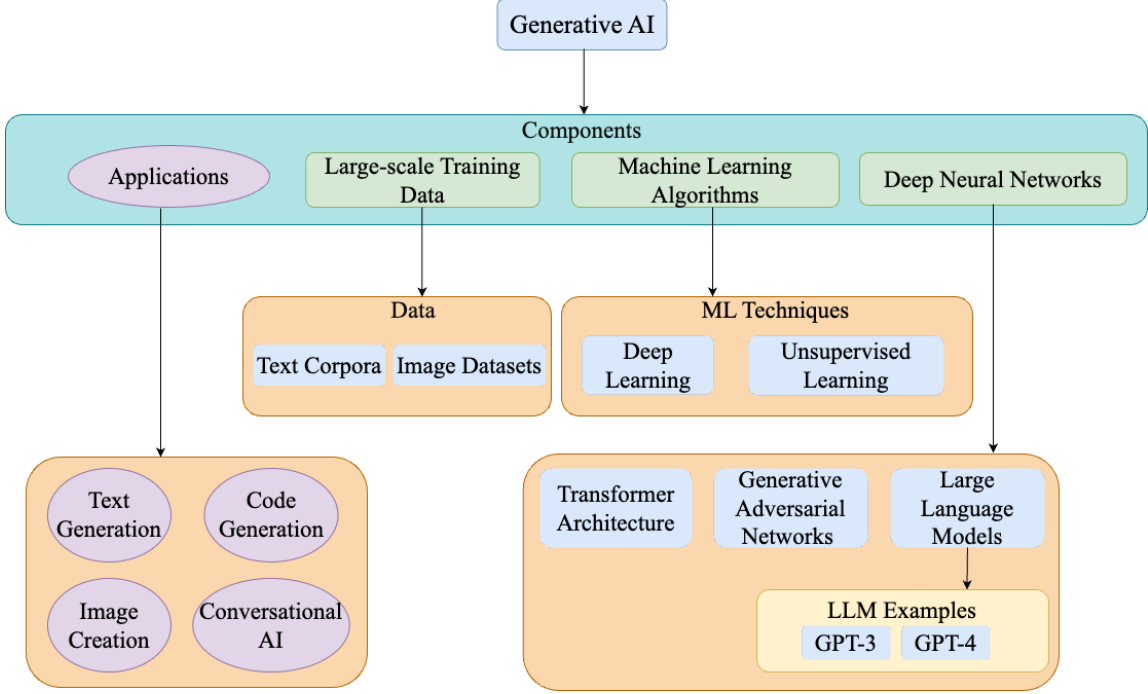


Figure 2: Concept of generative AI

These LLMs are trained on vast datasets, enabling them to generate text, images, or other forms of content similar to that generated by humans by understanding the context provided. One of the most notable applications of GenAI is ChatGPT, a conversational AI model that can engage in dialogue, answer questions, and assist with a wide range of tasks across various domains. This transformative technology continues to revolutionize industries by offering coherent, contextually relevant, and diverse outputs across multiple sectors [16]. These networks are trained using sophisticated machine learning algorithms, with emphasis on deep learning techniques and including methods such as unsupervised learning. Large-scale training data consisting of vast text corpora and image datasets provide the foundation for these models to learn patterns and generate new content. The applications of GenAI, as shown in the diagram, span various domains, including text generation, image creation, code generation, and, notably, conversational AI. The diagram illustrates how the core components of GenAI—deep neural networks (including LLMs), advanced machine learning algorithms, and large-scale training data—converge to enable these

diverse applications, showcasing the technology’s potential to affect numerous fields and industries.

Figure 3 illustrates the intricate relationship between GenAI capabilities and the RE process. The diagram is structured into two main parts: the RE process and GenAI capabilities. The RE process is depicted with five key components: elicitation, analysis, specification, validation, and management. Each of these components is shown to receive input from various sources, such as stakeholder input for elicitation, domain knowledge for analysis, and change requests for management. On the GenAI side, the diagram showcases four primary capabilities: NLP, pattern recognition, predictive analytics, and automated documentation. Nested within these capabilities is a subgroup representing LLMs; these capabilities include text generation, context understanding, and language translation. The diagram illustrates how these GenAI capabilities directly influence and enhance various stages of the RE process. For instance, NLP is shown to support both elicitation and specification, whereas predictive analytics contributes to analysis and validation. Automated documentation is depicted as aiding both specification and management processes.

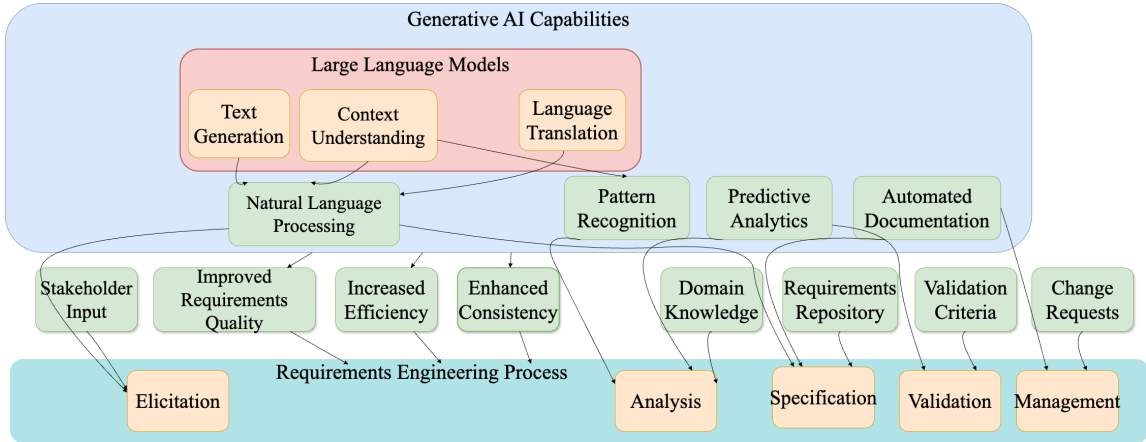


Figure 3: Overview of generative AI integration in requirements engineering

The integration of GenAI into the RE process is shown to yield several positive outcomes, including improved requirements quality, increased efficiency, enhanced consistency, and improved traceability. These outcomes are directly linked to the RE process, emphasizing the transformative effect of GenAI on RE practices. The development of GenAI has been propelled by advancements in machine learning architectures, particularly the transformer architecture that forms the backbone of modern LLMs. These models, trained on vast datasets, can understand and produce text similar to that produced by humans, leading to widespread applications

across industries such as healthcare, finance, entertainment, and education. GenAI’s ability to generate and manipulate complex data patterns has opened new avenues for automating and improving tasks that traditionally required considerable human intervention.

In the context of RE, the advent of GenAI introduces novel opportunities to enhance and transform RE processes. When the power of LLMs is leveraged, key RE activities—including requirements elicitation, analysis, specification, validation, and management—can be automated or augmented [14]. This possibility is clearly illustrated in Figure 3, where each component of the RE process is shown to benefit from specific GenAI capabilities. For example, GenAI can automatically generate requirements based on stakeholder input, as depicted by the connection between NLP and the elicitation process. The ability to detect inconsistencies or ambiguities in requirements documents is represented by the link between “pattern recognition” and the “analysis phase.” In addition, GenAI can assist in creating user stories, use cases, and other requirement artifacts, substantially reducing the time and effort required from RE practitioners. This capability is reflected in Figure 3 through the connection between “automated documentation” and “specification.” The diagram also highlights how GenAI contributes to the management aspect of RE, which is crucial for handling change requests and maintaining traceability throughout the project lifecycle. This relationship is represented by the link between “automated documentation” and the “management” component, as well as the “better traceability” outcome. By enhancing both the efficiency and effectiveness of RE activities, as shown by the multiple positive outcomes in the diagram, GenAI has the potential to improve software quality, mitigate project risks, and ultimately lead to more successful software development outcomes.

The rapid advancements in GenAI and its potential to transform RE make prioritizing research in this area imperative. Exploring how LLMs can be used to streamline and optimize RE tasks while ensuring the quality, consistency, and reliability of the generated requirements is critical. At the same time, addressing the limitations and challenges of integrating GenAI into the RE process, including concerns related to automation accuracy, contextual understanding, and stakeholder alignment, is also essential [17]. As software systems become increasingly complex and interconnected, managing and tracing requirements throughout the development lifecycle becomes a key challenge. GenAI holds the potential to substantially enhance requirements traceability, change impact analysis, and consistency checking, helping maintain clarity and cohesion in evolving systems. Investigating how LLMs can be used for these purposes may yield substantial benefits for project management, software maintenance, and overall system quality. In recent years, the integration

of GenAI into RE has garnered increasing attention. This integration presents opportunities to automate the generation and refinement of requirements, assist in analyzing complex specifications, and improve the consistency and completeness of requirements documentation. With leveraging of GenAI, human effort in these tasks can be dramatically reduced while improving accuracy and minimizing errors.

This systematic literature review seeks to explore the current state of research on the application of GenAI in RE. Specifically, it aims to examine and synthesize findings from 27 selected papers to identify the key contributions of GenAI toward improving RE practices, as well as the challenges and limitations that remain. Through a comprehensive analysis of the literature, this review highlights the strengths, weaknesses, and gaps in current research, offering a clear picture of how GenAI is transforming RE. In addition, this review proposes future research directions to guide advancements in this evolving field, with the expectation that the provided insights will help shape the next wave of innovation at the intersection of GenAI and RE.

The remainder of this paper is structured as follows: Section 2 presents the related work, providing context on AI in RE, GenAI in SE, and GenAI in RE. Section 3 outlines the research questions that guide this systematic literature review. Section 4 details the research methodology adopted for this study, including the search strategy, inclusion and exclusion criteria, and data extraction process. Section 5 presents the analysis and discussion of our findings, where each research question is addressed in turn and the current trends, predominant approaches, quality assessment, and future directions in GenAI for RE are explored. Section 6 discusses the threats to the validity of our study. Section 7 explores the challenges and future directions for GenAI in RE. Finally, Section 8 concludes the paper, summarizing the key insights and outlining future research opportunities.

2. Related Work

The application of AI in SE has emerged as a rapidly evolving field, with recent advancements in GenAI opening new frontiers and reshaping traditional practices. This section provides a comprehensive overview of the existing research landscape, exploring the transformative potential of AI and GenAI applications within the domains of SE and RE. We focus on key areas that are closely related to our research topic: AI in RE, GenAI in SE, and GenAI in RE. By contextualizing our study within this broader technological landscape, we aim to highlight the significance and timeliness of our research contributions.

Several systematic literature reviews have summarized the state-of-the-art in AI for SE. For example, Brar and Nandal [18] provide a comprehensive review of ma-

chine learning techniques applied to SE, covering topics such as defect prediction, effort estimation, and code smell detection. Their review highlights the potential of these techniques in improving software quality and development processes while also identifying challenges related to data availability, model interpretability, and generalizability. They focus on the use of deep learning in SE and highlight its applications in code generation, program repair, and software testing [19].

These AI-based approaches in SE lay the foundation for the application of more advanced GenAI techniques, such as LLMs, in various SE tasks. The success of AI in improving software development processes and quality sets the stage for exploring the potential of GenAI in the specific domain of RE.

2.1. AI in Requirements Engineering (AI for RE)

In the domain of RE, AI has emerged as a promising solution for automating and enhancing various RE tasks, addressing the challenges associated with the increasing complexity and scale of modern software systems. Feng et al. proposed an innovative AI-driven approach for requirements classification and prioritization, their approach significantly reduces the manual effort required in these critical processes [20]. This seminal work demonstrates the potential of AI in streamlining requirements management and enabling more effective decision-making in software development projects. In addition, AI techniques have been successfully used in requirements traceability, with studies like that of Rahimi et al. [21] showcasing improved accuracy in linking requirements to other software artifacts, such as design documents and test cases. These advancements highlight the transformative potential of AI in optimizing RE activities, ensuring requirement consistency, and facilitating effective communication among stakeholders. By leveraging the power of AI, researchers and practitioners can unlock new possibilities for automating and streamlining RE processes, ultimately leading to higher-quality software systems that better align with user needs and expectations.

Lovrencic et al. [22] have published a systematic literature review on the use of NLP in RE, covering topics such as requirements classification, information extraction, and requirements quality assessment. Their review highlights the potential of NLP techniques in reducing the manual effort required in RE tasks and improving the quality of requirements documents. However, it also identifies several challenges, such as the lack of RE-specific NLP tools and the need for domain-specific knowledge in applying NLP to RE. In another recent review, Marques et al. [23] focus specifically on the use of ChatGPT, an LLM, in RE. They discuss the potential applications of ChatGPT in requirements elicitation, analysis, and validation, as well as its limitations and challenges, such as the need for human oversight and the potential for

biased or inconsistent outputs.

Future research directions in AI for RE should focus on addressing these challenges. A pressing need exists to develop RE-specific AI tools that can better handle the complexities of requirements elicitation, analysis, and management. In addition, research efforts should be directed toward enhancing the domain adaptability of AI models, enabling them to process and interpret domain-specific requirements across various industries effectively. Improving the explainability and transparency of AI-generated outputs in RE is another critical area for future work because it directly affects the trustworthiness and adoption of these technologies in practice.

The application of AI techniques in RE has demonstrated significant potential in automating and enhancing various RE tasks. These advancements pave the way for the exploration of more sophisticated GenAI models, which could further revolutionize how requirements are elicited, analyzed, and managed.

2.2. Generative AI in Software Engineering (GenAI for SE)

The emergence of GenAI models, particularly LLMs, has ushered in a new era of possibilities in SE, enabling the automated generation of code, documentation, and other software artifacts. Mastropaolo et al. conducted a groundbreaking study on the use of Text-to-Text Transfer Transformer to support code-related tasks, highlighting the immense potential of generative models in enhancing software development processes [24]. This seminal work paves the way for more intelligent and efficient code generation techniques, reducing the manual effort required in software development. Similarly, Fried et al. introduced InCoder, a state-of-the-art generative model for code infilling and synthesis, and demonstrated its effectiveness in generating code snippets based on natural language descriptions [25]. This innovative approach showcases the potential of GenAI in enabling more natural and intuitive ways of expressing software requirements and specifications. However, challenges remain in ensuring the reliability and security of AI-generated code, as highlighted by Zong et al. [26]. These studies underscore the immense potential of GenAI in automating and augmenting various SE tasks while also emphasizing the need for further research to address the associated challenges and ensure the trustworthiness and robustness of GenAI-assisted software development.

A comprehensive research agenda by Nguyen-Duc et al. [27] has identified 78 open research questions across 11 areas of SE where GenAI can be applied. Their agenda covers a wide range of topics, including RE, software design, implementation, quality assurance, maintenance, processes, project management, professional competencies, education, macro aspects, and fundamental concerns of GenAI in SE. This research agenda highlights the need for further research to address challenges associ-

ated with GenAI-assisted software development. These challenges include industry-level evaluation, reliability and correctness concerns, data availability, explainability, and sustainability aspects. The agenda serves as a valuable resource for researchers and practitioners to guide future research and development efforts in this rapidly evolving field.

Another systematic literature review by Li et al. [28] focuses specifically on the use of LLMs for code generation. The review introduces a taxonomy to categorize and discuss recent developments in code LLMs, covering aspects such as data curation, latest advances, performance evaluation, and real-world applications. The authors provide a historical overview of the evolution of LLMs for code generation and offer an empirical comparison that uses widely recognized benchmarks to highlight the progressive enhancements in LLM capabilities. They identify critical challenges and promising opportunities regarding the gap between academia and practical development, emphasizing the need for a dedicated resource to continuously document and disseminate the most recent advances in the field.

These studies highlight the immense potential of GenAI in automating and augmenting various SE tasks while also emphasizing the need for further research to address the associated challenges and ensure the trustworthiness and robustness of GenAI-assisted software development. Our work builds upon these existing reviews, focusing specifically on the application of GenAI techniques in the context of RE. The successful application of GenAI techniques in various SE tasks, such as code generation and documentation, highlights the potential for similar approaches to be applied in the RE domain. The lessons learned and challenges identified in GenAI for SE provide valuable insights for researchers and practitioners exploring the use of GenAI in RE.

2.3. Generative AI in Requirements Engineering (GenAI for RE)

By examining these interconnected domains, our systematic literature review on GenAI for RE strategically positions itself at the convergence of AI, SE, and RE. This study builds upon the collective knowledge amassed in AI for SE, AI for RE, and GenAI for SE to offer a comprehensive analysis of the current landscape, challenges, and future trajectories of GenAI applications within the specialized domain of RE. This contextualization not only emphasizes the novelty and significance of our research but also illuminates its potential to advance the field of RE through the lens of cutting-edge GenAI technologies.

The application of GenAI in RE is a nascent field with immense potential to transform the way software requirements are elicited, analyzed, and validated. A recent study investigated the use of GenAI to automatically generate design practices

and evaluate their satisfaction of specific requirements [29]. This innovative approach showcases the potential of GenAI in assisting requirements engineers in exploring and evaluating alternative design solutions, thereby facilitating more creative and efficient requirements engineering processes. Moreover, ChatGPT, a state-of-the-art language model, has been used to automatically detect inconsistencies in natural language requirements, enhancing the requirements validation process [30]. This work highlights the potential of GenAI in improving the quality and consistency of software requirements, reducing the risk of errors and ambiguities.

Another important area of research explores the potential of ChatGPT to assist in requirements elicitation processes by evaluating the quality of requirements generated by the model and comparing them with those formulated by human RE experts [31]. This comparative analysis provides valuable insights into the strengths and limitations of GenAI in capturing and expressing user needs and expectations, paving the way for more effective human–AI collaboration in RE. Our study builds upon this growing body of research, aiming to systematically analyze the current state of GenAI applications in RE and provide a roadmap for future research and development in this transformative field.

Despite the promising applications of GenAI in RE, our comprehensive review has identified several critical challenges that need to be addressed for its widespread adoption and effective implementation. One significant challenge is ensuring the accuracy and reliability of GenAI-generated requirements, as highlighted by the inconsistency detection work of Fantechi et al. [30]. In addition, the comparative analysis of Ronanki et al. [31] reveals limitations in GenAI’s ability to fully capture the nuances of user needs and expectations, emphasizing the ongoing need for human expertise in the RE process.

Furthermore, our analysis has uncovered a range of challenges spanning technical, ethical, and practical dimensions:

- **Bias and Fairness [32]:** Ensuring that GenAI models do not perpetuate or amplify biases, particularly during requirements elicitation and analysis phases.
- **Ethical and Regulatory Concerns [33]:** Establishing ethical guidelines and regulatory frameworks specific to GenAI use in RE.
- **Security and Privacy [34, 35]:** Developing robust security protocols and privacy-preserving techniques for handling sensitive requirements data.
- **Interpretability and Explainability [36, 37]:** Enhancing the transparency and interpretability of GenAI models in RE to ensure stakeholder trust and accountability.

- **Computational and Economic Cost [38, 39]:** Addressing the significant computational resources and economic implications of deploying GenAI in RE.
- **Real-Time Processing [40]:** Developing GenAI systems capable of adapting to dynamically changing requirements and processing new information in real-time.
- **Hallucinations [41, 42]:** Mitigating the risk of hallucinations in GenAI outputs, which is crucial for maintaining the accuracy and reliability of generated requirements.
- **Reproducibility [43]:** Ensuring consistent and reproducible results in requirements generation and analysis to build stakeholder trust.
- **Controllability [44]:** Improving the precise control of GenAI models to align outputs with specific project or organizational needs.
- **Authorship and Copyright [45]:** Addressing the legal and ethical implications of AI-generated content in RE, including ownership and intellectual property rights.

Future research in GenAI for RE should focus on addressing these challenges and exploring several key areas. First, there is a need to develop more sophisticated GenAI models that can better understand and incorporate domain-specific knowledge in requirements generation and analysis. Second, research should explore ways to enhance the interpretability and traceability of GenAI-generated requirements, ensuring that stakeholders can understand and trust the AI’s outputs. Third, investigating optimal human–AI collaboration models in RE processes is crucial for leveraging the strengths of both human expertise and AI capabilities. In addition, developing techniques for bias mitigation, improving security and privacy measures, and establishing ethical guidelines will be essential for the responsible and widespread adoption of these technologies.

The studies mentioned above demonstrate the initial potential of Generative AI, particularly LLMs, such as the GPT series in RE. These applications span multiple stages of RE, including requirements elicitation, analysis, and validation. However, these studies represent only the beginning of this rapidly evolving field. Our study aims to fill the gap in comprehensive systematic literature reviews covering this specific area by providing a thorough analysis of the current state of research, identifying key challenges and opportunities, and outlining potential directions for future research in GenAI for RE.

In Section 5, based on the results of our systematic literature review, we will delve

deeper into the specific research trends, main technological methods, quality assessments, and future directions in this domain. We will explore how specific models like GPT-3 and GPT-4 are applied across various stages of RE, and how techniques such as prompt engineering and few-shot learning are being used to enhance the performance of these models in RE tasks. By building upon the growing body of research and addressing the identified challenges, we aim to systematically analyze the current state of GenAI applications in RE and provide a roadmap for future research and development in this transformative field. This interdisciplinary approach will be crucial in unlocking the full potential of GenAI in revolutionizing RE processes, leading to more efficient, accurate, and ethically sound software development practices.

3. Research Questions

RE is a critical phase in the software development lifecycle, focusing on eliciting, analyzing, specifying, and validating the requirements of a software system. The main activities in RE include requirements elicitation, analysis, specification, validation, and management. The increasing complexity of software systems and the growing demand for efficient and effective RE processes have led to explorations of advanced AI techniques, particularly GenAI, to support and enhance various RE activities.

The objective of this systematic literature review is to provide a comprehensive overview of the current state of research on the application of GenAI techniques in RE. We aim to identify the key trends, methodologies, and challenges in this emerging field by analyzing the existing body of knowledge. We further seek to provide insights into the potential of GenAI in addressing the limitations of traditional RE approaches and to propose future research directions to advance the field. To systematically investigate the application of GenAI in RE, we have formulated the following research questions (RQs):

3.1. *RQ1: What are the current research trends in applying GenAI to RE?*

This question aims to analyze the distribution and characteristics of published studies, including the venues of publication, temporal trends, and geographical distribution of research efforts. By examining these aspects, we seek to understand the evolving landscape of GenAI applications in RE research.

3.2. RQ2: What are the predominant approaches and techniques employed in current GenAI for RE research?

This question focuses on identifying and categorizing the specific methodologies, technologies, and strategies used in the selected studies. We aim to provide a comprehensive overview of the technical landscape, including prompt engineering techniques, model architectures, fine-tuning strategies, and other relevant approaches.

3.3. RQ3: How is the quality of current research in GenAI for RE evaluated?

This question addresses the critical aspect of research quality assessment. We will evaluate the quality of the reviewed papers by examining the effectiveness of their methodologies, the clarity of their research goals, and other relevant factors. This analysis will help identify best practices and potential areas for improvement in research quality.

3.4. RQ4: What are the main challenges in applying GenAI to RE, and what are the future research directions? How do these challenges and directions relate to the limitations of current research?

This research question addresses three crucial aspects of GenAI in RE: it identifies the primary challenges in implementing GenAI within RE practices, explores potential future research trajectories in this rapidly evolving field, and examines how these challenges and future directions are interconnected with the current limitations in research.

4. Research Methodology

To ensure a comprehensive and representative literature review, a systematic search strategy was implemented for retrieving relevant publications. Our research methodology aligns with the best practices in RE, as outlined in ISO/IEC/IEEE 29148:2018, particularly in the aspects of requirements classification and validation [8]. This alignment ensures that our approach is grounded in internationally recognized standards for RE processes.

Scopus was used as the main search engine because of its effectiveness in SE systematic literature reviews (SLRs) and its capability to export search results. Scopus encompasses numerous major publishers, including IEEE, ACM, Springer Nature, Wiley Blackwell, Taylor & Francis, and Elsevier. To maximize the retrieval of pertinent literature, the search was extended to ArXiv and Google Scholar using identical query parameters. Although ArXiv is not typically used in literature reviews because of its non-peer-reviewed content, it was included in the present search because of the

limited number of papers on GenAI for RE available in Scopus. Despite the lack of peer review, ArXiv papers often present innovative ideas and insights valuable for the review. Google Scholar was incorporated to capture a broader spectrum of relevant literature across various publishers. The search strategy employed diverse combinations of keywords and phrases related to GenAI and RE.

The search was confined to the past six years (2019–2024) to capture the most recent research advances in this domain. The year 2019 was selected as the starting point because of the significant emergence and development of GenAI technologies beginning that year, notably marked by the release of GPT-2, which represented a milestone in the advancement of LLMs.

Rigorous inclusion and exclusion criteria were established for the search process. Inclusion criteria encompassed (1) peer-reviewed journal articles and conference proceedings, (2) English-language publications, and (3) literature highly relevant to the application of GenAI in RE. Exclusion criteria comprised (1) non-peer-reviewed publications (with the exception of ArXiv, as previously noted), (2) literature irrelevant to the research topic, (3) duplicate research results, and (4) gray literature such as editorials, prefaces, and book reviews.

The initial search yielded 42 papers. Following a meticulous screening of titles and abstracts, 27 papers were ultimately included in this review. The screening process was conducted independently by three researchers, and any disagreements were resolved through discussion and consensus. All search results and screening processes were systematically managed and documented using Zotero, a reference management software. The selection and processing of the study as shown in Figure 4.

4.1. Search and Selection Process

4.1.1. Initial Search

The following query was executed on titles, abstracts, and keywords of papers, with the publication period limited to 2019–2024:

("Generative AI" OR "Generative Artificial Intelligence" OR "Large language model" OR "GPT") AND ("Requirement* Engineering")

This step resulted in the identification of 51 papers. The selected conferences and workshops from which the papers were retrieved are listed in Table 1.

4.1.2. Impurity Removal

Because of the nature of the involved data sources, the initial search results included elements that were not research papers, such as abstracts and international

Table 1: Major Conferences and Workshops Cited in Selected Papers and Their Abbreviations

Source	Acronym
<i>Artificial Intelligence and Engineering</i>	
Workshop on Artificial Intelligence and Model-driven Engineering	MAI
IEEE International Conference on Evaluation of Novel Approaches to Software Engineering	ENASE
<i>Requirements Engineering</i>	
IEEE International Requirements Engineering Conference	RE
IEEE International Requirements Engineering Conference Workshops	REW
<i>Business and Process Management</i>	
IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling	PoEM
International Conference on Business Process Management	BPM
<i>Systems and Data Engineering</i>	
IEEE International Symposium on Systems Engineering	ISSE
IEEE SoutheastCon	SECon
Proceeding of the International Conference on Intelligent Data Communication Technologies and Internet of Things	IDCIoT
<i>Other</i>	
IEEE Aerospace Conference	AeroConf

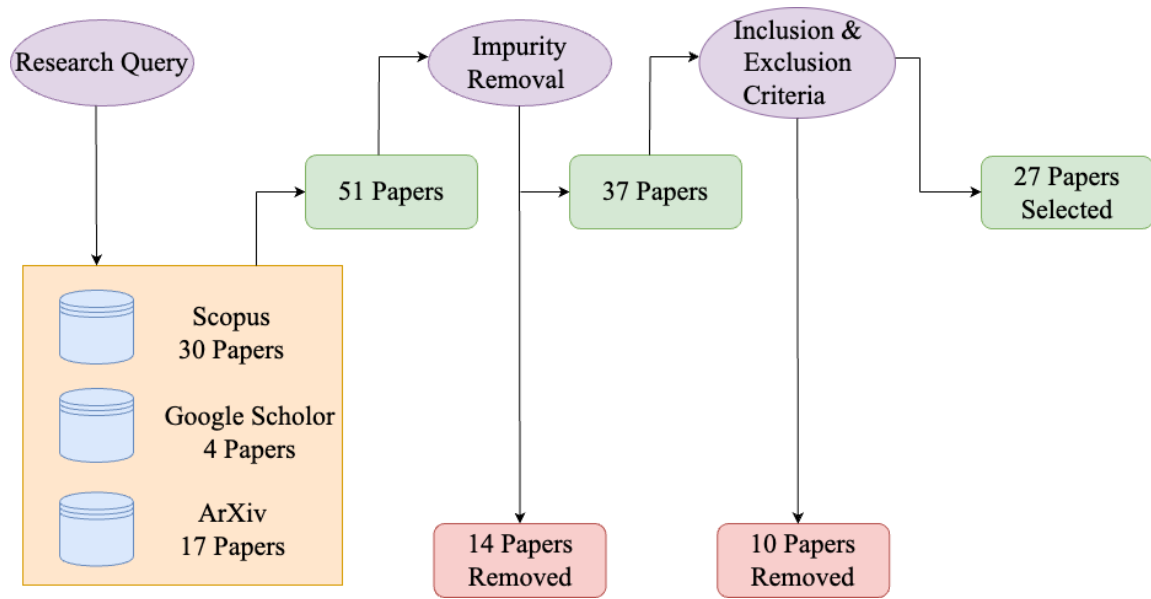


Figure 4: Paper selection and processing

standards. To ensure the relevance and quality of the literature, the selection criteria were refined to include only journal papers and conference papers. After this refinement, 37 papers remained in the dataset.

4.1.3. Inclusion and Exclusion Criteria

For each paper, three researchers independently vetted its inclusion in the SLR by applying the following criteria. Initially, the titles and abstracts were reviewed, followed by a full-text reading to determine the paper’s relevance to GenAI and RE. On the basis of the query definition, 27 papers were identified for inclusion.

- Inclusion Criteria:
 1. Papers addressing GenAI for RE
 2. Papers that are fully written in English
- Exclusion Criteria:
 1. Papers focusing on GenAI for SE but not specifically on RE
 2. Papers considering BERT as a GenAI or LLM
 3. Gray literature such as book chapters, Ph.D. theses, white papers, survey papers, and blogs

4.1.4. Data Extraction

To systematically address the research questions, a comprehensive data extraction process was conducted for each selected paper. Standardized data extraction forms were employed to ensure accuracy and consistency throughout this task. The extracted data, which encompassed essential details such as bibliographic information, research objectives, methodologies, and key findings, was meticulously organized and systematically stored in a comprehensively structured Microsoft Excel spreadsheet to facilitate efficient analysis and seamless access to the collected information. To further enhance the precision of the extracted data and facilitate seamless collaboration among the research team, the spreadsheet was subsequently transferred to Google Sheets. This cloud-based platform enabled the co-authors to efficiently review and validate the extracted information, minimizing the risk of errors or inconsistencies. The use of Google Sheets not only streamlined the data extraction process but also provided a user-friendly interface for displaying the collected data in a clear and concise tabular format. This approach substantially improved the overall efficiency and reliability of the data extraction phase because it enabled real-time collaboration and continuous quality control measures to be implemented throughout the review process. The extracted data is available online in our supplementary materials.¹

- Publication title
- Publication year
- Publication venue
- Requirements phases
- Quality characteristics of the system
- Model type
- Model parameters
- Prompt engineering techniques
- Whether fine-tuning was performed
- Metrics for evaluating the model
- Rating of quality issues
- Gaps and future work mentioned by authors

5. Analysis and Discussion

Building on the overview of GenAI applications in RE presented in Section 2.3, this section offers a detailed analysis of research trends, prevailing methodologies,

¹https://github.com/haowei614/GenAI4RE_SLR_Data

quality assessments, and potential future directions as identified through our systematic literature review. We conduct a comprehensive examination of the application of various GenAI technologies, including GPT series models (e.g., GPT-3, GPT-4), across the different stages of RE, such as elicitation, analysis, specification, validation, and management. Additionally, we assess the effectiveness of techniques like prompt engineering, few-shot learning, and chain-of-thought prompting in enhancing model performance. This in-depth analysis aims to provide a holistic perspective on how Gen AI is driving innovation and advancing RE practices.

5.1. RQ1: Publication Trends in GenAI for RE

Understanding the publication trends in GenAI for RE provides crucial insights into the evolution, focus areas, and emerging topics within this interdisciplinary field. By meticulously analyzing these trends, we can identify key contributors, shifts in research priorities, and the overall impact of published work. Our SLR initially encompassed publications from 2019 to 2024. However, an important observation emerged: the majority of relevant publications are concentrated in the period from 2023 to 2024. This timeframe represents a phase of rapid development and application of GenAI technologies, particularly in the domain of RE. To better illustrate this trend, we present the distribution of papers by publication type and year, focusing on the concentration of publications in 2023 and 2024.

Figure 5 illustrates the distribution of the 27 reviewed papers across three publication types—conference, workshop, and ArXiv—and three years—2022, 2023, and 2024. The data are summarized as follows:

- **Conference Papers (37.0%):** Of 10 conference papers, 1 was published in 2022, 5 in 2023, and 4 in 2024.
- **Workshop Papers (11.1%):** All 3 workshop papers were published in 2023.
- **ArXiv Papers (51.9%):** Of the 14 ArXiv papers, 4 were published in 2023 and 10 in 2024.

The substantial number of publications on ArXiv observed in 2024 in particular reflects a growing trend toward the rapid dissemination of research findings through preprint repositories. This practice facilitates quicker sharing and feedback within the research community, which is particularly important in the dynamic and fast-evolving field of GenAI for RE. The utilization of ArXiv by researchers underscores its role as a platform for the immediate dissemination of preliminary research outcomes. It enables scholars to gather early feedback and make necessary revisions

before submitting their work to peer-reviewed journals, thereby accelerating the research cycle and fostering collaborative advancements. These trends illustrate the dynamic nature of the GenAI for RE landscape and emphasize the increasing focus on leveraging GenAI technologies to address challenges in RE. They also highlight the need for continued collaboration between academia and industry to ensure that research findings are effectively translated into practical applications.

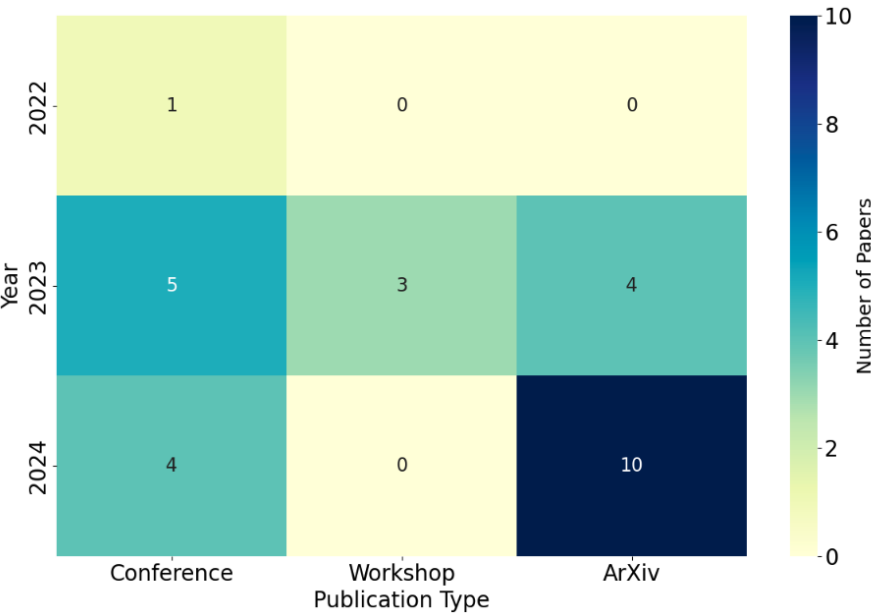


Figure 5: Distribution of papers by publication type and year

Main Findings for RQ1

Our findings reveal a marked increase in research on the application of GenAI in RE between 2023 and 2024, reflecting the rapid development of this emerging field. More than half of the reviewed papers are sourced from preprint platforms such as ArXiv, highlighting the importance of swift dissemination and early feedback. In addition, 37% of the studies are conference papers, demonstrating growing academic interest in this area. Together, these trends reflect the dynamic nature and increasing significance of GenAI applications in RE.

Takeaway Message 1

For researchers, it is recommended to strike a balance between publishing in peer-reviewed journals and leveraging preprint platforms to ensure both credibility and the rapid dissemination of new ideas. Researchers should also focus on exploring innovative GenAI applications in under-researched stages of RE. For practitioners, regularly following developments on platforms such as ArXiv is advisable, albeit with caution regarding non-peer-reviewed content. Given the fast-paced growth of this field, both communities should prepare for the accelerated integration of GenAI into RE and for fostering collaborations between academia and industry to bridge the gap between research and practical applications.

5.2. RQ2: Methodology Trends

The data extracted from each study are presented in Tables 2, 3, and 4. Table 2 provides a comprehensive overview of the 27 articles reviewed, published between 2022 and 2024, revealing a burgeoning body of work in the field of GenAI for RE. The temporal distribution of publications shows an exponential growth trend. In 2022, only 1 paper was published in this field. The number of published papers increased substantially to 12 in 2023 and then to 14 papers in 2024 (up to the time of this study’s completion). This rapid acceleration in publication frequency is indicative of several key factors. Primarily, it reflects an escalating academic interest in the application of GenAI to RE. It also suggests potential significant advancements in underlying AI technologies, particularly in the domain of LLMs. The trend also points to a growing recognition of the transformative potential of GenAI in addressing complex challenges within RE.

5.2.1. Overview of Reviewed Studies

To facilitate a more nuanced analysis and comprehension of the reviewed literature, we developed a systematic categorization framework. This framework encompasses several critical dimensions that enable a granular examination of each study’s focus and contributions. The key factors we identified and scrutinized are as follows:

- **Paper ID:** Each study was assigned a distinctive alphanumeric code, serving as a concise reference point throughout our analysis and discussion. Studies sourced from ArXiv are denoted by “A” followed by a sequential number (e.g., A1, A2, A3). Similarly, studies from conferences and workshops are denoted by “C” and “W,” respectively, followed by a number.

- **Author:** We documented the primary contributors to each piece of research, enabling the identification of influential researchers and research groups in the field.
- **Year:** The publication year was recorded to map the temporal evolution of AI applications in RE and to identify emerging trends.
- **Paper Type:** Studies were classified as either full papers or short papers on the basis of the submission guidelines for each conference.
- **RE Phase:** Examining the RE phase is crucial because the RE process lays the foundation for the entire software development lifecycle. The quality and effectiveness of RE activities directly affect the success of the project [46]. By analyzing how GenAI techniques are being applied across different stages of RE, we can gain valuable insights into the current state of research and identify areas where GenAI has the potential to make the most significant contributions. We categorized each study according to its primary focus within the RE lifecycle, including:
 - Requirements elicitation (Eli.): This phase involves gathering and discovering requirements from various stakeholders. It is a critical step because the quality of elicited requirements directly affects the overall success of the project [47].
 - Requirements analysis (Ana.): During this phase, the elicited requirements are analyzed to identify conflicts, inconsistencies, and dependencies. The goal is to achieve a coherent and feasible set of requirements [46].
 - Requirements specification (Spec.): This phase focuses on documenting the agreed-upon requirements in a clear, concise, and unambiguous manner. The specification serves as a contract between stakeholders and developers [48].
 - Requirements validation (Val.): The purpose of this phase is to ensure that the specified requirements meet the stakeholders' needs and expectations. Validation techniques, such as reviews and prototyping, are used to detect and correct errors early in the development process [48].
 - Requirements management (Man.): This phase spans the entire RE process and deals with managing changes to requirements, maintaining traceability, and ensuring consistency among related artifacts [49].

The five phases we have chosen to focus on are based on well-established RE process models in the literature. One of the most widely recognized frameworks is the RE process model proposed by Pohl [48], which divides RE into four core activities: elicitation, negotiation, specification, and validation. Our categorization closely aligns with Pohl’s model, with the addition of the management phase to encompass activities related to requirements change management and traceability.

- **GenAI Alignment with RE:** Investigating the alignment between GenAI techniques and specific RE tasks is essential for understanding how these advanced AI models can be effectively leveraged to support and enhance various aspects of the RE process. By examining the specific RE activities addressed by GenAI applications, we can gain insights into the current state of research and identify areas where GenAI has the greatest potential to make a positive impact.

The motivation behind this investigation is multifaceted. First, with the development of powerful language models such as GPT-3, the field of GenAI has seen rapid advancements in recent years [50]. The model has demonstrated remarkable capabilities in tasks such as natural language understanding, generation, and completion. Given the natural language-intensive nature of many RE activities, exploring how these GenAI techniques can be aligned with specific RE tasks to improve efficiency, quality, and consistency is important. In addition, the RE process is known to be complex, time-consuming, and prone to errors [46]. By identifying the specific RE tasks that can benefit from GenAI support, we can pave the way for the development of intelligent tools and methodologies that address the pain points in the RE process. For example, GenAI techniques could potentially assist in automating requirements elicitation, identifying inconsistencies during requirements analysis, or generating natural language requirements specifications. Finally, the alignment between GenAI and RE tasks can help bridge the gap between the AI and SE communities. By showcasing the potential applications of GenAI in the context of RE, we can foster cross-disciplinary collaboration and knowledge exchange, which, in turn, can lead to the development of more sophisticated and tailored GenAI solutions that cater to the specific needs of the RE process.

Upon further analysis, we classified the papers on the basis of their respective conferences. Among them, 25 are full papers and 2 are short papers, adhering to the specific guidelines of each conference. This predominance of full papers indicates a trend toward more comprehensive and in-depth research in this emerging field,

providing substantial evidence to review and understand the current trends and advancements in GenAI applications within RE. We subsequently categorized the papers according to their focus within the RE lifecycle. As illustrated in Figure 6, the distribution of research emphasis across RE phases is notably uneven:

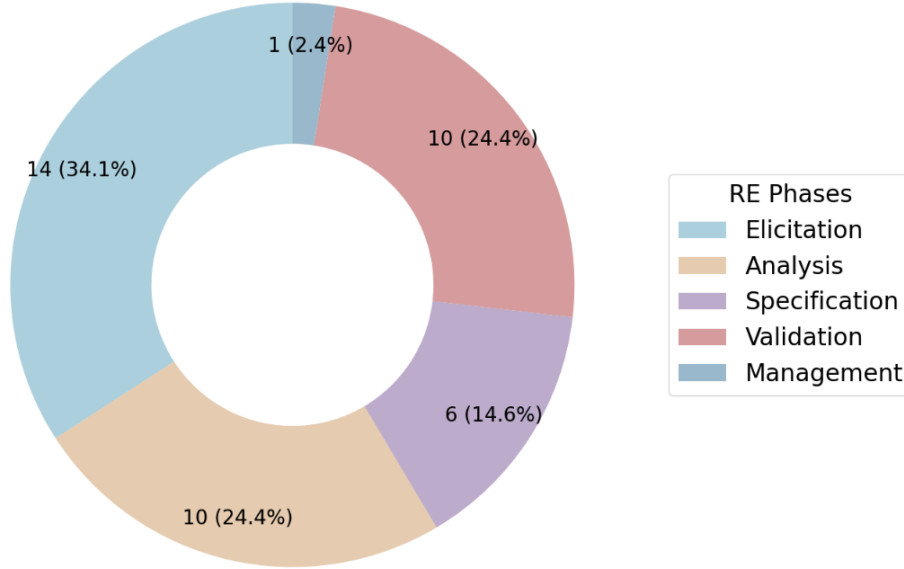


Figure 6: Distribution of requirements engineering phases

1. Requirements Elicitation: The majority of studies (51.9%, 14 out of 27 papers) focus on enhancing this phase. Examples include improving information retrieval (A1), domain model extraction (W1), and supporting elicitation quality assessment (A2).
2. Requirements Analysis: Studies such as A3 and C5 highlight the importance of analysis in transforming requirements into formal representations and generating goal models.
3. Requirements Specification: A substantial number of studies (22.2%, 6 papers) address specification, focusing on tasks such as summarizing contractual obligations (C1), automating inconsistency detection (C2), and generating interview scripts (W2).
4. Requirements Validation: This phase is explored in 37.0% (10 papers) of the studies, with notable examples like C7, which verify and identify code requirements.

5. Requirements Management: Notably, only one paper (3.7%) addressed this phase.

5.2.2. Comprehensive Analysis of GenAI Models in RE Applications

To provide a thorough understanding of the GenAI models used in RE, we have summarized our findings in Table 3 and 4. Table 3 offers a detailed exposition of the GenAI models used across various RE applications, elucidating the specific techniques and approaches adopted in each study. In addition, we have developed visual representations to enhance comprehension of our analysis. Figure 7 presents an overview of the preferred GenAI models in RE research, highlighting the prevalence and distribution of different AI architectures and frameworks. Figure 8 illustrates the focus areas within software quality characteristics that researchers have prioritized when applying GenAI to RE tasks. These visual aids complement our tabular data, offering a holistic view of the current landscape of GenAI applications in RE. They not only showcase the dominant trends in model selection but also reveal the quality attributes that researchers are most keen to address through AI-driven RE approaches.

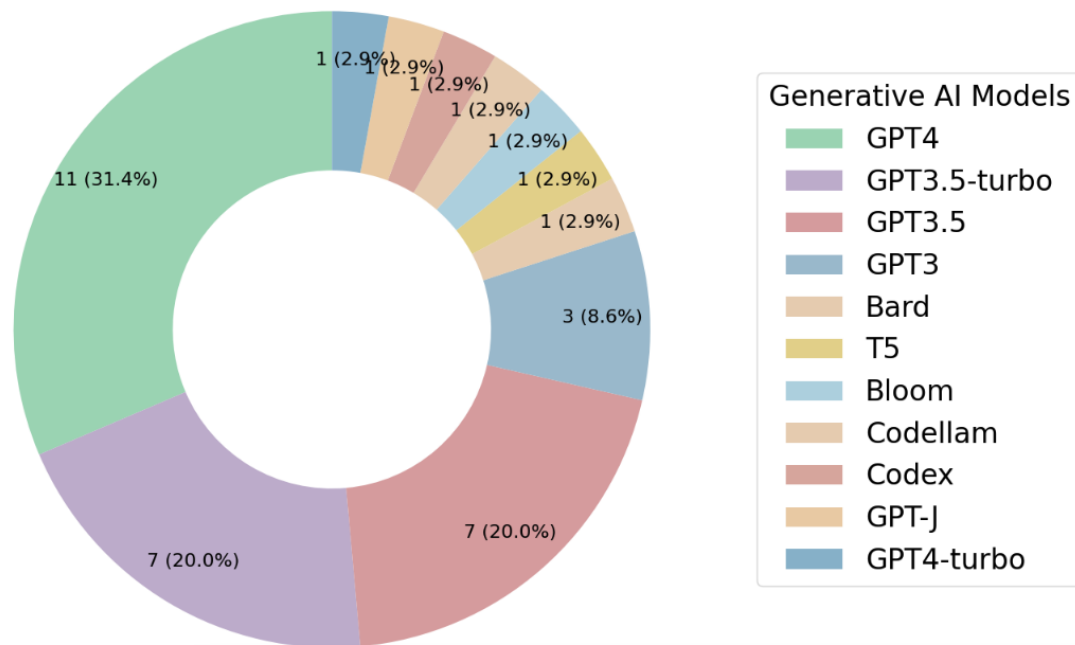


Figure 7: Distribution of generative AI models

Table 2: Overview of reviewed studies (F: Full papers; S: Short papers)

ID	Authors	Year	Type	RE Phase			GenAI Align with RE
				Eli.	Ana.	Spec. Val. Man.	
A1	J. Zhang et al. [51]	2023	F	✓	✓		Improve RE information retrieval
W1	S. Arulmohan et al. [52]	2023	F		✓		Enhance domain model extraction
A2	K. Ronanki et al. [53]	2023	F	✓			Supports elicitation quality assessment
C1	C. Jain et al. [54]	2023	F	✓	✓		Summarize contractual obligations
C2	A. Fantechi et al. [55]	2023	S			✓	Automatically detect inconsistencies
W2	B. Gorer et al. [56]	2023	F	✓			Generate interview scripts
C3	D. Clements et al. [57]	2023	F	✓			Automatically extract user characteristics
C4	N. Blasek et al. [58]	2023	F	✓			Simulate interviews, generate requirements
W3	B. Chen et al. [59]	2023	F	✓			Generate goal models and refine
A3	M. Cosler et al. [60]	2023	F		✓	✓	Translate requirements into temporal logic
C5	H. Mustroph et al. [61]	2023	F			✓	Process requirements into formal representations
C6	I. Grasler et al. [62]	2022	F	✓	✓		Generate requirements to train classifier
C7	J. O. Couder et al. [63]	2024	F			✓	Verify and identify code requirements
C8	J. U. Oswal et al. [64]	2024	F	✓	✓		Transform requirements into stories
C9	J. Peer et al. [65]	2024	F	✓	✓		Model-based design enhances RE practices
A4	M. Krishna et al. [66]	2024	F		✓	✓	Generate and validate SRS documents
A5	A. Nouri et al. [67]	2024	F			✓	Automate RE for safety requirements
A6	D. Jin et al. [68]	2024	F	✓	✓	✓	LLM agents automate and enhance RE tasks
A7	R. Feldt et al. [69]	2024	S	✓	✓	✓	Extract high-level user goals
A8	M. Ataei et al. [70]	2024	F	✓			Generate diverse user agents
C10	Z. Zhang et al. [71]	2024	F		✓		Improve user stories quality
A9	B. Wang et al. [72]	2024	F		✓	✓	Assist novice analysts in UML modeling
A10	A. El-Hajjani et al. [73]	2024	F	✓	✓		Improve requirements classification
A11	K. Ronanki et al. [74]	2023	F	✓	✓		Automate requirements classification and tracing
A12	M. Fazelnia et al. [75]	2024	F		✓		NLI for categorization and defect detection
A13	N. Feng et al. [76]	2024	F	✓		✓	Extract and suggest semantic relations
A14	S. Santos et al. [77]	2024	F			✓	Generate and evaluate design practices

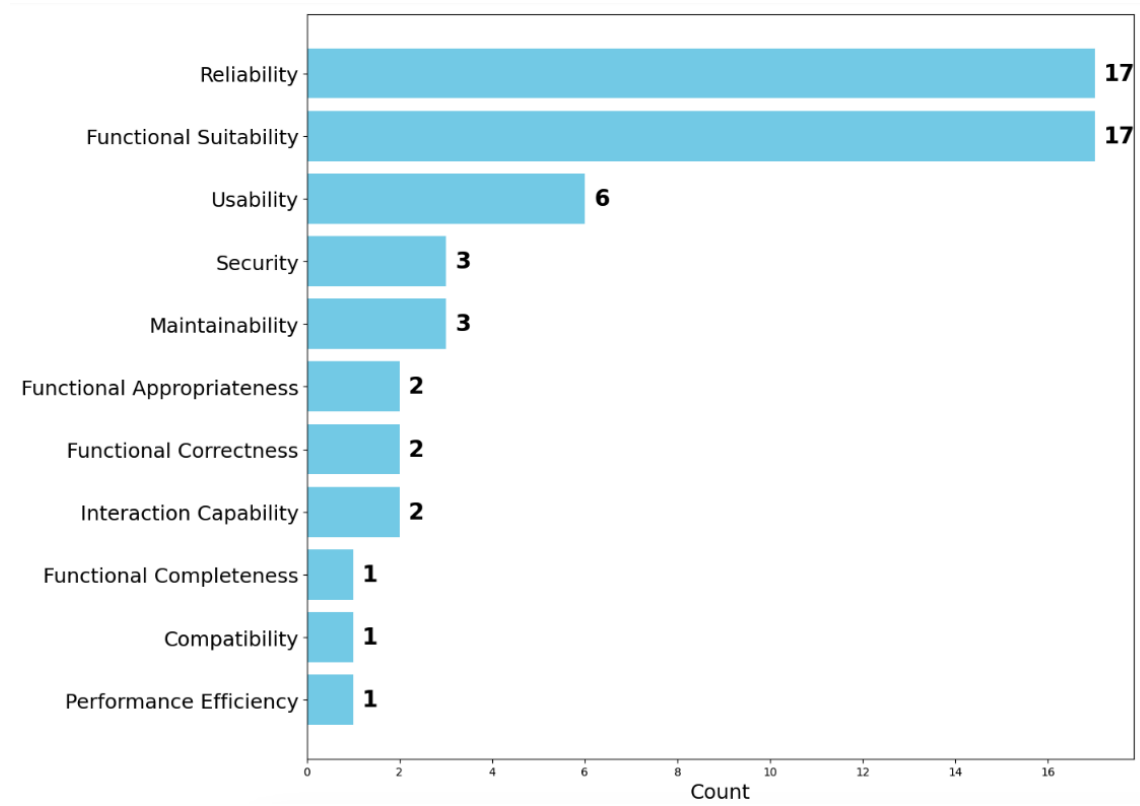


Figure 8: Distribution of software quality characteristics

Table 3 provides an in-depth overview of the GenAI models applied in RE across various studies. The table categorizes the studies by ID, characteristics addressed, models used, fine-tuning status, and specific parameters. This analysis aims to elucidate the contributions of each study, highlight the diversity of the AI models employed, and identify potential gaps and areas for future research. The analysis framework incorporates several key dimensions: study identification, characteristics addressed, GenAI models employed, fine-tuning status, and model parameters. Each study is assigned a unique identifier for reference throughout the analysis. The characteristics addressed outline the specific RE aspects or challenges focused on in each study. The GenAI models employed are documented, providing insight into the prevalent technologies in the field. The fine-tuning status indicates whether the AI models were used as-is or customized for specific RE tasks. Where available, details on model parameters are provided, offering insights into the scale and complexity of the AI systems used.

Our analysis framework incorporates the following key dimensions:

- **Paper ID**

- **Characteristics:** The ISO/IEC 25059 standard provides a well-defined quality model that outlines the key characteristics and sub-characteristics essential for evaluating software product quality. These characteristics include functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability, and portability [78]. By aligning our evaluation of GenAI models with this standard, we can ensure that we are considering a wide range of quality attributes that are relevant to the RE context. One of the primary reasons for adopting the ISO/IEC 25059 standard is to establish a common language and framework for assessing the quality of GenAI models across different studies and applications. Using a standardized approach, we can compare different GenAI techniques and identify best practices and areas for improvement. This standardization also enables better communication and collaboration among researchers and practitioners working on GenAI applications in RE.
- **Model(s):** This perspective is crucial for identifying the most prevalent and effective GenAI architectures used in RE contexts. By documenting the specific models, such as GPT-3, GPT-4, or alternative architectures, we can track the adoption and performance of different AI models across various RE tasks. This information can help researchers and practitioners make informed decisions about which models to leverage for their specific RE challenges and can guide future research efforts in developing and refining GenAI architectures for RE applications.
- **Fine-Tuning:** Examining the fine-tuning status of the models is important for understanding the level of customization and adaptation applied in each study. Fine-tuning techniques involve modifying a pre-trained model’s parameters to better suit a specific task or domain [79]. By distinguishing between studies that employ fine-tuning and those that use pre-trained models without modification, we can gain insights into the effectiveness of different adaptation strategies for RE tasks. This information can help identify best practices for tailoring GenAI models to the unique requirements and challenges of the RE process.
- **Parameter(s):** Documenting parameter(s) such as temperature settings and other relevant configuration details is vital for understanding the fine-tuning

approaches used to optimize model performance. Temperature is a key parameter in generative models such as GPT-3 because it controls the randomness and diversity of the generated outputs [80]. By recording these parameter settings, we can analyze how different configurations impact the effectiveness of GenAI models in various RE applications. This information can facilitate the identification of optimal parameter settings for specific RE tasks and contribute to the development of standardized fine-tuning strategies.

For model characteristics, “functional suitability” is identified as the most prevalent characteristic, appearing in 22 of 27 studies (e.g., A1, C3, and A14). This prevalence underscores the paramount importance placed on the practical applicability of AI models in RE tasks. “Reliability” is the second-most common characteristic, featured in 11 studies (e.g., W1, C7, and A9). This frequency reflects a strong emphasis on the consistency and stability of AI model outputs in RE contexts. Other notable characteristics include “usability,” “maintainability,” and “interaction capability.” These characteristics are critical for ensuring the effective application of AI models in real-world RE practices. The distribution of these characteristics across the studies provides insights into the multifaceted approach taken by researchers to address various aspects of AI model quality in RE applications.

Regarding model selection, GPT-series models (e.g., GPT-3, GPT-3.5, and GPT-4) are observed to dominate the field, indicating the preeminence of LLMs in RE tasks. Some studies have been found to use specialized variants such as GPT-3.5-turbo, possibly for enhanced performance in specific tasks. GPT-4 is identified as the most prevalent, followed by GPT-3 and GPT-3.5-turbo. A few studies have employed GPT-2 and other models (e.g., T5, Google Bard, Codex, Bloom, GPT-J, and Codellama). This distribution indicates that GPT-3.5 and GPT-4 series models currently dominate the RE field, potentially significantly enhancing the efficiency and quality of RE processes. Notably, several studies (e.g., A9, A10, and A14) used multiple models, suggesting comparative performance analyses or attempts to leverage complementary strengths of different models. This approach of using multiple models indicates efforts to comprehensively evaluate and optimize the application of GenAI in RE tasks.

With respect to fine-tuning, the majority of studies (20/27) used pre-trained models without fine-tuning. This tendency may indicate either robust capabilities of pre-trained models in RE tasks or a focus on out-of-the-box applicability. However, 7 studies are noted to have implemented fine-tuning, suggesting potential performance gains through model customization for specific RE tasks.

Regarding parameter settings, most of the studies do not specify detailed parameter settings, possibly because of the use of default configurations or the perception

that parameter details are not critical to the research outcomes. Among studies that provide parameter information, the temperature setting is the primary focus. The temperature parameter, typically ranging from 0 to 1, is a key factor controlling the randomness and creativity of generated text. Values closer to 0 are associated with more predictable and consistent text, whereas values closer to 1 are associated with more random and diverse text generation. Attention to temperature settings is helpful in understanding how researchers control model generation strategies and why specific values are chosen to achieve certain effects. This aspect is considered significant for reproducing research results and tuning model performance.

5.2.3. Prompt Engineering

Another critical aspect of the analysis is prompt engineering (a factor of paramount importance in leveraging GenAI models effectively), showcasing the field’s dynamic nature and potential for AI-driven innovation. The significance of prompt engineering is recognized in its capacity to enhance the quality of generated outputs through optimizing input prompts, controlling generated content, mitigating errors and biases, and adapting to diverse application scenarios and requirements.

To systematically evaluate the role and implementation of prompt engineering across the reviewed studies, the following key factors are analyzed; these factors are elaborated in Table 4:

- **Paper ID**
- **Learning Paradigm:** The “learning paradigm” category encompasses various approaches, including zero-shot, one-shot, few-shot learning, and chain of thought (COT). Analyzing the distribution of these paradigms in the reviewed papers is important for understanding the prevalent strategies used to enable GenAI models to learn and perform RE tasks with limited or no specific examples. This information can guide researchers and practitioners in selecting appropriate learning paradigms for their specific RE challenges and can inspire further research into novel approaches to enhance GenAI performance in low-data scenarios.

By articulating the thought process of the model, COT can provide valuable insights into the decision-making mechanisms of GenAI systems, enhancing their interpretability and trustworthiness [81]. Analyzing the prevalence and effectiveness of COT in the reviewed papers can inform future research efforts toward developing more transparent and explainable GenAI solutions for RE. The distribution of this factor in the reviewed papers is illustrated in Figure 9.

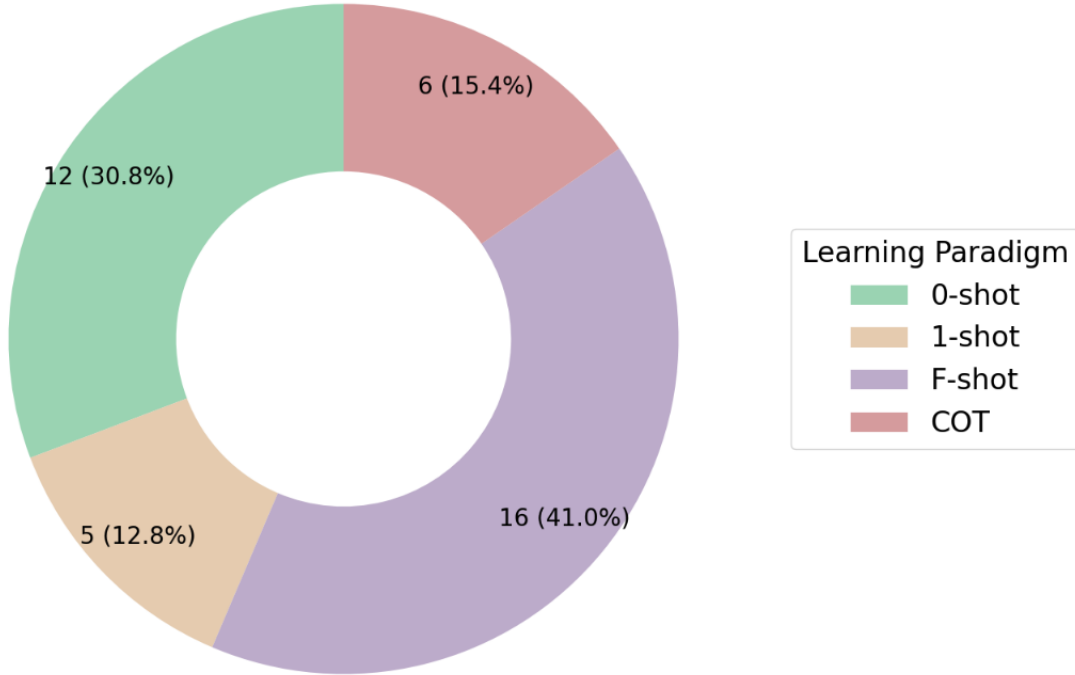


Figure 9: Distribution of learning paradigm

- **Prompt Type:** This category distinguishes between instruction-based prompts, question-based prompts, and other innovative formats used in the studies. Analyzing the distribution of these prompt types is important for understanding the most common and effective methods of presenting information to GenAI models in RE contexts. This information can guide researchers and practitioners in designing prompts that elicit the desired behavior from GenAI systems and can inspire further research into novel prompt engineering techniques tailored to specific RE tasks. The distribution of these prompt types is illustrated in Figure 10.
- **Task Specificity:** This category identifies the particular RE tasks or challenges addressed by prompt engineering in each study. By examining the focus areas of research efforts, we can gain insights into the RE activities most commonly targeted for improvement through GenAI applications. This information can help prioritize future research efforts and can guide practitioners in identifying GenAI solutions most relevant to their specific RE challenges.
- **Prompt Availability:** This category is essential for assessing the repro-

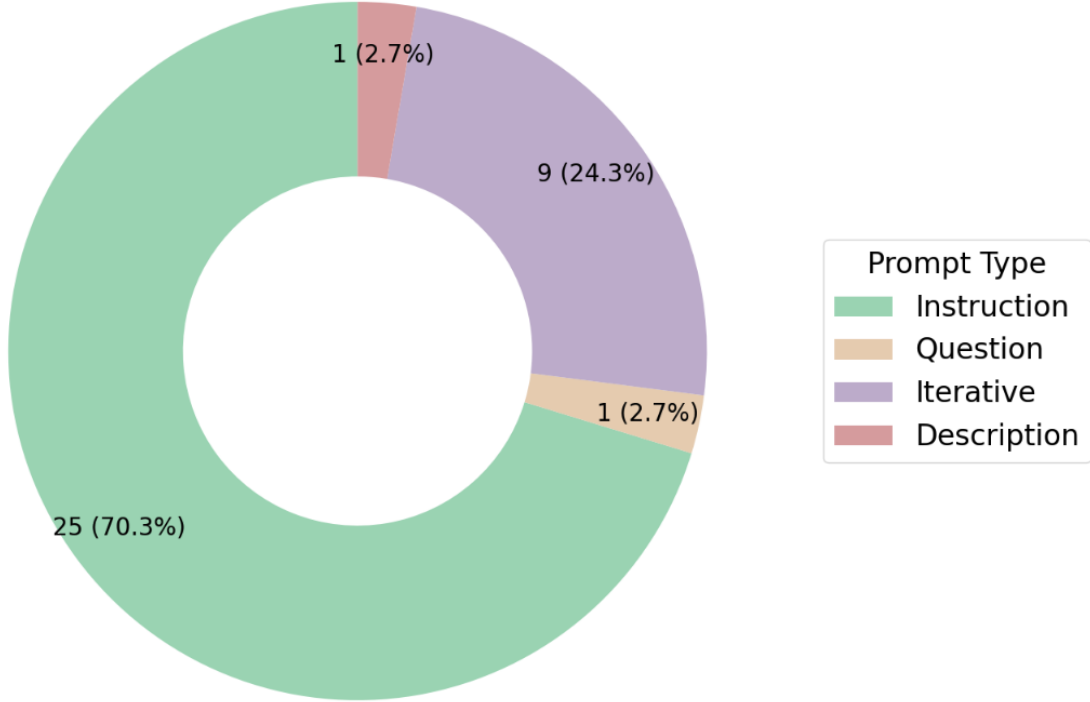


Figure 10: Distribution of prompt type

ducibility and transparency of the studies. The availability of the exact text of the prompts used in each study is crucial for enabling other researchers to replicate and build upon the findings [82]. By documenting the prevalence of prompt availability in the reviewed papers, we can identify potential gaps in research practices and can encourage authors to share their prompts to facilitate further analysis and advancement of prompt engineering techniques in RE contexts.

For learning paradigms, various approaches are used across the studies, reflecting ongoing efforts to optimize AI performance for specific RE tasks:

- **Zero-shot Learning:** This approach is used in studies such as A1 and C4, leveraging the model’s pre-existing knowledge without task-specific training examples.
- **One-shot and Few-shot Learning:** These methods are employed in studies like W1, A3, and C2, providing limited examples to guide the model’s output, balancing generalization with task-specific guidance.

- **Chain-of-Thought (COT):** Implemented in studies such as A2 and W2, COT prompts the model to generate intermediate reasoning steps, enhancing its capacity for complex task resolution.
- **Instruction (Inst.) and Iterative (Iter.):** These formats are widely adopted, as seen in A1 and W1, guiding the model through detailed instructions and iterative feedback loops, respectively.

The analysis reveals a clear preference for certain prompt formats: Instruction-based prompts are observed to dominate, featured in 22 of 27 studies. Iterative approaches are found to complement instructions in 7 studies, indicating a tendency toward dynamic prompt engineering. A minority of studies are noted to explore alternative formats, including questions and descriptive prompts, suggesting potential for further diversification in prompt design. The prevalence of instruction-based prompts underscores their efficacy in directing GenAI models for RE tasks, whereas the emergence of hybrid approaches is interpreted as signaling an evolution toward more sophisticated methodologies.

For task specificity, the studies are observed to address a wide spectrum of RE tasks, demonstrating the versatility of prompt engineering; requirements retrieval, classification, and verification are noted as key areas of focus. Domain model extraction is identified as another significant task. Elicitation response generation, inconsistency detection, user story quality enhancement, and safety RE are also recognized as important areas where prompt engineering is applied. This diversity is interpreted as highlighting the adaptability of prompt engineering techniques across various RE domains and challenges. The range of tasks addressed suggests the potential of prompt engineering to strongly impact multiple aspects of the RE process.

Regarding prompt availability, all examined studies are reported to have their prompts “Available,” indicating a significant tendency toward research transparency and reproducibility. This universal availability fosters collaborative innovation and enabling rigorous validation of results within the RE community. The consistent practice of making prompts available is interpreted as a positive development in the field, potentially facilitating more rapid advancements and improvements in prompt engineering techniques for RE applications.

The insights derived from this analysis underscore the profound influence of GenAI models on RE. The observed trends in model selection, application diversity, and parametric considerations are collectively interpreted as indicating a rapidly evolving domain. These advanced AI systems do not merely enhance functional suitability and operational efficiency but fundamentally reshape RE practices. The emphasis on parameter tuning and model adaptability across various RE tasks reveals

a nuanced approach to AI integration, balancing technological capability with task-specific optimization. Moreover, the application of these models in safety-critical domains demonstrates their potential to elevate the quality and reliability of software systems, addressing some of the most challenging aspects of modern software development. As the field progresses, the interplay between GenAI capabilities and the complex demands of RE processes is anticipated to drive innovation, potentially revolutionizing how requirements are elicited, analyzed, and managed across diverse industries. This evolving landscape not only presents opportunities for enhanced productivity and accuracy but also calls for continued research into the ethical and practical implications of GenAI-driven RE methodologies.

Main Findings for RQ2

On the basis of the analysis, current GenAI applications in RE primarily use the GPT-series models, especially GPT-4, with most research using pre-trained models rather than fine-tuned ones. Research predominantly focuses on requirements elicitation and validation phases, emphasizing functional suitability and reliability. In prompt engineering, instruction-based prompts are most prevalent, with researchers widely adopting zero-shot, few-shot learning, and COT paradigms. These techniques are applied to diverse RE tasks such as requirements retrieval, classification, and verification. Notably, all of the studies made their prompts available, reflecting a commitment to research transparency.

Table 3: Overview of Generative AI Models Applied in Requirements Engineering

ID	Characteristics	Model(s)	Fine-tuning	Parameter(s)
A1	Functional suitability, Reliability	GPT-3.5-turbo	✗	Temperature: 0
W1	Functional suitability, Maintainability	GPT-3.5-turbo	✗	Temperature: 0
A2	Functional suitability, Security	GPT-3.5	✗	-
C1	Functional suitability, Reliability	GPT-3,GPT-2,T5	✓	-
C2	Reliability, Maintainability	GPT-3.5	✗	-
W2	Functional Suitability, Interaction Capability	GPT-3,Bard	✗	-
C3	Functional Suitability, Interaction Capability	GPT-3	✗	Temperature: 0
C4	Functional suitability, Reliability	GPT-4	✗	-
W3	Performance Efficiency, Usability	GPT-4	✗	Temperature:0.5
A3	Functional Suitability, Usability	Codex, Bloom	✗	Temperature:0.2
C5	Security, Reliability	GPT-4	✗	-
C6	Functional Suitability, Maintainability	GPT-J	✗	-
C7	Functional Correctness, Reliability	GPT-3.5	✗	-
C8	Functional Suitability, Usability	GPT-3.5	✗	-
C9	Functional Completeness, Reliability	GPT-3.5	✗	-
A4	Functional Suitability, Reliability	GPT-4, CodeLlam	✗	-
A5	Reliability, Security	GPT-4	✗	-
A6	Functional Suitability, Reliability	GPT-3.5-turbo	✗	-
A7	Compatibility, Reliability	GPT-4	✗	-
A8	Functional Suitability, Usability	GPT-4-turbo	✗	-
C10	Usability, Reliability	GPT-3.5-turbo,GPT-4	✗	Temperature:1
A9	Functional Suitability, Usability	GPT-3.5, GPT-4	✗	-
A10	Functional Appropriateness, Reliability	GPT-3.5-turbo,GPT-4	✗	-
A11	Functional Suitability, Reliability	GPT-3.5-turbo	✗	-
A12	Functional Appropriateness, Reliability	GPT-3.5	✓	-
A13	Functional Suitability, Reliability	GPT-4	✗	-
A14	Functional Suitability, Reliability	GPT-3.5-turbo, GPT-4	✗	Temperature:0.7

Table 4: Overview of Prompt Engineering Techniques in Generative AI Applications

ID	Learning Paradigm	Prompt Format	Task Specificity	Prompt Availability
A1	0-shot	Inst.	Zero-shot requirements retrieval	Available
W1	1-shot	Inst.	Domain model extraction	Available
A2	COT	Quest.	Elicitation response generation	Available
C1	0-shot	Inst.	Contractual obligation summarization	Available
C2	1-shot	Inst.	Inconsistency detection	Available
W2	F-shot, COT	Inst.	Interview script generation	Available
C3	0-shot	Inst.	User characteristics extraction	Available
C4	0-shot	Inst.	Simulated expert interviews	Available
W3	F-shot	Inst.	Interactive goal model generation	Available
A3	F-shot	Inst.	Temporal logic translation	Available
C5	F-shot	Inst., Iter.	Automated compliance verification	Available
C6	F-shot	Inst., Iter., Desc.	Classifier training automation	Available
C7	0-shot, 1-shot	Inst., Iter.	Automated requirements verification	Available
C8	0-shot, 1-shot, F-shot	Inst.	Automated user stories transformation	Available
C9	0-shot, F-shot	Inst.	Automated requirements classification	Available
A4	F-shot	Inst.	Automated generation and validation	Available
A5	F-shot	Inst., Iter.	Engineering safety requirements	Available
A6	-	Inst., Iter.	Collaborative RE automation	Available
A7	-	Inst., Iter.	Goal-to-API mapping	Available
A8	0-shot, COT	Inst., Iter.	Diverse user agent generation	Available
C10	F-shot, COT	Inst., Iter.	User story quality enhancement	Available
A9	-	Inst., Iter.	UML modeling assistance	Available
A10	0-shot, F-shot	Inst.	Enhanced requirements classification	Available
A11	0-shot, F-shot, COT	Inst.	Requirements classification and tracing	Available
A12	0-shot, F-shot	Inst.	Entailment-based requirements analysis	Available
A13	F-shot	Inst.	Semantic relations extraction	Available
A14	0-shot, F-shot, COT	Inst.	Design practices generation and evaluation	Available

Inst.: Instruction; Iter.: Iterative; Quest.: Question; Desc.: Description

Takeaway Message 2

Researchers should prioritize exploring GenAI applications in understudied RE phases such as requirements management and investigate fine-tuning techniques for GPT-series models to enhance performance in domain-specific tasks. Developing more sophisticated prompt engineering techniques, designing comprehensive studies evaluating GenAI’s impact across the entire RE lifecycle, and maintaining prompt availability to promote research transparency are key to advancing the field.

Practitioners should consider adopting advanced models such as GPT-4, especially in early stages such as requirements elicitation and analysis. It is recommended to start with instruction-based prompts when implementing GenAI for RE tasks and to focus on enhancing functional suitability and reliability. Practitioners should also be prepared to use GenAI models without extensive fine-tuning and to stay informed about developments in COT and iterative prompt engineering techniques for more complex RE tasks.

5.3. RQ3 Quality Assessment

Once the paper list was constructed using the inclusion and exclusion criteria, a final corpus of 27 papers was obtained. As highlighted in Ref. [83], evaluating the quality of the studies included in a systematic literature review is crucial to ensuring the reliability of the evidence and the validity of the conclusions [84]. The following quality assessment (QA) criteria are established:

1. Clarity of the paper’s goal;
2. Effectiveness of the evaluation in meeting stated goals and objectives;
3. Detailed presentation of results;
4. Acknowledgment and discussion of limitations, findings, and suggestions for future study.

In the evaluation of each study, points were assigned to three assessment criteria on a scale from 0 to 1. A rating of Yes (Y) was given if the information was found to be precise and reliable, Partial (P) if the information was partially available, and No (N) if the information was absent. These ratings were scored as Yes = 1 (full point), Partial = 0.5 (half-point), and No = 0 (zero points). This method is a well-established approach for assessing the quality of studies and has been widely used in various systematic literature reviews [85, 86]. The application of this scoring system

is intended to provide a quantitative measure of each study’s quality across the established criteria. This approach enables a systematic and comparable assessment of the reviewed papers, facilitating the identification of high-quality research and potential areas of weakness in the literature. The use of a standardized scoring method is aimed at enhancing the objectivity and reproducibility of the QA process within the context of this systematic literature review.

Table 5: Quality Assessment of Reviewed Papers

ID	QI1	QI2	QI3	QI4	Total
A1, W1, A2, C1, W3, A3, C6, C7	1	1.0	1.0	1.0	4.0
C8, A4, A8, A10, A11	1	1.0	1.0	1.0	4.0
A12, A13, A14	1	1.0	1.0	1.0	4.0
C2, C5, C9	1	0.5	1.0	1.0	3.5
C3, A5	1	1.0	1.0	0.5	3.5
A6	1	1.0	0.5	1.0	3.5
W2, A7	1	0.5	0.5	1.0	3.0
C4	1	1.0	1.0	1.0	4.0
C10	1	0.5	1.0	1.0	3.5
A9	1	1.0	1.0	1.0	4.0
Average					3.79

Note: QI1–QI4 represent quality indicators used in the assessment process.
(1 = Yes, 0.5 = Partial)

From the data presented in Table 5, an average score of 3.79 is obtained, demonstrating a high quality of the reviewed papers. This high average reflects the stringent selection criteria used during the review process, which effectively ensured that only studies of substantial quality were included. Consequently, this outcome lends significant confidence to the reliability and validity of the conclusions drawn from these papers.

Main Findings for RQ3

The quality of current GenAI for RE research was evaluated using a systematic approach with four quality indicators: clarity of goals, effectiveness of evaluation, result presentation, and discussion of limitations. A standardized scoring system (0–1 points) was applied to 27 papers, resulting in a high average score of 3.79 out of 4. This high score indicates overall high quality of research in the field, reflecting stringent selection criteria and lending confidence to the reliability of findings in GenAI for RE studies.

Takeaway Message 3

Researchers should maintain high standards in research design and reporting, focusing on clear goals, robust evaluation methods, and comprehensive results presentation. They should use the quality assessment criteria as a guideline for future studies and strive for transparency to enhance reproducibility. Practitioners, by contrast, should critically evaluate GenAI for RE studies using these same criteria, prioritizing implementation of techniques from high-quality research while being cautious of findings from lower-scoring studies. This approach ensures that both the production and application of research in this field maintain high standards, ultimately leading to more reliable and effective GenAI solutions in RE practices.

5.4. RQ4: Gaps and Future Directions

Although GenAI has demonstrated immense potential in revolutionizing RE practices, a critical analysis of the current research landscape reveals several limitations that warrant further exploration. One prominent limitation is the predominant focus on leveraging LLMs, such as the GPT series, for requirements elicitation, analysis, and validation. This approach may encounter challenges when confronted with complex, domain-specific requirements, particularly in highly specialized or safety-critical domains, where GenAI may struggle to fully comprehend and capture nuanced domain knowledge and constraints.

In addressing these challenges and considering the various characteristics and properties of LLMs in the context of RE, our approach was guided by the comprehensive framework outlined in the survey by Naveed et al. [87]. This survey provides a structured overview of the key aspects and challenges in LLM development and application, which we have adapted to the specific context of RE. We particularly focused on the critical areas identified in their work, including computational costs,

bias and fairness issues, interpretability and explainability, and concerns related to safety and controllability, all of which have significant implications for RE practices. For instance, our discussion on bias and fairness was informed by their emphasis on the potential societal implications of these issues, which is particularly relevant in RE where biased requirements can lead to discriminatory systems. The issue of hallucinations in LLMs, a key point in their survey, guided our approach to addressing the challenge of generating incorrect or inconsistent requirements in the context of RE, where models need to provide reliable and accurate outputs across diverse projects and domains. Furthermore, our examination of safety and controllability challenges was enriched by the perspectives presented in their work, particularly in relation to the ethical deployment of LLMs. This is especially crucial in RE, where the outputs of these models can directly influence critical system specifications. By aligning our analysis with this established framework, we ensured a thorough and systematic examination of the multifaceted nature of LLMs in the specific context of RE.

In addition, a significant imbalance exists in the existing research, with a predominant concentration on the early stages of the RE process, such as requirements elicitation and analysis. Comparatively, there is a paucity of research exploring the later stages, including requirements management, evolution, and long-term maintenance. This disparity in research focus may lead to an incomplete understanding of GenAI’s applicability and potential across the entire RE lifecycle. Furthermore, concerns regarding the interpretability and traceability of GenAI outputs persist because understanding decision rationales and maintaining requirement sources are pivotal aspects of RE.

By aligning our analysis with this established framework, we ensured a thorough and systematic examination of the multifaceted nature of LLMs in the specific context of RE. Key challenges and limitations include:

- **Bias and Fairness:** Our analysis shows that existing research predominantly emphasizes the functional application of GenAI models in RE tasks, often overlooking the critical issue of bias. This overlooking of bias is particularly concerning during the requirements elicitation and analysis phases, where implicit biases in training data can be unintentionally amplified by models. These biases not only pose ethical concerns, such as perpetuating societal stereotypes, but also have important implications in high-stakes applications such as hiring or law enforcement [88, 89]. Addressing these challenges requires a multi-faceted approach, including the careful curation of training data and the development of fairness-aware algorithms [90]. Future research should prioritize techniques for identifying and

mitigating AI biases, ensuring that RE processes and applications remain both ethically sound and socially responsible.

- **Ethical and Regulatory Concerns:** Our review found that most studies did not thoroughly explore the ethical implications of applying GenAI to RE, even though requirements are foundational to the software development lifecycle. The integration of LLMs in RE tasks introduces unique ethical challenges, such as the potential for generating biased or harmful content, facilitating misinformation, or misinterpreting user intentions [33, 91]. As the role of GenAI in RE expands, an urgent need exists to establish ethical guidelines and regulatory frameworks tailored to the specific challenges of RE. Collaboration between researchers, policymakers, and industry stakeholders is essential to ensure that the use of GenAI in RE is responsible, aligned with societal norms, and adheres to ethical standards, helping mitigate risks while fostering innovation in the field.
- **Security and Privacy:** LLMs used in RE often process large volumes of data, including sensitive and confidential project requirements. Although some studies briefly mention security as an important characteristic in RE, few specifically addressed the critical concerns of ensuring data privacy and model security when using GenAI. In adversarial settings, malicious actors may attempt to manipulate outputs or extract confidential information from these models [34, 35]. As GenAI is increasingly applied to sensitive RE tasks, developing robust security protocols and privacy-preserving techniques tailored to handling sensitive requirements will be crucial for mitigating these risks. Ensuring both the integrity of the models and the confidentiality of the data they process should be a key research focus moving forward.
- **Interpretability and Explainability:** Despite the recognized importance of interpretability in RE, a substantial research gap exists in exploring the explainability of LLMs and GenAI systems in RE tasks [36]. As these models become more complex, understanding their decision-making processes for requirements generation and analysis grows increasingly challenging, raising concerns about reliability and accountability, especially in sensitive domains such as healthcare and law. Future RE research must focus on developing novel, RE-specific explainability techniques for LLMs, addressing how these models leverage pre-trained knowledge and in-context learning for RE tasks [37]. Improving the transparency and interpretability of GenAI models in RE is crucial for ensuring stakeholder trust and effective integration of these technologies into RE practices.

- **Computational and Economic Cost:** Training and deploying LLMs for RE demand significant computational resources, leading to substantial economic and environmental costs. Our review nonetheless found that most studies failed to address the computational and economic implications of applying GenAI in RE, which is an alarming oversight given the considerable expenses involved in training and maintaining these models. The power consumption associated with large-scale training is a growing concern, as is the concentration of LLM development within well-funded organizations, potentially exacerbating inequalities within AI research [38, 39]. Future research must prioritize evaluating the cost-effectiveness of GenAI technologies in RE and exploring more sustainable and economically viable alternatives to deploying these models in practice.
- **Real-Time Processing:** Although LLMs are increasingly expected to handle real-time processing tasks, such as interactive dialogues and decision-making in dynamic environments, most research in GenAI for RE remains focused on the earlier stages of RE, such as elicitation and analysis. Few studies have explored how GenAI can adapt to dynamically changing requirements or process new requirement information in real-time. Meeting the demands of real-time performance while maintaining accuracy and robustness poses an ongoing challenge, particularly as models scale and tasks become more complex [40]. Developing GenAI systems capable of responding to rapidly evolving project environments will be an essential direction for future research in RE.
- **Hallucinations:** Recent research has extensively characterized hallucinations in LLM outputs, categorizing them into input-conflicting, context-conflicting, and fact-conflicting types. Although various mitigation strategies have been proposed, including improved data curation, reinforcement of learning techniques, and leveraging external knowledge [41, 42], their applicability in high-stakes domains such as RE remains largely unexplored. Given the critical role that accurate requirements play in project success, reducing hallucinations and improving the precision of generated requirements are critical areas for future research. Because hallucinations can have significant consequences in RE tasks, more focused exploration of effective mitigation strategies is needed, including developing reliable automated evaluation metrics specific to RE, adapting existing techniques such as multi-agent interaction and uncertainty estimation to RE contexts, and exploring novel approaches that ensure LLMs can deliver reliable outputs in complex and accuracy-demanding scenarios. In addition, investigating the trade-offs between reducing hallucinations and maintaining model capabilities in RE applications is crucial for the practical implementation of LLMs.

- **Reproducibility:** Our review identified a significant gap in studies addressing the reproducibility of GenAI outputs in RE. In an RE environment, ensuring consistent and reproducible results in requirements generation and analysis is crucial for building stakeholder trust. The stochastic processes involved in LLM inference, combined with the sensitivity to small parameter changes, make it difficult to validate and rely on their responses in mission-critical tasks. Developing methods to enhance the reproducibility of GenAI outputs in RE, including standardizing evaluation metrics, providing detailed parameter settings, and exploring reproducibility under varied conditions, is essential [43]. These efforts should aim to ensure reliability and foster confidence in their use across high-stakes scenarios. In addition, as the field moves toward using more black-box commercial LLMs, addressing reproducibility challenges in these closed systems becomes increasingly important for RE applications.
- **Controllability:** Although some studies have explored prompt engineering techniques, the precise control of GenAI models to generate outputs that align with specific project or organizational needs remains under-researched, particularly in the context of RE. Controllability is increasingly recognized as a critical feature of LLMs; however, many studies do not delve deeply into improving control over model outputs. This represents a key research gap, especially as these models are applied in sensitive or unpredictable environments [44, 92]. Future research must focus on enhancing the controllability of GenAI in RE tasks to ensure that outputs are tailored to specific, often stringent, requirements.
- **Authorship and Copyright:** Surprisingly, our review did not find any studies that explored ownership and copyright issues in the context of GenAI-generated requirements documents. As these technologies become more widely applied in RE practice, the ambiguity surrounding the legal status of AI-generated content is expected to grow [93]. Relatively few papers have tackled the complexities of authorship and intellectual property, despite the rising proficiency of LLMs in generating text. Clarifying ownership rights and developing legal frameworks that balance intellectual property protection with fostering innovation in AI-generated content is an important area for interdisciplinary research between legal and technical experts [45].

To address these gaps and pave the way for future advancements, research on GenAI in RE should progress in several critical directions. First, investigating more effective methods for integrating domain knowledge into GenAI models is a pressing need. This research may involve developing domain-specific fine-tuning techniques or

constructing domain-specific knowledge graphs to enhance AI’s understanding of the intricacies and nuances within specific domains. Second, research should be extended to encompass the full RE lifecycle, with a particular emphasis on exploring GenAI’s potential in requirements evolution, conflict detection, and consistency maintenance.

Another crucial avenue for future research is the development of more robust human–AI collaboration frameworks. This research extends beyond merely improving the output quality of GenAI models; it necessitates the design of intuitive interaction interfaces that empower requirements engineers to effectively guide, validate, and refine AI-generated outputs. Fostering a synergistic collaboration between AI’s creative capabilities and human experts’ judgment can unlock the potential for achieving superior RE outcomes.

In addition, concerns regarding the interpretability and traceability of GenAI outputs persist because understanding decision rationales and maintaining requirement sources are pivotal aspects of RE. The development of guidelines for responsible AI deployment and the establishment of ethical standards and regulatory compliance mechanisms should form an integral part of future research endeavors.

By addressing these challenges and pursuing these research directions, we can unlock the full potential of GenAI in revolutionizing RE processes, leading to more efficient, accurate, and ethically sound software development practices. This interdisciplinary approach will be crucial to ensuring that the adoption of GenAI-assisted RE practices aligns with societal values and promotes trust in the technology. By proactively addressing these concerns, researchers can ensure that the adoption of GenAI-assisted RE practices aligns with societal values and promotes trust in the technology.

Furthermore, establishing comprehensive evaluation frameworks to assess the performance of GenAI in RE is imperative. These frameworks should transcend traditional metrics of accuracy and completeness, incorporating factors such as comprehensibility, consistency, and adaptability. Long-term empirical studies are also essential to gauge the impact of GenAI-assisted RE on overall software project quality and success rates, providing valuable insights into the real-world efficacy of these approaches.

Lastly, it is imperative for researchers to confront the ethical and legal implications associated with the application of GenAI in RE. These implications encompass critical aspects such as data privacy, bias mitigation, and decision accountability. The development of guidelines for responsible AI deployment and the establishment of ethical standards and regulatory compliance mechanisms should form an integral part of future research endeavors. By proactively addressing these concerns, researchers can ensure that the adoption of GenAI-assisted RE practices aligns with

societal values and promotes trust in the technology.

Main Findings for RQ4

Current research on GenAI for RE reveals significant limitations and challenges that hinder the realization of its full potential. A primary concern is the overreliance on LLMs, which struggle with complex domain-specific requirements, especially in specialized or safety-critical domains. This overreliance is compounded by an imbalanced research focus that favors early RE stages while neglecting later phases of the RE lifecycle.

Persistent issues with the interpretability and traceability of GenAI outputs, coupled with a lack of comprehensive evaluation frameworks, further complicate the field. The research highlights critical challenges, including bias and fairness concerns, ethical and regulatory issues, security and privacy risks, and high computational costs. Additional challenges involve real-time processing difficulties, the potential for hallucinations in AI-generated content, reproducibility issues, limited model controllability, and unresolved questions about authorship and copyright.

These findings underscore the need for a more holistic approach to GenAI in RE. Future efforts should address the full lifecycle of RE, improve the handling of specialized domain knowledge, and ensure transparency and accountability in AI-generated outputs. Tackling these multifaceted challenges is crucial for realizing the true potential of GenAI in revolutionizing RE practices.

Takeaway Message 4

For Researchers

1. **Advancing GenAI Models and Lifecycle Coverage:** Develop comprehensive GenAI models that integrate domain-specific knowledge and span the entire RE lifecycle, emphasizing complex scenarios and later stages.
2. **Enhancing Interpretability and Evaluation:** Develop transparent and traceable AI systems with comprehensive evaluation frameworks to assess both immediate performance and long-term impact, enhancing overall trust in GenAI outputs for RE.
3. **Addressing Ethical and Technical Challenges:** Address critical GenAI challenges in RE by focusing on ethical considerations, bias mitigation, security, and performance optimization, while advancing methods for reproducible, controllable, and real-time AI processing.
4. **Aligning with Societal Values:** Promote the responsible integration of GenAI in RE by balancing technical innovations with ethical considerations and broader societal impacts, ensuring sustainable and beneficial advancements in the field.

For Practitioners

1. **Transforming RE Practices Through Human-AI Collaboration:** Develop robust frameworks for human-AI collaboration in RE, leveraging GenAI as a tool to augment and enhance human expertise rather than replace it.
2. **Implementing Ethical and Gradual Adoption Strategies:** Implement GenAI in RE through a phased, compliance-driven approach, starting with proven areas and gradually expanding while adhering to industry standards and regulations.
3. **Conducting Empirical Studies and Shaping Industry Standards:** Conduct comprehensive real-world project studies to understand the long-term impact of GenAI in RE, using insights to actively shape industry best practices and guidelines for responsible AI adoption.
4. **Balancing Innovation with Practical Considerations:** Optimize RE processes by leveraging GenAI to improve efficiency, while proactively addressing challenges such as computational costs, data privacy, and process integrity, ensuring a balanced approach to navigating technological complexities.

6. Threats to Validity

In this systematic literature review, we conducted a comprehensive assessment of potential threats to the validity of the findings, encompassing internal, external, and construct validity aspects of the research. By critically examining these threats, we aim to provide a transparent and rigorous evaluation of the limitations and strategies employed to mitigate them. Internal validity, which pertains to the robustness and integrity of the research design and execution, is subject to two primary threats. First, despite the implementation of a systematic literature search and screening method, the inherent risk of inadvertently omitting relevant studies remains. This threat arises from the possibility of studies being indexed in databases not covered by our search strategy or studies using alternative terminology not captured by our search strings. To mitigate this threat, we conducted a comprehensive search across multiple reputable databases and applied a meticulously crafted set of inclusion and exclusion criteria to ensure the identification of pertinent literature. Second, the process of data extraction and analysis is susceptible to subjective judgments and potential biases introduced by individual researchers. To address this concern, we used a standardized data extraction template to maintain consistency and implemented a rigorous cross-checking procedure involving three independent researchers to minimize bias and ensure the reliability of the extracted data.

The external validity of the study, which pertains to the generalizability and applicability of the findings to wider contexts, is subject to two primary limitations that may circumscribe the extent to which the results can be extrapolated beyond the specific research setting. The temporal scope of the study, focusing on literature published between 2019 and 2024, may not fully capture the comprehensive state of GenAI applications in RE. This limitation is particularly relevant given the rapid evolution and proliferation of GenAI technologies in recent years. Consequently, the findings of this review may not entirely reflect the most recent advancements and innovations in the field. In addition, the generalizability of the conclusions drawn from this study may be limited by the specific characteristics and contexts of the included studies, such as the domain of application, the scale of the projects, and the cultural or organizational settings in which the research was conducted. These limitations underscore the importance of interpreting the findings with caution and considering the specific contextual factors when applying the insights to different scenarios. Construct validity, which relates to the definition and measurement of research concepts, presents inherent challenges in the domain of GenAI applications in RE. Although widely accepted quality assessment criteria were adopted to evaluate the included studies, we acknowledge that these standards may not be universally applicable or entirely comprehensive in capturing the nuances and specificities of

this emerging field. The rapid evolution of GenAI technologies and their application in RE necessitates the continuous refinement and adaptation of evaluation frameworks to ensure their relevance and effectiveness. In addition, the proposed analysis framework in this study, while based on a thorough examination of the literature and expert consultation, may not exhaustively encompass all the significant aspects and dimensions of GenAI applications in RE. This limitation highlights the need for ongoing research and discourse to identify and incorporate additional factors that can influence the effectiveness and impact of these technologies in practice.

To mitigate these threats to validity, a multipronged approach was employed. First, a rigorous and systematic literature search and screening process was adopted, where multiple databases were adopted and well-defined inclusion and exclusion criteria were applied to minimize the risk of omitting relevant studies. Second, cross-checking by multiple researchers was implemented to ensure the reliability and consistency of data extraction and analysis, reducing the influence of individual biases. Third, the temporal and technological limitations of the study were explicitly acknowledged, emphasizing the need for continuous updates and expansions of the review as the field progresses. Fourth, the use of widely accepted quality assessment criteria was complemented by a critical reflection on their potential limitations and the recognition of the need for tailored evaluation frameworks specific to GenAI applications in RE. Finally, the classification framework was iteratively refined through a combination of literature analysis and expert consultation, aiming to capture the most salient aspects of the field while acknowledging the potential for further enhancements.

7. Roadmap for Advancing GenAI in RE

The rapid evolution of GenAI and its increasing application in RE necessitates a structured approach to guide future research efforts. This roadmap outlines key areas of focus, potential challenges, and promising directions for advancing the field of GenAI in RE.

7.1. *Advancing Model Capabilities and Domain Adaptation*

To enhance the effectiveness of GenAI in RE, future research should focus on improving model capabilities and their adaptation to specific domains:

- **Enhancing Domain-Specific Knowledge Integration:** Fine-tune LLMs with domain-specific datasets, create and maintain domain-specific knowledge graphs, and explore methods for dynamic knowledge updating.

- **Expanding GenAI Applications Across RE Lifecycle:** Investigate GenAI applications in requirements management, evolution, validation, verification, and reuse across projects.
- **Optimizing Computational Efficiency:** Investigate model compression, optimization, federated learning, and efficient fine-tuning strategies.

7.2. Enhancing Human-AI Collaboration and Interpretability

As GenAI becomes more integrated into RE processes, ensuring effective collaboration between humans and AI systems is crucial:

- **Improving Interpretability and Transparency:** Develop explainable AI techniques, visualization tools, and methods to maintain traceability between GenAI-generated artifacts and their sources.
- **Enhancing Human-AI Collaboration in RE:** Design intuitive interfaces and interaction paradigms, develop adaptive GenAI models that learn from human feedback, and investigate cognitive aspects of human-AI collaboration.

7.3. Addressing Ethical, Legal, and Security Considerations

As the use of GenAI in RE grows, it is essential to address associated ethical, legal, and security challenges:

- **Ethical and Legal Considerations:** Develop frameworks for ethical GenAI use, investigate privacy-preserving techniques, and explore legal implications of GenAI-generated artifacts.
- **Enhancing Security and Reliability:** Develop robust testing and validation frameworks, and explore methods to ensure consistency, reliability, and protection against adversarial attacks.

7.4. Standardizing Evaluation and Benchmarking

To effectively assess and compare different GenAI approaches in RE, standardized evaluation frameworks are necessary:

- **Developing Standardized Evaluation Frameworks:** Create benchmark datasets and tasks, develop comprehensive metrics, and establish guidelines for conducting and reporting empirical studies on GenAI applications in RE.

This roadmap provides a structured guide for future research efforts in applying GenAI to RE. As the field evolves, researchers should focus on integrating these efforts, assessing their long-term impact, and addressing challenges related to reproducibility, controllability, real-time processing, bias mitigation, fairness, security, and privacy. Investigating solutions to reduce computational and economic costs, conducting extended empirical research on the impact of GenAI-assisted RE, and assessing the societal implications of widespread GenAI use in RE are also critical areas for future research.

8. Conclusion

This systematic literature review provides a comprehensive analysis of the current state and future directions of GenAI applications in RE. Our rigorous examination of 27 papers published between 2022 and 2024 reveals a rapidly evolving field with significant potential to transform RE practices, while also highlighting critical challenges that need to be addressed. The analysis underscores the predominant use of LLMs, particularly the GPT series, across various RE tasks. These models have shown promising results, especially in requirements elicitation, analysis, and validation phases. The high average quality score of 3.79 across the reviewed papers demonstrates the rigor and relevance of current research in this domain, indicating a strong foundation for future advancements. Our findings clearly indicate that GenAI is making substantial contributions to enhancing functional suitability, reliability, and efficiency in RE processes, offering innovative solutions to longstanding challenges in the field. However, our review also reveals several limitations and challenges in the current research landscape. A notable imbalance exists in research focus, with a concentration on early stages of the RE process, particularly requirements elicitation and analysis. By contrast, GenAI applications in later stages such as requirements management, evolution, and long-term maintenance has been underexplored. This disparity suggests a critical gap in understanding how GenAI can support the full RE lifecycle.

The application of GenAI in complex, domain-specific, or safety-critical areas remains a substantial challenge. Current models often struggle with nuanced, specialized requirements, highlighting the need for more sophisticated approaches to domain knowledge integration. In addition, ensuring the interpretability and traceability of AI-generated outputs poses a persistent challenge and is crucial for maintaining transparency and accountability in RE processes. Our review also identified critical issues that demand urgent attention from the research community. These include concerns about bias and fairness in AI-generated content, ethical and regulatory considerations, security and privacy risks associated with processing sensitive

requirements data, and the high computational costs of deploying large AI models. The potential for hallucinations in AI-generated content, difficulties in real-time processing and adaptation to changing requirements, reproducibility challenges, and limited controllability of AI models in RE tasks further complicate the practical application of GenAI in RE. Despite these challenges, the potential of GenAI to revolutionize RE practices remains strong. The field is at a critical juncture, poised for breakthrough advancements that could address these limitations and unlock new possibilities in RE. The high quality of current research provides a strong foundation for these future developments, suggesting a promising trajectory for the field.

CRedit authorship contribution statement

Haowei Cheng designed the overall research process and authored the main manuscript. **Jati H. Husen** provided support for conceptualization, investigation, and validation. **Sien Reeve Peralta** and **Bowen Jiang** assisted with validation efforts. **Nobukazu Yoshioka** and **Naoyasu Ubayashi** provided methodology, investigation and validation. **Hironori Washizaki** contributed to the conceptualization and offered supervision and guidance throughout the study. All authors reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Link to data is available in the manuscript.

Acknowledgments

This work was supported by JST SPRING (grant number JPMJSP2128) and the JST-Mirai program (grant number JPMJMI20B8).

References

- [1] R. Pressman and D. Bruce R. Maxim, *Software Engineering: A Practitioner's Approach*. McGraw-Hill Education, 2014.
- [2] A. L. Lederer and J. Prasad, "Causes of inaccurate software development cost estimates," in *Journal of Systems and Software*, vol. 31, 1995, pp. 125–134.
- [3] S. Group, "Benchmarks and assessments-virtual success ladder benchmark," in *Chaos Report*, 2020. [Online]. Available: <https://www.standishgroup.com/benchmark>
- [4] A. Svyatkovskiy, S. K. Deng, S. Fu, and N. Sundaresan, "Intellicode compose: code generation using transformer," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2020, p. 1433–1443.
- [5] X. Chen, Y. Zhao, Q. Wang, and Z. Yuan, "Multi: Multi-objective effort-aware just-in-time software defect prediction," *Information and Software Technology*, vol. 93, pp. 1–13, 2018.
- [6] M. Motwani, S. Sankaranarayanan, R. Just, and Y. Brun, "Do automated program repair techniques repair hard and important bugs?" in *Proceedings of the 40th International Conference on Software Engineering*. Association for Computing Machinery, 2018, p. 25.
- [7] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *Proceedings of the Conference on The Future of Software Engineering*. Association for Computing Machinery, 2000, p. 35–46.
- [8] "Iso/iec/ieee international standard - systems and software engineering – life cycle processes – requirements engineering," *ISO/IEC/IEEE 29148:2018(E)*, pp. 1–104, 2018.
- [9] K. Pohl, *Requirements Engineering: Fundamentals, Principles, and Techniques*, 1st ed. Springer Berlin, Heidelberg, 2010.
- [10] K. E. Wiegers and J. Beatty, *Software requirements*. Pearson Education, 2013.
- [11] B. H. Cheng and J. M. Atlee, "Research directions in requirements engineering," in *Future of Software Engineering (FOSE '07)*, 2007, pp. 285–303.

- [12] X. Franch, “Data-driven requirements engineering: A guided tour,” in *Evaluation of Novel Approaches to Software Engineering*, R. Ali, H. Kaindl, and L. A. Maciaszek, Eds. Cham: Springer International Publishing, 2021, pp. 83–105.
- [13] K. Zamani, D. Zowghi, and C. Arora, “Machine learning in requirements engineering: A mapping study,” in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 116–125.
- [14] N. Marques, R. R. Silva, and J. Bernardino, “Using chatgpt in software requirements engineering: A comprehensive review,” *Future Internet*, vol. 16, no. 6, 2024. [Online]. Available: <https://www.mdpi.com/1999-5903/16/6/180>
- [15] B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” 2020. [Online]. Available: <https://arxiv.org/abs/2002.04087>
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [17] A. M. Dakhel, A. Nikanjam, F. Khomh, M. C. Desmarais, and H. Washizaki, *An Overview on Large Language Models*, A. Nguyen-Duc, P. Abrahamsson, and F. Khomh, Eds., 2024.
- [18] P. Brar and D. Nandal, “A systematic literature review of machine learning techniques for software effort estimation models,” in *Proceedings of the 5th International Conference on Computational Intelligence and Communication Technologies (CCICT)*, 2022, pp. 494–499.
- [19] H. Tian, K. Liu, A. K. Kaboré, A. Koyuncu, L. Li, J. Klein, and T. F. Bissyandé, “Evaluating representation learning of code changes for predicting patch correctness in program repair,” in *Proceedings of the 35th International Conference on Automated Software Engineering*. Association for Computing Machinery, 2021, p. 981–992.
- [20] P. Talele and R. Phalnikar, “Software requirements classification and prioritisation using machine learning,” in *Machine Learning for Predictive Analysis*, 2021, pp. 257–267.
- [21] M. Rahimi, M. Mirakhorli, and J. Cleland-Huang, “Automated extraction and visualization of quality concerns from requirements specifications,” 2014, pp. 253–262.

- [22] S.-C. Necula, F. Dumitriu, and V. Greavu-Şerban, “A systematic literature review on using natural language processing in software requirements engineering,” vol. 13, no. 11, 2024.
- [23] N. Marques, R. R. Silva, and J. Bernardino, “Using chatgpt in software requirements engineering: A comprehensive review,” vol. 16, no. 6, 2024. [Online]. Available: <https://www.mdpi.com/1999-5903/16/6/180>
- [24] A. Mastropaolo, S. Scalabrino, N. Cooper, D. Nader Palacio, D. Poshyvanyk, R. Oliveto, and G. Bavota, “Studying the usage of text-to-text transfer transformer to support code-related tasks,” in *Proceedings of the 43rd International Conference on Software Engineering (ICSE)*, 2021, pp. 336–347.
- [25] D. Fried, A. Aghajanyan, J. Lin, S. Wang, E. Wallace, F. Shi, R. Zhong, W. tau Yih, L. Zettlemoyer, and M. Lewis, “InCoder: A generative model for code infilling and synthesis,” 2023. [Online]. Available: <https://arxiv.org/abs/2204.05999>
- [26] S. Zong, A. Ritter, G. Mueller, and E. Wright, “Analyzing the perceived severity of cybersecurity threats reported on social media.” Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 1380–1390. [Online]. Available: <https://aclanthology.org/N19-1140>
- [27] A. Nguyen-Duc, B. Cabrero-Daniel, A. Przybyłek, C. Arora, D. Khanna, T. Herda, U. Rafiq, J. Melegati, E. Guerra, K.-K. Kemell, M. Saari, Z. Zhang, H. Le, T. Quan, and P. Abrahamsson, “Generative artificial intelligence for software engineering – a research agenda,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.18648>
- [28] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim, “A survey on large language models for code generation,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.00515>
- [29] S. Santos, T. Breaux, T. Norton, S. Haghighi, and S. Ghanavati, “Requirements satisfiability with in-context learning,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.12576>
- [30] A. Fantechi, S. Gnesi, L. Passaro, and L. Semini, “Inconsistency detection in natural language requirements using chatgpt: a preliminary evaluation,” in *Proceedings of 31st International Requirements Engineering Conference (RE)*, 2023, pp. 335–340.

- [31] K. Ronanki, C. Berger, and J. Horkoff, “Investigating chatgpt’s potential to assist in requirements elicitation processes,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.07381>
- [32] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” 2022. [Online]. Available: <https://arxiv.org/abs/1908.09635>
- [33] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of ai ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, p. 399, 2019.
- [34] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, “Extracting training data from large language models,” 2021. [Online]. Available: <https://arxiv.org/abs/2012.07805>
- [35] Y. Li, T. Baldwin, and T. Cohn, “Towards robust and privacy-preserving text representations,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 25–30.
- [36] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, “Explainability for large language models: A survey,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.01029>
- [37] S. Huang, S. Mamidanna, S. Jangam, Y. Zhou, and L. H. Gilpin, “Can large language models explain themselves? a study of llm-generated self-explanations,” 2023. [Online]. Available: <https://arxiv.org/abs/2310.11207>
- [38] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for deep learning in nlp,” 2019. [Online]. Available: <https://arxiv.org/abs/1906.02243>
- [39] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623. [Online]. Available: <https://doi.org/10.1145/3442188.3445922>
- [40] T. Team, “Generative ai apps,” 2024, accessed: 2024-08-27. [Online]. Available: <https://www.timeplus.com/post/generative-ai-apps>

- [41] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, L. Wang, A. T. Luu, W. Bi, F. Shi, and S. Shi, “Siren’s song in the ai ocean: A survey on hallucination in large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.01219>
- [42] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.00661>
- [43] A. Belz, S. Agarwal, A. Shimorina, and E. Reiter, “A systematic review of reproducibility research in natural language processing,” in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, 2021, pp. 381–393. [Online]. Available: <https://aclanthology.org/2021.eacl-main.29>
- [44] E. K. Chong, “The control problem [president’s message],” *IEEE Control Systems Magazine*, vol. 37, no. 2, pp. 14–16, 2017.
- [45] J. Ihalainen, “Computer creativity: artificial intelligence and copyright,” *Journal of Intellectual Property Law & Practice*, vol. 13, no. 9, pp. 724–728, 03 2018.
- [46] D. Méndez Fernández and S. Wagner, “Naming the pain in requirements engineering: A design for a global family of surveys and first results from germany,” vol. 57, pp. 616–643, 2015.
- [47] C. Pacheco, I. García, and M. Reyes, “Requirements elicitation techniques: a systematic literature review based on the maturity of the techniques,” *IET Software*, vol. 12, no. 4, pp. 365–378, 2018.
- [48] K. Pohl, *Requirements Engineering: Fundamentals, Principles, and Techniques*, 1st ed. Springer-Verlag Berlin Heidelberg, 2010.
- [49] S. Jayatilleke and R. Lai, “A systematic review of requirements change management,” *Information and Software Technology*, vol. 93, pp. 163–185, 2018.
- [50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and

- D. Amodei, “Language models are few-shot learners,” 2020. [Online]. Available: <https://arxiv.org/abs/2005.14165>
- [51] J. Zhang, Y. Chen, N. Niu, Y. Wang, and C. Liu, “Empirical evaluation of chatgpt on requirements information retrieval under zero-shot setting,” 2023. [Online]. Available: <https://arxiv.org/abs/2304.12562>
- [52] S. Arulmohan, M.-J. Meurs, and S. Mosser, “Extracting domain models from textual requirements in the era of large language models,” in *Proceedings of the 26th International Conference on Model-Driven Engineering Languages and Systems*. MDEIntelligence, 2023.
- [53] K. Ronanki, C. Berger, and J. Horkoff, “Investigating chatgpt’s potential to assist in requirements elicitation processes,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.07381>
- [54] C. Jain, P. R. Anish, A. Singh, and S. Ghaisas, “A transformer-based approach for abstractive summarization of requirements from obligations in software engineering contracts,” in *Proceedings of the 31st International Requirements Engineering Conference (RE)*, 2023, pp. 169–179.
- [55] A. Fantechi, S. Gnesi, L. Passaro, and L. Semini, “Inconsistency detection in natural language requirements using chatgpt: a preliminary evaluation,” in *Proceedings of the 31st International Requirements Engineering Conference (RE)*, 2023, pp. 335–340.
- [56] B. Görer and F. B. Aydemir, “Generating requirements elicitation interview scripts with large language models,” in *Proceedings of the 31st International Requirements Engineering Conference Workshops (REW)*, 2023, pp. 44–51.
- [57] D. Clements, E. Giannis, F. Crowe, M. Balapitiya, J. Marshall, P. Papadopoulos, and T. Kanji, “An innovative approach to develop persona from application reviews,” in *Proceedings of the 18th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2023)*. SCITEPRESS – Science and Technology Publications, Lda., 2023, pp. 701–708.
- [58] N. Blasek, K. Eichenmüller, B. Ernst, N. Götz, B. Nast, and K. Sandkuhl, “Large language models in requirements engineering for digital twins,” in *Companion Proceedings of the 16th IFIP WG 8.1 Working Conference on the Practice of Enterprise Modeling and the 13th Enterprise Design and Engineering Working Conference*, 2023.

- [59] B. Chen, K. Chen, S. Hassani, Y. Yang, D. Amyot, L. Lessard, G. Mussbacher, M. Sabetzadeh, and D. Varró, “On the use of gpt-4 for creating goal models: An exploratory study,” in *Proceedings of the 31st International Requirements Engineering Conference Workshops (REW)*, 2023, pp. 262–271.
- [60] M. Cosler, C. Hahn, D. Mendoza, F. Schmitt, and C. Trippel, “nl2spec: Interactively translating unstructured natural language to temporal logics with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2303.04864>
- [61] H. Mustroph, M. Barrientos, K. Winter, and S. Rinderle-Ma, “Verifying resource compliance requirements from natural language text over event logs,” in *Business Process Management*, 2023, pp. 249–265.
- [62] I. Gräßler, D. Preuß, L. Brandt, and M. Mohr, “Efficient extraction of technical requirements applying data augmentation,” in *Proceedings of International Symposium on Systems Engineering (ISSE)*, 2022, pp. 1–8.
- [63] J. O. Couder, D. Gomez, and O. Ochoa, “Requirements verification through the analysis of source code by large language models,” in *SoutheastCon 2024*, 2024, pp. 75–80.
- [64] J. U. Oswal, H. T. Kanakia, and D. Suktel, “Transforming software requirements into user stories with gpt-3.5 -: An ai-powered approach,” in *Proceedings of the 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, 2024, pp. 913–920.
- [65] J. Peer, Y. Mordecai, and Y. Reich, “Nlp4ref: Requirements classification and forecasting: From model-based design to large language models,” in *2024 IEEE Aerospace Conference*, 2024, pp. 1–16.
- [66] M. Krishna, B. Gaur, A. Verma, and P. Jalote, “Using llms in software requirements specifications: An empirical evaluation,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.17842>
- [67] A. Nouri, B. Cabrero-Daniel, F. Törner, H. Sivencrona, and C. Berger, “Engineering safety requirements for autonomous driving with large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.16289>
- [68] D. Jin, Z. Jin, X. Chen, and C. Wang, “Mare: Multi-agents collaboration framework for requirements engineering,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.03256>

- [69] R. Feldt and R. Coppola, “Semantic api alignment: Linking high-level user goals to apis,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.04236>
- [70] M. Ataei, H. Cheong, D. Grandi, Y. Wang, N. Morris, and A. Tessier, “Elicित्रon: An llm agent-based simulation framework for design requirements elicitation,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.16045>
- [71] Z. Zhang, M. Rayhan, T. Herda, M. Goisauf, and P. Abrahamsson, “Llm-based agents for automating the enhancement of user story quality: An early report,” in *Agile Processes in Software Engineering and Extreme Programming*, 2024, pp. 117–126.
- [72] B. Wang, C. Wang, P. Liang, B. Li, and C. Zeng, “How llms aid in uml modeling: An exploratory study with novice analysts,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.17739>
- [73] A. El-Hajjami, N. Fafin, and C. Salinesi, “Which ai technique is better to classify requirements? an experiment with svm, lstm, and chatgpt,” 2024. [Online]. Available: <https://arxiv.org/abs/2311.11547>
- [74] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, and C. Berger, “Requirements engineering using generative ai: Prompts and prompting patterns,” 2023. [Online]. Available: <https://arxiv.org/abs/2311.03832>
- [75] M. Fazelnia, V. Koscinski, S. Herzog, and M. Mirakhorli, “Lessons from the use of natural language inference (nli) in requirements engineering tasks,” 2024. [Online]. Available: <https://arxiv.org/abs/2405.05135>
- [76] N. Feng, L. Marsso, S. G. Yaman, I. Standen, Y. Baatartogtokh, R. Ayad, V. O. de Mello, B. Townsend, H. Bartels, A. Cavalcanti, R. Calinescu, and M. Chechik, “Normative requirements operationalization with large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.12335>
- [77] S. Santos, T. Breau, T. Norton, S. Haghighi, and S. Ghanavati, “Requirements satisfiability with in-context learning,” 2024. [Online]. Available: <https://arxiv.org/abs/2404.12576>
- [78] *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI-based systems*, International Organization for Standardization Std. ISO/IEC 25 059:2019, 2019.

- [79] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 328–339.
- [80] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, “The curious case of neural text degeneration,” 2020. [Online]. Available: <https://arxiv.org/abs/1904.09751>
- [81] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, “Emergent abilities of large language models,” 2022. [Online]. Available: <https://arxiv.org/abs/2206.07682>
- [82] E. Filatovas, L. Stripinis, F. Orts, and R. Paulavičius, “Advancing research reproducibility in machine learning through blockchain technology,” *Informatica*, vol. 35, no. 2, p. 227–253, 2024. [Online]. Available: <https://doi.org/10.15388/24-INFOR553>
- [83] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, “Systematic literature reviews in software engineering—a systematic literature review,” in *Information and software technolog*, 7-15, pp. 1–7.
- [84] Kanewala, Upulee, Bieman, and J. M, “Testing scientific software: A systematic literature review,” vol. 56, no. 10. Elsevier, 2014, pp. 1219–1232.
- [85] A. Dadwal, H. Washizaki, Y. Fukazawa, T. Iida, M. Mizoguchi, and K. Yoshimura, “Prioritization in automotive software testing: Systematic literature review,” in *Proceedings of the 6th International Workshop on Quantitative Approaches to Software Quality (QuASoQ)*, 2018.
- [86] E. Pretel, A. Moya, E. Navarro, V. López-Jaquero, and P. González, “Analysing the synergies between multi-agent systems and digital twins: A systematic literature review,” in *Information and Software Technology*, vol. 174, 2024, p. 107503.
- [87] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, “A comprehensive overview of large language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2307.06435>

- [88] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, “Language (technology) is power: A critical survey of “bias” in NLP,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 5454–5476.
- [89] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, “Improving fairness in machine learning systems: What do industry practitioners need?” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2019, p. 1–16.
- [90] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, “On the opportunities and risks of foundation models,” 2022. [Online]. Available: <https://arxiv.org/abs/2108.07258>
- [91] M. Whittaker, K. Crawford, R. Dobbe, G. Fried, E. Kaziunas, V. Mathur, S. M. West, R. Richardson, J. Schultz, O. Schwartz *et al.*, *AI now report 2018*. AI Now Institute at New York University New York, 2018.
- [92] R. Yampolskiy, “On controllability of artificial intelligence,” in *IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021)*, 2020.
- [93] J.-M. Deltorn and F. Macrez, “Authorship in the age of machine learning and artificial intelligence,” *Centre for International Intellectual Property Studies*

(CEIPI) Research Paper No. 2018-10, 2018, forthcoming in: Sean M. O'Connor (ed.), *The Oxford Handbook of Music Law and Policy*, Oxford University Press, 2019. [Online]. Available: <https://ssrn.com/abstract=3261329>