Towards AI-centric Requirements Engineering for Industrial Systems

Sarmad Bashir
sarmad.bashir@{mdu.se,ri.se}
Mälardalen University & RISE Research Institutes of Sweden
Västerås, Sweden

ABSTRACT

Engineering large-scale industrial systems mandate an effective Requirements Engineering (RE) process. Such systems necessitate RE process optimization to align with standards, infrastructure specifications, and customer expectations. Recently, artificial intelligence (AI) based solutions have been proposed, aiming to enhance the efficiency of requirements management within the RE process. Despite their advanced capabilities, generic AI solutions exhibit limited adaptability within real-world contexts, mainly because of the complexity and specificity inherent to industrial domains. This limitation notably leads to the continued prevalence of manual practices that not only cause the RE process to be heavily dependent on practitioners' experience, making it prone to errors, but also often contributes to project delays and inefficient resource utilization. To address these challenges, this Ph.D. dissertation focuses on two primary directions: *i)* conduct a comprehensive focus group study with a large-scale industry to determine the requirements evolution process and their inherent challenges and *ii)* propose AI solutions tailored for industrial case studies to automate and streamline their RE process and optimize the development of large-scale systems. We anticipate that our research will significantly contribute to the RE domain by providing empirically validated insights in the industrial context.

KEYWORDS

requirements engineering, industrial automation, language models

ACM Reference Format:
Sarmad Bashir. 2024. Towards AI-centric Requirements Engineering for Industrial Systems . In *2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion '24), April 14–20, 2024, Lisbon, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3639478.3639811

1 INTRODUCTION

In the manufacturing industry, specifically within the railway domain, RE establishes a foundation to ensure the reliability of the software system under development. This is achieved by documenting clear, concise, and testable system-level requirements. Such systems consist of multiple interacting sub-systems or modules, which include, but are not limited to, traction, propulsion control, doors, and brakes [1]. Developing such sub-systems and their specific features depends on the user-centered requirements. This objective is realized by a systematic progression of RE stages, initiating with the elicitation of high-level customer requirements, extending to detailed feasibility analysis, and concluding with their verification and strategic allocation to sub-system modules for development [5]. In this context, the engineers and stakeholders must dedicate considerable time and effort to perform such RE tasks. In the industrial sector, the RE process predominantly depends on manual practices. This leads to inconsistencies, prolongs the feedback cycle and can adversely impact the quality of the system under development.

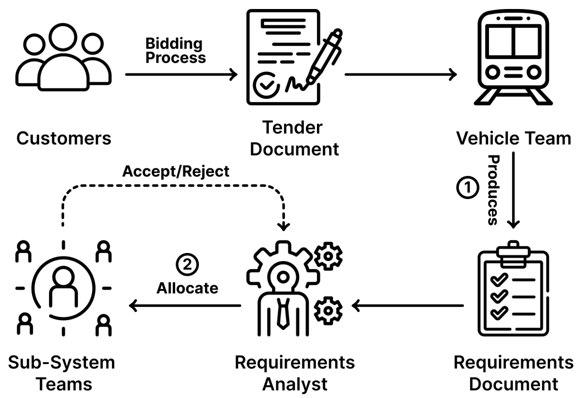

Figure 1: Overview of the requirements flow

Figure 1 shows an overview of the requirements management process observed in the railways' domain, focusing on the practices of our industrial partner Alstom[1]. The process commences with traditional tender calls initiated by customers seeking bids to procure specific products. A typical large tender document includes 600 to 1800 individual segments of natural language text encompassing high-level technical specifications (requirements), supporting information, and contractual obligations. The effective response to tender calls demands detailed risk analysis and time estimation to deliver the end products. Therefore, extracting high-level requirements from tender documents is crucial for enabling successful project acquisition and aiding in later RE activities. Currently, within the process, the vehicle team performs the task of manual labeling to distinguish between requirements and general

[1]https://www.alstom.com/alstom-sweden

ICSE-Companion '24, April 14–20, 2024, Lisbon, Portugal

ACM ISBN 979-8-4007-0502-1/24/04
https://doi.org/10.1145/3639478.3639811

information in tender documents (1). Once the agreed-upon customer requirements are extracted, they are allocated to multiple sub-system teams for implementation and rigorous testing. The requirements are assigned manually through a requirement management tool[2], where the relevant team can accept or reject them, especially if they are unrelated to their scope of work (2). If a requirement is incorrectly assigned to a sub-system team, the team will reject it, causing the repetition of the manual allocation process until the error is rectified. This results in extended feedback cycles, contributing to delays in project completion and an increase in time and effort.

While the effectiveness of current RE practices in developing highly reliable systems across various industries is evident, there is a growing imperative for enhanced support in the RE process. In this regard, the advent of advanced NLP and large language models (LLMs) can efficiently optimize complex industrial RE practices [6], especially when dealing with large amounts of textual artifacts. This is because requirements are predominantly expressed in natural language for developing large-scale systems across various industries. Therefore, context-driven NLP approaches offer scalable solutions that can support RE activities, such as identifying requirements from tender documents and allocating them to different sub-systems teams for development. Such AI-driven solutions can facilitate the project acquisition workflow and smart management of requirements in the industry, leading towards efficient development of large-scale systems.

Related Work. In the last decade, AI-based technologies have evolved rapidly and significantly adopted in various domains for multiple tasks [2, 13, 24, 36]. Within the realm of RE, multiple studies have proposed NLP-based solutions to support various stages of the RE process. This includes detecting linguistic issues in requirements [20, 21], classifying requirements into different categories [3, 23, 33], and establishing traceability links between requirements [26, 31]. A recent mapping study by Zhao et al. [36] comprehensively analyzed over 350 NLP solutions and their applications in various phases of RE, demonstrating the growing interest and potential in the area. However, the study also indicates that only 7% of the proposed solutions are evaluated in real-world industrial contexts. This leads to a significant challenge in adapting such approaches to support the RE process in industrial settings. According to the survey by Fernández *et al.* [19], only 16% of the companies are using automated techniques for the RE process. The issue arises primarily because existing proposed approaches are often evaluated on public datasets annotated by students or researchers (e.g., PROMISE [14]) and fail to scale effectively to the specific industrial domains' needs and complexities.

As a result, there is a pressing need to develop and empirically evaluate various AI-driven solutions that effectively capture the nuances of the industrial domain concerning the RE tasks. In this regard, such solutions can enable adaptability in industrial settings to streamline multiple activities within the RE process. Furthermore, practitioners operating in a requirements-centered context, such as the railways domain, can significantly benefit from these AI-tailored solutions.

[2]https://www.ibm.com/se-en/marketplace/requirements-management

2 RESEARCH HYPOTHESIS

The research hypothesis is formulated as follows:

> In developing large-scale industrial systems, the reliance on manual RE practices makes the development process error-prone, leading to scalability issues, project completion delays, and suboptimal resource utilization. By integrating NLP-based LLMs as automated tools, there is a potential to refine the RE process substantially. This enhancement is expected to mitigate project delays and foster more efficient management of resources and time in the RE process.

To test our hypothesis, we developed AI-driven solutions tailored to address the industry challenges within the RE process, specifically in identifying requirements and allocating them to multiple sub-system teams. For this purpose, we employed five recent representative tender documents from the railway domain. We empirically evaluated various classification methods to determine the most effective approach, from traditional machine learning and deep learning models to state-of-the-art LLMs. The results indicate that the proposed solutions are viable for real-world applications and establish a foundation for the early stage of technological adaption for automated RE processes within the industry.

3 METHODOLOGY

The proposed methodology employs a hybrid approach. Initially, a qualitative study will delve into the evolution of requirements and the challenges encountered in the RE process within an industrial context. Subsequently, the research conducts an empirical evaluation, employing multiple AI techniques in industry-focused case studies to automate RE tasks. This section will introduce two key research directions: the evolution of requirements, their challenges in the railway industry, and the proposed solutions to address these challenges.

A. Requirements evolution and RE practices at scale

Problem: Large-scale industries, such as railways, are characterized by complex, multi-faceted systems that involve various teams and stakeholders with different needs and perspectives [1]. It involves multiple activities, from the initial requirements elicitation, development, and testing phases to the deployment of systems. Each stage consists of unique requirement specifications that dynamically evolve because of technological advancements, regulatory shifts, and stakeholder needs. Understanding the evolution of requirements across various systems will facilitate the identification of distinct RE practices among different teams. Furthermore, similar patterns can be established to report common challenges and opportunities to improve the requirements management process throughout its entire lifecycle.

Proposed Approach: Following the guidelines of Breen [10], our approach is to conduct a focus group study, a qualitative research technique, with our industrial partner. This method is designed to collect diverse insights and perspectives from practitioners working across geographically distributed teams to develop their respective

sub-systems. The study aims to report an in-depth understanding of the dynamics, challenges, and strategic approaches inherent in the evolution of requirements. We anticipate that the results can inform the development of tailored RE tools and address management strategies for developing and standardizing RE practices within an industrial context.
Expected timeline: We plan to complete this work in six months.

B. Challenges within the RE process in the railway industry

Currently, two key areas have been identified through informal interviews and discussions with our industrial partner. We treated them as industrial case studies and evaluated the proposed solutions by following the guidelines of Runeson et al. [27] for reporting. Below, we discuss them in detail.

Case Study A. Requirements Identification from tender documents

Problem: Requirements specifications in tender documents serve as the central artifacts within the RE process as they outline the essential characteristics, qualities, and safety standards required for the system-to-be. In practical situations, the analysts are required to manually review such extensive documents, a process that is both prone to errors and time-consuming. Multiple studies have focused on automating requirements identification and classification tasks in the literature [3, 18]. However, these studies typically use traditional language models or single-sentence as a unit for the classification. In our case, the requirements or other supporting information consist of text fragments expanding over multiple sentences, making the existing evaluation of language models inadequate. Therefore, we rigorously evaluated various traditional and state-of-the-art language models for identifying requirements, utilizing open-source and industrial datasets. The work aims to automate RE task, i.e., distinguishing between requirements and other supporting information from tender documents within an industrial context.
Proposed Approach: The task is treated as a binary classification problem. We first apply multiple NLP techniques such as tokenization, part-of-speech tagging (POS), lemmatization, and stop-word removal to preprocess the datasets. Since the preprocessing of textual data can impact the performance of the language models, we trained them with and without preprocessing methods. To achieve efficient feature representation, we employed a lexical method, i.e., term frequency-inverse document frequency (tf-idf) and semantic techniques (FastText[3] and GloVe[4]) for training language models. Consequently, we constructed pipelines for the empirical evaluation of language models. These include traditional ML methods, fine-tuned LLMs such as the Bidirectional Encoder Representations from Transformers model (BERT) [17] and its variants, and few-shot classifiers.
Results: The pipelines are evaluated on five industrial tender documents and the open-source Dronology dataset [15]. We employed five-fold stratified cross-validation to assess the language models' generalizability. In both datasets, the BERT uncased language model

[3]https://fasttext.cc
[4]https://nlp.stanford.edu/projects/glove

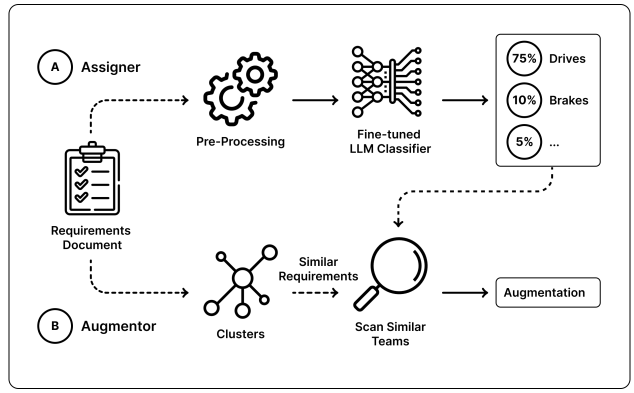

Figure 2: Overview of the Requirements Allocation approach

demonstrated better performance than other pipelines. The best average recall measure reported for industrial and open-source datasets was 82% and 88%, respectively.
Expected timeline: This part of the work has been published in REFSQ [8].

Case Study B. Requirements Allocation to multiple sub-system teams

Problem: In large-scale industrial systems, the platforms under development are typically decomposed into multiple interacting sub-systems. This approach promotes architectural modularity and maintainability [29]. In this context, the sub-system teams are responsible for the development, testing and integration of the associated requirements, thereby providing enhanced control over the entire RE process. In current practices, the requirements analyst performs the requirements allocation manually, a process that is not only time-intensive but subject to errors. In the literature, only a few studies have focused on requirements allocation [4, 12]. Furthermore, the previous work does not augment the classification process with additional information that could further support well-informed requirements allocation.
Proposed Approach: The approach consists of two modules to automate the requirements allocation task: the `Assigner` and `Augmenter` (as shown in the figure 2). The `Assigner` module is a classification pipeline that first applies text preprocessing techniques to remove the noisy elements from textual requirements and then utilizes a language model to predict the relevant sub-system team. We selected the language model based on empirical evaluation of various classification techniques. The `Augmenter` module is designed to generate supplementary supporting information to complement the predictions generated by the assigner module. It employs lexical similarity-based measurements to generate augmentations by retrieving the most likely sub-system teams. The `Augmenter` pipeline creates the clusters and searches for the most closely related requirements utilizing the same requirements document as `Assigner`. Subsequently, it assesses if the teams predicted by the `Assigner` correspond with the teams of the most similar requirements, as identified based on lexical feature analysis. This can be regarded as a complementary channel to provide enhanced guidance to the

requirements analyst in making informed decisions about requirements allocation.

Results: We independently evaluate the `Assigner` and `Augmentor` modules on six recent and representative requirement documents. The documents were already annotated with ground truth on 15+ team allocations of the requirements. For `Assigner` module, we considered multiple classification pipelines and achieved the best average recall score of 67% with the BERT uncased language model. We also assessed the domain's vocabulary coverage in BERT-based LLMs and their variations. On average, 65% of the datasets' unique words were out-of-vocabulary (OOV) in pre-trained LLMs. Despite the ability of such LLMs to handle OOV words, their presence impacts the performance of downstream classification tasks. For the `Augmentor` module, we selected the most effective pipeline from the `Assigner` and then evaluated the generated augmentions (additional information). This is essential because the two modules are interdependent; the `Augmentor` relies on the predictions made by the `Assigner` to create its augmentations. On average, in 31% cases, the `Augmentor` outputs the correct team when the `Assigner` incorrectly predicts the team.

Expected timeline: This work has been published in RE [7].

C. Generative LLMs for RE tasks

Problem: LLMs are initially trained on a wide range of open-source datasets. Utilizing transfer learning, such models have demonstrated exceptional performance in downstream tasks like requirements classification on domain-specific datasets [16, 22]. However, LLMs like BERT and its variants are discriminative for classification tasks and cannot provide additional explanations for their deterministic predictions. This is because the architecture and training objectives of BERT LLMs do not inherently support the generation of human-like rationales or explanations. Such rationales are crucial for the adaptability of LLMs in an industrial context, as they enhance the trustworthiness of LLM predictions. As a result, this allows a more efficient human-in-the-loop approach for decision-making and sanitizing the final outputs [9]. In this regard, the generative LLMs, including GPT-3 [11] and LLaMA [30], can address the challenges of discriminative LLMs for requirements classification tasks. Generative LLMs can generate classification decisions and human-like rationales based on the prompting approach [35]. While state-of-the-art generative LLMs exhibit capabilities in various software evolution and maintenance tasks, such as extracting relevant information from bug reports [34] and generating code [25], their application and evaluation in the RE domain have not been explored.

Proposed Approach: We plan to perform rationale-augmented learning with generative LLMs for requirements classification tasks in industrial settings. Specifically, the tasks are based on two industrial cases: the first involves identifying the severity of GitHub issues and the second entails allocating high-level system requirements to multiple teams (similar to Case Study B). In this regard, we plan to devise an approach based on a chain-of-thoughts prompting [32] with in-context learning to support the classification tasks. The chain-of-thoughts prompting—a series of intermediate reasoning steps—provides an efficient way to generate the explanation behind the classification of requirements. The method leverages the generative LLMs' ability to learn from a few examples within the prompt to generate natural language rationales for the given task effectively. The initial aim is to conduct a comprehensive ablation study to assess the effectiveness of the different components in the chain-of-thoughts prompt for generative LLMs. Furthermore, we will conduct a large-scale industrial evaluation to assess the ability of the proposed approach. The evaluation will include the classification decisions and the supporting rationales. The plan also includes empirically evaluating the requirements classification decisions of both discriminative and generative LLMs.

Expected Timeline: This part of the work will be completed in the next 1.5 years.

4 CONCLUSION

RE is a widely recognized discipline marked by well-established concepts to deliver highly reliable industrial systems. However, the ongoing technological advancements and growing system complexities demand additional support in the RE process to develop large-scale systems efficiently. In this regard, researchers have focused on applying AI techniques to automate various tasks in the RE domain. Nevertheless, such solutions often encounter challenges in adaptability within industrial contexts. This is because most existing approaches are evaluated on public datasets and fail to capture the complexities inherent in the industrial domain. Moreover, most of the current AI solutions, while useful, do not offer supporting explanations or rationales for the RE tasks necessary for well-informed decision-making in the industry. This Ph.D. work delves into these challenges and delivers AI-tailored solutions that are developed to augment multiple RE activities. The progress in ongoing research and plans aims to address the identified challenges encountered in the industrial RE process. Specifically, this work proposes AI-based solutions to augment multiple essential RE activities in the studied context, i.e., identifying requirements from tender documents and allocating requirements to different sub-system teams for development. The dissertation outcomes will support the stakeholders working in the requirements-centered context by alleviating their workload. Consequently, this contributes to minimizing project delays and improving resource utilization and time management efficiency while developing large-scale systems.

Future Work. While the current research primarily utilizes deterministic AI approaches, i.e., BERT LLMs, to support the identified industrial use cases, the plan is to expand the exploration by evaluating and integrating generative LLMs. This is because of the generative LLMs' ability to provide additional rationales explaining the decision process for a specific task. This advancement will significantly enhance the practitioners' confidence in the results generated by such solutions. Moreover, it will enable the adoption of AI-based approaches within industrial settings to support the efficient development of large-scale systems.

Acknowledgement. This work is partially funded by the AIDOaRt (KDT) and SmartDelta [28] (ITEA) projects. The author extends gratitude to his advisors and the team at Alstom for their unwavering support.

REFERENCES

[1] Muhammad Abbas, Alessio Ferrari, Anas Shatnawi, Eduard Enoiu, Mehrdad Saadatmand, and Daniel Sundmark. 2023. On the relationship between similar requirements and similar software: A case study in the railway domain. *Requirements Engineering* 28, 1 (2023), 23–47.

[2] Muhammad Abbas, Mehrdad Saadatmand, Eduard Enoiu, Daniel Sundamark, and Claes Lindskog. 2020. Automated reuse recommendation of product line assets based on natural language requirements. In *International Conference on Software and Software Reuse*. Springer, 173–189.

[3] Sallam Abualhaija, Chetan Arora, Mehrdad Sabetzadeh, Lionel C Briand, and Michael Traynor. 2020. Automated demarcation of requirements in textual specifications: a machine learning-based approach. *Empirical Software Engineering* 25 (2020), 5454–5497.

[4] Ethem Utku Aktas and Cemal Yilmaz. 2020. Automated issue assignment: results and insights from an industrial case. *Empirical Software Engineering* 25, 5 (2020), 3544–3589.

[5] Bruno M Arnaut, Denise B Ferrari, and Marcelo Lopes de Oliveira e Souza. 2016. A requirements engineering and management process in concept phase of complex systems. In *2016 IEEE International Symposium on Systems Engineering (ISSE)*. IEEE, 1–6.

[6] Chetan Arora, John Grundy, and Mohamed Abdelrazek. 2023. Advancing Requirements Engineering through Generative AI: Assessing the Role of LLMs. *arXiv preprint arXiv:2310.13976* (2023).

[7] Sarmad Bashir, Muhammad Abbas, Alessio Ferrari, Mehrdad Saadatmand, and Pernilla Lindberg. 2023. Requirements Classification for Smart Allocation: A Case Study in the Railway Industry. In *2023 IEEE 31st International Requirements Engineering Conference (RE)*. IEEE, 201–211.

[8] Sarmad Bashir, Muhammad Abbas, Mehrdad Saadatmand, Eduard Paul Enoiu, Markus Bohlin, and Pernilla Lindberg. 2023. Requirement or not, that is the question: A case from the railway industry. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 105–121.

[9] Daniel M Berry. 2021. Empirical evaluation of tools for hairy requirements engineering tasks. *Empirical Software Engineering* 26, 6 (2021), 111.

[10] Rosanna L Breen. 2006. A practical guide to focus-group research. *Journal of geography in higher education* 30, 3 (2006), 463–475.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[12] Agustin Casamayor, Daniela Godoy, and Marcelo Campo. 2012. Functional grouping of natural language requirements for assistance in architectural software design. *Knowledge-Based Systems* 30 (2012), 78–86.

[13] Raffaele Cioffi, Marta Travaglioni, Giuseppina Piscitelli, Antonella Petrillo, and Fabio De Felice. 2020. Artificial intelligence and machine learning applications in smart production: Progress, trends, and directions. *Sustainability* 12, 2 (2020), 492.

[14] Jane Cleland-Huang, Sepideh Mazrouee, Huang Liguo, and Dan Port. 2007. NFR. https://doi.org/10.5281/zenodo.268542

[15] Jane Cleland-Huang, Michael Vierhauser, and Sean Bayley. 2018. Dronology: An incubator for cyber-physical system research. *arXiv preprint arXiv:1804.02423* (2018).

[16] Gouri Deshpande, Behnaz Sheikhi, Saipreetham Chakka, Dylan Lachou Zotegouon, Mohammad Navid Masahati, and Guenther Ruhe. 2021. Is bert the new silver bullet?-an empirical investigation of requirements dependency classification. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 136–145.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[18] Andreas Falkner, Cristina Palomares, Xavier Franch, Gottfried Schenner, Pablo Aznar, and Alexander Schoerghuber. 2019. Identifying requirements in requests for proposal: A research preview. In *Requirements Engineering: Foundation for Software Quality: 25th International Working Conference, REFSQ 2019, Essen, Germany, March 18–21, 2019, Proceedings 25*. Springer, 176–182.

[19] D Méndez Fernández, Stefan Wagner, Marcos Kalinowski, Michael Felderer, Priscilla Mafra, Antonio Vetrò, Tayana Conte, M-T Christiansson, Des Greer, Casper Lassenius, et al. 2017. Naming the pain in requirements engineering: Contemporary problems, causes, and effects in practice. *Empirical software engineering* 22 (2017), 2298–2338.

[20] Alessio Ferrari and Andrea Esuli. 2019. An NLP approach for cross-domain ambiguity detection in requirements engineering. *Automated Software Engineering* 26, 3 (2019), 559–598.

[21] Alessio Ferrari, Gloria Gori, Benedetta Rosadini, Iacopo Trotta, Stefano Bacherini, Alessandro Fantechi, and Stefania Gnesi. 2018. Detecting requirements defects with NLP patterns: an industrial experience in the railway domain. *Empirical Software Engineering* 23 (2018), 3684–3733.

[22] Tobias Hey, Jan Keim, Anne Koziolek, and Walter F Tichy. 2020. Norbert: Transfer learning for requirements classification. In *2020 IEEE 28th International Requirements Engineering Conference (RE)*. IEEE, 169–179.

[23] Zijad Kurtanović and Walid Maalej. 2017. Automatically classifying functional and non-functional requirements using supervised machine learning. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*. Ieee, 490–495.

[24] Seemab Latif, Sarmad Bashir, Mir Muntasar Ali Agha, and Rabia Latif. 2020. Backward-Forward Sequence Generative Network for Multiple Lexical Constraints. In *Artificial Intelligence Applications and Innovations: 16th IFIP WG 12.5 International Conference, AIAI 2020, Neos Marmaras, Greece, June 5–7, 2020, Proceedings, Part II 16*. Springer, 39–50.

[25] Jia Li, Ge Li, Yongmin Li, and Zhi Jin. 2023. Enabling Programming Thinking in Large Language Models Toward Code Generation. *arXiv preprint:2305.06599* (2023).

[26] Anas Mahmoud and Grant Williams. 2016. Detecting, classifying, and tracing non-functional software requirements. *Requirements Engineering* 21 (2016), 357–381.

[27] Per Runeson and Martin Höst. 2009. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering* 14 (2009), 131–164.

[28] Mehrdad Saadatmand, Muhammad Abbas, Eduard Paul Enoiu, Bernd-Holger Schlingloff, Wasif Afzal, Benedikt Dornauer, and Michael Felderer. 2023. SmartDelta project: Automated quality assurance and optimization across product versions and variants. *Microprocessors and Microsystems* 103 (2023), 104967.

[29] Richard W Selby. 2005. Enabling reuse-based software development of large-scale systems. *IEEE Transactions on Software Engineering* 31, 6 (2005), 495–510.

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[31] Wentao Wang, Nan Niu, Hui Liu, and Zhendong Niu. 2018. Enhancing automated requirements traceability by resolving polysemy. In *2018 IEEE 26th International Requirements Engineering Conference (RE)*. IEEE, 40–51.

[32] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[33] Jonas Winkler and Andreas Vogelsang. 2016. Automatic classification of requirements based on convolutional neural networks. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*. IEEE, 39–45.

[34] Ting Zhang, Ivana Clairine Irsan, Ferdian Thung, and David Lo. 2023. Cupid: Leveraging ChatGPT for More Accurate Duplicate Bug Report Detection. *arXiv preprint arXiv:2308.10022* (2023).

[35] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2023. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029* (2023).

[36] Liping Zhao, Waad Alhoshan, Alessio Ferrari, Keletso J Letsholo, Muideen A Ajagbe, Erol-Valeriu Chioasca, and Riza T Batista-Navarro. 2021. Natural language processing for requirements engineering: A systematic mapping study. *ACM Computing Surveys (CSUR)* 54, 3 (2021), 1–41.