Thesis for The Degree of Licentiate of Engineering

Enhancing Requirements Engineering Practices Using Large Language Models

Krishna Ronanki

Division of Interaction Design & Software Engineering
Department of Computer Science & Engineering
Chalmers University of Technology and Gothenburg University
Gothenburg, Sweden, 2024

“na karmaṇāmanārambhānnaiṣkarmyaṃ puruṣo’śnute
na ca saṃnyasanādeva siddhiṃ samadhigacchat”
- Bhaghavadh Geetha (3:4)

Abstract

Background: Large Language Models (LLMs) offer users natural language interaction, technical insights and task automation capabilities. However, the systematic integration of LLMs within Requirements Engineering (RE) processes presents unique challenges and limitations. A need to develop and understand mechanisms to integrate LLMs in an efficient and responsible manner has been observed in current state-of-the-art literature. This requires scientific inquiry on three key LLM aspects: 1. Technical proficiency in assisting with RE tasks and processes, 2. Ethical and regulatory constraints on their usage, and 3. Artefacts and processes that enable the systematic integration of LLMs in an efficient and responsible manner within organisations.

Objective: This thesis investigates the technical capabilities and ethical/ regulatory constraints to address aspects 1 and 2 before collecting preliminary evidence that motivates further researcher on aspect 3 to enable the systematic integration of LLMs to assist users within RE processes.

Method: A multi-methodology approach, combining quasi-experiments, interviews, a survey and a case study was employed to gather empirical data on LLMs' technical abilities and user experiences in assisting users within RE tasks and processes. A tertiary review followed by a meta-analysis of the literature on ethical AI guidelines and frameworks was conducted to identify and understand the constraints involved in the using LLMs for RE tasks and processes in practice.

Findings: The results of the empirical experiments revealed that LLMs are capable of performing technical tasks like generating requirements, evaluating the quality of user stories, binary requirements classification and tracing interdependant requirements with varying levels of performance. The comparative analysis of ethical AI guidelines and frameworks revealed the constraints and requirements involved in the use of LLMs in practice. The industrial case study and the survey resulted in ten recommendations for the responsible and systematic integration of LLMs for RE in practice.

Conclusion: In conclusion, a need for a human-LLM task delegation framework has been observed, the importance of validating LLMs in real-world-like environments has been highlighted, the under-emphasised human and organisational aspects have been brought to the forefront. Future work needs to delve deeper into identifying and mitigating the challenges associated with the adoption and integration of LLMs. This includes the need to study the organisational barriers, deciding factors, and artefacts that influence and enable the responsible and systematic adoption of LLMs.

Keywords Large Language Models, Requirements Engineering, Prompt Engineering, Trustworthy AI, Multi-methodology

Acknowledgment

First and foremost, I extend my deepest gratitude to my supervisor, Professor Dr. Christian Berger, for his unwavering support and motivation throughout my journey. His immense patience and the wealth of knowledge he imparted have been invaluable, shaping my academic experience in profound ways. Without his guidance, this thesis would not have been the same.

I am also indebted to my co-supervisors, Dr. Jennifer Horkoff and Dr. Beatriz Cabrero-Daniel, whose kind support, insightful feedback, and invaluable advice have played a crucial role in shaping my research acumen. Working with them has been a privilege, and I am grateful for their mentorship.

I extend my appreciation to my examiner, Professor Dr. Alejandro Russo, the CSE PhD admin team, as well as the other members of the research school, for their constructive feedback and administrative assistance throughout this process.

I extend special appreciation to Wardah, Ricardo, Ranim, Dr. Hamdy, Malsha, Razan, Dr. Mazen, Dr. Sushant, Cristy, Habib, Dr. Khaled, and all my colleagues turned friends at the IDSE division. I deeply appreciate your presence in celebrating my tiny successes and your unwavering support during challenging moments.

Lastly, my deepest gratitude goes to my parents, Padmavathi Ronanki and Thirumala Rao Ronanki, whose unwavering love and support have been my anchor throughout my life. Their encouragement and sacrifices have been instrumental in my academic achievements, and I am forever grateful for their belief in me.

This research was supported by the Vinnova project ASPECT.

List of Publications

Appended publications

This thesis is based on the following publications:

[A] K. Ronanki, C. Berger, J. Horkoff.
"Investigating ChatGPT's Potential to Assist in Requirements Elicitation Processes"
Proceedings of the 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2023.

[B] K. Ronanki, B. Cabrero-Daniel, C. Berger.
"ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box?"
Proceedings of the Agile Processes in Software Engineering and Extreme Programming (XP'23) – Workshops, 2023.

[C] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, C. Berger.
"Requirements Engineering using Generative AI: Prompts and Prompting Patterns"
Proceedings of the Effective Software Development with Generative Artificial Intelligence, 2023.

[D] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, C. Berger.
"RE-centric Recommendations for the Development of Trustworthy (er) Autonomous Systems"
Proceedings of the First International Symposium on Trustworthy Autonomous Systems (TAS'23), 2023.

[E] K. Ronanki, B. Cabrero-Daniel, J. Horkoff, C. Berger.
"Practitioners' User Experiences of LLM-assisted RE: An Industrial Case Study"
Under Submission at the Journal of Information and Software Technology (IST).

Other publications

The following publications were published during my PhD studies. However, they are not appended to this thesis, due to contents overlapping that of appended publications or contents not related to the thesis.

[a] K. Ronanki.
"Towards an AI-centric Requirements Engineering Framework for Trustworthy AI"
Proceedings of the IEEE/ACM 45th International Conference on Software Engineering: Companion Proceedings (ICSE-Companion), 2023.

[b] K. Ronanki, B. Cabrero-Daniel, C. Berger.
"Prompt Smells: An Omen for Undesirable Generative AI Outputs"
Proceedings of the Third International Conference on AI Engineering — Software Engineering for AI (CAIN 2024), 2024.

Research Contribution

I bear the primary responsibility for the planning, design, execution, and composition of this thesis. I served as the primary contributor in the planning, co-designing, data gathering, data analysis, and authoring of Paper A, Paper B, Paper C, Paper D and Paper E.

Contents

Abstract iii

Acknowledgement v

List of Publications vii

Personal Contribution ix

1 Introduction 1
- 1.1 Research Questions . . . 2
- 1.2 Scope of the Thesis . . . 3
- 1.3 Background . . . 4
  - 1.3.1 Requirements Engineering . . . 4
  - 1.3.2 Artificial Intelligence for Requirements Engineering . . . 4
  - 1.3.3 Large Language Models . . . 5
    - 1.3.3.1 Retrieval Augmented Generation . . . 5
    - 1.3.3.2 Temperature . . . 6
  - 1.3.4 Prompt Engineering . . . 6
  - 1.3.5 Trustworthy AI . . . 7
- 1.4 Research Methodology . . . 8
  - 1.4.1 Philosophical Paradigms . . . 8
  - 1.4.2 Research Paradigms . . . 9
  - 1.4.3 Experiments . . . 10
  - 1.4.4 Tertiary Review . . . 11
  - 1.4.5 Case Study . . . 11
  - 1.4.6 Other Data Collection Techniques . . . 12
    - 1.4.6.1 Interviews . . . 12
    - 1.4.6.2 Surveys . . . 12
- 1.5 Summary of Findings . . . 12
  - 1.5.1 Technical Proficiency of LLMs in Assisting Users with RE Tasks and Processes . . . 13
  - 1.5.2 Ethical and Regulatory Constraints on the Use of LLMs 15
  - 1.5.3 LLM's Adoption in Industrial Practice . . . 16
- 1.6 Threats to Validity . . . 17
- 1.7 Discussion . . . 18
- 1.8 Conclusion and Future Work . . . 20

Bibliography 23

Chapter 1

Introduction

In the field of Requirements Engineering (RE), there has been a growing interest in the application of Artificial Intelligence (AI) techniques, particularly Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL). These methods have been employed for a variety of RE tasks and processes such as classification of requirements, prioritisation, tracing, detection of ambiguity, and modelling [1, 2].

It has been empirically observed that natural language is the most commonly used medium for drafting requirements in the industrial context. This significant synergy between natural language and requirements has led to the emergence of NLP in the field of RE [2].

Natural Language Processing for Requirements Engineering (NLP4RE) is an approach that utilises NLP tools and methodologies in RE processes to aid human analysts in performing various tasks on textual requirements [2]. Natural Language Generation (NLG), one of the most critical yet complex sub-fields of NLP [3], is a technique employed to generate coherent phrases, sentences, and paragraphs in a natural human language [4].

Despite the advancements in NLP4RE, there is a noticeable lack in the application of NLG-based method in support of RE tasks [2]. Moreover, the majority of existing state-of-the-art approaches for supporting RE tasks and processes rely on ML/DL. These methods necessitate substantial amounts of task-specific labelled training data. However, the scarcity of open-source, RE-specific labelled data presents a challenge for RE researchers and practitioners in developing, training, and testing advanced ML/DL models [1] for their application in RE tasks and processes.

One potential solution to overcome the need of large quantities of high quality labelled training data is using pre-trained generative AI models like Large Language Models (LLM). Utilising LLMs for performing RE tasks can potentially remove the need for large amounts of labelled data.

Recent LLMs, despite being trained on a broad general domain data corpus [5], are demonstrating increasingly impressive capabilities when performing a wide range of tasks [6], including their significant enhancements in NLP tasks [7]. The GPT-3.5 LLM has demonstrated a surprising proficiency in specific technical tasks [8]. LLMs are also observed to increase developer productivity [9] when used for software engineering (SE) tasks.

In light of these observations, a working hypothesis was formulated to investigate the potential application of LLMs in the field of RE.

Working Hypothesis: The inherent ability of LLMs to generate natural language text and their emergent capabilities can be leveraged to assist users in performing RE tasks and processes.

Research Gap:

However, the realisation of this working hypothesis introduces several challenges. Firstly, a comprehensive understanding of LLMs, especially their competences, is yet to be achieved. This lack of understanding is further exacerbated by the ambiguity surrounding their suitability and the extent of their applicability within complex activities like RE. Therefore, further investigation into LLMs' performance within RE tasks is required.

Secondly, the use of LLMs is governed by strict regulations, such as the EU AI Act. These regulations aim at ensuring the safe and ethical use of AI systems, including LLMs. There is a need to understand the effects and impact these regulations have on the use of LLMs.

Lastly, the integration of LLMs into RE workflows is not a simple process. It requires the development of systematic mechanisms that allow for responsible adoption of the LLMs which adhere to the relevant and applicable regulatory constraints. This involves considering technical, ethical, and organisational implications.

Research Goal:

These challenges, while significant, motivate further research to facilitate the responsible and systematic adoption of LLMs, especially within industrial practice. So, in order to find reliable evidence supporting the working hypothesis, these three key aspects need to be addressed which were formulated as the primary research goals of this thesis. The three research goals outlined below serve as the fundamental basis for fulfilling the objective of this thesis.

- **Goal-1:** Investigate the technical proficiency of LLMs in assisting users with RE tasks and processes.
- **Goal-2:** Explore the ethical and regulatory constraints on the use of LLMs within industrial practice.
- **Goal-3:** Identify and synthesise artefacts and processes that enable the adoption of LLMs within organisations.

The primary objective of this thesis is to provide evidence and insights that address **Goal-1** and **Goal-2**, and some preliminary evidence that warrants further investigation to address **Goal-3**.

1.1 Research Questions

Formulating clear research questions is essential to guide this thesis towards fulfilling its intended objective. To that end, the following research questions were formulated to achieve the specified research goals:

RQ1 To what extent can LLMs support users in the specific RE tasks and processes?

- **RQ1.1** What is the quality of LLM-generated requirements in comparison to human RE expert-formulated requirements?
- **RQ1.2** What is the LLM's performance in user story quality evaluations?
- **RQ1.3** What is the LLM's performance in binary requirements classification?
- **RQ1.4** What is the LLM's performance in identifying inter-dependant requirements?

RQ2 What are the observed effects of the applied prompt engineering techniques on the LLM-generated outputs?

RQ3 What are the key ethical principles to consider when applying LLMs to RE tasks?

RQ4 What are the recommendations for practitioners and organisations concerning the responsible integration of LLMs in practice?

1.2 Scope of the Thesis

This section specifies the boundaries within which the this research operates, offering a clear context for the RQs posed above. By outlining the specific parameters and limitations of the thesis, this section ensures readers accurately gauge the applicability and generalisability of the findings presented within Section 1.5.

Within RQ1, five RE tasks and processes were examined. RQ1.1 focuses on requirements elicitation, comparing the requirements generated by ChatGPT (GPT-3.5) with those formulated by human RE experts for a generic AI system that must comply with the EU AI Act's trustworthiness criteria. RQ1.2 involves the quality evaluation of a set of user stories using ChatGPT (GPT-3.5), sourced from an open-source database presented in Tõemets [10]. RQ1.3 and RQ1.4 evaluate the performance of the LLM (GPT-3.5 accessed via API) in performing binary requirements classification and requirements traceability on the PROMISE [11] and PURE [12] datasets respectively.

For RQ2, the prompt engineering techniques taken into consideration were one-shot prompting, a variation of the few-shot prompting [13] and five prompt patterns out of the sixteen presented by White et al. [14]. The LLM-generated outputs were taken from the same tasks as considered in RQ1.2, RQ1.3 and RQ1.4.

The EU AI Act was still in the draft stage as of February 2023, which is when the results to answer RQ1.1 and RQ3 were being collected. Since then, the EU AI Act was updated, addressing of the use of general-purpose AI models, and ratified as law by the European Council [15].

For RQ4, the RE tasks and processes taken into consideration were generating user stories and stakeholder analysis documentation. All the results presented, analysed and discussed within this thesis fall under the scope of the above mention parameters.

1.3 Background

This section outlines the essential concepts from relevant literature that form the foundation of this thesis. It offers a detailed overview of the core topics relevant to this study.

1.3.1 Requirements Engineering

RE is essential in software development, ensuring that the final product meets stakeholders' needs and expectations while delivering value [16]. However, it is often prone to challenges such as gathering accurate inputs, analysing ambiguous and conflicting information, and integrating RE processes with broader software development activities [16, 17]. To overcome these challenges and limitations, RE has undergone several transitions over the years, from traditional methods to agile approaches emphasising increased automation [18].

Traditional RE focuses on comprehensive documentation and upfront requirements gathering, which can provide a clear roadmap but bears the potential risk of misalignment of stakeholder expectations including potentially superfluous requirements. On the other hand, Agile RE prioritises flexibility and speed. It employs an iterative, incremental framework that enables rapid product delivery, fosters constant communication between customers and developers, and allows continuous refinement of requirements and timely delivery of systems that meet customer expectations, thereby enhancing business value [19]. To enhance the efficiency of such agile development practices, labour-intensive tasks are usually automated as a practical means of streamlining processes [20].

1.3.2 Artificial Intelligence for Requirements Engineering

With the rapid advancements in AI, various AI techniques are increasingly employed to enhance the efficiency of SE tasks, including RE. This fusion of AI and RE, known as AI for RE (AI4RE), has garnered significant attention from researchers to address the challenges associated with RE practices [21].

Zhao et al. [2] conducted a systematic mapping study that surveys the landscape of NLP4RE research to provide a holistic understanding of the field. The mapping study not only provides a collection of the literature in NLP4RE but also propose conceptual framework of NLP4RE that encompasses both RE and NLP aspects through categorisation, synthesis and conceptualisation of the main theoretical concepts. This highlights the synergy between the fields of NLP and RE and provides an in-depth analysis of the existing gaps and the evolution of NLP4RE over the years.

Mehraj et al. [21] conducted a tertiary review to understand how AI assists RE practices. They analysed 28 secondary studies from 2017 to September 2023 about using AI in RE tasks such as elicitation, classification, analysis, specification, management, and tracing. Their results reveal a trend of combining NLP techniques with ML models and a surge in using LLMs for RE. The study also identified challenges of AI4RE related to ambiguity, language, data, and evaluation. These findings strengthen the working hypothesis of this thesis further.

Zhang et al. [22] empirically evaluate how ChatGPT performs on requirements analysis tasks to derive insights into how generative LLMs, influence the

research and practice of NLP4RE. In further work, Zhang et al. [23] empirically evaluate ChatGPT's performance on requirements information retrieval (IR) tasks. They posit their evaluation provides preliminary evidence for designing and developing more effective requirements IR methods based on LLMs.

These studies provide an overview on the current research landscape and helped in designing the experiments relevant to addressing **Goal-1**.

1.3.3 Large Language Models

LLMs are AI models, often built upon deep learning techniques like transformer architectures, that can produce useful natural language outputs [24]. This led to them being employed in various language-related tasks such as text generation [25], question answering [26], translation [6], summarisation [27], and sentiment analysis [28]. Figure 1.1 depicts an abstracted version of a transformer model architecture presented by Vaswani et al. [29].

The primary components of a transformer model architecture include an encoder and a decoder. Both the encoder and decoder have sub-components like self-attention mechanism, multi-head attention, feed-forward network and layer normalisation [30]. These sub-components have been abstracted within Figure 1.1 from the original transformer model architecture presented by Vaswani et al. [29]. The encoder is responsible for transforming the input text (a sequence of tokens) into a sequence of continuous vectorised representations, while the decoder's function is to convert this vectorised representations back into output tokens [29,30]. A positional encoding is added within the transformer to the input embeddings to provide information about the position of each token in the sequence. The decoder output is then converted into output probabilities using the learned linear transformation and softmax function.

LLMs achieve their impressive performance partly due to being trained on extensive datasets, built upon this architecture. For instance, GPT-3 is trained on 45TB of text data and boasts 175 billion parameters [31]. In February 2023, Meta launched the open-sourced LLaMA, trained on 1.4 trillion tokens across various model sizes ranging from 7 billion to 65 billion parameters [32]. The third generation of this model, LLaMA-3, made available in April 2024, is trained on 15 trillion tokens for the 8 billion and 70 billion parameter variants [33].

1.3.3.1 Retrieval Augmented Generation

While LLMs have achieved remarkable success, they still grapple with substantial challenges. These are particularly evident in tasks that are specialised or require extensive knowledge [34]. Notably, they tend to generate misleading or false information termed as "hallucinations" [35] when dealing with inquiries that exceed their training data or necessitate up-to-date information.

To address this limitation, recent research has explored the integration of both parametric and non-parametric memory into LLMs. Parametric memory refers to the learned parameters within the model, while non-parametric memory typically involves accessing external knowledge sources such as large text corpora or databases. One approach to combining these types of memory is through a method called Retrieval Augmented Generation (RAG). RAG endows pre-trained parametric-memory generation models, such as transformers, with

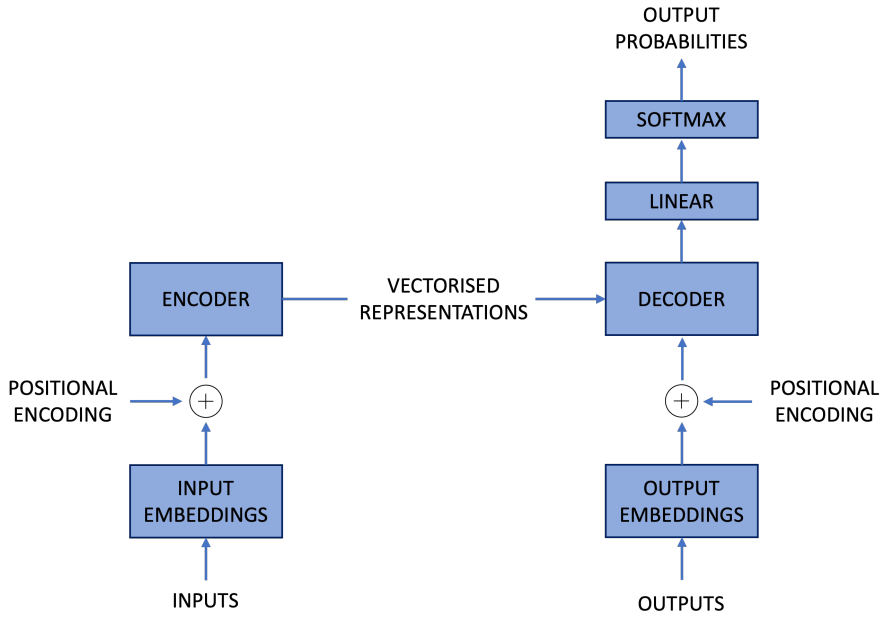

Figure 1.1: Schematic Architecture of a Transformer Model

non-parametric memory by incorporating a dense vector index of external knowledge. This integration is achieved through a general-purpose fine-tuning approach, allowing the model to access extensive knowledge during inference without additional training [36].

1.3.3.2 Temperature

Temperature of a LLM is a parameter that controls the randomness of the model's output. When adjusting the temperature setting of a LLM, which ranges from 0.0 to 1.0, one is essentially controlling the randomness of the model's responses. Higher temperatures (e.g., 0.8 or 1.0) result in more diverse and "creative" responses, while lower temperatures (e.g., 0.2 or 0.5) produce more focused and deterministic responses [37–40].

1.3.4 Prompt Engineering

The natural language text inputted within a prompt to interact with the LLMs has been observed to have a significant effect on the downstream task performance of the LLM model [6]. The emerging practice of using carefully selected and composed sentences to achieve the desired output is called prompt engineering [41]. Prompt engineering allows the model's users to express their intent in plain language, rather than a specially designed programming language [42].

Prompt engineering is still more of an art than science. It is a rapidly evolving field of study that involves a high degree of creativity, intuition, and iterative refinement [43]. This raises concerns about the long-term viability of

prompt engineering evaluation methods. Current evaluation methods involve continuous experimentation with these rapidly evolving prompting techniques to assess their relative performance within a corpus of techniques. But these comparative evaluation methods do not ensure the replicability of such experimental results in complex real-world scenarios. As the model's output is directly influenced by the quality of the prompts, the potential implications of low maturity in prompt engineering could prevent LLMs from reaching their full potential or result in unexpected and biased responses [6, 44].

However, a few techniques have been empirically validated in experimental settings to improve the LLMs performance. These techniques include zero-shot prompting, few-shot prompting and chain-of-thought prompting techniques.

In its most basic form, zero-shot prompting involves using natural language sentences to convey the "problem to be solved" or the "expected output", without providing any examples [42]. It is a technique modelled after zero-shot learning, which directly applies previously trained models to predicting both seen and unseen classes without using any labelled training instances [45].

Few-shot prompting builds on zero-shot prompting by conveying the "problem to be solved" or the "expected output" using a few demonstrations of the task (examples) at inference time as conditioning [13].

Chain-of-thought prompting is a technique that is observed to enhance the reasoning capabilities of a LLM. Using this technique, the user constructs the prompts in a way that lets the model generate a coherent series of intermediate reasoning steps that leads to the final answer for the task at hand [46].

Going beyond exemplar prompting techniques, White et al. [14] present prompt design techniques for software engineering in the form of patterns to automate common software engineering activities using LLMs. Prompt patterns serve as a means of knowledge transfer, similar to software patterns, by offering reusable solutions to common problems related to generating output and engaging with LLMs. They establish a framework for documenting prompt patterns that can be adapted to various domains, providing a systematic approach for structuring prompts to tackle a range of issues. The five prompt patterns selected and applied within studies addressing RQ1.3, RQ1.4 and RQ2, the **Cognitive Verifier**, **Context Manager**, **Persona**, **Question Refinement** and **Template** patterns were proposed in this study.

1.3.5 Trustworthy AI

Due to enhanced access to large volumes of data generated all the over the world, we observe an accelerated use of AI in our daily lives [47, 48]. As the adoption of AI across various fields, such as healthcare, transportation, and education, is growing, the need for AI to be trustworthy is also increasing every day.

Trustworthy AI can be defined as a conceptual framework that ensures that the development and implementation of technically and socially robust AI systems adhere to all the applicable laws and regulations, and conform to general ethical principles [49]. The concept of trustworthy AI can foster the realisation of AI's full potential by establishing trust in the development and implementation of AI to maximise its benefits and reduce its risks at the same time [50].

Since the development and use of general purpose AI models like LLMs is strictly regulated under the EU AI Act, it is therefore more crucial than ever for AI systems and AI-based software to be not only technically robust but also compliant with relevant development guidelines and regulations [48].

1.4 Research Methodology

In this section, an overview of the employed research design is provided, followed by detailed explanations of the research methods for each of the appended publications and how they contributed to addressing the research goals by helping to answer the relevant research questions. Figure 1.2 is a visual representation of the research design employed for this thesis. It depicts how the elements that constitute this thesis (the working hypothesis, the research gap, the research goals, the research questions, and the appended papers) are connected.

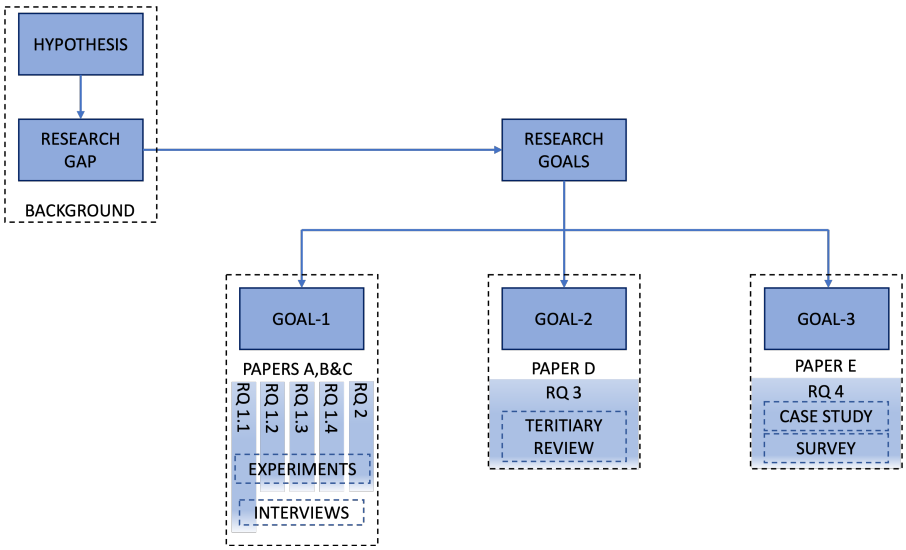

Figure 1.2: Research Design Overview

1.4.1 Philosophical Paradigms

There are four distinguished philosophical paradigms that a researcher's worldview can be classified under: (1) post-positivist, (2) constructivist, (3) advocacy/participatory, and (4) pragmatist [51]. The post-positivist worldview focuses on collecting and presenting evidence that indicates a failure to reject a hypothesis rather than prove the hypothesis. The constructivist worldview, on the other hand, focuses on seeking subjective reality surrounding the human interpretation of the observed phenomenon. The advocacy/participatory worldview is intertwined with political influences that impact the phenomenon under study, while the pragmatist worldview advocates for the use of mixed methods within the scientific inquiry without being limited to any single philosophical

paradigm [52]. The choice of which paradigm to adopt depends on the type of research study being conducted and the research goal that needs to be addressed [53]. Subsequently, the adopted paradigm influences the research methods employed to gather reliable evidence [54]. In empirical software engineering, pragmatism is the dominant philosophical paradigm [52] and this is reflected within this thesis, given that each of the stated research goals were addressed using different worldviews.

Goal-1 focuses on assessing the viability of leveraging LLMs' inherent technical abilities in performing and assisting with various RE tasks and processes. A post-positivist worldview is well suited here to collect reliable evidence that indicates a failure to reject the formulated working hypothesis.

Goal-2 focuses on identifying the regulatory and ethical constraints involved in the use of LLMs. A participatory worldview would be well-suited to understand the impact of various ethical AI guidelines on the use of LLMs within a human-centric process such as RE.

A constructivist worldview, which focuses on exploring phenomenological interpretations of human experiences is a suitable approach in addressing **Goal-3**, which focuses on studying organisational aspects of LLMs' adoption.

The most promising solution to address all three research goals would be to adopt a pragmatist worldview, which enables the use of mixed methods without being limited to a particular philosophical paradigm. According to the ACM SIGSOFT empirical standards for software engineering research, a study that uses two or more approaches to collect and/or analyse data to corroborate, complement and expand research findings [55] is called a multi-methodology study. Following the pragmatist worldview, this thesis employs four different research methods across five research articles to address the three research goals.

In this multi-methodology study, the first step was to formalise the working hypothesis presented in Section 1 and identify the research gap by screening the relevant background literature. The research goals and questions, presented in Section 1 and 1.1 respectively, were synthesised to address the research gap and collect reliable evidence that makes rejecting the formalised working hypothesis difficult.

1.4.2 Research Paradigms

Software engineering research is acknowledged as a fundamentally empirical discipline [56]. In empirical studies, there are two types of research paradigms, each with different approaches. *Exploratory research*, also referred to as *qualitative research*, is concerned with studying objects in their natural setting and letting the findings emerge from the observations. This implies that a flexible research design [57] is needed to adapt to changes in the observed phenomenon. *Explanatory research*, also referred to as *quantitative research*, is primarily concerned with collecting quantitative data to identify a cause-effect relationship between the object(s) under study. This required a fixed research design [57], often conducted through setting up controlled experiments [58].

1.4.3 Experiments

An experiment in software engineering is an empirical investigation involving the manipulation of one or more potentially influential factor(s), or variable(s), of the studied setting. If humans were to apply various treatments to the object(s) under study, they would be human-oriented experiments. Quasi-experiments are a variation of experiments where the choice of applied treatments is based on the characteristics of either the subjects or objects involved rather than randomisation [58]. To address **Goal-1** of this thesis, several experiments were conducted, with a post-positivist worldview. The subject within these experiments was the LLM. The datasets used in these experimental studies served as the objects of investigation, while the RE tasks performed by the LLM were the treatments applied.

Paper A employed a two-stage experimental setup where the first stage involved using a LLM to generate requirements. The second stage was to conduct two rounds of interview-based surveys, each with different objectives. The objective of this study is to compare the quality of the requirements generated by the LLM with the quality of the requirements formulated by human RE experts, collected through interviews. A different set of interviews was conducted to evaluate the quality of these requirements using a different set of domain experts. This mixed-methods approach [59] to collect data enabled acquiring quantified measures of the quality of requirements, allowing for a direct comparison between human-formulated and LLM-generated requirements. The results of the second round of interviews with expert evaluators and the following comparison helped in answering **RQ1.1**.

In **Paper B**, an experimental setup involving the use of a LLM to perform user story quality evaluation was designed. To start the experiment, a set of user stories was randomly selected from a dataset and their quality was evaluated manually. The human evaluations served as the ground truth against which the LLM's performance was measured. Next, the LLM was prompted to perform the quality evaluation on the same set of user stories using the same criteria as human evaluations. Finally, the data from the human evaluations and the LLM evaluations were compared by computing the overall precision, recall, and F-score [60], which helped in measuring the LLM's robustness and consistency, answering **RQ1.2**. Since the "best of three" approach while prompting was compared with two other approaches, this study also helped in partially answering **RQ2**.

In **Paper C**, the experimental setup involved selecting five prompt patterns to craft input prompts that enable the automation of two RE tasks: (i) binary classification of functional requirements and non-functional requirements, and, (ii) requirements traceability. The study was initiated by conducting a series of controlled experiments using a LLM to automate the selected RE tasks. Each experiment involved crafting an input prompt using one of the five selected prompt patterns and then automating both the RE tasks. The performance metrics for these experiments were retrieved and compared with the ground truth derived from the datasets used within the experimental setup. This process was repeated for all five prompt patterns. These performance metrics enabled in identifying which of the selected prompt patterns worked best for each of the RE tasks, answering **RQ1.3**, **RQ1.4** and partially answering **RQ2**.

Paper A, **Paper B** and **Paper C** together helped in addressing **Goal-1** through the conducted quasi-experiments.

1.4.4 Tertiary Review

A tertiary review is a systematic review of systematic review studies conducted to answer wider research questions [61]. It is conducted where a number of systematic reviews of the topic of study already exist. Systematic reviews, and subsequently tertiary reviews, are research methodologies that emerged in the aftermath of the push by Kitchenham et al. [62] for adopting evidence-based software engineering within software engineering research [61]. In this context, evidence is defined as "a synthesis of best quality scientific studies on a specific topic or a research question" [61].

To address **Goal-2**, **Paper D** involved conducting a tertiary review and a meta-analysis on selected ethical AI frameworks to assess the feasibility and identify the challenges in operationalising the principles presented within the ethical AI guidelines. This was accompanied by conducting a comparative analysis of selected ethical AI guidelines (IEEE-EAD [63], UNESCO's AI guidelines [64], OECD AI guidelines [65] and the EU AI Act draft [49, 66]) to identify and understand the regulatory constraints in developing and deploying trustworthy AI systems.

The comparative analysis highlights the constraints on the use of AI, applicable to the use of LLM for RE tasks and processes, answering **RQ3** to address **Goal-2**.

1.4.5 Case Study

The case study is a research method that is well-suited for studying contemporary phenomena that are hard to study in isolation [67]. Since the integration of LLMs to assist RE tasks and processes in an industrial setting is a relatively contemporary phenomenon as of 2024, **Paper E** presents the results of a industrial case study conducted based on the guidelines for conducting and reporting case study research in software engineering by Runeson and Höst [67], with a constructivist worldview. The primary unit of analysis of this case study is a workshop. The objective of the workshop was to enable participants to experiment with the concept of LLM-assisted RE in a practical setting. The case study involved a mixed methods approach [59] to collect data on the user experiences and perceptions surrounding the use of LLMs' assistance in RE tasks and processes through semi-structured interviews [68].

The first round of interviews was conducted before the workshop was held to gather data on the participants' and the case company's expectations on integrating LLMs into their RE processes. The second round of interviews was conducted after the workshop. The data collected from the interviews was analysed using content analysis [69] to generate insights. These insights were used to synthesise ten recommendations for the use and integration of LLMs to assist users in their RE tasks and processes in industrial contexts. These recommendations were later validated through an online survey with software practitioners from various industries, answering **RQ4**.

1.4.6 Other Data Collection Techniques

In order to address the three main objectives of this thesis, the above presented research methods were employed. As part of these methods, interviews and surveys were administered to gather data. The data collected through these methods contributed significantly to the research findings and conclusions.

1.4.6.1 Interviews

Interviews are one of the two most commonly used data collection methods for qualitative data within empirical software engineering [58, 70]. Advantages of interview-based surveys include achieving a higher response compared to e-mail or online surveys, and the decreased number of inconclusive responses such as "do not know" or "no answer" [58]. Software has a significant impact on almost every aspect of society, making ethics in software engineering research essential. When humans are used as data sources, researchers must respect and protect them from harm. To ensure this, all the interviews conducted within the studies contributing to this thesis were done so by following the ethical interview checklist presented by P.E. Strandberg [71].

As mentioned above, **Paper A** involved the use of two rounds of semi-structured interviews [68], each with a different objective. These interviews helped in answering **RQ1.1** and contributed to addressing **Goal-1**.

Paper E, too, involves the use of two rounds of semi-structured interviews [68] as part of a case study. The insights generated by analysing the interview data contributed to synthesising ten recommendations, answering **RQ4**.

1.4.6.2 Surveys

Surveys are designed to provide broad overviews and help develop general conclusions [58]. They have three main purposes: descriptive, explanatory, and explorative [72]. Descriptive and explanatory surveys can apply empirical findings to many projects or organisations, making them useful for large-scale research in real-world projects [58, 73].

To investigate the extent of generalisability of the recommendations synthesised from the case study as part of **Paper E**, an online survey was also administered. The target population of the survey were software practitioners from various industries who have experience using a LLM. The survey was conducted to collect practitioners' agreement with the ten recommendations. The participants were asked to rate your agreement with the presented statement on a scale of 1-5, where 1 is "strongly disagree", 2 is "somewhat disagree", 3 is "neutral", 4 is "somewhat agree" and 5 is "strongly agree".

1.5 Summary of Findings

This section summarises and presents the results and findings of the appended research articles and puts them into the context of answering the thesis' research questions. Subsections 1.5.1-1.5.3 present the thesis' research questions, the findings and contributions that helped in answering the relevant research question and the appended paper the findings were derived from in detail.

1.5.1 Technical Proficiency of LLMs in Assisting Users with RE Tasks and Processes

RQ1- To what extent can LLMs support users in the selected requirements engineering tasks and processes?

This research question is answered using the findings of the four sub-research questions presented below. The answers to these four sub-research questions present evidence collected through the quasi-experiments for four different RE tasks and processes. While positive results for the application of LLMs in four RE tasks/processes do not mean LLMs are useful for all RE tasks, these results make it hard to reject the formalised working hypothesis, as intended.

RQ1.1- What is the quality of LLM-generated requirements in comparison to human RE expert-formulated requirements?

In **Paper A**, the quality of the LLM-generated requirements was systematically evaluated by a set of five RE experts in direct comparison with human RE expert-formulated requirements for the same system. The results show that LLM-generated requirements received higher scores on most of the assessed quality properties, according to expert evaluators, compared to human-formulated requirements. The findings indicate that LLMs can assist requirements analysts in, at least, making requirements more abstract, atomic, consistent, and understandable. But this does not mean they are flawless. Despite the average scores of LLM-generated requirements for unambiguity and feasibility being higher than human-formulated requirements in the majority of instances, generating ambiguous and unfeasible requirements is a clear shortcoming identified in using the LLM to elicit requirements.

RQ1.2- What is the LLM's performance in user story quality evaluations?

In **Paper B**, the consistency and reliability of the LLM in evaluating user stories using a set of predetermined criteria was evaluated. Based on the findings presented in this study, the LLM demonstrated a consistent agreement rate throughout the evaluations. The results indicate that the LLM was consistent with itself in 61% of the evaluations, meaning it gave the same response for a given pair {criteria, user story} in three separate executions 61% of the times. Furthermore, the LLM's assessments were consistent in at least two out of three runs in 83% of the evaluations. In terms of agreement rate with human evaluators, LLM's performance was relatively stable across the three assessments.

RQ1.3- What is the LLM's performance in binary requirements classification?

In **Paper C**, one of the two RE tasks performed using the LLM was binary requirements classification. Based on the findings presented in this study, the precision of LLM in performing the task ranged between 66% and 86% across five prompt patterns and three temperature settings. The highest precision was achieved using the **Context Manager** pattern at a LLM temperature setting of 1.0. This is an interesting observation as the majority of the following higher precision scores were achieved at lower temperature settings. The recall ranged between 49% and 91% across all five prompt patterns and three different LLM temperature settings. The highest recall was achieved using the **Template**

pattern at a LLM temperature setting of 1.0. The accuracy ranged between 69% and 78% across all five prompt patterns and three different LLM temperature settings. The highest accuracy was achieved using the **Cognitive Verifier** pattern at a LLM temperature setting of 0.0.

RQ1.4- What is the LLM's performance in identifying inter-dependant requirements?

In **Paper C**, the other RE task performed using the LLM was identifying inter-dependant requirements. Based on the findings presented in this study, the precision of LLM in performing the task ranged between 26% and 46% across all five prompt patterns and three different LLM temperature settings. The recall ranged between 51% and 77% across all five prompt patterns and three different LLM temperature settings. The accuracy ranged between 85% and 92% across all five prompt patterns and three different LLM temperature settings.

The precision, recall, and F1 scores, which serve as crucial indicators of the model's performance, consistently exhibit higher values in the binary requirements classification task as compared to the tracing of dependent-requirements task. Observations from the presented results underscore the model's ability to discern and classify binary requirements effectively, demonstrating a higher precision in isolating relevant instances, and a better recall in identifying all pertinent cases. A major observed deviation from this is the higher accuracy scores the model achieved for all patterns for the requirements traceability task in comparison to the binary requirements classification task. This rather significant deviation might be the result of the model being more adept at accurately predicting false negatives than identifying true positives, true negatives and false positives. This leads to a scenario where predicting non-existing trace links between two requirements within the SRS documents results in higher accuracy scores. This is why the focus of the analysis was based more on precision, recall and F-score measures than accuracy.

To answer RQ1, the findings from several quasi-experimental studies suggest that LLMs are capable of assisting users in various RE tasks and processes. The performance, however, varies based on the specific task performed and the prompt engineering techniques employed.

RQ2- What are the observed effects of the applied prompt engineering techniques on the LLM-generated outputs?

The 61% consistency of the LLM evaluations in **Paper B** led to the use of a "best of three" approach while prompting inputs to perform the user story quality evaluation using the LLM. In this approach, the LLM is required to give the same response in at least two out of three attempts. This approach resulted in a higher agreement rate between human evaluations and the LLM evaluations.

Based on the analysis of the results presented in **Paper C**, the **Cognitive Verifier** pattern and the **Question Refinement** pattern were found to be better suited for the binary requirements classification task at any temperature setting compared to the **Context Manager**, the **Template** and the **Persona** patterns. Similarly, the **Question Refinement** pattern is the most consistent

and reliable pattern amongst the five for tracing dependent requirements followed by the **Cognitive Verifier** and the **Persona** pattern. The **Context Manager** pattern and the **Template** pattern are the least reliable pattern for this task.

Overall, the **Question Refinement** pattern, shows consistent results across both the classification and requirements traceability tasks. The **Cognitive Verifier** pattern and the **Persona** pattern obtained higher performance scores in binary requirements classification, although their performance in tracing dependent requirements was reduced. The **Context Manager** pattern was found to have a greater degree of variability for both tasks.

To answer RQ2, choosing the right prompt engineering strategy and prompt patterns result in a significant improvement in the performance of the LLM for the chosen task. However, the deciding factors behind choosing the right prompt pattern should be based on continuous experimentation to identify the better suited patterns for the chosen task.

1.5.2 Ethical and Regulatory Constraints on the Use of LLMs

RQ3- What are the key ethical principles to consider when applying LLMs to RE tasks?

As part of **Paper D**, a comparative analysis of various ethical AI guidelines like IEEE-EAD [63], UNESCO's AI guidelines [64], OECD AI guidelines [65] and the EU AI Act (AIA) draft [49, 66] was performed. This comparative analysis involved understanding each of the ethical AI principles that are applicable to the development and deployment of AI systems.

Based on the results of the **Paper D**, the AIA was observed to provide the most comprehensive coverage of the addressed and recommended principles. The European Commission's (EC) High-level expert group on artificial intelligence (AI HLEG) has identified four ethical principles that form the basis for trustworthy AI which include *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability* [49, 66]. To foster trustworthy AI in practice, they believe that these seven key requirements must be continuously assessed and managed throughout the entire lifecycle of an AI system: *human agency and oversight*, *technical robustness and safety*, *privacy and data governance*, *transparency*, *diversity, non-discrimination, and fairness*, *societal and environmental well-being*, and *accountability*.

To answer RQ3, the use of LLM's assistance to help users with their RE tasks and processes involves compliance with the AIA, which in turn implies continuously assessing and managing the seven key requirements presented as long as they are being used. Non-compliance with these seven requirements might incur significant financial and legal repercussions.

1.5.3 LLM's Adoption in Industrial Practice

RQ4- What are the recommendations for practitioners and organisations concerning the responsible integration of LLMs for RE in practice?

Paper E gathered empirical evidence on the efforts to integrate LLMs to assist users in their RE processes within an industrial setting. The insights focused on six aspects surrounding ((i) the potential for LLMs to be used as "AI assistants", (ii) the importance of "user friendliness" of the LLM-based tools, (iii) necessary human oversight and validation mechanisms, (iv) the expected evolution of engineers' role in LLM-assisted processes, (v) the need for restructuring processes to leverage LLMs, and (vi) the prompt engineering practices that leverage LLMs effectively.

Based on these findings, ten recommendations, labelled **R1-R10** (accessible in **??**) were synthesised to help users of the case company with the responsible use and integration of LLMs into RE tasks and processes. Recommendations **R3** to **R7** are targeted for the LLM users, on an individual level. On the other hand, **R1**, **R2**, and **R8** to **R10** are meant to guide the integration and adoption of LLMs, at a team or organisational level.

These ten recommendations were later validated through survey with software practitioners from various industries outside of the case company to examine the extent of their generalisability. The survey results indicate the most of the recommendations synthesised from the case study are generalisable beyond the contexts of the case company, with nine out of the ten recommendations received a favourable response from the survey participants. When the recommendations are ranked based on the percentage of favourable responses (a survey score of 4 or 5) received, organisational level recommendations received more favourable responses from the survey users compared to individual user level recommendations.

A mapping of these ten recommendations to the seven key requirements for trustworthy AI was also performed in **Paper E** to understand how the recommendations can contribute to the responsible integration and adoption of LLMs for RE in practice. These ten recommendations align with four of the seven key requirements for trustworthy AI. Recommendations **R3**, **R4**, **R5**, **R8** and **R9** align with the *human agency and oversight* as well as *accountability* requirements. **R6**, **R7** & **R10** align with the *technical robustness* requirement. **R1** & **R2** align with the *technical robustness* and *privacy & data governance* requirements. This points towards the need to take a "priori" approach rather than a "post-hoc" approach when complying with the AIA requirements as their alignment with the recommendations post-hoc did not ensure the complete coverage of the AIA requirements.

To answer RQ4, ten recommendations, labelled R1-R10, were synthesised to help for users and organisations with the responsible integration and adoption LLMs for RE in industrial practice. There is a need to take a "priori" approach rather than a "post-hoc" approach when complying with the AIA requirements.

1.6 Threats to Validity

While the thesis presents a variety of insights to address multiple aspects of the application of LLMs within RE, it is essential to acknowledge and address the potential threats to the validity of the findings. The guidelines provided by Runeson and Höst for qualitative research analysis in software engineering [67] were followed to discuss the potential threats to the validity of this research study.

Internal Validity:

Changes or updates to the GPT models used within the quasi-experimental studies in **Paper A**, **Paper B** and **Paper C** could affect the results achieved. Transparency about the details of models used was maintained in an effort to mitigate the threat of generalisability based on the achieved results.

The participants may have provided responses they believe are socially acceptable or desirable rather than their true opinions or experiences. To mitigate this, the interview participants from **Paper A** and **Paper E** were reminded of the confidentiality of their responses to encourage honesty. There is also a risk of the Hawthorne effect where the participants may have altered their behaviour or responses simply because they are being observed or interviewed, leading to artificial results.

Coming to tertiary reviews, studies with positive results are more likely to be published, leading to an overrepresentation of positive findings in the literature. To mitigate this, the tertiary review conducted in **Paper D** was supplemented with a review of grey literature to counteract the overrepresentation of positive findings.

Ambiguous or leading questions within the survey questionnaire may elicit misleading responses from the survey participants. To mitigate this, the survey questions in **Paper E** were constructed using a clear, precise and neutral language to avoid leading the participant towards a particular response. Multiple researchers involved in the study reviewed the questionnaire to check for potential bias and ambiguity. The survey questionnaire was also pilot tested with a control group sampled using convenience sampling [74].

External Validity:

The extent to which the findings of experiments from **Paper A**, **Paper B** and **Paper C** can be generalised to real-world settings is an external validity threat. The experiments were conducted in controlled environments that do not accurately reflect the complexities of software development in the real world.

The sample of interviewees within **Paper A** and **Paper E** may not be representative of the broader population to different extends. Cultural differences between the interviewer and interviewee is another threat that may impact the interpretation of responses or the effectiveness of communication. In an effort to mitigate this, the interview questions were phrased using a clear, precise and neutral language and the interviews were conducted while following this ethical interview checklist [71].

When performing tertiary reviews, if publication bias is present in the primary studies, the review's findings may not accurately reflect the true distribution of effects in the literature. Since the tertiary review performed in **Paper D** was supplemented with grey literature findings, publication bias

was reduced, leading to less biased review findings and a reduced distribution effect.

The survey sample within **Paper E** may not be entirely representative of the broader population of software practitioners who have experience using LLMs for RE.

Construct Validity:

If the concepts being measured in the interview are not clearly defined or understood by both the interviewer and the interviewee, the validity of the measurements can be compromised. In **Paper A**, Neither ChatGPT nor the requirements elicitation interview participants were provided definitions for Accuracy & Robustness, Safety, Non-discrimination, Transparency & Explainability, Accountability, Privacy & Security, Regulations, and Human Agency & Oversight.

To ensure that constructs (concepts, tools, artefacts) studied in **Paper E** accurately represent the generalised phenomena of interest, a survey was conducted to validate the findings beyond the operational constraints of the case. However, these constructs may in turn be interpreted differently by different respondents. To mitigate this, the survey was pre-tested to ensure that the questions were understood as intended.

Conclusion Validity:

Using inappropriate statistical techniques or failing to account for confounding variables can compromise the validity of conclusions. The validity of conclusions was reinforced by comparing and contrasting the findings of the experimental studies from **Paper A**, **Paper B** and **Paper C** with the relevant state-of-the-art literature.

Misinterpretation or incorrect recording of participants' responses can affect the conclusions drawn from the data. To reduce recording errors, the recordings of the interviews conducted as part of **Paper A** and **Paper E** were compared with the interview transcripts and validated through member checking [75].

Variability in methodologies, measures, and participant characteristics among primary studies can make it challenging to draw meaningful conclusions from the tertiary review. To mitigate this threat, a meta-analysis was conducted in **Paper D** to synthesise findings from different studies and assess variability.

1.7 Discussion

This section analyses the research findings by comparing and contrasting them with existing literature wherever necessary within the broader context of the research problem this thesis addresses. The implications of the results for the academic body of knowledge and industrial practice are also presented.

The learnings from of the quasi-experimental studies can be summarised as follows. First, the type of the task being performed affects the performance of the LLM when assisting a user with an RE task. Tasks vary in their suitability for execution with the assistance of LLMs. This indicates towards a need for a systematic taxonomy or assessment metric to decide what tasks can and should be handed over to LLMs and which of them should not be. This will ensure that tasks are assigned based on the capabilities of the LLMs, thereby

optimising their use and preventing potential misuse or over-reliance on LLMs for tasks they are not suited for.

Second, prompt engineering was identified to be another factor that plays a significant role in the downstream task performance of LLMs. Choosing the right prompt engineering strategy through experimentation has been observed to improve the performance of the LLMs on specific tasks.

Apart from that, three things were identified as confounding factors that have a significant effect on performance of the LLM: the choice of the model (e.g; GPT-3.5, GPT-4, BERT, LLaMA), the prompt engineering techniques employed and the design of the selected task to be performed using the LLM. For the results of any experimental studies involving the use of LLMs to assist users, these three confounding factors need to be diversified to get any generalisable conclusions. That is to say, the experiments need to be conducted using multiple LLM models, employing multiple prompt engineering techniques and standardising the design of the task using LLM. On top of this, the deployment context also needs to be taken into consideration. It is impossible to replicate the results from a controlled experiment in an industrial setting even if the identified confounding factors were managed. Taking all of these into consideration while designing and performing future experimental studies might help in achieving results that tend to stay relevant for a longer period of time. This will help in identifying and addressing any confounding factors early on, ensuring a smoother technology transfer process when the LLMs are eventually integrated into the industry.

The findings of the tertiary review study indicate the necessity to continuously assess and manage the seven key requirements underlined by the AIA. However, the state-of-practice landscape lacks tools, artefacts and processes that support the required activities that are necessary to be compliant with the AIA and responsibly employ LLMs.

The industrial case study led to synthesising ten recommendations to facilitate the responsible and efficient integration of LLMs into RE tasks and processes. One of the major implications of this study is the need to develop artefacts and processes that operationalise human oversight and validation.

Shifting the focus to look at the overall research landscape within the field, the underlying process of LLMs when performing a task is observed to be driven by associative reasoning rather than analogical reasoning [76]. That means LLMs are capable of identifying semantic similarities and patterns between objects, and associate them to generate the requested information. However, what they are not very good at is applying critical reasoning to solve problems from a previously unseen context based on their understanding of the semantic similarities and patterns between objects.

However, this does not align with the results reported from the performed quasi-experimental studies. LLMs have been observed to be good at generating requirements and performing user story quality evaluation in direct comparision with human RE experts and NLP tools. These tasks typically require analogical reasoning to not only identify the semantic similarities between RE domain concepts but also understand the context of the problem that is to be solved and respond accordingly. The requirements traceability task, which requires identifying semantic similarities and patterns between requirements, was the task at which the LLM demonstrated the lowest performance. This observation

is in contradiction with the fact that LLMs are supposed to be good at performing associative reasoning-based tasks.

The purpose of this discussion is not to assert that LLMs are capable of analogical reasoning, but rather to challenge the prevailing perceptions about concepts such as how LLMs process the user's input prompts and perform downstream tasks. Drawing a comparison between the functioning of LLMs and human cognitive operations does not seem equitable. This perspective also prompts the question of whether it is necessary for LLMs to perform tasks that humans are already efficient at, simply because they have the capability to do so. It is crucial to view LLMs not as replacements for humans, but as catalysts that enhance the human productivity within the software development activities.

All of the above-mentioned implications need to be taken into consideration when working on addressing **Goal-3**. There is a need for developing organisational artefacts and processes that support the adoption and integration of LLMs within software development companies in a responsible and trustworthy manner. This is important because the success of any technology adoption not only depends on the technology itself but also on how well it is accepted and used by the people in the organisation. Considering these aspects can lead to better user support and more effective integration of LLMs into existing workflows. This involves working on developing a trustworthy human-LLM collaboration framework and synthesising interaction mechanisms, processes, and artefacts that support this framework.

1.8 Conclusion and Future Work

This thesis provides evidence and insights that address **Goal-1** and **Goal-2**, and presents some preliminary evidence that motivates further investigation addressing **Goal-3**. To that end, a multi-method research design was employed to formulate a working hypothesis and four research questions.

The answers to RQ1 and RQ2 suggest that LLMs are capable of performing and assisting users in various RE tasks and processes with varying levels of performance. For instance, the nature of the task and the prompt engineering have a significant impact on the performance of the LLM. Moreover, the results of RQ3 indicate that the use of LLMs to help users with their RE tasks and processes needs to comply with the requirements of the AIA. This involves continuously assessing and managing the seven key requirements outlined in the AIA. Thus, ten recommendations to guide users and organisations with the responsible integration of LLMs in RE workflows are presented, addressing RQ4.

The findings presented and discussed within this thesis have various implications. First,the necessity for a systematic taxonomy for task delegation to LLMs has been underscored, challenging the extent and suitability of their applicability within software engineering. Second, the importance of validating LLMs in real-world-like environments, key to address the gap between academic and industrial practices. Furthermore, the under-emphasised human and organisational aspects have been brought to the forefront, highlighting the need for a holistic approach to the integration of LLM-based tools and

services. Lastly, the practice of comparing LLMs with human cognition has been discussed, drawing attention to their fundamental differences and the potential for unrealistic expectations and misuse.

The research presented in this dissertation has laid the groundwork for further exploration into the integration of LLMs within software development organisations. However, several avenues remain open for future investigation.

Future work needs to delve deeper into the challenges associated with the adoption and integration of LLMs. This needs to involve a systematic identification of these challenges and the development of strategies to address them. An integral part of this process will be an in-depth study of organisational barriers and deciding factors that influence the adoption of LLMs in software engineering processes. In addition, an industry-centric investigation into the technological and organisational readiness level for LLMs adoption within software engineering is also important. This will also involve the development of a maturity model to continuously assess the readiness and performance levels of LLMs and the organisational artefacts that support the use of the LLMs. The goal is to propose best practices for LLM adoption. The final goal of this research direction is to work on the development of a trustworthy human-AI framework that enables the responsible adoption and integration of LLMs in software engineering organisations. This framework will serve as a guide for organisations seeking to leverage the power of LLMs in their software development processes efficiently and in a responsible way.

Bibliography

[1] Alhoshan, Waad and Ferrari, Alessio and Zhao, Liping, "Zero-shot Learning for Requirements Classification: An Exploratory Study," *Information and Software Technology*, p. 107202, 2023.

[2] Zhao, Liping and Alhoshan, Waad and Ferrari, Alessio and Letsholo, Keletso J and Ajagbe, Muideen A and Chioasca, Erol-Valeriu and Batista-Navarro, Riza T, "Natural Language Processing for Requirements Engineering: A Systematic Mapping Study," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–41, 2021.

[3] Ji, Ziwei and Lee, Nayeon and Frieske, Rita and Yu, Tiezheng and Su, Dan and Xu, Yan and Ishii, Etsuko and Bang, Ye Jin and Madotto, Andrea and Fung, Pascale, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[4] Khurana, Diksha and Koli, Aditya and Khatter, Kiran and Singh, Sukhdev, "Natural Language Processing: State-of-the-Art, Current Trends and Challenges," *Multimedia Tools and Applications*, pp. 1–32, 2022.

[5] Shen, Yiqiu and Heacock, Laura and Elias, Jonathan and Hentel, Keith D and Reig, Beatriu and Shih, George and Moy, Linda, "ChatGPT and Other Large Language Models Are Double-edged Swords," p. 230163, 2023.

[6] Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and others, "Language Models are Few-shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[7] Haque, Mubin Ul and Dharmadasa, Isuru and Sworna, Zarrin Tasnim and Rajapakse, Roshan Namal and Ahmad, Hussain, ""I think this is the most disruptive technology": Exploring Sentiments of ChatGPT Early Adopters using Twitter Data," *arXiv preprint arXiv:2212.05856*, 2022.

[8] Choi, Jonathan H and Hickman, Kristin E and Monahan, Amy and Schwarcz, Daniel, "ChatGPT Goes to Law School," *Available at SSRN*, 2023.

[9] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer, "The Impact of AI on Developer Productivity: Evidence from Github Copilot," *arXiv preprint arXiv:2302.06590*, 2023.

[10] T. Tõemets, "Analysing the Quality of User Stories in Open Source Projects," Ph.D. dissertation, University of Tartu, 2020.

[11] J. Cleland-Huang, S. Mazrouee, H. Liguo, and D. Port, "NFR," 2007.

[12] A. Ferrari, G. O. Spagnolo, and S. Gnesi, "PURE: A Dataset of Public Requirements Documents," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017, pp. 502–505.

[13] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language Models are Unsupervised Multitask Learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[14] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," 2023.

[15] C. of the EU | Press release, *Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI*, 2024. [Online]. Available: https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/?utm_source=substack&utm_medium=email

[16] K. Wiegers and J. Beatty, *Software Requirements*. Pearson Education, 2013.

[17] D. M. Fernández, S. Wagner, M. Kalinowski, M. Felderer, P. Mafra, A. Vetrò, T. Conte, M.-T. Christiansson, D. Greer, C. Lassenius *et al.*, "Naming the Pain in Requirements Engineering: Contemporary Problems, Causes, and Effects in Practice," *Empirical software engineering*, vol. 22, pp. 2298–2338, 2017.

[18] M. A. Umar and K. Lano, "Advances in Automated Support for Requirements Engineering: A Systematic Literature Review," *Requirements Engineering*, pp. 1–31, 2024.

[19] A. Batool, Y. H. Motla, B. Hamid, S. Asghar, M. Riaz, M. Mukhtar, and M. Ahmed, "Comparative Study of Traditional Requirement Engineering and Agile Requirement Engineering," in *2013 15th International Conference on Advanced Communications Technology (ICACT)*, 2013, pp. 1006–1014.

[20] Y. Li, J. Keung, Z. Yang, X. Ma, J. Zhang, and S. Liu, "SimAC: Simulating Agile Collaboration to Generate Acceptance Criteria in User Story Elaboration," *Automated Software Engineering*, vol. 31, no. 2, p. 55, 2024.

[21] A. Mehraj, Z. Zhang, and K. Systä, ""A Tertiary Study on AI for Requirements Engineering"," in *Requirements Engineering: Foundation for Software Quality*, D. Mendez and A. Moreira, Eds. Cham: Springer Nature Switzerland, 2024, pp. 159–177.

[22] Zhang, Jianzhang and Chen, Yiyang and Niu, Nan and Liu, Chuang, "A Preliminary Evaluation of ChatGPT in Requirements Information Retrieval," *arXiv preprint arXiv:2304.12562*, 2023.

[23] J. Zhang, Y. Chen, N. Niu, Y. Wang, and C. Liu, "Empirical Evaluation of ChatGPT on Requirements Information Retrieval Under Zero-Shot Setting," 2023.

[24] Angela Fan and Beliz Gokkaya and Mark Harman and Mitya Lyubarskiy and Shubho Sengupta and Shin Yoo and Jie M. Zhang, "Large Language Models for Software Engineering: Survey and Open Problems," 2023.

[25] T. Goyal, J. J. Li, and G. Durrett, "News Summarization and Evaluation in the Era of GPT-3," *arXiv preprint arXiv:2209.12356*, 2022.

[26] Nakano, Reiichiro and others, "WebGPT: Browser-assisted Question-answering with Human Feedback," *arXiv preprint arXiv:2112.09332*, 2021.

[27] Xie, Qianqian and Luo, Zheheng and Wang, Benyou and Ananiadou, Sophia, "A Survey on Biomedical Text Summarization with Pre-trained Language Model," *arXiv preprint arXiv:2304.08763*, 2023.

[28] Kheiri, Kiana and Karimi, Hamid, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," *arXiv preprint arXiv:2307.10234*, 2023.

[29] 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*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[30] A. Vogelsang and J. Fischbach, "Using Large Language Models for Natural Language Processing Tasks in Requirements Engineering: A Systematic Guideline," 2024.

[31] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen *et al.*, "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models," *arXiv preprint arXiv:2303.10420*, 2023.

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and Efficient Foundation Language Models," *arXiv preprint arXiv:2302.13971*, 2023.

[33] M. AI, *Introducing Meta Llama 3: The most capable openly available LLM to date*, 2024. [Online]. Available: https://ai.meta.com/blog/meta-llama-3/

[34] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large Language Models Struggle to Learn Long-tail Knowledge," in *International Conference on Machine Learning*. PMLR, 2023, pp. 15 696–15 707.

[35] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv preprint arXiv:2309.01219*, 2023.

[36] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2021.

[37] P.-H. Wang, S.-I. Hsieh, S.-C. Chang, Y.-T. Chen, J.-Y. Pan, W. Wei, and D.-C. Juan, "Contextual Temperature for Language Modeling," 2020.

[38] C. Wang, S. X. Liu, and A. H. Awadallah, "Cost-Effective Hyperparameter Optimization for Large Language Model Generation Inference," 2023.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," 2015.

[40] M. Renze and E. Guven, "The Effect of Sampling Temperature on Problem Solving in Large Language Models," 2024.

[41] V. Liu and L. B. Chilton, "Design Guidelines for Prompt Engineering Text-to-Image Generative Models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, ser. CHI '22. New York, NY, USA: Association for Computing Machinery, 2022. [Online]. Available: https://doi.org/10.1145/3491102.3501825

[42] Fiannaca, Alexander J. and Kulkarni, Chinmay and Cai, Carrie J and Terry, Michael, "Programming without a Programming Language: Challenges and Opportunities for Designing Developer Tools for Prompt Programming," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3544549.3585737

[43] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[44] 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

[45] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to Detect Unseen Object Classes by Between-class Attribute Transfer," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 951–958.

[46] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning

in Large Language Models," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[47] Iyer, Lakshmi Shankar, "AI-Enabled Applications Towards Intelligent Transportation," *Transportation Engineering*, vol. 5, p. 100083, 2021.

[48] Kaur, Davinder and Uslu, Suleyman and Rittichier, Kaley J and Durresi, Arjan, "Trustworthy Artificial Intelligence: A Review," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–38, 2022.

[49] H.-L. E. G. on Artificial Intelligence, *Ethics Guidelines for Trustworthy Artificial Intelligence (AI)*, 2019. [Online]. Available: www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[50] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy Artificial Intelligence," *Electronic Markets*, vol. 31, pp. 447–464, 2021.

[51] J. W. Creswell, "Research Designs. Qualitative, Auantitative, and Mixed Methods Approaches," 2009.

[52] K. Petersen and C. Gencel, "Worldviews, Research Methods, and their Relationship to Validity in Empirical Software Engineering Research," in *2013 Joint Conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, 2013, pp. 81–89.

[53] R. Valerdi and H. L. Davidz, "Empirical Research in Systems Engineering: Challenges and Opportunities of a New Frontier," *Systems Engineering*, vol. 12, no. 2, pp. 169–181, 2009.

[54] S. F. Brown, "Naivety in Systems Engineering Research: Are We Putting the Methodological Cart Before the Philosophical Horse," in *7th Annual Conference on Systems Engineering Research (CSER 2009)*. Citeseer, 2009.

[55] Paul Ralph and others, "Empirical Standards for Software Engineering Research," 2021.

[56] Y.-G. Guéhéneuc and F. Khomh, *Empirical Software Engineering*. Cham: Springer International Publishing, 2019, pp. 285–320. [Online]. Available: https://doi.org/10.1007/978-3-030-00262-6_7

[57] J. W. Anastas, *Research Design for Social Work and the Human Services*. Columbia University Press, 2000.

[58] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in software engineering*. Springer Science & Business Media, 2012.

[59] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, *Selecting Empirical Methods for Software Engineering Research.* London: Springer London, 2008, pp. 285–311. [Online]. Available: https://doi.org/10.1007/978-1-84800-044-5_11

[60] D. M. Berry, J. Cleland-Huang, A. Ferrari, W. Maalej, J. Mylopoulos, and D. Zowghi, "Panel: Context-Dependent Evaluation of Tools for NL RE Tasks: Recall vs. Precision, and Beyond," in *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 2017, pp. 570–573.

[61] B. Kitchenham and S. Charters, "Guidelines for Performing Systematic Literature Reviews in Software Engineering," vol. 2, 01 2007.

[62] B. A. Kitchenham, T. Dyba, and M. Jorgensen, "Evidence-based Software Engineering," in *Proceedings. 26th International Conference on Software Engineering.* IEEE, 2004, pp. 273–281.

[63] IEEE, *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. IEEE, 2017.* IEEE, 2021. [Online]. Available: http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[64] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000381137

[65] OECD, *Artificial Intelligence in Society.* OECD, 6 2019. [Online]. Available: https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en

[66] C. European Commission, Directorate-General for Communications Networks and Technology, *EUR-Lex - 52021PC0206 - EN - EUR-Lex*, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[67] Runeson, Per and Höst, Martin, "Guidelines for Conducting and Reporting Case Study Research in Software Engineering," *Empirical Software Engineering*, vol. 14, pp. 131–164, 2009.

[68] Robson, Colin, *Real World Research: A Resource for Social Scientists and Practitioner-Researchers.* Blackwell, 1993. [Online]. Available: https://cir.nii.ac.jp/crid/1130282271686871424

[69] Krippendorff, Klaus, *Content Analysis: An Introduction to its Methodology.* Sage publications, 2018.

[70] S. Hove and B. Anda, "Experiences from Conducting Semi-structured Interviews in Empirical Software Engineering Research," in *11th IEEE International Software Metrics Symposium (METRICS'05)*, 2005, pp. 10 pp.–23.

[71] Strandberg, Per Erik, "Ethical Interviews in Software Engineering," in *2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 2019, pp. 1–11.

[72] E. R. Babbie, "Survey Research Methods," 1973.

[73] M. Ciolkowski, O. Laitenberger, S. Vegas, and S. Biffl, "Practical Experiences in the Design and Conduct of Surveys in Empirical Software Engineering," 2003.

[74] S. Baltes and P. Ralph, "Sampling in Software Engineering Research: A Critical Review and Guidelines," *Empirical Softw. Engg.*, vol. 27, no. 4, jul 2022. [Online]. Available: https://doi.org/10.1007/s10664-021-10072-8

[75] Seaman, C.B., "Qualitative Methods in Empirical Studies of Software Engineering," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 557–572, 1999.

[76] C. E. Stevenson, M. ter Veen, R. Choenni, H. L. J. van der Maas, and E. Shutova, "Do Large Language Models Solve Verbal Analogies Like Children Do?" 2023.