Chippy Chandra Pillai

Bridging The QA Gap: A Lean Project Management Framework for Migrating to AI-Powered Testing

School of Technology and Innovations

Strategic Project Management

Master's Programme in Industrial Engineering and Management

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovation
Author: Chippy Chandra Pillai
Title of the thesis: Bridging The QA Gap: A Lean Project Management Framework for Migrating to AI-Powered Testing: Possible specifying subheading
Degree: Master of Science in Technology
Degree Programme: Strategic Project Management
Supervisor: Jari Ruokalainen
Year: 2026 **Pages:** 90

ABSTRACT:

Software has come to play a major role in the operations of most contemporary organisations, in particular in those areas where reliability and accuracy are main issues. With the growth in the complexity of software, the traditional Quality Assurance (QA) practices, particularly script-based automation has been found to struggle with overheads in maintenance, limited scalability and inefficiencies. Meanwhile, the potential alternative can be AI-powered testing tools, which offer both adaptive features and better efficiencies. But the way toward the legacy QA automation to AI-powered test tools has not been paved easily and is fraught with both technical and organisational issues.

This paper attempts to explore how a transition like this can be effectively handled in a systematic way. The research problem is addressed by developing a Lean QA Migration Framework (LQMF), which is grounded in Lean thinking and supported by concepts related to organisational change and technology adoption. The research is justified based on Design Science Research approach, with an empirical case study. Semi-structured interviews with practitioners in the industry and thematic analysis facilitated by NVivo have been used to collect and analyze the data respectively.

The results suggest that the source of inefficiencies in current QA processes are usually located along repetitive maintenance activities and lack of process visibility. It is also noted that organisational barriers to the adoption of AI are mainly the ones associated with resistance to change and with issues related to trustworthiness and reliability. Moreover, AI testing tools are observed to be more used as augmentation mechanisms, and not fully autonomous solutions, in practice.

These insights suggest that the proposed framework can offer a stepwise method through which organisations can detect inefficiencies within their processes, the tools to use, and the adoption to undertake in a controlled manner. The conclusion drawn is that the crucial element in the successful transformation of QA does not solely rely on technological facilities but also the systematic procedures as well as organisational preparedness.

KEYWORDS: AI–powered testing, QA automation migration, Lean Project Management, Design Science Research, technology adoption

Contents

Figures

Tables

Abbreviations

QA	Quality Assurance
AI	Artificial Intelligence
API	Application Programming Interface
LQMF	Lean QA Migration Framework
TAM	Technology Acceptance Model
DSR	Design Science Research
TMMI	Test Maturity Model Integration
CMMI	Capability Maturity Model Integration
NDA	Non Disclosure Agreement
UTAUT	Unified Theory of Acceptance and Use of Technology
CI	Continous Integration
CD	Continuous Delivery
DevOps	Development and Operations
IT	Information Technology
LLM	Large Language model
ML	Machine Learning
NLP	Natural Language Processing
POC	Proof Of Concept
SDET	Software Development Engineer in Test
UI	User Interface
VSM	Value Stream Mapping
VM	Virtual Machine

1 Introduction

1.1 Background

Software applications are important in every sector, particularly in the banking, healthcare and government sectors. With the growth of the complexity of the software systems, it has become a critical challenge to ensure software systems' integrity, reliability and security. In this regard, Quality Assurance (QA) is of great importance to ensure software systems behave accordingly to their requirements.

Over the years, with the emergence of Agile and DevOps, modern Quality Assurance (QA) has evolved from being manual to fully automated and data-driven software testing. This kind of approach motivated the adoption of continuous delivery. Automation scripting tool such as selenium, Cypress, among others have been widely adopted to make the process of testing faster and more efficient (Kothamali, 2025). This was very beneficial for organisations with a large regression suite. However, automation via scripts was inherently problematic, although this was only revealed over time. Automated tests were extremely delicate as they used particular locators and scripts. They can break when a single line of code is changed, and need to be fixed by hand. In dynamic environments, this is an ongoing process(Leotta et al., 2015). Nidagundi & Novickis (2016) examined this cycle in the context of lean transformation in software testing and found the cost of maintaining scripts increases over time as the size of the application and test suite increases.

Meanwhile, the advances in machine learning (ML) and artificial intelligence (AI) bring innovations in software testing. Instead of failing due to a change in application, testing solutions adapt themselves to the structure of the application. The best feature is self-healing automation, if an element moves, the application knows what to look for and automatically updates itself. Apart from Self-healing capabilities, an AI driven platform provides Intelligent testing and Visual AI testing(Amalfitano et al., 2024; Garousi et al., 2020).

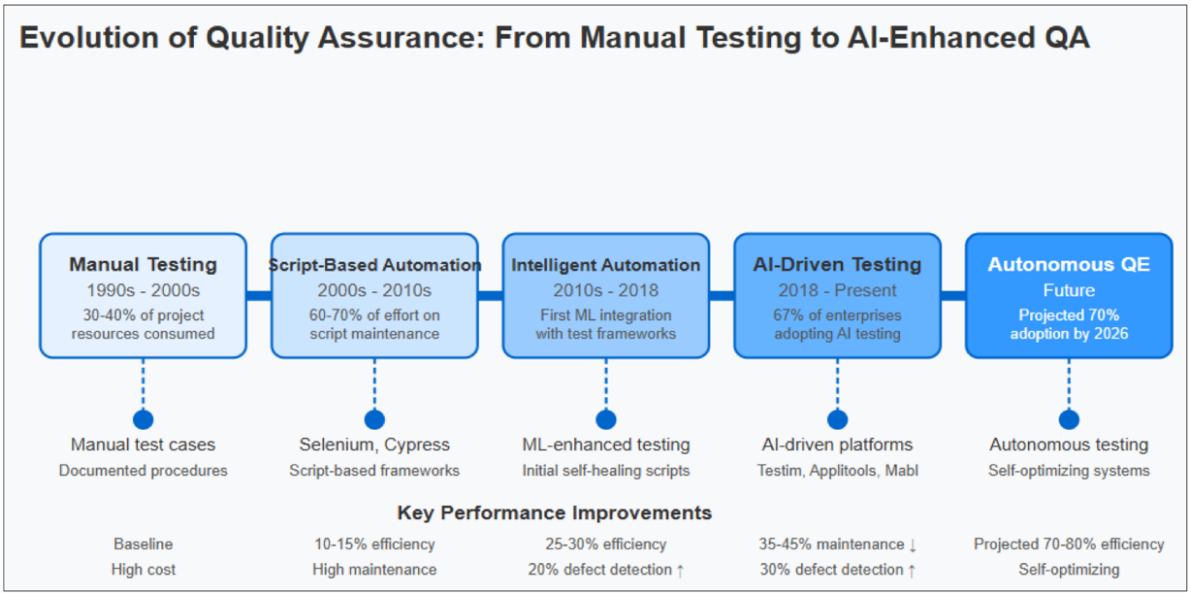

Figure 1. Evolution of QA source:(Kothamali, 2025)

In practice, these platforms report results in lower test maintenance costs and higher defect detection rates than traditional automation strategies and workflows. These are not only incremental gains. They are a game-changer in the cost / benefit of QA automation, but the transition from a traditional system to an automated platform powered by AI is more than a change in technology. The team needs to be retrained. The process needs to be redesigned. Existing test assets need to be assessed, converted or replaced, and decisions need to be made including what tool to use, how to prioritise the migration, and how to ensure the new system is beneficial. Without a clear strategy to manage this, many migrations do not achieve their full potential or are very disruptive to the teams.

Therefore, applying Lean principles in QA transformation could provide a solution to manage the transition. Reviewing the existing testing practice and highlighting the options of reducing testing waste and creating benefit can help organisation plan a migration that will reduce disruption and potential new technology opportunities. This study is supported with an empirical case study of AquilaTest Inc., as well as industry practitioners.

The aim of the study is to uncover the obstacles that organisations encounter when using AI-powered testing tool, and offer an analysis of the steps using Lean principles and procedures needed to facilitate the transition from legacy QA automation to AI-powered testing.

1.2 Research problem and significance

The research problem this thesis addresses is clear in outline but complex in practice. Although software development has become Increasingly automated, organisations still struggle with high maintenance costs, lack of scalability and efficiency in software testing. Meanwhile, AI-based testing has the potential to remove these issues. But it is not a straightforward process to shift from traditional automation frameworks to AI-powered testing. The company has to deal with challenges such as tool-interface, training, confidence on AI-powered decision-making, and integration with the current processes.

One of the issues is a lack of a process by which this can be done. Organisational governance needs to be in place, otherwise technical debt can be built into the new system as it did with the legacy system. Without success measures, it can be difficult to convince stakeholders that the system is effective, leading to the loss of organizational support for sustaining change. (Hevner et al., 2004).

Most existing research focuses on how to deal with the technical aspects of applying the AI testing tools, but not much has been done on how to organise the processes to migrate. This leaves project teams with little support as they go through this transition.This creates a gap in research on how to manage the journey from the traditional QA automation technologies to AI to fix problems with business process backlog and gain effectiveness. In sectors like aviation, financial, healthcare, etc., failure to pick the defect would be costly compared to other sectors. Developing the current QA capability in this area in a structured way is not just a matter of professional honor

but also an important aspect of competitive advantage. Transitioning in established organizations to AI is one of the most critical challenges of the era, and the principles learned on the migration to QA have relevance beyond the realm of software testing(Brocke et al., 2020).

1.3 Research questions and objectives

The research question is based on dimensions of the migration challenges such as waste diagnosis, tool selection, change management, and governance. The main research question is

“How can an organisation effectively manage the migration from legacy QA automation to AI-powered testing to eliminate the process waste and sustain long-term quality improvement?”

The above question has three components that cover the demands on the framework. The word “effectively” requires that the framework be actionable and produce measurable outcomes. The phrase “eliminate the process waste” anchors the framework in lean thinking and commits the research to a diagnostic approach that must precede technology selection. Lastly, the phrase “sustain longterm quality improvement” requires a framework to address governance and continuous improvement as its main concerns.

To address the research question, the study pursues the following objective:

(1) To examine the limitations and challenges associated with legacy QA automation in the selected case organisation.
(2) To analyse the types of process waste present in the current QA workflow using lean principles.
(3) To explore organisational readiness and approach on AI-powered testing
(4) To analyse practitioner experience related to QA transformation

(5) To develop a Lean QA Migration Framework(LQMF)that supports structured transition

1.4 Scope of the study

This research focuses on the process and project management aspects of transitioning from legacy QA automation to AI- powered testing. The study does not focus on an AI algorithm. Instead, it examines how an organisation can manage the transition in a structured and efficient manner. Functional and API testing are given primary emphasis, whereas other testing areas receive comparatively less attention. In contrast, End-to-End testing remains a significant challenge for AI-powered testing tools, thereby limiting their effectiveness in addressing such a comprehensive testing scenario.

The empirical component of the study is based on insights from industry practitioners, including a case organisation involved in AI-powered testing and professionals working in an enterprise environment. While these perspectives provide valuable insight, the study does not attempt to generalize these findings across all industries.

1.5 Significance of the study

This study contributes to both academic research and practical. From an academic perspective, it addresses a gap in understanding how organisations can manage the transition to AI-powered testing from a process perspective. While the existing literature focuses on technical capabilities, this research focuses on the importance of structured migration strategies.From a practical perspective, this study offers guidance for organisations seeking to modernize their QA practices. The proposed framework provides a step-by-step approach that can help teams identify inefficiencies, plan migration activities, and adopt new testing technologies in a controlled manner.

2 Literature review

This chapter provides an overview of the bodies of knowledge that the Lean QA Migration Framework draws upon. It spans four interrelated themes: legacy test automation structural constraints, AI-powered testing capabilities and adoption, Lean management in the software sector and software digital transformation, and the theoretical underpinnings of managing technology-driven changes in organisations. The aim of the chapter is not to survey all papers related to these areas. It is selective and purposeful: every section explains the theory behind a particular design decision in the LQMF and Section 2.6 visualises the five research gaps the framework is meant to resolve.

2.1 Legacy QA automation and structural limitations

Agile and DevOps delivery practices completely transformed the role of software testing. A seminal conceptual research on agility in ISD was undertaken by Conboy (2009), who proposed that agility in ISD is characterised by the constant ability of an organisation to rapidly generate change in an IT system, accepting change and uncertainty. It's definition is new on testing. Verification can no longer be done at the end of the development process — it needs to be automated, continuous and as close as possible to any code change. The solution seemed to come in the form of test automation: a suite could run hundreds of scenarios, whereas a human tester would take days with only a handful, and it could do so without being touched inside a CI/CD pipeline.

This method worked under stable conditions quite nicely. It found itself in a difficult situation and, to this day, remains difficult in most environments where software is developed today: with applications that constantly evolve, interfaces that change with each sprint, delivery pipelines that merge changes several times a day. Structural problem is locator brittleness. Automated scripts locate UI elements with locators, which are exact references to where and how an element is located in a page's structure.

If the structure is modified, even as a side-effect of another code change, the locator will also become invalid and the test will fail. The failure does not indicate a bug in the software: it indicates that the software is not in the same form as the script assumed. This cycle of breaking and fixing happens over and over again in code bases that change all the time (Leotta et al., 2015).

This adds to the maintenance burden which is not an indicator of substandard engineering practice. It is one of structural properties of locator dependent paradigm. Leotta et al. (2024) addressed this issue with an empirical comparison of the various types of automation approaches (namely NLP-based, programmable and capture-and-replay), reaching the same conclusion that none of the existing script-based approaches eliminates maintenance overhead, but that it merely shifts it. It is shifted depending on the approach, but no one paradigm takes it away. In their analysis of practitioner accounts in the grey literature, Ricca et al. (2021) identified script maintenance as the most consistently and strongly mentioned challenge in all categories of tools in real-world automation practice. What comes out of all the evidence is not a failure of script-based automation, but it was doing a job that it was needed to do and it is still doing it in the right contexts. The claim is more specific: As the delivery speed rises and application change frequency rises, the ratio of maintenance effort and value of defect-detection falls progressively. After a certain stage the automation suite starts to work against the team instead of for the team.

LQMF Phase 2 (Waste Identification) is meant to quantify this deterioration prior to any migration decision. All seven Lean waste types are represented in the legacy automation context: Overproduction, Waiting, Inventory, Motion, Overprocessing, Defects, Transport. Scripts that are not maintained, but cover functionality not in production are Inventory. Full regression suite runs on each minor change are overprocessing. The most expensive waste is when flaky test signals cause the team to lose faith in the automation output and defects make it to production. The analytical bedrock for all the migration is making these costs visible and specific – in engineering hours and release cycles.

2.2 AI-powered testing: capabilities, evidence, and adoption reality

Artificial Intelligence (AI) has introduced new capabilities in software testing, enabling more adaptive intelligent automation. Techniques such as Natural language processing(NLP) and machine learning are increasingly used to generate, execute, and optimize test cases. Amalfitano et al. (2024) conducted a tertiary study, a structured review of existing systematic reviews covering AI in software testing research published between 1995 and 2021. They found that the field was developed across four main areas- Automated test case generation, defect prediction, test prioritization and self-healing automation. Each addresses a known limitation of script-based approaches.

Self-healing automation is the most practically relevant capability for organisations migrating from legacy tools. Rather than relying on fixed locators, self-healing platforms learn the structure of an application and update their element references automatically (Ricca et al., 2021). The test continues to execute correctly without manual intervention. Garousi et al. (2024) conducted a systematic review and empirical assessment of AI-powered testing tools and found genuine capability in self-healing and risk-based test prioritization, while noting that claims about fully automated end-to-end test generation remain aspirational in most production contexts.

NLP-based and ML-based test generation represent a second significant capability stream. Fontes & Gay (2023) subsequently conducted a systematic mapping of machine learning integration into automated test generation, confirming that ML-based approaches improve test coverage and reduce the manual effort required for test design, though the maturity of deployable implementations varies considerably across test types and application domains. Ayenew & Wagaw (2024) reviewed more recent NLP approaches that incorporate transformer architecture and found measurable improvements in general test quality, though dependency on well-structured requirement documentation remains a practical constraint.

Large language models have added a new dimension to test generation research. Wang et al. (2024) conducted a comprehensive survey of software testing with a large language model, published in IEEE Transactions on Software Engineering. Their analysis covered test case generation, test oracle inference, test suite maintenance, and bug reproduction, finding that LLMs demonstrate genuine utility across all four areas but introduce new challenges around hallucination, nondeterminism, and the difficulty of validating whether a generated test captures the intended behaviour. The survey identified the integration of LLMs into existing test workflow as the most productive near term research and practice direction rather than as standalone generation engines. Allala et al., (2019) had demonstrated a technically coherent pathway from natural language requirements to executable test cases, and the LLM work Wang et al. (2024) represents a substantial extension of that direction into practically deployable territory.

2.3 Aspiration-adoption Gap

Despite the technical progress described above, AI testing adoption in industry lags considerably behind the attention it receives in research and in commercial markets. Karhu et al. (2025) performed a systematic mapping study of industry context, empirical research on AI adoption in software testing published since 2020. Their conclusion is striking; AI has not yet made a significant breakthrough in software testing, unlike in activities such as code generation. Thematic analysis of the available studies revealed a consistent pattern, strong strategic aspiration at the leadership level and limited, uneven adoption at the practitioner level. Industry surveys cited in their review that over 75% of organisations identified AI-driven testing as a strategic priority for 2025, while only 16% reported having adopted it in practice.

The gap is not primarily technological. The tools exist; their capabilities are real. Wang et al., (2024) noted that while LLMs raise expectations further, with practitioners reading about ChatGPT-style generation and expecting immediate applicability, the practical

integration of LLM capabilities into production testing workflows is more demanding than demonstrations suggest. Ramos et al. (2025) found in their review of AI-augmented software engineering outcomes that human and process dimensions(team skills, workflow integration, governance) determine outcomes at least as much as the technical capabilities of the selected tool. Ozkaya (2023) writing in IEEE software, argued that realizing the benefits of AI-augmented development requires deliberate attention to how tools are embedded into engineering workflow, because beneficial integration does not happen automatically.

These findings have direct implications for the framework design: LQMF is not primarily a tool selection guide. It is a migration management guide. The aspiration adoption gap documented by Karhu et al. (2025) is evidence that organisational and process dimensions of the transition. The framework emphasis on waste diagnosis before tool selection, on phased implementation, and on sustained governance reflects this diagnosis.

2.4 Lean management applied to software and digital transformation

Lean thinking, as articulated by Womack & Jones (1996), is built around the systematic identification and elimination of waste, which means any activity that consumes resources without creating any value for the customers. The five principles of lean are defining value, mapping the value stream, enabling flow, establishing pull, and pursuing perfection. Value stream mapping is the practical diagnostic tool developed by Rother & Shook (1999) that translates these principles into a method for tracing a process end to end and making visible every point where time and effort are lost without adding value. While VSM was developed for manufacturing contexts, the underlying logic of tracing activities from input to output and classifying each as value adding or waste applies directly to software testing processes.

The translation to software was demonstrated Poppendieck (2007) by those who mapped the seven canonical Lean waste types to software development contexts. Overproduction in software is writing code or tests that are never used. Waiting is the time spent on environment setup, building queues, and delayed reviews. Inventory is accumulated unfinished work, including unmaintained test scripts and unexecuted test cases. The mapping is not perfectly neat in every detail, but the conceptual correspondence is strong enough to make Lean diagnostic methods genuinely useful in the QA context. Hicks (2007) extended this further, showing that waste analysis applies equally to knowledge-intensive, information-mediated processes where the inputs and outputs are data and decisions rather than physical goods.

Lima et al. (2023) investigated the implementation of Lean project management within the digital transformation context, proposing a socio-technical framework that integrates Lean principles with digital change management. Their study found that Lean implementation in a digitally transforming organisation must address both the process and the people dimensions simultaneously, a finding that directly informs the LQMF's design, which incorporates enablement (phase 6) and governance (phase 7) as integral parts of the migration rather than afterthoughts. Liu et al. (2025) confirmed in an industry 4.0 that Lean-derived waste analysis method retains its economic productivity in technology-mediated, real-time environments, which is precisely the kind of environment that AI-powered testing infrastructure creates.

Liker (2004), In the Toyota Way, extended Lean from a process improvement toolkit into a comprehensive management system. His argument that the technical tools of Lean are necessary but insufficient, that organisations adopting the tools without the underlying management philosophy achieve short-term gains that erode over time, is directly relevant to QA migration. Building a testing function that continuously improves requires the kind of leadership development and cultural foundation that the Toyota Way addresses. The Lean principle of Kaizen, continuous improvement as an organisational habit rather than a periodic project, is the operational foundation of phase 7 of the LQMF.

What the existing literature does not provide is a study that assembles Lean theory, digital change management, and QA specific research into a structured, phased migration framework. The individual literatures are available, but integration is not. This is the theoretical contribution the LQMF makes.

2.5 Organisational change management and technology acceptance

A QA migration involves changing the tools that engineers use everyday, the processes that surround those tools, and in many cases the professional identities of the people doing the work. Managing these changes successfully requires a theory that operates at two levels: the organisational level, where leadership must create conditions for sustained change, and the individual level, where each engineer must decide to adopt a new way of working.

Kotter's (1996) eight step model operates at the organisational level, describing the conditions that must be established for major change to succeed: a sense of urgency, a guiding coalition, a strategic vision, broad communication, enable action by removing barrier, short-term wins, consolidated gains, and Anchor the change in culture. The model's most useful feature for QA migration is its diagnostic power, which reveals why migrations fail. An organisation that skips urgency creation encounters resistance that training cannot overcome. Organisations that do not generate short-term wins lose leadership support before the transition reaches scale. The LQMF incorporates Kotter's logic by treating waste analysis as the urgency-creation mechanism, making the cost of the current approach visible in terms that leadership can understand, and by designing the pilot phase to produce credible early gains before the full migration commits(Kotter, 1996).

Davis & Granić (1989) has proposed the Technology Acceptance Model as identifying perceived usefulness and perceived ease of use as the two primary psychological

determinants of adoption. An engineer who does not see how the AI testing platform improves their day to day work resist adoption regardless of the platform's objective capabilities. Karhu et al. (2025) documented this pattern directly: the adoption gap they identified corresponds to a TAM failure, leadership perceives usefulness and commits strategically, while practitioners at the point of implementation often do not, and the gap between the two is what prevents adoption from translating into practice. Venkatesh et al. (2003) extended this framework, adding social influence and facilitating conditions as additional predictors of adoption. These factors are addressed in LQMF phase 6, which sequences enablement around demonstrating usefulness in the organisation's own context before demanding commitment, and on building practitioner-level advocates before attempting organisation-wide adoption.

2.6 Existing QA transformation frameworks and their limitations

Several established frameworks support the assessment and improvement of software testing capability, and it is important to understand both what they offer and where they fall short. Garousi & Veenendaal (2022) provided a global status report on TMMI (Testing Maturity Model Integration) covering worldwide trends in TMMI assessments and certifications. Their data showed that the most common level of maturity among TMMI-certified organisations is Level 3(Defined), representing a standardized, Institutionalized testing process with documented practices. The TMMI framework provides organisations with a vocabulary for describing current testing capability and a roadmap for advancing through five defined maturity levels. It is strong on process definition and measurement, and Garousi and van Veenendaal's data confirm its value as a widely adopted benchmark for testing capability.

However, TMMI and comparable frameworks such as CMMI and Forsgren et al. (2018) Accelerate share a fundamental characteristic that limits their usefulness for the problem this thesis addresses: they define destination states rather than transition paths. An organisation can use TMMI to assess its current maturity level and define a target

level yet receive no structured guidance on how to manage the migration from one to the other, especially when that migration involves changing the fundamental technology platform of the testing function. These are improvement roadmaps, not migration management frameworks. The LQMF occupies the space they leave empty; it provides a phased, structured guide for managing the journey itself, from diagnostic waste analysis through tool selection, phased implementation, team enablement, and governance.

2.7 Research gap and Contribution to this study

The five sections above outline the theoretical and empirical context from which the LQMF is developed. The review identifies a specific and consistent gap, i.e., the elements of an integrated comprehensive QA migration framework exist in the literature but have not been integrated, such as Lean waste diagnosis, AI testing tool evaluation criteria, change management sequencing, and technology adoption theory. They run in parallel streams and are highly developed within their individual streams, but not linked in a manner a practitioner managing a real migration could follow. Table 2.2 outlines the five gaps this study addresses and provides an explanation of why each gap is important.

The fifth gap – lack of an integrated and sequenced migration framework – is the one the LQMF directly addresses. The previous four make it clear why it has not yet been constructed: AI testing research and Lean management research have not yet been interconnected; change management theory has not been applied specifically to the context of QA technology migration; and the adoption barriers reported by Karhu et al. (2025) have been described and characterised but not yet solved with a methodology for structured transition.

Table 1. Research gap identified using literature and how it connects to LQMF

Research Gap	Key Source(s)	Why It Matters for the LQMF
AI testing research is technically strong but organisationally silent	Garousi et al. (2024); Amalfitano et al. (2023); Ricca et al. (2021); Karhu et al. (2025)	Studies rigorously evaluate tool capabilities but do not address how organisations manage adoption, sequence the migration, or sustain improvement after deployment. Technical knowledge exists; migration knowledge does not.
The aspiration-adoption gap is large, persistent, and underexplored	Karhu et al. (2025); Wang et al. (2024); Ramos et al. (2024)	A large majority of organisations name AI testing as a strategic priority while a small minority have adopted it in practice. LLM-based tools raise aspirational expectations further without providing the organisational guidance needed to close the gap.
Lean management has been applied to software but not to QA technology migration	Womack & Jones (2003); Poppendieck (2007); Hicks (2007); Lima et al. (2023); Liu et al. (2025)	Lean waste analysis demonstrably applies to knowledge-intensive and digitally transforming processes. No study uses Lean as the diagnostic and design architecture for a structured QA technology migration framework.
Change management and technology acceptance theory are not integrated into QA transformation guidance	Kotter (1996); Davis (1989); Venkatesh et al. (2003)	General change and TAM/UTAUT models provide relevant theory at the organisational and individual level. They have not been assembled with waste analysis and tool selection into a migration-specific phased framework.
No integrated, sequenced migration framework exists in the literature	All reviewed literature	The theoretical and empirical building blocks — Lean diagnosis, AI tool selection criteria, change management, governance — exist separately. Their integration into a single, usable, phased guide is the specific contribution this thesis makes.

The Design Science Research methodology outlined in Chapter 3 is the right approach to this kind of gap. Hevner, March, Park, and Ram (2004) established DSR as the right paradigm when the research goal is to design an artifact that solves a real, specified problem for a defined class of users — and when the artifact does not yet exist. That artefact is the LQMF. The theoretical basis for its implementation is taken from literature reviewed in this chapter. Its empirical validation in Chapter 5 is based on primary data

from seven practitioners who are representative of the class of organisations that the framework is designed for. The evidence in this chapter supports the claim that the framework provides a genuine gap, while the evaluation in Chapter 5 supports the claim that the framework fills that gap adequately.

3 Research methodology

3.1 Research approach

This chapter explains how the research was conducted and justifies the methodological choices made at each stage. It explains the research approach, data collection methods, case study design, and data analysis techniques used to develop the proposed LQMF. It begins by situating the study within a philosophical position, then introduces Design Science Research (DSR) as the governing methodology and explains why it applies.

3.2 Design Science Research approach

DSR is a research paradigm developed for information systems by (Hevner et al., 2004). The goal of the study is to create an artifact rather than to describe or explain an existing phenomenon. Their seven principles that the artifact must address an important problem, make a knowledge contribution, be rigorously evaluated and be communicated to both academic and practitioner audiences, shape how this chapter describes the methodology and how Chapter 5 structures the evaluation.

(Gregor & Hevner, 2013) classified DSR contributions as invention, improvement, or exaptation. The LQMF sits across improvement and exaptation: it takes established knowledge from Lean management, AI testing research and change management theory and assembles it into an integrated migration framework that does not exist in the current literature.(Nielsen, 2020) argued that clear problematisation articulating precisely what is missing, is essential for DSR rigor, the research gap mentioned in chapter 2 serves this purpose. (Vom Brocke et al., 2020) proposed that well-formed DSR questions must specify a problem class, an artifact type, and an improvement goal; the research question used here satisfies all three. (Vom Brocke et al., 2020) required real world relevance, established here through seven practitioner interviews. (Winter, 2008) argued that rigour and relevance must be pursued simultaneously, a principle that

shapes both the theoretical grounding in Chapters 4 and the empirical validation in Chapter 5.

3.3 Research design: empirical case study

This study is grounded in a case organization, AquilaTest Inc., which specializes in AI-powered testing solutions. The organization offers tools and services that strive to improve test automation using capabilities that include self-healing scripts, intelligent test generation, and minimized maintenance overhead.

AquilaTest Inc. was selected as the case organization due to its active involvement in developing and implementing AI-driven QA solutions. This positioning presents a helpful perspective where the potential and constraints of AI-enabled testing can be viewed.

Although the study also involves the viewpoints of the external practitioners, the case organization remains the main empirical prism through which the problems of QA transformation are analyzed. This enables the research to be anchored to real world conditions of implementation as opposed to hypothetical assumptions alone.

3.4 Data collection

3.4.1 **Secondary literature**

A structured literature review covering five thematic areas- legacy automation limitations, AI testing capabilities and adoption, Lean management in software contexts, organisational change management, and existing QA frameworks provides the theoretical foundation for the LQMF. Sources were selected on three criteria: peer review was required, systematic reviews and mapping studies were preferred over single study empirical claims for factual assertions.

3.4.2 **Primary data: semi-structured interviews**

Seven semi-structured expert interviews constitute the primary empirical component of the research. Semi-structured interviews were chosen because they allow the researcher to ensure theoretically important topics are covered through a prepared guide while giving participants room to raise issues the guide had not anticipated, a balance that is particularly valuable in DSR evaluation, where the goal is not only to confirm the framework's assumptions but to discover what it may have missed(Hevner et al., 2004).

Participants were selected across two perspectives. Supply side of participants, 2 participants from AquilaTest AI testing platform who bring the knowledge of how migrations are designed and supported from the tool provider's vantage point. Demand side participants are industry practitioners whose expertise and experience of QA gave insight on challenge and tool transitions that supported the creation of LQMF. Each participant is described by role, experience, Interview Duration and perspective as described in Table 2. No individual or organisation is identified by name in the body text of the thesis.

Table 2. List Of Participants

S.No	Participants	Role	Experience	Interview Duration	Perspective
1	Participant 1/ P1	QA Team lead	10	43 min	Demand- Side
2	Participant 2/P2	SE Team Lead	20	45 min	Demand- Side
3	Participant 3/P3	Senior Specialist Test Automation	25	45 min	Demand- Side
4	Participant 4/P4	Software Tester	3	45 min	Demand- Side
5	Participant 5/P5	Co founder and VP Of AI testing tool	6	35 min	Supply-Side
6	Participant 6/P6	Engineering Team Lead	10	40 min	Demand- Side
7	Participant 7P7	CEO and Founder Of AI testing tool	6	50 min	Supply-Side

All interviews were conducted remotely via Teams and Google Meet which ranged from 40 to 50 min, and recorded with participants consent.

3.5 Data analysis: nvivo based thematic analysis

All the interview transcripts were imported into NVivo qualitative analysis software and subjected to structured thematic analysis. NVivo forces systematic engagement with every coded segment across all participants, and its matrix coding query function enables structured comparison of supply side and demand side responses on each theme, making the analysis auditable, replicable, and defensible.

The analysis followed an inductive deductive approach. Deductively, the seven LQMF phases provided an initial organising structure. The responses were according to which phase are confirmed , challenged or added to. Inductively, codes were also generated directly from data, allowing themes to emerge that the framework had not anticipated. Table 3 describe all 5 stages and information in this study.

Table 3. Five stage coding and their information

Coding Stages	**Info**
Import	Transcripts uploaded to Nvivo as individual Text sources with speaker label retained
Open Coding	Coding of all transcript provided around 91 codes covering challeneges, AI perceptions, adoption barriers and framework phase responses
Axial Coding	Related codes merged into 14 intermediate categories- eg locator break, flaky test under legacy automation pain points
Selective Coding	categories mapped against the seven LQMF phases to identify confirmation, challenges and additions mapped to 7 final themes
Verification	Nvivo matrix coding query cross verify the themes by participants role.

The process produced 91 codes and 7 parent codes which turned to 7 themes as shown in Table 4, each theme has direct implication for the LQMF's design and evaluation in chapter 5.2.

Table 4. codes and files based on Interview participants

S.NO	Parent code	Files	Codes
1	AI Adoption Barrier	7	22
2	Current AI Capabilities in Practices	4	17
3	Current QA & Automation Practices	5	22
4	Framework phase feedback	4	5
5	Legacy Automation Challenge	5	12
6	QA Role Evolution	6	9
7	Trust building Mechanism	4	4

3.6 Identified themes

Based on Parent code, we have clustered under higher level themes as showin in Table 5 to describe bigger trends seen among participants.

Table 5. Themes based on Code

S.No	Theme
1	Script Maintenance as Operational Waste
2	Cultural and Trust barrier to AI adoption
3	AI augumentation Within Current Maturity Limits
4	Strategic Shift in QA role
5	Risk Reduction through Phased Commitment
6	Evidence based Trust in Regulated Environment
7	Organisational scale as a Distinct Migration Challenge

At this point, the themes are addressed as observations, which are not yet comprehended regarding the framework.

3.7 Ethical consideration

The research has three ethical issues: informed consent and data governance, NDA governance, and bias.

The seven interviewees have given consent. They were advised of the research questions, and how their feedback would be used, that the interview would be audio recorded and they could withdraw from the research at any time, or request that particular statements were not used. An Interview data is safely stored and confidential to researcher. NVivo files (containing raw transcripts) cannot be disseminated.

Aquilatest interview recordings are under a Mutual Non-Disclosure Agreement (NDA). No confidential information (client names, finance, processes, technology) is included in the thesis. The NDA does not restrict academic use of practitioner feedback; it restricts the publication of commercial confidential information which is not needed for the research.

The supply-side participants have a potential commercial bias as founders of an AI testing Company. This is mitigated in two ways. First, the five demand-side participants are independent of AquilaTest and offer an independent view. Second, the NVivo matrix coding query was used to compare supply-side and demand-side responses on each theme, and highlighted differences rather than aggregates. These Perspective differs particularly on the current maturity of AI testing capabilities and these are reported in Chapter 5.

3.8 Evaluation criteria

(Hevner et al., 2004)were adamant about this: the criteria used for evaluating an artifact must be stated up-front, rather than defined to suit the outcome of the evaluation. Table

6. sets out the four criteria used to evaluate the LQMF in Chapter 5.4 and a very brief description of each criterion. The NVivo analysis makes a contribution to each of them

Table 6. Evaluation Criteria for LQMF

Criterion	What It Asks	How It Is Assessed in Chapter 6
Completeness	Does the framework cover all migration dimensions without leaving actionable gaps?	Benchmarking vs TMMi, Forsgren's Accelerate, and Kotter; NVivo-coded interview gaps
Coherence	Are the seven phases logically sequenced, with consistent theoretical grounding throughout?	Theoretical review; NVivo matrix query of phase-level practitioner feedback
Utility	Do the phases offer tangible steps, decisions, and outputs that a practitioner can actually follow?	Three migration scenario applications; NVivo-coded clarity feedback per phase
Generalisability	Does the framework hold across different organisation sizes, sectors, and maturity levels?	Scenario diversity; NVivo attribute variation across five organisational contexts

A note on what is not assessed: the LQMF is not assessed in terms of having been implemented in a real organisation. This is a valid way of evaluating it, but it requires access to the field over time, something that is beyond the scope of this thesis. The alternatives available - and those used - are theoretical benchmarking and expert review of the framework by practitioners; both approaches are recognised by (Hevner et al., 2004) as types of evaluation for a first-generation design artifact. Failure to implement in the field is discussed in Chapter 6.5 as a limitation.

4 The Lean QA Migration Framework

This chapter shows the complete Lean QA Migration Framework. Each of the seven stages is outlined their objectives, the activities they require, the decision gate between them and the next phase, and the outputs they produce. With every stage, the theoretical rationale of the particular stage, the design decisions taken is discussed.

4.1 Theoretical pillars of framework

The LQMF is based on four theoretical pillars that are not limited to a single phase but active in various phases. The pillars and how they apply at the phases are summarised in Table 7. The decision to use are four distinct theoretical bodies instead of one unified theory reflects the structure of the problem: QA migration involves in a waste diagnosis process (Lean), technology management (Toyota Way), cultural change (Kotter), and individual adoption (TAM/UTAUT). None of the four can be explained by any single theory. Attempting to force the framework into a single theoretical container would produce a less rigorous result, not a disciplined one.

Table 7. Theoretical Pillars of LQMF

Lean Thinking	**The Toyota Way**	**Change Management**	**Technology Acceptance**
Womack & Jones (1996)	Liker (2004) Principles 8,9,10,14	Kotter (1996)	Davis (1989)
Define value, map waste,create flow, establish pull,pursue perfection	test technology; grow leaders; develop people; become a learning organisation	Create urgency; build coalition;generate wins; anchor change in culture	Perceived usefulness, ease of use, social influence determine adoption
Applied in Phases 1, 2, 5,7	Applied in Phases 4, 6, 7	Applied in Phases 2, 4, 7	Applied in Phase 6

4.2 Framework view

The seven phases are presented in a sequence in Figure 2. The sequence is not arbitrary, it is a deliberate response to the specific failure modes that unstructured QA migrations typically exhibit. The majority of unsuccessful migrations start with the tool selection and do not diagnose what is wrong with the existing process (Phases 2 and 3 reversed), or they start to scale without proving its value in a controlled environment (Phase 5 attempted without Phase 4), or they invest in technology without investing in the people who are going to use it (Phase 6 omitted). The order is not intended to create a bureaucratic process, but rather to avoid these certain failures.

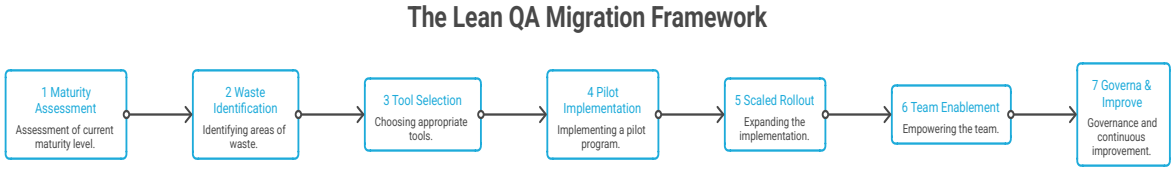

Figure 2. Lean QA Migration Framework

Although, The LQMF is shown as a sequence, more practical practitioners will be aware that actual migrations are not always going to follow a perfectly linear sequence. Phase 4 (pilot) occasionally divulges facts that lead to reviewing Phase 3 (tool selection). Phase 6 (enablement) occasionally brings cultural resistance that was not apparent in Phase 1 (maturity assessment). The decision gates that exist between phases are designed to slow that type of discovery and give a structured home, not to prevent it. The framework supports both within and between phases; but it resists skipping the phases.

4.3 Phase 1- maturity assessment

4.3.1 **Theoretical grounding**

Lean Principle 1 (Define Value): an organisation can never eliminate the waste that it has not yet defined. The results of the assessment are interpreted with the help of TMMi maturity levels (Garousi & Veenendaal, 2022)

The correct identification of the present position of the organisation is a step towards transformation initiative. Though this may seem self-evident, empirical results of the present study show that such systematic considerations are often disregarded in practice. In fact, organisations tend to embark on migration acts on the basis of assumptions, as opposed to facts driven approaches.

Participant 3, who has worked in testing with over 25 years experience and employed in SDET within one of the largest banking organisations in Finland, characterized a legacy testing environment, composed of a collection of various structures used by various teams of people over the decades, every team organises as they want and there is not a single way to do everything . Entering a migration in this kind of an environment without thorough evaluation is not going to speed up the process; in fact, this may further propagate the same inefficiencies in a new technological environment.

Phase 1 evaluates present QA maturity on five dimensions: definition of process and documentation; share of current testing that is automated and quality of automation; breadth and trustworthiness of CI/CD integration; team maturity - technical expertise and cultural willingness to change; and maturity of governance - quality measures are. as a collection, comprehension, and action. These dimensions are not measured against some abstract ideal but to the requirement of delivery by the organisation. A team that deploys multiple times a day require a different quality infrastructure as compared to a team that releases quarterly. The TMMi five- level scale offers a standardized reference by which the results are interpreted. Garousi and van Veenendaal(2022) discovered that

Level 3 (Defined) is the most widely used level of certification in the world. The majority of organisations that are thinking about a migration are at Level 2 to Level 3 - they have automation, but it is in patches, and processes are present but not necessarily used. Being aware of this, as opposed to making assumptions about it, defines the degree of ambition that the first pilot can plausibly have and the level of change leadership investment that the migration requires.

4.3.2 **Decision gate**

The team should be capable of answering the following questions before proceeding to the next step: Is the current maturity level understood in enough depth level to establish a particular and realistic target state? If not ,Has the evaluation has proven to be more complicated than supposed? continue exploring the issue by delving into the depths before transitioning to Phase 2.

4.3.3 **Phase output**

(1) Report of maturity assessment with evidence and rating covering five dimensions.
(2) Current-state baseline — what QA is currently delivering versus what the business needs it to deliver
(3) Post-migration QA function target maturity level, rationale based on delivery requirements.
(4) Preliminary migration risk register identifying the most likely failure modes of the particular organisation.

4.4 Phase 2-waste identification

4.4.1 **Theoretical grounding**

Lean waste taxonomy (Womack & Jones, 1996) adapted to QA automation contexts. (Hicks, 2007) on Lean Information Management to eliminate waste. (Poppendieck, 2007) about Lean software development. Waste analysis precede tool selection, not follows it.

It is quite natural and common at the beginning of the process to move the selection of tools. AI testing platforms are already appealing in nature: they are observable, can be easily compared, can be demonstrated, and they are constantly developing in a highly competitive environment. The process of waste analysis is relatively slow, less interesting and requires a critical and honest analysis of why the current system cost high. But it is also where the most initiatives of transformation end up failing before making any significant positive change. organisations that take on the AI testing platforms, before having even defined and analysed the specific types of waste that may contribute to inefficiency, have a high chance of making decisions based on the vendor positioning and marketing stories rather than on the contextual fit. This often lead to the platform it chose addresses the problem that do not even exist, leaving the most significant areas of inefficiencies unsolved.

With the aim to reduce this risk, the seven categories of Lean wastes are a structured analytical prism by which one may diagnose inefficiencies in QA automation situations. These categories, as operationalised in Table 8, are based upon the underpinning Lean literature, such as (Womack & Jones, 1996), (Poppendieck, 2007) and (Hicks, 2007). It should be noted that these frameworks were not initially designed to be used in QA automation environments, hence, its application in this study was by interpretive consideration in the mapping of the generalised waste constructs into domain orientated symptoms. The mappings thus made are the perceived meaning based on the empirical data of the interviews as well as available in the literature. Alternative

mappings are defensible and the table should be treated as an analytical starting point rather than a definitive taxonomy.

Table 8. Lean waste types mapped to QA context

Lean Waste Type	QA Manifestation	Severity	Consequence for Migration
Overproduction	Writing test cases that are never executed, or that cover functionality no longer present in the application	Medium	Accumulates test debt; inflates the maintenance overhead of already-burdened scripts
Waiting	Queuing for test environments, build completion, or peer review approvals before testing can begin	High	Delays feedback loops; slows CI/CD pipeline velocity; increases cost of defect detection
Inventory	Backlog of unmaintained scripts, obsolete test plans, and unexecuted regression suites that no one is certain are still valid	High	Creates false confidence in coverage; maintenance burden compounds with every sprint
Motion	Engineers switching between disconnected tools — test management platform, CI dashboard, bug tracker, spreadsheet	Medium	Context switching reduces analytical focus; increases error rate in manual tracking steps
Overprocessing	Running full regression suites for minor changes when risk-based or change-impact selection would adequately cover the risk	High	Execution time grows disproportionately; slows pipeline; wastes cloud compute budget
Defects	Escaped bugs reaching production due to inadequate coverage, flaky test signals, or unreliable environment behaviour	High	Highest-cost waste category; triggers regulatory obligations in governed sectors; damages user trust
Transport	Handoffs between QA, development, and product management teams without shared context, tooling, or traceability	Medium	Information loss at each handoff; duplicate effort; misaligned expectations across functions

Table 8 is supported by the practicality of the findings of participant interviews. Participant 1 (P1) who has more than 10 years of experience managing QA in various

fields described the Defects waste category in terms that go beyond inconvenience. Particularly, the constant maintenance of locator brittleness, which compromise the trust in automated testing. The automation suite can no longer serve as a useful decision-support aid when the perception of the test results as the reliable indicator of the system quality stops. When it happens like this, the actual worth of the automation investment is erased regardless of the magnitude or reach of executed test scripts.

”*UI keeps changing every release or every feature getting added, so the maintenance of these test cases are exhausting, you need to keep updating the elements and then keep checking it, verifying back and forth”(P1).*

Participant 3 spoke about how what is a moderately maintainable test suite on a small project becomes an overcomplicated nightmare once the suite gets large enough - and there is no clear line, it just increasingly fails to do what was intended. This theme provides empirical justification for the waste audit in LQMF Phase 2 and takes it beyond being abstract Lean terminology.

4.4.2 **Decision gate**

The team should be able to respond to the following prior to moving to the next stage: Have the top three categories of waste identified, discussed with the engineering team and provided with a preliminary cost estimation in engineering hours per sprint? Otherwise, Phase 3 tool choice will be maximising on the incorrect targets.

4.4.3 **Phase output**

(1) Waste identification register - all seven categories assessed for presence, frequency and severity ranking.

(2) Estimate cost on three type of wastes in engineering hours per sprint or per release cycle.
(3) List of priorities of the types of waste the chosen AI tooling needs to deal with to justify the cost of the migration investment.
(4) A common ground among the team members as to why migration is necessary - the waste register is the urgency document.

4.5 Phase 3- tool selection

4.5.1 **Theoretical grounding**

The phase 3 is based on The Toyota Way Principle 8 that implies adopting technologies that are proven, reliable and congruent with organisational capabilities and human systems. The principle can be supplemented by the evaluative criteria of AI testing tools suggested by (Garousi et al., 2024), emphasizing the critical evaluation of the tools to differentiate between the theoretical and operational performance. To be consistent with the thinking of Lean, the choice of tools cannot be made without the analysis of waste clearly informed by it.

The choice of AI testing tools in the context of the present market is always complicated, and an understanding of its complexity is a precondition in itself to make good decisions. A common trend, Garousi et al. (2024) in what is one of the more careful systematic reviews of AI-powered testing tools to date, found a consistent and troubling pattern: tool documentation routinely overstates the degree of autonomy the tool can actually deliver in production environments. Demonstrations features by vendor view of AI testing defined by a seamless end-to-end test generation directly out of requirement document, extensive self-healing at scale with zero maintenance, and zero false positives. Nevertheless, these representations seldom match with the realities faced by organisational deployment, whereby the tools have to work with legacy systems, limited environments, and diverse engineering practices.

This gap is supported by the practical experience of Participant 5 , an employee of an AI testing platform that is a customer success expert and has witnessed a lot of client implementation was explicit about this:

"*What is claimed to be happening under the name of AI testing in vendor demos (autonomous generation of massive end-to-end test coverage from a document that captures the requirements) is not what is happening in practice".*

What AI platform can do reliably in production today is create codeless test creation, synthetic data generator, faster configuration, Visual AI testing and reduced maintenance overhead through selfhealing etc.. These are practical and effective abilities. They represent an alternative value statement to complete autonomy, and the criteria by which one picks particular tools have to be dialed down to what the technology actually produce as opposed to what the marketing is claims to do. Undoubtedly a sure way to disappoint and abort adoption is to choose a tool, based on aspirational claims, and then apply it to a real codebase.

Phase 3 operationalises the choice of the tools based on the results of the Phase 2 (waste analysis) as the evaluation filter. The key point is not which platform has the broadest range of features, but which platform best meets the major categories of waste characterized above that are too cost-intensive. As an example, the locator maintenance can be the cause of the greatest loss of efficiency; in this case, the strong self-healing capability is the status quo, and then other features become secondary. On the other hand, synthetic data capability and evidence transparency are of greater importance in the highly regulated settings where test data generation and governance is of great concern.

Table 9 is a systematic mapping of the 9 selection criteria of the present situation of AI testing platform capabilities. The framework is informed with the help of systematic

review literature (Amalfitano et al., 2024; Fontes & Gay, 2023; Garousi et al., 2024; Karhu et al., 2025; Wang et al., 2024) and the insights that can be achieved based on the supply-side interview data. The table is not a vendor scorecard but an indication of what can be achievable in real-world practice by a state-of-the-art AI testing tool, and it is presented as a field-level feature benchmark.

Table 9. Phase 3 selection criteria

Selection Criterion	AI Core Capability	Relevance to LQMF Phase	Literature and Interview Grounding
Self-healing automation	Deployed	Directly targets Inventory and Waiting waste (Phase 2). Removes locator-breaking as the primary source of script maintenance overhead. Non-negotiable criterion when maintenance cost is the dominant waste category identified in Phase 2.	(Amalfitano et al., 2024): self-healing is the most practically significant AI testing capability stream. (Ricca et al., 2021): locator instability is the most cited automation frustration across grey literature. Participant 7: AI frees engineers to focus on thinking through what can go wrong rather than repairing broken scripts.
Codeless test creation	Deployed	Lowers the skill barrier to adoption, accelerating Phase 6 enablement without requiring scripting expertise. Reduces Motion waste by removing context switching between scripting environments.	(Garousi et al., 2024): codeless interfaces are among the few AI testing capabilities that have reached reliable production-grade deployment. Participant 5: codeless automation is the primary productivity gain for organisations without large automation engineering teams.
ML-based test case generation (edge cases, negative flows)	Partial deployment	Supplements manual test design. Addresses Overproduction waste by generating edge-case and negative-flow scenarios that engineers routinely omit under sprint pressure. Useful but not yet reliable enough to be a primary selection criterion.	(Fontes & Gay, 2023): ML-based generation improves test coverage and reduces manual effort, though maturity varies across test types and application domains. (Ayenew & Wagaw, 2024): transformer-based approaches show measurable quality improvement but require structured requirement inputs.

LLM-assisted test oracle inference and bug reproduction	Emerging	Supports Phase 6 capability development — builds analytical QA thinking rather than replacing it. Not yet sufficiently reliable for primary Phase 3 selection; a consideration for future framework revisions.	(Wang et al., 2024): LLMs demonstrate genuine utility in oracle inference and bug reproduction but hallucination and non-determinism risks complicate production integration. (Karhu et al., 2025): LLM-raised expectations are partly responsible for widening the aspiration-adoption gap.
Synthetic data generation	Deployed	Critical selection criterion for regulated-sector organisations (Phase 3). Addresses Theme T6 trust barrier — enables realistic test coverage without exposing real customer or patient data to the test environment.	Participant 5: synthetic data largely resolves the test data trust problem in banking and hospitality sector clients. Garousi et al. (2024): data management remains a persistent challenge in AI testing deployments; built-in synthetic generation reduces this friction significantly.
Risk-based test prioritisation	Deployed	Directly addresses Overprocessing waste (Phase 2). Enables targeted rather than full regression suites without reducing defect detection confidence. Accelerates CI/CD pipeline velocity and reduces compute cost per release cycle.	Garousi et al. (2024): risk-based prioritisation alongside self-healing is one of two AI testing capabilities with confirmed production-grade deployment across the tools reviewed. Amalfitano et al. (2023): test prioritisation is a distinct and growing AI testing capability stream.
CI/CD pipeline integration	Deployed	Enables Lean Flow in Phase 5 scaled rollout. Migration should not require introducing additional tooling layers. Absence of native CI/CD integration is a disqualifying criterion for organisations with established delivery pipelines.	(Ramos et al., 2025): successful AI testing adoption is consistently associated with deep workflow integration rather than standalone tool deployment. (Ozkaya, 2023): beneficial integration does not happen automatically and must be assessed explicitly during tool selection.

Execution evidence generation (logs, screenshots, video)	Deployed	Prerequisite for regulated-sector procurement in Phase 3. Provides the Phase 7 governance audit trail. Without auditable evidence of test execution, AI-generated results cannot be verified by compliance-governed stakeholders.	P5: trust in AI-generated test results builds when outputs include traceable evidence — screenshots, execution logs, and run records. Consistent with Theme T6 finding across all NVivo demand-side data. Garousi et al. (2024): evidence transparency cited as a key differentiator among production-ready platforms.
End-to-end autonomous test generation from requirements	Not at production maturity	The capability most organisations describe when asked what they want from AI testing. Currently not a realistic primary selection criterion. Platforms claiming full autonomous generation from a requirements document are overstating their current capability relative to what production deployments deliver.	Garousi et al. (2024): autonomous end-to-end generation remains aspirational in most production contexts. P5: 'that level of maturity has not yet been reached — AI currently augments tester judgment rather than replacing test design.' Karhu et al. (2025): the gap between this aspiration and actual adoption is the defining challenge of the 2025-2026 period.

The nine criteria give the difference in the maturity in the capability landscape. Some of these criteria reflect stable features that are probably scaled and in production. Others are partially developed features with some performance and reliability differing drastically depending on platform, implementation situation and line of application. Notably, one of the criteria, such as autonomous end-to-end test generation based on requirements, will be aspirational. Despite its prominence in vendor narratives, current evidence indicates that this capability has not yet achieved consistent, production-level maturity at scale. The last row is included deliberately. Any evaluation of a tool that describes only what the tool can do and not what the tool can do at any point in time is not a selection framework it is a sales document, and organisations who rely on one to make their procurement decisions will pay that optimism during deployment.

Two practical notes on how to conduct Phase 3. To begin with, any tool assessment that lacks a proof-of-concept in the organisations actual environment - against its actual test suite, connected to its CI/CD pipeline - is incomplete. Vendor demonstrations are always under optimal conditions with prepared data and with a maintained codebase. The only means of evaluating how a platform will perform in a legacy environment with inconsistent test structure is to run that platform in a legacy environment under controlled conditions, before committing. Second, the assessment should be done with a clear understanding of the observation made by Participant 2 concerning the velocity of the tool landscape:

"*If you pick a tool today, it might be something that you don't actually need anymore in three months because something better has appeared, or it has been deprecated, or in the worst case your whole need is gone.*"

It does not mean that we should wait indefinitely before making a choice of tools. It is a plea to choose on the basis of what the tool can achieve in the current waste categories in Phase 2, and not based on aspirational roadmaps or projected future capabilities.

4.5.2 **Decision gate**

The team should be in a position to respond to the subsequent question before advancing on to the next stage:Does the tool you have chosen show demonstrably that it will help in solving the top three waste points identified in Phase 2? Have you developed a proof-of-concept in the environment of the organisation (not a vendor sandbox) before Phase 4?.

4.5.3 **Phase output**

(1) Weighted tool evaluation scorecard with criteria derived directly from Phase 2 waste analysis
(2) Proof-of-concept report - tool tested on a representative selection of the existing test suite in the real world.
(3) Chosen tool with recorded reasons of its choice, as well as recognized current limitation.
(4) Initial integration plan — CI/CD touchpoints, team access configuration, licensing structure, data privacy assessment

4.6 Pilot implementation

4.6.1 **Theoretical grounding**

Toyota Way Principle 8 (test technology properly before going at large scale). (Kotter, 20) Step 6: create and share short term wins. The initial step of the change management process is not a technical dress rehearsal, but a pilot.

P2 explained the tool selection limitation of most medium and large organisations: no specific internal ability to compare a variety of tools at once, and the risk of choosing the inappropriate platform, in terms of sunk time, disrupted workflows and lost team confidence, is not commensurable. Therefore, a structured pilot prior to the commitment to the organisation is not an optional methodological choice. It is the sole viable way of imparting the decision risk to a level where commitment can be made possible.

"There is no internal expertise to trial multiple tools at once and the implications of selecting the wrong tool are significant. (Participant 2)"

However, when referring to the pilot as a risk reduction mechanism, we undersell it. The initial step of a learning loop is the pilot in the Lean sense. It produces data that cannot be replaced, on the actual behaviour of the tool within the particular context of the organisation: what types of tests are capable of self-healing and what generate false positives, where CI/CD integration causes unforeseen friction, how long an engineer

new to the platform takes to become independently productive, what documentation is lacking, and most importantly, who are the engineers who turn into natural advocates and who remain sceptical.

Kotter (1996) identified the creation of short-term wins among the most significant requirements of maintaining organisational change in the long and often discouraging middle section of a major transition,. A pilot that takes eight to twelve weeks to run on a limited scope - one product area, one team, one discernible slice of the current test suite - and induces visible, measurable maintenance time reduction or defect detection or pipeline acceleration provides precisely the type of evidence that the leadership and sceptical engineers require. Leadership needs to have a good investment direction; engineers need to have a good sense that the new direction will help them in their daily work and not to put their expertise aside.

The QA Competence Lead of a high-growth technology company, P6, made an observation concerning the planning aspect of Phase 4 that is easy to overlook:

“If you can't explain it in the paper and pen to our fellow new colleagues — what the test is actually doing and why — then robots can execute but they can't be very good at executing if you give them poor instructions. (P6)"

The implication is significant. Running a good pilot is not only about configuring the platform and measuring execution times. It is about forcing the engineering team to think more carefully about what they are actually trying to verify, and why. That conceptual discipline is not a by-product of a well-run pilot ,it is one of its primary contributions. Teams that emerge from a good pilot do not only have a configured AI testing platform; they have a sharper collective understanding of what their tests are for.

4.6.2 **Decision gate**

Before proceeding to the next phase, the team must be able to answer:

Did the pilot produce measurable, documented improvement on at least two of the Phase 2 waste categories? Do at least two engineers who participated in the pilot feel confident enough with the platform to train colleagues in Phase 6?

4.6.3 **Phase output**

(1) Pilot evaluation report — quantified results against Phase 2 waste categories; honest assessment of where the tool underperformed
(2) Friction log — documented list of integration problems, configuration gaps, and unexpected tool behaviours encountered during the pilot
(3) Updated migration risk register incorporating the specific failure modes revealed by the pilot
(4) Champion engineers identified, the people whose confidence and advocacy will anchor Phase 6 enablement

4.7 Phase 5- scaled rollout

4.7.1 **Theoretical grounding**

Lean Principle 3 (Make Flow): value must flow without batch processing, waiting, or interrupting. Lima et al. (2023) on socio-technical aspects of digital transformation Lean implementation.

In principle, scaling a successful pilot is the easiest aspect of the migration. As a matter of fact, it is commonly where migrations come to a halt. This is not normally a technical reason, but an organisational one. The description of the testing environment in the banking institution given by P3 provides the best example of why enterprise scaling is a problem in itself. The organisation has hundreds of testers spread out across dozens of product teams, some of which have a decade or fifteen years of experience with the same testing method. What proved effective in the pilot - one team, one product area, a fairly modern codebase - may not necessarily be re-used in another team in a different part of the same organisation with different technical constraints, different team norms, and a different relationship with the QA function. At some point, the dependency

mapping among the teams and coordination of the migration sequencing become real project management problems, rather than technical.

Lima et al. (2023) suggested that Lean implementation in digitally transforming organisations should consider the socio-technical aspects of the process, which are the technical configuration options regarding which teams should migrate first, in what sequence, and on what timeline; and the organisational decisions on who owns the migration in each team and what support is available when things go wrong. The Phase 2 waste analysis should determine the rollout sequencing and not by convenience or political presence. Groups with the greatest maintenance load and the greatest desire to change should migrate the sooner - since the change will be most conspicuous there, and conspicuous improvement will support leadership commitment and interest in colleagues well into the middle section of the rollout.

Lean Flow here refers to developing a driving, flowed-out migration that retains the teams flowing across the phases without leaving huge waiting line or requiring more support capacity internally than is presently available. A well structured rollout is self-reinforcing: migrated teams themselves turn into trainers and advocate for team that are about to migrate

4.7.2 **Decision gate**

Before moving to the next step, the team should answer: Are the rollout sequence motivated by waste priority and not organisational convenience? Is there sufficient internal support capacity such as trained engineers, documented solutions, and clear escalation paths , to sustain each wave of adoption before the next begins?

4.7.3 **Phase output**

(1) Rollout roadmap - team-by-team sequence of migration with a timeline, dependencies, and support assignments.
(2) Updated integration documentation adjusted to the technical situation of each new team.
(3) Adoption metrics dashboard CI/CD pipeline time, defect detection rates, maintenance time.
(4) Wave-by-wave issue logs into Phase 6 training content.

4.8 Phase 6- team enablement

4.8.1 **Theoretical grounding**

The Technology Acceptance Model informs Phase 6 by the view that perceived usefulness and perceived ease of use are the key determinants of technology adoption as opposed to the sole influence of technical capability. The Unified Theory of Acceptance and Use of Technology extends this view to incorporate social influence and facilitating conditions as other determinants of adoption behaviour. To supplement these frameworks, Toyota Way Principles 9 and 10 are about developing leaders and teams who not only know how systems operate, but why they are created in the way that they are.

Training and enablement are part of the seven stages of the framework, but this area is least commonly taken seriously in practice. Although the process of procuring the tools usually guarantees the financial investment and the process of integrating CI/CD integration is typically awarded with specific engineering work, the element of training is often diminished to a minimum intervention that is usually a brief workshop with accompanying documentation due to the premise that the adoption will come naturally. The interview information constantly questions this vision, suggesting that this is not a way to generate meaningful and lasting adoption.

Participant 7 recognised organisational mindset as a greater impediment to technical integration issues. The aversion towards AI testing tools is usually not based on a reasonable analysis of technical inferiority; on the contrary, it is anchored in undermining the traditional professional identities. Experienced engineers who have widely work in structures like Selenium or Robot Framework might see the emerging tools as weakening their acquired knowledge. The resistance of this type cannot be resolved by only feature demonstrations and procedural training. Rather, it fades away over time and practitioners can find that the new tools transform their jobs in a way that does not eliminate their skills but instead leads back to more valuable types of work, less about maintenance-focused work and more about analysis

"People have to understand that they have to scale up. Now I can focus on more interesting things — thinking through what can go wrong — rather than repairing the same scripts every sprint. (P7)"

The proposed observation is in line with the main postulations of the Technology Acceptance Model. Perceived usefulness and perceived ease of use are economies, not intrinsic characteristics of a technology, but constructed socially by experience and context. An application that has proven to be valuable in a pilot context, especially one that is enshrined by trusted colleagues, would be more prone to be viewed as helpful. Conversely, any tool presented by top-down requirements together with prescriptive training resources will have a lower chance of attaining the actual acceptance. The UTAUT model also supports this understanding, suggesting that social influence plays a critical role: adoption is much faster when the people note the successful adoption of new tools by respected peers in the workplace.

Here, the discovery of champion engineers in the previous stages serves as a structure mechanism of facilitation of adoption. Their work is not just a technical cut and paste

exercise but by showing practice and viable advocacy they establish that social context to which adoption becomes self-reinforcing instead of coerced.

Liker Principles 9 and 10 offer another insight with regards to the significance of profound organisational knowledge. Good progress can be sustained when those in leadership have a true understanding of what they are leading, and group members will be aware of the reasoning behind process and technology decisions. When applied to Phase 6, this means that training should go beyond training on procedures to the level of training concepts. Practitioners must understand why AI testing platforms function as they do: for example, how self-healing mechanisms infer application structure, the current limitations of NLP-driven test generation, and the role of synthetic data in ensuring compliance and trust in regulated environments. Participant 6's observation applies directly:

"If you put more effort into the planning phase — if you can explain in plain language, with paper and pen, what the test is actually doing — then the robots can execute it well. The quality of the instruction is what limits the quality of the execution."

In line with this, the goal of Phase 6 is not only to become more proficient with the chosen AI testing platform, but to become more competent in general on reasoning about quality. Training programmes must teach a better comprehension of what meaningful testing is in the organisational context, to enable the practitioners to develop more useful, strong and value-oriented test strategies. In this respect, the effective enablement can be signalled not only by the metrics of the use of tools themselves, but by the apparent change in the way the teams conceptualise and engage in quality assurance.

4.8.2 **Decision gate**

The team needs to respond to the following one before moving to the next stage: Can atleast 80% of affected engineers write, run and interpret AI-generated tests without resorting to an expert? Have the Phase 4 champion engineers supervised at least one internally enabled training?

4.8.3 **Phase Output**

(1) Role-differentiation training programme - various depth and focus with QA engineers, developers and engineering managers.
(2) Internal knowledge base -, documented solutions to the identified friction points in Phase 4 and in wave of rollout in Phase 5.
(3) Champion engineer network - formally recognised internal champions with time allocated to serving colleagues.
(4) Adoption health measures - frequency of use of the tool, escalation rates , peer confidence measures, voluntary adoption by choice without mandating team.

4.9 Phase 7- governance and continous improvement

4.9.1 **Theoretical grounding**

Kaizen - Continous Incremental improvement as a habit of the organisation. Toyota Way Principle 14: be a learning organisation by continuous reflection and Kaizen. Kotter (1996) Step 8: influence the change of culture in such a way that the change lasts longer than the migration project.

Majority of migration projects have a termination date. The LQMF does not. The discipline of Phase 7 is the difference between a successful migration and a durable improvement. This is because failure mode that was noted by Kotter (1996) stating that

victory has to be declared prior to the integration of new behaviours in the system is more prevalent in technical migrations. The pilot was successful, the rollout, the tool is deployed and the project is closed. And then, slowly, and with no one ever realizing the process is occurring, engineers would begin to work around the new system where it forms friction, governance reviews are overlooked when there is sprint pressure, and the waste that the migration is meant to eliminate begins to creep back. Not a metaphor, Kotter anchoring is the eighth step of change. This is a particular prerequisite to make sure that the practices the migration introduced are dealt with as normal operating procedure rather than a special project.

Kaizen holds this by establishing an expectation, rather than a procedure. Continuous improvement is not an annual process of the organisation that fits into a cycle of strategic planning. Every Engineering team does it, in little forms, as it is a usual aspect of their functioning.In QA terms, this entails a perspective specifically looking at the testing process - not product, not sprint velocity, but testing process itself - with the question: when did waste renetered into the system since we last checked? The proposed waste review of quarterly suggested in.

Phase 7 reiterates the Table 8 matrix on the present state. The review of the annual framework inquires of the LQMF phases themselves whether the team still works as such, or not, and amends them accordingly. An inflexible system which may not be updated with practice is not a framework, it’s a constraint. The description of the QA activity given by Participant 6, is directed on what sustainable governance would appear to be like in practice:

“QA engineers are working like assistants in a way — one person embedded across multiple teams, helping with testing strategy, helping with tooling, pair testing, pair programming. The quality responsibility is distributed rather than centralised. (P6)"

This distributed model change reminds what governance means. When the QA role is centralised, it is easy to govern: the QA team owns the quality and has liability over quality. A distributed model means that governance should be oriented towards developing and maintaining the quality capacity of teams that lack a full-time QA engineer - which requires a different form of oversight, which is focused on thinking quality and analytical skill instead of the performance metrics and number of defects.

4.9.2 **Decision gate**

The team needs the ability to answer the following before the next stage: Is the work improvement integrated into the team's working rhythm - sprint retrospectives, quarterly reviews - rather than managed as a separate project with its own timeline .Does each active improvement initiative have a named owner?

4.9.3 **Phase output**

(1) Quarterly waste review report - Table 8 re-applied to current state; new types of waste identified and prioritised.
(2) Practice board - visible, active record of ongoing improvement efforts that have named owners, schedules, and measures of results.
(3) Annual LQMF review - record changes in phases of the framework where practice has evolved beyond the description of the current stage.
(4) Long-term quality metric baseline - multi-quarter showing the reduction in the burden of maintenance, reduction in defect detection, and increased pipeline velocity are taking hold.

5 Evaluation and Discussion

In this chapter, the Lean QA Migration Framework is analysed according to the four requirements outlined in Chapter 3 completeness, coherence, utility and generalisability. It uses three sources of evidence: empirical data on the NVivo structured analysis of seven expert interviews, benchmarking to familiar QA and change management models, and the LQMF applied to three typical migration situations, which encompass the spectrum of organisational situations, a variety of which are portrayed in the interview data. The chapter ends by discussing what the findings, collectively, imply about the place where the framework is strongest and weakest, how the research fills a gap in the existing literature, and the locations of the greatest limitations.

5.1 Empirical finding- nvivo matrix analysis

The NVivo analysis led to the emergence of seven themes out of 91 initial codes in 7 intermediate categories. Those themes were named and described in chapters 3 and 4; this section explores what the distribution of references across participants and perspectives can tell us about the reliability and extent of each theme. Two figures are displayed: Figure 3 displays the number of references by each theme per individual participant; Figure 4 displays the supply-side versus demand-side comparison in aggregate. Both characters are based on the actual matrix coding query results in NVivo.

Topic	P1	P2	P3	P4	P5	P6	P7
AI Adoption Barrier	2	4	1	4	2	4	4
Current AI Capabilities	4	0	4	0	6	0	3
Current QA & Automation Practices	4	4	2	4	0	5	0
Framework phase feedback	1	0	1	0	0	2	1
Legacy Automation Challenge	6	2	1	2	0	1	0
QA Role Evolution	2	2	1	1	1	2	0
Trust building Mechanism	1	1	1	0	0	1	0
0- Minimum 1-3= Moderate 4-6= Strong							

Figure 3. NVivo Matrix Coding Query. Reference counts represent the number of coded passages per participant per theme. P1–P7 correspond to the participant profiles in **Table 2**. Colour coding: Blue = strong (4+ references); Yellow = moderate (1–3); grey = no references.

The first visible pattern in Figure 3 is The concentration of zeros among supply-side participants, on these themes it is indicative of the design of the supply-side interviews rather than lack of awareness. The supply side interviews were based on the capabilities of the AI platform, the dynamics of adoption, and the experience of customer migration - the terrain P5 and P7 was in the best position to do the interviews. Interview design questions that focus more on the day-to-day experience of having to maintain legacy scripts and have their position as the QA role were primarily directed to demand-side participants whose occupational context places them in a more direct position to answer the questions about that experience.

The practitioners on the demand side produced eleven mentions of Legacy Automation Challenge and fourteen to Current QA and Automation Practices with no prodding to those two themes in the supply-side interviews, is nevertheless analytically important. The vividness, the specificity with which P1, P2, P3, P6, and others explained the maintenance burden - the tiring locator repair cycles, the Inventory of scripts the validity of which cannot be determined, the gradual loss of team faith in the automation package - indicate a set of problems are being experienced on the practitioner side and are not naturally coming to the surface in research that faces vendors. This has a first-order implication into the framework: the insistence of the LQMF on Phase 2 (Waste Identification) preceding Phase 3 (Tool Selection) is a priority that is most visible on the demand side and is therefore underweighted structurally, in advice generated primarily through vendor or market-analyst lenses.

The members of the demand-side group, P1 (including six references) is the highest number of references per participant on any theme. This is the profundity as well as the straightforwardness of the story of P1 how to manage the enterprise QA pipelines in numerous fields over more than a decade, and a specific and quoteable account of the

locator maintenance cycle. P2 and P4 both fulfill four references on AI Adoption Barrier, the highest number of demands on the theme in both directions, at very different angles: P2 through the prism of tool commitment anxiety in a large retail organisation, and P4 through the experience of implementing automation as something new in a small consultancy environment.

Matrix By perspective

	A : Interview Participants:Perspective = Demand Side	B : Interview Participants:Perspective = Supply side
1 : AI Adoption Barrier	11	6
2 : Current AI Capabilities	8	9
3 : Current QA & Automation Practices	14	0
4 : Framework phase feedback	2	1
5 : Legacy Automation Challenge	11	0
6 : QA Role Evolution	6	1
7 : Trust building Mechanism	3	0

Figure 4. NVivo Matrix Coding Query- Reference Counts by Participant Perspective

Figure 4 provides three analytical meaningful patterns. To start with, there are two themes that are only represented in the demand-side data: Current QA and Automation Practices (14 demand-side, 0 supply-side) and Legacy Automation Challenge (11 demand-side, 0 supply-side). This distribution is that these themes are used to capture lived practitioner experience as opposed to externally constructed or strategically framed narratives. They mirror the operational reality and constraints that characterize the problem space that the Lean QA Migration Framework (LQMF) is designed to address. The fact they are not included in supply-side data is analytically significant: it implies that vendor-focused viewpoints do not systematically interact with the underlying inefficiencies of the existing QA systems. That is why, in many cases, the vendor advice to adopt AI testing only starts at a later stage, i.e., at Phase 3 (tool selection), not at Phase 2 (waste diagnosis). In practice, vendors pull toward positioning their solutions, whereas the LQMF is explicit grounded on problem diagnosis.

Second, four themes, namely, QA Role Evolution (6 demand-side, 1 supply-side), Framework Phase Feedback (2 demand-side, 1 supply-side), Framework Phase Feedback (2 demand-side, 1 supply-side), and AI Adoption Barrier (11 demand-side, 6 supply-side) are more on the demand side but also include a small amount of supply-side representation. One of them, in particular, is the theme of AI Adoption Barrier. Both the demand-side and supply-side actors come to the same conclusion of identifying mindset and trust as the key barriers to adoption, but conceptualise these barriers differently. Participant 7 (P7) presents the issue as one of organisational preparedness and scaling, highlighting the fact that the human aspect of adoption requires as much attention as the technical integration. Participant 2 (P2) on the other hand describes the barrier as a commitment risk, which is determined by the fact that the AI tooling environment is rapidly changing and that the uncertainty on making the decision that may become obsolete in a short period is a lot. These views are not opposing, but they reflect complementary interpretations of the same underlying phenomenon, at different organisational levels and time perspectives.

Third, there is only one theme, Current AI Capabilities (8 demand-side, 9 supply-side), that is widely shared in both datasets. This overlap indicates that there is a relatively stable comprehension of the current strength of AI testing tools by participants with firsthand experience with such tools. Another slight yet analytically significant difference can, however, be found inside this common thread. Participant 5 (P5) makes a technically based assessment, clearly stating that autonomous end-to-end test generation based on requirements is not as mature as it can be at production level. Participant 7 (P7), on the other hand, frames in a more productivity-focused way, highlighting the incremental efficiency benefits enabled by the current tools. Both descriptions hold true, but they prefigure different areas of the landscape of capability. The matrix query displays this divergence without trying to harmonize it, thus maintaining the delicacy of the information. It is interesting to note that the Phase 3 tool selection criteria in LQMF is calibrated to match P5 more closely, in terms of its technically specific and constraint-aware view.

Another requirement is needed based on the nature of demand-side AI experience embodied in this theme. Some demand-side participants noted involvement with AI tools in some form or other, usually in the most accessible and individual-level forms, such as conversational AI systems (e.g., ChatGPT, copilot) to write test scaffolding, or code assistance tools (e.g., GitHub Copilot) to generate test scaffolding. This kind of use signifies productivity enhancement and improvement on an individual level rather than implementation of fully integrated AI testing systems at the organisational scale. Specifically, participants P2 and P4 spoke about the interactions with AI that do not require the central QA infrastructure. This difference is analytically important: although these experiences offer plausible information about the capabilities of AI, they do not represent the realities of complexities related to the implementation of AI at enterprise-level, such as integrating a system, requiring governance, and the coordinated change management at the team-level.

The given observation is consistent with the results provided by Wang et al. (2024) since they observe that since the large language model (LLM)-based tools became widely accessible, it has led to the increased familiarity and anticipation of the tool among the practitioners; yet this observation is accompanied by the lack of providing the corresponding exposure of the large-scale deployment issues to the practitioners. Consequently, the gap between perceived capability and operational reality does exist and therefore it must be explicitly accounted both in the analytical frameworks and implementation strategies.

5.2 Interview themes mapped to phases

5.2.1 **Script Maintenance as operational waste(phase 2)**

The legacy automation challenge theme generated 11 demand side references , the highest demand side references compared to other. All demand-side participants

referred to locator maintenance a problem. Participant 1 characterized updating scripts for each user interface (UI) update as exhausting

"UI keeps changing every release or every feature getting added, so the maintenance of these test cases are exhausting, you need to keep updating the elements and then keep checking it, verifying back and forth".

Participant 3 spoke about how what is a moderately maintainable test suite on a small project becomes an overcomplicated nightmare once the suite gets large enough - and there is no clear line, it just increasingly fails to do what was intended. This theme provides empirical justification for the waste audit in LQMF Phase 2 and takes it beyond Lean terminology. The waste register which is generated in Phase 2, is not the result of an academic exercise, it is the evidence base, that makes the case of migration within the organisation and it is the specification document which defines what the migration must deliver.

5.2.2 **Cultural and trust barrier to ai adoption(phase 6)**

With 17 total references across all seven participants, AI adoption barrier is most discussed theme among demand and supply side. The supply-side participants both observed that the challenge in adopting AI is not using the tool - it is getting people to use the tool.

Participant 7: "*observed that adaptation will take time and that resistance comes from discomfort with uncertain territory rather than rational objection to the technology itself".*

Participant 5 characterised organisational reluctance in the regulated financial and health care industries more as a fear of giving up control than a conscious calculation of risk. The demand-side participants provided another perspective: Participant 2

“a tool environment that was evolving so rapidly that it was a risk in mid-life to get committed to anyone them”.

All three categories of resistance directly correlate to Phase 6 design choices: enablement ordered around proven usefulness before demanded commitment, champion engineers building social proof and a pilot phase that resulted in the evidence that replaces commitment anxiety with warranted confidence.

5.2.3 **AI augmentation within current maturity limits (phase 3)**

This theme (8 versus 9references) can truly be seen as a convergence point of this current state of AI testing capability. Participant 5 articulated this direct statement of maturity gap

“What is claimed to be happening under the name of AI testing in vendor demos (autonomous generation of massive end-to-end test coverage from a document that captures the requirements) is not what is happening in practice”.

Instead, they are achieving significant aided productivity improvement: codeless test generation, synthetic test data, rapid configuration, low test maintenance through self-healing. Aligning with what AI testing platforms can offer with demonstrative accuracy today, meaningful productivity augmentation, generates more realistic selection criteria, and provides more lasting post-deployment satisfaction.This discovery directly influences Phase 3 of the LQMF.

5.2.4 **Strategic shift in the qa role (phase 6)**

The role of QA Role Evolution has produced 6 demand-side references and 1 supply-side reference. P6 gave the best and most quoteable version of the change:

"If we can't explain it in paper and pen to our fellow new colleagues — what the test is doing and why — then robots can execute but they can't be very good at executing if you give them poor instructions. The planning phase is where the real work is now. "

P5 ascertained the same of the progression of the supply-side approach: AI testing will not eradicate the need to have QA expertise - it will simply change its form, shifting it towards the test design thought process and quality strategy. The implication in relation to Phase 6 is that the training programme should instill a critical thinking with regard to quality, and not just operational proficiency in the new platform. Engineers that descend through Phase 6 to have the privilege of knowing what their tests are and why such tests are important would get more out of an AI testing platform than engineers that can navigate the interface but are unable to explain the purpose of their tests.

5.2.5 **Risk reduction through phased commitment(phase 4)**

Participant 2's description of the problem in tool selection was eloquent:

“there is no internal expertise to trial multiple tools at once and the implications of selecting the wrong tool are significant”.

In this context, an pilot phase before final commitment is not an option - it is the only reasonable way to reduce risk to a level at which commitment is possible. The pilot is the mechanism through which organisations can shift off the informed commitment anxiety that P2 elaborated and into the evidence- based confidence that would render full migration possible. Eliminating Phase 4 of the process and moving directly through tool selection to scaled rollout is the same thing as eliminating the only source of information that can answer the most critical question: will this work in us, in our environment, with our team?

5.2.6 **Evidence based trust in regulated environment(phase3)**

The explanation of what regulated-sector clients need to achieve before they can adopt AI testing tools, as provided by P5, explains a selection criterion that does not emerge in an otherwise purely technical evaluation. Even the ability question is not whether the platform can generate correct test results - it is whether it can generate auditable evidence that it generated correct test results in a specific run, and by a specific test configuration, and against a specific application state. synthetic data generation and execution evidence (screenshots, logs, traceable video) is the prerequisite feature for banks, healthcare for procurement. In Table 9 of Phase 3 includes both criteria that explicitly state this reason.

5.2.7 **Organisational scale as a distinct migration challenge**

The most transparent example of an instance of why the scaling assumption that is implicit in most migration frameworks is given by P3 concerning the testing environment of a banking institution, in which even 1990s legacy software still in active production, hundreds of testers, a Lego-piece modular tooling philosophy assembled on the need side. It is assumed that what has been successful in terms of team can be replicated in terms of more resources on an enterprise scale. In the account given by P3, at some level of complexity, the coordination problem turns the issue completely:

“The best teams are teams with developers and testers co-located, working together. But at the scale that we work at, you cannot assume that. Each team has its own legacy, its own setup. The migration problem is not the same problem across all of the teams. (P3)"

This has a tangible implication on Phase 5: in a large organisation, the rollout roadmap may not be able to assume that the migration of each team is going to be a simple replica

of the pilot. It should consider each team as a bounded context and having a waste profile of its own, its own technical constraints, and its own change management needs.

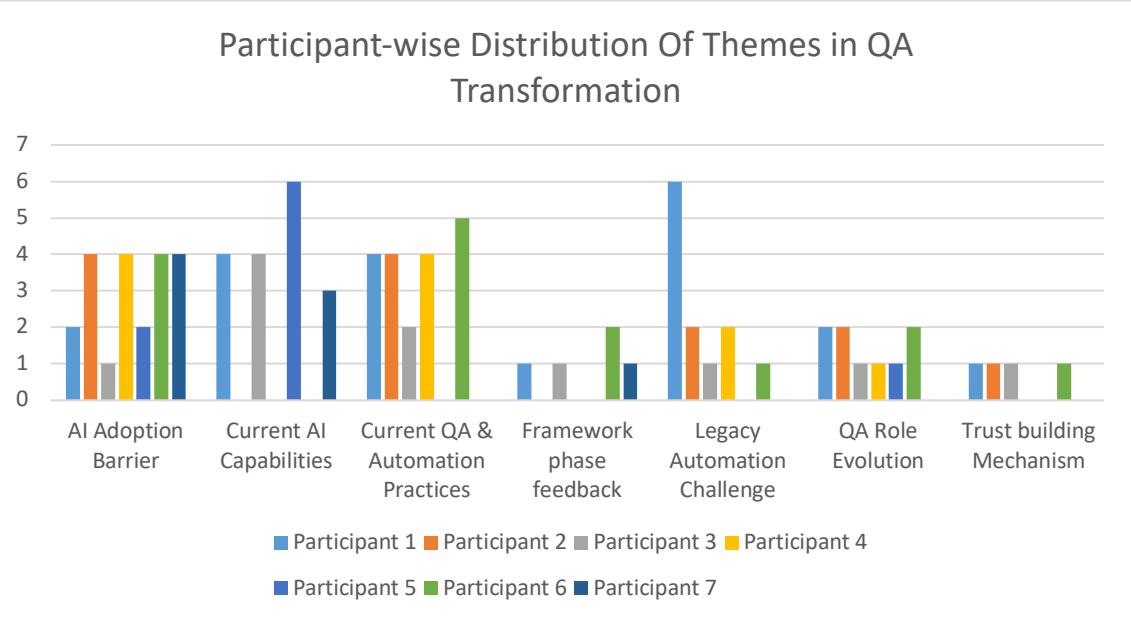

Figure 5. Participant wise distribution of themes

5.3 Benchmarking against established framework

Table 10 compares the LQMF to four existing models that cater to either the management of QA transformation or the management of the organisation change. The rationale is to evaluate the completeness criterion; whether the LQMF addresses dimensions that existing framework have identified as important, and whether it occupies a space that is genuinely underserved.

The benchmarking indicates a pattern which is analytically consistent with the gap in research identified in Chapter 2. Each of the established frameworks is either too broad in scope (CMMI), too narrow in focus (Forsgren et al., 2018) or offers organisational change guidance without the process and technology-specific content of a migration that a migration requires (Kotter). The nearest complement is TMMi, which defines the

level of maturity with precision and even with a common language to describe the capabilities of the QA process. TMMi can be used by an organisation to calibrate (current) maturity foundation and determine a realistic target state, and then manage the transition between the two points using LQMF.These frameworks are complementary to each other rather than competing.

Table 10. LQMF against the four criteria

Framework	What It Offers	What It Leaves Unaddressed	How the LQMF Relates
TMMi	Five maturity levels for testing processes; Level 3 (Defined) is the most common globally; strong on process definition and measurement	Defines destination states but provides no structured transition path between them; technology-neutral; no AI testing guidance	LQMF Phase 1 uses TMMi-aligned maturity assessment as its baseline; Phases 2–7 provide the migration journey that TMMi defines the destination for but does not describe how to manage
Forsgren et al. (2018) Accelerate	Quantifies the link between test automation, deployment frequency, and change failure rate across 23,000+ organisations. Identifies automated testing as a key capability of high-performing DevOps teams. Strong on performance metrics and CI/CD pipeline architecture.	Describes what high DevOps performance looks like rather than how to migrate from legacy automation to achieve it. No waste diagnosis, change management framework, team enablement model, or post-migration governance guidance.	LQMF provides the migration path that Accelerate implicitly requires but does not describe. Phases 2 and 3 address the automation modernisation that Accelerate's metrics assume is in place; Phases 6 and 7 address the human and governance dimensions its framework omits entirely.
Kotter's 8-Step Model	Comprehensive organisational change sequence covering urgency, coalition, vision, communication, wins, consolidation, and culture anchoring	Generic change management model not tailored to technology migration or QA-specific dynamics; provides no waste diagnosis or tool selection guidance	LQMF incorporates Kotter's logic throughout: Phase 2 creates urgency, Phase 4 generates wins, Phase 7 anchors change. LQMF adds the QA-specific technical and process content Kotter does not provide

CMMI (Capability Maturity Model Integration)	Broad process improvement framework covering development, services, and acquisition; recognised across global enterprise procurement	Too broad for QA-specific application; no AI testing guidance; improvement roadmap not migration management framework	LQMF is narrower and more specific — designed for QA technology migration — and operationally more detailed, with concrete activities, decision gates, and outputs per phase

5.4 Formal evaluation against the four criteria

Table 11 gives formal evaluation of LQMF in relation to the four criteria in Chapter 3 with supporting evidence and rationale provided as to the evaluation

Table 11. LQMF formal evaluation against four criteria

Criterion	Assessment	Evidence and Reasoning
Completeness	**Satisfied**	All four necessary migration dimensions are addressed: process diagnosis (Phase 2), technology selection (Phase 3), change management (Phases 4 and 6), and post-migration governance (Phase 7). The NVivo Framework Phase Feedback theme confirmed no participant identified a dimension the framework omits. The enterprise scale limitation of Phase 5 is acknowledged as a design boundary rather than an unacknowledged gap.
Coherence	**Satisfied**	Each phase produces output that directly serves as input to the next phase. The waste analysis in Phase 2 defines the Phase 3 evaluation criteria; the Phase 3 tool selection defines what Phase 4 pilots; the Phase 4 pilot produces the evidence that Phase 6 uses to build practitioner confidence. The sequencing reflects Lean's diagnosis-before-intervention principle and the TAM insight that perceived usefulness must be demonstrated before adoption can be expected. Practitioner feedback confirmed the sequencing logic rather than challenging it.
Utility	**Partially satisfied**	Each phase includes defined objectives, concrete activities, a decision gate, and documented outputs — giving practitioners enough specificity to act. Indirect evidence of utility comes from the alignment between how participants described migration challenges and the phases designed to address them. Direct utility evidence requires field implementation, which is the primary recommendation for future research. The claim of utility is therefore reasoned and plausible, not empirically demonstrated.

Generalisability	**Satisfied within scope**	The five demand-side organisations span IT services, retail, banking, consulting, and high-growth technology across Finland, India, and Bulgaria. The core themes appear consistently across this range. The principal scope limitation — very large enterprises with highly heterogeneous legacy environments — is explicitly acknowledged in both Phase 5 and Chapter 6. Organisations significantly larger than the contexts represented in the interviews should treat Phase 5 as requiring substantial adaptation.

The fact that the utility criterion was met partially is reason enough to comment briefly on the matter. The utility case is based on consistency between the descriptions of the participants of the migration challenges and the phases to which the LQMF devotes the migration challenges to address them ,a form of construct validity. It does not rest on evidence that the implementation of the framework leads to improved migration outcomes as compared to its non-implementation. Such evidence can only be applied longitudinally to a field and is not covered in this thesis. The statement that is here made is more humble, and more justifiable, that the framework is logically well-constructed, theorised, and designed in the ways which practitioners recognised as representing the actual contours of the problem they face. Whether it works better than alternative in the practice is an important and open question.

5.5 Scenario based validation

To explore further the applicability and the boundary conditions of the LQMF conceptual applicability was tested on three different organisational situations. These situations are not case studies nor do they directly describe a particular participant organisation. They have built analytical programs, each depicting a identifiable category of organisation that would realistically take part in a QA migration. Every scenario is approached as an evaluative exercise, not only to determine where the framework supports the migration effectively but also to identify where it has recognized limitations or needs to be adapted

to. This dual framing is consistent with the evaluation criteria set forth in Chapter 3: a framework that only confirms to its own applicability, not being evaluated seriously.

5.5.1 **Scenario 1- legacy heavy enterprise**

A large organisation with a lot of legacy systems that have undergone multiple technology decisions through the decades, a heterogeneous testing toolset assembled together out of a series of technology choices rather than planned architecture, hundreds of testers spread across semi-autonomous teams, and a large backlog of unmaintained scripts whose validity may be questionable.

In this respect, Phase 2 (Waste Identification) and Phase 4 (Pilot Implementation) come into focus and emphasize their significance further. The waste register that will be generated at Phase 2 does not just happen to be a diagnostic but something that will make the case on the need to embark on migration to leadership and in one that is rich in legacy will most certainly result in the documentation of a level of accumulated wastage that is difficult to quantify unless approached meticulously. In the absence of this evidence base, migration proposals in large organisations can often fall short at the investment justification stage. The pilot role of Phase 4 is equally important: the decision to make a full migration is risky enough on its own, and when the cost of a failed migration is high, the risk of pilot error is disproportionately large.

The LQMF also seems to be best suited to helping facilitate a gradual, phased transition process within this context , which is the approach most likely way of minimising the level of disruption and sustain leadership commitment across a multi-year programme. Nevertheless, the amount of coordination needed in large-scale enterprises is complex and has not been entirely taken into account by the present framework. Phase 5 (Scaled Rollout) assumes that the pilot findings can be generalized to a relatively stable level when applied to a subsequent teams. This assumption is not uniformly true in those organisations whereby teams have essentially different tooling stacks, codebases of

different ages, and varying degrees of automation maturity. The migration of each team may need to be managed as an autonomous sub-project of its own with its own waste profile, having its own champion engineer and with its own Phase 3 tool evaluation - across which the governance and coordination overhead is multiplied in ways not fully detailed in the current Phase 5 guidance of the framework. This is recognised as a design boundary and is listed in Chapter 6 as a priority in developing future frameworks.

5.5.2 **Scenario 2 - agile mid-sized organisation.**

An organisation of 150 to 400 employees that have mature practices in Agile-delivery, have established CI/CD pipelines, have cross-functional product teams, and already have automation (usually Cypress or Selenium) with reasonable but suboptimal reliability. The biggest areas of concern include locator maintenance overhead and the difference in automation coverage and pace of delivery of new features.

In this respect, the LQMF improves the organisational practices that are already in place compared to the other two scenarios. The various phases like pilot implementation and scaled roll out may be directly integrated into existing sprint and planning cycles instead of a separate migration project structure. The champion engineer model of Phase 4 projects onto the informal influence networks that already exist in mature Agile teams. The governance mechanisms of Phase 7 (waste reviews quarterly, practice boards) are very similar to the culture of retrospective and continuous improvement which defines well-functioning Agile organisations.

In this respect, the framework cannot be seen as a transformation tool, but more as a structuring tool to an improvement, which the organisation is already culturally prepared to make. Underinvestment as opposed to resistance is the risk in this scenario: already reasonably functional teams may not feel the urgent need to prioritize the migration, and Phase 2 waste identification may reveal problems that are real but not urgent enough to cause a compelling need to act with immediate effect. The Phase 2

decision gate, which requires that the top three categories of waste get quantified in engineering hours before Phase 3 commences, is specifically created to deal with this tendency. This general understanding of the existence of maintenance overhead having been converted to a definite cost figure creates the urgency that Kotter (1996) identified as a precondition to the sustained change, even in improvement-oriented organisations.

5.5.3 **Scenario 3 — Regulated Environment**

An organisation in any sector where there exist formal regulatory obligations, such as banking, healthcare, insurance, or patient-safety. Test execution should be auditable, test data should be handled with a lot of care and any new tooling must go through the procurement processes that involve compliance assessment. The organisation might already automate, of differing degrees of maturity, but constraints imposed by the regulatory environment are not addressed by generic migration frameworks.

Trust and traceability are not sub-tinges in controlled settings - they are integral conditions of admission. The introduction of governance mechanisms (Phase 7) and explicit mandates on the execution evidence generation capability and synthetic datasets generation ability (non-negotiable Phase 3 selection criteria) seem to directly address these requirements. In particular, the insistence of the framework on a structured pilot (Phase 4) prior to scaled commitment is especially well adapted to regulated settings where the cost of a failed scaled public rollout is much higher than in less regulated areas. The pilot makes available the internal evidence base, that procurement and compliance functions must have, before full organisational deployment is given the green light.

There is however, one limitation which deserves a candid recognition. The degree at which AI-generated test output can be fully audited in the strictest of regulatory environments is an open question in the field - not just in this framework. In this case, in particular, when autonomous end-to-end test generation does not yet reach

production maturity, the observation by P5 that the production maturity of autonomous end-to-end test generation is yet to achieve is particularly relevant here: unless AI creates a test battery based on a requirements document without human verification of each test case, the auditability of such a test-generation process may not satisfy regulators who demand documented human sign-off on testing scopes. The LQMF attempts to address this by the insistence of Phase 3 on the generation capability of evidence and the mechanisms of governance review, which Phase 7 requires, but cannot answer a field-wide question of maturity that even the insistence of Phase 3 can not possibly resolve. Organisations within heavily-regulated sectors ought to make the selection of Phase 3 tools an iterative process and maintain strong communication with their compliance and legal functions through the migration.

5.6 Discussion

The results of the evaluation put collectively in support of a well-founded claim which is both cautionary and justifiably cautious. The Legacy QA Migration Framework (LQMF) presents a clearly identified and under-researched problem by offering a theoretically based approach, with practical structure. This faith is substantiated by the correspondence of themes determined in the analysis of the NVivo, gaps determined in the study of benchmarking, and the phases of the structure that will start to use these gaps. Such a statement is however qualitative because the framework is yet to be proved through real-life implementation. Although practitioners view the framework as rational and helpful, its ability to yield better migration outcomes in comparison with ad hoc approaches can only be established in future field practice.

Identification of the aspiration-adoption gap in AI testing, which Karhu et al. (2025) emphasize as one of the key empirical findings of this study, can be consistently supported by the interview data. Though three-quarters of organisations rate AI testing as a strategic priority, only one-sixth have actually assigned AI testing into practice. The results indicate that such a gap is not caused by any technological constraints, but

instigated by managerial and organisational issues. Although the tools and the technical capabilities are easily accessible, there is no defined orgnisational methodology to manage their organisational change, process integration, and governance requirements that they need to manage in order to successfully adopt them. To fill this gap LQMF is uniquely created to offer such structured pathway.

The results of the interview also justify the fact of the existence of this gap. All participants, including the demand-side practitioners, reported that they use AI tools on a regular basis in their daily work, specifically the tools like Copilot to generate and provide test support. Such an extensive use generates an illusion of AI preparedness on a personal scale. Nonetheless, this does not lead to a level of preparedness to adopt AI testing to enterprise-levels. The demands of deploying AI testing platforms at scale such as governance, CI/CD integration, synthetic data configuration and team-level changes management are all topics that are much more complex. Thus the difference between personal AI usage and AI use in the organisation is a pivotal element behind the aspiration-adoption discrepancy.

The implications of this distinction, on Phase 6 (Enablement) of the LQMF are important. The results suggest that training programs are not to presume that familiarity with the AI tools identify organisational readiness. Rather, enablement activities need to explicitly consider how to bridge between the use of AI at the individual level and its application in the enterprise level in nature.

The next valuable lesson is associated with the existing maturity of AI testing technologies. One participant who had a direct involvement in the industry (P5) pointed out that capabilities of AI based testing tools are usually exaggerated in the vendor narrative. Other participants who had acquired hands-on experience also supported this view. Frameworks, organisations and procurement processes that tune to what can be delivered today by AI testing (instead of to the vendor roadmap or aspirational demonstration) will better set more realist expectations, will encounter less

disappointment when deploying it, and will build more sustainable adoption cultures. This principle is designed around Phase 3 of the LQMF that focuses on realistic evaluation and choosing tools.

Lastly, the results demonstrate the changing nature of the role of QA professionals. The fact that the jobs in QA will be more conceptual in terms of the goals of testing, as opposed to simply implementing them in practice. Although the LQMF is just starting to tackle this transition by its enablement phase, it is not entirely able to address the bigger issue of redefining QA roles in an AI-driven environment. This is a key area of future research and organisational emphasis.

6 Conclusion

This chapter summarizes the main findings, outlines what contributions this thesis makes, what its limitations are, and indicates the most significant directions that should be explored in the future under the impact of this work.

6.1 Answer to the research question

How can organisations effectively manage the migration from legacy QA automation to AI-powered testing in order to eliminate process waste and sustain long-term quality improvement?

The solution is the Lean QA Migration Framework - a seven step model that covers migration of initial diagnosis to the sustained governance. The answer is based on 3 principles. First, the waste needs to be identified prior to the selection of tools: Phase 2 produces the evidence base of what Phase 3 needs to achieve, and organisations that reverse the sequence routinely select platforms that seek to address the wrong problems. Second, the pilot needs to be regarded as an act of change management, not a technical exercise: the aspiration-adoption gap — 75 percent strategic intent, 16 percent actual adoption (Karhu et al., 2025) is an act of failure on the human and process dimensions of adoption, rather than a failure of the technology. Third, the governance should not be project based but permanent: Phase 7 exists because improvement is not pegged on the regular working cycles rhythm. The combination of Kaizen and Kotter's step 8 provides the mechanism through which the gains of the migration can be made lasting.

6.2 Research objective

Chapter 1 outlines all three objectives of the research. The literature review in Chapter 2 identifies five distinct gaps and determines that no previous study combines Lean management, AI testing research, and change management theory together into a unified QA migration framework. Chapter 4 conveys that framework, complete seven phases, which are theoretically-based, decision-gates, and documented outputs. Chapter 5 judges it against NVivo matrix analysis, its comparison with TMMi, Forsgren et al. (2018) Accelerate, Kotter, CMMI, and compares it to all four pre-specified criteria.

6.3 Contributions

The five contributions that this thesis has are summarised in Table 12. The primary contribution is the LQMF itself — a designed artifact in the DSR sense (Hevner et al., 2004). Whether it will assist organisations more effectively in managing QA migrations than they do without it, a question that field implementation will answer.

Table 12. Five contributions summary

Type	Contribution	Description
Theoretical	Framework Design	The LQMF is the first integrated, phased migration framework to apply Lean waste analysis as the diagnostic architecture for a QA technology transition. It assembles existing theoretical knowledge — Lean Thinking, The Toyota Way, Kotter, Davis — into a coherent, sequenced structure that no existing framework provides.
Theoretical	Waste Taxonomy	Table 8 constitutes an original mapping of the seven Lean waste types to QA automation contexts. While Poppendieck (2007) and Hicks (2007) established that Lean waste applies to software and knowledge work, no prior study applies the taxonomy specifically to QA migration as a structured diagnostic tool.

Empirical	Aspiration-Adoption Gap	The interview data confirms and extends Karhu et al.'s (2025) finding that the adoption gap is primarily managerial. The NVivo matrix analysis adds specificity: the gap is concentrated in the waste diagnosis and enablement dimensions, not the technology selection dimension, which existing vendor guidance already addresses.
Empirical	Maturity Limit	The supply-side acknowledgement (P5) that autonomous end-to-end test generation has not reached production maturity, confirmed by demand-side experience, provides empirical grounding for a claim that the academic literature documents but practitioners frequently underestimate.
Practical	Phase-by-Phase Guide	Each LQMF phase includes defined objectives, concrete activities, a decision gate, and documented outputs — making the framework directly applicable to practitioners without requiring translation through additional implementation guidance.

6.4 Practical implications

There are three audiences whose takeaways of this research are different. To the QA managers and engineering directors: begin with the waste register rather than tool evaluation. Phase 2 is the definition of what the migration has to bring; ignoring it results in optimisation for the wrong problems. To CIOs and technology leaders: budget the human aspects of the migration process, such as training, time spent by champion engineers, change management, etc., on the same level as the technical ones. Phase 6 is the phase that has been most underinvested and has contributed the most to unsuccessful adoption. To AI testing vendors: precise capability positioning is more commercially durable than aspirational marketing. Organisations that choose platforms on realistic descriptions become long term customers; those that choose based on overstated claims become disappointed ones.

6.5 Limitation

The five most important limitations are directly given in Table 13.

Table 13: Limitation of study

Limitation	Description and Implication
No field implementation	The LQMF has not been tested in a live migration. The utility criterion is partially satisfied — the framework is logically coherent and practitioners recognised it as addressing real problems — but the claim that it produces better migration outcomes than an unguided approach cannot be made without longitudinal field evidence.
Interview sample size	Seven participants across five organisations is sufficient for theoretical validation and framework evaluation through expert review, but does not support statistical generalisability. Patterns identified in the data should be treated as analytically plausible rather than statistically representative.
Supply-side concentration	Two of seven participants are co-founders of an AI testing company. Despite the mitigation strategy described in Chapter 3, this concentration means the supply-side perspective is better represented than its share of the total sample implies.
Geographic scope	Most participants and the majority of the literature reviewed are from European or North American enterprise contexts. Organisations in South-East Asia, Latin America, the Middle East, or other regions with different regulatory frameworks, talent markets, and tooling ecosystems may face constraints that the LQMF does not fully anticipate.
Temporal validity	AI testing capability is developing rapidly. The tool evaluation in Chapter 4 (Table 9) and the capability assessments in Chapters 2 and 5 reflect the state of the field in 2025–2026. Practitioners consulting this thesis in subsequent years should verify that specific capability claims and competitive positions remain current before applying Phase 3 criteria.

6.6 Future research

Table 14 lists five areas in which the research could be developed in the future. The top priority is a longitudinal field implementation study - applying LQMF to a live migration and evaluating the outcomes in 12 to 24 months. It is the sole path to bridging the utility gap that can be bridged and moved to through expert review alone.

Table 14. Future research

Research Direction	Description and Rationale
Field implementation study	Longitudinal case study applying the LQMF to one or more live QA migrations, tracking waste reduction, adoption rates, and quality improvement outcomes over 12–24 months. This is the primary gap in the current evidence base and the most direct route to demonstrating utility empirically.
Enterprise-scale adaptation	In-depth study of very large organisations (1,000+ testers) to develop Phase 5 guidance that reflects the dependency complexity, heterogeneous tooling, and distributed team structures that make enterprise migration qualitatively different from the contexts primarily represented in this study.
QA professional identity and role transition	Qualitative research into how QA engineers experience the transition from execution-focused to strategy-focused roles as AI tools absorb routine automation. The T4 theme suggests this is both a significant personal and organisational challenge that the LQMF addresses partially but does not resolve.
Regulated sector migration study	Dedicated study of QA migration in banking, healthcare, or government contexts — sectors where the trust and evidence requirements identified in T6 are most acute and where the consequences of migration failure are most significant.
Framework maturation through iteration	Subsequent thesis or practitioner research updating the LQMF phases on the basis of field implementation feedback, capturing both what the framework predicted accurately and what it failed to anticipate. Design science research artefacts are expected to improve through iterative evaluation.

The gap this thesis attempts to address was never about technology. The limitations of legacy automation are well documented. The capbilities of AI-powered testing are increasingly understood. What has been lacking is a systematic, theoretically grounded path between the two, something which informs organisations not only where to go but how to manage the journey. The LQMF is that path. It is a first generation artifact and has the drawbacks that come with that status The question the following stage of research has to answer is whether it is practical or not.

6.7 AI limitation

This thesis is undertaken with the assistance of artificial intelligence tools to enhance the quality and coherence of the essays. To improve grammar and linguistic flow the Quill Bot paraphrasing tool and Grammarly were utilized. The original idea, content and analysis that the researcher had were preserved in the thesis. The AI tools were employed in refining language only. Moreover, all AI-generated suggestions were thoroughly checked and corrected to ensure that they are suitable in terms of academic standards and the thesis purpose. The writing process is improved with these tools, but they never take the place of intellectual work that is carried out by the author.

References

Allala, S. C., Sotomayor, J. P., Santiago, D., King, T. M., & Clarke, P. J. (2019). Towards Transforming User Requirements to Test Cases Using MDE and NLP. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, 350–355. https://doi.org/10.1109/COMPSAC.2019.10231

Amalfitano, D., Faralli, S., Hauck, J. C. R., Matalonga, S., & Distante, D. (2024). Artificial Intelligence Applied to Software Testing: A Tertiary Study. *ACM Computing Surveys*, *56*(3), 1–38. https://doi.org/10.1145/3616372

Ayenew, H., & Wagaw, M. (2024). Software Test Case Generation Using Natural Language Processing (NLP): A Systematic Literature Review. *Artificial Intelligence Evolution*, 1–10. https://doi.org/10.37256/aie.5120243220

Conboy, K. (2009). Agility from First Principles: Reconstructing the Concept of Agility in Information Systems Development. *Information Systems Research*, *20*(3), 329–354. https://doi.org/10.1287/isre.1090.0236

Davis, F. D., & Granić, A. (1989). *The Technology Acceptance Model: 30 years of TAM*. Springer International Publishing AG.

Fontes, A., & Gay, G. (2023). The integration of machine learning into automated test generation: A systematic mapping study. *Software Testing, Verification and Reliability*, *33*(4), e1845. https://doi.org/10.1002/stvr.1845

Forsgren, N., Humble, J., & Kim, G. (2018). *Accelerate: The science behind DevOps: building and scaling high performing technology organizations*. IT Revolution.

Garousi, V., Bauer, S., & Felderer, M. (2020). NLP-assisted software testing: A systematic mapping of the literature. *Information and Software Technology*, *126*, 106321. https://doi.org/10.1016/j.infsof.2020.106321

Garousi, V., Joy, N., Jafarov, Z., Keleş, A. B., Değirmenci, S., Özdemir, E., & Zarringhalami, R. (2024). *AI-powered software testing tools: A systematic review and empirical assessment of their features and limitations* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2409.00411

Garousi, V., & Veenendaal, E. V. (2022). Test Maturity Model Integration: Trends of Worldwide Test Maturity and Certifications. *IEEE Software*, *39*(2), 71–79. https://doi.org/10.1109/MS.2021.3061930

Gregor, S., & Hevner, A. R. (2013). Positioning and Presenting Design Science Research for Maximum Impact1. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research1. *MIS Quarterly*, *28*(1), 75–106. https://doi.org/10.2307/25148625

Hicks, B. J. (2007). Lean information management: Understanding and eliminating waste. *International Journal of Information Management*, *27*(4), 233–249. https://doi.org/10.1016/j.ijinfomgt.2006.12.001

Karhu, K., Kasurinen, J., & Smolander, K. (2025). *Expectations vs Reality—A Secondary Study on AI Adoption in Software Testing* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2504.04921

Kothamali, P. R. (2025). *AI-Powered Quality Assurance: Revolutionizing Automation Frameworks for Cloud Applications*. *5*.

Kotter, J. P. (20). *Leading change* (Nachdr.). Harvard Business School Press.

Leotta, M., Ricca, F., Marchetto, A., & Olianas, D. (2024). An empirical study to compare three web test automation approaches: NLP-based, programmable, and capture&replay. *Journal of Software: Evolution and Process*, *36*(5), e2606. https://doi.org/10.1002/smr.2606

Leotta, M., Stocco, A., Ricca, F., & Tonella, P. (2015). Using Multi-Locators to Increase the Robustness of Web Test Cases. *2015 IEEE 8th International Conference on Software Testing, Verification and Validation (ICST)*, 1–10. https://doi.org/10.1109/ICST.2015.7102611

Liker, J. K. (2007). *Toyota way: 14 management principles from the world's greatest manufacturer* (First edition). McGraw-Hill Education.

Lima, B. F., Neto, J. V., Santos, R. S., & Caiado, R. G. G. (2023). A Socio-Technical Framework for Lean Project Management Implementation towards Sustainable Value in the Digital Transformation Context. *Sustainability*, *15*(3), 1756. https://doi.org/10.3390/su15031756

Liu, M., Sun, M., Zhang, X., Ge, M., & Hu, J. (2025). Real-time shop floor operations improvement based on dynamic value stream mapping and hybrid simulation in Industry 4.0: An economic perspective. *International Journal of Production Research*, *63*(18), 6776–6800. https://doi.org/10.1080/00207543.2025.2487921

Nidagundi, P., & Novickis, L. (2016). Introduction to Lean Canvas Transformation Models and Metrics in Software Testing. *Applied Computer Systems*, *19*(1), 30–36. https://doi.org/10.1515/acss-2016-0004

Nielsen, P. A. (2020). Problematizing in IS Design Research. In S. Hofmann, O. Müller, & M. Rossi (Eds.), *Designing for Digital Transformation. Co-Creating Services with Citizens and Industry* (Vol. 12388, pp. 259–271). Springer International Publishing. https://doi.org/10.1007/978-3-030-64823-7_24

Ozkaya, I. (2023). The Next Frontier in Software Development: AI-Augmented Software Development Processes. *IEEE Software*, *40*(4), 4–9. https://doi.org/10.1109/MS.2023.3278056

Poppendieck, M. (2007). Lean Software Development. *29th International Conference on Software Engineering (ICSE'07 Companion)*, 165–166. https://doi.org/10.1109/ICSECOMPANION.2007.46

Ramos, T., Dean, A., & McGregor, D. (2025). AI-Augmented Software Engineering: Revolutionizing or Challenging Software Quality and Testing? *Journal of Software: Evolution and Process*, *37*(2), e2741. https://doi.org/10.1002/smr.2741

Ricca, F., Marchetto, A., & Stocco, A. (2021). AI-based Test Automation: A Grey Literature Analysis. *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 263–270. https://doi.org/10.1109/ICSTW52544.2021.00051

Rother, M., & Shook, J. (2018). *Learning to see: Value-stream mapping to create value and eliminate muda* (Version 1.5 ; 20th Anniversary Edition). Lean Enterprise Inst.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward A Unified View1. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Vom Brocke, J., Hevner, A., & Maedche, A. (Eds.). (2020a). *Design Science Research. Cases*. Springer International Publishing. https://doi.org/10.1007/978-3-030-46781-4

Wang, J., Huang, Y., Chen, C., Liu, Z., Wang, S., & Wang, Q. (2024). Software Testing With Large Language Models: Survey, Landscape, and Vision. *IEEE Transactions on Software Engineering*, *50*(4), 911–936. https://doi.org/10.1109/TSE.2024.3368208

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, *17*(5), 470–475. https://doi.org/10.1057/ejis.2008.44

Womack, J. P., & Jones, D. T. (1996). *Lean thinking: Banish waste and create wealth in your corporation*. Simon & Schuster.

Appendices

Appendix 1 a. Interview Questionnaire- Demand Side

Current QA State

1. How is QA automation currently structured within your team or project?
2. What percentage of your regression testing is automated?
3. What are the biggest challenges you face with the current automation framework?
4. How long does your full regression test suite take to execute end-to-end?
5. How reliable are your automated test results? Do you frequently experience false failures (flaky tests)?

Waste and Pain Points

1. Where do you see delays or bottlenecks in your QA process?
2. Are there repetitive manual activities that could potentially be reduced?
3. Do testers spend time on low-value tasks (e.g., fixing scripts instead of improving coverage)?
4. Are there types of testing your team knows they should be doing but currently can't — and if so, what's stopping them?

Organisational and Team Dynamics.

1. How do your QA engineers feel about their current tools ?
2. When new team members join QA, how long does it realistically take them to become productive contributors to the automation suite?
3. How well do your developers and QA engineers collaborate on testing?
4. If you proposed switching to a completely new testing platform, what would be the main sources of resistance?

AI Awareness & Readiness

1. Has your organisation explored or discussed AI-powered testing tools?
2. What concerns or risks do you associate with adopting AI in QA?
3. Do you believe your current team has the skills required to adopt AI-powered testing?
4. What would a testing tool need to do differently from what you have today to make a significant positive difference to your team's output?
5. What concerns would you have about adopting AI in your testing process — around control, transparency, skill gaps, or anything else?

Strategic and Business Alignment

1. What does a "successful QA transformation" look like to you in concrete, measurable terms over the next 12 to 24 months?
2. What support would project teams need during such a transition?

Appendix 1 b. Interview Questionnaire- Supply Side

1. What motivates organisations to adopt AI-powered testing platforms?
2. What challenges do organisations face when transitioning from traditional testing to AI-powered testing?
3. How does AI testing change the role of QA engineers?
4. What are the biggest barriers organisations face when implementing AI testing tools?
5. What strategies help organisations successfully adopt AI-powered testing platforms?
6. In your experience, what skills will become most important for QA professionals as AI tools become more widely used?

Appendix 2. Codebook

Name	Sources	References
AI Adoption Barrier	5	9
Challenge of Transition	1	1
Challenge of Trust	1	1
Challenge organisation face	1	1
Commitment anxiety	1	1
MIndset and Barrier Issue	1	1
Mindset problem	1	2
Trust concern	1	1
Unwilling to use AI	1	1
Current AI Capabilities	4	14
AI Benefit Of Integrationg AI Testing Tool	1	1
AI helps the QA	1	1
AI is faster	1	1
AI Testing tool Benefits	1	1
AI tool for testing	1	1
AI Usage in company	1	2
Benefit Of AI	1	2
Copilot usage	1	1
Current Ai Assisted tool	1	1
Current AI Usage	1	2
Synthetic Data generator	1	1

Name	Sources	References
Current QA & Automation Practices	5	14
Agile Manifestation	1	1
Agile practices	1	1
AUtomation Limitation	1	1
Automation tool	1	1
Collaboration Between Team	1	1
continous deployment	1	1
Current absence of automation in project and wish to introduce it	1	1
Manual regression sessions	1	1
QA position in Company	1	1
Role and career info	1	1
Roles and skills	1	1
Team setup	1	1
Test process & lifecycle	1	2
Current QA tooling	3	3
Tooling	1	1
tooling in Team	1	1
Value of migration frameworks and limits	1	1
Desire for more automation of basic flows	1	1
Developer also does testing	1	1
Duplication of code	1	1
Evolution of QA	1	1
Expectation of AI	1	1
Exploratory testing	1	1

Name	Sources	References
Flaky behaviour	1	1
Framework phase feedback	0	0
Furturistic Team	1	1
Future expectation Of QA	1	1
Future of QA	1	2
Future of QA And AI transition	1	1
Future of QA huld	1	1
Governance and trust	1	1
Human-in-the-loop principle and need to read AI-generated tests	1	1
Jobs at risk	1	1
Lack of structure	1	1
Legacy Automation Challenge	4	10
Automation Setup Clearly	1	1
Challenge of not accessible to production	1	1
Environment instability	1	1
Flaky test	1	1
Locator change	1	1
Repeatitive Work	1	1
Skill Gap and slow Work	1	1
Team owning the work	1	2
Timer issue	1	1
LLM Usage	1	1
Mindset challenge by organisation	1	1
Motivation to adopt AI Powered Testing Tool	1	1

Name	Sources	References
Need for Common	1	1
No single tool to do Everything	1	1
No time to explore AI	1	1
OP landscape	1	1
Organisational scale challenges	0	0
Organisational challenges	1	1
Past experience in Robot Framework automation	1	1
Perception of AI as “scary” but interesting	1	2
POC approach needed	1	1
Proper Documentation by Organisation	1	1
POC approach	1	1
QA Role Evolution	5	7
Training for AI	1	1
Training in company	1	1
Training to learn AI	1	1
Training, communities of practice, GenAI testing tribe	1	1
Upskill	1	1
Upskilling the team	1	1
View on structured migration	1	1
QA structured in company	1	2
Regression workload	1	1
Repeatitive Testing	1	1
Role oF AI tool CEO	1	1
Role, experience and domain	1	1
roles and response	1	1

Name	Sources	References
Roles Responsibilities	1	1
Scaling test execution	1	1
Shortfall of AI testing tool	1	1
Skepticism about fully autonomous AI testing tool or self-healing tools	1	1
Skills lost or not used	1	1
Successful way of AI adoption	1	1
Sucessful Transformation to AI	1	1
Team concerns about AI replacement	1	1
Team organisation and collaboration	1	1
Trust building Mechanism	2	2
AI trust issue	1	1
Conditions for adopting AI testing tools	1	1