CHALMERS
UNIVERSITY OF TECHNOLOGY

Taming the Invoice Review Monster

Exploring the Potential of AI to Identify Cost Deviations in The West Link Project

Master's thesis in Management and Economics of Innovation

Jonathan Krantz | Henrik Stenfelt

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
DIVISION OF INNOVATION AND R&D MANAGEMENT
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

MASTER'S THESIS 2026

Taming the Invoice Review Monster

Exploring the Potential of AI to Identify Cost Deviations in The West Link Project

JONATHAN KRANTZ HENRIK STENFELT

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Technology Management and Economics
Division of Innovation and R&D Management
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026

Taming the Invoice Review Monster
Exploring the Potential of AI to Identify Cost Deviations in The West Link Project
JONATHAN KRANTZ
HENRIK STENFELT

Supervisor: Christoffer Rönndahl, Division of Innovation and R&D Management
Examiner: Petra Bosch-Sijtsema, Division of Innovation and R&D Management

Master's Thesis 2026
Department of Technology Management and Economics
Division of Innovation and R&D Management
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: The image on the front cover was provided by the Swedish Transport Administration.

Taming the Invoice Review Monster
Exploring the Potential of AI to Identify Cost Deviations in The West Link Project
JONATHAN KRANTZ
HENRIK STENFELT
Department of Technology Management and Economics
Chalmers University of Technology

Abstract

Cost overruns and administrative complexity represent persistent challenges in large-scale infrastructure megaprojects. This master's thesis investigates the potential of using Artificial Intelligence (AI) to identify cost deviations within the invoice review process, focusing on the Swedish Transport Administration (Trafikverket) and its West Link project. Currently, manual review processes suffer from a critical "capability gap", where the overwhelming volume of documentation forces a sampling-based approach that fails to detect cost deviations.

Through a qualitative case study involving semi-structured interviews with project controllers and managers, this research identifies the bottlenecks created by informational asymmetries between commissioning organizations and dominant contractors. The findings suggest that AI, specifically through automated data processing and pattern recognition, can bridge this gap, allowing project owners to transition toward the role of a "strong owner". The thesis proposes a framework for integrating AI into existing workflows to enhance the invoice review function in projects with cost-plus contracts. By shifting from manual sampling to automated, comprehensive verification, organizations can better manage the administrative complexities inherent in large infrastructure projects.

Keywords: Artificial Intelligence (AI), Infrastructure Megaprojects, Cost Deviations, Invoice Review, Construction Management, Trafikverket, Västlänken, Procurement Governance.

Acknowledgements

This master's thesis was conducted during the spring of 2026 as part of the program in Management and Economics of Innovation at Chalmers University of Technology, in collaboration with the Swedish Transport Administration.

We would like to begin by expressing our gratitude to the entire team at the Swedish Transport Administration. In particular, we would like to thank our main contact, Lillian Brunbäck, for her guidance during this project.

Moreover, we would like to extend our sincere thanks to our supervisor at Chalmers University of Technology, Christoffer Rönndahl. He has provided us with consistent support and valuable guidance.

Finally, we would like to thank all respondents, to whom we are grateful for their time and for sharing their insights about the West Link project and their work. Without the generous input from these dialogues, this report would not have been possible in its current form.

Jonathan Krantz & Henrik Stenfelt, Gothenburg, June 2026

List of Acronyms

AI	Artificial Intelligence
AP	Accounts Payable
API	Application Programming Interface
CAATs	Computer Assisted Audit Techniques
CNN	Convolutional Neural Network
CoT	Chain of Thought
EDI	Electronic Data Interchange
ERP	Enterprise Resource Planning
GDPR	General Data Protection Regulation
GDP	Gross Domestic Product
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GRN	Goods Received Notice
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LLM	Large Language Model
ML	Machine Learning
NDA	Non-Disclosure Agreement
NER	Named Entity Recognition
OCR	Optical Character Recognition
OSL	Offentlighets- och Sekretesslagen
PDF	Portable Document Format
PO	Purchase Order
RPA	Robotic Process Automation
RQ	Research Question
SEK	Swedish Krona
SQL	Structured Query Language
STA	Swedish Transport Administration
TF-IDF	Term Frequency-Inverse Document Frequency
USD	United States Dollar
VAT	Value Added Tax
XAI	Explainable Artificial Intelligence
XML	Extensible Markup Language
ÄTA	Ändring, Tillägg, Avgående

List of Figures

List of Tables

Contents

1

Introduction

This chapter introduces the fundamental challenges of cost control in large-scale infrastructure projects, using The Swedish Transport Administration's (Trafikverket, hereafter referred to as STA) West Link project as the empirical setting. It outlines the background of the problem, defines the specific research focus regarding manual invoice review, and presents the research questions and limitations that guide the remainder of the study.

1.1 Background

Large infrastructure projects are characterized by exceptional complexity, long planning horizons, and significant exposure to cost escalation. Flyvbjerg, Holm, and Buhl (2002) analyzed 258 transport infrastructure projects in 20 countries and found that costs were underestimated in nine out of ten cases, with actual costs on average exceeding forecasts by 28%. Similar conclusions have been drawn from the literature on so-called "megaprojects", which are large, complex, and often politically charged public investments. In this type of project, cost overruns and schedule delays are the norm rather than the exception (Flyvbjerg, Bruzelius, & Rothengatter, 2003). Furthermore, Eliasson (2025) analyzed Swedish transport infrastructure projects from 2004 to 2022 and found a pattern of considerable cost escalation during early planning stages. According to Eliasson (2025), these escalations are primarily driven by frequent changes in project content and scope, as well as the premature lock-in of project decisions before costs and benefits have been thoroughly assessed. This early lock-in distorts project selection by reducing the incentive to find cost-efficient solutions and creating opportunities for project beneficiaries to intentionally underestimate costs to secure approval. Hence, Eliasson (2025) shows that the distribution of cost changes is highly skewed. While many projects remain close to initial estimates, these causal factors drive a long right tail, with a subset of projects escalating dramatically.

Sweden's largest ongoing railway project shows tendencies of these dynamics. The West Link is an eight-kilometer underground railway connecting Gothenburg's main station to the southern parts of the city, with over six kilometers in tunnel and three new underground stations at Centralen, Haga, and Korsvägen (Trafikverket, n.d.). The project was first conceived in 2001 and was formally defined through a feasibility study in 2002. It then entered full planning and rail planning stages across the following decade. Construction began in 2018 and is currently ongoing at several sites in central Göteborg, with a partial opening planned at Station Centralen in 2026, and a full opening of the link expected around 2030 (Trafikverket, n.d.). The West Link ranks among the most resource-intensive infrastructure investments in Swedish history, with an original budget of Swedish Krona (SEK) 20 billion at 2009 price levels. The budget estimates have since been revised significantly upward. As of 2023, STA's assessment placed the probable total cost at SEK 25.7 billion in 2009 price levels, with the likely range extending from SEK 24.1 to 27.1 billion, representing a cost increase of approximately 20 to 35% relative to the original estimate. At 2023 price levels, this translates to a probable total cost of SEK 36.5 billion. In September of 2025, STA further revised its cost estimates to SEK 49.3 billion at current price levels (Trafikverket, n.d.). The West Link exemplifies a project in which early estimates were significantly below the realized cost, cost escalation occurred at the planning and execution stages, and the scale and complexity of the undertaking have made effective cost control a challenge. This trajectory is consistent with the patterns described by Eliasson (2025) and Flyvbjerg et al. (2002).

The The West Link sub-project at Korsvägen stands out in terms of cost deviation. A cost deviation can be defined as a cost item that exceeds previous estimates. An independent review of the sub-project by Delray Consulting AB (2025) found that it exceeded its original cost estimate by SEK 4.1 billion. According to the authors of the review, this was partly driven by insufficient supporting documentation to invoices, weak traceability between invoice lines and the contractual basis, unclear contract terms that affected how costs were claimed and reviewed, and inadequate pre-approval review processes. Additionally, the review documents how multi-tier subcontracting arrangements contributed to cost-driving markups and reduced transparency, undermining STA's ability to maintain effective cost oversight (Delray Consulting AB, 2025).

This challenge is not unique to the West Link. Suprapto, Bakker, Mooi, and Hertogh (2016) argue that the cost deviations in megaprojects of this kind are driven by the contractual structures used in projects of this type. When uncertainty and scope complexity make fixed-price contracts impractical, commissioning organizations such as STA typically rely on cost-plus ("löpande räkning") arrangements. In such arrangements, the contractor is compensated for the actual costs incurred, subject to the terms of the contract, as well as a pre-defined profit margin that is proportional to the costs (Suprapto et al., 2016). This contract form shifts the risk distribution to the commissioning organization, who must verify the allowability and correctness of each claimed cost, rather than the contractor absorbing cost uncertainty. Eriksson and Lind (2015) describe this through a moral hazard lens, arguing that cost-plus contracts create conditions where contractors may have limited incentives to control costs, with their profit growing proportionally to the costs incurred. Also, the commissioning organization's ability to monitor is constrained by information asymmetry, meaning that the contractor has direct insight into work composition and actual costs while the commissioning organization must infer them from documentation and inspections (Eriksson & Lind, 2015). Suprapto et al. (2016) add that the results of megaprojects under such arrangements are shaped not only by formal contract terms, but by relational quality and the degree of collaborative governance between the commissioning organization and the contractor. Additionally, supplier structures with subcontracting chains further reduce the commissioning organization's visibility into cost composition, since markups and indirect cost allocations can accumulate across tiers in ways that are difficult to trace without strong documentation routines and systematic review (Suprapto et al., 2016). In a project as complex as the West Link, with several major contracts, multiple subcontracting tiers, and a project duration spanning multiple decades, these dynamics are amplified both in terms of scale and frequency.

Invoice review is an important control activity available to the commissioning organization under cost-plus arrangements. It is the process through which the commissioning organization assesses whether invoices are allowable under the contract, supported by sufficient evidence, correctly attributed, and within the agreed scope (Artsyl, 2025). Eriksson et al. (2019) observe that long-duration infrastructure projects challenge the ability of organizations to recall past behaviors to make new decisions, i.e., organizational memory. This can be seen through the increasing difficulty for reviewers to verify decisions and costs that were committed to years earlier, unless clear and traceable documentation has been maintained throughout (Eriksson et al., 2019). This problem of organizational memory loss creates ambiguity in how contractual terms are interpreted and what constitutes sufficient evidence. These ambiguities tend to benefit the party with better records and deeper familiarity with historical decisions (Eriksson & Lind, 2015). Poppo and Zenger (2002) argue that in complex long-term relationships of this kind, formal controls and relational governance function not as substitutes but as complements. In this view, one can see it as a contract setting clear expectations for what should be delivered, while trust provides the flexibility needed for teams to collaborate on complex problems. This is relevant because it indicates that by using the rules and price lists from a contract, the owner may gain the transparency needed to prevent administrative friction from damaging the trust required for project collaboration.

A recent review of data-driven fraud detection research in public procurement by Schneider dos Santos, Machado dos Santos, Castro, et al. (2025) found that fraud and anomaly detection systems are rarely deployed in the real world, and there is typically a shortage of suitable datasets to work with. The study also underscores that practical tools capable of operating under real-world governance and data constraints are underdeveloped. Additionally, the volume of transactions involved in large infrastructure projects requires exhaustive manual review of every invoice line, which is not feasible through manual methods when thousands of lines are being processed monthly, and traditional sampling approaches are insufficiently targeted to consistently identify the most consequential deviations (Jans & Hosseinpour, 2019). These conditions create a need for scalable, systematic, and risk-sensitive methods of structuring invoice data, identifying anomalies, and directing the reviewer's attention toward items with the highest expected risk relative to the available effort (Jans & Hosseinpour, 2019). The continuous auditing literature frames this as a problem of review by exception, that is, when transaction volumes are high, assurance must be organized around risk. This allows human review to focus on the deviations that are most likely to represent genuine problems rather than spread uniformly across all items (Jans & Hosseinpour, 2019). For this logic to be implementable in practice, research suggests it must address three design conditions. First, it must prevent exception overload, that is, the volume of alerts itself becomes unmanageable and undermines the utility of the system (Svanberg et al., 2025). Second, it must produce outputs that are explainable and traceable, since public-sector audit contexts require that decisions can be documented and defended (C. A. Zhang, Cho, & Vasarhelyi, 2022). Third, it must be feasible with the available data, which, in invoice review, depends on reliably structuring the content of line items and supporting documents across various supplier formats (Yashwant, Dubey, Paikray, & Thulsiram, 2025).

The literature on invoice review automation typically highlights traditional rule-based automation, such as Robotic Process Automation (RPA), as effective in addressing the problems outlined above. This is true for repetitive tasks and structured data, but it struggles with unstructured information, such as complex line-item descriptions and varying document layouts. These types of documents are common in construction invoicing and require a different approach (Kanaparthi, 2023). In this regard, the emergence of AI, specifically machine learning (ML) and large language models (LLMs), represents an opportunity to address the limitations of manual verification and classical rule-based automation. This technological evolution marks a transition from deterministic systems to probabilistic models capable of reasoning under uncertainty (Anwar & Akeel, 2026).

Historically, audit automation relied on Computer Assisted Audit Techniques (CAATs), which codified human knowledge into static logic. This method ensured consistency, but failed to discover new patterns or adapt rapidly to changing business contexts (Houseblend, 2025). The subsequent introduction of statistical ML addressed this rigidity by utilizing supervised and unsupervised algorithms to identify hidden anomalies within large datasets. Yet these models introduced a "black box" problem, in which the rationale behind risk flags remained unclear to human reviewers (C. A. Zhang et al., 2022). The late developments of LLMs represent a third developmental phase. LLMs distinguish themselves by being able to process the semantic context of transactions, contracts, and pricing arrangements, such as the intent behind a cost claim or the interpretation of a contract clause. Thereby, LLMs can help in closing the gap between statistical probability and human understanding (C. A. Zhang et al., 2022). This transformation of raw text into structured logic enables the transition from manual sampling to full-population testing, allowing the commissioning organization to monitor the entire transaction flow for anomalies (Rozario & Issa, 2020). This allows the analytical focus to shift from simple error checking to the detection of subtle anomalies and process deviations. Additionally, AI-based anomaly detection can use learned patterns to surface unusual combinations of prices, quantities, and descriptions that would be difficult to detect manually (Tang et al., 2020). To manage the subsequent volume of potential alerts, active learning frameworks enable the system to refine its detection thresholds based on human feedback, progressively reducing false positives over time (Svanberg et al., 2025). LLMs further enable working with unstructured justifications, for example, by linking flagged invoice lines to relevant clauses and generating reviewer-ready explanations that support auditable decisions (C. A. Zhang et al., 2022; Houseblend, 2025). In summary, the outlined requirements and developments in AI point toward a risk-based, AI-supported approach as a plausible and potentially effective complement to manual review processes.

Against this background, this thesis investigates how AI and automated data analysis can support STA in reviewing invoices in large cost-plus infrastructure projects, using the West Link as the empirical context.

1.2 Problematization and Research Focus

As established in the background, invoice review under cost-plus arrangements is a governance activity whose effectiveness depends on the depth and consistency with which each claimed cost is assessed against the contract, the supporting documentation, and the agreed scope. Under this definition of effectiveness, exhaustive manual review becomes highly ineffective in large infrastructure projects under cost-plus arrangements (Kanaparthi, 2023; Eriksson & Lind, 2015). In large projects like the West Link, the review team receives massive volumes of individual supporting documents each month, spread across consolidated invoices from multiple contractors and subcontracting tiers. The consequence is that the review often cannot be organized as comprehensive checking, but as sampling, where reviewers direct attention toward items flagged on the basis of size, novelty, or area (Rozario & Issa, 2020). This means that a substantial share of invoice lines is processed with limited scrutiny, and that patterns of systematic over-billing or contractual non-compliance at the subcontractor level may not be detected at all.

Continuous auditing research suggests that when transaction volumes are high, assurance cannot rely on checking every item, but must instead be organized as review by exception, where human attention is directed toward the deviations most likely to represent genuine problems (Jans & Hosseinpour, 2019; Svanberg et al., 2025; P. Li, Chan, & Kogan, 2016). For this logic to function in practice, three linked design conditions must be satisfied simultaneously. The first condition is that the system must prevent exception overload. If automated detection produces more alerts than reviewers can process, the volume of flags itself becomes unmanageable and undermines the utility of the system

(Svanberg et al., 2025). Assuming an at-capacity invoice review team under the manual sampling approach, a tool that simply surfaces more items without prioritizing them would substitute one form of overload for another. The second condition is that outputs must be explainable and evidence-linked. In a public-sector audit context, decisions must be documentable and defensible, meaning that any flagged item must be traceable to a contract term, a supporting document, or a recorded approval (C. A. Zhang et al., 2022). This is important since the difficulty is not only identifying a potential discrepancy, but being able to link it clearly enough to the contractual basis to justify withholding payment or requesting a correction (Anwar & Akeel, 2026; C. A. Zhang et al., 2022). The third condition is that the approach must be feasible on the data actually available. Invoice review in construction settings involves heterogeneous document formats, multi-tier subcontractor invoices, and semi-structured line-item descriptions across varying supplier templates, and weak extraction from these sources undermines both anomaly detection and the ability to document the decision trail (Yashwant et al., 2025). The literature suggests that the three outline conditions must hold. The empirical data collection in this study examines whether and how these manifest in STA's specific context, and whether additional conditions emerge.

Understanding how these and potentially additional conditions can be satisfied by an AI-based invoice review system in STA's context requires a grounded understanding of how invoice review is currently performed in practice. The specific decision points reviewers rely on, the evidence they require, the recurring traceability gaps, and the data constraints imposed by existing systems and workflows are not documented in the academic literature for cost-plus construction contexts of this kind. Thus, STA's West Link project serves as the empirical case through which these questions are investigated. Its scale, contractual structure, and the control challenges associated with it make it a representative setting for studying invoice review under conditions where AI-based support is both necessary and difficult to implement.

Based on this, the research focus is twofold. The first part uses the West Link as the empirical context to develop a practitioner-grounded understanding of STA's current invoice review process, covering its decision points, pain points, traceability challenges, and data constraints. Thus, the main limitations and control gaps can be understood. The second part translates these findings into a feasible first-stage concept for AI-based support of invoice review in the context of STA. In line with this problematization and research focus, two research questions have been formulated.

Table 1.1: An overview of the research questions

RQ	Description
1	What are the main limitations and control gaps in the invoice review process that contribute to cost deviations in large public cost-plus infrastructure projects?
2	How could a conceptual framework for AI-based invoice review be designed to improve the identification of cost deviations in a commissioning organization's invoice review process in large public cost-plus infrastructure projects?

1.3 Research Scope and Delimitations

This thesis is concerned with STA's perspective, i.e., the commissioning organization, in large infrastructure projects operating under cost-plus arrangements, and does not address the contractor's internal cost management or procurement practices. Because of this, the findings reflect a bounded perspective focused on project ownership, and the results should be interpreted as such rather than a comprehensive account of the entire supply chain. The research also focuses specifically on the mechanics of the invoice review process itself, rather than investigating other external or project-level drivers of cost deviations in detail. The empirical analysis is further bounded by its single-case design, meaning that the findings are derived exclusively from STA's West Link project and are not intended to be statistically generalizable across organizations or project types.

The work does not constitute a financial audit of any project or supplier, nor does it seek to establish or prove the existence of specific fraud in any dataset. Furthermore, it does not deliver a deployed production system. More broadly, this is not a technical or computer science report. The thesis does not evaluate specific algorithms, benchmark model architectures, or produce deployable code. The contribution is management- and governance-oriented. It identifies organizational requirements, structural barriers, and a conceptual framework for integrating AI-based tools into existing workflows. Assessments of AI capabilities draw on published benchmarks and existing literature rather than empirical technical testing. In addition, the achievable scope is constrained by data access, confidentiality requirements, and the availability of structured information related to contracts, approvals, and supporting documentation. Also, the scope is limited to AI-based solutions rather than broader governance-related or project control tools.

2

Theoretical Background

This chapter introduces both the theoretical frameworks and the technological background concepts that will guide the study's analysis and the development of a first-stage framework for AI-supported invoice review in cost-plus projects. The chapter covers four areas:
(i) Drivers of Cost Deviations in Large Infrastructure Projects: This section explains why megaprojects are systematically prone to cost overruns. It explores psychological and strategic drivers, such as optimism bias and strategic misrepresentation, the impact of power asymmetry in consolidated supplier markets, and the information asymmetry and moral hazard inherent in cost-plus arrangements.
(ii) Invoice Review as a Governance Tool: This area positions invoice review within the broader purchase-to-pay process and highlights its critical role in preserving organizational memory over long project durations. It breaks down traditional manual workflows and identifies recurring performance gaps, such as exception handling, approval bottlenecks, and the cognitive overload caused by complex line-item verification.
(iii) Design Constraints and Requirements: This section outlines the necessary conditions any automated review system must satisfy to be viable in practice. It details the need for risk-based assurance to prioritize attention and prevent exception overload , the requirement for explainable outputs to maintain trust and support human-in-the-loop decision-making , and the critical role of a reliable data foundation.
(iv) Contemporary Methods for Automated Invoice Review: Finally, this section introduces the applied technological background concepts necessary for understanding modern solutions. It traces the evolution from traditional optical character recognition (OCR) and rule-based three-way matching to the emerging capabilities of Large Language Models (LLMs). Furthermore, it covers advanced architectural frameworks, such as Retrieval-Augmented Generation (RAG) and Agentic AI workflows, exploring how modern systems handle unstructured data and execute multi-step verification tasks autonomously.

2.1 Drivers of Cost Deviations in Large Infrastructure Projects

2.1.1 Project Governance and Strategic Behavior

As established in the background, cost deviations are creating a practical need for control approaches that remain effective under conditions of high project complexity (Flyvbjerg et al., 2002). Because cost underestimations occur so frequently, in nine out of ten cases according to Flyvbjerg et al. (2002), they cannot be dismissed as technical errors or bad luck. This makes it evident that the problem of cost overruns is at least partly rooted in governance and human behavior during the planning and implementation phases. Flyvbjerg et al. (2003) describe a paradox for infrastructure "megaprojects", where they have been growing in size but continue to fail in performance. Flyvbjerg (2008) argues that there are two key drivers of cost overruns: strategic misrepresentation and optimism bias. Strategic misrepresentation is described as the economic incentive to underestimate the cost to win project approval over competitors. In the context of a large infrastructure project, this could manifest as contractors intentionally bidding low, creating an information asymmetry where the commissioning organization relies on misleading documentation. Meanwhile, optimism bias is a psychological phenomenon in which planners tend to underestimate risks and, therefore, costs. Flyvbjerg (2008) advocates for "Reference Class Forecasting" to mitigate this problem of strategic misrepresentation and optimism bias. This means using data from past similar projects in order to benchmark cost estimates. However, Flyvbjerg (2018) contrasts this with Hirschman's principle of the "hiding hand". This theory holds that both the costs and benefits of projects are initially underestimated, but that technical project difficulties often lead to problem-solving that results in benefit overruns that outweigh cost overruns (Flyvbjerg, 2018). In spite of this, Flyvbjerg (2018) argues that the hiding hand occurs in only about 20% of projects and is thus more of an exception, and that planning fallacies such as optimism bias and strategic misrepresentation are the typical outcome.

2.1.2 Power Asymmetry in Complex Cost-Plus Projects

Megaprojects are defined as large-scale, complex ventures typically costing between USD 500 million and 1 billion or more, characterized by high uncertainty, long durations, and significant political stakes (Brunet, 2021). By some, a project can also be classified as "mega" if its budget represents 0.01% to 0.02% of a government's GDP (Hu, Chan, Le, & Jin, 2015). Complexity in this context arises from the interplay between multiple public and private stakeholders and the transformational nature of the infrastructure (Brunet, 2021). To further explain this concept, Geraldi, Maylor, and Williams (2011) propose a framework that divides project complexity into five distinct dimensions: structural, uncertainty, dynamics, pace, and socio-political complexity. Understanding these dimensions is important, as they dictate how an organization must respond in terms of managerial capacity, strategic choice, and process design (Geraldi et al., 2011).

In these settings, the asymmetry of power often shifts toward suppliers due to three theoretical drivers. Power asymmetry, as defined by Zarei, Hui, Duffield, and Wang (2017), refers to an imbalance in power between project actors, where one party holds greater authority or influence than its working partners. Importantly, Zarei et al. (2017) argue that power in project delivery should not be understood only as formal authority, but as a combination of authority and competence. Authority, according to Zarei et al. (2017) refers to the formal ability to approve, fund, direct, or reject decisions, while competence refers to the knowledge, expertise, information, and practical capability required to make informed decisions. One source of power asymmetry is capability gaps, in which project owners frequently struggle to become "strong owners" and lack the strategic and commercial capabilities required to manage high-level supplier relationships (Winch, 2014; Winch & Leiringer, 2016). Another driver is information asymmetry, where contractors possess superior insight into actual work composition and costs, while the commissioning organization must rely on documentation and inspections (Eriksson & Lind, 2015). Market oligopoly is another driver. This is highlighted by the fact that while the construction industry features a high number of small firms, the market often behaves as an oligopoly at the regional level (Uusitalo, Peltokorpi, Seppänen, & Alhava, 2024). This is further supported by an analysis of the Swedish construction market by Grahn (2024), who found that it can be described as an oligopoly. These few, large, dominant companies are described as having a larger impact on the market in terms of pricing and setting standards (Grahn, 2024). Viewed through the lens of Porter's five forces, this type of consolidation among suppliers increases the bargaining power of suppliers, granting them the leverage to dictate terms and potentially drive up project costs (Porter, 1980).

The degree of power asymmetry is further influenced by how project actors make sense of the governance framework. Sensemaking is the process through which individuals or groups interpret novel and ambiguous situations (Brunet, 2021). In megaprojects, actors at different levels operate with conflicting priorities. At the institutional level, actors focus on legitimacy and ensuring procedures are acceptable to the public (Brunet, 2021). Meanwhile, at the organizational level, they prioritize accountability through legal and procedural devices (Brunet, 2021). At the project level, efficiency and maximum productivity are prioritized (Brunet, 2021).

Tensions arise because the heavy documentation required for legitimacy, often described by practitioners as a "monster, because it's big, it's heavy to manage", can paralyze project-level efficiency (Brunet, 2021, p. 410). This fragmented sensemaking creates "gray areas" in contract interpretation. When project actors prioritize "cutting corners" to maintain efficiency, they may bypass formal controls, effectively ceding more power to the contractor who exploits these interpretative gaps to drive cost deviations (Pinto, 2014).

2.1.3 Information Asymmetry in Cost-Plus Contracts

Cost-plus contracts are often used when uncertainty and scope complexity make fixed pricing difficult (Suprapto et al., 2016). In these settings, the commissioning organization's ability to ensure correct payment depends on documentation quality and on consistent decisions in invoice review, since the commissioning organization must verify costs with limited direct observability into underlying work and cost structures (Eriksson & Lind, 2015). A central challenge is therefore information asymmetry (Eriksson & Lind, 2015). The contractor typically has better insight into actual work and cost composition, while the commissioning organization must rely on evidence, traceability, and review routines to judge allowability and correctness. The client also has the entire project to consider, whereas the contractors only have their part (Eriksson & Lind, 2015).

Eriksson and Lind (2015) describe this challenge through a moral hazard lens in construction procurement. They discuss how commissioning organizations use different strategies to reduce exposure, including formal monitoring. They also argue that monitoring becomes harder and less effective in large, complex projects, creating a practical need for control approaches that remain workable under such conditions. Suprapto et al. (2016) add that outcomes in major projects are shaped by relational mechanisms and collaboration quality, not only by contract form.

Public procurement is frequently described as exposed to fraud and corruption risks, but research also shows that detection in practice is constrained by limited dataset availability, limited real-world deployment, and the need for explainable and auditable outputs (Schneider dos Santos et al., 2025; Jans & Hosseinpour, 2019; C. A. Zhang et al., 2022). Schneider dos Santos et al. (2025) map data-driven fraud detection research in public procurement and show that many studies focus on collusion and favoritism, while also highlighting limited real-life deployment and limited availability of suitable datasets.

2.2 Invoice Review as Governance Tool

2.2.1 The Purpose of Invoice Review

Invoice review, invoice verification, invoice validation, or accounts payable (AP) processing, is the process of ensuring that invoices are accurate and legitimate before approving them for payment (Artsyl, 2025). This process includes cross-referencing invoice details with other documentation and contractual agreements, such as purchase orders (PO) and delivery receipts. According to Neuvonen (2015), invoice review is manually performed in many organizations and is often considered the most resource-intensive part of finance departments. Figure 2.1 shows where in the purchase-to-pay cycle the invoice review usually takes place. The purchase-to-pay process itself encompasses all activities from an organization purchasing goods and services to paying for them.

Figure 2.1: Invoice review as part of the purchase-to-pay process (Neuvonen, 2015)

As Eriksson et al. (2019) observe, long infrastructure projects "challenge the collective memory of the organization", making it difficult for current auditors or reviewers to verify decisions that were made several years ago, unless they have access to clear documentation. This loss of "organizational memory" creates ambiguities in the contractual agreements and may lead to delays and conflicts between the commissioning organization and the supplier (Eriksson et al., 2019). Hence, invoice review not only serves as a tool for financial control but also as a governance tool to mitigate the risk of knowledge loss in the project over time.

Poppo and Zenger (2002) highlight that a common theoretical tension in collaborative projects is whether formal controls, such as a rigorous invoice review, weaken the trust between the commissioning organization and the supplier. However, Poppo and Zenger (2002) also argue that formal contracts and relational governance act as complements rather than substitutes.

2.2.2 Traditional Invoice Processing Workflows

Figure 2.2 illustrates what the invoice verification process usually looks like in modern AP teams, according to Artsyl (2025). The workflow typically begins with the receipt of an invoice, which is commonly received in four different formats (Neuvonen, 2015):

1. Paper invoice
2. Electronic invoice
3. E-mail invoice
4. EDI invoice

The next step in the process includes an initial verification, manually checking the invoice for completeness and compliance, and capturing key metadata such as supplier, date, and currency amount (Artsyl, 2025). This often also includes manually entering details into an accounting or finance system (Bikute, 2024). Thereafter, the invoice is compared with PO and contracts. The goal here is to confirm that the information on the invoice matches what was agreed, and common discrepancies include price, quantity, missing receipt, and incorrect tax amount (Artsyl, 2025). If there is a discrepancy, it is flagged and put into a resolution queue to be handled by the right owner. The next step is validating invoice details to confirm they adhere to internal systems and policies and are ready to be approved and processed for payment. Common details to verify in this step include bank details, payment terms, approval limits, and required documentation (Artsyl, 2025). In the event of an issue, structured exception workflows should capture the reason and the requested corrections. Here, maintaining an auditable trail is important to reduce rework (Artsyl, 2025).

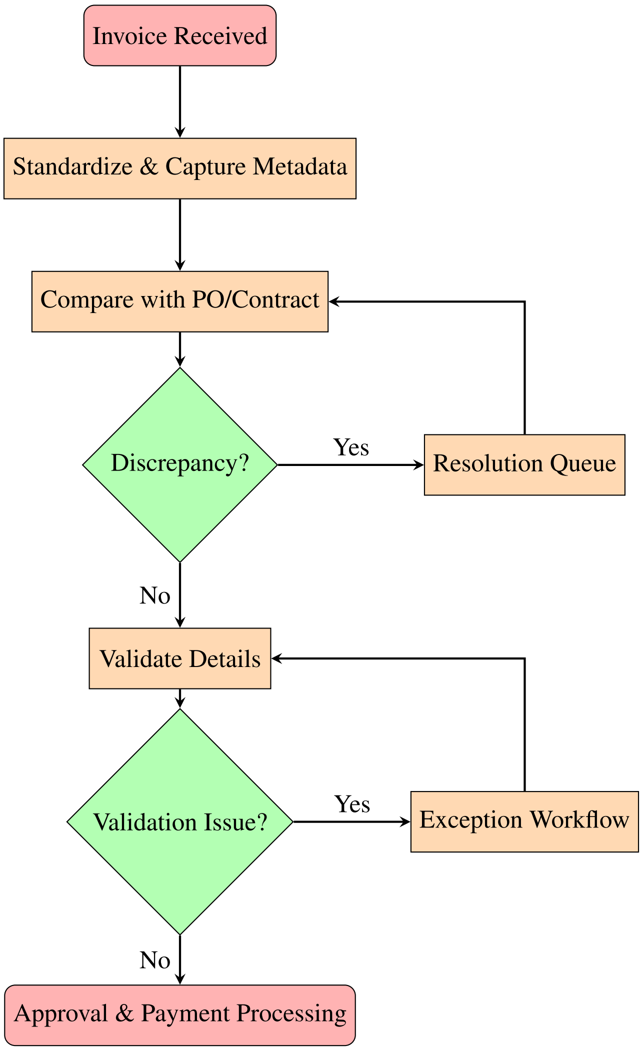

Figure 2.2: Steps involved in invoice verification (Artsyl, 2025)

In addition to the technical verification of invoices, i.e., determining if the data is correct, there is also a challenge in ensuring that the cost is allowable or reasonable for the project. This is known as reimbursability verification. According to Bajari and Tadelis (2001), there is a fundamental difference between fixed-price contracts and other cost-sharing contracts, such as cost-plus contracts. Whereas a fixed-price contract does not require measuring of construction costs, this is needed for cost-plus contracts (Bajari & Tadelis, 2001). Bajari and Tadelis (2001) highlight that for complex cost-plus projects where changes happen often, the buyer needs to determine what costs are reasonable before reimbursing the contractor. Bajari and Tadelis (2001) also argue that the contractors have superior information about the costs of changes. One can therefore conclude that the buyer faces information asymmetry when determining the reimbursability of these expenses. This is also supported by Winch (2010), who argues that managing construction projects is partly defined by the lack of information needed for decision-making.

2.2.3 Performance Gaps and Pain Points

According to benchmarking reports from Bartolini (2024), IOFM (2021), IFOL (2020), and Medius (2025), traditional invoice processing shows performance gaps across organizations. The gaps concern a few recurring areas, most notably, exception work, manual handling, approval bottlenecks, and supplier communication overhead. Figure 2.3 shows which operational issues AP teams most frequently report as limiting performance. A high share of respondents point to a high percentage of exceptions (53%) and slow invoice or payment approvals (41%) as the main constraints, followed by fraud risk (29%). Lower but still material shares report staffing shortages (24%) and limited visibility into invoice and payment data (22%) (Bartolini, 2024).

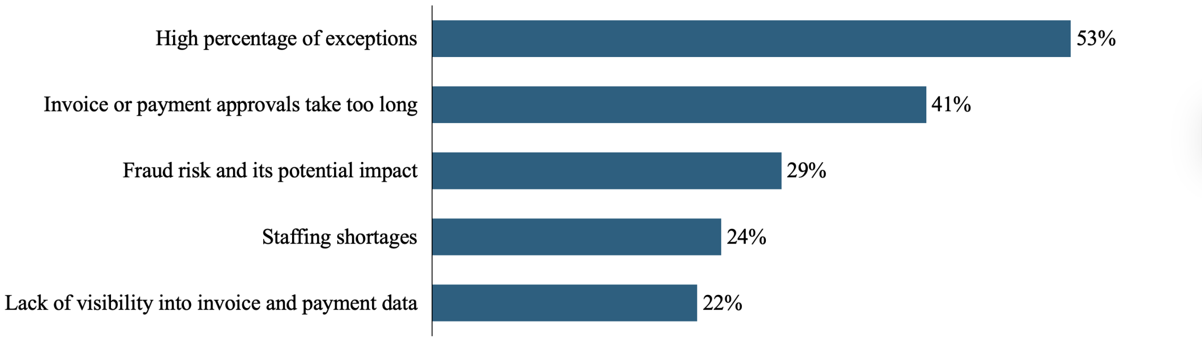

Figure 2.3: Reported constraints on invoice processing performance, share of respondents citing each issue (Bartolini, 2024)

Bartolini (2024) suggests that differences in results are driven by the frequency with which invoices deviate from the intended processing path and require manual intervention. The author reports market-level baseline values of USD 9.40 cost per invoice, 9.15 days processing time, a 14.0% exception rate, and a 32.6% straight-through processing rate. The straight-through processing rate is defined as the share of invoices processed without manual intervention, which functions as an overall outcome metric for how well capture, matching, routing, and rules work together. These values suggest that manual handling and exception management account for a significant portion of the total workload in many AP functions. Best-in-class organizations process invoices at USD 2.78 per invoice, while all others report USD 12.88, and processing time ranges from 3.1 days to 17.4 days. The same benchmarking comparison shows an exception rate of 9.0% for best in class compared with 22.0% for all others, and straight-through processing of 49.2% compared with 23.4%.

Cycle time gaps show that approvals and routing are a key constraint. Medius (2025) reports a non-PO invoice cycle time of 7.4 days on average versus 2.1 days best in class, and an approval time of 3.3 days on average versus 1 day best in class (Medius, 2025). These differences are consistent with upstream automation gaps that increase manual touches, including lower touchless capture and lower automatic routing in the average group (Medius, 2025). Cost gaps follow the same pattern with IFOL (2020) reporting an average cost per invoice of USD 7.90 for limited automation, USD 4.31 for moderate automation, and USD 3.40 for significant automation. This supports the link between automation coverage and lower processing cost.

Manual handling affects data quality in traditional invoice processing, mainly because information is re-entered, interpreted, or routed with incomplete context. Bartolini (2024) describes manual and paper-based processing as more error-prone and links this to time spent on manual data entry and exception resolution rather than straight-through flow. Kanaparthi (2023) further argues that duplicate invoice payments are a frequent occurrence in the traditional invoice review. This error is largely a consequence of the highly resource-consuming nature of manually reviewing invoice lines (Kanaparthi, 2023). Beyond general processing delays, the verification of individual line items is identified as a major time sink. This bottleneck is especially prevalent when organizations manage huge supplier bills that include pages of line-item data. The challenge is often driven by a mismatch in catalogs, where buyers and sellers use different terminology for the same goods, creating discrepancies between invoice descriptions and POs. Because traditional processes frequently struggle with these unstructured variations and shortened names, the process remains labor-intensive and highly prone to mistakes (Kanaparthi, 2023). Bartolini (2024) further argues that quality issues in manual invoice review appear as exceptions and rework. When exception rates are higher, more invoices require investigation, clarification, and correction, which increases cycle time and resource use. Using IFOL's benchmarking categories, moving from limited automation to significant automation is associated with a reduction in average cost per invoice from USD 7.90 to USD 3.40, a difference of USD 4.50 per invoice (IFOL, 2020). This can be explained by manual handling increasing the number of touches per invoice through re-entry, correction, and exception resolution. At the same time, higher automation reduces those touchpoints by keeping more invoices on a straight-through path (Bartolini, 2024).

2.3 Design Constraints and Requirements

2.3.1 Introduction to Automated Invoice Processing

In this paper, automated invoice processing is defined as the automation of the core tasks required to move an invoice from receipt to a state of being booked, approved, and payable, with a documented audit trail. IOFM (2021) proposes a practical framework for decomposing different levels of automation into discrete process steps. The different steps that can be automated are outlined in Figure 2.4 and include invoice receipt, data capture, invoice approval, PO-to-invoice matching, and payment processing. Kanaparthi (2023) builds on this by describing a system that utilizes AI-based modules and micro-services to streamline the pipeline and reduce the manual labor typically required for these stages. IOFM (2021) defines automation maturity as the count of automated steps, where low automation means no more than two steps automated, significant automation means three steps, end-to-end automation means at least four steps, and “no automation” means none of these steps are automated. In addition, Kanaparthi (2023) distinguishes between traditional automation and AI-driven systems. Traditional approaches, such as RPA or rule-based scripts, are often limited to routine, repetitive tasks involving structured data. In contrast, AI-driven automation is designed to handle unstructured data and cognitively demanding tasks, such as complex line-item descriptions that classic scripts cannot process (Kanaparthi, 2023).

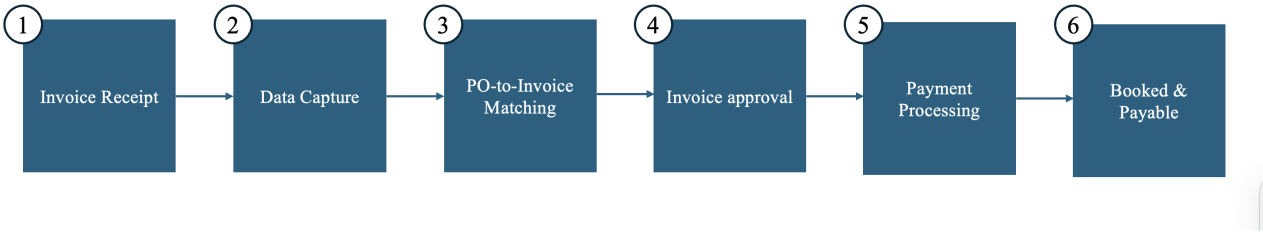

Figure 2.4: The invoice process (IOFM, 2021)

To further define the automation of invoice processing, the following two boundaries are important. First, automation includes the capture and structuring of invoice information, where Medius (2025) defines the "touchless capture rate” as the percentage of invoices captured without human involvement. Second, automation includes routing and approvals, where Medius (2025) defines the “automatic routing rate” as the percentage of invoices automatically routed to the correct approver. In addition to these metrics, Bartolini (2024) defines “straight-through processing rate” as the share of invoices processed without manual intervention, which functions as an overall outcome metric for how well capture, matching, routing, and rules work together. Kanaparthi (2023) notes that achieving this state requires the system to handle the entire pipeline, from receipt through payment, with "little or no human interaction". This is concretized through confidence thresholds, where the system automatically validates a step only if the AI’s prediction confidence exceeds a predefined minimum (Kanaparthi, 2023). According to the definitions and boundaries outlined, the automation of invoice processing in this study is therefore viewed as a means to reduce manual touches and exceptions, shorten lead times, and improve the predictability of results.

The effectiveness of reducing these manual touches is fundamentally tied to the distinction between rule-based logic and cognitive processing. Traditional automation, typically characterized by RPA, excels at executing repetitive tasks within structured data environments but remains limited by a reliance on rigid, predefined scripts (Gotthardt et al., 2020; Kanaparthi, 2023). Kanaparthi (2023) argues that AI-driven systems exceed these limitations by addressing the cognitive burden of processing unstructured data, such as variable document layouts and complex item descriptions. This is achieved through multi-modal models that integrate text and spatial features to interpret documents in a way classic scripts cannot (Appalaraju, Jasani, Kota, Xie, & Manmatha, 2021). Such systems move beyond simple extraction to perform advanced tasks, such as predicting accounting and VAT codes based on the underlying economic nature of a transaction (Bardelli, Rondinelli, Vecchio, & Figini, 2020). By utilizing confidence thresholds to validate these predictions, AI-driven automation enables the high levels of straight-through processing required for an autonomous pipeline (Kanaparthi, 2023).

2.3.2 Risk-Based Assurance and Scaling Logic

A shared idea across auditing and control research is that, when volumes are high, assurance should be organized around risk. Continuous auditing is one expression of this, in which information systems support more frequent monitoring and direct attention to deviations that warrant investigation. While traditional methods rely on manual sampling, modern frameworks advocate for systematic, full-population analysis. This comprehensive scope is necessary to identify rare but critical anomalies, such as duplicate payments, across the entire transaction dataset (Rozario & Issa, 2020).

However, transitioning to full-population scanning creates the risk of exception overload. Jans and Hosseinpour (2019) emphasize that the practical aim of continuous auditing is to focus human effort, not to investigate every flag the system generates. If automation produces an overwhelming number of alarms, it reduces usability and can degrade the trust required for appropriate reliance on the system. Therefore, detection and prioritization must be treated as entangled components of the same process (Lee & See, 2004).

To manage high volumes, Svanberg et al. (2025) frame the problem in terms of exception prioritization and suggest using thresholding mechanisms to control the flow of alerts. This is a filtering technique used to prioritize and limit alerts (Svanberg et al., 2025). This is operationalized through frameworks that assign suspicion scores to each exception based on the likelihood of error or fraud, ensuring auditors focus on the most important issues first. The value of these systems for invoice review lies in the better allocation of limited human resources (P. Li et al., 2016).

Achieving this level of scalability requires reengineering manual audit routines into methodical procedures specifically designed for automation rather than simply digitizing existing workflows (Alles, Kogan, & Vasarhelyi, 2024). Kanaparthi (2023) further notes that an autonomous pipeline relies on confidence thresholds to validate automated steps. By filtering out low-risk transactions, risk-based assurance provides the scaling logic necessary to handle high-volume invoice processing without a linear increase in manual oversight.

2.3.3 Explainability and Trustworthiness

Outputs produced by an automated system must be defensible, meaning they must be understandable, evidence-linked, and consistent with governance needs. In other words, they must be trustworthy. Jans and Hosseinpour (2019) treat the auditor as an active participant in the monitoring loop, supporting a human-in-the-loop framing in which human judgment remains central. Lee and See (2004) similarly argue that the human operator has the ultimate responsibility in a human-automation partnership, highlighting that trust acts as the primary attitude guiding reliance when complexity and unanticipated situations make a complete understanding of the automation impractical. Because the operator is the ultimate bearer of responsibility, the automation must be designed to be trustworthy by conveying information about its purpose, process, and performance, thereby allowing the auditor to calibrate their reliance rather than accepting outcomes blindly (Lee & See, 2004).

Explainability serves as the critical mechanism for trust calibration in complex systems where the inner workings are hidden, making a complete understanding impractical. C. A. Zhang et al. (2022) identify a lack of explainability as a barrier to adopting AI in auditing and positions explainable AI as a way to improve transparency and support audit evidence and documentation needs. This barrier aligns with the concept of disuse, which refers to failures that occur when people reject automation's capabilities because they do not trust the agent (Lee & See, 2004). To resolve this, systems must be designed to be trustworthy by revealing their purpose, processes, and performance. Svanberg et al. (2025) similarly discuss interpretability in exception outputs, such as providing local explanations that show why a specific item was flagged. In invoice review terms, explainability is the practical mechanism that allows a reviewer to justify follow-up, link the item to contract basis and approvals, and document the decision trail without extra effort (C. A. Zhang et al., 2022). Providing information about the basis of the automation's behavior is more effective than providing information about outcomes alone (Lee & See, 2004). This transparency enables the joint human-automation system to achieve performance superior to either the human or the automation acting in isolation (Lee & See, 2004).

2.3.4 The Data Foundation and Feasibility

Risk-based assurance and explainable outputs require reliable data. In invoice review, information appears in semi-structured formats across invoices, line items, and attachments. Yashwant et al. (2025) describe consistency across vendor templates as a challenge in information extraction and suggests that evaluation should include line-item complexity and generalization in addition to field accuracy. Extraction that lacks precision affects both the detection of deviations and the ability to trace the decision trail.

Data quality supports audit evidence and the credibility of decisions. C. A. Zhang et al. (2022)

note that transparency in AI relies on inputs that allow for documentation, as unclear processing can introduce bias. Explainable methods provide transparency over inputs and decisions. Lee and See (2004) connect trust in automation to the provision of information regarding system performance. Bao, Ke, Li, Yu, and Zhang (2020) indicate that the selection of raw data through theory-guided methods reduces overfitting and supports fraud detection compared to models based on financial ratios.

The use of AI in audit firms is associated with changes in audit quality and fee structures (Fedyk, Hodson, Khimich, & Fedyk, 2022). Johri et al. (2026) report anomaly detection accuracy above 87% and forecasting errors below 5% when complete datasets are used. Consequently, Jans and Hosseinpour (2019) describe the role of data as a requirement for exception handling within continuous auditing processes. Incompleteness and inconsistencies in the data affect system performance; Johri et al. (2026) emphasize the use of complete inputs for dataset analysis, while Abela (2024) describes how gaps in records limit the identification of risks.

The variety of invoice documents influences the technical feasibility of extraction. Saout, Lardeux, and Saubion (2024) note that invoice extraction requires the integration of optical character recognition with layout analysis to identify physical and logical structures. Variations in layout and formatting across different vendors lead to differences in extraction outcomes (Yashwant et al., 2025). Systematic implementation requires a structured approach to data management. Anwar and Akeel (2026) establish that data governance and integration form the foundational layer of a tiered audit workflow architecture. This layer serves as the basis for subsequent analytics and application phases in automated systems.

2.4 Contemporary Methods for Automated Invoice Review

2.4.1 Automated Data Capture and Line-Item Extraction

Automating data capture and line-item extraction are two distinct parts of the invoice review process. Data capture includes reading the invoice as a scanned or native PDF, and outputs fields such as invoice ID, customer, and total amount (Laujan, n.d.). In their study, Alkhaled and Fei (2023) proposed implementing deep learning to build an automated invoice processing system. Their proposed model, which combines advanced deep learning computer vision techniques, Optical Character Recognition (OCR), and a Convolutional Neural Network (CNN), was able to extract data such as invoice number, date, and total amount from invoices. This model achieved 88.62% accuracy on the 1,000 images it was tested on (Alkhaled & Fei, 2023). A line item is a part of an invoice that details the goods or services purchased (Thomas, 2025). This could include a description, the quantity, and the price paid. Hence, line-item extraction is a specific sub-task of data capture, which involves capturing line items and collecting them in a structured format, such as a table. Table extraction is described as more complex by Saout et al. (2024) due to the high variation in table formats. To address this, automated methods increasingly rely on Graph Neural Networks (GNNs), enabling invoices to be modeled spatially and less dependent on the layout format (Saout et al., 2024). Because data capture and line-item extraction are closely related, tools such as invoice OCR APIs or software can perform both steps (Thomas, 2025). OCR is a technology that converts images of text into machine-readable text, which can be used to extract data from documents such as invoices, as noted in Mindee (2025). However, they argue that LLMs can also be used for this, but that they are more expensive, slower, and less reliable for structured documents than OCR APIs. Instead, using LLMs for complex reasoning or summarizing, and OCR APIs for data extraction would be more cost-efficient (Mindee, 2025). Saout et al. (2024) also discuss using Named Entity Recognition (NER), or other natural language processing techniques, as a method for extracting various data from the invoice after having used OCR technology for digital format conversion.

Amari et al. (2024) tested different OCR models on a dataset in their study and compared their accuracy (exact match) and similarity (difference between two sequences) between the invoice and the extracted data. It was found that Microsoft Azure OCR is the highest-performing cloud-based solution available in terms of both accuracy and similarity. This is shown in Table 2.1. Amari et al. (2024) also propose using multimodal transformers such as LayoutLMv3 to combine the extraction of text, document layout, and visual data. This means that OCR models can be used to feed multimodal models in order to identify key fields and data in invoices by understanding both the content and its position.

Table 2.1: Comparison of OCR Models (Amari et al., 2024)

Model	Accuracy	Similarity
EasyOCR	0.284	0.719
Tesseract	0.306	0.569
Amazon Textract	0.345	0.513
Google OCR	0.576	0.892
Microsoft Azure OCR	**0.633**	**0.915**

2.4.2 Automated Validation and "3-Way Matching"

Automated validation of invoices can be defined as using rules and algorithms to ensure accuracy, completeness, and consistency of invoice details (Medius, n.d.). More specifically, it includes comparing the invoice data to supporting documents. 3-Way matching is one common method for this, where you match the invoice data to information from two other documents: a PO, and either a shipping receipt or notice of goods received (GRN) (Stampli, 2024). The invoice can be approved and processed for payment if the information matches; otherwise, the difference in information needs to be investigated. The specific information that is compared is header data (e.g., payment terms and company name) and line data (e.g., number of hours worked, descriptions of services delivered, and prices) (Stampli, 2024). Neuvonen (2015) developed a framework for automated invoice processing, where a 40-56% improvement in processing time and 70-80% reduction in error rates was anticipated, compared to manual invoice processing. In the 3-way matching used here, a 100% match was required between the invoice and PO to be automatically matched. It should be noted that this was tested specifically in manufacturing and healthcare settings. Kanaparthi (2023) examines how AI can be used for invoice validation and line item matching, and semantic similarity is highlighted as part of the approach to address common data quality issues. Methods such as lexical normalization can be used to handle typos and abbreviations in the invoice data, and TF-IDF vectors and cosine similarity can be used to match descriptions semantically between invoices and POs (Kanaparthi, 2023).

2.4.3 Automated Anomaly Detection and Risk Scoring

Manual invoice processing, relying on sampling techniques and individual skills, is limited in its capability to capture complex errors or fraud patterns (Singhal, 2021). AI or ML can be used to improve this process by providing higher coverage and learning new anomalies from the data (Singhal, 2021). This is also supported by Tang et al. (2020), who argue that anomalies and malicious attacks are difficult to notice through traditional invoice systems or manual methods. Instead, they propose an ML anomaly detection method for an e-invoice system. In their study, the authors found that the proposed methods effectively detect malicious attacks and anomalies. The accuracy ranged from 76-98% depending on specific implementations and algorithms, such as the k-means clustering algorithm and the Skip-gram based correlation analysis algorithm (Tang et al., 2020). Houseblend (2025) argues that LMMs and Generative AI can be used to generate anomaly hypotheses for invoices based on historical data, semantically analyze invoice descriptions, and summarize findings for auditors or managers. However, they highlight that based on scientific studies by Jin et al. (2024), the right prompting and training data are crucial for detecting hidden patterns in the invoice data.

2.4.4 The Accuracy of Large Language Models

LLMs are deep learning systems trained on large volumes of text data to understand and generate human language across a broad range of tasks (Zhao et al., 2023). Early LLM architectures relied on scaling parameters and training data to achieve general language comprehension (Zhao et al., 2023). The introduction of transformer models enabled the processing of sequential data to execute broad natural language tasks (Minaee et al., 2024). More recently, development has shifted toward specialized models for distinct domains to improve performance in specific applications (Yang, Zhao, Liu, & Jiang, 2025), alongside architectures designed to support deliberate, structured logical reasoning rather than heuristic outputs (Z.-Z. Li et al., 2025).

LLMs can extract data from unstructured business documents by bypassing the limitations of traditional template-based systems. Before LLMs, extracting information from invoices required rigid templates and OCR, which often fail when encountering layout variations (Saout et al., 2024). Modern automated systems for data gathering are increasingly incorporating LLMs to apply semantic comprehension and visual named-entity recognition. This allows for mapping of contextual relationships without requiring explicit programming rules (Khanchandani, Thakur, Shetty, Reddy, & Behera, 2026). Multimodal language models can process document images directly, enabling extraction accuracy that often matches or exceeds that of traditional text-based parsing (Berghaus, Berger, Hillebrand, Cvejoski, & Sifa, 2025; Shen, Yuan, Ghosh, Mai, & Dahlmeier, 2026). According to Yashwant et al. (2025), implementing these models within extraction frameworks consistently improves the reliable retrieval of targeted fields from diverse invoice formats.

A benchmark by Shen et al. (2026) extends the evaluation to multimodal LLMs, comparing image-only, OCR-only, and combined input strategies across two real-world document datasets. The results indicate that no single input strategy dominates consistently across models, and that the performance gap between open-source and proprietary models is considerable. Google's Gemini 1.5 Pro achieved the highest mean scores across all three input strategies, while models such as Claude 3.5 Sonnet and GPT-4o perform competitively in OCR-assisted conditions, suggesting that the choice of input modality interacts with model architecture in ways that are relevant to practical deployment decisions (Shen et al., 2026). The scores reported are F1 scores measured at the field level. This means that they reflect how accurately the model extracts specific target fields such as amounts, dates, and line item descriptions. Shen et al. (2026) argue that for sufficiently capable models, OCR preprocessing may be unnecessary and counterproductive. Their error analysis shows that OCR-only inputs are dominated by schema-ambiguity errors, whereas image-only inputs yield fewer total errors overall. Shen et al. (2026) attribute this to the implicit OCR capability that web-scale pretraining confers on large multimodal models, which allows them to recover text, reading order, and layout structure directly from images, preserving spatial cues that external OCR pipelines can distort or discard.

Language models can also analyze unstructured text to align financial narratives with structured accounting taxonomies. Financial auditing involves cross-referencing narrative disclosures against rigid reporting standards and numerical data (Y. Wang et al., 2025). LLMs identify semantic similarities to recommend relevant text passages for legal and financial compliance, a method that outperforms traditional zero-shot text matching systems (Hillebrand et al., 2023). Moreover, these models can process raw transaction data and unstructured financial statements to detect errors or map them to specific compliance violations (R. Wang, Liu, Zhao, Li, & Zhang, 2025; A. G. Kim, Muhn, & Nikolaev, 2024). The ability to synthesize data across multiple documents enables LLMs to support end-to-end financial workflows (R. Wang, Jiang, Yang, Li, et al., 2025), suggesting their potential for linking invoice lines to contract terms and supporting documentation at scale.

Furthermore, E. W. Kim, Shin, Kim, and Kwon (2025) show that a framework combining an LLM with a domain knowledge database can identify contractual risks and produce outputs comparable to those of experienced contract reviewers. Wong, Zheng, Su, and Tang (2024) similarly find that knowledge-augmented language models effectively identify risk clauses in construction contracts, outperforming standard models on domain-specific tasks. These findings are relevant to the invoice review context examined in this thesis, since the task of verifying whether a claimed cost is allowable under a cost-plus contract requires the same kind of reasoning across contract language, agreed terms, and submitted evidence.

Also, LLMs can weigh information from multiple heterogeneous sources simultaneously to reach a defensible judgment in high-ambiguity decisions. Z.-Z. Li et al. (2025) argue that foundational LLMs rely on fast, heuristic-driven outputs that perform well on routine tasks but fall short in scenarios requiring multi-step analysis (Z.-Z. Li et al., 2025). Reasoning LLMs such as OpenAI's o1 and DeepSeek-R1 are designed to process information step-by-step in a manner more closely resembling how a human expert would work through a problem. The performance gap between these two classes of models is substantial on tasks that demand extended logical deduction. On the AIME 2024 benchmark for complex multi-step mathematical reasoning, OpenAI-o1 scores 79.2% against GPT-4o's 9.3%, and on GPQA-Diamond, a graduate-level multi-domain reasoning benchmark more analogous to contract judgment, o1 scores 75.7% against GPT-4o's 49.9% (Z.-Z. Li et al., 2025). In financial contexts, R. Wang, Liu, et al. (2025) show that LLMs can manage full-pipeline workflows involving simultaneous reasoning across contracts, transaction data, and supporting documents (R. Wang, Jiang, et al., 2025). This would indicate that a reasoning LLM is better suited to flagging likely violations and surfacing the relevant contract clauses and evidence for a reviewer, whereas a foundational LLM operating heuristically is more prone to missing subtle or multi-conditional non-compliance. That said, even reasoning LLMs underperform human experts on tasks requiring numerical precision and extended deduction (Z. Zhang, Cao, & Liao, 2025), which reinforces the human-in-the-loop design condition discussed in Section 2.3.

LLMs exhibit limitations that are consequential for their deployment in professional invoice review contexts. Benchmark evaluations indicate that while LLMs perform well on text-based queries, they perform worse than human experts in scenarios requiring extended logical deduction or numerical reasoning (Z. Zhang et al., 2025). This limitation has implications for two of the design conditions identified in Section 2.3. It reinforces the need for explainable outputs, since a model that cannot reliably account for its reasoning cannot produce the evidence-linked flags necessary for auditable decisions (C. A. Zhang et al., 2022). Additionally, it supports the case for a human-in-the-loop architecture, in which the LLM handles semantic matching and anomaly flagging while human reviewers retain responsibility for final judgment. Deployment in environments involving sensitive project data further requires rigorous governance frameworks to manage algorithmic outputs and ensure that model behavior remains consistent with organizational accountability requirements (Wondra, 2025).

2.4.5 Agentic AI Best Practices

Agentic AI refers to systems that utilize LLMs to direct and execute multi-step tasks with a degree of autonomy. Joshi (2025) describes agentic AI as autonomous entities capable of perceiving their environment, making decisions, and taking actions to achieve specific goals. The authors also note that such systems are increasingly being used in high-complexity domains such as financial verification, risk management, and fraud detection. This autonomy is what distinguishes agentic systems from conventional automation, which relies on deterministic, rule-based processes. Where a traditional rules engine operates like a checklist, flagging outcomes based on preset criteria, an agentic system functions more like an experienced analyst that evaluates context, recognizes subtle patterns, and reaches judgments even in situations where clear-cut rules are absent (OpenAI, n.d.).

According to Anthropic (n.d.) and OpenAI (n.d.), there is an important architectural distinction within agentic systems between workflows and agents. Workflows are systems in which LLMs and tools are orchestrated through predefined code paths, while agents are systems in which an LLM dynamically directs its own processes, tool usage, and sequencing based on what it encounters during execution (Anthropic, n.d.). Workflows offer predictability and consistency for well-defined tasks, while agents are better suited to scenarios that require flexible, model-driven decision-making at scale (Anthropic, n.d.). For the majority of applied contexts, including structured document review, a workflow architecture with predefined sequencing is often more appropriate than a fully autonomous agent, as it allows for better oversight and auditability of each processing step (OpenAI, n.d.).

OpenAI (n.d.) proposes that agentic systems, as opposed to rules-based systems or a single agent instance, are best suited to workflows where three conditions apply. The first condition is that the decision-making is complex and involves nuanced judgment, exceptions, or context-sensitive reasoning. The second condition is that the rule sets are too dynamic or numerous to be maintained as static logic. The last condition is that the processes involve large volumes of unstructured or semi-structured data.

OpenAI (n.d.) identifies the model, tools, and instructions as the three important aspects of agent design. The model provides the reasoning and language understanding layer, tools provide access to external data and the ability to take actions, and instructions define the behavioral scope and constraints within which the agent operates. Anthropic (n.d.) frames these same pillars around the concept of an augmented LLM, a language model enhanced with retrieval, tool use, and memory. Anthropic (n.d.) identify this as the basic building block from which both workflows and agents are constructed. Jiang et al. (2023) similarly demonstrate how augmented LLM systems can improve over single-pass approaches by dynamically retrieving context across multiple steps, showing that iterative, multi-step information access substantially outperforms single-time retrieval in knowledge-intensive tasks.

Anthropic (n.d.) identifies three core principles to maintaining simplicity in agent design, prioritizing transparency by making the agent's reasoning steps explicit, and carefully designing the agent-tool interface through thorough documentation and testing. A key implication of the simplicity principle is that complexity should only be added when it demonstrably improves outcomes. According to the authors, this implies that one should begin with a single-step or single-agent implementation and expand to multi-step or multi-agent orchestration only when the simpler approach is proven to fall short (Anthropic, n.d.; OpenAI, n.d.). Jiang et al. (2023) support this principle through their finding that unnecessary retrieval steps can introduce noise and reduce performance, suggesting that retrieval should be triggered selectively based on model confidence rather than at a fixed frequency.

Guardrails constitute another foundational best practice, functioning as a layered defense mechanism to ensure the agent operates within safe, intended boundaries (OpenAI, n.d.). For document processing applications, these may include relevance classifiers that constrain the agent to its defined task domain, tool-level safeguards that assess the risk of an action before it is executed, and thresholding mechanisms that regulate the volume of exceptions directed toward human reviewers. Anthropic (n.d.) recommends that agents pause for human feedback at checkpoints or when encountering situations that exceed their defined operational bounds. OpenAI (n.d.) similarly emphasizes that human intervention remains a critical safeguard, particularly for high-risk actions or when the agent fails to resolve a task within expected parameters.

2.4.6 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a framework that retrieves relevant information from external knowledge bases before answering prompts with LLMs (Gao et al., 2023). This approach enhances the accuracy of the LLM's responses and reduces hallucinations according to Gao et al. (2023), making it especially beneficial for knowledge-intensive applications. Furthermore, RAG allows users to verify the accuracy of the provided answers because it can cite its sources, thereby increasing trust in the outputs (Gao et al., 2023). Despite these advantages, most RAG setups typically execute retrieval only once based on the initial input, limiting the model's effectiveness in scenarios requiring continuous generation of long texts, according to Jiang et al. (2023). Additionally, assessing the performance of these systems presents challenges, as end-to-end evaluation often lacks transparency regarding which specific retrieved document contributed to the final generated response (Salemi & Zamani, 2024). The resource-intensiveness of RAG systems, both in terms of time and computation, especially with large amounts of retrieval results, is also a limitation of using RAG (Salemi & Zamani, 2024).

3

Methodology

This chapter details the methodological approach used to conduct the study. It explains the choice of an abductive, single-case study design and then provides a breakdown of the overall research process. Furthermore, it outlines how qualitative data was collected and analyzed using a Gioia-style framework, and concludes with a discussion on research quality and the use of AI during the project.

3.1 Study Approach

In this study, an abductive research approach was used, as the purpose was to understand how AI-based methods can be utilized to improve invoice review in a context where the key mechanisms were not fully understood in advance. Dubois and Gadde (2002) describe abduction as a systematic combination of theory and empirical material where the researcher iteratively moves between the two to discover new relationships and explanations. This study aligns with the research, as the interviews were expected to reveal how risk is assessed in practice, what evidence is considered sufficient, and where traceability breaks down. Meanwhile, the literature was used in parallel to structure and refine the emerging framework and design requirements.

An inductive approach was not suitable because the study does not begin from open-ended observations aimed at constructing a new theory from the ground up, but rather enters with an existing problem context, prior frameworks, and a defined design objective. A purely deductive approach was equally unsuitable because it presupposes a set of fixed propositions to be systematically tested against empirical data, whereas the problem formulation in this study remained partly emergent throughout the research process. Instead, the study was designed as a qualitative interview-based investigation, which aligns with Blomkvist and Hallin (2015), who describe qualitative methods as including data gathering through interviews and observations. The interviews were appropriate here because the most critical inputs were practitioners' judgment, operational constraints, and governance requirements, which needed to be captured directly from the stakeholders to produce design requirements.

3.1.1 Case Selection and Empirical Setting

Yin (2018) argues that a single embedded case study is the appropriate design when the aim is to investigate a contemporary phenomenon in a real-world context where the boundaries between the phenomenon and its context cannot be clearly separated. This applies well to invoice review in a live infrastructure project, since the review practices, data constraints, and governance requirements are hard to separate from the organizational setting in which they operate. For these reasons, a single case study was chosen over a comparative approach. Furthermore, an embedded case study design was chosen rather than a holistic case study because the West Link consists of multiple sub-projects with partly different invoice review routines, making the internal variation within the case analytically relevant. The aim of this study is not to produce statistical generalizations across organizations, but to develop a grounded understanding of requirements and feasibility in a specific high-complexity setting, for which a single case provides stronger analytical leverage than a broader but shallower comparison across multiple sites (Yin, 2018). The West Link was selected as the empirical case for this study because it exemplifies the megaproject characteristics and cost escalation patterns outlined in Section 1.1.

Figure 3.1: Map of West Link projects (Trafikverket, n.d.)

Within the West Link, the empirical focus was primarily on the Haga and Korsvägen sub-projects. These sub-projects were selected because interviewees provided detailed insight into their invoice review routines, and because they illustrate variation within the overall case. Haga was used to understand the operational workflow through interviews and direct observation, while Korsvägen provided insight into a sub-project with documented cost-control challenges and more extensive review documentation.

Many of the sub-projects within the West Link operate under a cost-plus contract structure across several major contractors, with multi-tier subcontracting arrangements that reduce the commissioning organization's visibility into cost composition and create conditions where markups accumulate across tiers in ways that are difficult to trace without strong documentation routines (Trafikverket, n.d.; Suprapto et al., 2016). Monthly invoice volumes exceed what is feasible to review exhaustively by manual means, and the project's multi-decade duration compounds the problem of organizational memory loss described by Eriksson et al. (2019). The West Link also provides an empirically grounded starting point through the independent review of the Korsvägen sub-project conducted by Delray Consulting AB (2025), which documented a cost deviation of SEK 4.1 billion and attributed it in part to insufficient supporting documentation, weak traceability between invoice lines and contractual basis, and inadequate pre-approval review processes. By these factors, the West Link represents the conditions under which the previously identified design requirements are consequential, and under which an AI-based approach to invoice review appears both most needed and most difficult to implement.

The findings of this study are not intended to be statistically generalizable to a population of projects. They are instead analytically generalizable, meaning that the requirements framework and implementation concept developed here can be transferred to other large infrastructure projects operating under cost-plus arrangements with similar contractual structures, transaction volumes, and governance constraints (Yin, 2018).

3.2 Research Process

The research process was organized into two sequential and iterative stages, each comprising a data collection and analysis component. The stages were designed to address each research question, with the first stage addressing RQ1, and the second stage addressing RQ2. This is illustrated in Figure 3.2. Stage 1 aimed to develop a requirements framework for automated invoice review through exploratory interviews, a targeted literature review, and solution-targeted interviews that were iteratively refined through reference group feedback and consolidated in a reference group workshop. Stage 2 builds on stage 1 by using and refining the requirements framework to build a first-stage implementation framework for AI-supported invoice reviews. This was done iteratively through feedback from a reference group and by iteratively revisiting the literature. The reference group consisted of stakeholders at STA, including some of the respondents. Across both stages, analysis was conducted in parallel with data collection to synthesize findings, validate emerging interpretations, and inform adjustments to subsequent activities, thereby maintaining a structured process while enabling iterative learning and incorporating new insights as needed.

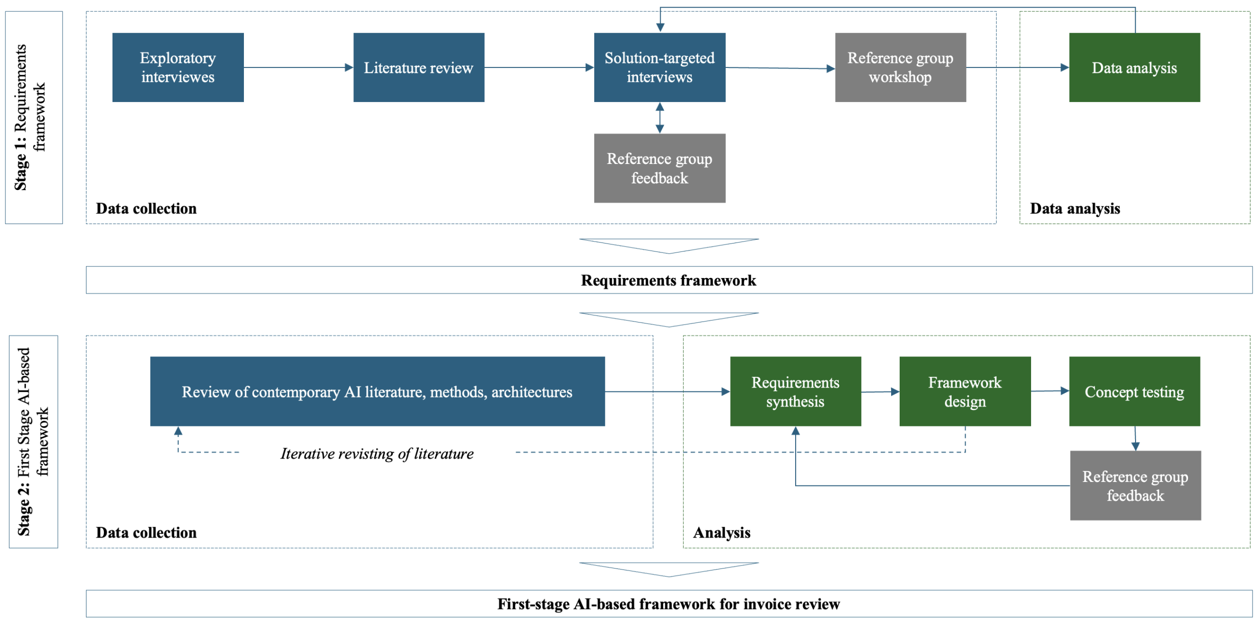

Figure 3.2: Research process

3.3 Literature Overview

The literature overview was conducted to build the theoretical background used in this study. The search targeted literature that helped explain invoice review in a cost-plus context, as well as contemporary invoice review automation best practices to help inform the proposed solution to STA.

Following the approach suggested by Bryman and Bell (2011), articles and books already known to the project or recommended by others were used first. From this reading, keywords were extracted, which were later used to expand. Searches were conducted across library resources and relevant electronic databases, including Google Scholar. Titles and abstracts were screened for relevance before full texts were retrieved and reviewed. Reference lists in key papers were used to identify additional sources, and the search was revisited to capture relevant newer publications during the project.

Because the study addresses a broad, applied topic with evolving boundaries, the review was not fully systematic. This is consistent with the discussion by Bryman and Bell (2011) about limitations of systematic reviews in such settings.

3.4 Data Collection

3.4.1 Interviews

A semi-structured interview approach was selected because it enabled coverage of key topics while still allowing flexibility to pursue relevant issues that emerged during the conversation. Furthermore, an interview guide was used to ensure that comparable areas were addressed between respondents. It was built such that the ordering and wording of the questions could vary, and additional probes could be introduced when the responses indicated that a topic required deeper exploration (Bryman & Bell, 2011). This design supported an emphasis on the interviewee's own framing of work practices and constraints, rather than imposing a rigid structure that could limit insight into how participants understood their context and responsibilities (Bryman & Bell, 2011)

Respondents were selected through purpose sampling, which means that participants were chosen based on their expected ability to contribute relevant and rich information about invoice review practices, contract interpretation, control activities, and support systems and information flows (Bryman & Bell, 2011). The respondent group included in stage 1 included roles spanning project management, finance, legal, and IT. Snowball sampling, as described by Bryman and Bell (2011), was also used by asking initial contacts to recommend additional participants with deeper involvement in the operational invoice review and related control processes. This allowed us to get a more granular understanding of the workflows and methods currently used and what parts could be automated.

Interviews were conducted as individual sessions and documented through audio recordings and subsequent transcription, with consent, which served as the primary basis for analysis to preserve detail and reduce reliance on memory-based reconstruction. When recording was not allowed, detailed notes were taken during the interview and written immediately afterward to minimize loss of information. Both authors attended each session, with one leading the discussion and the other focusing on note-taking, monitoring coverage of the interview guide, and identifying areas requiring clarification or follow-up probing (Bryman & Bell, 2011).

Practical arrangements were made to support uninterrupted discussion and usable documentation. Interviews were conducted in quiet and private settings, and each interview followed a structured close that allowed respondents to add topics not covered by the guide and provide final reflections (Bryman and Bell, 2011). The number of interviews was determined pragmatically based on coverage of relevant roles and diminishing returns in new information, and data collection continued until additional interviews no longer produced materially new insights relevant to the research questions, consistent with saturation as a stopping principle (Bryman & Bell, 2011). In this study, data saturation was considered achieved by the 15th interview. To maintain organizational anonymity, the term "Consultancy" is used in the Table 3.1 below to refer to independent external advisory firms specialized in cost verification, project review, and auditing within the infrastructure sector.

Table 3.1: Description of Interviews

Interviewee	Function	Role	Organization	Date	Length
1	Project Management	Finance and Chief of Staff	STA	2026-02-04	42 min
2	Project Management	Contract Support	STA	2026-02-06	48 min
3	Project Management	Purchasing / Calculations	STA	2026-02-09	36 min
4	Project Management	Staff	STA	2026-02-09	53 min
5	Project Management	Qualified Project Engineer	STA	2026-02-11	46 min
6	Consulting	Engagement Manager	Consultancy	2026-02-13	68 min
7	Project Management	Assistant Project Manager	STA	2026-02-13	51 min
8	Finance	Finance Manager Large Projects	STA	2026-02-19	49 min
9	Business Development	Business Developer	STA	2026-02-24	49 min
10	Consulting	CEO	Consultancy	2026-02-24	46 min
11	Digital Strategy	Digital Strategist	STA	2026-02-26	45 min
12	IT/Systems	System Administrator CDI	STA	2026-03-02	61 min
13	IT/Systems	RPA / UIpath / IT	STA	2026-03-03	56 min
14	Project Management	Qualified Project Engineer	STA	2026-04-02	56 min
15	Law	Legal Counsel	STA	2026-04-15	63 min

3.4.2 Observations

Observations of STA's invoice review workflow were conducted directly on-site. Bryman and Bell (2011) argue that observation is particularly valuable in qualitative research when the aim is to capture how work is actually performed, rather than how participants retrospectively describe or recall it. Since invoice review involves tacit routines, real-time judgment, and embedded system interactions that are difficult to articulate fully in an interview setting, direct observation provided a level of operational detail that would otherwise have been inaccessible.

The observation took the form of a 2.5-hour detailed walkthrough of the review process at Project Haga. This allowed real-time observation of the different checks carried out by project engineers ("projektingenjörer), which is the name of the role of the people at STA conducting the invoice review. This included the decision points, systems, and documentation they rely on at each stage. This allowed for observing the workflow as a continuous process rather than a set of discrete activities described in isolation. This hands-on component was valuable in developing a grounded understanding of the practical constraints involved, including the volume of supporting documentation, the interface of the systems used, and the reasoning required at key review steps. Furthermore, it allowed for detailed questioning of the various checks performed by project engineers, which was essential for understanding which parts of the process are feasible to automate with AI.

Consistent with the discussion by Bryman and Bell (2011) of non-participant observation in organizational settings, the authors maintained an observer role during the observation session and did not intervene in or influence the reviewers' work. Notes were taken during and immediately after the sessions and were subsequently used to triangulate and contextualize findings from the interviews, particularly in relation to the volume and fragmentation themes identified in the analysis.

3.4.3 Supporting Materials

Supporting materials consisted of invoice PDFs and their associated attachments, reflecting the day-to-day material processed by STA invoice reviewers. An initial sample of invoices was used to form a preliminary understanding of how much of the invoice population follows standardized formats versus heterogeneous, supplier-specific structures. The examples were illustrative and are used to ground interview discussions, map feasible links between invoice lines, contract terms, and supporting documentation, and derive constraints and requirements for an AI-based invoice review tool. In addition, internal documents and guidelines related to invoice review at STA was also analyzed. Use of these materials followed confidentiality requirements under a signed non-disclosure agreement (NDA).

3.5 Data Analysis

3.5.1 Literature Synthesis for the Theoretical Framework

The theoretical framework was constructed by synthesizing the selected literature into a set of concepts and design principles that are applied consistently throughout the report. The synthesis began by identifying the core mechanisms described in the literature relevant to invoice review under cost-plus contracts. These were then examined in relation to the empirical context of this study to assess which concepts carried direct implications for the design of an AI-supported review system. Concepts that appeared across multiple independent sources and that mapped onto recurring constraints in the empirical setting were given greater weight in the framework. This process resulted in a set of analytically grounded design requirements covering risk-based prioritization, audit-trail integrity, and human-in-the-loop validation, which serve as evaluative criteria for assessing both the findings and the proposed solution.

3.5.2 Gioia-style analysis

The interview material was analyzed using a Gioia-style method, which is a structured way to go from raw interview statements to a small set of clear themes, while keeping an explicit trail showing how conclusions are grounded in the data (Gioia, Corley, & Hamilton, 2013). The analysis also followed an abductive approach, meaning the interviews were repeatedly compared with relevant research literature to refine the understanding and focus areas (Gioia et al., 2013).

First, interview notes, supporting documentation, and transcripts were read, and concrete statements about how invoice review is done, what information is used, and what goes wrong were highlighted. Similar to the interview transcripts, the notes from the observation session were coded and integrated into the same data structure, ensuring that observed workflows were analyzed systematically alongside the stated routines. The statements were later turned into short labels that used the respondents' own wording where possible. These labels are called first-order concepts in the Gioia method. Second, patterns across interviews were identified. Similar first-order concepts were grouped into broader themes, for example, themes about traceability gaps, recurring documentation issues, reviewer workload and prioritization, or contract-related incentives. These are called second-order themes. Third, related themes were combined into a few higher-level categories that summarized the main mechanisms shaping invoice review in this context. These are called aggregate dimensions. The full chain from first-order concepts to second-order themes was documented to aggregate dimensions in a Gioia-style data structure so the reader can see how the final themes were derived from specific interview statements (Gioia et al., 2013). Figure 3.3 illustrates this data structure and served as the outline for the results chapter.

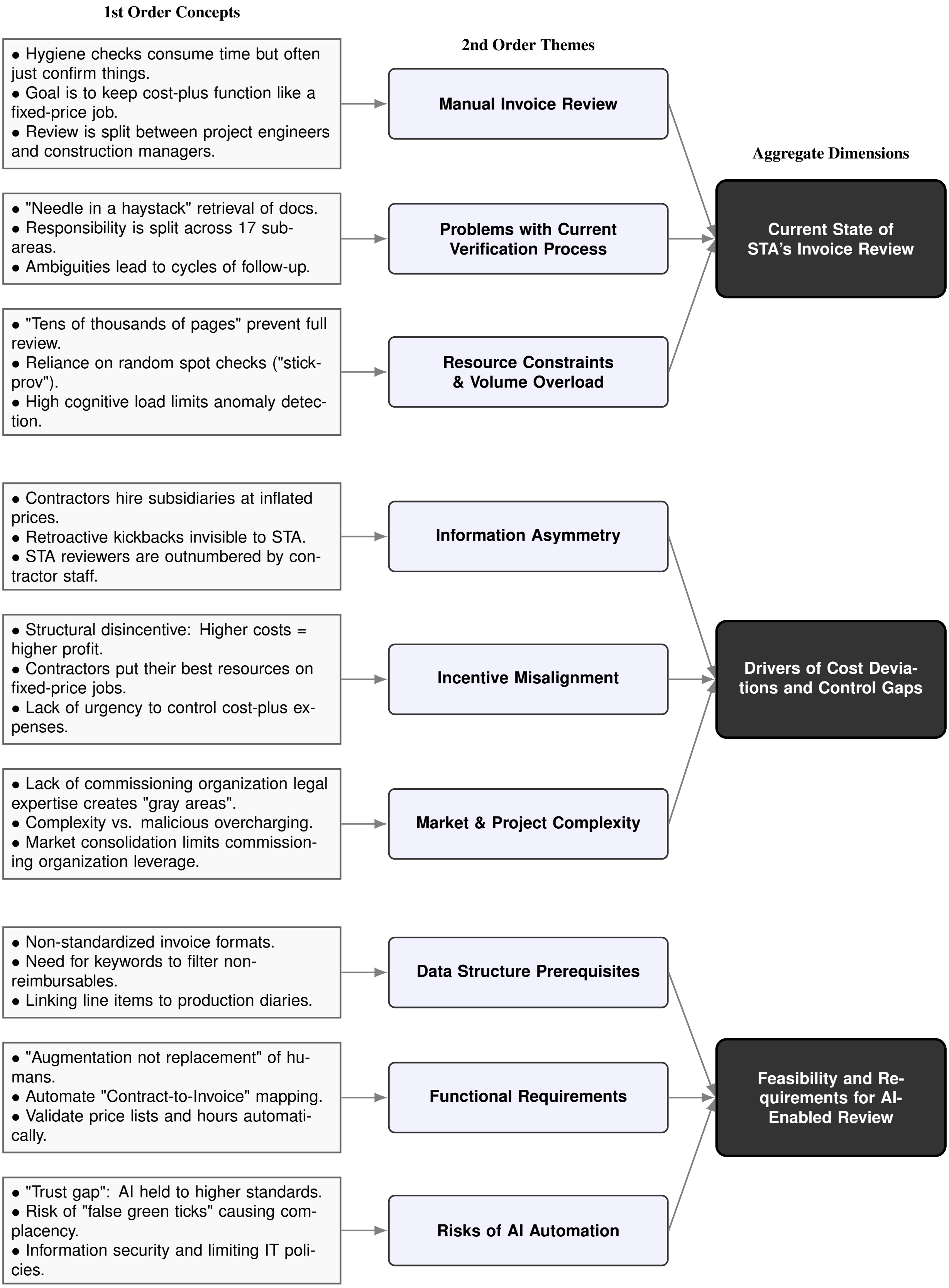

Figure 3.3: Data Structure

3.6 Research Quality

Bryman and Bell (2011) describe research quality in qualitative studies as being a function of reliability and validity. Reliability concerns whether the study's results would be consistent if the same approach were repeated, and validity concerns whether the conclusions are supported by the empirical material (Bryman & Bell, 2011). These aspects are often further segmented as internal and external. External reliability concerns whether another researcher could follow the same procedure and reach similar findings, while internal reliability concerns whether different observers would interpret the material consistently. Internal validity concerns the credibility of the interpretations, and external validity concerns whether findings can be transferred to other settings (LeCompte, Preissle, & Tesch, 1993).

Also, in qualitative research, quality is frequently evaluated through the concept of trustworthiness (Bryman & Bell, 2011). Trustworthiness consists of four criteria: credibility, transferability, dependability, and confirmability. In this study, credibility was established by triangulating data from multiple sources—combining interviews, direct observations, and internal document-based material to cross-check descriptions of workflows, controls, and recurring issues. A recognized risk was that many insights might solely reflect the specific organizational context of STA and its established ways of working. To address this and support transferability, the study provides a detailed contextual description of the West Link project and utilizes multiple roles to verify statements, allowing the reader to assess relevance and analytically generalize the findings to other similar infrastructure projects. Dependability was ensured by systematically applying the Gioia method to structure the analysis. Finally, confirmability was achieved by maintaining a transparent analytical chain within the data structure, explicitly linking the raw empirical data to the aggregate dimensions to demonstrate that the findings are strictly grounded in the respondents' actual accounts rather than researcher bias.

3.7 Use of Artificial Intelligence

Various LLMs, e.g., Claude, Gemini, and ChatGPT, were integrated into the research workflow as assistive tools. These AI systems functioned as collaborative sparring partners during the initial brainstorming phases. Throughout the drafting process, AI was utilized for continuous proofreading and comprehensive feedback. Furthermore, AI supported the literature review by assisting in the identification of relevant academic sources and keywords related to megaproject governance, information asymmetry, and contemporary automated data capture methods. By also assisting in the technical organization of raw qualitative data into structured formats like LaTeX tables, these tools allowed for a greater focus on the analysis of the findings. It is important to emphasize that all final judgments, interpretations, and the synthesis of empirical data remain the sole responsibility of the authors, with

AI functioning strictly as a mechanism for cognitive augmentation rather than content generation.

3.8 Ethical Considerations

This study followed established ethical guidelines by prioritizing participant protection and data security throughout the research process. STA provided access to internal materials under an NDA, and all participants provided voluntary, informed consent before any data collection. To ensure privacy, the study adhered to the STA's internal policies regarding cloud-based AI services and addressed legal concerns such as GDPR compliance regarding personal data. Sustainability was integrated into these ethical standards by focusing on the efficient use of public funds, improving the work environment through the automation of monotonous tasks, and preserving institutional memory by codifying tacit knowledge into the AI system.

4

Results

This chapter presents the empirical findings of the study by detailing the qualitative analysis of the semi-structured interviews conducted with project managers, contract specialists, and procurement officers at STA. These findings focus on the current workflow, the drivers of cost deviations, and the operational requirements for an AI-supported solution.

4.1 Current State of STA's Invoice Review

The first aggregate dimension characterizes the existing state of invoice review at STA. The analysis reveals a process designed to be rigorous but constrained by volume, fragmentation, and reliance on manual checks. This makes some cost deviations and invoice errors go unnoticed.

4.1.1 Manual Invoice Review

The manual invoice review process at STA is an important defense against cost overruns in cost-plus contracts. Interviewee 3 explained that this review essentially aims to keep cost-plus arrangements functioning somewhat like fixed-price jobs, ensuring that suppliers are only paid for what is contractually agreed upon. Interviewees 1 and 5 added that, beyond simple verification, the review also serves a broader cost-control function, helping the organization track, manage, and forecast project finances each month. Interviewee 5 described the invoice review work and the cost control as two distinct but connected parts of the invoice process. The respondent also indicated that the cost control function is where "things actually happen" and is the more central function. Interviewee 6 described the function of cost control as identifying where costs exceed the budget. Something she brought up that could be tracked on the contractor level is the ratio between the percentage of budget consumed and the percentage of the work that has been completed. If this ratio is high, e.g., exceeding 1, she explained that the contractor should be able to explain why, e.g., if there have been large initial purchases or if certain items have been more expensive than estimated in the tender.

The actual work of verifying invoice reimbursability is split into two types: desktop work and field verification. According to Interviewee 1 and 4, the Korsvägen project relies on a small core team of one to three dedicated invoice reviewers, called project engineers. These core reviewers handle hygiene checks, including verifying hourly rates and material amounts, confirming that approved personnel were used, and cross-referencing price lists. They act as coordinators and distribute invoices to a wider network of 10 to 15 people, primarily construction managers ("byggledare"), who perform fieldwork. These field managers are responsible for verifying that the claimed physical work was actually completed on-site and that the material amounts make sense. This invoice review organization is similar to the one at the Haga project, where 3 invoice reviewers coordinate with up to 8 construction managers.

This two-tier structure for the invoice review process is confirmed by an internal document from the Korsvägen project. The document clearly distinguishes between reviewers and project engineers and assigns each role separate responsibilities (Trafikverket, 2025a). The document further describes that the review is divided across 17 distinct sub-areas, ranging from civil works and concrete to installations and shared costs, each with one or more named responsible reviewers. In several sub-areas, up to three individuals share responsibility for reviewing a single cost category (Trafikverket, 2025a). This means that at any given month, an invoice from a single main contractor is, in practice, being assessed by a network of individuals across 17 parallel tracks, with the project engineer responsible for coordinating the findings into a single set of comments. Interviewee 14 also explained that supporting documents are distributed to construction managers based on their specific technical disciplines, such as concrete or groundwork. To optimize this process, project engineers absorb the administrative burden and contract verification. This deliberate split ensures that construction managers can dedicate their time exclusively to verifying the physical work in their areas of expertise.

Reviewers spend their time manually skimming through hundreds of PDF pages and production diaries. Interviewee 4 noted that reviewers often have to rely on their own judgment and experience, rather than following strict, quantifiable criteria. An example she brought up is whether a contractor's purchase of material is reasonable. However, Interviewee 5 described the general invoice process as a more standardized list of hygiene checks. In the Haga project, each of these is checked off sequentially in an Excel spreadsheet, according to Interviewee 5. Even though the Korsvägen project also follows the same checks, they do not check each one off in the same way, according to Interviewee 14. Instead, it serves more as a guiding principle for their invoice review rather than a list of mandatory checks. As described above, these checks are split into those conducted by construction managers and project engineers. Those performed by the construction managers are shown below in Table 4.1. To conduct these, Interviewee 5 explained that you need to be physically present at the construction site. He also highlighted that it is more of a qualitative judgment rather than the more binary checks performed by the project engineers.

Table 4.1: Review Checks Performed by Construction Managers

#	Description
1	**Clear Documentation?** Checks for clear supporting documents (protocols, diaries, measurements, test certificates) and verifies that costs relate to actual work recorded in the production diary. Invoices without this are stopped.
2	**Fair and Reasonable Cost?** Assesses the reasonableness and professional quality of the performed work, ensuring billed hours and material quantities align with the actual progress made during the month.
3	**Reimbursable by Contract?** Verifies if the cost is reimbursable under cost-plus principles, falls under the contractor's fee, or stems from defects/deviations requiring a liability investigation (referencing the list of cost agreements).
4	**Contract or ÄTA Work?** Determines whether the invoiced cost relates to standard contract work or ÄTA (Ändring, Tillägg, Avgående) work.
5	**Specific Contract Work?** Identifies the exact standard contract work package or activity the cost is associated with.
6	**Specific ÄTA Work?** Identifies the specific ÄTA work the cost is connected to and its current status.

Those review steps performed by the project engineers at STA are shown in Table 4.2, as described by Interviewee 5 and 14. Both made similar estimates of the time project engineers spend on each step as a % of the whole process.

Table 4.2: Review Checks Performed by Project Engineers

#	Review Step	Time Spent	Description
1	Documentation Sufficient?	20–30%	Assesses quality and presence of required supporting docs (diaries, receipts)
2	Contract Exists?	∼5%	Matches invoice marking with the financial system.
3	Reimbursement by Contract?	5–10%	Binary check against contract terms (e.g., included travel costs)
4	No. of Subcontractor Tiers?	10–15%	Tracks subcontractor layers to prevent fee multiplication
5	Correct Price?	∼30%	Verifies invoice against the price list.
6	Cost-plus Reimbursement?	5–10%	Checks contract specifics for cost-plus billing.
7	Reviewed by Contractor?	∼5%	Binary check of the approval log.

The first step requires qualitative judgment to assess if the correct supporting documents are present and of sufficient quality. It involves checking if daily logs are attached (either embedded in a PDF or as separate files) and ensuring receipts are provided for any expenses. Because documentation quality fluctuates, it can be difficult to define exactly which documents to look for. The second step involves locating the contract by matching the invoice markings with the company's financial system. While it takes only about 5% of the total review time on average, it can be time-consuming if the plan is outdated or agreements are not readily accessible. The third step is a binary check to verify what is included in the agreed pricing structure. Reviewers check the contract to determine whether elements such as travel costs or specific out-of-pocket expenses are permissible or already included in the hourly rate. This step takes 5–10% of the time and is relatively straightforward since most parameters are explicitly defined in the contract. The fourth step is tracking the chain of subcontractors, which is critical because overhead fees and margins essentially multiply with each added tier. While it takes up 10-15% on average, complex contractor chains can make it more time-consuming. The fifth step involves a direct comparison of the invoice line items against the agreed-upon price list. It is the most time-consuming single step, taking up 20-30% of the review time. In the sixth step, reviewers verify the specific contract details regarding cost-plus billing to ensure the cost is reimbursable according to cost-plus principles. This takes 5–10% of the review time and may require some qualitative judgment. The last step involves a binary check of the approval flow log to ensure the contractor has signed off. This is a minor step and takes roughly 5% of the time. Approval logs are often stored in different locations depending on the supplier, and vague approval comments (e.g., simply writing "OK") often require manual follow-up to confirm intent.

The internal review document also outlines the criteria that define an optimal verification document (Trafikverket, 2025a). These criteria include:

1. Contract works, and change orders must be reported separately
2. Quantities and unit prices must reference the attached supporting documents
3. Any supporting document without a delivery note must be highly specified
4. Hours worked must be traceable to a production diary
5. Agreements must exist for works exceeding SEK 300,000
6. The boundary between fixed-price, cost-plus, and other reimbursement categories must be clearly visible

These criteria define the standard that reviewers apply when working through their checklists, and their specificity illustrates why the process is so cognitively demanding. Each proof of payment must be assessed against multiple conditions simultaneously.

In Figure 4.1 below, the invoice review process is illustrated as described by the interviewees. It should be noted that the invoice review process differs between different projects within the West Link and within STA. Within the West Link, the process is standardized to some degree regarding what is reimbursable, but usage of IT systems and ways of working differ between sub-projects. Some interviews also mentioned that some sub-projects rely solely on sampling checks, while others review all underlying invoices. Also, the invoice formats received from suppliers differ across sub-projects. For example, while Interviewee 4 and 14 described how suppliers for the Korsvägen project upload supporting documents directly into a project portal, Interviewee 5 explained that at the Haga project, the review begins with a manual data dump via email. This email contains a consolidated package including PDF invoices, time diaries (TIM-dagbok), professional diaries (YA-dagbok), and an Excel summary linking back to the financial system, e.g., Medius.

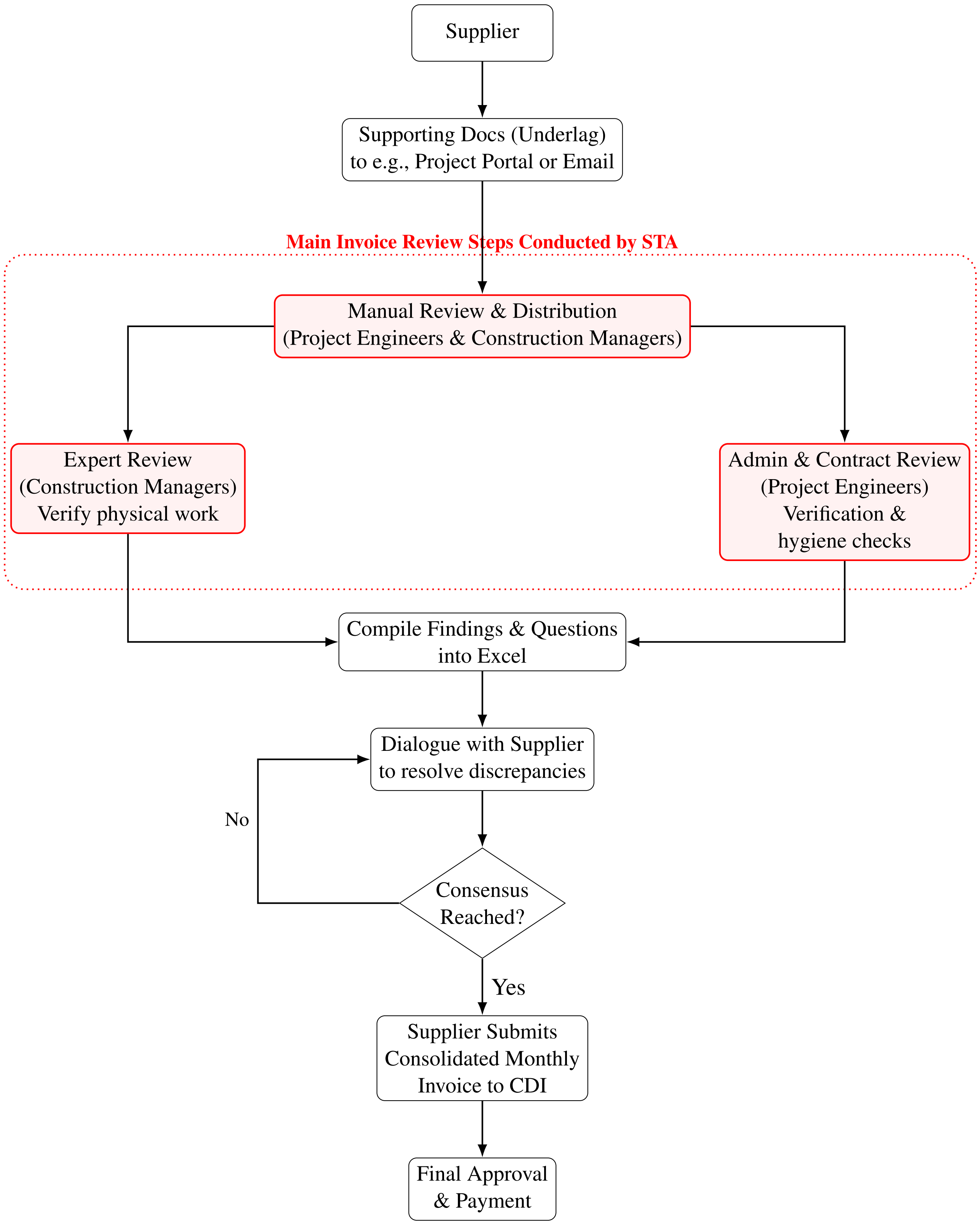

Figure 4.1: Illustration of the Invoice Review Process

The invoice submitted to the CDI system consists of a single monthly consolidation per contractor. Interviewee 14 clarified that this is intended for payment execution, not for review, as the actual verification of costs is performed on the underlying documentation before the consolidated invoice reaches CDI.

4.1.2 Problems with Current Verification Process

The verification process is characterized by significant fragmentation, with responsibility distributed across multiple roles, from invoice reviewers and project engineers to on-site construction managers. Interviewees 1, 2, and 8 all confirm that each sub-project adapts its own routines and depth of control. While this aims to leverage specific expertise, it creates information scattering and cognitive overhead.

Interviewee 4 described the workflow at the Korsvägen project as involving a small core of invoice reviewers and a wide network of verifiers: "Those of us working most hands-on are myself and three project engineers, but we also have ten to fifteen construction managers who help review the documentation". This fragmentation arguably makes it difficult to maintain a holistic view of costs. Interviewee 3 noted that invoices pass through contractors who verify quantities before they reach STA, resulting in large invoices that aggregate numerous sub-purchases.

The fragmentation also leads to a needle-in-a-haystack problem. Interviewee 4 emphasized that much of the reviewer's time is not spent on judgment but on locating the correct supporting document among thousands of pages: "The time-consuming part is the extensive documentation, supporting documents often include hundreds of invoices from their subcontractors". These hygiene checks, as Interviewee 5 described them, almost always result in a straightforward confirmation, i.e., the invoice looks correct, and payment proceeds. As he put it, price verification work "in the best case gives nothing, because it checks out". The concern is therefore not that hygiene checks frequently catch errors or that some fall through the cracks, but that performing them manually across thousands of verifications is time-consuming, which takes time away from more high-value tasks.

A further consequence of the verification process is what happens when questions are not resolved promptly. Interviewee 5 illustrated this by noting that if contractual ambiguity about which costs qualify as reimbursable is not clarified promptly, unresolved amounts accumulate over months and become difficult to address retroactively. This is seen in practice as either withheld payments or as a potential reclaim from the contractor. He therefore argued that the only good outcome is to resolve such questions immediately. However, as Interviewee 5 described, the capacity of invoice reviewers is largely consumed by routine hygiene checks, which he characterizes as pure volume work that ties up project engineers, leaving less room to address the more complex questions that require prompt resolution. Interviewee 3 confirms this dynamic from a different angle, noting that while STA can withhold payment when uncertain, doing so triggers a process where the supplier must return with additional supporting documentation, creating an ongoing cycle of follow-up rather than a clean resolution. Interviewee 4 noted that with the previous contractor on the Korsvägen project, over 50% of consolidated invoices received comments or required follow-up, and this recurred every single month throughout almost the entire project. Interviewee 6 framed the longer-term implications, arguing that if strict documentation standards are not enforced from the very start of a project, the commissioning organization loses the ability to build a self-sustaining review process and instead remains locked into a reactive, correction-oriented workflow throughout the project's lifetime.

4.1.3 Resource Constraints and Volume Overload

Another aspect of the current workflow that was frequently raised in the interviews is the mismatch between invoice volume and available human resources. This forces the organization to abandon comprehensive reviews in favor of sampling checks, creating a vulnerability in which errors can go undetected.

Interviewee 2 stated explicitly that "manually reviewing tens of thousands of pages monthly is practically impossible for STA". However, he stated that in some projects all underlying invoices are controlled, whereas in others sampling checks by construction managers are the only control. Interviewee 2 also noted that STA's invoice review organization is often understaffed, which is a reason why some cost deviations slip through the cracks, e.g., when suppliers hire their subsidiaries at non-competitive prices. Interviewee 5 put concrete numbers to this challenge, noting that in Haga's largest contract alone, between 2,000 and 4,000 verifications were processed monthly. He noted that completeness checks at this scale are handled by junior staff precisely because it was pure volume work, which raised implicit questions about consistency and quality across the full population of verifications.

Interviewee 5 noted that the largest bottleneck in the invoice review process is the frequency of edge cases and complex questions that require discussion with the supplier. These edge cases require legal competence, as highlighted by Interviewee 1: "We lack expertise in construction law specifically, what is and isn't okay to pay for... There are many gray areas". This interpretive ambiguity allows costs to drift. There is a tendency for these complex questions to be escalated upwards within STA due to the difficulty of finding people at suppliers with roles similar to those of the invoice reviewers. This takes up a lot of time for senior staff at STA, creating a bottleneck in the invoice review process. Interviewee 5 made this point explicitly, arguing that quality inevitably suffered at scale because project engineers got tired, and that this was a natural consequence of volume rather than a failure of the individuals involved. As he put it, a computer does not get tired after reviewing a thousand verifications, while a human reviewer inevitably does.

A related resource constraint is the burden of competence placed on new invoice reviewers. Interviewee 1 noted that the first task for any new reviewer is to read and understand the project's cost-plus contracts, which she described as complex and difficult to grasp. This onboarding burden further reduces the review team's effective capacity during staff turnover. Interviewee 1 also highlighted the time-consuming task of matching invoice amounts to the correct supporting documents across 1,500 to 2,000 verifications, including cases where the referenced contract cannot be found in the submitted material at all, requiring the reviewer to go back to the supplier.

Interviewee 7 identified what she described as a structural paradox in the resource constraint. To improve the quality of the main contractor's internal invoice review, the commissioning organization must maintain a diligent, consistent review process, since contractors only tighten their routines when they know deviations will be caught. However, the commissioning organization lacks the resources to sustain that level of scrutiny, so the two sides of the paradox reinforce each other. Interviewee 6 added a temporal dimension to this constraint, noting that when production is running at hundreds of millions per month, there is simply no capacity to pause and fix process problems. This makes the early stages of a project the only realistic window for establishing proper routines, and mid-project improvement is essentially impossible in practice.

Despite the resource-intensive operational effort conducted by the invoice review organizations, Interviewee 1 confirmed that they do not currently use performance metrics or key performance indicators (KPIs) to measure their success or efficiency.

4.2 Drivers of Cost Deviations and Control Gaps

The second dimension identifies the structural and behavioral mechanisms that drive cost deviations. The interviews suggest that these deviations are not just errors but are often driven by information asymmetry, misaligned incentives inherent in cost-plus contracts, and the high complexity of the project environment.

4.2.1 Information Asymmetry

A central driver of cost-control gaps is STA's lack of insight, due to complex supplier structures. Respondents expressed significant concern regarding the use of sister companies and subsidiaries, where visibility into actual costs is obscured.

Interviewee 2 highlighted this as a major risk: "Contracting firms often have internal units... The risk is that they hire these internal companies at inflated prices". He explained that this structure makes it difficult for STA to verify if the price is truly competitive, especially when "competitors rarely offer their best prices to a rival, favoring the internal company". The reason for this dynamic was not specified. Interviewee 7 confirmed this dilemma, noting that while they have tried to limit subcontracting tiers, they have not dared to ban intra-group purchases due to a fear of losing bidders.

Also, Interviewee 3 highlighted that the issue of suppliers hiring subsidiaries at inflated prices has mostly occurred at the Korsvägen project, rather than being widespread throughout the West Link.

A further form of information asymmetry was described by Interviewee 10, who highlighted the common practice of retroactive kickbacks. In practice, this works by the contractor claiming the full market price to the commissioning organization, while subsequently receiving a retroactive discount or rebate from the supplier that the commissioning organization never sees or benefits from. Interviewee 10 noted that these arrangements go by many names in the industry, such as annual volume discounts or marketing contributions, but that the underlying mechanism remains the same regardless of the name. Because the kickback occurs after the invoice has already been submitted and approved, it is invisible to the commissioning organization under normal review conditions.

Interviewee 2 added that information asymmetry does not only arise from supplier structures but also from how contractors behave after contracts are signed. He described a pattern in which, despite agreed prices and quantity schedules in the contract, contractors frequently argue that conditions have changed, using this as justification to shift costs from contracted fixed prices to cost-plus. Because the commissioning organization cannot easily verify whether conditions have actually changed, this creates an information gap that is difficult to close without deep and continuous monitoring. The resource imbalance between the commissioning organization and the contractor further enables this asymmetry in practice. Interviewee 2 noted that the commissioning organization is often heavily outnumbered, with perhaps one person on the commissioning organization side for every seven or eight on the contractor side. This means that even when the asymmetry is understood in principle, the commissioning organization simply lacks the capacity to investigate it systematically, and much of it goes undetected as a result. Interviewee 14 further emphasized that contractors bear primary responsibility for the accuracy of their billing. Likening the process to a consumer phone bill in which the provider is expected to track usage accurately, Interviewee 14 noted that STA should ideally find no deviations. All submitted invoices should already have been fully vetted by the contractor's internal review. Therefore, STA can withhold payments if the documentation is inadequate. However, it is STA's responsibility to prove that a certain cost or invoice is not to be paid, according to Interviewee 2. Interviewee 4 estimates that 20% of the underlying supporting documents are initially rejected or require follow-up comments from STA.

Interviewee 8 pointed to a further structural source of information asymmetry, noting that contracts at STA do not specify the format of supporting documentation, such as PDF, Excel, or Word. This means that from the very outset of a project, the commissioning organization has no contractual basis for demanding standardized evidence, leaving the information provided by the contractor largely on the contractor's own terms.

4.2.2 Incentive Misalignment

The cost-plus contract model itself was identified as a source of incentive misalignment. Because the contractor's profit grows in proportion to the costs incurred, there is a structural disincentive for cost control.

Interviewee 3 was clear about this dynamic: "Yes, it is an incentive. The more they increase the cost, the higher their absolute profit is". He further explained that while there are sometimes mechanisms to reduce the fee percentage if time/quality targets are missed, the underlying logic remains that higher costs yield higher absolute profits. Both Interviewee 3 and Interviewee 5 (who works on the Haga project) confirmed that STA's standard add-on fee for cost-plus contracts is 11%. However, to mitigate delays, this fee is reduced by 0.5 percentage points for every calendar month the contractor falls behind schedule, down to a minimum floor of 5%. However, Interviewee 5 noted that these penalties are primarily tied to schedule overruns, and only certain contracts allow the fee to be reduced due to quality failures. Interviewee 2 also mentioned that contractors often prioritize their best resources for fixed-price projects, where efficiency directly affects profitability, leaving cost-plus projects with less efficient and expensive resources. However, Interviewee 2 noted that cost-plus pricing is necessary for these types of projects due to the high risk associated with their complexity.

Interviewee 2 also described a behavioral dimension of this misalignment, noting that contractors tend to become comfortable in their supplier management over time, no longer strictly following agreed contracts but allowing pricing to drift gradually into running costs. This pattern is difficult to detect without continuous and systematic monitoring, and is enabled by the resource gap described in the previous section.

STA has attempted to address the incentive misalignment, in part, through contractual mechanisms. Interviewee 3 noted that wherever possible, STA tries to use fixed pricing to limit exposure, though he acknowledged this is not feasible for highly complex projects. Interviewee 5 described how STA relies on the so-called "omsorgsplikten", which states that work should be carried out as if it were a fixed-price contract, in the best technical and economic interest of the commissioning organization. However, Interviewee 7 noted that, in practice, STA has tended to rely too heavily on this clause as a governance mechanism, and that it is frequently contested rather than serving as a reliable control, since its interpretation is vague and open to dispute. Interviewee 14 added that there are no immediate consequences when contractors fail to adequately review their own invoices. He noted that unless administrative failures are repeated enough to constitute a formal breach of contract, STA cannot easily enforce legal penalties. Instead, their primary recourse is often reduced to applying pressure during meetings by explicitly pointing out the contractor's errors, hoping they will correct their behavior.

A further dimension of misalignment was highlighted by Interviewee 4, who argued that the real lever for cost savings lies upstream in the procurement process. Once a price list has been agreed and a purchase has been approved, there is little room left to question the cost at the invoice stage. Interviewee 4, therefore, suggested that pressing contractors harder to find the cheapest available solutions before purchases are made would generate more savings than increasing the depth of invoice review.

4.2.3 Market and Project Complexity

Beyond incentives and asymmetry, the complexity of the project environment also drives deviations. This includes both technical complexity and the ambiguity of contractual interpretations.

Interviewee 3 noted that cost overruns are often not malicious but a result of genuine complexity: "It is usually a matter of genuine complexity that was not originally accounted for, rather than anyone taking advantage". However, the exact split between genuine complexity, administrative leakage, and deliberate overcharging remains ambiguous. Across all interviews, only Interviewee 7 attempted to quantify how much of the overrun was actually preventable. She estimated that a perfectly carried out invoice review could potentially save 20-30% of the excessive costs. The complexity is also exacerbated by a consolidated market, as mentioned by Interviewee 3: "The supplier market is dominated by a few nation-wide firms". This limits STA's leverage. Interviewee 7 further noted that despite STA representing over 40% of the Swedish construction market, contractors still hold significant leverage in tenders for complex projects like the West Link. This is because so few actors are actually capable of executing projects of this scale and technical difficulty, meaning STA must design contract arrangements that are attractive enough to secure competent bidders, even if those arrangements are less favorable to the commissioning organization.

Interviewee 10 framed the problem partly as a matter of industry culture. He described how the construction industry is accustomed to handling very large sums of money and that an attitude of accepting whatever the outcome is is deeply embedded. He noted that a large Swedish contractor told him they only conduct systematic invoice follow-up for projects exceeding SEK 500 million, which he found remarkable. In his view, this cultural norm explains why so little systematic cost control is carried out, even in projects of the West Link's scale.

4.3 Feasibility and Requirements for AI-Enabled Review

The third dimension outlines the practical requirements for implementing AI support. Respondents shifted from discussing problems to proposing solutions, emphasizing that AI should not replace human judgment, but rather augment it by handling data structuring and hygiene checks.

4.3.1 Data Structure Prerequisites

A significant barrier to automation is the lack of standardized data. The interviewees frequently pointed to messy, scattered, and heterogeneous data as a major operational hurdle.

Interviewees 3 and 4 noted that invoice formats differ by supplier, creating a technical challenge for automated reading. Interviewee 5 emphasized that "data is scattered and formats are not standardized", which complicates the linkage between invoices and supporting documents such as "production diaries... and transport notes". Interviewee 4 suggested that a key prerequisite is the ability to screen for specific keywords or items that are contractually non-reimbursable, such as "hand tools", which requires the system to structure unstructured line-item descriptions.

According to Interviewee 6, an automated system must be able to process a wide variety of artifacts, including tender specifications, signed contracts, verification records, diary sheets, and approval logs. Interviewee 4 adds that while invoices are mainly in Swedish and English, exceptions like German do occur, adding another layer of complexity. Interviewee 10 described the incoming data as massive volumes of unstructured, unstandardized, and messy documents.

Interviewee 13's current attempt to handle the unstructured PDF layer provides a clear sense of both the scale of the problem and the limits of existing tools. They are using a UiPath module called Document Understanding, which requires a person to manually mark up example invoices and train the model to identify specific fields. Interviewee 13, who demonstrated the tool in practice, said the central question they are still working through is whether the model can handle an invoice it has never seen before, with a layout it was not trained on, or whether it simply fails. This uncertainty means the current approach requires significant ongoing human effort and is not reliable across the different invoice formats that come in. Interviewee 13 also said he believes an LLM would do the extraction job better than the current approach, handling more formats more reliably and at higher volumes, but that the team has not been able to test this yet.

While standardizing invoices would solve many manual review problems, there is a financial trade-off. Interviewee 3 stated that demanding standardized formats would directly cost STA money. Interviewee 1 confirmed this, explaining that requiring higher quality and transparency on invoices and supporting documents means contractors will have to hire third-party administrative resources, a cost they will ultimately pass on to STA.

Because each supplier formats their invoices differently, traditional rules-based programming is unscalable. Interviewees 6 and 10 noted that this requires unique, rule-based computer scripts for each supplier. Interviewee 6 argued that while a Python script performs better for purely structured mass data, AI is potentially a more generalizable solution to the industry's lack of data standards. Interviewee 10 agreed, noting that while LLMs still struggle when a document is highly unclear, their ability to understand diverse formats is rapidly improving. Interviewee 8 also suggested that AI could bypass the need to forcefully standardize all incoming invoices.

A further problem flagged by Interviewee 13 concerns the quality of the data that suppliers actually submit. Contractors frequently send in billing descriptions that are completely unstructured. In some cases, a lump sum of hundreds of thousands of kronor arrives with nothing more than a note reading, e.g., "according to agreement with Bengt". On top of this, there are almost no unique identifiers for individual consultants in the data. Interviewee 13 noted that if eight people named, e.g., Johan Johansson, are active across different projects, the system makes it appear as if a single person has billed an impossible number of hours, rendering any automated check on consultant-level patterns unreliable. Resolving this requires STA to place much greater demands on contractors regarding how they structure and submit their data.

Despite AI's flexibility, certain structural demands must still be met for successful automation. Interviewee 10 argued that STA must place harder demands on main contractors regarding invoice standardization. He outlined several key prerequisites:

1. Exact role names must be consistent between the contract and the invoice. For example, if a contract specifies an "electrician", the invoice must use that exact term instead of abbreviations or alternate spellings.

2. All documents must be digital originals, not scanned copies

3. Main contractors must present costs broken down by specific suppliers or subcontractors

4. Main contractors must implement their own invoice controls before forwarding them to STA

4.3.2 Functional Requirements

All the interviewees agreed that an AI system should act as a support tool for human reviewers, rather than replacing them entirely. Interviewee 7 stressed that the tool should flag anomalies and guide human attention, rather than automating the entire process. Interviewee 8 also shared this view, stating that while AI is great at spotting patterns, it should not make the final call: "It will never be able to make a judgment... Who is responsible for judging whether the invoice is correct? AI just gives you the pattern". Therefore, the system must always keep a human in the loop to approve the final payments, according to the interviewee. Interviewee 3 was also grounded in the technology's initial limitations, noting: "There will definitely be errors when it reads and checks. So you have to do the review yourself initially". This was also raised by Interviewee 8: "The work of an AI system will still need to be manually reviewed, e.g., through human-sampled checks".

The core functional requirement is the automation of repetitive, rule-based verification. When comparing the construction managers' review with the project engineers', Interviewee 5 found the project engineers' review more suited to automation. He noted that "it is possible to automate a great deal of this" and that it would save significant time. Breaking down the project engineer's workflow reveals a clear divide in the feasibility of automation. According to Interviewee 5, step number 5, which includes price verification, is the most obvious candidate for automation. This step simply requires extracting data from PDFs and matching it against established Excel price lists. Interviewee 5 described this as the "lowest hanging fruit", estimating that AI could reduce 100 hours of manual work down to 10 minutes, potentially saving roughly SEK 1 million annually. While Interviewee 14 also thought it was a step that could be automated, he did not think it was the easiest to automate since there are edge cases, such as hidden discounts, that make the check difficult and require more judgment. Interviewee 5 also highlighted that verifying the reimbursement form (such as checking if travel costs are included in the hourly rate) and confirming cost-plus parameters are binary checks that are highly feasible to automate. Interviewee 14 argued that step 2 is the easiest to automate, since it is simply a matter of matching invoice numbers across two documents. Also, Interviewee 5 believed this step is easy to automate, provided the purchasing plan is updated in the system. He argued that if there's ambiguity regarding the contract number, which there sometimes is, an automation system might be hard to implement for this step.

Conversely, Interviewee 5 identified several checks that are highly resistant to automation. Verifying the quality of supporting documentation is difficult because the AI must evaluate not only the presence of a document, but also whether it is the correct type (e.g., a specific receipt for an expense), and the quality of these submissions fluctuates significantly. Tracking the number of subcontractor tiers is also complex but crucial to automate, as the contractor fee essentially doubles with each added tier, according to both Interviewee 5 and 14. However, Interviewee 7 mentioned that a digital system called Infobric was recently implemented which according to her makes this step easy to automate. There is however still high uncertainty regarding how this would be automated using the Infobric system, as opposed to the other review steps labeled as easy. Finally, verifying contractor attestations is hard to automate because approval logs are stored in varying locations depending on the supplier, and vague comments like "OK" often require qualitative human follow-up to confirm intent. easy In Table 4.3 below, an overview of Interviewee 5's and 14's estimates of automation feasibility for each of the project engineer's review steps is shown. The interviewees categorized each step into easy or hard, which is to be seen as an indicative relative scale of automation feasibility rather than a definitive, objective metric.

Table 4.3: Automation Feasibility by Review Step

#	Review Step	Time Spent	Difficulty	Main Challenge
1	Documentation Sufficient?	20–30%	**Hard**	Requires qualitative judgment of document types.
2	Contract Exists?	∼5%	**Easy**	Dependent on purchasing plans being updated in the financial system.
3	Reimbursement by Contract?	5–10%	**Easy**	Binary check, but requires accurate extraction.
4	No. of Subcontractor Tiers?	10–15%	**Easy***	Tracing complex, multi-tier supplier relationships that are often not explicitly recorded. Easy according to Interviewee 7 due to new implemented system
5	Correct Price?	∼30%	**Easy**	Extracting unstructured data from PDFs to match against existing price lists.
6	Cost-plus Reimbursement?	5–10%	**Easy**	Interpreting specific contract clauses and billing parameters.
7	Reviewed by Contractor?	∼5%	**Hard**	Approval logs are stored in varying locations, and vague comments require human follow-up.

The internal review documentation highlights areas that an end-to-end automated system would need to handle (Trafikverket, 2025a). For consultant and staff invoices, the project engineer must verify whether each resource has been formally agreed with STA and entered into the specialist hourly rate register, whether the billed hourly rate matches the agreed rate for that individual, whether rates in any additional subcontracting tier also comply with their respective agreements, and whether the rate increase, if any, is supported by documentation and consistent with the contract. These checks are currently performed manually against a maintained Excel file and a separate resource and competence plan. This means that the reference data for rate verification already exists in structured form Trafikverket (2025a). This implies that the gap is not the absence of a reference but the manual effort required to cross-check every line against it.

Interviewee 3 wanted the AI to be able to directly link incoming invoices to their underlying contracts: "If there is a contract number, it could check against the quote or agreement to ensure prices match". He emphasized that exact price matching is crucial because manually verifying hours and granular prices is difficult. Interviewee 4 added that it would be highly valuable to use AI to screen for contractually non-reimbursable costs. She gave a practical example, that the AI could search for terms like "hammers and wrenches", as basic hand tools are usually expenses the supplier should cover themselves.

Interviewee 13 reinforced this point from a more technical angle, arguing that AI has strong potential specifically for qualitative judgments of this kind. He gave concrete examples, such as flagging suspicious cost categories like tool purchases, and checking whether hourly rates for specific role codes match what has been agreed in complex contracts. In his view, these are exactly the kinds of tasks where an LLM adds value beyond what simple rule-based scripts can do.

Interviewee 2 also highlighted the system's role in finding hidden markups. He pointed out that previous script-based systems were used to detect "back bonuses", hidden kickbacks, or dual discount systems where the commissioning organization misses out on savings. He envisioned an AI that continuously verifies "that prices match agreed levels, quantities are reasonable, and valid contracts exist".

To be effective, the AI must handle messy, unstandardized documents at scale. Interviewee 4 noted it would be "invaluable to process it through a system rather than manually scrolling through every detail of a five-page invoice". Interviewee 2 added that an AI solution should trade a small degree of accuracy for massive scale, allowing it to handle "tens of thousands of pages monthly".

Interviewee 8 suggested a practical workflow requirement, that the AI should be the "first point of contact" within STA's invoice system, using traffic-light logic for unstructured data. As Interviewee 8 envisioned it, the system would output a "Red flag: wait, something doesn't quite match, double check it", or a "Green flag: everything seems to be in order, do a sample check". Furthermore, Interviewee 8 emphasized that the system must continuously learn. When a human corrects the AI, it needs to incorporate that information for the next batch of invoices.

By automating tedious desktop work, the AI would free up human experts to perform tasks that save money. Interviewee 6 explained that while an AI can match agreed prices, human experts (like construction managers) are needed to investigate if a purchase was actually necessary or to analyze field production methods, because "that is often where the really big savings are". Interviewee 7 agreed, noting that while physical verification must remain manual, using AI to check lists acts as a vital "early warning system". Rather than just finding small mathematical errors, the real value of AI is highlighting systematic cost escalations early on.

Interviewee 11 broadened the picture of what success should look like for any digital investment of this kind. She argued that the business case for an AI tool need not rest solely on financial savings. Demonstrating higher quality, building trust in STA as an organization, eliminating monotonous work tasks, and improving the work environment are all valid outcomes. She also highlighted fulfilling legal requirements as a public authority as an important consideration, noting that this dimension is sometimes overlooked when building the case for new digital tools.

STA is already taking steps toward these functional requirements. Interviewee 9 explained that the IT department is running a pilot project using the RPA tool UiPath for the Ostlänken project. The goal is to machine-read invoices, pick out dates, amounts, and names, and structure the data into readable reports. Interviewee 8 confirmed this initiative but reiterates that while the tool processes the data, it still requires manual intervention to judge validity. Both Interviewee 8 and 9 believed this exact functionality can and should be scaled up to all large projects—including the West Link, since the fundamental process of invoice review remains the same regardless of the project's geography. Interviewee 13 added an important detail about how this pilot is structured and where it is headed. He was clear that the primary goal of the current RPA project is to extract the raw data from PDF files and load it into a Structured Query Language (SQL) database in a structured format. No advanced analysis happens at this stage. The longer-term vision, however, is to move that structured data into a broader data platform where both business intelligence tools and AI models can be used on top of it. In Interviewee 13's framing, the current extraction work is the necessary foundation, and LLM-based analysis is the second layer that becomes possible once that foundation is in place.

4.3.3 Risks of AI-Automation

While AI presents significant opportunities for automated invoice review, the interviewees identified several risks. The primary concerns were not about the technology's capabilities, but rather information security, operational costs, and human psychology.

The most immediate barrier to AI adoption is data privacy. Interviewee 9 noted that current AI-usage at STA is highly restricted when it comes to generic, external LLMs such as ChatGPT or Copilot. The moment internal, sensitive, or personal data is involved, strict confidentiality rules block its use. Interviewee 4 confirmed this reality, stating that no invoice reviewers are currently using AI tools because they simply do not know if it is allowed under current guidelines. Interviewee 8 explicitly highlighted information security as a primary risk, stressing that transparency and the compliant handling of sensitive data are absolutely crucial before any AI system can be trusted with STA's information. Also, Interviewee 11 raised the concern that the policy at STA regarding what can be uploaded to LLMs is quite strict, and that all files uploaded via another digital tool need to be security cleared first. A document internal to STA providing employees with guidance on the use of external AI services confirms this (Trafikverket, 2025b). The guide explicitly states that internal information or confidential data should not be entered into cloud-based AI services such as ChatGPT, Copilot, or similar services. All data handling must therefore happen internally to protect information from leaking and to maintain the country's civil preparedness in the event of a crisis or conflict. To handle this, STA is building a fully local on-premise version of UiPath and developing its own internal AI platform, called STA Labs, where models can be trained or fine-tuned on their own servers. Interviewee 13 acknowledged that this is significantly more resource-intensive and slower-moving than simply buying compute capacity through cloud services.

Interviewee 10 highlighted this issue from a vendor's perspective. Even though STA's invoices are technically public records, contractors often claim business secrecy over specific pricing details. This makes it risky to upload unmasked invoices to standard, cloud-based LLMs. To comply with these strict security requirements, Interviewee 10 explained that his consultancy firm sometimes has to run AI models locally rather than in the cloud, or rely on strict non-disclosure agreements that heavily anonymize data and delete it immediately upon request. Interviewees 12 and 13 confirmed this point, stating that while invoices received by STA are indeed public records and must be archived for 7 years, contractors can formally request business secrecy over pricing details during the early years of a contract. In practice, this means that even a document that is technically public may need to have all its prices masked before it can be processed through an AI system, which significantly complicates any attempt at automated price verification.

Interviewee 15, a public law attorney at STA, added some more precise legal context to this question. The interviewee explained that invoices and supporting documents in cost-reimbursement contracts are almost always public records the moment they reach STA and that they often contain confidential information. Particularly itemized prices protected under OSL (Offentlighets- och sekretesslag (2009:400)) 31:16. To upload such material to a cloud-based AI service, STA would need legal grounds to break confidentiality, such as OSL 10:2 (necessary disclosure) or 10:2a (technical processing/storage), but both must be applied restrictively and require careful appropriateness assessments. The cleanest path forward, according to the interviewee, would be to include an explicit contract clause informing contractors upfront that documents may be processed by AI, effectively establishing formal consent under OSL 10:1 and avoiding legal disputes after the fact. Although, the interviewee did add that contractors would likely not be on board with this immediately without a proper understanding of its actual implications. To this, Interviewee 7 proposed a solution that STA should contractually mandate that the contractor is responsible for telling STA which documentation and invoices are not pulbic and not allowed to be processed in external AI tools.

A further legal risk that has not yet been resolved internally concerns the General Data Protection Regulation (GDPR). Interviewees 12 and 13 noted that there are active discussions at STA about whether it is permissible to aggregate data on individual consultants' working hours across projects in order to detect irregularities. The concern is that doing so may constitute processing of personal data in a way that conflicts with GDPR requirements. This question has not been settled and represents a real constraint on one of the more valuable use cases for an AI system, namely identifying consultants who appear to be billing implausible numbers of hours across multiple projects simultaneously.

Another practical risk is the AI's direct cost. Interviewee 10 pointed out that while the technology is absolutely mature enough to read and capture invoice data, doing so with LLMs is expensive because you pay per token (the amount of text processed). Given the massive volume of files STA handles, running every invoice through an LLM would cause costs to increase rapidly, depending on how tokens are managed. As a workaround, Interviewee 10 suggested using AI solely to write traditional, rule-based code (e.g., Python), which can then process invoices for free.

Once an AI system is implemented, operational risk arises from how human reviewers interact with it. Interviewee 8 strongly warned against overreliance on the system. He described a scenario in which a stressed, overworked reviewer sees a "green flag" from the AI and just clicks "approve" to save time, only to realize two months later that the AI made a mistake. He emphasized that AI should only flag patterns, not replace the human judgment required to actually validate a claim. Interviewee 4 brought up a similar limitation, that AI might be able to match numbers, but it struggles to judge whether the billed work was actually reasonable in a real-world context.

Interviewee 5 pointed out a psychological double standard regarding errors that could hinder AI adoption. While people accept that human reviewers might only be 95% to 98% accurate after six hours of tedious work, an AI is expected to be completely flawless. He noted that the tolerance for a machine giving a "false green check" is significantly lower than for a human making a mistake. However, Interviewee 4 argued against this demand for absolute perfection, arguing that even an imperfect AI system is better than the current status quo. She pointed out that simply reducing the time reviewers spend manually scrolling through every detail of five-page PDFs "would be worth its weight in gold", proving that a system does not need to be flawless to drastically improve the workflow.

Interviewee 13's provided a concrete, practical approach to managing this risk. His system applies a confidence threshold of around 80%, meaning that if the model is sufficiently confident in its extraction, it proceeds automatically. If confidence falls below that level, the document is routed to a human reviewer through an interface called "Action Center". There, the reviewer manually points out where the correct information is, and the model learns from that correction for future invoices. Interviewee 13 noted that this continuous human-in-the-loop retraining is necessary precisely because suppliers frequently change their invoicing systems and layouts, requiring the model to keep adapting. He also observed that this confidence threshold approach may not be straightforwardly replicable with a brute-force LLM, which is an important design consideration for any system built on top of language models.

Some interviewees argued that the bottleneck for AI automation is rarely the software itself, but instead the bureaucracy. Interviewee 9 stated that the hardest part of rolling out STA's new RPA tool is not the tech, but rather building the surrounding data structures, changing daily workflows, and "getting everyone on board". Interviewee 10 also argued for this, that the barrier to change isn't cost or technology, it is "old habits". He pointed out a direct conflict of interest, noting that there are consultants whose entire job revolves around manually approving invoices, and they are understandably reluctant to see their work automated away. However, this sentiment does not appear to be shared by STA's internal staff. Interviewee 14 noted that project engineers would strongly welcome an automated tool, stating that such an implementation would "fly" and that staff would be highly appreciative of being able to offload the repetitive administrative burden. However, he cautioned that the actual barrier to adoption is not internal resistance, but the inherent complexity of the review process itself. Automating these workflows requires navigating a complex intersection of legal, economic, and production-related variables.

Furthermore, Interviewee 10 observed that STA is highly fragmented. Sometimes, obvious efficiency measures are missed simply due to poor internal communication. Because of this bureaucracy, he argued that attempting a massive, organization-wide AI rollout is too risky, and efforts should instead focus on developing specific, sub-project pilots. This is also a concern Interviewee 11 raised, specifically that there are many slow processes and approvals that need to be addressed for an organization-wide digital solution to be rolled out. This was described as an IT council that takes about a month, during which aspects such as the business case impact and the tool's information security are evaluated. The fact that the IT department at STA, called IKT, was also described as understaffed by Interviewee 11 further supports the view that this type of bureaucracy poses a risk. This is even clearer when she described her view on the overall IT work and AI-agenda at STA: "It is fragmented and comes from many different directions. The IT department also struggles with prioritizing what is important for STA as a whole". On top of this, Interviewee 11 also argued that graphical processing unit (GPU) availability might be one of the largest risks for implementing an AI system for invoice review, since all AI usage should run on STA's own GPUs according to her.

5

Discussion

5.1 The Failure of Manual Review and its Root Causes

The findings from the West Link project point to a governance problem that is structural in origin and difficult to resolve through incremental change and technological adoption alone. This discussion interprets those findings against the literature explored in Chapter 2. It examines why some of the conditions observed at STA could be predicted by theory, how the proposed framework aligns with and diverges from prior literature, and which conditions would need to change before the governance potential of AI can be realized in practice.

5.1.1 The Predictability of Governance Failures

The budget escalation at the West Link is striking in scale but not in kind. As established, cost overruns of this magnitude are statistically normal in major public infrastructure projects, driven by a combination of honest planning errors and deliberate misrepresentation at the approval stage (Flyvbjerg et al., 2002). What the findings at STA add to this picture is an account of how the overrun mechanism operates during project execution rather than at the planning stage. The problem is not only that the original estimate was wrong. It is the contract structure and the daily review process that contain features that allow costs to drift upward continuously. In addition to this, the technical complexity of the project is also a driver of cost overruns. This happens through the gradual shifting of items from fixed-price to cost-plus, the layering of subcontractor markups through intra-group transactions, and the accumulation of unresolved contractual ambiguities that are individually small but collectively significant.

Flyvbjerg (2008) distinguishes optimism bias, which is the tendency to underestimate project complexity, from strategic misrepresentation, a deliberate choice to understate costs in order to secure project approval. In practice, the two forces are difficult to separate in execution-phase cost management. Interviewees at STA expressed genuine uncertainty about whether the cost escalations they observed reflected deliberate manipulation, the unpredictability of underground rail construction in a dense urban environment, or administrative leakage and a lack of invoice review. In fact, across all interviews, only Interviewee 7 attempted to estimate this split when asked. The interviewee explained that a perfectly carried out invoice review could potentially save 20-30% of excessive costs, but she still signaled high uncertainty and that it was more of a guess than an estimate. That ambiguity is itself theoretically significant. It implies that a governance mechanism designed solely to detect fraud will be poorly suited to an environment where the boundary between opportunism and complexity is contested and unclear. One implication of this is arguably that the AI solution should be positioned as instruments for surfacing questions rather than for determining answers.

5.1.2 Power Asymmetry and the Normalization of Inadequate Review

In order to understand the root cause of the systemic failure of the manual review process, one must also consider the power asymmetry present in megaprojects such as the West Link.

Some respondents highlighted the genuine complexity of the West Link as a fundamental cause to the cost escalations. Applying Geraldi et al. (2011) dimensions of project complexity to the West Link case suggests that structural complexity and uncertainty complexity are the most prevalent dimensions. Structural complexity is visible in the project's many interdependent contracts, sub-projects, subcontracting tiers, and fragmented review responsibilities, which make cost traceability difficult and weaken STA's ability to maintain a holistic view of invoice flows. Uncertainty complexity is also central, since the cost-plus contract form is used precisely because scope, ground conditions, and future changes cannot be fully specified in advance. By contrast, pace complexity appears less dominant, as the West Link is a long-duration project rather than one primarily characterized by rapid execution pressure. Socio-political complexity is present through STA's public-sector accountability requirements and dependence on a concentrated supplier market, but in this study it mainly becomes relevant because it reinforces the structural and uncertainty-related challenges already embedded in the project.

According to respondents, this complexity limits the number of contractors that can potentially have end-to-end responsibility of carrying out sub-projects within the West Link. This thereby arguably shifts the power asymmetry in favor of contractors. Theoretical frameworks emphasize that the Swedish construction market, to some degree, functions as a regional oligopoly (Grahn, 2024). This market consolidation shifts bargaining power heavily toward dominant suppliers, giving them the leverage to dictate terms (Porter, 1980). However, in the context of STA, this dynamic was somewhat unexpected. Viewed through Porter's framework, one might expect STA to possess significant buyer power, given that it is essentially the sole customer for infrastructure megaprojects and represents over 40% of the Swedish construction market. Despite this expected market dominance, interviewed project stakeholders noted that STA's actual leverage is limited. The technical complexity and the large scale of projects like the West Link mean that only a handful of consolidated, nationwide firms are capable of submitting bids. These firms also appear to be highly diversified, with STA accounting for only a fraction of total revenues. One can therefore argue that despite the presence of a large number of international contractors, the extreme technical difficulty and scale of the project limit the number of truly capable bidders. This lack of practical competition probably neutralizes STA's theoretical buyer power, forcing the agency to design contract arrangements that are attractive enough to secure competent bidders, even if those terms limit their own control.

Furthermore, project owners often face a capability gap, struggling to act as "strong owners" in high-level supplier relationships (Winch & Leiringer, 2016). This theoretical gap is evident in the West Link's resource imbalance, where STA reviewers are significantly outnumbered by contractor staff, resulting in volume overload. The capability gap is also evident through the lack of knowledge in e.g., construction law, as described by one interviewee. Faced with tens of thousands of pages, the heavy documentation becomes a "monster" as Brunet (2021) describes it, which can paralyze project-level efficiency. Consequently, the power asymmetry between STA and the contractors can be seen as one reason why reviewers are forced into a sampling approach. This makes the invoice review at STA function as "cutting corners" to maintain efficiency, thereby ceding more power to the contractor by allowing them to exploit interpretative gaps (Pinto, 2014). Hence, it seems that the way the invoice review is "cutting corners" acts as a short-term solution, but ultimately works against STA long-term.

5.1.3 Information Asymmetry and the Limits of Documentary Evidence

The cost-plus contract structure used at the West Link is theoretically justified. Bajari and Tadelis (2001) demonstrate that fixed-price contracts become impractical when project scope cannot be reliably specified in advance, which is the normal condition in complex underground construction according to respondents. Transferring cost risk to the client is appropriate when the client is better placed to absorb it. The problem is that this rational choice also creates the information asymmetry that Eriksson and Lind (2015) describe as moral hazard. This moral hazard is evident in the West Link case because contractors earn profits proportional to total costs. As a result, the structural incentive to control those costs is weakened. The empirical findings provide detailed evidence of how this plays out. Contractors hire their own subsidiaries at internally set prices, creating transactions that are technically documented but not independently priced. Retroactive bonuses and back-payments between affiliated entities occur entirely outside the invoice data visible to STA. However, viewing these mechanisms strictly through the lens of moral hazard arguably creates an expectation of widespread, deliberate exploitation, since the contract structure makes it theoretically possible (Eriksson & Lind, 2015). Surprisingly, the empirical data contradict this assumption. While there is, without doubt, an incentive to inflate costs, respondents clarified that the severe exploitation of subsidiary pricing seen at the Korsvägen sub-project, which led to a cancellation of the contract, was a rare exception rather than a widespread norm across the West Link. This unexpected uniqueness of opportunistic behavior aligns with Suprapto et al. (2016), who argue that megaproject outcomes under cost-plus arrangements are heavily moderated by relational quality and collaborative governance, preventing standard moral hazard from running rampant in every sub-project. This suggests that the quality of the client-contractor relationship and the desire to maintain long-term market standing may be stronger deterrents to cost deviations than the formal contract and the incentive structures. STA's large size, its dominance of the megaproject market, and its 40% share of the Swedish construction market likely reinforce this dynamic by making it vital for contractors to maintain a long-term relationship with the client.

Schneider dos Santos et al. (2025) argue that detecting procurement fraud through data analysis is constrained not by algorithmic capability but by data availability and the transparency of complex supplier structures. The interviewees described retroactive bonuses and intra-group pricing arrangements between contractors and their own subsidiaries that are settled entirely outside the invoices submitted to STA, and therefore fall beyond the reach of any invoice-level analysis. While Schneider dos Santos et al. (2025) frame the issue as a general data availability problem, it can be found in a specific contractual structure in the West Link case. An implication of this is arguably that AI-assisted invoice review should not be positioned as a fraud detection system in the strong sense, but rather as a supporting tool for surfacing potential areas for further investigation.

A theoretical concept that is important to the West Link case is institutional memory. Eriksson et al. (2019) argue that rigorous review in long-term infrastructure projects serves to document how contractual edge cases are resolved, creating a precedent record that becomes essential when later disputes arise. When invoice review involves sampling, this memory function is arguably reduced. Ambiguous cases pass through without resolution and, over time, compound into disputes that are too large and too old to address effectively.

5.2 Structural Barriers to Implementation

The theoretical case for AI-assisted invoice review is supported by the literature. The empirical findings confirm the case but reveal a gap between what is technically desirable and what is currently deployable. That gap has three main components: data infrastructure, information security, and human factors. These are organizational and contractual problems that technology adoption alone cannot resolve.

5.2.1 The Data Infrastructure Deficit

To successfully implement an automated invoice review system, an organization must first have a reliable and structured data foundation. As established by Anwar and Akeel (2026), structured data governance and system integration constitute the mandatory foundational layer for any automated audit workflow architecture. However, the empirical findings demonstrate that STA's current operational reality fundamentally lacks this baseline.

The literature clearly identifies unstructured data as a primary technical bottleneck. Yashwant et al. (2025) and Saout et al. (2024) emphasize that high variations in layout and formatting across different vendor templates impede accurate data and table extraction. This theoretical challenge manifests directly at STA. Interviewees consistently highlighted that invoice data arrives scattered across multiple IT systems in non-standardized formats.

While a large portion of routine invoices are processed automatically via standardized formats, this is for the sake of payment processing. The complex construction invoices, which are those for cost-plus contracts that require rigorous review, trap critical line-item details within unstructured PDF attachments. Hence, it seems that STA has been able to create a structured data infrastructure for its AP function to handle routine invoices, but this way of working has not been applied to the invoice review needed for the cost-plus projects.

Furthermore, the deficit is not just a matter of file formats, but of data quality and completeness. As an example, contractors frequently submit unstructured billing descriptions, such as large lump-sum claims with informal references rather than detailed breakdowns. Another example is that the data often lacks unique identifiers for individual consultants, making it impossible to reliably track cumulative billed hours across multiple projects for a single individual. These empirical observations align closely with the theoretical assertions of Abela (2024) and Johri et al. (2026), who argue that gaps, incompleteness, and inconsistencies in records severely limit risk identification and the accuracy of anomaly detection systems.

Because the underlying data lacks strict standardization, traditional rules-based automation (such as RPA) becomes unscalable. As noted by interviewees, addressing the varied layouts through traditional programming would require writing and maintaining custom scripts for every individual supplier. While AI, specifically LLMs, offers a more flexible method for extracting unstructured data without rigid templates, it cannot invent missing information or verify undocumented costs.

Based on this data infrastructure deficit, there are several implications for STA. Most importantly, a fully automated, end-to-end invoice review seems practically impossible until upstream data inputs and delivery methods are standardized. The empirical findings highlight stark contrasts even between just two of the West Link's sub-projects. For instance, Haga receives documentation via email data dumps, while Korsvägen relies on the project portal. Since one interviewee highlighted that differences are even larger between STA's projects, attempting to engineer an autonomous solution across such a fragmented data landscape would probably require customized programming for almost every sub-project, making it highly resource-intensive and impractical. However, this infrastructural fragmentation is a barrier to end-to-end automation, not for an assistive tool where the project engineer uploads the data itself. Because the core steps of the invoice review process itself are relatively consistent across sub-projects, an upload-driven AI tool could effectively augment the reviewer without requiring a massive, unified IT overhaul.

However, data availability is a major bottleneck, regardless of whether an automation solution is fully integrated or upload-based. While AI can act as a bridge by probabilistically reading unstructured PDFs and bypassing the need for rigid vendor templates, it is not a cure for fundamentally unstructured data practices and a lack of data quality. If a contractor submits a lump-sum invoice without line-item details, the AI will fail just as a human reviewer would. Therefore, the successful implementation of AI-supported review requires strict, contractually enforced data governance from the start. As suggested by interviewees, STA should place harder demands on contractors, such as requiring digital originals, consistent role nomenclature, and explicit sub-tier breakdowns, before the operational benefits of AI can be fully realized.

5.2.2 Information Security Constraints on Deployment

The deployment of an AI-supported invoice review system faces a potential barrier regarding data privacy and information security. The theoretical literature emphasizes that deploying AI in environments involving sensitive project data requires governance frameworks to ensure the model behavior is permitted (Wondra, 2025). Furthermore, AI systems must be transparent and provide explainable outputs to maintain an auditable decision trail (C. A. Zhang et al., 2022).

At STA, these theoretical requirements clash with operational security guidelines. The empirical findings reveal that current information classification rules strictly prohibit the use of generic, cloud-based LLMs for processing internal or sensitive data. While invoices submitted to public authorities are public records, contractors sometimes invoke business secrecy over specific pricing details during active contract periods. On top of this, STA's own policy means that every invoice uploaded to an external cloud-based solution must be security-cleared first. Hence, processing this information through external commercial APIs in these situations is not an option. One alternative is to use cloud-based LLMs, but only for documentation that is a matter of public record. This assumes that security clearing every invoice would not act as a bottleneck in this process, which seems unlikely. As discussed by two of the interviewees, getting contractual consent from the contractor to process invoices through cloud-based LLMs is also a potential solution. Another is to use STA's on-premises AI platform, which it is currently building. However, interviewees have noted that the capacity of this on-premises system is limited and that it restricts access to the most advanced, commercially available models.

Analyzing these security constraints reveals that while data privacy is a concern, the actual severity of this barrier remains uncertain. The empirical findings indicate some internal confusion at STA regarding current AI policies, and it remains unclear exactly how often contractors actually invoke business secrecy to protect their pricing in practice. Because of this ambiguity, information security can be seen as a constraint rather than a problem, as it can be overcome through an on-premise AI solution. The true challenge is not whether this barrier can be bypassed, but rather at what cost. Both options, relying on an under-resourced on-premise platform or filtering which specific documents are cleared for public cloud LLMs, risk reducing the overall efficiency impact of the automation. Ultimately, navigating these security constraints would likely also make the AI system more costly and resource-intensive to implement, forcing STA to weigh strict data sovereignty against optimal processing capabilities and financial return.

5.2.3 Trust Calibration and the Risk of Automation Complacency

Psychologically, Lee and See (2004) warn of trust calibration issues in automation. This mirrors the interviewees' concerns regarding "automation complacency", where an overworked reviewer might blindly trust a "false green tick", as well as the double standard that accepts human fatigue but expects AI to be flawless.

One risk that concerned interviewees, described by C. A. Zhang et al. (2022) as automation complacency, is that a fatigued reviewer may treat a green flag as a license to approve an invoice without checking whether it is warranted. Even a highly accurate system will produce false negatives, as implied by Z. Zhang et al. (2025). Therefore, an operator who has learned to trust the system's outputs will systematically miss the cases the system fails to catch. Addressing this requires deliberate design decisions about how outputs are presented to reviewers, how frequently the system's performance is audited, and whether workload conditions allow reviewers to apply genuine rather than nominal oversight. These are organizational and management questions that fall outside the scope of any technical deployment.

This implies that the primary roadblocks to AI adoption are not computational, but infrastructural and behavioral. If the commissioning organization cannot contractually compel upstream data standardization or ensure secure processing within local environments, the theoretical capabilities of advanced AI remain practically constrained.

5.2.4 Organizational Barriers

The ultimate bottleneck for AI automation at STA is organizational rather than technical. A critical barrier is the complete absence of performance metrics. Bartolini (2024) emphasizes "straight-through processing rates" and exception frequencies as fundamental indicators of AP maturity. Similarly, the benchmarking report by Medius (2025) highlights cycle times as a core baseline. However, respondents confirmed that STA's invoice review organization operates without any such KPIs. This absence creates a capability gap for the commissioning organization, a vulnerability described by Winch and Leiringer (2016). Without quantitative baselines, STA cannot construct a viable business case for AI or measure its return on investment.

Implementation is further hindered by a deficit in AI knowledge and contractual competence. Reviewers frequently lack the specialized construction law expertise needed to confidently navigate contractual gray areas. Simultaneously, they face ambiguity about strict information security guidelines and which data AI tools can legally process. In addition, the adoption and use of AI tools among interviewees is limited to non-existent. Lee and See (2004) argue that trust is the primary attitude guiding reliance on automation. Consequently, when users lack knowledge about an AI system's boundaries and security compliance, it leads to disuse, a state where the automation is systematically rejected by its intended users. At STA, this knowledge gap is compounded by IT council evaluations and an understaffed IT department that stifles proactive digital initiatives. One potential mechanism for overcoming this barrier is to actively encourage low-risk, hands-on exposure to generic LLM tools within the review organization. By gaining practical experience, reviewers can build an intuitive understanding of the technology's capabilities and limitations, effectively calibrating their trust and mitigating the risk of systematic disuse.

Finally, one interviewee at a consultancy firm mentioned that those whose daily work revolves around manual invoice processing possess a natural, self-preserving incentive against automation, which was characterized as "old habits". This operational friction could be explained by the theory of fragmented sensemaking (Brunet, 2021). At the institutional level, STA prioritizes rigorous IT security and procedural legitimacy, resulting in heavy bureaucratic demands. Conversely, the project level prioritizes efficiency and the practical execution of invoice approvals. When institutional bureaucracy clashes with project-level resistance to changing daily workflows, obvious efficiency measures are missed, making organization-wide AI rollouts highly risky without unified governance. However, this does not seem to be grounded in empirical findings, since none of the STA staff interviewed shared the view of there being a self-preserving incentive against automation. The general perception was that an automation solution would be well received by the staff.

5.3 A Proposed Framework for AI-Supported Review

5.3.1 Governance Enabled by AI

A common theoretical tension is whether strict formal controls weaken relational trust in collaborative projects. While Poppo and Zenger (2002) discuss governance in a general business sense, their complementarity logic is relevant to how AI can support project oversight. In their framework, formal tools are used to define expectations and deliveries, while trust is the essential ingredient for flexible collaboration. In the context of the West Link, a rigorous invoice review may strengthen transparency for the commissioning organization by ensuring that what is being paid for actually matches the contractual expectations. Rather than the review signaling a lack of trust, using AI to objectively settle the contract of the project could potentially removes a source of administrative friction. The introduction of AI presents a method to enforce these controls comprehensively. Modern frameworks suggest that shifting from manual sampling to full-population analysis using AI enables organizations to systematically detect rare but critical anomalies.

Empirically, the project's stakeholders explicitly desire a system built on this logic, emphasizing an "augmentation, not replacement" model. They identified that automating repetitive "hygiene checks", such as verifying hourly rates against price lists and confirming approved resources, would free up human experts to perform the high-value field verifications that actually save money. This aligns with the human-in-the-loop framework of Lee and See (2004), which emphasizes that humans must retain ultimate responsibility.

This indicates that implementing AI in cost-plus megaprojects should not be viewed as deploying an automated decision-maker, but rather an advanced governance filter. By leveraging AI to verify contractual baselines across the entire invoice population, the commissioning organization can effectively enforce the formal contract without signaling distrust, shifting the human focus from data matching to complex problem-solving. To resolve the gap between the theoretical ideals of continuous auditing and the operational reality at STA, an AI implementation framework must be highly structured. Anwar and Akeel (2026) establish that data governance and integration form the foundational layer of any tiered audit workflow architecture. In practice, this aligns with the empirical findings that AI cannot simply be overlaid onto messy workflows. A functional, AI-supported invoice review system for large infrastructure projects requires a phased architecture that addresses data prerequisites, cognitive automation, thresholding, and trust calibration. This proposed framework indicates that successful AI integration in cost-plus megaprojects cannot rely solely on the technological deployment of advanced models. It requires a systemic overhaul combining strict upstream contractual data mandates with targeted, threshold-based cognitive automation. By designing the AI specifically to act as an explainable filter rather than an autonomous decision-maker, STA can bypass the inherent security and cost limitations of LLMs while systematically neutralizing the information asymmetry that currently plagues manual review.

5.3.2 Generalizability and Boundary Conditions

Before examining the framework in detail, it is worth considering where its applicability is strongest and where it is likely to break down. The conditions at the West Link, cost-plus contracts, multi-tier subcontracting, high invoice volumes, limited owner capacity, and concentrated contractor markets, are common features of large public infrastructure projects in many countries. The framework should therefore be relevant to other commissioning organizations operating in similar structural conditions, not only in Sweden. However, the specific combination of data systems, security requirements, and supplier relationships will differ, and the framework's phasing will need to adapt accordingly.

Organizations that already enforce digital originals and consistent data standards from their contractors, which is increasingly common in public-sector procurement in northern Europe, are in a much stronger position to deploy AI processing components directly. Organizations facing the data infrastructure deficit described in Section 5.2 will need to invest substantially in Phase 1 before later phases deliver meaningful value. There is also a maturity dimension. The framework assumes a commissioning organization with enough technical capacity to configure and maintain an automated system. Smaller or less technically resourced organizations may find the prerequisite investment prohibitive. For such organizations, the more immediately relevant finding may be the upstream contractual requirements, which can be adopted as procurement conditions regardless of whether AI is eventually deployed. In Figure 5.1 below, the proposed solution framework is illustrated. It consists of three distinct phases, and each is described in detail in the following sub-sections.

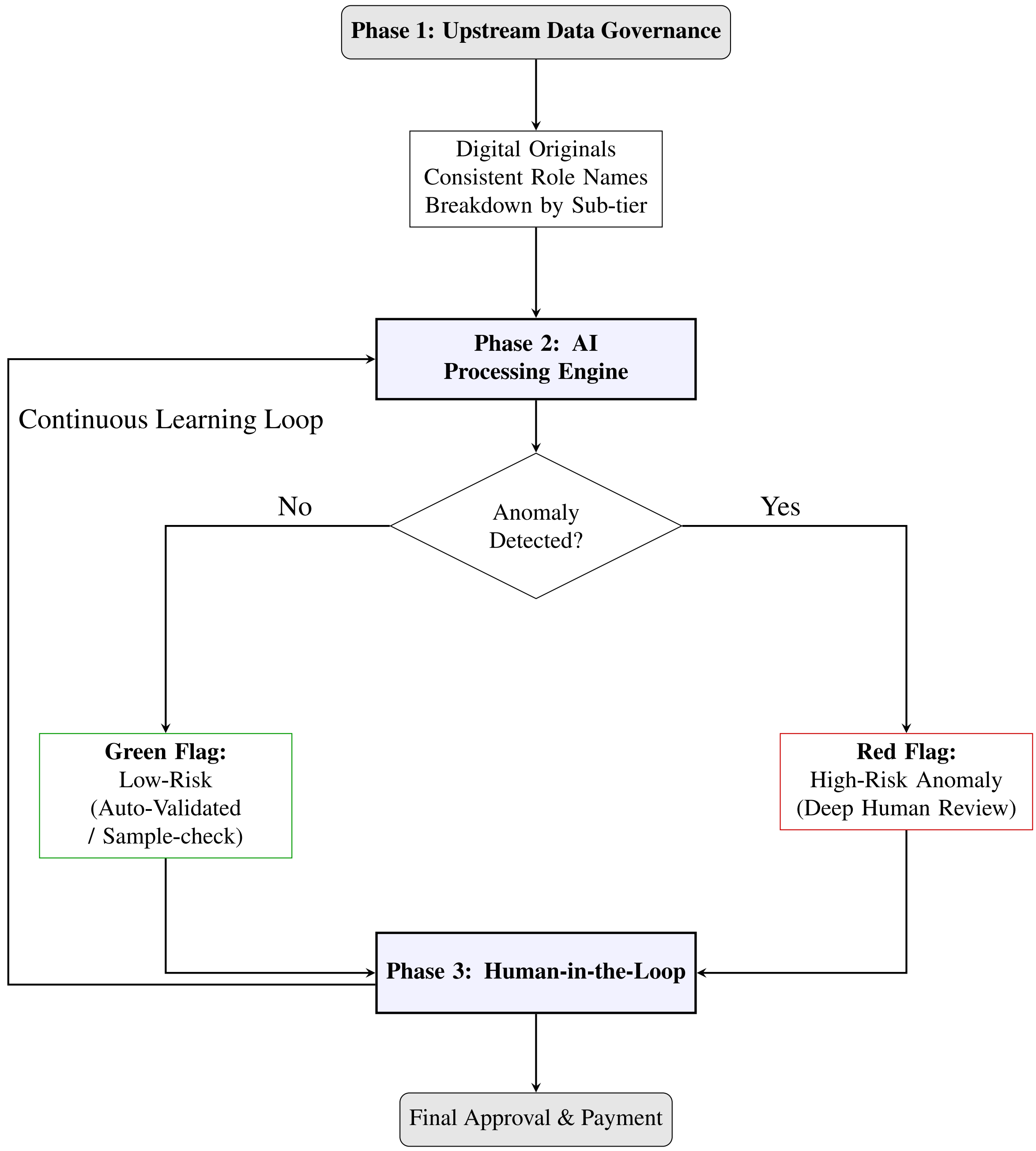

Figure 5.1: Proposed Framework for AI-Supported Invoice Review at STA

5.3.3 Phase 1: Upstream Data Governance

The literature notes that incompleteness and inconsistencies in data severely limit the identification of risks. Empirically, respondents emphasized that STA must place harder demands on main contractors before AI can be effective. This upstream governance layer must enforce specific prerequisites, such as exact role names consistent between the contract and the invoice, and all documents must be submitted as digital originals rather than scanned copies. By contractually enforcing these standards, the commissioning organization creates the structured data environment necessary for the following automated extraction. However, it is clear from the interviews that if this data governance requirement results in increased costs for the contractors, it is likely that the cost will be passed on to STA. Although this might be a minor cost, it would be beneficial to conduct a cost-benefit analysis of this, possibly in collaboration with a supplier, to accurately estimate the long-term financial implications. One could also argue that this risk is enlarged by STA's relatively weak bargaining power in relation to the contractors, as established by the power asymmetry between them.

5.3.4 Phase 2: AI Processing Engine

The high-level design of the AI processing engine, illustrated in Figure 5.2, is based on the practical feasibility of automating each review check and the architectural principles recommended in the literature for AI systems applied to structured document review.

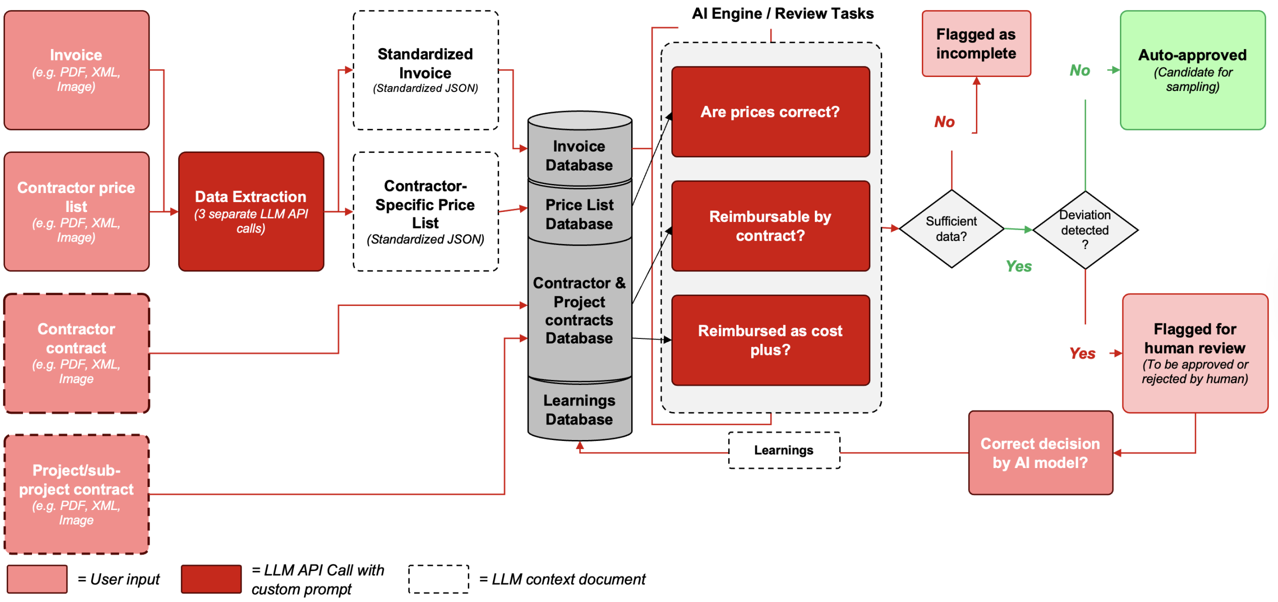

Figure 5.2: Phase 2: First-stage framework for AI-supported invoice review

To arrive at a feasible first-stage scope, each of the seven checks that make up the project engineer's review process was assessed along two dimensions grounded in the empirical findings. The first was what share of total review time each check consumes, and the second was how difficult it would be to automate reliably. This assessment indicates that price verification accounts for roughly 30% of total review time for one invoice and is well-suited for AI-augmented automation. It is essentially a matching task between invoiced line items and a contractor-specific price list that already exists as a separate reference document. The contractual reimbursability check and the cost-plus reimbursement check each consume between 5 and 10% of review time and are similar in terms of feasibility. This is because they both rely on binary, document-level signals rather than contextual field judgment. By contrast, the supporting documentation check at approximately 25% of review time and the subcontracting chain check at 10 to 15% of review time are both judged difficult to automate. The documentation check requires an understanding of which documents are even relevant for a given invoice type, and the subcontracting chain check requires tracing multi-tier supplier relationships that are not consistently recorded in the invoice data.

The three checks highlighted in Figure 5.2, namely price correctness, contract reimbursability, and cost-plus reimbursability, represent an overlap between high time cost and high automation feasibility, which is an area where AI can deliver a reduction in manual effort without introducing unreliable outputs. This is consistent with the principles put forward by Anthropic (n.d.) and OpenAI (n.d.) that complexity should be added only when it demonstrably improves outcomes, and that implementation should begin with a single-step approach before expanding to more orchestrated designs.

Anthropic (n.d.) distinguishes between workflows, where LLM calls are orchestrated through predefined code paths, and agents, where the model dynamically directs its own tool use and sequencing. For a structured, well-defined task such as invoice review against contractual criteria, a workflow architecture is arguably more appropriate. This is because it offers greater predictability and stepwise auditability than financial verification requires. It therefore limits the risk that the model will take unintended actions on high-stakes documents. The three checks map onto three predefined, sequential LLM API calls, one per document type, specifically the invoice, the contractor price list, and the attest diary, as shown in Figure 5.2. Each call uses a constrained prompt that limits the model's reasoning to the relevant task.

The data extraction layer that precedes the checks utilizes the RAG framework outlined in section 2.4.6, applying the augmented LLM architecture described by Anthropic (n.d.), where the base language model is enhanced with retrieval and structured document access. Rather than passing entire documents incorporated into a single prompt, the framework uses separate extraction modules for each source, producing standardized JavaScript Object Notation (JSON) outputs that populate a set of internal databases. This reduces context overload and the risk of hallucination by limiting each model call to a narrowly defined retrieval task, and it makes the origin of every extracted data point more traceable. Kanaparthi (2023) notes that AI can use semantic similarity to match descriptions between invoices and POs in ways that rigid, template-based OCR systems cannot, which is particularly relevant for price verification, where line-item descriptions may vary in wording across contractors.

According to Svanberg et al. (2025), moving from sample-based to full-population review creates a risk of exception overload. Svanberg et al. (2025) address this through thresholding mechanisms that control alert flow and direct attention toward genuine anomalies. Figure 5.2 attempts to incorporate this logic through a binary output logic. Where all three checks pass without a detected deviation, the invoice receives a green flag and is routed for standard sample-checking by a human reviewer. Where any check detects a discrepancy, a red flag halts processing and triggers a deeper human review. This design reflects the guardrail principle identified by OpenAI (n.d.), that systems should be designed to pause and defer to humans when outcomes exceed their defined operational bounds. In this case, the threshold is not a confidence score but a detected deviation from a contractual reference, which is both easier for reviewers to interpret and more defensible in an audit context.

Figure 5.2 is explicitly labeled a first-stage framework. The three checks it covers are those where the data prerequisites are most manageable, and the automation logic is most deterministic, yet hard to automate through rules-based programs. The checks that are excluded are not abandoned, but deferred. Their exclusion reflects the principle of expanding scope only when simpler implementations have proven stable (Anthropic, n.d.; OpenAI, n.d.), and the empirical recognition that those checks require a degree of contextual reasoning, and in some cases physical field presence, that goes beyond what a language model can reliably provide at this stage.

5.3.5 Phase 3: Explainability and Human-in-the-Loop Validation

Jans and Hosseinpour (2019) and Lee and See (2004) strongly advocate for a human-in-the-loop framing where the human operator retains ultimate responsibility. The empirical findings mirror this exact sentiment, with respondents universally agreeing that AI should act as an augmentation tool to flag patterns, not a replacement that makes final judgments. For the human reviewer to properly calibrate their trust and avoid the risk of "false green ticks", the system's outputs must be explainable and evidence-linked. C. A. Zhang et al. (2022) identify explainable AI as a method to improve transparency and support audit documentation. Finally, the framework must be iterative, meaning that when a human corrects the AI, the system must continuously learn and incorporate that feedback for future invoices.

Beyond immediate anomaly detection, the continuous learning loop could serve another function by acting as an artificial institutional memory. The empirical findings reveal that STA's current invoice review relies on the tacit knowledge and subjective experience of individual reviewers to interpret contractual edge cases. This reliance on personal judgment is probably part of an explanation for the observed organizational fragmentation, where different sub-projects within the West Link have developed different routines and control processes. This could also be a problem since, for decade-long projects like the West Link, human turnover inevitably degrades the ability to recall historical precedents and informal agreements, called organizational memory loss by Eriksson et al. (2019). By codifying human corrections and routine judgments into the AI system over time, the tacit knowledge can be transformed into more of a permanent, standardized organizational asset. This would not only enable a unified standard of review across disparate sub-projects, but it could also neutralize the informational advantage contractors currently gain when experienced commissioning staff leave the project.

6

Conclusion

This chapter concludes the study by synthesizing the empirical findings and theoretical insights to answer the research questions. It presents the core takeaway regarding the proposed AI-supported framework for invoice review, discusses the limitations and boundary conditions of the research design, and concludes with actionable suggestions for future research within megaproject governance.

6.1 Synthesis of Findings

This study utilized an abductive, qualitative single-case design focused on STA's West Link project to investigate the mechanisms driving cost deviations in large-scale infrastructure investments. Through semi-structured interviews with project stakeholders, shadowing observations, and an iterative literature review, the research aimed to understand the limitations of manual invoice review under cost-plus arrangements. Consequently, the study sought to develop a feasible, first-stage framework for integrating AI to automate the invoice review process.

Regarding the first research question (RQ1), the findings reveal that cost deviations are largely permitted by structural vulnerabilities and control gaps rather than solely unpredictable technical complexity. STA's manual invoice review process suffers from severe volume overload, forcing reviewers to abandon comprehensive, full-population verification in favor of fragmented sampling. This resource constraint intensifies the inherent information asymmetry of cost-plus contracts, where contractors possess superior insight into underlying costs, such as hidden intra-group pricing and retroactive kickbacks. Together with a structural incentive misalignment, where a contractor's absolute profit scales with total costs incurred, these limitations allow for gradual cost escalation and the exploitation of contractual gray areas.

To address these control gaps (RQ2), this study proposes a three-phase framework for AI-supported invoice review. Crucially, the empirical findings and theoretical synthesis demonstrate that AI should not be deployed as an autonomous decision-maker, but as an advanced filter designed for "augmentation, not replacement". The framework consists of three phases:

- **Phase 1: Upstream Data Governance:** Recognizing that AI cannot resolve fundamentally absent or chaotic data, the framework requires STA to contractually mandate standardized data inputs, including digital originals, consistent role nomenclature, and explicit sub-tier breakdowns.

- **Phase 2: AI Processing Engine:** The system targets the most time-consuming yet feasible to automate verification steps, specifically price list matching, cost-plus reimbursability checks, and contract reimbursability checks. By utilizing a workflow of constrained LLMs enhanced with retrieval tools, the system can extract and verify unstructured data across massive volumes.

- **Phase 3: Human-in-the-Loop Validation:** To prevent automation complacency and exception overload, the framework employs a binary thresholding logic. Invoices passing all automated checks receive a "green flag" for standard sampling, while identified discrepancies trigger a "red flag" requiring deep human review.

Ultimately, this study concludes that integrating AI into megaproject governance requires more than technological deployment, as an overhaul of data prerequisites and human trust calibration is just as important. By adopting this framework, commissioning organizations can shift from manual sampling to comprehensive verification, effectively mitigating information asymmetry and taking a decisive step toward becoming "strong owners" in complex supplier relationships.

6.2 Limitations

The primary limitation of this approach is its exclusive focus on STA's perspective. Excluding contractors limits the understanding of upstream data structuring and invoicing practices, leaving the friction between project-level execution and institutional documentation demands partially unexplored.

Despite lacking statistical generalizability due to the single-case design, the findings achieve high analytical generalizability. The structural conditions observed at the West Link, such as high uncertainty, long durations, and the necessity of cost-plus contracts due to scope complexity, are defining characteristics of infrastructure megaprojects globally according to both Brunet (2021) and Suprapto et al. (2016). Therefore, one can argue that the identified control gaps and the framework for AI-supported review are transferable to other large public contracting organizations operating under similar governance constraints.

Finally, due to strict information security regulations restricting access to unmasked invoice data, the AI processing engine could not be empirically tested. While the framework establishes the necessary architectural prerequisites, its technical effectiveness and the actual risk of exception overload would require validation. This could be done through, e.g., a secure, live pilot implementation.

6.3 Suggestions for Future Research

Building on the findings and limitations of this study, three primary avenues for future research are recommended to advance the integration of AI in megaproject governance. First, since strict information security regulations prevented live testing during this study, future research should implement the proposed Phase 2 AI processing engine within a secure, on-premise IT infrastructure. A quantitative pilot study is necessary to measure the actual reduction in manual review time, benchmark the accuracy of the LLM in matching unstructured invoice data, and assess the real-world risk of exception overload.

Second, as this study focused exclusively on the commissioning organization's perspective, subsequent research should investigate the supply chain's readiness to adopt the upstream data governance mandates outlined in Phase 1. Studies are needed to evaluate the financial and administrative implications for contractors when forced to provide standardized, digital-original documentation and explicit sub-tier breakdowns.

Finally, as AI systems take over routine verification tasks, longitudinal studies should observe how project engineers calibrate their trust over time. Research is needed to develop organizational strategies and interface designs that effectively mitigate automation complacency, ensuring that AI-generated flags do not degrade the quality of critical human oversight in the long term.

References

Abela, R. (2024). *A qualitative assessment of the approaches adopted by malta-based audit firms for the detection of money laundering: Insights from semi-structured interviews* (Master of Science in Insurance Risk Management). University of Malta, Msida, Malta.

Alkhaled, L., & Fei, N. Y. (2023). Automated invoice processing system. In *2023 ieee international conference on industrial engineering and engineering management (ieem)* (p. 0188-0192). doi: 10.1109/IEEM58616.2023.10406704

Alles, M. G., Kogan, A., & Vasarhelyi, M. A. (2024). Principles and problems of audit automation as a precursor to continuous auditing. *International Journal of Accounting Information Systems*. (In Press/Early Access)

Amari, A., Makni, M., Fnaich, W., Lahmar, A., Koubaa, F., Charrad, O., … Douss, R. Y. (2024). An efficient deep learning-based approach to automating invoice document validation. In *2024 ieee/acs 21st international conference on computer systems and applications (aiccsa)* (p. 1-8). doi: 10.1109/AICCSA63423.2024.10912544

Anthropic. (n.d.). *Building effective agents.* `https://www.anthropic.com/engineering/building-effective-agents`.

Anwar, A., & Akeel, M. O. (2026). Integrating artificial intelligence in audit workflow: Opportunities, architecture, and challenges: A systematic review. *Preprints.org*. (Review, Posted: 27 January 2026) doi: 10.20944/preprints202601.2060.v1

Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). Docformer: End-to-end transformer for document understanding. *arXiv preprint arXiv:2106.11539*.

Artsyl. (2025, December 26). *Invoice verification: Process and standards of invoice checking.* Retrieved from `https://www.artsyltech.com/invoice-verification-as-part-of-invoice-processing` (Accessed: February 2, 2026)

Bajari, P., & Tadelis, S. (2001). Incentives versus transaction costs: A theory of procurement contracts. *The RAND Journal of Economics*, *32*(3), 387–407. Retrieved 2026-02-12, from `http://www.jstor.org/stable/2696361`

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded u.s. firms using a machine learning approach. *Journal of Accounting Research*, *58*(1), 199-

235. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-679X.12292 doi: https://doi.org/10.1111/1475-679X.12292

Bardelli, C., Rondinelli, A., Vecchio, R., & Figini, S. (2020). Automatic electronic invoice classification using machine learning models. *Machine Learning and Knowledge Extraction*, *2*(4), 617–629. doi: 10.3390/make2040033

Bartolini, A. (2024, 05). *The state of epayables 2024: Money never sleeps* (Research report No. PAX-NA-SRR-2406-2585). Ardent Partners Ltd. (Underwritten by Bottomline)

Berghaus, D., Berger, A., Hillebrand, L., Cvejoski, K., & Sifa, R. (2025). *Multi-modal vision vs. text-based parsing: Benchmarking llm strategies for invoice processing.* Retrieved from https://arxiv.org/abs/2509.04469

Bikute, M. (2024, April 3). *Best practices in invoice processing workflow: Maximising efficiency.* Kefron. Retrieved from https://kefron.com/2024/04/best-practices-invoice-processing-workflow-efficiency/?switch=rw

Blomkvist, P., & Hallin, A. (2015). *Method for engineering students: Degree projects using the 4-phase Model* (1st ed.). Lund: Studentlitteratur. Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:mdh:diva-41032

Brunet, M. (2021). Making sense of a governance framework for megaprojects: The challenge of finding equilibrium. *International Journal of Project Management*, *39*(4), 406-416. Retrieved from https://www.sciencedirect.com/science/article/pii/S0263786320300703 (Trust and Governance in Megaprojects) doi: https://doi.org/10.1016/j.ijproman.2020.09.001

Bryman, A., & Bell, E. (2011). *Business research methods* (3rd ed.). Oxford, UK: Oxford University Press.

Delray Consulting AB. (2025, August). *Granskningsrapport avseende projekt korsvägen* (Tech. Rep.). Nacka, Sweden: Delray Consulting AB. (Date: 2025-08-29)

Dubois, A., & Gadde, L.-E. (2002). Systematic combining: an abductive approach to case research. *Journal of Business Research*, *55*(7), 553-560. Retrieved from https://www.sciencedirect.com/science/article/pii/S0148296300001958 (Markets as Networks) doi: https://doi.org/10.1016/S0148-2963(00)00195-8

Eliasson, J. (2025). Cost overruns of infrastructure projects – distributions, causes and remedies. *Transportation Research Part A: Policy and Practice*, *198*, 104532. Retrieved from https://www.sciencedirect.com/science/article/pii/S0965856425001600 doi: https://doi.org/10.1016/j.tra.2025.104532

Eriksson, P. E., & Lind, H. (2015). *Moral hazard and construction procurement: A conceptual framework.* Retrieved from https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-159237 (KTH Royal Institute of Technology website)

Eriksson, P. E., Volker, L., Kadefors, A., Lingegård, S., Larsson, J., & Rosander, L. (2019).

Collaborative procurement strategies for infrastructure projects: a multiple-case study. *Proceedings of the Institution of Civil Engineers - Management, Procurement and Law*, *172*(5), 197-205. Retrieved from https://www.sciencedirect.com/science/article/pii/S1751430419000254 doi: https://doi.org/10.1680/jmapl.19.00016

Fedyk, A., Hodson, J., Khimich, N., & Fedyk, T. (2022, September). Is artificial intelligence improving the audit process? *Review of Accounting Studies*, *27*(3), 938–985. doi: 10.1007/s11142-022-09697-x

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, *16*(1), 3–21. Retrieved from https://doi.org/10.1080/09654310701747936 doi: 10.1080/09654310701747936

Flyvbjerg, B. (2018, March). Planning fallacy or hiding hand: Which is the better explanation? *World Development*, *103*, 383–386. Retrieved from http://dx.doi.org/10.1016/j.worlddev.2017.10.002 doi: 10.1016/j.worlddev.2017.10.002

Flyvbjerg, B., Bruzelius, N., & Rothengatter, W. (2003). *Megaprojects and risk: An anatomy of ambition*. Cambridge, UK: Cambridge University Press.

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, *68*(3), 279–295. Retrieved from https://doi.org/10.1080/01944360208976273 doi: 10.1080/01944360208976273

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., … Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey.* Retrieved from https://arxiv.org/abs/2312.10997

Geraldi, J., Maylor, H., & Williams, T. (2011, 08). Now, let's make it really complex (complicated): A systematic review of the complexities of projects. *International Journal of Operations & Production Management*, *31*(9), 966-990. Retrieved from https://doi.org/10.1108/01443571111165848 doi: 10.1108/01443571111165848

Gioia, D., Corley, K., & Hamilton, A. (2013, 01). Seeking qualitative rigor in inductive research. *Organizational Research Methods*, *16*, 15-31. doi: 10.1177/1094428112452151

Gotthardt, M., Koivulaakso, D., Paksoy, O., Saramo, C., Martikainen, M., & Lehner, O. (2020). Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. *ACRN Journal of Finance and Risk Perspectives*, *9*, 90–102.

Grahn, E. (2024). *Quality control of infrastructure construction in the swedish market* (Bachelor's thesis). Karelia University of Applied Sciences.

Hillebrand, L., Berger, A., Deußer, T., Dilmaghani, T., Khaled, M., Kliem, B., … Sifa, R. (2023). *Improving zero-shot text matching for financial auditing with large language models.* Retrieved from https://arxiv.org/abs/2308.06111

Houseblend. (2025, November 24). *Netsuite anomaly detection: Using AI for vendor invoices.* Retrieved from https://www.houseblend.io/articles/netsuite-vendor-bill-anomaly-detection (Houseblend)

Hu, Y., Chan, A. P. C., Le, Y., & Jin, R. (2015). From construction megaproject management to complex project management: Bibliographic analysis. *Journal of Management in Engineering*, *31*(4), 04014052. Retrieved from https://ascelibrary.org/doi/abs/10.1061/%28ASCE%29ME.1943-5479.0000254 doi: 10.1061/(ASCE)ME.1943-5479.0000254

IFOL. (2020). *Accounts payable performance benchmark report 2020: How to use key metrics to measure your ap performance improvement journey* (Benchmarking report). Institute of Financial Operations & Leadership (IFOL).

IOFM. (2021). *World class ap performance: Efficiency benchmarking metrics* (Benchmarking report). Institute of Finance & Management (IOFM), Diversified Communications.

Jans, M., & Hosseinpour, M. (2019). How active learning and process mining can act as continuous auditing catalyst. *International Journal of Accounting Information Systems*, *32*, 44-58. Retrieved from https://www.sciencedirect.com/science/article/pii/S1467089517300180 doi: https://doi.org/10.1016/j.accinf.2018.11.002

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., … Neubig, G. (2023). Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7969–7992). Association for Computational Linguistics.

Jin, H., Papadimitriou, G., Raghavan, K., Zuk, P., Balaprakash, P., Wang, C., … Deelman, E. (2024). *Large language models for anomaly detection in computational workflows: from supervised fine-tuning to in-context learning.* Retrieved from https://arxiv.org/abs/2407.17545

Johri, A., Sayal, A., Chong, K. M., Khoja, M., Chaithra, N., Jha, J., & Tyagi, N. (2026, January). Enhancing audit quality and reducing costs: the impact of AI in banking and financial services. *Frontiers in Artificial Intelligence*, *8*. doi: 10.3389/frai.2025.1718854

Joshi, S. (2025). Advancing innovation in financial stability: A comprehensive review of ai agent frameworks, challenges and applications. *World Journal of Advanced Engineering Technology and Sciences*, *14*(2), 117–126. doi: 10.30574/wjaets.2025.14.2.0071

Kanaparthi, V. K. (2023). Examining the plausible applications of artificial intelligence & machine learning in accounts payable improvement. *FinTech*, *2*(3), 461–474. Retrieved from https://doi.org/10.3390/fintech2030026 doi: 10.3390/fintech2030026

Khanchandani, K., Thakur, A., Shetty, A., Reddy, C., & Behera, R. (2026). *Automated invoice data extraction: Using llm and ocr.* Retrieved from https://arxiv.org/abs/2511.05547

Kim, A. G., Muhn, M., & Nikolaev, V. V. (2024). *Financial statement analysis with large language models.* Retrieved from https://arxiv.org/abs/2407.17866

Kim, E. W., Shin, Y. J., Kim, K. J., & Kwon, S. (2025). Development of an automated construction contract review framework using large language model and domain knowledge. *Buildings*, *15*(6), 923. Retrieved from https://doi.org/10.3390/buildings15060923 doi: 10.3390/buildings15060923

Laujan. (n.d.). *Invoice data extraction – document intelligence - foundry tools.* Microsoft Learn. Retrieved from https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/prebuilt/invoice?view=doc-intel-4.0.0 (Accessed: 2026-02-03)

LeCompte, M. D., Preissle, J., & Tesch, R. (1993). *Ethnography and qualitative design in educational research* (2nd ed.). San Diego, CA: Academic Press.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. doi: 10.1518/hfes.46.1.50.30392

Li, P., Chan, D. Y., & Kogan, A. (2016). Exception prioritization in the continuous auditing environment: A framework and experimental evaluation. *Journal of Information Systems*, *30*(2), 135–157. doi: 10.2308/isys-51220

Li, Z.-Z., Zhang, D., Zhang, M.-L., Zhang, J., Liu, Z., Yao, Y., ... Liu, C.-L. (2025). *From system 1 to system 2: A survey of reasoning large language models.* Retrieved from https://arxiv.org/abs/2502.17419

Medius. (n.d.). *Ap automation explained: What is accounts payable automation?* https://www.medius.com/glossary/what-is-accounts-payable-automation/. (Accessed: 2026-02-05)

Medius. (2025). *The accounts payable automation kpis* (Benchmarking report). Medius. (Benchmarks based on anonymized and aggregated invoice processing data from the Medius customer community)

Minaee, S., Mikolov, T., Nikzad, N., Chennubhotla, M., Darvishi, P., & Gao, J. (2024). *Large language models: A survey.* Retrieved from https://doi.org/10.48550/arXiv.2402.06196

Mindee. (2025, July 9). *LLM vs OCR API: Cost comparison for document processing in 2025.* Retrieved from https://www.mindee.com/blog/llm-vs-ocr-api-cost-comparison (Accessed: 2026-02-04)

Neuvonen, J. (2015). *Developing purchase invoice processing through workflow automation* (Master's thesis). Lappeenranta University of Technology, Industrial Engineering and Management, Lappeenranta, Finland.

OpenAI. (n.d.). *A practical guide to building ai agents.* https://openai.com/business/guides-and-resources/a-practical-guide-to-building-ai-agents/.

Pinto, J. K. (2014). Project management, governance, and the normalization of deviance. *International Journal of Project Management*, *32*(3), 376-387. Retrieved from https://

www.sciencedirect.com/science/article/pii/S0263786313000835 doi: https://doi.org/10.1016/j.ijproman.2013.06.004

Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal*, *23*(8), 707–725. Retrieved 2026-02-11, from http://www.jstor.org/stable/3094289

Porter, M. E. (1980). *Competitive strategy: Techniques for analyzing industries and competitors*. New York: Free Press. (Republished with a new introduction, 1998)

Rozario, A. M., & Issa, H. (2020). Risk-based data analytics in the government sector: A case study for a u.s. county. *Government Information Quarterly*, *37*(2), 101457. doi: 10.1016/j.giq.2020.101457

Salemi, A., & Zamani, H. (2024). Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th international acm sigir conference on research and development in information retrieval* (p. 2395–2400). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3626772.3657957 doi: 10.1145/3626772.3657957

Saout, T., Lardeux, F., & Saubion, F. (2024). An overview of data extraction from invoices. *IEEE Access*, *12*, 19872–19886. doi: 10.1109/ACCESS.2024.3360528

Schneider dos Santos, E., Machado dos Santos, M., Castro, M., et al. (2025). Detection of fraud in public procurement using data-driven methods: a systematic mapping study. *EPJ Data Science*, *14*(52). Retrieved from https://doi.org/10.1140/epjds/s13688-025-00569-3 doi: 10.1140/epjds/s13688-025-00569-3

Shen, J., Yuan, P., Ghosh, A., Mai, Y., & Dahlmeier, D. (2026). *Ocr or not? rethinking document information extraction in the mllms era with real-world large-scale datasets.* Retrieved from https://arxiv.org/abs/2603.02789

Singhal, R. (2021, jan 12). *Improving invoice anomaly detection with AI and machine learning.* Ericsson. Retrieved from https://www.ericsson.com/en/blog/2021/1/improving-invoice-anomaly-detection-with-ai-and-machine-learning (Accessed: 2026-02-06)

Stampli. (2024, may 1). *Streamline AP processes with 3-way match automation.* https://www.stampli.com/blog/ap-automation/3-way-match-automation/.

Suprapto, M., Bakker, H. L., Mooi, H. G., & Hertogh, M. J. (2016). How do contract types and incentives matter to project performance? *International Journal of Project Management*, *34*(6), 1071-1087. Retrieved from https://www.sciencedirect.com/science/article/pii/S0263786315001349 doi: https://doi.org/10.1016/j.ijproman.2015.08.003

Svanberg, J., Öhman, P., Samsten, I., Neidermeyer, P., Rana, T., & Danielson, M. (2025). Addressing the exception prioritization problem in continuous auditing systems with thresholding. *Intelligent Systems in Accounting, Finance and Management*, *32*(4), e70022. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.70022` (e70022 ISAFM-24-024) doi: https://doi.org/10.1002/isaf.70022

Tang, P., Qiu, W., Huang, Z., Chen, S., Yan, M., Lian, H., & Li, Z. (2020). Anomaly detection in electronic invoice systems based on machine learning. *Information Sciences*, *535*, 172-186. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0020025520302693` doi: https://doi.org/10.1016/j.ins.2020.03.089

Thomas, E. (2025, July 8). *Invoice OCR line items extraction.* Mindee. Retrieved from `https://www.mindee.com/blog/invoice-ocr-line-items-extraction` (Accessed: 2026-02-04)

Trafikverket. (n.d.). *Västlänken.* `https://www.trafikverket.se/vara-projekt/projekt-i-vastra-gotalands-lan/vastlanken/`. (Accessed: 2026-02-13)

Trafikverket. (2025a, September). *Fakturagranskning e05 korsvägen.* Internal document.

Trafikverket. (2025b, December). *Pr guide för externa ai-tjänster* (Internal PM). Trafikverket.

Uusitalo, P., Peltokorpi, A., Seppänen, O., & Alhava, O. (2024, 11). Towards systemic transformation in the construction industry: a complex adaptive systems perspective. *Construction Innovation*, *24*(7), 341-368. Retrieved from `https://doi.org/10.1108/CI-01-2024-0015` doi: 10.1108/CI-01-2024-0015

Wang, R., Jiang, J., Yang, C., Li, B., et al. (2025). *Finmaster: A holistic benchmark for mastering full-pipeline financial workflows with llms.* Retrieved from `https://arxiv.org/abs/2505.13533`

Wang, R., Liu, J., Zhao, W., Li, S., & Zhang, D. (2025). *Automating financial statement audits with large language models.* Retrieved from `https://arxiv.org/abs/2506.17282`

Wang, Y., Wang, K., Yang, S., Patel, J., Zhao, J., Mo, F., … Nie, J.-Y. (2025). *Finauditing: A financial taxonomy-structured multi-document benchmark for evaluating llms.* Retrieved from `https://arxiv.org/abs/2510.08886`

Winch, G. (2010). *Managing construction projects: An information processing approach* (2nd ed.). Chichester, UK: Wiley-Blackwell.

Winch, G. (2014). Three domains of project organising. *International Journal of Project Management*, *32*(5), 721-731. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0263786313001518` doi: https://doi.org/10.1016/j.ijproman.2013.10.012

Winch, G., & Leiringer, R. (2016). Owner project capabilities for infrastructure development: A review and development of the "strong owner" concept. *International Journal of Project Management*, *34*(2), 271-281. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0263786315000356` doi: https://doi.org/10.1016/j.ijproman.2015.02.002

Wondra, M. (2025). Auditing large language models. *SSRN Electronic Journal*. Retrieved from `https://ssrn.com/abstract=5675847` doi: 10.2139/ssrn.5675847

Wong, S., Zheng, C., Su, X., & Tang, Y. (2024). Construction contract risk identification based on knowledge-augmented language models. *Computers in Industry*, *157-158*(3). Retrieved from `https://doi.org/10.1016/j.compind.2024.104082` doi: 10.1016/j.compind.2024.104082

Yang, C., Zhao, R., Liu, Y., & Jiang, L. (2025). *Survey of specialized large language model.* Retrieved from `https://arxiv.org/abs/2508.19667`

Yashwant, S., Dubey, A., Paikray, P., & Thulsiram, G. (2025). *Invoice information extraction: Methods and performance evaluation.* Retrieved from `https://arxiv.org/abs/2510.15727`

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Thousand Oaks, California: SAGE Publications.

Zarei, H., Hui, F. K. P., Duffield, C., & Wang, G. (2017). The risk of power imbalance in project delivery: A study of large victorian public infrastructure projects. *Journal of Risk Analysis and Crisis Response*, *7*(2), 53–63. Retrieved from `https://doi.org/10.2991/jrarc.2017.7.2.1` doi: 10.2991/jrarc.2017.7.2.1

Zhang, C. A., Cho, S., & Vasarhelyi, M. (2022). Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems*, *46*, 100572. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1467089522000240` (2021 Research Symposium on Information Integrity Information Systems Assurance) doi: https://doi.org/10.1016/j.accinf.2022.100572

Zhang, Z., Cao, Y., & Liao, L. (2025). Xfinbench: Benchmarking llms in complex financial problem solving and reasoning. In *Findings of the association for computational linguistics: Acl 2025.* Retrieved from `https://arxiv.org/abs/2508.15861`

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Wen, J.-R. (2023). *A survey of large language models.* Retrieved from `https://doi.org/10.48550/arXiv.2303.18223`

Appendix 1: Interview Guide

Bakgrund och upplägg

- Kort om vilka vi är, examensarbetet, syftet med intervjun och hur materialet används
- Be om tillåtelse att spela in
- Har du några frågor om vårt projekt innan vi börjar?

Bakgrund om personen och mandat

- Beskriv din roll och ditt ansvar kopplat till kontrakt, inköp, uppföljning och fakturagranskning
- Hur länge har du arbetat med den här typen av projekt?
- Vilka beslut eller moment är du själv involverad i, och vilka ligger utanför ditt mandat?
- Vilka roller är normalt involverade när en faktura ska bedömas och godkännas?
    - Vem sätter sista godkännandet i praktiken?

Kontraktsmodellen och varför den används

- Beskriv kontraktsmodellen för Västlänken och varför just denna används
- Löpande räkning
    - När och varför används detta?
    - Vilka risker ser du med modellen jämfört med fast pris?
- Skiljer sig kontraktsmodellen eller tillämpningen mycket beroende på leverantör eller delprojekt?

Ersättningsbarhet och gråzoner

- Hur tydligt är det för alla parter vad som är ersättningsbart och vad som ingår i fast del respektive rörlig del?
- Vilka typer av gråzoner orsakar mest diskussion eller merarbete?
    - Exempel på kostnadstyper ni inte ska betala, men som ändå dyker upp

    - Exempel på definitioner som blir otydliga i praktiken, såsom handverktyg och liknande

- Hur avgör ni i praktiken hur tuffa ni kan vara när ni vill neka betalning eller kräva komplettering?

    - Vad behöver ni ha på fötterna för att hålla inne betalning?

Inköpsprocess och upphandling

- Hur går upphandlingen av leverantörer till?

- Vilka krav ställs vid upphandling kopplat till ekonomiadministration, redovisning, underlag och spårbarhet?

- Finns det lärdomar från tidigare upphandlingar där kraven var för otydliga eller för svåra att följa upp?

Organisation och arbetssätt i fakturagranskningen

- Hur är fakturagranskningen organiserad i projektet?

    - Finns en samordnande roll för granskningen?

    - Hur många personer deltar typiskt i granskningen, och vilka typer av roller är det?

- Hur ser en typisk granskningsdag eller vecka ut för er?

- Granskar ni rad för rad eller använder ni stickprov?

    - Vad avgör vad som hamnar i stickprov eller djupgranskning?

- Hur hanterar ni oklarheter?

    - Vilka kontaktar ni för att verifiera rimlighet, exempelvis byggledning eller produktionsledning?

Processen från leverantör till betalning

- Från att leverantör väljs till att fakturor börjar komma in

    - När uppstår första kraven på struktur och underlag?

- Från att faktura inkommer till att betalning görs
    - Faktureras Trafikverket före eller efter utfört arbete i löpande räkning, och hur påverkar det kontrollen?
- Vilka manuella steg finns i processen, och vilka är mest tidskrävande?

Underlag, verifikat och spårbarhet

- Vilka krav ställs på fakturor och underlag idag?
    - Vad måste finnas med för timmar, datum, utförare och verifikat?
- Vad brister oftast i praktiken?
    - Saknas verifikat, är de svåra att hitta, eller är kopplingen till fakturarad otydlig?
- När ni har väldigt många verifikat, vad är det som tar tid?
    - Att hitta rätt verifikat
    - Att koppla rätt verifikat till rätt kostnad
    - Att hitta avtalsstöd när en rad hänvisar till avtal, sida eller villkor
- Vilka typer av referenser hade gjort störst skillnad för spårbarheten?
    - Standardiserade referenser per rad
    - Entydiga IDn till verifikat
    - Hänvisning till avtalsdel och klausul

Kravställning mot entreprenören

- Vilka krav kan ni ställa för att göra granskningen enklare?
- Vilka praktiska nackdelar ser du med hårdare krav på underlag?
    - Exempelvis att det kan bli dyrare på grund av mer administration hos leverantören

- Kontrollerar entreprenören sina fakturor innan de skickas?
    - Vad är kravet idag?
    - Hur följer ni upp att de faktiskt har kontrollerat, inte bara markerat ok?

Systemstöd och dataflöde

- Vilka system används i processen?
- Var ligger underlagen, och hur söker ni i dem?
- Finns det moment där ni tvingas hoppa mellan system för att hitta avtal och verifikat?
- Vilken data skulle du vilja ha mer strukturerad än idag?

Uppföljning och styrning

- Hur följer ni kostnadsutveckling och prognos i projekten?
    - Vad rapporterar entreprenören, och hur ofta?
- Hur kopplas fakturagranskning till uppföljning och prognosarbete?

Problemformulering och rotorsaker

- Vilka är de största problemen med nuvarande process?
- Vad tror du är orsakerna till problemen?
    - Kompetens och stöd för entreprenadjuridik och kontraktstolkning
    - Otydliga gränsdragningar i kontraktet
    - Svår spårbarhet och stora mängder underlag
    - Bristande egenkontroll hos entreprenören

AI och automatisering

- Tror du att processen går att förbättra med AI?
  - Var i flödet skulle det ge mest effekt?
  - Vilka typer av avvikelser eller felmönster skulle vara viktigast att hitta?
- Vilka risker ser du med AI implementering?
  - Felaktiga avslag
  - Bristande spårbarhet och motivering
  - Fel ansvarsfördelning mellan roller

Mätning och KPIer

- Vad mäter ni idag för att följa fakturagranskningen?
- Vad skulle ni vilja mäta men inte kan idag?
  - Träffsäkerhet i granskningen
  - Riskexponering kopplad till gråzoner
  - Tid som läggs på att leta verifikat jämfört med faktisk bedömning

DEPARTMENT OF TECHNOLOGY MANAGEMENT AND ECONOMICS
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se