Aalto-yliopisto
Aalto-universitetet
Aalto University

Master’s Programme in Economics & Business Administration

AI Workforce of the Future

Enhancing Qualitative Organizational Data Analysis through Integrated GraphRAG and Multi-Agent AI Systems with comparison to Baseline AI

Ali Amaan

Master’s Thesis
2025

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Ali Amaan

Title of Thesis AI Workforce of the Future

Programme Master's Programme in Economics & Business Administration

Major International Design Business Management **Code** BIZ223

Thesis Supervisor Prof. Henri Schildt

Collaborative Partner Department of Management Studies at Aalto University

Date 31.12.2025 **Number of Pages** 90 **Language** English

Abstract

The exponential growth of unstructured data presents a critical paradox for organizations and researchers: while the volume of rich qualitative information, such as interview transcripts and internal reports, is expanding, the capacity to analyse it using rigorous, interpretive methodologies remains constrained by human cognitive limits. Traditional qualitative data analysis (QDA) is inherently unscalable, while standard Artificial Intelligence (AI) solutions, including Retrieval-Augmented Generation (RAG), often fail to capture the structural nuance and global context required for deep inquiry. This thesis addresses this gap through a Design Science Research Method (DSRM) study. The research designs, implements, and evaluates a novel artifact: a web-based multi-agent system empowered by Graph Retrieval-Augmented Generation (GraphRAG). This system structures unstructured text into knowledge graphs to support complex reasoning and employs a staged autonomy model to balance AI efficiency with human interpretive control. Evaluated through quantitative analytics and semi-structured interviews with organizational practitioners, the study demonstrates that while higher levels of agentic autonomy increase analytical depth, they introduce significant coordination costs. The findings suggest that "Human-in-the-Loop" (HITL) mechanisms are not merely safeguards but essential components for organizational trust. This research contributes to the intelligent agentic systems of the future by offering design principles for AI systems that scale qualitative and quantitative data analysis.

Keywords Multi-Agent Systems (MASs), Knowledge Graph, Data Analysis, Qualitative Data Analysis (QDA), Organizational Research, AI Agentic Systems, AI Agents, Design Thinking, Iterative Design, Graph Retrieval-Augmented Generation (GraphRAG).

Table of Contents

List of Tables

List of Figures

Preface & Acknowledgements

This master's thesis was conducted as a part of the Master's Degree Programme in IDBM at Aalto University. I want to thank my supervisor, Professor Henri Schildt, for his tremendous trust, support, direction, and research independance throughout the process.

This thesis is dedicated to Islam, my way of life, which has always pushed me to strive for the best.

Ali Amaan
31st December 2025
Otaniemi, Espoo, Finland

Symbols and Abbreviations

AI	Artificial Intelligence
GraphRAG	Graph-based Retrieval Augmented Generation
MAS	Multi-Agent System
RAG	Retrieval Augmented Generation
ML	Machine Learning
LLM	Large Language Model
DB	Database
DAS	Distributed Agent Systems
API	Application Programming Interface
KG	Knowledge Graph
MCP	Model Context Protocol
HITL	Human in the Loop
AGI	Artificial General Intelligence
SOTA	State of the Art
QDA	Qualitative Data Analysis
NLP	Natural Language Processing
DSRM	Design Science Research Method
UI	User Interface
UX	User Experience

1 Introduction

1.1 Background

The digital era has ushered in a period of unprecedented information expansion, fundamentally altering the landscape of organizational knowledge management and academic research. We are currently witnessing a "data deluge" of such magnitude that it threatens to overwhelm traditional analytical capacities. According to a report by IDC in 2023, the global datasphere is projected to swell to 175 zettabytes by 2025, a nearly five-fold increase from the 33 zettabytes recorded in 2018. This exponential growth is not merely a function of increased storage capacity but a reflection of the pervasive digitization of human interactions, data pipelines, and mechanical operations. (Heeg, 2023)

Crucially, the composition of this data presents a specific challenge. It is estimated that between 80% and 90% of all new enterprise data generated today is unstructured (Heeg, 2023). This category includes a diverse array of formats that defy the rigid logic of relational databases: narrative text files, internal communication logs (chat, email), audio recordings from customer service interactions, textual customer reviews, video surveillance footage, and the continuous stream of sensor data from the Internet of Things (IoT). Unlike structured data, which is in the form of clean rows and columns, unstructured data is messy, context-dependent, and semantic-rich.

The high rate of digitization of organization processes has created masses of qualitative and quantitative data never present before, whether in customer service processes or internal documents and multilingual contexts (Cuevas et al., 2024). Although some of these traditional forms of qualitative data analysis have allowed researchers to formulate knowledge bases regarding organizations in terms of processes, business opportunities, and further insights, as well as decision making processes. However, those methods are not very useful in scaling alongside today data volumes and complexity (Herhausen et al., 2025). At the same time, generative artificial intelligence developments can offer both opportunities and challenges.

The emergence of Large Language Models (LLMs) has offered a potential bridge for this gap. However, standard implementation approaches, such as naive Retrieval-Augmented Generation (RAG), have proven insufficient for high-stakes qualitative inquiry. While effective for simple fact retrieval, naive RAG systems often suffer from the "lost in the middle" phenomenon and lack a structural understanding of how themes are interconnected across large corpora. They treat text as isolated fragments rather than a cohesive web of meaning, failing to support the "global" reasoning required to answer complex research questions. (Han et al., 2025)

The novel approach to these challenges discussed in this thesis is the combination of the multi-agent AI systems with hybrid retrieval of vector search (traditional RAG) and graph retrieval augmented generation (GraphRAG) and in an agentic manner (OpenAI, 2025a; Sarmah et al., 2024). The capacity of GraphRAG to model relational knowledge by indexing and retrieving it using graphs supplements the concept of MASs that solve problems based on specific roles of the particular agents (Gusarov et al., 2025; Han et al., 2024; Hu et al., 2024). Although both technologies have shown both promise on their own their potential remains untapped about qualitative organizational settings (Han et al., 2025). Simultaneously, multi-agent architectures allow for the decomposition of analytical tasks (such as coding, searching, and executing) into specialized agent workflows. This thesis posits that combining these technologies within a controlled environment can demonstrate a comparative experience of baseline to a functional MAS model of the AI workforce of the future.

1.2 Research Objectives

The primary problem identified in this study is the analytical gap: the lack of methodological tools that can leverage the computational power of state-of-the-art generative agentic AI techniques to analyse large-scale unstructured datasets without sacrificing the interpretive control and contextual validity required by organizational and academic standards. Accordingly, the objective of this thesis is to design, implement, and evaluate an multi-stage agentic AI system augmented with agentic GraphRAG system tailored for qualitative data analysis and comparative study of different stages of autonomy of agentic AI systems. Following the Design Science Research Methodology (DSRM), this study moves beyond theoretical speculation to construct a functional artifact, in the form of a web-based platform, and report its analytical stats. The entire research methodology will be accompanied by qualitative insights from industry professionals that regularly interface with AI.

The main research question in this investigation is as follows: How is the state of depth, usability, and methodological rigour of data analysis in organisational settings especially with current AI system and what is the outlook on future AI systems?

This question branches into four subordinate inquiries:

- **RQ1:** What architectural requirements enable effective knowledge graph traversal alongside distributed agentic flows for organizational tasks?

- **RQ2:** How do organizations perceive agentic AI based systems and how proactively are they adopting latest AI technologies?

- **RQ3:** How do organizations or roles measure success of agentic AI systems?

- **RQ4:** What practical constraints emerge when deploying such systems in real organizational settings, particularly regarding data, security, accuracy, and other concerns?

1.3 Scope & Limitations

The **scope** of the study is confined to opensource dataset analysis within the frames of publicly available organizationally relevant data. The most significant conceptual limits are the usage of the current graph database systems and LLM models as opposed to the development of new models. This research is focused at the intersection of Human-Computer Interaction and Artificial Intelligence. The scope of the artifact implementation is limited to the analysis of textual dataset (e.g., interview transcripts, industry reports, etc. in certain file type formats). While the broader category of unstructured data includes audio and video, this study focuses on text to isolate the effectiveness of graph-based reasoning and agentic workflows. The empirical evaluation is conducted within an organizational context, involving practitioners who actively manage or analyse data or have regular interface with AI systems for their daily tasks or decision-making purposes.

Despite definition of the scope, **limitations** must be acknowledged. First, the study employs an exploratory Design Science approach. The artifact is a research prototype intended to demonstrate feasibility and elicit feedback, not a production-ready enterprise software solution. Second, the evaluation sample size is small with seven participants. While consistent with the qualitative nature of DSR evaluation intended to identify design utility, the findings regarding organizational adoption are indicative rather than statistically generalizable in the context of agentic AI systems. Third, there is an inherent epistemological limitation. The system is designed to augment human intelligence, but it cannot replicate the lived experience or cultural intuition of a human researcher. The "truth" of the analysis remains an interpretive construction of the user; the AI serves only to create the structural scaffolding for this interpretation. Finally, the system relies on existing Foundation Models (e.g., GPT-5). Consequently, it inherits the known biases of these models regarding language, prompts, and cultural context, which may affect the nuance of thematic extraction in other contexts or arrangements.

1.4 Structure of the Thesis

Structurally, this thesis progresses from establishing theoretical foundations to empirical validation with incorporation of qualitative study results.

Following this introduction, Chapter 2 synthesizes literature on agentic AI, multi-agent collaboration mechanisms, GraphRAG, and AI's role in data analysis and qualitative research. Chapter 3 details the mixed-methods methodology, including prototype development stages and evaluation frameworks. Chapter 4 presents findings from technical assessments and user studies, while Chapter 5 discusses conclusive implications for organizational research practices and AI system design. The following discussion section also outlines potential practical applications and future research directions in AI-augmented methodologies in other contexts as well following similar principles.

2 Literature Review

2.1 Research Trends & Background

The high pace of technological development in the sphere of artificial intelligence has opened unparalleled opportunities in the process of improving qualitative research methodologies, especially in the organizational setting. Other structures are becoming more popular in Graph Retrieval-Augmented Generation (GraphRAG), multi-agent AI systems, have shown strong potential to solve complex analytical problems that demand the provision of both structural knowledge description and multi-agent abilities of solving problems(Han et al., 2025; Peng et al., 2026) . Nonetheless, the cross-section of these technologies has not been studied extensively in the context of organizational research and thus there is a severe gap in the existing literature and practice.

This literature review explores the prevailing situation of research in three main areas including GraphRAG technologies, multi-agent AI systems and the qualitative method of research within organizations. The systematic analysis of recent developments and rising trends presented in this review shows that there are substantial opportunities of integration and methodological issues that revolve around transparency, evaluation and actual implementation of an organization are possible (Abdul Rahman et al., 2025; Brailas, 2025).

2.2 Multi-Agent AI Systems (MASs)

2.2.1 Conceptual Foundations & Definitions

Multi-Agent Systems (MAS) in the context of agentic AI is a model of distributed artificial intelligence that characterizes a group of autonomous agents working in a common environment to address complex problems that would not be possible by single agents in terms of rigorous thinking and output alignment (Goonatilleke and Hettige, 2022; Mohan Singh, 2025). Fundamentally speaking, the multi-agent system is a framework of orchestrating team environments at the agentic AI abstraction level. This includes autonomous AI agents capable of perceiving their environments, interpreting information, and acting in ways that help them meet user given task objections and subsequently tasks delegated by other AI agents (Herrera et al., 2020). The basic assumption of the MAS is the thought of emergent intelligence: the belief that the complex behaviours and solutions should be parts of the self-interaction of the simpler autonomous components.

MAS have theoretical underpinnings that heavily rely on distributed artificial intelligence (DAI), game theory and organizational theory (Goonatilleke and Hettige, 2022). At the heart of appreciation of MAS lies agent autonomy as an ability of an agent to act without direct supervision by humans or other agents. This autonomy can be denoted as being characterized by four properties: reactivity (able to perceive and respond to the changes in the environment), proactivity (agent centred behaviour), social ability (interaction with other agents), and the capability of learning (adaptation according to experience) (Binyamin and Ben Slama, 2022).

The other key conceptual pillar of the MASs is coordination mechanisms. These processes inform agents to coordinate, exchange in-formation and cooperate to achieve a set of goals and cope with resource scarcity and potential disputes. Basic to the coordination mechanisms in MASs is the fact that they are quite different than the centralized systems in that they have to accommodate the distributed nature of decision-making and the fact that there is a potential of misunderstood goals. (Y. Liu et al., 2025)

2.2.2 Evolution & Historical Development

MASs have existed (or may exist) in a variety of different phases, each possessing a unique theoretical focus and technological potential. Precursors The first conceptual bases date to the 1970s when distributed problem-solving systems were developed, and it was realized that distributed solutions might be useful in solving complex computational problems. The origin era (1980s-1990s) was characterized by the development of the main theoretical bases, such as the Contract Net Protocol, which allocates tasks and the creation of the agent architecture, i.e. the Belief-Desire-Intention (BDI) model. It was at around this time that researchers were more interested in coordination protocols and fruitful communication languages, and came up with standards like the Knowledge Query and Manipulation Language (KQML), and later still, the Foundation for Intelligent Physical Agents (FIPA) standards. (Goonatilleke and Hettige, 2022)

The period of expansion (1990s-2000s) was the period of tremendous growth in the field of theoretical sophistication as well as in practical use. It was the time of the de-development of more advanced agent architectures, combination of learning functions and market-based agency. Investigations as the phase progressed were also more related to problems of scalability and multi-agent system methodology engineering.

When paired with machine learning and LLMs, the modern (2010s-present) era has seen the implementation of multi-agent models based on them (Lin et al., 2025). During this time there has been the development of LLM-based MASs where LLMs act as the cognitive system of individual agents, with higher capabilities of both natural language communication and reasoning. LLM integration has essentially changed the environment of MAS; now agents are able to carry out more complicated collaborative activities utilizing natural language. (Mitchell, 2021)

2.2.3 Architectural Structures & Architectural Patterns

There are various architectural patterns that are used by the modern MASs depending on the types of problems and operational needs. The hierarchical architecture is one of the most common design patterns which structures the agents into various layers with various modes of abstraction and authority. In hierarchical systems, strategic planning and coordination are usually involved in the high-level, and the tactical execution with the environmental interaction are implemented in the low-level agents.

The other technique is the federated architecture which is based on peer-to-peer coordination of autonomous agents. Such systems are especially appropriate to the context in which one agent has no overall information or authority and hybrid decision-making and negotiation is necessary. The federated approach has been particularly useful in cross-organizational use cases where two or more autonomous organizations are required to work together and at the same time have independence. Figure 1 shows an example of this.

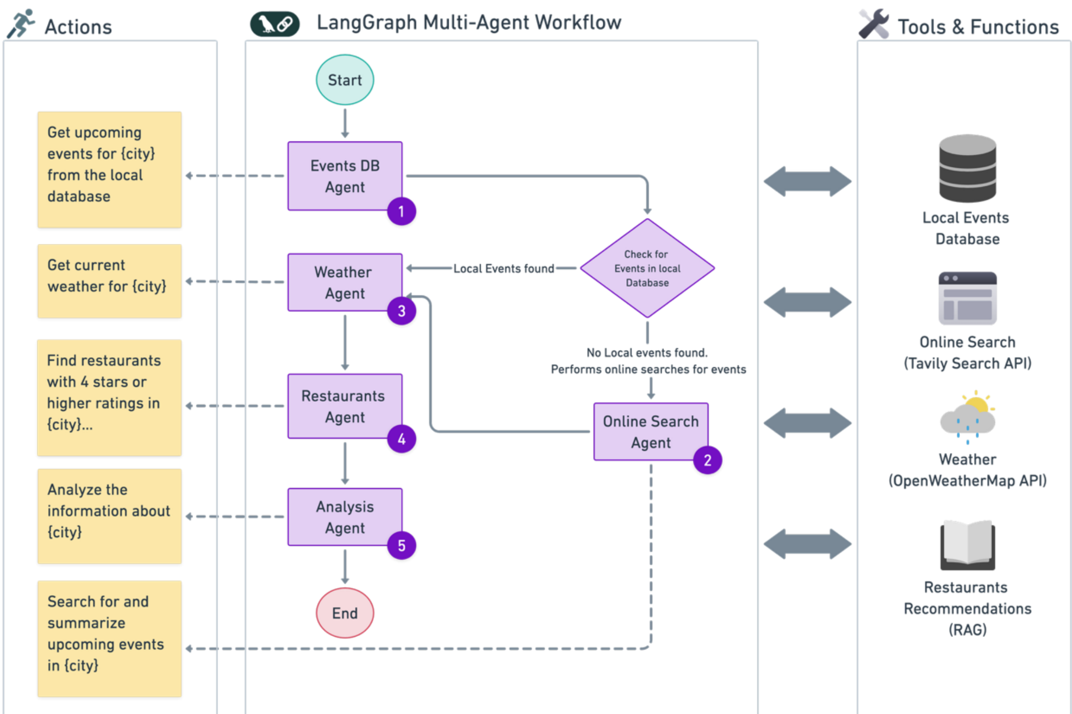

Figure 1 Multi-Agent Workflow High-Level Example (Boaventura, 2025)

Team-based organizational models are the development of the multi-agent architecture design in the realm of major improvement. Such models are based particularly on human organization structure, which applies the concept of role specialization, mobilization and formation of dynamic groups, and adaptive hierarchies. It has been proved that team-based strategies could dramatically enhance the performance of the complex, dynamic environments as the quick adaptation and effective use of resources could be achieved.

The most recent architecture development is the use of LLM-based architectures in MASs. Neither of these structures involves the use of reasoning or communication abilities of LLMs, but with their help these structures can support agents in complex cooperative activities. An example of these innovations is the ReAct planning model and Mixture of Agents architecture that proves the significant enhancements of role allocation and decision-making process. (Wang et al., 2024; Wei et al., 2025; Yao et al., 2022)

2.2.4 Coordination, Communication Mechanisms.

Successful performance of MASs is rooted in the sound coordination and communication schemes that can allow agents to coordinate their actions and exchange information. Communication protocols are used to describe agent-to-agent communication, and they specify the syntax, semantics and pragmatics of message communication. The FIPA Agent Communication Language (ACL) has become a common standard offering a platform of expressive and standardised agent communicative statement.

Negotiation systems can be a complex form of coordination language enabling the agents to achieve an agreement in situations characterized by conflicting interests or resource limitations. The game-theoretic methods have been especially useful in this regard, offering mathematical models to study and maximize the results of the negotiations. Multi-issue negotiation schemes enable agents to do various forms of bargaining that incorporate various dimensions of assent.

Economic policing mechanisms are market-based coordination mechanism that involve use of economic concepts to produce distributed coordination, and this is not centralized. This is in contrast to the methods that coordination is an economic challenge with the agents purchasing and selling services or resource in an auction or other market structure. Studies have shown that in most situations market-based methods can successfully lead to almost optimal resource allocation, and agent autonomy.

Another significant category of coordination mechanisms is associated with consensus algorithms in situations when agents need to reach some shared states or decisions. These algorithms are aimed at addressing the problem of distributed decision making under the conditions of information delays and partiality, and even agent failure.

2.2.5 Learning and Adaptation in Multi-Agent Environments

MASs have been identified as highly effective in learning and adapting as features. Multi-agent reinforcement learning (MARL) has become one of the new paradigms where agents can enhance their performance by interacting with the environment and other agents. Uniqueness of MARL systems versus single-agent learning MARL systems have distinctive issues, such as non-stationarity (where the environment is changing as the other agents learn), partial observability, and multi-agent coordination of learning.

Evolutionary methods of learning in MASs are inspired by biological evolution to help the optimization of behaviours of agents and configurations of systems. Such approaches are also especially useful in the case of problems with multiple objectives or complex fitness landscapes in which traditional methods of optimization prove difficult. Evolutionary algorithms allow the agents to dynamically adjust their strategies based on the dynamics in the strategies of other agents, which results in ever more advanced behaviours.

Social learning systems allow the agent to gain experiences and knowledge of other agents within the system. The mechanisms have the capability to dramatically speed up the process of learning because they enable agents to exploit the experience of the system as a whole and not just based on their own experience. Social learning strategies entail different knowledge sharing techniques, learning through imitation, and collaborative filtering.

2.2.6 Existing Applications and Case Studies

MASs have been used effectively in a myriad of fields with success before proving their relevance and usefulness. Financial applications are one of the most developed segments of applications and MASs are deployed to provide algorithmic trading, risk management and portfolio optimization (Xiao et al., 2024). Additionally, Yu et al. (2024) also demonstrate a FinCon framework which incorporates a manager-analyst communication hierarchy based on the organizational structure of investment firms to attain enhanced performance in the financial decision-making duties.

The use of scientific research has been of great interest especially with the use of automated research assistants. A exploration of it has also been done by Ampel and Ullman (2025) for information systems research. The DORA AI Scientist system illustrates the potential of hierarchical teams of AI agents to scientific exploration, using hierarchical teams of AI agents to conduct literature review, hypothesis generation and research report generation (Naumov et al., 2025). These types of systems use special agents in the various stages of the research process, including data collection, analysis and synthesis.

The practical usefulness of MASs to complex operations settings is evident in academic, industrial, and manufacturing applications. Such applications will usually be in the area of coordination of autonomous systems, supply chain management, and production optimization. Studies and experiments have shown that multi-agent solutions may lead to high operational efficiency gains with reported up to 20-40% gains in enterprise efficiency.(Binyamin and Ben Slama, 2022; Chen et al., 2025; González-Cancelas et al., 2025; Luo et al., 2025; Martinez-Gil and Gil-Magraner, 2025; Papageorgiou et al., 2025)

One of the new application domains for MASs is the collaborative software development, whereby, the MASs are being implemented to automate the different activities in the software development cycle (Liu et al., 2024). These systems make use of special agents to analyse the requirements, generate the code, test and deploy the code, and show huge enhancement in the productivity of the development process and the quality of the code.

2.3 Graph Retrieval-Augmented Generation (GraphRAG)

2.3.1 Conceptual Foundations & Theoretical Framework

First introduced by Edge et al. (2024), GraphRAG is an important extension of the information-retrieval and natural language generation problem, to overcome essential drawbacks of the earlier Retrieval-Augmented Generation (RAG) systems (Han et al., 2025). GraphRAG is based on the structural attributes of graphical knowledge representations that will allow more intricate retrieval of information and contextual reasoning. In contrast to RAG methods that use a document as an isolated element in a vector space, GraphRAG directly represents the degree of interrelations between entities, concepts and documents in the form of a graph (Xiang et al., 2025).

GraphRAG is on a theoretical basis of graph theory and knowledge representation. Graphs form an intuitive structure in which the information can be encoded through relational data on complex relations: nodes are the entities of the concepts and the relationship represented by an edge. This architecture allows flexibility of GraphRAG systems to make multi-hop reasoning through a variety of relationships they traverse to extract total context of query responding.

The other essential theoretical component of GraphRAG systems is the knowledge graph embedding techniques. The methods allow the conversion of discrete graph structures into continuous vectors representations without destruction of structural and semantic properties. Its embedding model allows one to compute similarities effectively and permits to combine neural models with graph-based retrieval models.(Ferrari et al., 2022; Yang et al., 2025)

GraphRAG frameworks have the query processor component that acts as an interface between natural language query and graph-based operations of retrieval. The advanced query processing methods use natural language understanding to find the relating entities, relationships, and query patterns to facilitate the process of graph traversal. This element plays a major role in the process of filling the semantic gap between the information requirements of the user and graph-based knowledge representation.

2.3.2 Design of Architectural Components and System

GraphRAG are commonly made of five major architectural blocks namely query processor, retriever, organizer, generator, and data source (Peng et al., 2026). The retriever component uses advanced graph traversal algorithms to search the knowledge graph to locate subgraphs of information that are of interest. However, in contrast to the conventional methods of retrieval by vectors, the graph-based retrieval is capable of representing the multi-hop connections and situational dependencies necessary to perform delicate reasoning processes. A typical GraphRAG implementation architecture is displayed in figure 2.

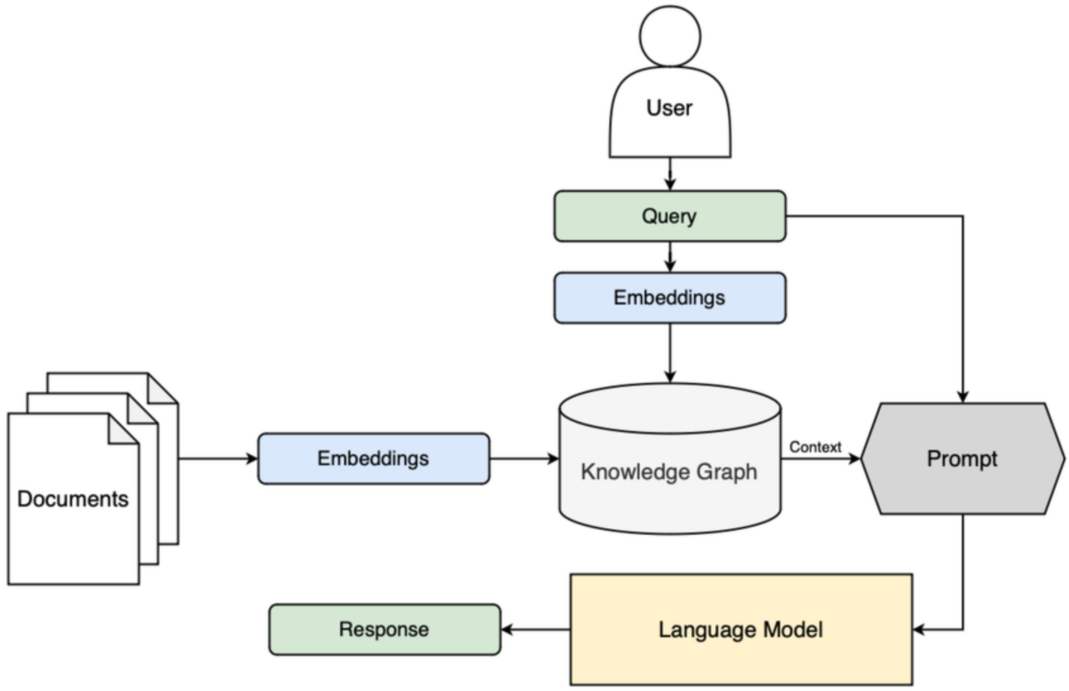

Figure 2 Typical GraphRAG Architecture (Roll et al., 2025)

The organizer component is very important to incorporate retrieved information among the various graph nodes and edges into coherent contextual representations. This component has to deal with the issue of information integration of the heterogeneous graph representations but remain relevant and coherent. State-of-the-art implementations of the advanced organizer take advantage of the retrieved information by utilizing attention mechanisms and graph neural networks to weigh and integrate the retrieved items.

The final stage of the GraphRAG pipeline is graph-augmented generation in which retrieved context represented as a graph is combined with LLMs to generate context-relevant and correct responses. Integration is a special problem with regards to traditional RAG since graph-based context is usually structured relational information that should be effectively merged into the generation process.

The data source component includes the knowledge graphs and their methodology of construction. GraphRAG systems today are frequently based on automated knowledge graph construction methods that rely on textual corpora depicting LLMs to identify entities and relationships between them. This automated strategy has greatly decreased the expenditure and complexity of setting up GraphRAG systems and has maintained high quality knowledge representation. (Dehal et al., 2025; Peng et al., 2026; Yang et al., 2025)

2.3.3 Construction of Graphs & Extraction of Knowledge

The quality and coverage of the underlying knowledge graphs is very crucial to the efficacy of GraphRAG systems. Knowledge graph construction Automated knowledge graph construction has become a major enabler of scalable GraphRAG deployment. The use of LLMs in entity extraction, relationship identification, and graph population has become popular using textual input sources.

The automated graph construction is established on triple extraction methods, which can find subject-predicate-object relations in natural language passages. More advanced methods of extraction use specialized language models which are trained on information extraction missions and obtain high precision and recall in entity and relationship recognition. A more extreme example of text extraction is the GraphRAG system which applies highly elaborate extraction methods by using specific medical vocabularies and ontologies to guarantee clinical precision.

Another important issue in knowledge graph construction is graph schema design. Intense schemas have to be trade-offs between expressiveness and computational efficiency, proving to be detailed enough to allow complex reasoning and at the same time maintain manageable graph structures. As an example, Abu-Salih and Alotaibi (2024) in figure 3 demonstrate a well schemed knowledge graph in the education context as part of their research. Schema design decisions can be motivated by domain specific concerns, so there may be specialized vocabularies and types of relationships associated with specific application realms. (Hogan et al., 2021)

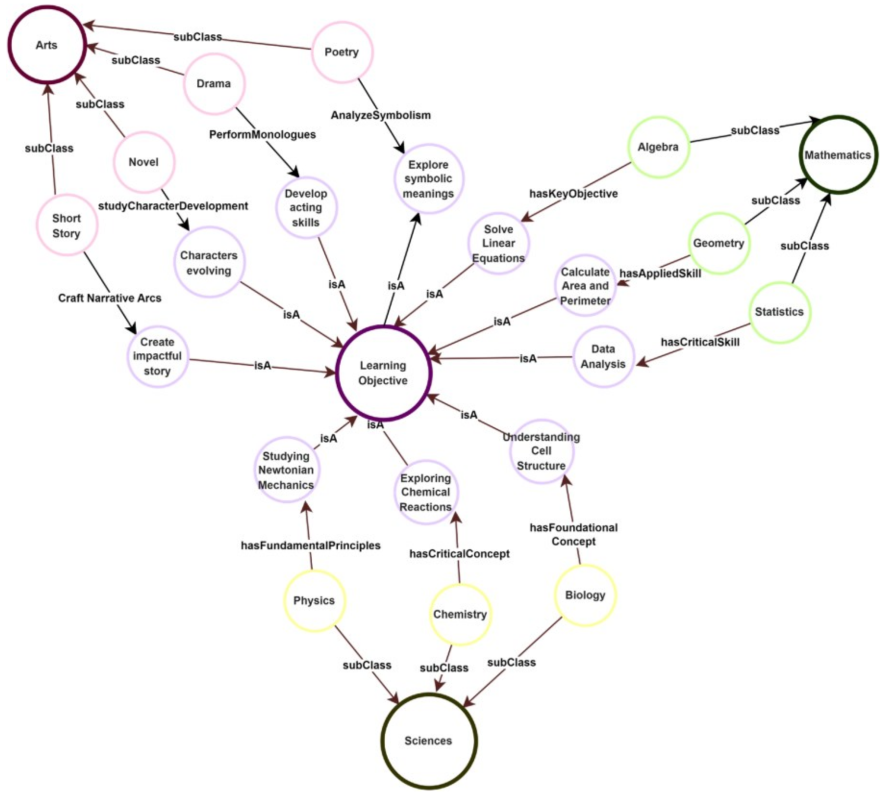

Figure 3 Education Curriculum Graph Representation Example (Abu-Salih and Alotaibi, 2024)

Constructed knowledge graphs need to have quality assurance and checking mechanisms to ascertain their reliability. Such mechanisms involve consistency checking, completeness testing and checking accuracy based on authoritative sources. Automated validation algorithms use statistical algorithms and rule systems to detect errors in the constructed graphs and correct them.

2.3.4 Retrieval Algorithms and Graph Traversal Strategies

GraphRAG systems use advanced graph traversal algorithms adding dynamically chosen retrieval strategies on query features, and the structure of the graph. Breadth-first search (BFS) and depth-first search (DFS) offer basic traversal designs, although more sophisticated designs are used in advanced systems, which include random walks, personalized PageRank, and graph attention. (Ismaeel and Zebari, 2025)

Query-adaptive traversal is another important addition in GraphRAG retrieval strategy. These methods examine query semantics to compute superior traversal designs, and modify the exploration approach because of the kind of information needed. Adaptive graph traversal as illustrated in the PolyG framework has substantial improvements in effectiveness and efficiency (R. Liu et al., 2025).

GraphRAG systems combine multi-hop reasoning algorithms that allow the system to collect information across a series of relationships in a graph. Such algorithms are required to be very careful in the extent of exploration employed versus the cost of the computations, because the size of the set of possible paths increases exponentially with the number of hops. The sophisticated methods use beam search and pruning techniques to handle this complexity at the expense of quality of retrieval.

Semantic filtering processes play a role in the process of ensuring that the information retrieved is relevant to the initially asked query despite multi-hop viewing. These processes take advantage of semantic similarity scale and relevancy rating to determine the most relevant subgraphs of information. The Dynamic Similarity-Aware BFS (DSA-BFS) method can be an example of advanced filtering algorithms that avoid losing relevance and can be used to effectively explore the area. (Prakash Jadhav, 2022)

2.3.5 Integration with LLMs

The integration of GraphRAG with LLMs presents both significant opportunities and substantial technical challenges. Context injection strategies determine how graph-based retrieved information is presented to language models for generation. Simple approaches concatenate retrieved information with the original query, while more sophisticated methods employ structured prompting and template-based formatting to preserve graph structure information. (Xiang et al., 2025)

Graph-aware attention mechanisms represent an advanced integration approach that enables language models to directly process graph-structured information. These mechanisms extend traditional transformer attention to operate over graph structures, allowing the model to selectively attend to relevant graph nodes and edges during generation. Such approaches require specialized model architectures but can achieve superior performance on graph-based reasoning tasks. (Q. Zhang et al., 2025)

Fine-tuning strategies for GraphRAG involve adapting language models to better process graph-based contextual information. The GraphRAFT approach by Clemedtson and Shi (2025) demonstrates how retrieval-augmented fine-tuning can improve model performance on knowledge graph question answering tasks. These techniques require careful balancing of general language capabilities with specialized graph reasoning abilities.

Hallucination mitigation represents a critical challenge in GraphRAG systems, as the complexity of graph-based context can sometimes lead to inaccurate or inconsistent generation. Advanced systems employ fact verification mechanisms and consistency checking to identify and prevent hallucinated content. The GraphRAG-FI framework demonstrates sophisticated filtering and integration strategies that reduce over-reliance on retrieved information while maintaining accuracy (Guo et al., 2025).

2.3.6 Domain-Specific Applications and Case Studies

GraphRAG has demonstrated significant value across diverse application domains, each presenting unique challenges and opportunities. Medical applications represent one of the most mature and successful domains for GraphRAG deployment. The Medical GraphRAG system showcases sophisticated medical knowledge graph construction and retrieval, achieving superior performance on medical question answering benchmarks while ensuring clinical accuracy. Medical GraphRAG systems must address unique challenges including medical terminology complexity, evidence-based reasoning, and safety-critical accuracy requirements. (Wu et al., 2024)

Legal applications demonstrate the potential of GraphRAG for complex document analysis and legal reasoning as well as complex legal ontologies. Legal GraphRAG systems must navigate intricate regulatory frameworks, case law relationships, and statutory hierarchies. These applications require sophisticated entity extraction techniques capable of identifying legal concepts, precedents, and regulatory relationships. The complexity of legal language and reasoning presents ongoing challenges for automated systems. (de Martim, 2025)

Logistics and industrial applications showcase GraphRAG's effectiveness in technical knowledge management and complex networks analysis (Naganawa et al., 2025). Document GraphRAG systems have been deployed for manufacturing quality assurance, demonstrating consistent performance improvements over traditional RAG approaches (Knollmeyer et al., 2025). These applications often involve complex technical documentation, equipment relationships, and procedural knowledge that benefit from graph-based representation.

2.3.7 Performance Evaluation and Benchmarking

Evaluating GraphRAG systems requires sophisticated benchmarking approaches that assess both retrieval quality and generation accuracy. The GraphRAG-Bench benchmark developed by Xiao et al. (2025) represents a comprehensive evaluation framework specifically designed for graph-based retrieval-augmented generation. This benchmark emphasizes multi-hop reasoning capabilities, domain-specific knowledge requirements, and logical coherence assessment.

Retrieval evaluation metrics for GraphRAG systems extend traditional information retrieval measures to account for graph structure and multi-hop reasoning. Context relevance, information completeness, and structural coherence represent key evaluation dimensions that distinguish GraphRAG assessment from traditional RAG evaluation. Advanced evaluation frameworks employ expert annotation and ground truth construction to ensure reliable performance assessment. (Tang et al., 2024)

Generation quality assessment in GraphRAG systems requires evaluation of factual accuracy, coherence, and relevance in graph-enhanced responses. Traditional generation metrics such as BLEU and ROUGE may not adequately capture the unique aspects of graph-based generation. Specialized metrics focusing on entity accuracy, relationship consistency, and multi-hop reasoning correctness are increasingly employed.

Comparative analysis studies have consistently demonstrated GraphRAG's advantages over traditional RAG approaches, particularly for complex reasoning tasks. Studies report improvements ranging from 10-15% on challenging multi-hop question answering datasets. The benefits are most pronounced for queries requiring cross-document reasoning, temporal analysis, and complex relationship traversal. (Knollmeyer et al., 2025; Peng et al., 2026; Xiao et al., 2025)

2.4 Qualitative Research & Analysis using AI

2.4.1 Traditional Qualitative Research Methodologies

Qualitative research methodologies have long served as cornerstone approaches for understanding complex organizational phenomena, providing rich insights into human experiences, behaviours, and social processes within organizational contexts. Grounded theory methodology represents one of the most influential qualitative approaches in organizational research, offering systematic procedures for developing theory from empirical data. Originally developed by Glaser and Strauss in 1967, grounded theory has evolved into multiple variants including Glaserian, Straussian, and constructivist approaches, each offering different epistemological orientations and analytical procedures. (Lim, 2025)

Phenomenological research methodologies provide another fundamental approach to organizational qualitative research, focusing on understanding lived experiences and the meanings individuals attribute to their organizational experiences. Phenomenological approaches encompass multiple traditions including Husserl Ian descriptive phenomenology, Heideggerian interpretive phenomenology, and interpretive phenomenological analysis (IPA). These methodologies are particularly valuable for exploring complex organizational phenomena such as leadership experiences, change processes, and cultural dynamics. (Lim, 2025)

Ethnographic methodologies enable researchers to gain deep understanding of organizational cultures, practices, and social processes through immersive observation and participation. Organizational ethnography has proven particularly valuable for understanding complex work practices, informal organizational structures, and the cultural dynamics that shape organizational behaviour. These approaches typically involve extended fieldwork and participant observation to develop comprehensive understanding of organizational phenomena. (Lim, 2025)

Case study methodologies provide frameworks for in-depth investigation of specific organizational phenomena within their real-world contexts. Case studies are particularly valuable for understanding complex organizational processes, strategic decisions, and change initiatives where contextual factors play crucial roles. Multiple case study designs enable researchers to identify patterns and develop theoretical insights across different organizational contexts. (Lim, 2025)

2.4.2 The Integration of AI in Qualitative Research

The integration of artificial intelligence technologies into qualitative research represents a paradigmatic shift that is fundamentally transforming how researchers approach data collection, analysis, and interpretation. Natural Language Processing (NLP) integration has emerged as one of the most significant developments, enabling researchers to process vast quantities of textual data while maintaining the interpretive depth characteristic of qualitative inquiry. Recent studies demonstrate that NLP-assisted qualitative analysis can reduce project time by 120+ hours while maintaining analytical rigor and interpretive quality (Abram et al., 2020).

LLM applications in qualitative research have shown remarkable potential for enhancing various aspects of the research process. ChatGPT and similar models have been successfully employed for initial coding, thematic analysis, and theory development. Research demonstrates that LLMs can effectively support grounded theory analysis, with some studies showing perfect concordance between AI-assisted and traditional analytical approaches. However, these applications require careful consideration of the human-AI collaboration dynamic to ensure that technological augmentation enhances rather than replaces human interpretive expertise. (Kabir et al., 2025; Kalanda and Cheboi, 2025)

Multi-modal AI applications are expanding the possibilities for qualitative research by enabling analysis of diverse data types including text, audio, video, and visual materials. These capabilities are particularly valuable for organizational research that often involves multiple data sources and complex contextual information. The integration of multi-modal AI enables researchers to conduct more comprehensive analysis while maintaining the holistic perspective essential to qualitative inquiry. (Xu et al., 2025)

Automated coding and thematic analysis represent some of the most mature AI applications in qualitative research. The Automated Qualitative Assistant (AQUA) platform demonstrates how AI/ML can effectively support qualitative coding while maintaining the interpretive control essential to qualitative research. (Lennon et al., 2021)

2.4.3 AI-Enhanced Data Collection and Processing

AI technologies are transforming qualitative data collection processes, enabling researchers to gather and process data more efficiently while maintaining the depth and nuance characteristic of qualitative inquiry. AI-powered interviewing systems demonstrate significant potential for enhancing data collection quality and efficiency. Research comparing AI-powered chatbot interviews with traditional online surveys shows that AI-driven approaches generate significantly higher participant engagement and response quality as measured by informativeness, relevance, specificity, and clarity. (H. Zhang et al., 2025)

Real-time transcription and processing capabilities enable researchers to focus more attention on interview dynamics and participant responses rather than technical data capture concerns. Advanced transcription systems incorporating speaker identification and emotion recognition provide richer data for subsequent analysis. These capabilities are particularly valuable for organizational research involving multiple stakeholders or complex group dynamics.

Sentiment analysis and emotional intelligence applications enable researchers to identify emotional patterns and affective dimensions in qualitative data that might be missed in traditional analysis. These capabilities are particularly valuable for organizational research focusing on employee experiences, customer satisfaction, or change management processes. Advanced systems can identify subtle emotional patterns across large datasets while preserving the interpretive flexibility essential to qualitative research. Figure 4 explains a very detailed spectrum of agent autonomy and QDA tasks quite aptly. (Kalanda and Cheboi, 2025)

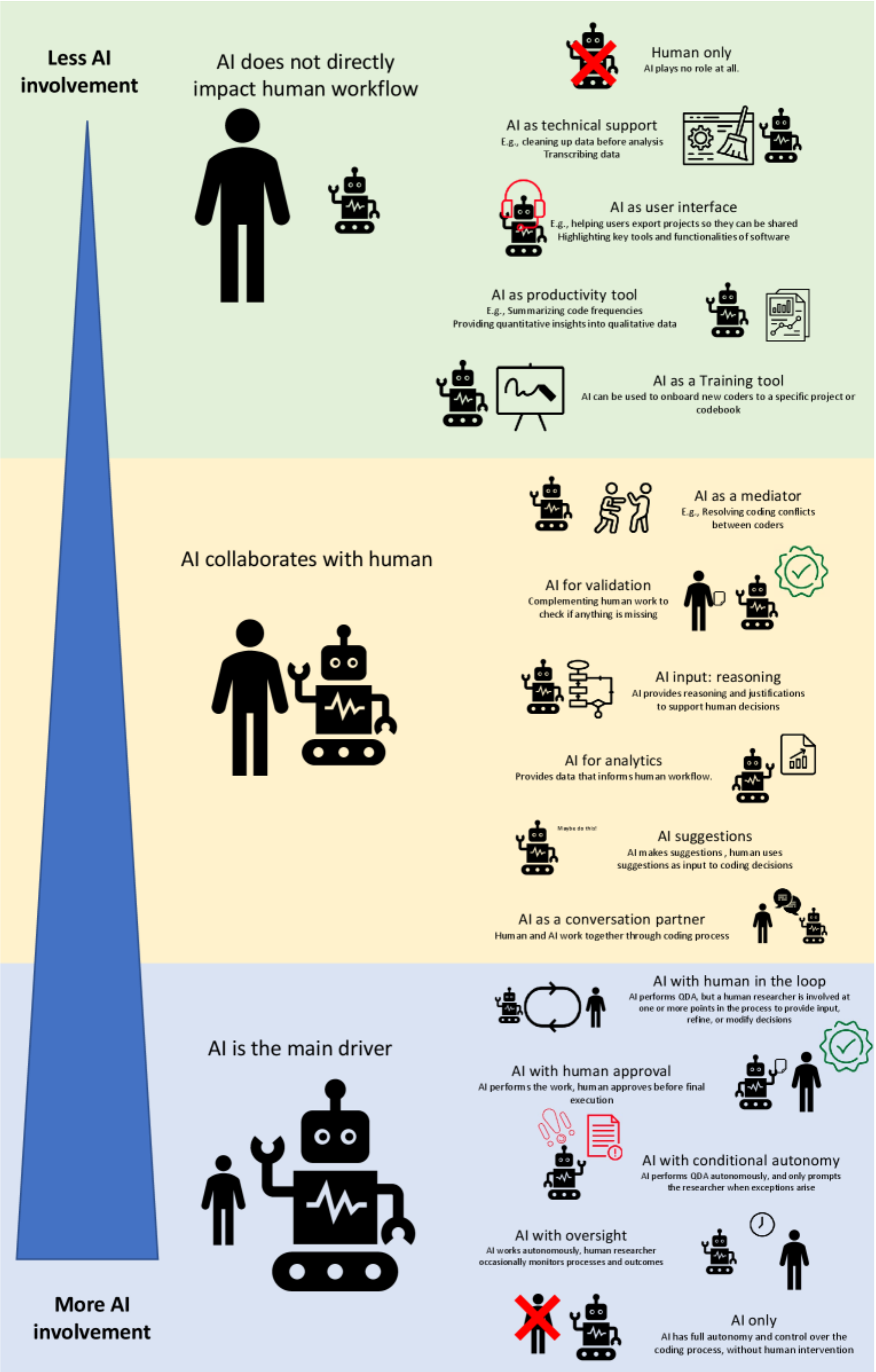

Figure 4 3 Stage AI Involvement Spectrum for QDA (Kirsten et al., 2025)

2.4.4 AI Applications in Specific Organizational Domains

Customer service analysis represents one of the most mature applications of AI-enhanced qualitative research in organizations. AI systems are increasingly deployed to analyse customer interactions, identify service quality issues, and understand customer experience patterns. Research demonstrates that AI-enhanced analysis of customer service data can provide insights that improve operational efficiency, personalization, and customer satisfaction (Klén, 2024). These applications often involve analysis of large volumes of customer interaction data across multiple channels and touchpoints.

Legal document analysis showcases sophisticated AI applications for processing complex organizational documentation. AI systems are being deployed for contract analysis, compliance monitoring, and legal risk assessment. Research demonstrates significant time savings and accuracy improvements when AI is applied to legal document analysis tasks. These applications require sophisticated natural language understanding capabilities to navigate complex legal terminology and reasoning. (de Martim, 2025)

Human resources and employee experience analysis leverage AI technologies to understand workforce dynamics, engagement patterns, and organizational culture. AI-enhanced analysis of employee feedback, performance reviews, and organizational communications provides insights into employee satisfaction, retention factors, and organizational effectiveness. These applications require careful consideration of privacy and ethical concerns while providing valuable insights for organizational development. (Dutta and Kannan Poyil, 2024; Murugesan et al., 2023)

Marketing and brand analysis employ AI technologies to analyse customer feedback, social media interactions, and market research data (Jain and Kumar, 2024). AI-enhanced qualitative analysis enables organizations to understand brand perception, customer preferences, and market positioning more comprehensively (Mitchell, 2021). These applications often involve analysis of large-scale, multi-modal data from diverse sources requiring sophisticated integration and synthesis capabilities.

2.4.5 Methodological Innovations and Hybrid Approaches

The integration of AI into qualitative research has spawned innovative hybrid methodologies that combine the strengths of human interpretation with AI computational capabilities. These approaches recognize that AI augmentation should enhance rather than replace human interpretive expertise. HITL systems maintain human control over key analytical decisions while leveraging AI for data processing and pattern identification. (Goyal et al., 2024)

Collaborative analysis frameworks enable researchers to work synergistically with AI systems throughout the research process. These frameworks typically involve AI systems providing initial coding and pattern identification, followed by human refinement, interpretation, and theory development. Research demonstrates that such collaborative approaches can maintain or improve analytical quality while significantly reducing time requirements. (Jun, 2025)

Multi-perspective analysis approaches leverage AI capabilities to examine qualitative data from multiple analytical perspectives simultaneously. These approaches can identify patterns and themes that might be missed by single-perspective analysis while maintaining the interpretive depth characteristic of qualitative research. AI systems can process the same dataset using different theoretical frameworks or analytical approaches, providing researchers with multiple interpretive lenses.

Iterative refinement methodologies employ AI systems to support ongoing analytical refinement and theory development. These approaches recognize that qualitative analysis is inherently iterative, with understanding developing through successive analytical cycles. AI systems can support this iterative process by providing rapid reanalysis capabilities and pattern identification across analytical iterations.

2.5 State of Integration of MASs in QDA

2.5.1 Current State of Integration

The intersection of multi-agent AI systems with qualitative research in organizational contexts represents an emergent and largely unexplored research domain. Current literature reveals significant gaps in the systematic integration of multi-agent capabilities with qualitative research methodologies, despite considerable potential for synergistic applications. Most existing applications of AI in qualitative research rely on single-agent systems or centralized AI platforms with similar architecture as shown in figure 5, missing opportunities for distributed and collaborative approaches that could enhance both analytical depth and breadth.

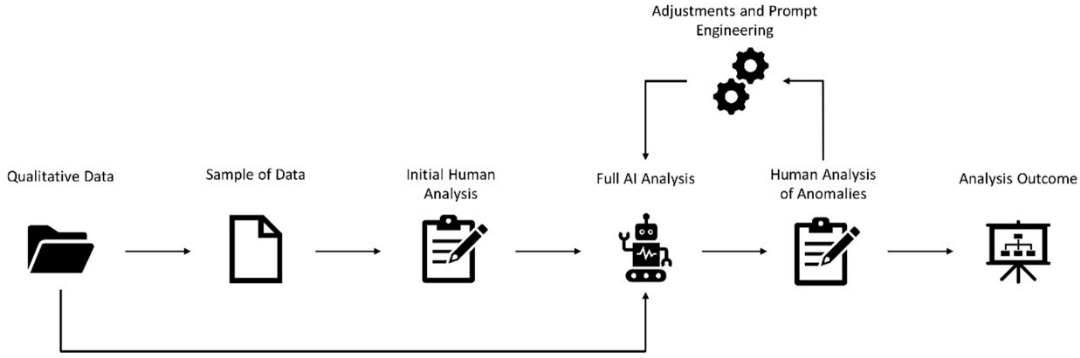

Figure 5 Flow Chart of AI Assisted QDA Process Example (Nicmanis and Spurrier, 2025)

Limited empirical studies have explicitly examined multi-agent approaches to qualitative organizational research. The few existing examples focus primarily on educational applications or technical system evaluations rather than comprehensive qualitative research methodologies. This limitation represents a significant gap given the natural alignment between multi-agent coordination capabilities and the collaborative nature of qualitative research teams. (Xu et al., 2025)

Conceptual frameworks for multi-agent qualitative research remain underdeveloped, with most existing literature treating AI assistance as a single-agent augmentation rather than a distributed, collaborative system. The DORA AI Scientist system represents one of the few examples of multi-agent approaches to research processes, but its focus remains primarily on quantitative and bibliometric analysis rather than deep qualitative investigation (Naumov et al., 2025).

Methodological integration challenges persist due to the fundamental differences between traditional qualitative research paradigms and multi-agent system architectures. Qualitative research emphasizes human interpretation, contextual understanding, and emergent meaning-making, while MASs focus on distributed computation, coordination protocols, and algorithmic decision-making. Bridging these paradigmatic differences requires sophisticated theoretical and practical frameworks that have yet to be fully developed.

2.5.2 Potential Applications and Use Cases

Distributed thematic analysis represents a promising application area where MASs could significantly enhance qualitative research capabilities. Multiple specialized agents could work collaboratively to identify themes, patterns, and relationships across large qualitative datasets while maintaining the interpretive depth essential to qualitative inquiry. Each agent could specialize in different analytical perspectives or theoretical frameworks, providing multi-faceted analysis that surpasses single-agent capabilities. (Xu et al., 2025)

Cross-organizational research scenarios present natural applications for multi-agent qualitative research systems. When research involves multiple organizations, agencies, or stakeholders, MASs could facilitate distributed data collection and analysis while respecting organizational boundaries and privacy requirements. Each organization could deploy its own agent that collaborates with other agents while maintaining local data control and analytical autonomy. (Xu et al., 2025)

Multilingual and cross-cultural studies could benefit significantly from multi-agent approaches where different agents specialize in specific languages, cultural contexts, or regional knowledge. This specialization could enable more nuanced cross-cultural analysis while maintaining the contextual sensitivity essential to qualitative research. Agents could collaborate to identify universal patterns while preserving culturally specific insights. (Lingard and Klasen, 2025; Xia et al., 2024)

Longitudinal organizational studies requiring sustained data collection and analysis across extended time periods could leverage MASs for continuous monitoring and analysis. Different agents could specialize in different temporal aspects or organizational dimensions while maintaining analytical coherence across time. This approach could provide insights into organizational change processes that are difficult to capture through traditional qualitative methods.

2.5.3 Technical Architecture Considerations

Hybrid human-AI coordination architectures must balance the need for systematic coordination with the flexibility and creativity essential to qualitative research. These architectures must accommodate both structured analytical processes and emergent, creative insight development. Current proposals suggest hierarchical coordination structures with human researchers maintaining strategic oversight while AI agents handle tactical analytical tasks (Olatunji Ottun et al., 2025). Anthropic's reasoning architecture demonstrates this well in figure 6.

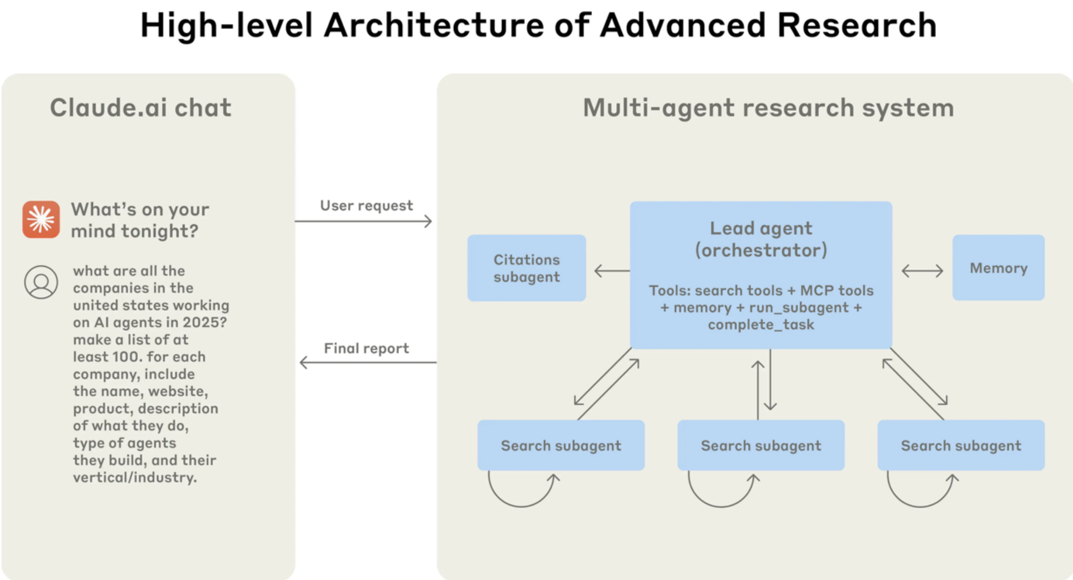

Figure 6 Advanced Research MAS Architecture by Anthropic (Hadfield et al., 2025)

Knowledge representation and sharing mechanisms become critical in multi-agent qualitative research systems where different agents must share insights, interpretations, and emerging theoretical understanding. Traditional knowledge representation approaches may be insufficient for capturing the nuanced, contextual understanding characteristic of qualitative research. Graph-based knowledge representation, as employed in GraphRAG systems, may provide more suitable frameworks for sharing qualitative insights among agents.

Privacy and confidentiality protection requires sophisticated technical approaches in MASs where sensitive organizational data may be processed by multiple AI agents (Huo et al., 2025). Federated learning approaches may provide technical solutions that enable collaborative analysis while maintaining data locality and privacy protection. These approaches allow agents to learn from distributed data without requiring centralized data sharing.

Real-time collaboration capabilities are essential for multi-agent qualitative research systems that must support dynamic, iterative analytical processes. Unlike batch processing systems, qualitative research requires ongoing collaboration and refinement as understanding develops. Technical architectures must support synchronous collaboration while maintaining analytical coherence across distributed agents (Olatunji Ottun et al., 2025).

2.6 Identified Research Opportunities

The comprehensive literature review reveals significant gaps at the intersection of multi-agent AI systems, GraphRAG, and QDA. These gaps represent substantial opportunities for advancing both theoretical understanding and practical applications in organizational contexts.

2.6.1 Gap 1: Absence of Integrated Multi-Agent GraphRAG Frameworks for Qualitative Data Analysis

The most critical gap identified is the fundamental lack of integrated frameworks that systematically combine multi-agent AI capabilities with GraphRAG's knowledge traversal advantages for qualitative organizational tasks. While MASs demonstrate sophisticated coordination capabilities and GraphRAG shows superior performance in knowledge-intensive reasoning tasks, no existing research has developed comprehensive frameworks that leverage both paradigms synergistically for QDA.

Current literature predominantly focuses on single-agent AI augmentation of qualitative research, missing the collaborative intelligence potential that distributed agent systems could provide. The few existing multi-agent applications in research contexts primarily address quantitative or bibliometric analysis rather than deep qualitative investigation, creating a significant methodological void.

This gap is particularly pronounced in organizational contexts where complex qualitative tasks such as customer service analysis, legal document processing, and multilingual dataset assessment could benefit substantially from both distributed intelligence and graph-based knowledge integration. The absence of such integrated approaches limits the scalability and depth of qualitative analysis in organizational settings.

2.6.2 Gap 2: Limited Empirical Validation of Multi-Agent QDA Methodologies

Despite significant theoretical potential, there exists a critical shortage of empirical validation studies examining the effectiveness of multi-agent approaches to organizational qualitative research. Current research lacks systematic comparisons between traditional qualitative methods, single-agent augmentation, and multi-agent collaborative approaches across different organizational contexts and task types.

The literature reveals that most AI-enhanced qualitative research studies focus on single-agent systems with limited evaluation of collaborative multi-agent configurations. This empirical gap is particularly concerning given the complexity of organizational qualitative research, which requires nuanced understanding of human experiences, cultural contexts, and organizational dynamics that may benefit from distributed analytical approaches.

Furthermore, existing evaluation frameworks for AI systems are predominantly designed for single-agent or technical performance metrics, lacking specialized methodologies for assessing multi-agent collaborative qualitative analysis. This methodological limitation impedes the development of robust empirical evidence for multi-agent qualitative research approaches.

2.6.3 Gap 3: Inadequate Technical Architectures for Multi-Agent Systems with Ability to do QDA

Existing multi-agent system architectures are inadequately designed for the dynamic, iterative, and contextually sensitive requirements of qualitative research processes. Current coordination mechanisms, knowledge representation approaches, and human-AI interaction designs are primarily optimized for task-oriented applications rather than the exploratory, emergent nature of qualitative inquiry.

The literature reveals significant limitations in existing multi-agent coordination protocols for handling the fluid, iterative analytical processes characteristic of qualitative research. Traditional coordination mechanisms may be insufficient for managing the creative, non-linear exploration required for qualitative insight development while maintaining analytical coherence across distributed agents.

Additionally, current knowledge representation and sharing mechanisms in MASs are poorly suited for capturing and exchanging the nuanced, contextual understanding characteristic of qualitative research findings. This technical limitation impedes the development of effective collaborative qualitative analysis systems that can maintain interpretive richness while enabling distributed processing.

2.6.4 Gap 4: Insufficient Multi-Level Evaluation Methodologies for Complex AI Systems

Current evaluation methodologies lack comprehensive multi-level assessment frameworks that can systematically evaluate AI systems across different operational modes (baseline, HITL, fully autonomous). Existing evaluation approaches typically focus on single-mode assessment or technical performance metrics, missing the complex interplay between human expertise and AI capabilities in organizational contexts.

The literature reveals limited research on comparative evaluation frameworks that can assess AI system effectiveness across multiple operational configurations while accounting for organizational context, task complexity, and human-AI collaboration dynamics. This methodological gap is particularly significant for organizational applications where system effectiveness depends critically on the balance between human expertise and AI automation.

Furthermore, existing evaluation frameworks inadequately address the unique challenges of assessing qualitative research outcomes, which require consideration of interpretive validity, contextual relevance, and theoretical contribution rather than purely quantitative performance metrics. This limitation impedes the development of robust evaluation methodologies for AI-enhanced qualitative research systems.

2.6.5 Gap 5: Limited Understanding of Organizational Implementation Barriers and Success Factors

Despite growing interest in AI adoption in organizational contexts, there remains insufficient understanding of the specific barriers and success factors for implementing complex AI systems like integrated multi-agent GraphRAG frameworks in organizational research activities. Current literature focuses primarily on technical capabilities rather than organizational readiness, change management requirements, perception analysis, and implementation challenges.

Additionally, current literature provides limited guidance on stakeholder engagement, training requirements, and organizational culture considerations necessary for successful implementation of sophisticated AI research systems. This implementation knowledge gap represents a significant barrier to translating research advances into practical organizational applications, use cases, and tasks.

2.6.6 Summary

Table 1 summarizes the gaps identified in the literature review alongside the current situation with a mapping on toward further research opportunities.

Table 1 Summary of Gaps Identified in the Literature Review

Gap Category	***Specific Gap***	***Current State***	***Research Opportunity***
Methodological Integration	Absence of integrated multi-agent GraphRAG frameworks for qualitative analysis	Single-agent systems dominate; no comprehensive integration approaches exist	Develop synergistic frameworks combining distributed intelligence with graph-based knowledge traversal
Empirical Validation	Limited empirical validation of multi-agent qualitative research methodologies	Theoretical potential exists but systematic comparative studies lacking	Conduct multi-context comparative studies across operational modes and task types
Technical Architecture	Inadequate architectures for multi-agent qualitative research systems	Existing coordination mechanisms unsuited for iterative, exploratory qualitative processes	Create specialized coordination protocols and knowledge representation for qualitative inquiry
Evaluation Methodology	Insufficient multi-level evaluation frameworks for complex AI systems	Single-mode assessment dominates; limited consideration of human-AI collaboration dynamics	Establish comprehensive multi-level assessment methodologies across operational configurations
Implementation Science	Limited understanding of organizational implementation barriers and success factors	Focus on technical capabilities rather than organizational readiness and change management	Investigate organizational factors facilitating successful adoption of complex AI research systems
Cross-Domain Application	Lack of domain-specific applications in organizational contexts	Limited deployment in customer service, legal analysis, and multilingual assessment contexts	Develop and evaluate domain-specific applications across diverse organizational contexts
Human-AI Collaboration	Insufficient frameworks for optimizing human-AI collaboration in qualitative research	Limited understanding of optimal collaboration patterns and decision authority distribution	Investigate optimal human-AI collaboration configurations for qualitative research tasks

3 Research Methodology

This chapter outlines the research methodology that will be used to design, implement, and evaluate a qualitative data analysis AI-driven system in the organization. Since it is the gist of the thesis, it clarifies the methodological decisions whose purpose is to answer the goals and gaps found in Chapters 1 and 2. This chapter will elaborate on the justification of the adopted method, how the research artifact was designed, developed, and implemented, and how the study used procedures to conduct qualitative analysis by interviewing experts. In that way, it will reveal the strong and the weak aspects of the research design and document methodological transparency and rigor.

The problem of the proposed research studied in this thesis is fundamentally socio-technical, and it concerns the interaction of the further AI systems, human researchers, and organizational settings (Olan et al., 2022). The available literature has been either concerned with technical implementations in MASs and retrieval-augmented generation, or on qualitative research approaches, which highlight the interpretation power of humans. Nevertheless, very little has been done that would offer these perspectives as one evaluative system. This limitation is overcome by the present study, which uses a Design Science Research Methodology (DSRM) approach that has an integrated mixed-method assessment because it will include quantitative data-based model assessments and qualitative two-phase semi-structured interviews.

Development of research environment, design artifacts, including system architecture drawings, agent workflow representations, data flow charts, codebase implementation, and user interfaces are included as methodological tools as opposed to outputs. The ability to be included usually guarantees the standardization of the conditions under which the evaluation occurs and defines the traceability of the system behaviour as well as the ability of the research process to be repeated. The chapter therefore gives a detailed description of what, how and why these decisions were fit to solve the research gaps identified.

3.1 Rationale for a Multi-Level Methodology

This thesis explicitly has the methodological approach that is guided by the five research gaps that have been found in literature review. These shortcomings are: **(1)** the lack of combined multi-agent GraphRAG systems, **(2)** empirical verification of multi-agent AI systems in qualitative research, **(3)** vices in technical architectures in terms of exploratory qualitative inquiry, **(4)** lack of multi-level evaluation paradigms, and **(5)** quality of implementation-oriented insights of organizational adoption.

To embrace the above gaps in totality, the analysis uses a multi-level approach, which cuts across the research environment development, quantitative analysis, and qualitative analysis. The central concept of this methodology is the creation and test of a prototype system that can work on five levels of AI autonomy (from agentic minimal to fully agentic multi-agent work with an optional HITL control). Depending on the autonomy configurations that need to be compared methodically, this staged design directly addresses Gaps 2 and 4.

Gap 1 will be fulfilled with the development of an integrated multi-agent GraphRAG architecture, which unites an agent coordination, as well as graph-based retrieval-augmented generation. This research is a paradigm operationalization and integration within a single organizational-based system, unlike other studies, which have been carried out focusing on these paradigms separately, but have yet to accomplish the QDA activities related to thematic coding, text corpus analysis, and instinct comparison.

Gap 3 is accordingly addressed by the fact that the HITL processes are introduced into selected autonomy levels which provide the possibility to master human interpretability and produce output based on efficiency. Instead of substituting human judgement, the system is created to supplement it with contextual retrieval, recommendations, and systematic analytical processes.

Gap 4 and gap 6 are resolved with the help of the modular technical architecture of the system and its demonstration to organizational practitioners. The architecture, together with practitioner interviews, enables the exploration of qualitative data in an iterative, non-linear way. Whereas, the practice-related insights of usability, trust, governance, and adoption readiness (dimensions that are often not considered in technical evaluations) are derived through the qualitative analysis part.

3.2 Overall Research Design

This study follows the double diamond framework (Design Council, 2005) with Design Science Research Methodology (DSRM), which is particularly suitable for research involving the creation and evaluation of an innovative artifact intended to solve a real-world problem (vom Brocke et al., 2020). DSRM emphasizes the iterative development of artifacts, their empirical evaluation using multiple forms of evidence, and the generation of design knowledge that is both theoretically grounded and practically relevant (Huseynli et al., 2022). Figure 7 and figure 8 visualize both the concepts at a high level.

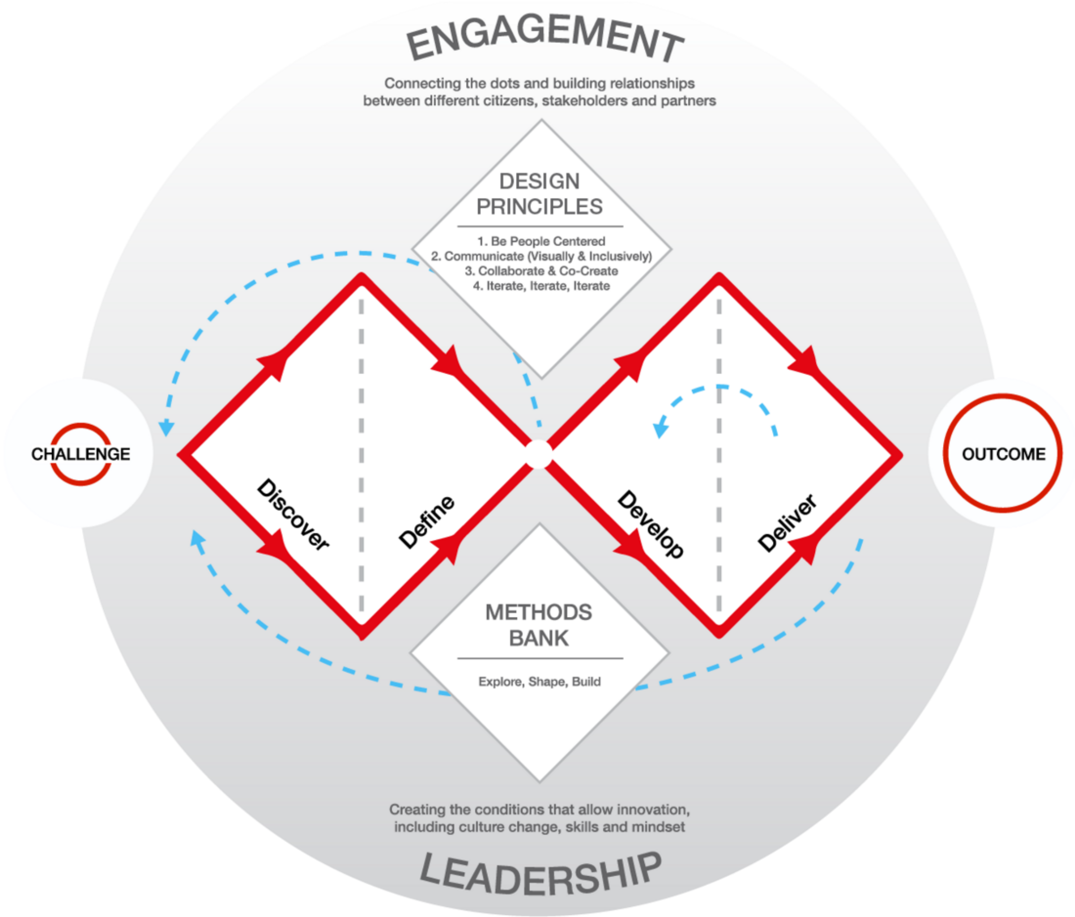

Figure 7 Double Diamond Design Framework (Design Council, 2005)

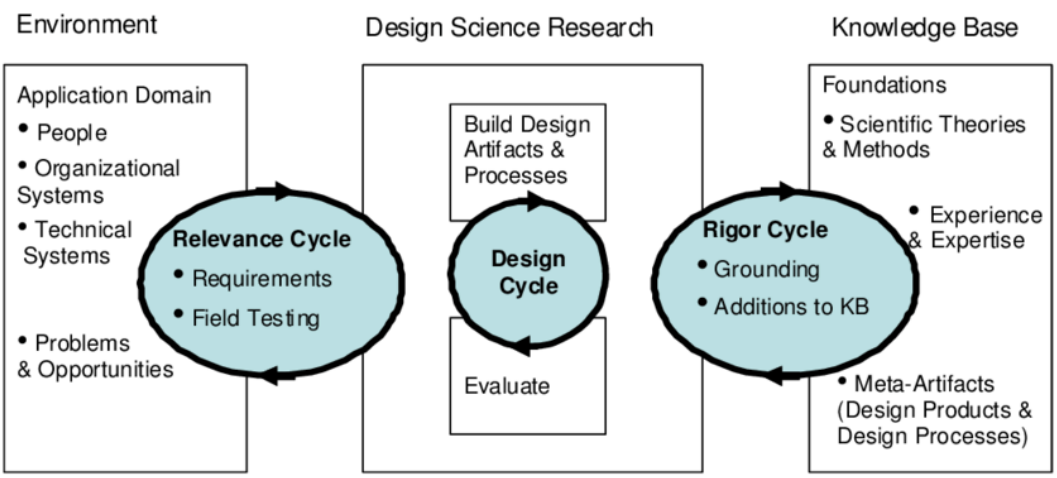

Figure 8 Design Science Research Method Overview (Zhou and Chen, 2025)

The research design comprises three tightly integrated components:

I. **Artifact Design and Implementation:** The development of a web-based multi-agent GraphRAG system, including system architecture, agent coordination mechanisms, staged autonomy configurations, and user interface components. Also uses double diamond iterative development approach.

II. **Quantitative Evaluation:** This study does not involve the development or training of new foundation models. Instead, it evaluates the behaviour of an AI-based system built on existing LLMs through controlled use of datasets. These datasets are used to benchmark retrieval quality, assess consistency of agent workflows, and stress-test system behaviour across different architectural and autonomy configurations.

III. **Qualitative Evaluation Through Two-Phase Interviews:** Semi-structured interviews with organizational AI practitioners conducted in two phases: an initial phase exploring existing AI practices and constraints, and a second phase evaluating the proposed artifact.

This embedded mixed-methods design supports triangulation by integrating complementary forms of evidence, thereby strengthening the validity and reliability of the findings.

Table 2 Research Questions & Methodological Mapping

Research Question	***Focus***	***Methodological Approach***
RQ1: *What architectural requirements support agentic qualitative analysis?*	System design	Artifact design, architecture diagrams, agent workflow analysis, system logging
RQ2: *How do organizations perceive and evaluate agentic AI for qualitative data analysis?*	Organizational perspective	Phase 1 semi-structured interviews, thematic analysis
RQ3: *How is success defined and assessed across different autonomy configurations?*	Evaluation criteria	Phase 2 interviews, derived success dimensions, comparison across autonomy stages
RQ4: *What technical and organizational constraints affect adoption (e.g., data, security, accuracy)?*	Constraints & risks	Interview data, risk/constraint coding, technical checks and system logs

3.3 Design Science Research Methodology Applied

3.3.1 Problem Identification and Motivation

Organizations are turning towards qualitative data like reviews, interviews, surveys, reports, and open ended as a source of data to guide strategic decision making. Although AI systems are scalable and efficient, solutions are not usually transparent, aligning interpretively with human researchers, and providing tools to manage autonomy. This forces a conflict between scalability, quality, and interpretive rigor, which encourages the creation of an AI system that may facilitate qualitative inquiry without impairing human processes of deriving valuable meaning. (Hadfield et al., 2025; Luzolo et al., 2024; Xiang et al., 2025; Xu et al., 2025)

3.3.2 Objectives of the Solution

The proposed system will meet the following objectives of the research:

I. To construct a unified multi-agent GraphRAG system to construct and analyse datasets.

II. To permit the selection and variation of AI autonomy under control by configurations which are staged.

III. To maintain human interpretive power through the HITL systems.

IV. To offer accountability of AI reasoning and agent action.

V. To measure the technical performance of agentic operations empirically.

All the objectives are specific to a particular research gap which guarantees the alignment of theory and design with the evaluation.

3.3.3 Implementation and Design of Artifacts

The research artifact applies as a web-based modular system where it is a frontend user interface, a backend systems layer, various AI agents' orchestration, and a GraphRAG knowledge layer. Several diagrams, which are discussed in this chapter, are used in documenting the system architecture to promote methodological transparency and reproducibility.

A system overview diagram, as in figure 9, is used to show the high-level elements and their interactions so that the scope and boundaries of the artifact are defined.

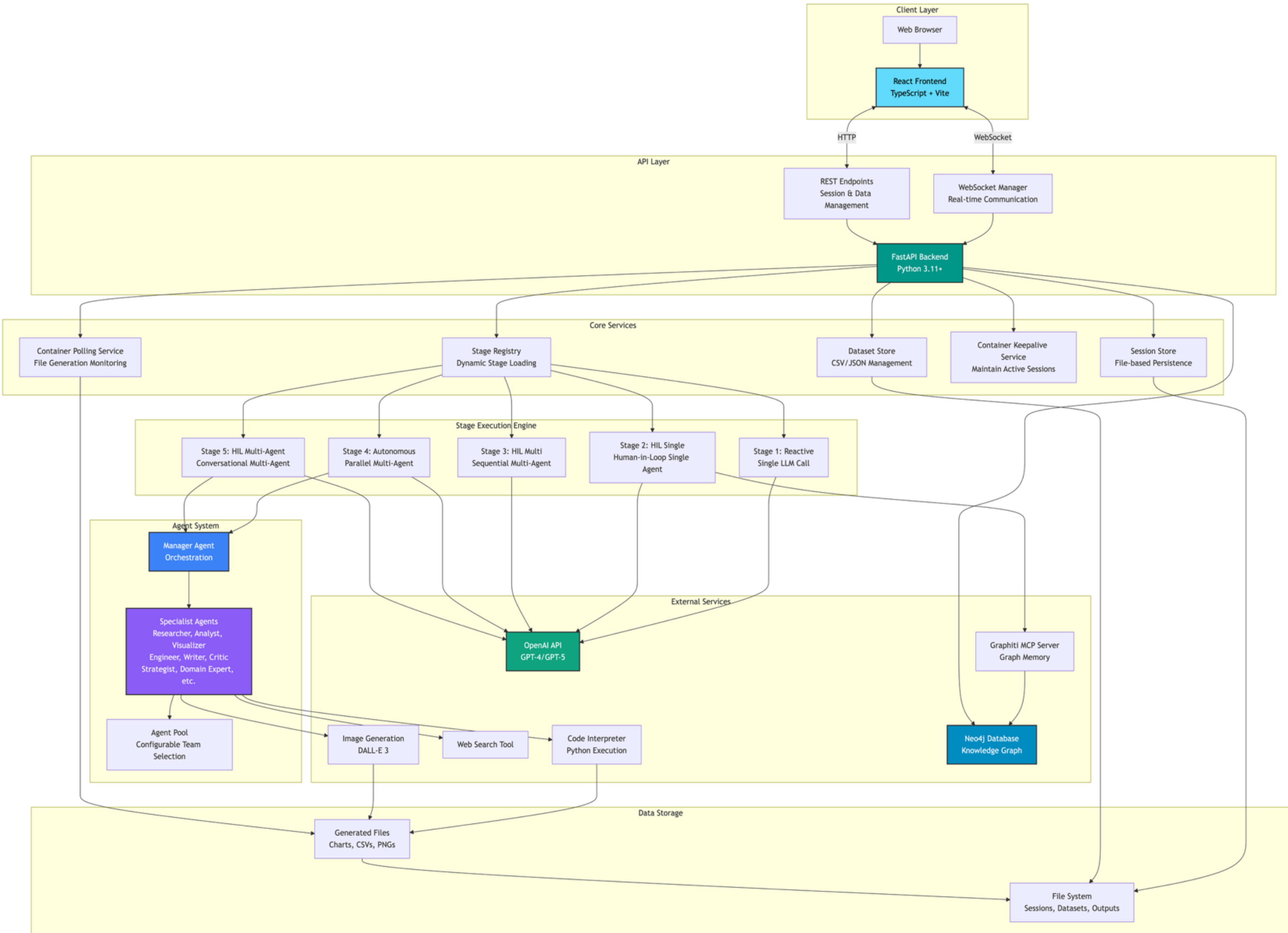

Figure 9 System Architecture Overview of the Research Environment

The research artifact is architected as a multi-tier, web-based modular system designed to balance high-scale agentic autonomy with human-centric interpretive rigor. At the Client Layer, a high-fidelity interface developed with React, TypeScript, and Framer Motion facilitates real-time interaction through 34 custom UI components. This layer communicates via HTTP and WebSockets with a FastAPI backend, ensuring that the Agent Activity Log and internal "Thinking" steps are streamed to the user with zero latency to maintain accountability (Objective IV). This real-time visibility is essential for bridging the gap between AI scalability and qualitative interpretive rigor.

The core of the platform is the Stage Execution Engine, which operationalizes the system's primary innovation: staged autonomy (Objective II). Through a centralized Stage Registry, the system permits the variation of AI control across five distinct configurations, ranging from simple reactive LLM calls (Stage 1) to a fully conversational Human-in-the-Loop multi-agent environment (Stage 5). When higher stages are selected, a Manager Agent serves as the primary orchestrator, delegating complex qualitative tasks to a configurable team of Specialist Agents (such as the Researcher, Analyst, Visualizer, etc.) who utilize external tools like Python code execution and web search to ground their analysis. (OpenAI, 2025b, 2025c, 2025a)

To ensure structural transparency and data persistence, the architecture integrates a robust Data and Knowledge Layer. A GraphRAG system, powered by a Neo4j Database, constructs a knowledge graph of the qualitative data to maintain deep context and relationship-based insights (Objective I). This knowledge layer directly maintained by an MCP server that handles all the interactions with the knowledge graph including initializing schema, adding nodes and relationships, and maintaining it with adding or deleting knowledge based on natural human inputs (GetZep, 2025; The Linux Foundation, 2025). All analytical outcomes are finalized in a Data Storage layer that manages session persistence and the export of generated artifacts, such as CSVs and charts. The technical scale of this implementation is evidenced by a codebase of 29,137 lines and the integration of 50 unique software packages, providing a stable environment for empirical performance measurement (Objective V). More information is also detailed in Appendix A. (Amaan, 2025)

Appendix G, H, and I contain the frontend architecture, the backend architecture, and the data flow diagram of the Agentic AI research environment we have developed. These diagrams are critical elements of the methodology framework because they delineate the conditions of the experimentation in which the evaluation will take place and allow future researchers to replicate results and methodologies.

3.3.4 Staged Autonomy Design of Agentic Systems

It has five levels of AI autonomy, each of which corresponds to a specific structure of agent capabilities, memory, access to tools, and human control, as detailed in table 3. These configurations are formalized via agent system diagrams and stage workflow diagrams in which every stage corresponds to an experimentally defined clear-cut experimental condition.

The staged design is a key to finding an answer to the research questions concerning autonomy, trust, and interpretive control. Through a controlled manipulation of the degree of autonomy, the methodology will make possible a quantitative and qualitative comparison to determine the success of various stages, which the literature has found lacks the necessary multi-level evaluation paradigms. Appendix J displays the 5 agentic stages in a more visual way in the form of flowcharts.

Table 3 Operational Definition of AI Autonomy Stages

Stage	***Tools***	***Memory***	***HITL***	***Success Definition***	***Logs Captured***
Stage 1	None	None	No	Accurate response to user query	Prompt + response
Stage 2	Retrieval, analysis tools	Session memory	No	Relevant retrieval + coherent output	Tool usage + outputs
Stage 3	Same as Stage 2	Session memory	Yes	Human-approved analytical steps	Approval actions + logs
Stage 4	Multi-agent tools	Shared memory	No	Coordinated agent output	Agent messages + actions
Stage 5	Same as Stage 4	Shared memory	Yes	Human-approved multi-agent result	Full agent + approval logs

3.4 Research Data and Quantitative Evaluation

This study does not involve the development or training of new foundation models. Instead, it evaluates the behaviour of an AI-based system built on existing LLMs through controlled use of open-source datasets (Appendix C). These datasets are used to benchmark analytical quality, assess consistency of agent workflows, and stress-test system behaviour across different architectural and autonomy configurations.

The testing processes involve standardized agentic runs to evaluate the capability of AI models to recall pertinent contextual data, produce similar thematic codes and contribute to analytical thinking. Assessment is done on the levels of autonomy to measure the effect of agent coordination and HITL mechanisms on technical performance indicators.

The performance measures are statistical in terms of system performance including response latency, files/artefacts produced, and number of intermediate thinking or tool call steps measured. These measures give objective results about the behaviour of the systems and supplement qualitative results obtained through the interviews of the experts.

3.5 Qualitative Study with Two Phase Interviews

3.5.1 Participant Selection

Participants are selected using purposive sampling to ensure relevance to the research objectives. Eligible participants are organizational practitioners with direct experience in using, deploying, managing, or evaluating outputs of AI systems for data analysis, decision support, or research-related tasks.

A total of seven participants were selected, consistent with established practices in exploratory design science and expert interview studies. Recruitment is conducted through professional networks, and all participants provide informed consent prior to participation.

3.5.2 Participant Background Overview

The table 4 below presents a high-level overview of the participants of the qualitative study.

Table 4 Qualitative Study Participants' Overview

Participant	*Current Role & Organization*	*Primary Domain*	*Years of Experience*	*Previous Background / Notable Experience*
P1	DevOps Engineer, Nokia	Software engineering / DevOps (telecom software lifecycle automation)	~3+ years at Nokia	Former Network Engineer at Huawei in multiple countries
P2	AI / Machine Learning Expert at Granlund	AI, Machine Learning, Computer Vision, Generative AI	~4 years + projects' experiences	Experience in construction industry, startups, and leading technical teams
P3	PR, Marketing, & Communications Expert at IG Group	Corporate communications, reputation & media management in large tech firms	~5 years	Works with major global companies; extensive use of AI for workflow automation and GEO/SEO related tasks
P4	Business Development & Finance Professional, Industrial Water Treatment Company	Business development, finance, industrial water systems	~7–8 months in current role	Recent graduate; internships at Daimler Trucks (US) and Bosch (Germany)
P5	Data & AI Lead, Sailpoint	Data engineering, enterprise identity security analytics & visualization	~2.5 years	Works on data visualization and search tools within data platform teams
P6	Data Integration Engineer, Wärtsilä	Data engineering / integration	~2 years	Master's in Data Science; prior experience in fintech (Turkey) and internship in Finland;
P7	PhD Researcher, Aalto University	Geoinformatics, Remote Sensing, Space Data Science	PhD Year 2 + prior industry experience	Works with UAV, satellite, and airborne datasets; former role at ICEYE Finland

3.5.3 Phase 1: Organizational Context

Phase 1 interviews will examine current applications of AI, governance systems, analytical processes and perceived barriers. This step will constitute a baseline which will be used to determine the suitability of the proposed system. This phase also gauges the AI usage of the participant in an organizational context and their experience in interacting with such systems. The interview questions are listed in Appendix E.

3.5.4 Phase 2: Artifact Evaluation

At Phase 2, the system is used to engage the participants through varying degrees of autonomy on standardized consistent opensource datasets, prompts, and visual artifacts. Architecture diagrams and agent workflows and UI elements are taken to guarantee some common understanding and evaluation conditions.

Interview data are analysed using thematic analysis, with codes derived inductively and iteratively refined. Trustworthiness is supported through consistent interview protocols, transparent documentation of analysis steps, and triangulation with system logs and quantitative observations. The interview questions are listed in Appendix F.

3.6 Data Analysis

To establish the study, the mixed-method data analysis strategy aligned with DSRM Research is used, involving quantitative analysis of the system performance and qualitative data analysis of the interviews using the thematic data analysis strategy. This will allow assessing the technical behaviour of the proposed system and evaluating the suitability of the system both in the aspect of interpretation and organization.

Quantitative analysis is performed with the help of organizational-style open source datasets that are contextually relevant to organizational tasks. Descriptive metrics are used to assess performance at the different levels of autonomy and include retrieval relevance and consistency of generated thematic codes, response latency, and error patterns that were observable. These measures are employed in a comparative and not inferential way, which helps to assess the impact of raising degrees of autonomy on system behaviour and analytical dependability.

Thematic analysis is used to analyse qualitative data of the Phase 1 and Phase 2 interviews. The theme analysis of phase 1 is aimed at determining the existing organizational AI practices, the organization governance and the fundamental analytical processes. Phase 2 considers how the proposed system is perceived by the participants and what categories are used to understand how much they can trust the new system, have transparency, interpretive control, and appropriateness of various autonomy levels.

Quantitative and qualitative results are combined showing design implications and to evaluate trade-offs in autonomy, human control, and organizational viability, which is a direct solution to the problem of multi-level evaluation gap, which the literature references.

3.7 Research Rigor & Trustworthiness

The rigor of research is provided by methodological triangulation, standard evaluation processes and clear documentation of artifacts. It not only increases the validity of findings but also enhances the credibility of the results achieved through a combination of the dataset-based performance evaluation, interview data, and design artifacts.

The consistency of datasets, prompts, environment version, and demonstration protocols in all evaluations aid to support credibility. Phase 2 interviews make use of architecture diagrams, agent workflow representations and UI elements, as we attempt to have a common understanding of how the system functions by the participants.

A high level of dependability is made by providing descriptive documentation of system architecture, autonomy phases, and evaluation guidelines so that this can be recreated or expanded in the future by other researchers. The systematic coding of data collected during the interview and a clear recognition of the dual role that the researcher plays as the system designer confirm the aspect of confirmability.

Although statistical generalization is not a goal of the study, transferability is fulfilled by presenting the key characteristics of the participants, the system, and conditions of evaluation in clear ways.

3.8 Ethical Conflicts & Limitations

The approval of ethics is received according to the institutional guidelines. Every respondent gives informed consent and gets informed about his/her right to withdraw at any time. All the data of participants and the affiliate organization remain anonymous, and all data of the interviews are safely stored. System exhibitions do not use organizational data, which is sensitive and proprietary.

There are several weaknesses to be noted. The qualitative analysis will use the small sample of seven (7) subjects, which suits the exploratory design science study, but restricts generalization. The system is tested as a research prototype and is not evaluated as production system and the demos are made in controlled locally hosted run. Also, both positions are taken by the researcher as a designer and a evaluator which can bring bias although this would be managed by standard processes and systematic analysis.

3.9 Key Takeaways

This chapter detailed the hybrid-methodological research. It involves DSRM, integrating qualitative expert insights with quantitative system evaluation. The primary research artifact is a modular Agentic AI environment developed to facilitate staged AI autonomy across five levels, bridging AI scalability and human interpretive rigor. The empirical framework utilized a dual-phase approach: Phase 1 identified QDA related pain points through semi-structured interviews with seven specialists, while Phase 2 encapsulated their insights through a user study demonstrating the Agentic AI research environment. Quantitively, performance is recorded through 20 controlled experimental runs across four diverse opensource datasets. By analyzing the pattern between agentic autonomy and analytical outputs, including run time, tool-call frequency, and file production, this methodology provides a transparent foundation for assessing the organizational feasibility and accountability of multi-agent systems in qualitative inquiry.

4 Results

This chapter will present the findings of empirical review of the proposed multi-agent AI GraphRAG based system that was designed based on Design Science Research Methodology (DSRM) as described in Chapter 3. As per the research design, the findings will be organized in such a way that they show the qualitative as well as descriptive quantitative evidence collected during two stages of assessment among organizational practitioners.

The evaluation along with the methodology were done as outlined in the methodology chapter it is based on semi-structured interviews with seven participants who were selected based on their organizations using or actively assessing AI-based tools of analysis. The assessment was in a two phase format:

- **Phase 1 (pre-demonstration)** presented already extant organizational AI practices, attitudes, governance limits, and expectations in regard of agentic AI systems.

- **Phase 2 (during and post-demonstration)** involved responses of the participants post/during demonstration of being interacted with the agentic AI research environment, developed as a part of the autonomy, transparency, usability, trust, and potential to adopt.

In line with the **hybrid approach** of methodology, this chapter combines:

- Thematically analysed interview transcripts to obtain the **qualitative** findings to capture interpretive insights, perceptions, and contextual constraints; and

- Descriptive **quantitative** results critical to system architecture, agent coordination and interface-level analysis of the system architecture, agent interaction and user interface/user experience factors are the results of structured observations, indicators of system interaction, and system architecture indicators.

The qualitative findings offer a detailed insight into the participants and/or organization's perception, the trusts, and the adoption factors whereas the quantitative findings accompanied these perceptions by demonstrating how participants engaged with the architectural aspects, the agent activities, and the user interface features of the system. These results, combined, provide a multi-level analysis of the research environment/artifact, which is in line with the concept of DSRM regarding utility and relevance.

The results are presented according to the research questions of the study in a manner that allows tracking the existence of the material findings to the research questions, the methodological decisions, and bridging the gaps between the pieces of the concept map developed in the literature review. It is thus a linkage between the system design and implementation that has been provided in the preceding chapters.

4.1 Quantitative Results

This section includes a quantitative measure of the descriptive characterization of the system operation behaviour across five stages (Stage 1 - Stage 5) based on four representative datasets (Airline, Airline 100, Spotify and Amazon) (Appendix C). Based on the DSRM, referenced in Chapter 3, the aim is to describe the effect of architectural decisions, autonomy, and HITL control on system behaviour, as opposed to purporting to statistical performance.

4.1.1 Experimental Controls and Methodology

To ensure the transparency and reproducibility of these results, the following constant variables were maintained across all x20 runs as given in table 5.

Table 5 Experiment Runs' Control Setup

Control Category	***Constant Variable / Definition***	***Research Purpose***
Primary Prompt	A standardized command was used for all runs: *“Analyse the dataset and create a report.”*	Ensures that variations in output are due to agentic autonomy levels rather than prompt engineering.
System Configuration	"Reasoning Depth" was held at **Low** for all stages to maintain a consistent computational baseline. Only GPT-5 was used for all LLM calls.	Standardizes the "thoughtfulness" and resource consumption of the underlying model.
HITL Protocol	For Stages 3 and 5, the agent always selected the **recommended option** with no additional user instructions.	Removes human bias and variability from the experimental performance data.
Team Selection	Stages 4 and 5 used a fixed team: **Manager, Data Analyst, Visualizer, and Researcher**.	Validates the multi-agent orchestration layer under a controlled organizational structure.
Run Time (s)	Total wall-clock duration from initial submission to final report output.	Measures technical performance and efficiency (Objective V).
Thinking/ Tool Steps	The cumulative count of internal reasoning turns and active tool calls recorded in the activity log.	Quantifies the complexity and depth of agentic reasoning (Objective IV).
Files Produced	The total number of saved deliverables (e.g., .md, .pdf, .csv, .png, .svg).	Evaluates the tangible output density of each autonomy stage.

Three operational metrics are reported:

I. **Run time (seconds)**: recorded time from prompt submission to final report generation with final message sent.

II. **Thinking/tool-call steps**: total recorded reasoning turns and tool invocations.

III. **Files produced**: number of saved output artifacts (reports, tables, visualizations)

4.1.2 Stage-Level Summary Across Datasets

Table 6 provides a stage-level summary of runtime, steps, and artifact output aggregated across the four datasets.

Table 6 Stage-Level Summary of Controlled Agentic Analysis Runs

Stage	Mean Runtime (s)	Median Runtime (s)	Runtime Range (s)	Mean Steps	Mean Files	Files Range
Stage 1	37.5	37.5	31–44	2.5	1.0	1–1
Stage 2	142.5	145.0	112–168	23.75	4.25	3–5
Stage 3	475.5	479.0	440–504	30.25	7.25	6–8
Stage 4	1297.8	1293.5	1146–1458	82.25	12.25	11–13
Stage 5	2488.0	2479.5	2204–2789	91.75	14.50	9–18

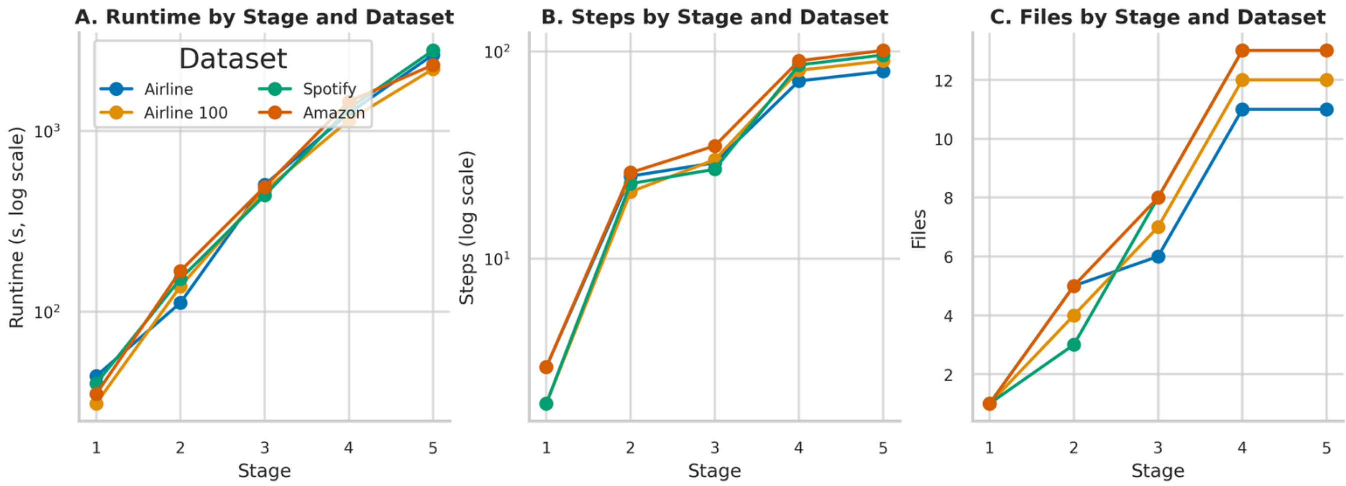

Figure 10 Dataset-specific Results by Autonomy Stage

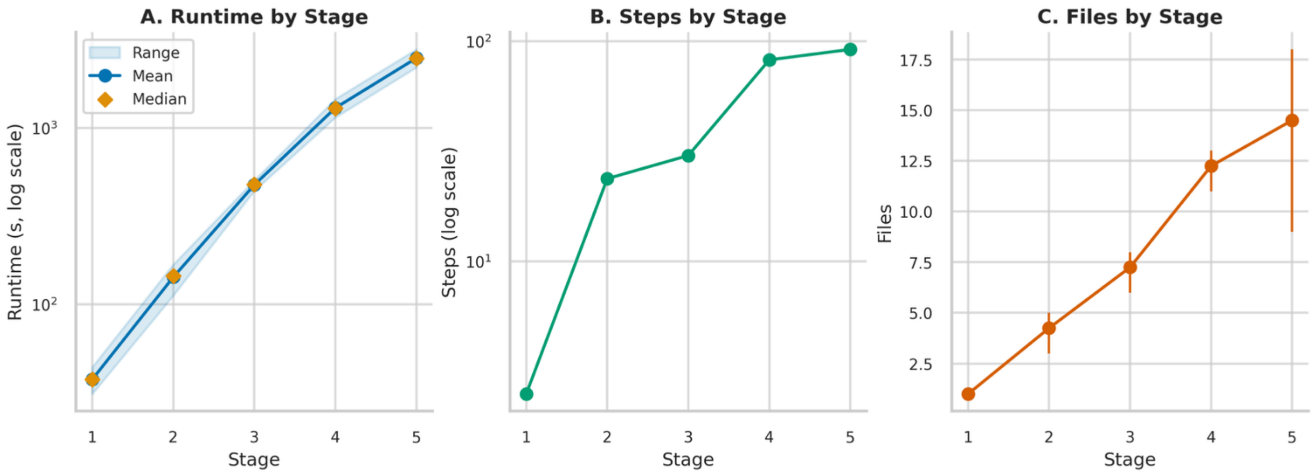

Figure 11 Stage-level Results by Autonomy Stages (Aggregated)

A clear monotonic trend is observed in figure 10 and 11: higher autonomy corresponds to longer runtimes, more interaction steps, and richer artifact production. The most pronounced increase occurs between Stage 1 and Stage 2, reflecting the transition from a reactive chatbot to tool-based analytical execution. Multi-agent stages 4/5 exhibit the highest workload due to coordination and parallel agentic activity.

4.1.3 Raw Results by Dataset & Stage

To illustrate dataset-specific behaviour, table 7 below reports the raw runtime, step count, and file output for each dataset at each autonomy stage.

Table 7 Raw Results by Dataset (Runtime / Steps / Files)

Stage	***Airline***	***Airline 100***	***Spotify***	***Amazon***
Stage 1	44 / 3 / 1	31 / 2 / 1	40 / 2 / 1	35 / 3 / 1
Stage 2	112 / 25 / 5	138 / 21 / 4	152 / 23 / 3	168 / 26 / 5
Stage 3	504 / 29 / 6	468 / 30 / 7	440 / 27 / 8	490 / 35 / 8
Stage 4	1255 / 72 / 11	1146 / 81 / 12	1332 / 86 / 13	1458 / 90 / 13
Stage 5	2640 / 80 / 14	2204 / 90 / 9	2789 / 96 / 18	2319 / 101 / 17

Within each autonomy stage, runtime varies with dataset complexity. Datasets with more information, richer categorical structure, and broader segmentation demands (Spotify, Amazon) consistently produce longer runtimes and more interaction steps. However, these dataset effects remain secondary to the dominant influence of autonomy level.

4.1.4 Human-in-the-Loop (HITL) Overhead

A core design question addressed in this thesis concerns the cost of human oversight. By fixing HITL decisions to the system-recommended option, the benchmark isolates the structural overhead introduced by approval checkpoints. Table 8 and 9 calculate the subsequent cost deltas of HITL.

Table 8 Stage 3 – Stage 2 (Single Agent HITL) Comparison

Dataset	***Runtime Δ (s)***	***Steps Δ***	***Files Δ***
Airline	+392	+4	+1
Airline 100	+330	+9	+3
Spotify	+288	+4	+5
Amazon	+322	+9	+3
Mean	**+333**	**+6.5**	**+3.0**

Table 9 Stage 5 – Stage 4 (Multi Agent HITL) Comparison

Dataset	***Runtime Δ (s)***	***Steps Δ***	***Files Δ***
Airline	+1385	+8	+3
Airline 100	+1058	+9	-3
Spotify	+1457	+10	+5
Amazon	+861	+11	+4
Mean	**+1190.25**	**+9.5**	**+4.5**

HITL checkpoints introduce measurable overhead, particularly in multi-agent configurations where managerial approval must integrate multiple specialist outputs. While runtime and steps increase substantially, artifact counts also increases in Stage 5, indicating that HITL affects process cost as well as output volume due to more informed and directed execution.

4.1.5 Alignment with Research Questions

The quantitative findings directly support the study's research questions:

- **RQ1 (Architectural requirements):** Increasing steps and runtime across stages demonstrate the operational impact of tool integration, agent coordination, and graph-based workflows.
- **RQ2 (Organizational perception and adoption)**: Measured HITL overhead empirically explains participants' preference for staged autonomy and controlled execution.

- **RQ3 (Measuring success of agentic systems)**: Artifact volume and workflow richness provide tangible indicators of analytical productivity beyond response speed.

- **RQ4 (Practical deployment constraints)**: Runtime growth and coordination costs highlight real-world trade-offs between autonomy, control, and efficiency.

4.2 Qualitative Results

This section presents the qualitative findings derived from a two-phase series of semi-structured interviews conducted with seven participants (P1–P7) from organizations actively using or evaluating AI-based analytical tools. The analysis followed a reflexive thematic analysis approach, progressing from familiarization and open coding to theme development and refinement.

The two phases served distinct but complementary purposes. Phase 1 (pre-demonstration) explored existing organizational practices, expectations, and concerns regarding AI systems. Phase 2 (during and after system demonstration) examined how these perceptions evolved when participants directly interacted with the proposed multi-agent, staged-autonomy system. Rather than treating the phases as isolated datasets, the analysis emphasizes perceptual shifts and continuities across phases, enabling a deeper understanding of how experiential exposure shapes organizational evaluation of agentic AI.

4.2.1 Overview of the Qualitative Analysis

The qualitative analysis consists of:

- Seven participants (P1–P7) from diverse organizational and professional contexts.

- Phase 1 interviews focusing on current AI usage, governance practices, and perceived limitations.

- Phase 2 interviews conducted during and immediately after live demonstration of the proposed system, including its autonomy stages, agent workflows, and transparency mechanisms.

This longitudinal structure enabled direct comparison between anticipated concerns and experience-based judgments, strengthening the interpretive validity of the findings. The composition of 7 diverse participants ensured a breadth of perspectives on AI-enabled workflows, risk and governance considerations, and practical integration constraints. Experience levels ranged from early-career professionals to a doctoral researcher with prior industry exposure, enabling triangulation across operational, analytical, and strategic lenses.

Data collection followed a two-phase, longitudinal protocol. Phase 1 interviews established current AI usage, governance practices, and perceived limitations. Phase 2 interviews were conducted during and immediately after a live demonstration of the proposed system, covering the 5 autonomy stages, agent workflows, and transparency mechanisms. All 14 interviews were audio recorded and transcribed verbatim, then anonymized for analysis. The paired structure supports direct comparison between anticipated concerns and experience-based judgments, enhancing interpretive validity in the subsequent thematic analysis.

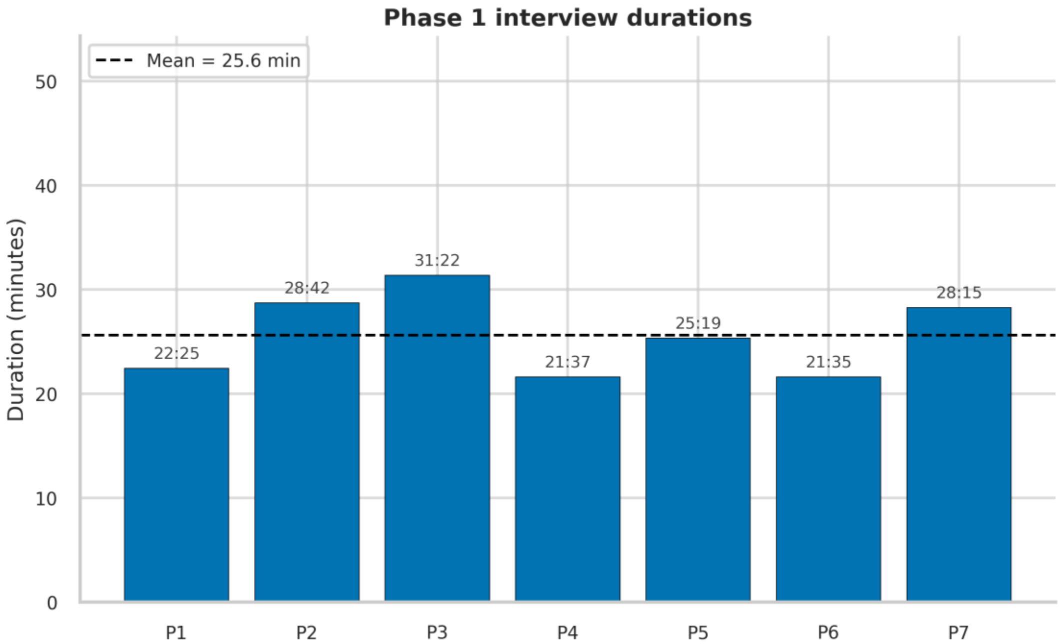

Figure 12 Phase 1 Semi-Structured Interviews' Duration

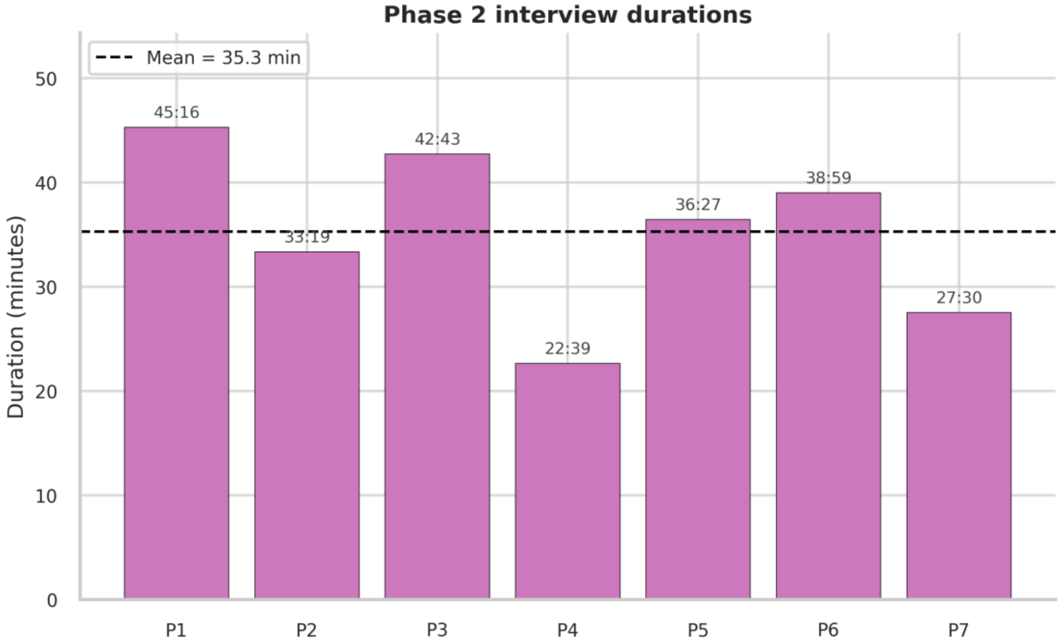

Figure 13 Phase 2 Semi-Structured Interviews' Duration (with Live Demonstration)

Interview duration patterns align with the design intent. Phase 1 averaged 25:36, while Phase 2 averaged 35:16. On average, interviews were 9:40 longer in Phase 2 (+37.7% relative to Phase 1), with 6 of 7 participants speaking longer and one slightly shorter (−0:45).

Table 10 Interview Duration Metrics Summary by Phase

Phase	***n***	***Total Time***	***Mean***	***Median***	***Min***	***Max***	***SD***	***IQR (Q1–Q3)***
Phase 1	7	2:59:15	25:36	25:19	21:35	31:22	3:55	22:01–28:28
Phase 2	7	4:06:53	35:16	36:27	22:39	45:16	8:06	30:24–40:51

The increased mean and wider dispersion in Phase 2 are consistent with deeper probing elicited by the live system demonstration, including think-aloud reactions to autonomy transitions and follow-up questions about explainability, oversight, and integration. Notably, the largest increases were observed for P1 (+22:51) and P6 (+17:24), both engineering-focused roles, reflecting heightened engagement with operationalization and data flow details during the demo.

4.2.2 Phase 1 Findings: Organizational AI Practices and Constraints (Before Demonstration)

Theme 1: AI Is Used Primarily for Efficiency, Not Interpretive Authority

(RQ1: How do organizations currently use AI in analytical workflows?)

In Phase 1, participants consistently described AI as an efficiency-oriented support tool rather than an independent analytical actor. Common applications included summarization, data preprocessing, reporting assistance, and exploratory pattern identification. Across participants, final interpretation, judgment, and decision-making were described as firmly human-led.

Participants emphasized that while AI could accelerate routine tasks, it lacked the contextual sensitivity required for interpretive authority. Several noted that AI outputs were often descriptive rather than explanatory, requiring analysts to manually interrogate results to understand underlying causes or implications. As one participant remarked, AI "helps you get started, but it doesn't finish the thinking for you."

This positioning of AI reflects an organizational norm in which analytical responsibility remains human and AI is framed as a productivity enhancer rather than a collaborator. This framing became a critical reference point in Phase 2, as participants evaluated whether the demonstrated system respected or challenged this boundary.

Theme 2: Trust Is Conditional and Governed by Verification Practices

(RQ2: What factors influence trust in AI-supported analysis?)

Before system exposure, trust in AI was described as conditional, contextual, and continuously negotiated. Participants reported that AI outputs were routinely verified through manual checks, secondary tools, or expert review, especially in high-stakes or regulated contexts. Trust was influenced by several factors: task criticality, data sensitivity, clarity of model behaviour, and perceived consequences of error.

Several participants stressed that AI results were "never taken at face value," particularly when reasoning processes or data provenance were unclear. Trust was therefore less about confidence in AI accuracy and more about confidence in the ability to detect and correct errors. These trust heuristics formed the baseline against which participants later evaluated the system's transparency and control mechanisms during the demonstration.

Theme 3: Governance, Privacy, & Risk Limit Autonomous AI Adoption

(RQ1, RQ4)

Phase 1 interviews revealed strong organizational constraints shaping AI adoption. Participants frequently referenced data protection regulations, internal governance policies, and restrictions on external data processing. Fully autonomous systems were perceived as particularly risky when connected to sensitive datasets or operational processes.

Participants expressed concern that agentic systems capable of acting independently could introduce unintended consequences if not carefully bounded. This led to skepticism toward high-autonomy AI, regardless of its technical sophistication. These concerns became more concrete in Phase 2, as participants evaluated whether the demonstrated system meaningfully addressed governance and accountability requirements.

4.2.3 Phase 2 Findings: Perceptions During and After System Demonstration

Phase 2 interviews were conducted during and immediately after live interaction with the system. Participants observed and interacted with different autonomy stages, agent coordination patterns, and transparency features, allowing them to ground their evaluations in direct experience rather than abstraction.

Theme 4: HITL Control as a Trust-Stabilizing Mechanism

(RQ2: How do different autonomy levels affect trust and acceptance?)

During the demonstration, participants repeatedly returned to the importance of HITL configurations after witnessing stage 3 and 5 runs. While many acknowledged the efficiency potential of autonomous multi-agent operation, they emphasized that approval checkpoints fundamentally changed their comfort level. HITL stages were described as enabling "shared responsibility," where AI could propose actions or insights, but humans retained final authority. This configuration was seen as especially important for complex, ambiguous, or high-risk analytical tasks.

Fully autonomous stages were not rejected outright, but participants framed them as suitable mainly for exploratory or low-impact scenarios. Where Phase 1 skepticism toward autonomy was largely precautionary, Phase 2 evaluations were more differentiated, suggesting that experienced autonomy feels less threatening when paired with visible control mechanisms. Stage 3 and 5 were majorly the most preferred autonomy stages and in some cases the former more than the latter. However, this was also a deductive and described understanding that stage 5 with HITL and autonomous team versus a stage 3 simple agentic AI interaction similar to other prevalent AI tools was hinted to be a comprehensibility overload.

Theme 5: Multi-Agent Collaboration as Both Capability and Cognitive Challenge

(RQ3: How is multi-agent collaboration perceived by users?)

Participants generally perceived multi-agent collaboration as a powerful feature, particularly for tasks involving parallel reasoning or multiple analytical perspectives. Seeing agents specialize and coordinate during the demonstration helped participants conceptualize how complex qualitative tasks could be distributed.

However, this capability also introduced concerns about cognitive overload. Several participants reported difficulty understanding system state when multiple agents were active simultaneously in delegated tasks in parallel. This highlighted a tension between analytical depth and user comprehensibility.

Participants emphasized the importance of orchestration mechanisms, such as managerial agents or summarized outputs, to prevent users from becoming overwhelmed. In Phase 1, multi-agent AI was discussed abstractly; in Phase 2, participants could articulate specific usability challenges tied to real interaction.

Theme 6: Transparency and Explainability as Enablers of Confidence

(RQ2, RQ3)

Transparency features observed during the demonstration, such as agent logs, thinking streams, structured workflows, and traceable outputs, were widely cited as confidence-building. Participants noted that seeing how results were generated made it easier to trust, critique, and refine AI outputs. Transparency was particularly valued when AI behaviour diverged from expectations. Rather than undermining trust, visible reasoning allowed participants to diagnose issues and remain engaged with the system. What Phase 1 participants described abstractly as "needing transparency" became, in Phase 2, a concrete requirement tied to specific UI and architectural features.

Theme 7: Adoption Depends on Usability, Training, & Organizational Fit

(RQ4: What factors influence organizational adoption of agentic AI systems?)

Despite generally positive reactions, participants identified several barriers to adoption during Phase 2. These included the learning curve associated with advanced autonomy stages, the need for user training, and challenges integrating the system into existing workflows.

Participants emphasized that adoption would require gradual rollout, clear onboarding, and alignment with established organizational roles. The system was viewed as promising, but only if introduced incrementally and with sufficient support or training in organizational contexts. Phase 1 concerns about adoption were speculative, whereas Phase 2 discussions were grounded in practical interaction experiences.

Table 11 RQ Based Qualitative Themes & Insights

Research Question (RQ)	*Qualitative Themes*	*Phase 1 Evidence (Before Demo)*	*Phase 2 Evidence (During Demo)*
***RQ1**: What architectural requirements enable effective knowledge graph traversal alongside distributed agentic flows for organizational tasks?*	**Theme 4:** HITL Control as a Trust-Stabilizing Mechanism **Theme 5:** Multi-Agent Collaboration as Capability and Cognitive Challenge **Theme 6:** Transparency and Explainability as Enablers of Confidence	Architecture discussed implicitly; participants focused on outcomes rather than system structure. Limited articulation of architectural needs beyond “reliability” and “control.”	Demonstration enabled participants to explicitly identify architectural needs such as visible agent roles, coordinated workflows, traceable reasoning, and transparent knowledge traversal as prerequisites for usability and trust.
***RQ2**: How do organizations perceive agentic AI-based systems and how proactively are they adopting latest AI technologies?*	**Theme 1:** AI Used for Efficiency, Not Interpretive Authority **Theme 2:** Trust Is Conditional and Governed by Verification Practices	Organizations perceived agentic AI as experimental and risky. Adoption described as cautious, fragmented, and constrained by governance and risk considerations.	Perceptions became more nuanced; agentic AI was viewed as feasible under staged autonomy. Adoption was seen as possible through incremental rollout and HITL configurations.
***RQ3**: How do organizations or roles measure success of agentic AI systems?*	**Theme 2:** Trust and Verification Practices **Theme 6:**	Success loosely defined in terms of efficiency and output usefulness, with no formal success criteria for agentic systems.	Success reconceptualized as multi-dimensional, including transparency, interpretability, decision confidence, and reduced

	Transparency and Explainability		cognitive workload alongside efficiency.
***RQ4**: What practical constraints emerge when deploying such systems in real organizational settings, particularly regarding data, security, accuracy, and other concerns?*	**Theme 3:** Governance, Privacy, and Risk Limit Adoption **Theme 7:** Adoption Depends on Usability, Training, and Organizational Fit	Constraints discussed abstractly, focusing on compliance, data security, and organizational risk. Fully autonomous systems viewed as problematic.	Constraints articulated more concretely during demonstration, including learning curve, integration effort, need for training, and governance mechanisms to manage autonomy safely.

4.2.4 Integrated Discussion Across Phases

Across both phases, a clear pattern emerges organizational perceptions of agentic AI evolve significantly through experiential engagement. Phase 1 interviews reveal cautious, governance-driven attitudes shaped by abstract risk considerations. Phase 2 interactions transform these concerns into more nuanced evaluations, where design choices, such as staged autonomy, transparency, and HITL control, mediate trust and acceptance. The demonstration acted as a sense-making mechanism, enabling participants to articulate requirements and constraints with greater specificity.

4.3 Key Takeaways

This chapter reported the outputs of the empirical analysis of the suggested multi-agent using GraphRAG-based analytical system in accordance with the DSRM in Chapter 3. The assessment was made up of qualitative thematic analysis of practitioner interviews and descriptive quantitative measures of system example conduct under different levels of staged autonomy where a multi-level view of system performance is coupled with an organizational perception.

Based on the qualitative findings of two phased semi-structured interviews on seven respondents, it became evident that the current use of AI in organizations is not an autonomous decision-maker but a support (efficiency-enhancing) tool. As before system demonstration, the respondents highlighted the importance of governance, a calibration of trust, and human interpretive authority as the core constraints defining the adoption of AI. After experiencing interconnectedness with the proposed system, respondents registered higher regard to the analytical capacity of multi-agent interaction and at the same time solidified the necessity of HITL controlling, transparency, and usability. In both stages, trust proved conditioned and contingent depending on clarification, verification behaviours, as well as that which conformed to organizational processes.

The quantitative findings offered a complementary viewpoint of operations which defined the way behaviour of the system scales over five stages of autonomy and four datasets. The analysis revealed a definite monotonic correlation between the level of autonomy and the system workload: more advanced stages of autonomy included more closing procedures, more interaction processes and more enriched artifact findings. Introducing the use of tools, a multi-agent environment, GraphRAG capabilities, and HITL checkpoints all introduced quantifiable overhead, especially with fully autonomous and multi-agent environments. Simultaneously, the wider and more structured the analytical output in terms of increased analytical depth was the same time created by higher autonomy stages, not inefficiency per se.

Notably, the quantitative results empirically supported several themes discovered in the literature analysis. The measured HITL overhead was in line with how participants rated control costs and there was an increase in artifact production in multi-agent stages and this matched perceived analytical richness and concerned cognitive load in some instances. The complexities on dataset-level also demonstrated the interaction between real-world data complexity and architectural and autonomy decisions.

Combined, the findings show that staged autonomy offers a feasible and organizationally independent channel towards the implementation of agentic AI systems. Instead of preferring a fully autonomous implementation, the results obtained indicate that graduated autonomy, where implementation is facilitated by transparency, human supervision, and orchestration, is more consistent with organizational trust, governance, and adoption needs.

It is thus a robust empirical basis of the next chapter of the discussion, which puts the findings into perspective in terms of existing literature, theoretical implications of human-AI collaboration, and advice in the design and implementation of agentic AI systems in organizations.

5 Discussion & Conclusion

This thesis aimed to explore the effectiveness of agentic AI systems, especially those involving knowledge graph traversal and distributed multi-agent processes, to be designed and implemented successfully in actual organization situations. The research also studied the human, organizational and governance factors that influence the adoption of staged-autonomy analytical systems through the evaluation and development of a staged-autonomy analytical system. Combining architectural design, controlled experimentation, and two-phase qualitative inquiry, the research offers a comprehensive idea of what makes it possible to transition agentic AI between the promise as a concept and the usefulness as an organization component.

Architecturally, the observations show that to achieve successful agentic AI systems, one needs more than autonomy or even computing power. The system of structured orchestration, and role-specialized agents, plus the direct visibility of execution trajectories, proved to be the key aspect of allowing meaningful knowledge extraction graphs and coordinated streams of analysis. Staging nature was essential in demonstrating the effects of capability expansion, in terms of reactive chat interaction to autonomous and HITL multi-agent teams, on system behaviour in quantifiable manners. Instead of viewing autonomy as a dichotomous characteristic, the thesis issues a spectrum of autonomy, with each phase having different trade-offs of speed, analytic profoundness, coordination cost, and governance control.

The quantitative findings are a clear indication that during runtime, the number of interaction steps, and the number of artifacts significantly increase with the increment in autonomy and agentic complexity. Stages that used tools and varied in agents added analytical outputs such as multiple reports, tables, and visualizations with increased costs of computation and coordination. Even human-in-the-middle checkpoints that brought no corrective human action were showing perceivable overhead, and thus this may stress that monitoring is not an insignificant element of system behaviour. These results have placed quantitative performance measures not as the sole measures of efficiency, but as expressions of workload, mental effort, and organizational cost of the various combinations of autonomy.

These measurements are supplemented by the qualitative results that offer background to the interpretation of agentic AI behaviour and its response in organizations. Before exposure to the system, people mostly perceived AI as a support system with a high efficiency score, instead of an independent analysis system that would make decisions and trust was closely related to the verification practice, requirements of the governance, and data sensitivity. After the demonstration of systems, the opinions shifted to cautious optimism especially with the combination of autonomy that is accompanied by transparency and human control. The perception of multi-agent collaboration power was very strong, but maybe it was so strong that one could not easily be able to fend it off in the absence of clear coordination, summarization, and interface support. In both stages, the concept of trust was always a measured and contingent attribute, and not a natural attribute of automation.

Put together, both the qualitative and quantitative findings lead to one general conclusion: the effectiveness of organizational adoption of agentic AI will similarly rely less on the optimization of autonomy and more on providing the behaviour of systems with human expectations, institutional limitations, and already existing work processes. The mechanisms of transparency that draw a balance between innovation and accountability in an organization include staged autonomy, HITL control, and transparent execution processes with each of this having an additional layer of UI/UX applicability. In this thesis, what the system under consideration has shown is that agentic AI can be potentially useful as an addition to analytical work, if it is designed to not usurp human interpretive authority, facilitate a gradual adoption process, or reveal its internal mechanisms in a manner that makes sense intuitively.

The research adds value to the research area by empirically describing the behaviour of agentic AI systems in terms of the level of autonomy followed and in terms of basing such behaviour in actual organizational perceptions. It contributes to the existing knowledge in terms of demonstrating how architectural choices, interaction design, and governance issues are intertwined in their effect on system performance and user trust. Although the quantitative results are designed to be descriptive in nature, with single-run observations, they can form the basis of future benchmarking at a large scale, repeated testing, and evaluation metrics that are quality-driven. This study can be further developed into the future with the addition of longitudinal research in organizations, more extensive sample of participants, and more intensive evaluation of quality of output and quality of decisions made with added scrutiny by human oversight and validation methods.

To sum up, these arguments prove that agentic AI systems will be most feasible not as completely autonomous substitutes to replace human analysis, but as formal, transparent, and manageable partners. Organizations may slowly adopt to use staged autonomy and socio-technical alignment to benefit the analytical power of agentic AI and ensure trust, accountability, and operational integrity.

6 References

Abdul Rahman, H., Masri, N.A., Husaini, A., Shaharuddin, M.Y., Zaim, S.N.N., 2025. Integration of Machine Learning and Artificial Intelligence Techniques for Qualitative Research: The Rise of New Research Paradigms. Int J Qual Methods 24. https://doi.org/10.1177/16094069251347545

Abram, M.D., Mancini, K.T., Parker, R.D., 2020. Methods to Integrate Natural Language Processing Into Qualitative Research. Int J Qual Methods 19. https://doi.org/10.1177/1609406920984608

Abu-Salih, B., Alotaibi, S., 2024. A systematic literature review of knowledge graph construction and application in education. Heliyon 10, e25383. https://doi.org/10.1016/J.HELIYON.2024.E25383

Amaan, A., 2025. Agentic-AI-Environment [WWW Document]. URL https://github.com/ali-amaan/Agentic-AI-Environment (accessed 31.12.25).

Ampel, B.M., Ullman, S., 2025. Multi-Agent Systems for Information Systems Research: A Framework for Collaborative AI-Augmented Inquiry, in: Proceedings of the International Conference on Information Systems (ICIS) TREOS.

Bhojani, J., Pitroda, K., 2025. Airline Reviews [WWW Document]. URL https://www.kaggle.com/datasets/juhibhojani/airline-reviews (accessed 11.1.25).

Bilal, W., 2025. Spotify Global Music Dataset (2009–2025) [WWW Document]. URL https://www.kaggle.com/datasets/wardabilal/spotify-global-music-dataset-20092025 (accessed 11.1.25).

Binyamin, S.S., Ben Slama, S., 2022. Multi-Agent Systems for Resource Allocation and Scheduling in a Smart Grid. Sensors 22, 8099. https://doi.org/10.3390/s22218099

Boaventura, A., 2025. Build a Multi-Agent System with LangGraph and Mistral on AWS | Artificial Intelligence [WWW Document]. URL https://aws.amazon.com/blogs/machine-learning/build-a-multi-agent-system-with-langgraph-and-mistral-on-aws/ (accessed 11.2.25).

Brailas, A., 2025. Artificial Intelligence in Qualitative Research: Beyond Outsourcing Data Analysis to the Machine. Psychology International 7, 78. https://doi.org/10.3390/psycholint7030078

Chen, J., Chen, Yingguo, Pham, D.T., Song, Y., Wu, J., Xing, L., Chen, Yingwu, 2025. A Large Language Model-based Multi-Agent Framework to Autonomously Design Algorithms for Earth Observation Satellite Scheduling Problem. Engineering. https://doi.org/10.1016/j.eng.2025.10.027

Clemedtson, A., Shi, B., 2025. GraphRAFT: Retrieval Augmented Fine-Tuning for Knowledge Graphs on Graph Databases.

Cuevas, A., Scurrell, J. V., Brown, E.M., Entenmann, J., Daepp, M.I.G., 2024. Collecting Qualitative Data at Scale with Large Language Models: A Case Study 1.

de Martim, H., 2025. An Ontology-Driven Graph RAG for Legal Norms: A Structural, Temporal, and Deterministic Approach.
Dehal, R.S., Sharma, M., Rajabi, E., 2025. Knowledge Graphs and Their Reciprocal Relationship with Large Language Models. Mach Learn Knowl Extr 7, 38. https://doi.org/10.3390/make7020038
Design Council, 2005. The Double Diamond - Design Council [WWW Document]. URL https://www.designcouncil.org.uk/our-resources/the-double-diamond/ (accessed 11.2.25).
Dutta, D., Kannan Poyil, A., 2024. The machine/human agentic impact on practices in learning and development: a study across MSME, NGO and MNC organizations. Personnel Review 53, 791–815. https://doi.org/10.1108/PR-09-2022-0658
Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R.O., Larson, J., 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization.
Ferrari, I., Frisoni, G., Italiani, P., Moro, G., Sartori, C., 2022. Comprehensive Analysis of Knowledge Graph Embedding Techniques Benchmarked on Link Prediction. Electronics (Basel) 11, 3866. https://doi.org/10.3390/electronics11233866
González-Cancelas, N., Vaca-Cabrero, J., Camarero-Orive, A., 2025. Multi-Agent System for Smart Roll-on/Roll-off Terminal Management: Orchestration and Communication Strategies for AI-Driven Optimization. Applied Sciences 15, 6079. https://doi.org/10.3390/app15116079
Goonatilleke, S.T., Hettige, B., 2022. Past, Present and Future Trends in Multi-Agent System Technology. Journal Européen des Systèmes Automatisés 55, 723–739. https://doi.org/10.18280/jesa.550604
Goyal, N., Chang, M., Terry, M., 2024. Designing for Human-Agent Alignment: Understanding what humans want from their agents, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1–6. https://doi.org/10.1145/3613905.3650948
Guo, K., Shomer, H., Zeng, S., Han, H., Wang, Y., Tang, J., 2025. Empowering GraphRAG with Knowledge Filtering and Integration 25450–25464. https://doi.org/10.18653/v1/2025.emnlp-main.1293
Gusarov, A., Volkova, A., Khrulkov, V., Kuznetsov, A., Maslov, E., Oseledets, I., 2025. Multi-Agent GraphRAG: A Text-to-Cypher Framework for Labeled Property Graphs.
Han, H., Ma, L., Shomer, H., Wang, Y., Lei, Y., Guo, K., Hua, Z., Long, B., Liu, H., Aggarwal, C.C., Tang, J., 2025. RAG vs. GraphRAG: A Systematic Evaluation and Key Insights.
Han, H., Wang, Y., Shomer, H., Guo, K., Ding, J., Lei, Y., Halappanavar, M., Rossi, R.A., Mukherjee, S., Tang, X., He, Q., Hua, Z., Long, B., Zhao, T., Shah, N., Javari, A., Xia, Y., Tang, J., 2024. Retrieval-Augmented Generation with Graphs (GraphRAG).
Heeg, R., 2023. Possibilities and limitations, of unstructured data - Research World [WWW Document]. URL

https://researchworld.com/articles/possibilities-and-limitations-of-unstructured-data (accessed 11.28.25).
Herhausen, D., Ludwig, S., Abedin, E., Haque, N.U., de Jong, D., 2025. From words to insights: Text analysis in business research. J Bus Res 198, 115491. https://doi.org/10.1016/j.jbusres.2025.115491
Herrera, M., Pérez-Hernández, M., Kumar Parlikad, A., Izquierdo, J., 2020. Multi-Agent Systems and Complex Networks: Review and Applications in Systems Engineering. Processes 8, 312. https://doi.org/10.3390/pr8030312
Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J.E.L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C.N., Rashid, S.M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A., 2021. Knowledge Graphs. ACM Comput Surv 54. https://doi.org/10.1145/3447772
Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., Zhao, L., 2024. GRAG: Graph Retrieval-Augmented Generation 4145–4157. https://doi.org/10.18653/v1/2025.findings-naacl.232
Huseynli, M., Bub, U., Ogbuachi, M.C., 2022. Development of a Method for the Engineering of Digital Innovation Using Design Science Research. Information 13, 573. https://doi.org/10.3390/info13120573
Ismaeel, A.Z., Zebari, I.M.I., 2025. Comparing Traversal Strategies: Depth-first Search vs. Breadth-first Search in Complex Networks. Asian Journal of Research in Computer Science 18, 60–73. https://doi.org/10.9734/ajrcos/2025/v18i2562
Jain, R., Kumar, A., 2024. Artificial Intelligence in Marketing: Two Decades Review. NMIMS Management Review 32, 75–83. https://doi.org/10.1177/09711023241272308
Jun, C., 2025. AI-Driven Human-Computer Interaction and Intelligent Agents: Impact on New Quality Productive Forces. ResearchGate. https://doi.org/10.13140/RG.2.2.33318.79688
Kalanda, B.F., Cheboi, A.J., 2025. Artificial Intelligence in the Analysis of Unstructured Qualitative Data: A Literature Review. Adv Soc Sci Res J 12, 199–205. https://doi.org/10.14738/assrj.1208.19286
Karkavelraja, J., 2025. Amazon Sales Dataset [WWW Document]. URL https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset (accessed 11.1.25).
Kirsten, E., Buckmann, A., Lassak, L., Borgert, N., Mhaidli, A., Becker, S., 2025. From Assistance to Autonomy -- A Researcher Study on the Potential of AI Support for Qualitative Data Analysis.
Klén, E., 2024. Artificial intelligence in customer service Frameworks and tools for identifying prominent implementation avenues: Benefits, best practices, and overcoming barriers.
Knollmeyer, S., Caymazer, O., Grossmann, D., Knollmeyer, S., Caymazer, O., Grossmann, D., 2025. Document GraphRAG: Knowledge Graph Enhanced Retrieval Augmented Generation for Document Question Answering Within the Manufacturing Domain. Electronics 2025, Vol. 14, 14. https://doi.org/10.3390/ELECTRONICS14112102

Lennon, R.P., Fraleigh, R., van Scoy, L.J., Keshaviah, A., Hu, X.C., Snyder, B.L., Miller, E.L., Calo, W.A., Zgierska, A.E., Griffin, C., 2021. Developing and testing an automated qualitative assistant (AQUA) to support qualitative analysis. Fam Med Community Health 9, 1287. https://doi.org/10.1136/FMCH-2021-001287
Lim, W.M., 2025. What Is Qualitative Research? An Overview and Guidelines. Australasian Marketing Journal 33, 199–229. https://doi.org/10.1177/14413582241264619
Lin, Y., Lin, F., Cai, G., Chen, H., Zou, L., Liu, Y., Wu, P., 2025. Evolutionary Reinforcement Learning: A Systematic Review and Future Directions. Mathematics 13, 833. https://doi.org/10.3390/math13050833
Liu, J., Wang, K., Chen, Y., Peng, X., Chen, Z., Zhang, L., Lou, Y., 2024. Large Language Model-Based Agents for Software Engineering: A Survey.
Liu, R., Jiang, H., Yan, X., Tang, B., Li, J., 2025. PolyG: Adaptive Graph Traversal for Diverse GraphRAG Questions. arxiv.org.
Liu, Y., Lo, S.K., Lu, Q., Zhu, L., Zhao, D., Xu, X., Harrer, S., Whittle, J., 2025. Agent design pattern catalogue: A collection of architectural patterns for foundation model based agents. Journal of Systems and Software 220, 112278. https://doi.org/10.1016/j.jss.2024.112278
Luo, Z., Li, D., Wan, J., Wang, S., Wang, G., Cheng, M., Li, T., 2025. Multi-agent collaboration mechanisms based on distributed online meta-learning for mass personalization. J Ind Inf Integr 46, 100852. https://doi.org/10.1016/j.jii.2025.100852
Martinez-Gil, F., Gil-Magraner, E., 2025. Multi-Agent Reinforcement Learning for creating intelligent agents in social networks-oriented role playing games. Entertain Comput 54, 100941. https://doi.org/10.1016/j.entcom.2025.100941
Mitchell, A., 2021. Multi-Modal AI Models: Combining Text, Vision, and Speech. SSRN Electronic Journal 21, 789–797.
Mohan Singh, 2025. Multi-agent systems: the future of distributed AI platforms for complex task management. World Journal of Advanced Research and Reviews 26, 048–055. https://doi.org/10.30574/wjarr.2025.26.3.1985
Murugesan, U., Subramanian, P., Srivastava, S., Dwivedi, A., 2023. A study of Artificial Intelligence impacts on Human Resource Digitalization in Industry 4.0. Decision Analytics Journal 7, 100249. https://doi.org/10.1016/j.dajour.2023.100249
Naganawa, H., Hirata, E., Yamada, A., Naganawa, H., Hirata, E., Yamada, A., 2025. Implementing a Knowledge Management System with GraphRAG: A Physical Internet Example. Electronics 2025, Vol. 14, 14. https://doi.org/10.3390/ELECTRONICS14244948
Naumov, V., Zagirova, D., Lin, S., Xie, Y., Gou, W., Urban, A., Tikhonova, N., Alawi, K., Durymanov, M., Galkin, F., Chen, S., Sidorenko, D., Korzinkin, M., Scheibye-Knudsen, M., Aspuru-Guzik, A., Izumchenko, E., Gennert, D., Pun, F.W., Zhang, M., Kamya, P., Aliper, A., Ren, F., Zhavoronkov, A., 2025. DORA AI Scientist: Multi-agent Virtual

Research Team for Scientific Exploration Discovery and Automated Report Generation. bioRxiv 2025.03.06.641840. https://doi.org/10.1101/2025.03.06.641840
Nicmanis, M., Spurrier, H., 2025. Getting Started with Artificial Intelligence Assisted Qualitative Analysis: An Introductory Guide to Qualitative Research Approaches with Exploratory Examples from Reflexive Content Analysis. International Journal of Qualitative Methods 24. https://doi.org/10.1177/16094069251354863;REQUESTEDJOURNAL:JOURNAL:IJQA;WEBSITE:WEBSITE:SAGE;WGROUP:STRING:PUBLICATION
OpenAI, 2025. Temporal Agents with Knowledge Graphs [WWW Document]. URL https://cookbook.openai.com/examples/partners/temporal_agents_with_knowledge_graphs/temporal_agents (accessed 11.26.25).
Papageorgiou, G., Sarlis, V., Maragoudakis, M., Tjortjis, C., 2025. Hybrid Multi-Agent GraphRAG for E-Government: Towards a Trustworthy AI Assistant. Applied Sciences 15, 6315. https://doi.org/10.3390/app15116315
Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., Tang, S., 2026. Graph Retrieval-Augmented Generation: A Survey. ACM Trans Inf Syst 44, 1–52. https://doi.org/10.1145/3777378
Prakash Jadhav, S., 2022. Graph Traversal and its Applications. International Journal of Computer Science and Mobile Applications 10, 1–7. https://doi.org/10.5281/zenodo.6838779
Rashid, A. Bin, Kausik, M.A.K., 2024. AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances 7, 100277. https://doi.org/10.1016/j.hybadv.2024.100277
Roll, D.S., Kurt, Z., Li, Y., Woo, W.L., 2025. Augmenting Orbital Debris Identification with Neo4j-Enabled Graph-Based Retrieval-Augmented Generation for Multimodal Large Language Models. Sensors 25, 3352. https://doi.org/10.3390/s25113352
Sarmah, B., Hall, B., Rao, R., Patel, S., Pasquali, S., Mehta, D., 2024. HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction. Proceedings of ACM Conference (Conference'17) 1.
Tang, Y., Yang, Y., Kong, H., 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries.
vom Brocke, J., Hevner, A., Maedche, A., 2020. Introduction to Design Science Research 1–13. https://doi.org/10.1007/978-3-030-46781-4_1
Wang, Junlin, Wang, Jue, Athiwaratkun, B., Zhang, C., Zou, J., 2024. Mixture-of-Agents Enhances Large Language Model Capabilities. 13th International Conference on Learning Representations, ICLR 2025 40752–40771.
Wei, X., Dong, Y., Wang, X., Zhang, X., Zhao, Z., Shen, D., Xia, L., Yin, D., 2025. Beyond ReAct: A Planner-Centric Framework for Complex Tool-Augmented LLM Reasoning.

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., Grau, V., 2024. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation 1–12.
Xiang, Z., Wu, C., Zhang, Q., Chen, S., Hong, Z., Huang, X., Su, J., 2025. When to use Graphs in RAG: A Comprehensive Analysis for Graph Retrieval-Augmented Generation.
Xiao, Y., Dong, J., Zhou, C., Dong, S., Zhang, Q., Yin, D., Sun, X., Huang, X., 2025. GraphRAG-Bench: Challenging Domain-Specific Reasoning for Evaluating Graph Retrieval-Augmented Generation.
Xiao, Y., Sun, E., Luo, D., Wang, W., 2024. TradingAgents: Multi-Agents LLM Financial Trading Framework.
Xu, Q., Amjad, N., Giles, G., Cumming, A., Hermesky, D., Wen, A., Kwak, M.J., Kim, Y., 2025. A Multi-Agent Large Language Model Framework for Automated Qualitative Analysis.
Yang, W., Some, L., Bain, M., Kang, B., 2025. A comprehensive survey on integrating large language models with knowledge-based methods. Knowl Based Syst 318, 113503. https://doi.org/10.1016/j.knosys.2025.113503
Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y., 2022. ReAct: Synergizing Reasoning and Acting in Language Models. 11th International Conference on Learning Representations, ICLR 2023.
Yu, Y., Yao, Z., Li, H., Deng, Z., Jiang, Y., Cao, Y., Chen, Z., Suchow, J.W., Cui, Z., Liu, R., Xu, Z., Zhang, D., Subbalakshmi, K., Xiong, G., He, Y., Huang, J., Li, D., Xie, Q., 2024. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. Adv Neural Inf Process Syst 37. https://doi.org/10.52202/079017-4354
Zhang, H., Wu, C., Xie, J., Lyu, Y., Cai, J., Carroll, J.M., 2025. Harnessing the power of AI in qualitative research: Exploring, using and redesigning ChatGPT. Computers in Human Behavior: Artificial Humans 4, 100144. https://doi.org/10.1016/J.CHBAH.2025.100144
Zhang, Q., Chen, S., Bei, Y., Yuan, Z., Zhou, H., Hong, Z., Chen, H., Xiao, Y., Zhou, C., Dong, J., Chang, Y., Huang, X., 2025. A Survey of Graph Retrieval-Augmented Generation for Customized Large Language Models.

Appendix A

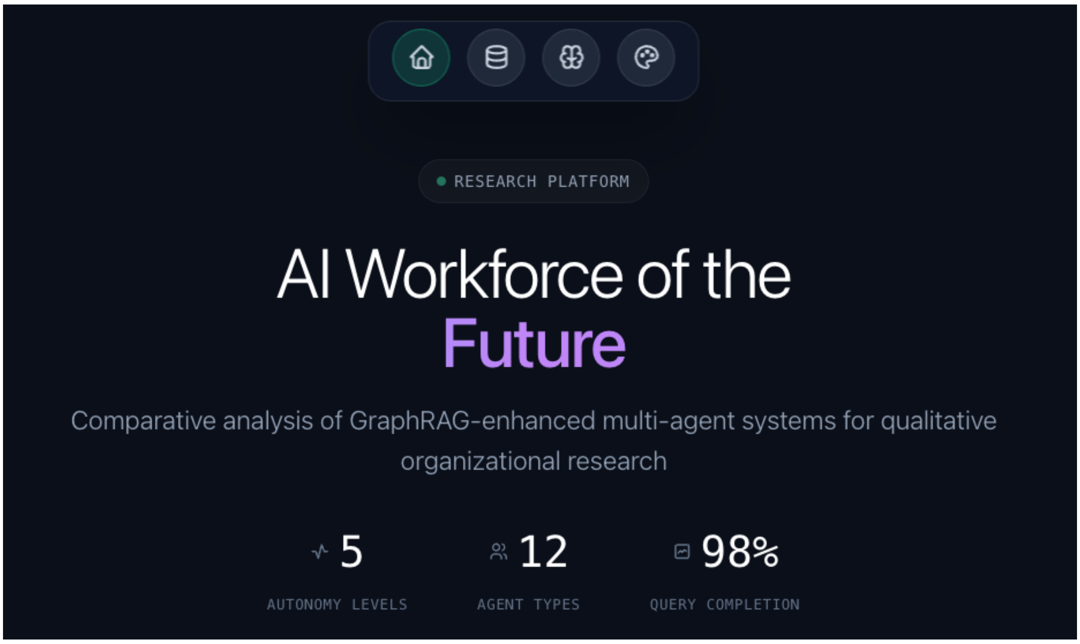

Figure 14 Home of Agentic AI Environment (Core Research Artifact)

Codebase available at: https://github.com/ali-amaan/Agentic-AI-Environment (Amaan, 2025)

Metric Category	*Value*	*Description & Qualitative Impact*
Development	**6 Months**	Iterative development cycle following the DSR framework for problem identification and artifact design.
Research Foundation	**7 Interviews**	Qualitative data used to guide the interpretive requirements and HITL interface design.
System Flexibility	**5 Autonomy Stages**	Levels of agentic control, ranging from reactive prompting to multi-agent team orchestration.
UI/UX Complexity	**34 UI Components**	Modular, reusable frontend elements designed for transparency and human interpretability.
System Connectivity	**24 API Endpoints**	Integrated backend services connecting the agentic reasoning layer to the data management layer.
Software Scale	**29,137 LOC (App)** (AI Assisted Development)	Total Lines of Code (Frontend and Backend), excluding dependencies, representing the custom build.
Package Ecosystem	**50 Unique Packages**	Integrated libraries spanning LLM orchestration, data processing, and graph database management.

Appendix B

Table 12 Select UI/UX Artifacts & Design Rationale

UI/UX Artifact Name	Description	UI/UX Rationale & Research Alignment (Section 4.3.2)
Data Page & Upload Component	A centralized library for dataset management (CSV, JSON, Excel) with metadata preview (Name, Description, Size).	**Objective V:** Enhances organizational feasibility by providing a structured data management layer. The "Loaded" status provides immediate system feedback.
Dataset Preview Modal	Detailed view of the loaded dataset, showing row/column counts, data size, file format, and a scrollable sample data table.	**Objective IV:** Ensures accountability and transparency. Allows researchers to verify data integrity and schema alignment before initiating AI analysis.
Reasoning Depth Selector	A configuration menu allowing users to toggle between "Quick," "Balanced," and "Deep" reasoning modes.	**Objective II:** Directly implements staged autonomy. It allows the user to control the computational intensity and "thoughtfulness" of the agentic process.
Team Selection Modal	A modular interface to assemble a specialist team (Manager, Researcher, Analyst, etc.) with quantified skill bars.	**Objective I:** Visualizes the Multi-Agent system architecture. The skill bars and "Agent Role" cards define the operational boundaries of each specialized agent.
Basic Agentic Environment	A clean, "Reactive" (Stage 1) state of the workspace with a command bar and system connectivity status.	**Objective II:** Represents the baseline of the autonomy spectrum. The "Suggestions" chips reduce cognitive load and guide the start of the qualitative inquiry.
Runtime Agent Execution (Activity Log)	Real-time execution view showing "Thinking" steps, active code blocks, and a live side-panel activity log.	**Objective IV:** Provides a granular audit trail of AI reasoning. By surfacing "Thinking" steps, the system moves away from "black box" AI toward interpretive transparency.
Human-in-the-Loop (HITL) Component	An expanded decision-support interface offering multiple analysis options (e.g., Option 1 vs. Option 2) with risk assessments.	**Objective III:** The core of the HITL system. It empowers the human to steer the agent's trajectory, maintaining interpretive power over the final strategic output.
Multi-Agent Team Execution	High-complexity view (Stage 5) showing a Manager agent delegating tasks to multiple specialists (Analyst, Writer, etc.).	**Objective I & II:** Demonstrates full orchestration. The UI manages the complexity of parallel workflows while keeping the user informed of who is "Working" vs. "Thinking."
Detailed Agent Activity Log	A vertical timeline of system events, API requests, and stream starts with precise timestamps and trace toggles.	**Objective IV:** Accountability and reproducibility. Essential for researchers to debug the agentic flow and audit the "chain of thought" for qualitative rigor.

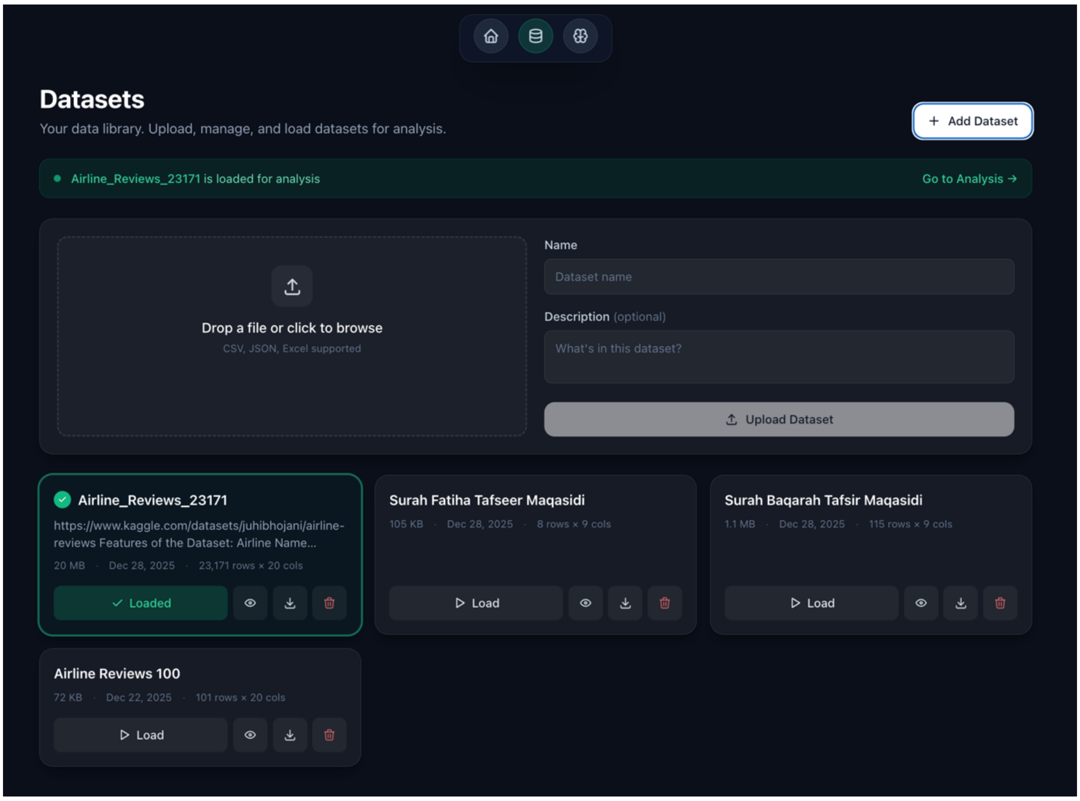

Figure 15 Data Page & Upload Component

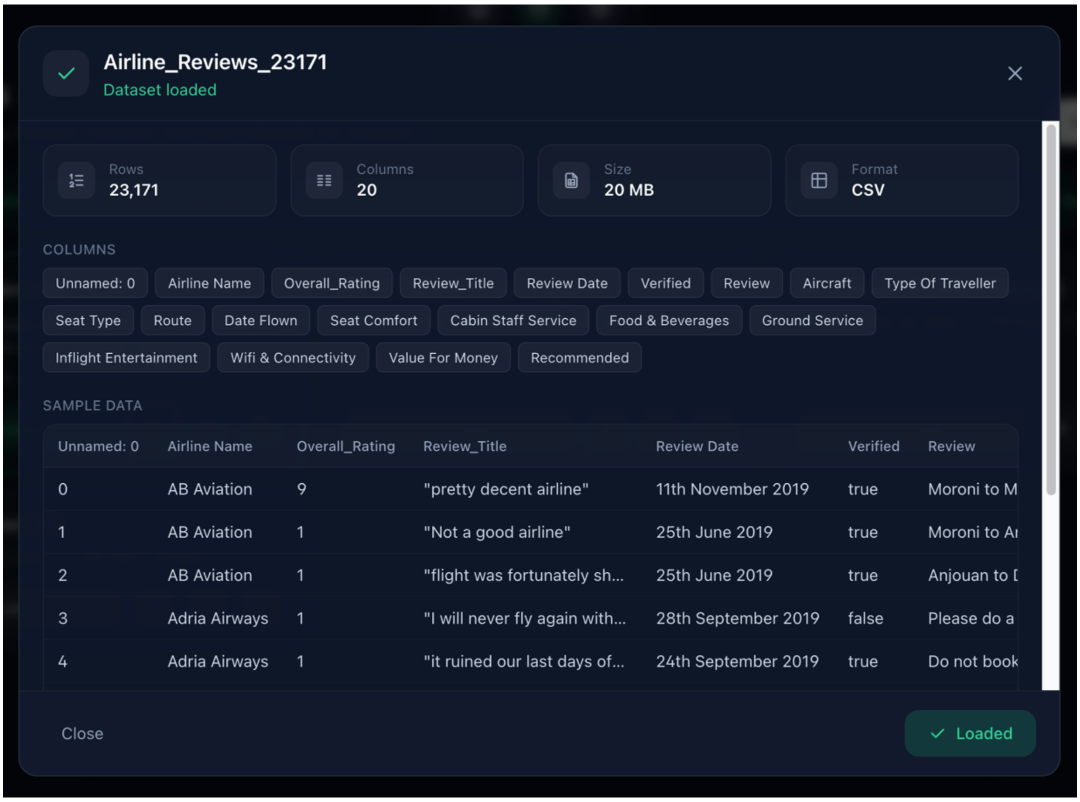

Figure 16 Dataset Preview Modal

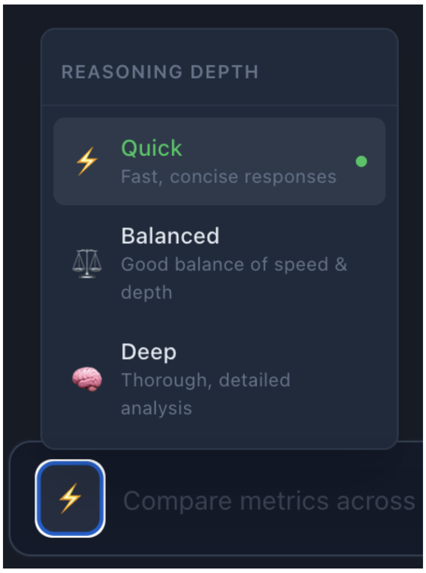

Figure 17 Reasoning Depth Selector

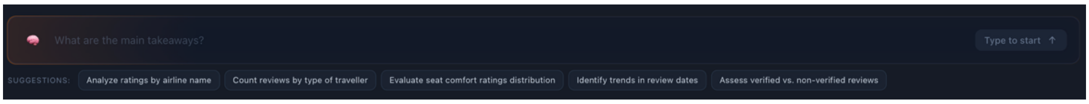

Figure 18 Input Bar Component

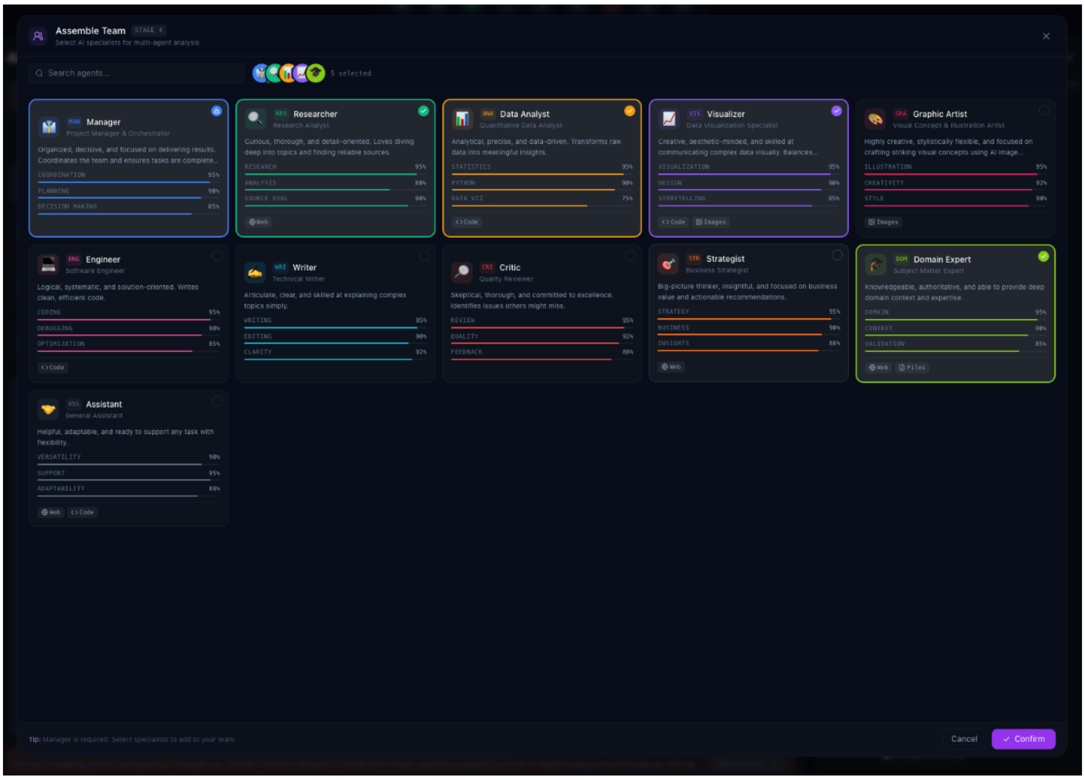

Figure 19 Reasoning Depth Selector

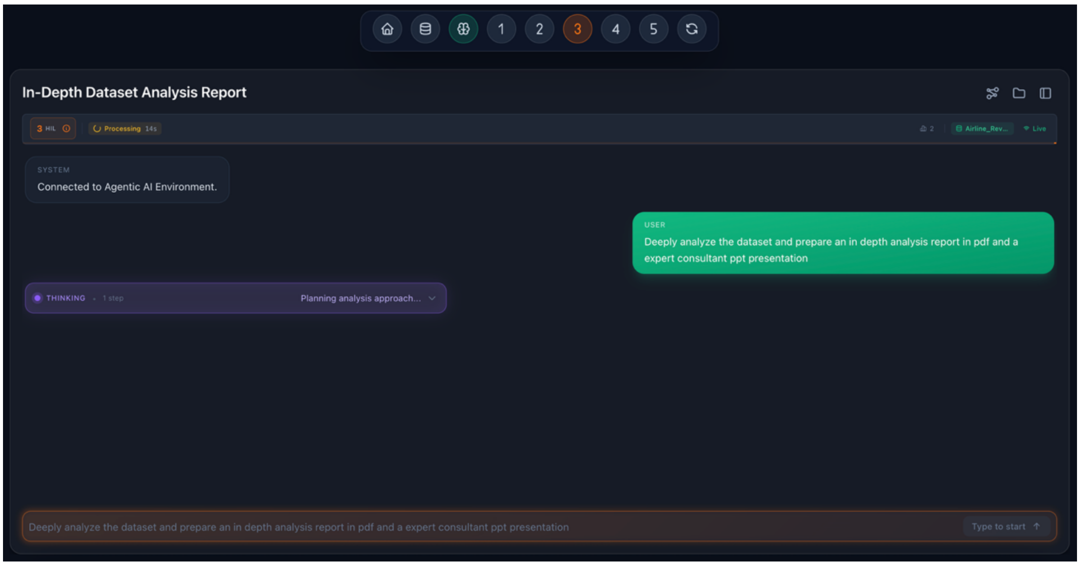

Figure 20 Basic Agentic Environment (Stage 3 Running)

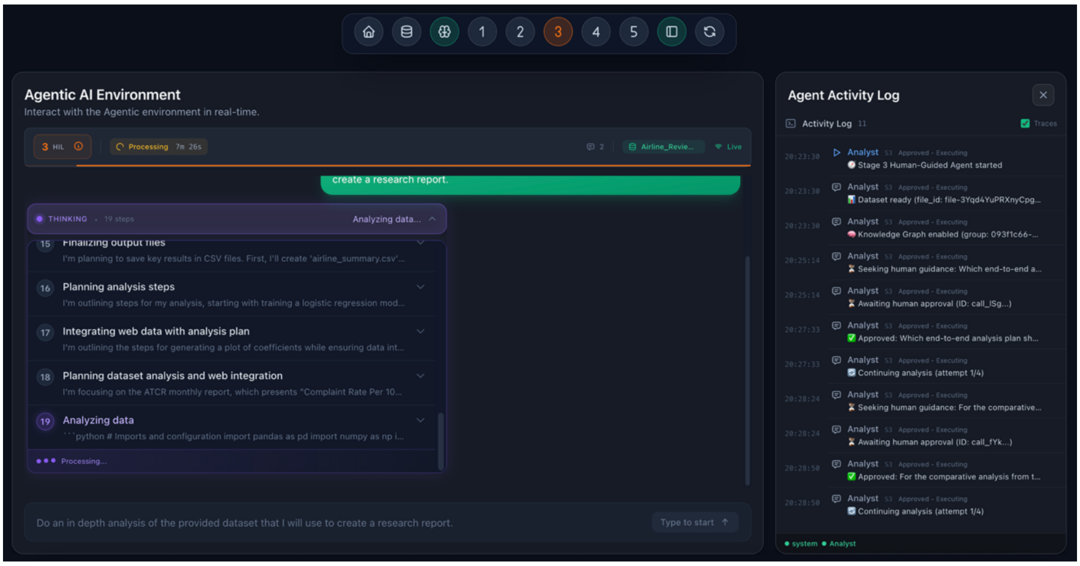

Figure 21 Runtime Agent Execution (Activity Log) (Stage 3 Detailed Running)

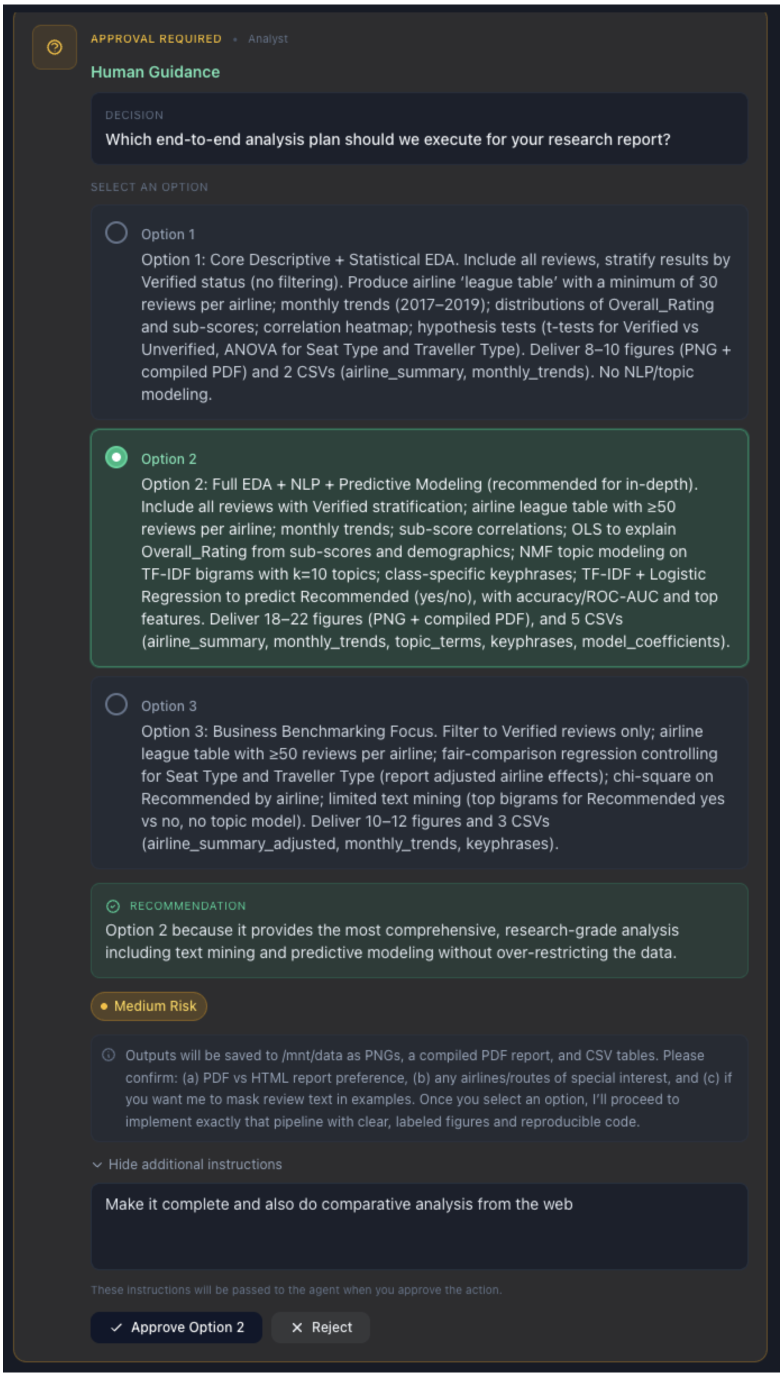

Figure 22 Human-in-the-Loop (HITL) Component

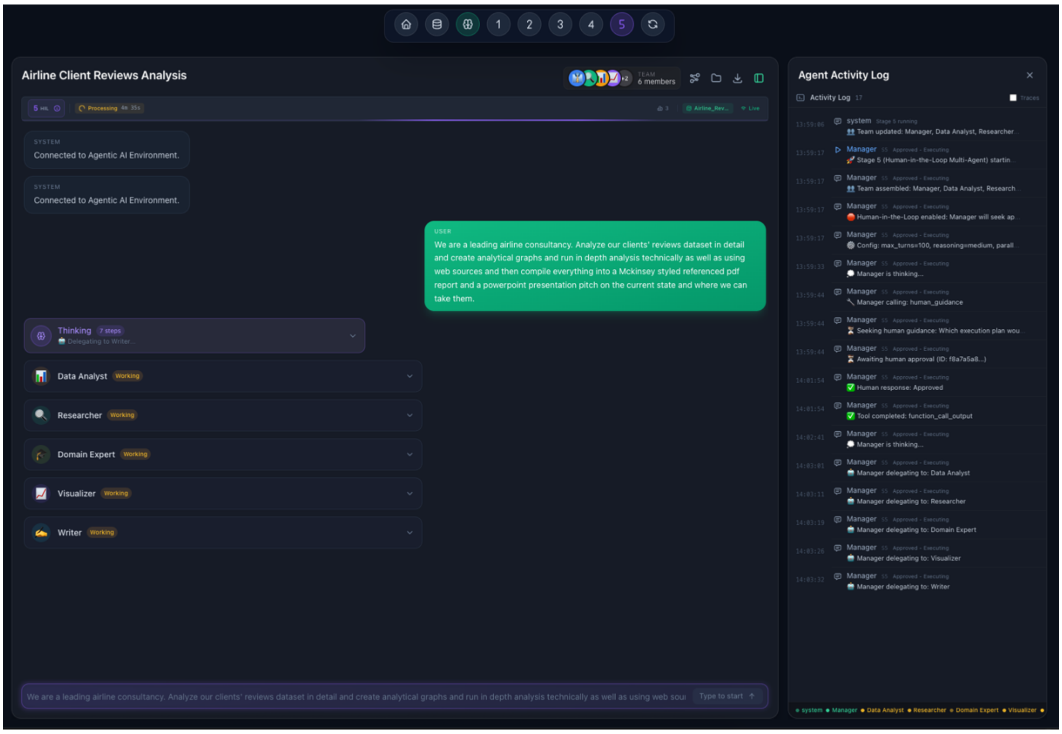

Figure 23 Multi-Agent Execution (Stage 5 Running)

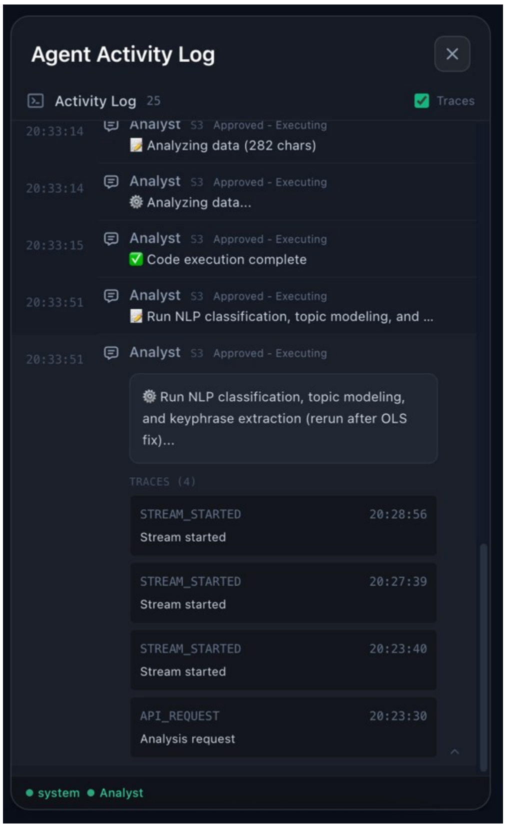

Figure 24 Detailed Agent Activity Log Panel

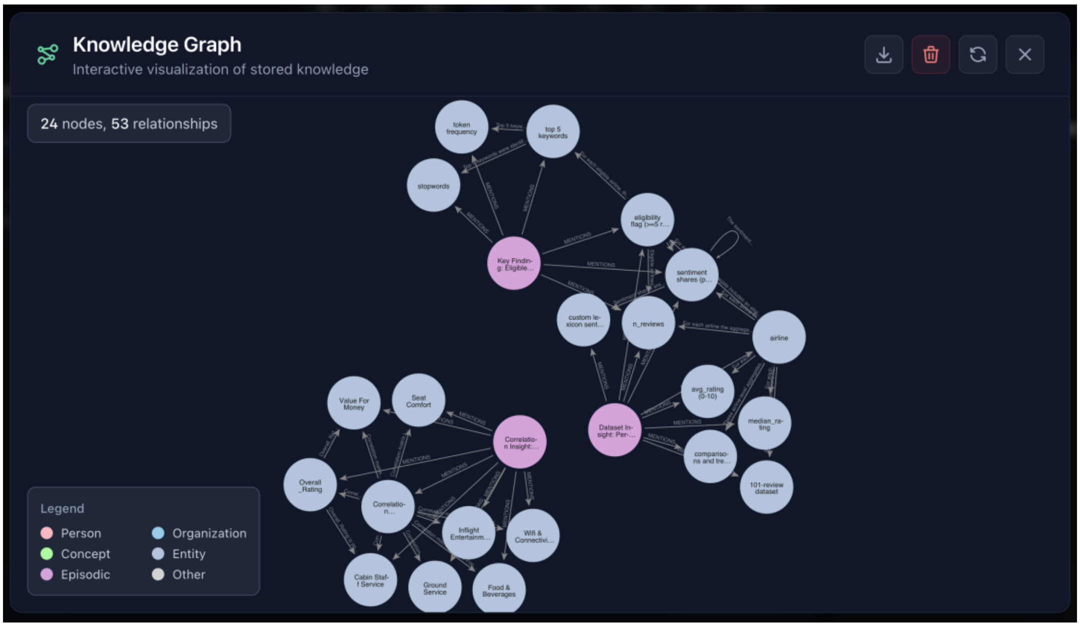

Figure 25 Knowledge Graph Visualization Modal Component to Support GraphRAG

Appendix C

Table 13 Datasets' Information & Statistics

Dataset	Domain	Rows	Columns	Size	Format	Source
Airline 100	Aviation/Reviews Dataset	101	20	72 KB	CSV	(Bhojani and Pitroda, 2025)
Amazon	E-commerce/Reviews Dataset	1,465	16	4.5 MB	CSV	(Karkavelraja, 2025)
Spotify	Music Streaming Dataset	8,582	15	1.4 MB	CSV	(Bilal, 2025)
Airline	Aviation/Reviews (top 100) Dataset	23,171	20	20 MB	CSV	(Bhojani and Pitroda, 2025)

Airline / Airline 100 Dataset
Purpose: Airline customer feedback dataset for initial analysis and testing
Key Features:

- Overall ratings and detailed reviews
- Service-specific ratings: Seat Comfort, Cabin Staff Service, Food & Beverages, Ground Service, Inflight Entertainment, WiFi & Connectivity, Value For Money
- Metadata: Aircraft type, Type of Traveller, Seat Type, Route, Date Flown
- Verification status and recommendation indicators

Analysis Cases: Customer sentiment analysis, service quality assessment, comparative airline analysis
Note: Both are same datasets except that Airline 100 is a 100 row subset of the large dataset.

Amazon Dataset
Purpose: E-commerce product and customer review dataset containing 1,000+ Amazon products
Key Features:

- Product information: ID, name, category, pricing (actual and discounted prices, discount percentages)
- Ratings and rating counts
- Product descriptions
- User-generated content: Review titles and detailed review content
- User identifiers and product/image links

Analysis Cases: Product recommendation systems, sentiment analysis of reviews, pricing analysis, customer behaviour studies

Spotify Dataset
Purpose: Comprehensive music streaming data extracted from Spotify's public API, covering music trends from 2009 to 2025
Key Features:

- Track-level: ID, name, number, popularity, duration (minutes), explicit content flag
- Artist-level: Name, popularity, follower count, genres
- Album-level: ID, name, release date, total tracks, album type

Use Cases: Music trend analysis, artist popularity tracking, genre evolution studies, temporal pattern analysis, recommendation system development

Appendix D

The table 11 represents 20 controlled experimental runs designed to measure system performance across the five defined autonomy stages. This is the basis of the quantitative results.

Table 14 System Performance Result by Autonomy Stage and Dataset

Stage	*Dataset*	*Run time (s)*	*Thinking/tool-call steps*	*Files produced*
1	Airline	44	3	0
1	Airline 100	31	2	0
1	Spotify	40	2	0
1	Amazon	35	3	0
2	Airline	112	25	5
2	Airline 100	138	21	4
2	Spotify	152	23	3
2	Amazon	168	26	5
3	Airline	504	29	6
3	Airline 100	468	30	7
3	Spotify	440	27	8
3	Amazon	490	35	8
4	Airline	1255	72	11
4	Airline 100	1146	81	12
4	Spotify	1332	86	13
4	Amazon	1458	90	13
5	Airline	2640	80	14
5	Airline 100	2204	90	9
5	Spotify	2789	96	18
5	Amazon	2319	101	17

Appendix E

Phase 1 Interview Questions: Current State of AI Usage (10 Questions)

These questions explore the participant's current reality. They are designed to identify the need for looking at users AI usage, Knowledge Graphs, UI/UX, pain points related to data connectivity, accuracy, and context loss.

Q1. Can you walk me through your current process for analyzing large sets of unstructured data? Where do you feel the biggest bottleneck is right now?

Rationale: Open-ended discovery of the "Process" workflow. Identifies if the bottleneck is retrieval (Technology) or synthesis (Cognitive Load).

Q2. When you look at your organization's data, do you feel like you can easily see the connections between different departments or projects, or does the information feel 'stuck' in separate islands?

Rationale: Indirectly probes for Data Silos. If they say "islands," it confirms the need for a Knowledge Graph architecture to traverse relationships.

Q3. Have you ever tried to use AI to find patterns across different documents? Did the AI seem to 'remember' the context from one document to the next, or did it treat them as completely separate items?

Rationale: Probes for Context Fragmentation. "Treating them as separate" indicates a vector-based failure mode that MAS/Knowledge Graphs aim to solve.

Q4. How often do you find yourself double-checking the AI's work because you don't trust the data it's referencing? What kinds of errors are you most worried about?

Rationale: Assesses Data Accuracy and Trust. High verification needs suggest a failure in current retrieval mechanisms (RAG).

Q5. Thinking about the AI tools you currently use, what is the most frustrating or good thing about the interface (UI)?

Rationale: UI/UX & Explainability. Addresses the "Black Box" problem. If the UI doesn't show the "why," adoption drops.

Q6. If you had a magic wand and could fix one thing about how your current systems handle complex queries, what would it be?

Rationale: Elicits the Ideal State. Often reveals a desire for "reasoning" or "connectivity" which aligns with MAS capabilities.

Q7. How easy or difficult is it for your team to adopt new AI tools? Do people generally feel excited, or are they overwhelmed by the learning curve?

Rationale: Organizational Culture and Effort Expectancy. High overwhelm suggests a need for better UX in agentic systems.

Q8. Who in your organization is currently responsible if an automated system gives bad output? Is that clear to everyone, or is it a grey area?

Rationale: Governance & Accountability. Critical for Stage 4 autonomy. "Grey area" responses indicate high risk for autonomous deployment.

Q9. Do you feel your current tools allow you to dig deep enough into the 'why' behind the data, or do they mostly just give you high-level summaries?

Rationale: Interpretive Depth. Validates the core thesis that current LLMs lack depth.

Q10. Are there any strict rules or security concerns that currently stop you from connecting all your data sources together for the AI to access?

Rationale: Environmental/Regulatory Constraints (GDPR, Privacy). This identifies barriers to the "Agentic Mesh."

Appendix F

Phase 2 Interview Questions: Reaction to the 5-Stage Agentic Demo (10 Questions)

These questions are asked after showing the user the progression from Chatbot to Autonomous and Human in the Loop MAS. They focus on User Experience (UX), Trust, and Perception.

Methodology: The participant watches a running practical demonstration run of each autonomy stage in the developed AI environment. Questions are asked after each stage or as a bridge between them to measure the shift in perception, thoughts, and analysis.

Stage 1: Reactive Baseline
Context: Simple chat. No tools. No memory.

Q1. We've all seen this before. In your work, how quickly does this 'chat-only' experience become redundant or remains useful?

Rationale: Baseline Frustration. Establishes the limitation of "Generative" vs "Agentic" AI.

Transition: From Chat to Action (Stage 1 → 2)

Q2. The next AI system we will see can actually use tools (like code execution and web search) on its own. What is the one task you would be terrified to let an AI do without your supervision?

Rationale: Perceived Risk. Identifies the "Red Lines" for autonomy before the user even sees the capability.

Stage 2: Tool-Augmented Agent (Autonomous)
Context: Single agent. Uses tools (Search, Graph) autonomously.

Q3. This agent just executed a complex search and analysis without asking you for help. Did that speed make you feel productive, or did it make you anxious that you couldn't check its work?

Rationale: Automation Bias vs. Anxiety. Tests the user's "Comfort with Autonomy" in a single-agent context.

Transition: From Autonomy to Human-in-the-Loop (Stage 2 → 3)

Q4. That last agent was fast. Would you be willing to sacrifice that speed if the agent was forced to 'pause and ask permission' before it finalized the report?

Rationale: The Trust-Efficiency Trade-off. Measures the value placed on "Human Agency" vs "Computational Efficiency."

Stage 3: HIL Single-Agent
Context: Single agent. Proposals presented for approval.

Q5. In this version, you had to click 'Approve' for every major step. Did this feel like a safety net, or did it feel like you were just micromanaging a junior employee?

Rationale: UI/UX Friction. "Human-in-the-loop" can often lead to "Alert Fatigue" or cognitive overload.

Transition: From Individual to Team (Stage 3 → 4)

Q6. So far, we've only used one agent. Real organizational problems are complex. Do you think adding more AI agents to work together would make the analysis better, or just more chaotic?

Rationale: Perception of Multi-Agent Complexity. Probes the user's mental model of "Swarm Intelligence" vs "Too Many Cooks."

Stage 4: Autonomous Multi-Agent
Context: Manager + Specialists. Parallel execution. Group chat coordination. (Sheridan Level 8)

Q7. You just saw several agents talking to each other to solve the problem. Did seeing them 'collaborate' make you trust the answer more, or did it feel like a 'Black Box' you couldn't control?

Rationale: Social Proof in AI. Does the "multi-agent debate" pattern act as a trust signal for humans?

Transition: From Black Box to Glass Box (Stage 4 → 5)
Stage 5: HIL Multi-Agent
Context: Full MAS. Maximum transparency. Checkpoint oversight.

Q8. Now in this system you had a detailed interaction with multiple agents with you own involvement in decision making as well. How do you feel about this transparent or more in control approach?

Rationale: Cognitive Load vs. Transparency. The "Transparency Paradox"—too much information can decrease trust or usability.

Q9. Which of the five stages felt the most 'natural' to you as a professional? Where was the balance of human control and AI help just right?

Rationale: Optimal Autonomy Identification. Identifies the "Sweet Spot" on the Sheridan scale for this specific professional domain.

Q10. If you were designing the AI workforce of the future, what would you add, change, or subtract to make these agents feel like teammates rather than software?

Rationale: Anthropomorphism & Teaming. Moves beyond "User Interface" to "Teaming Interface."

Appendix G

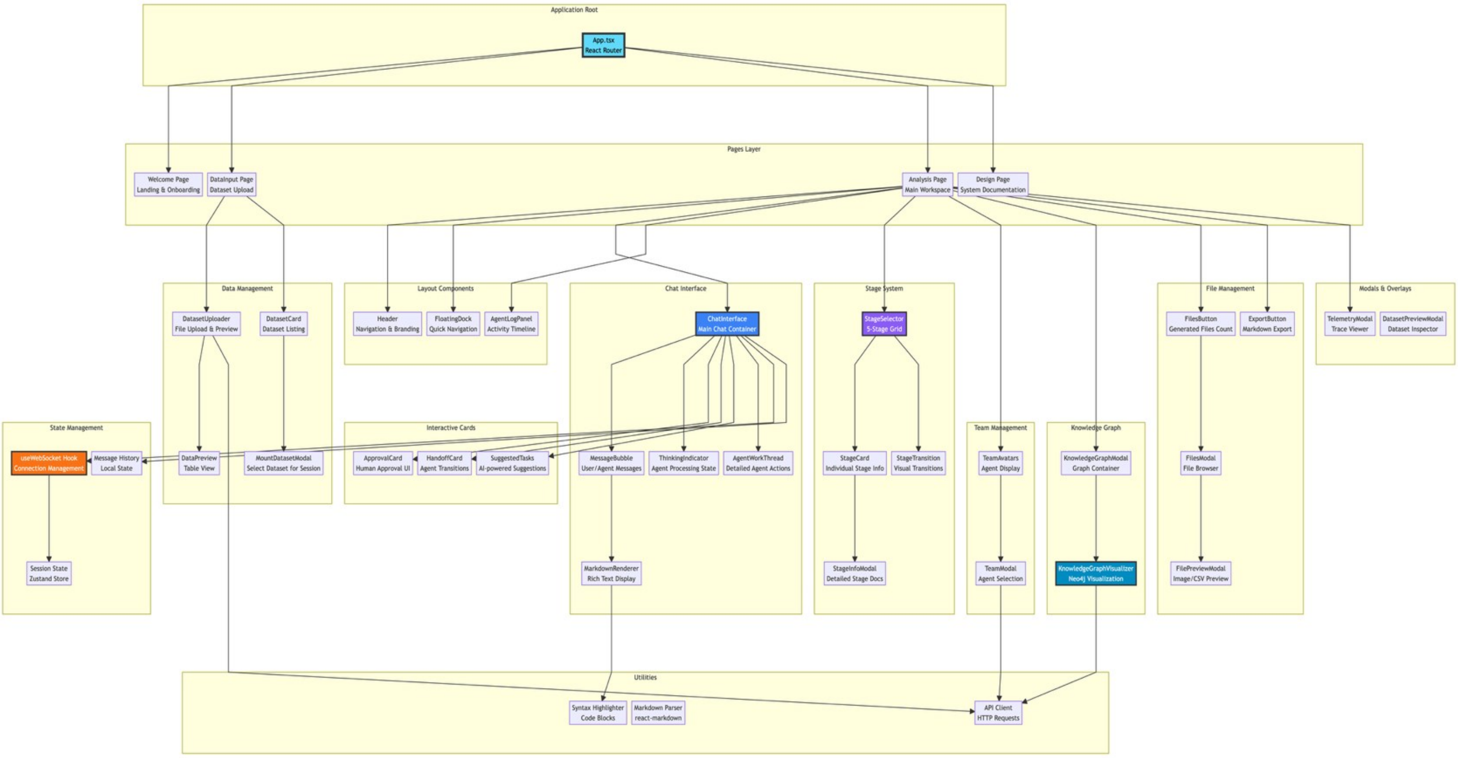

Figure 26 Frontend Architecture of Agentic AI Research Environment

Appendix H

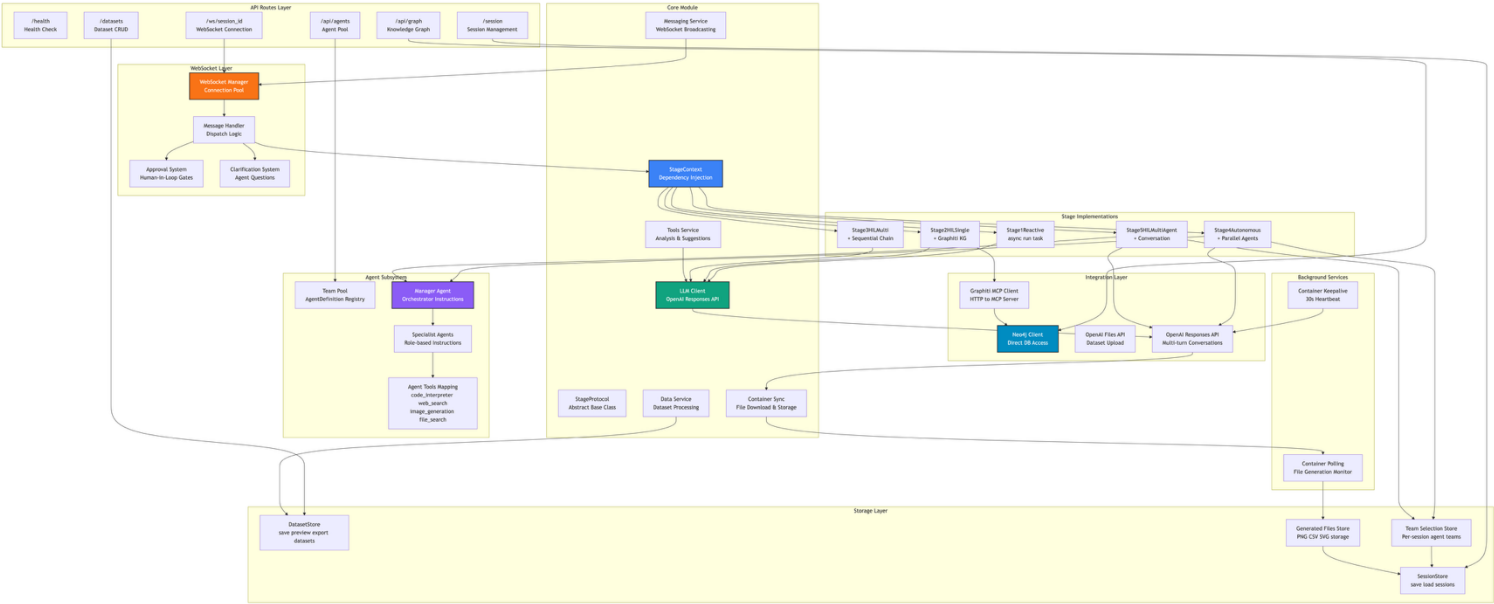

Figure 27 Backend Architecture of the Agentic AI Environment

Appendix I

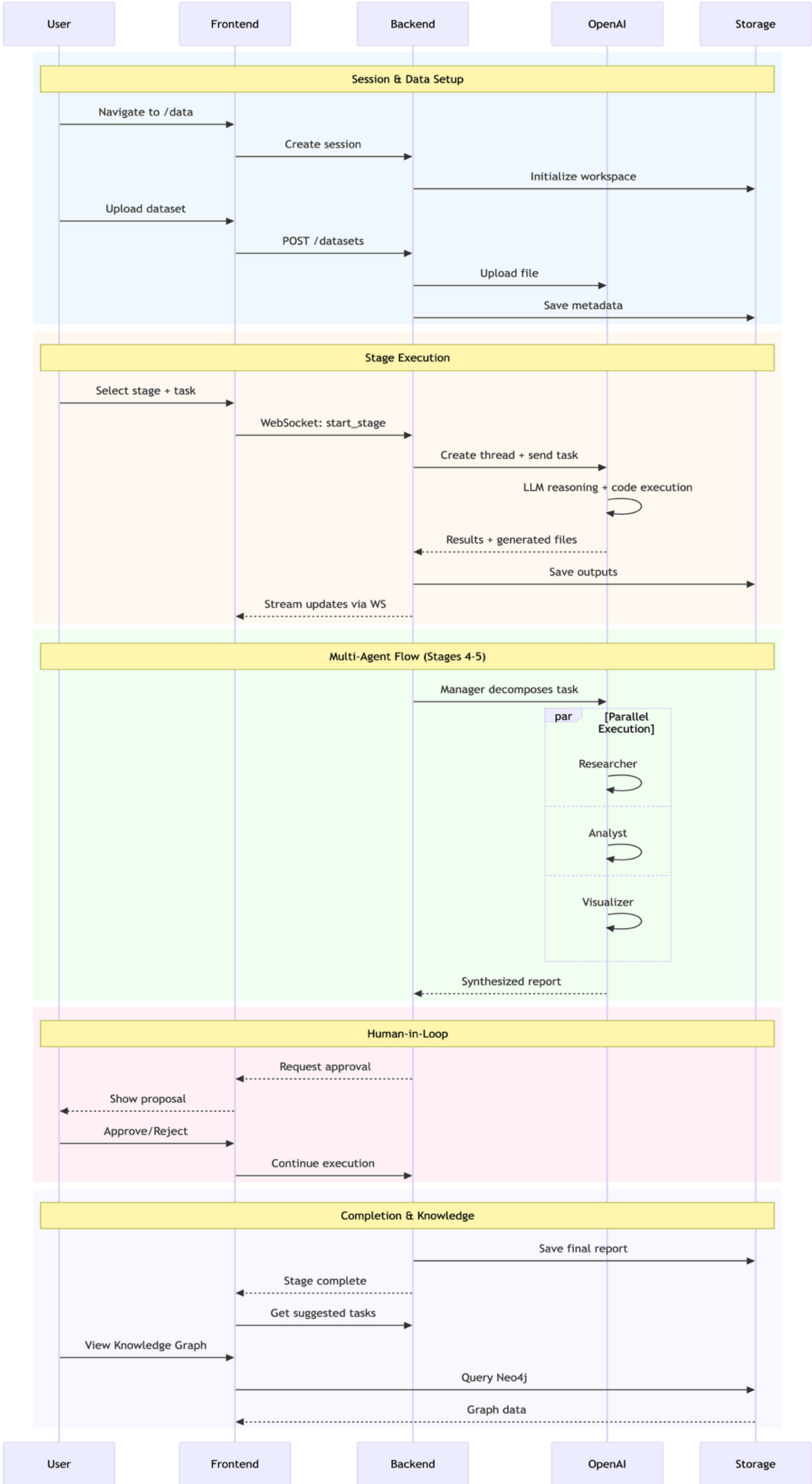

Figure 28 Data Flow Diagram of the Agentic AI Environment

Appendix J

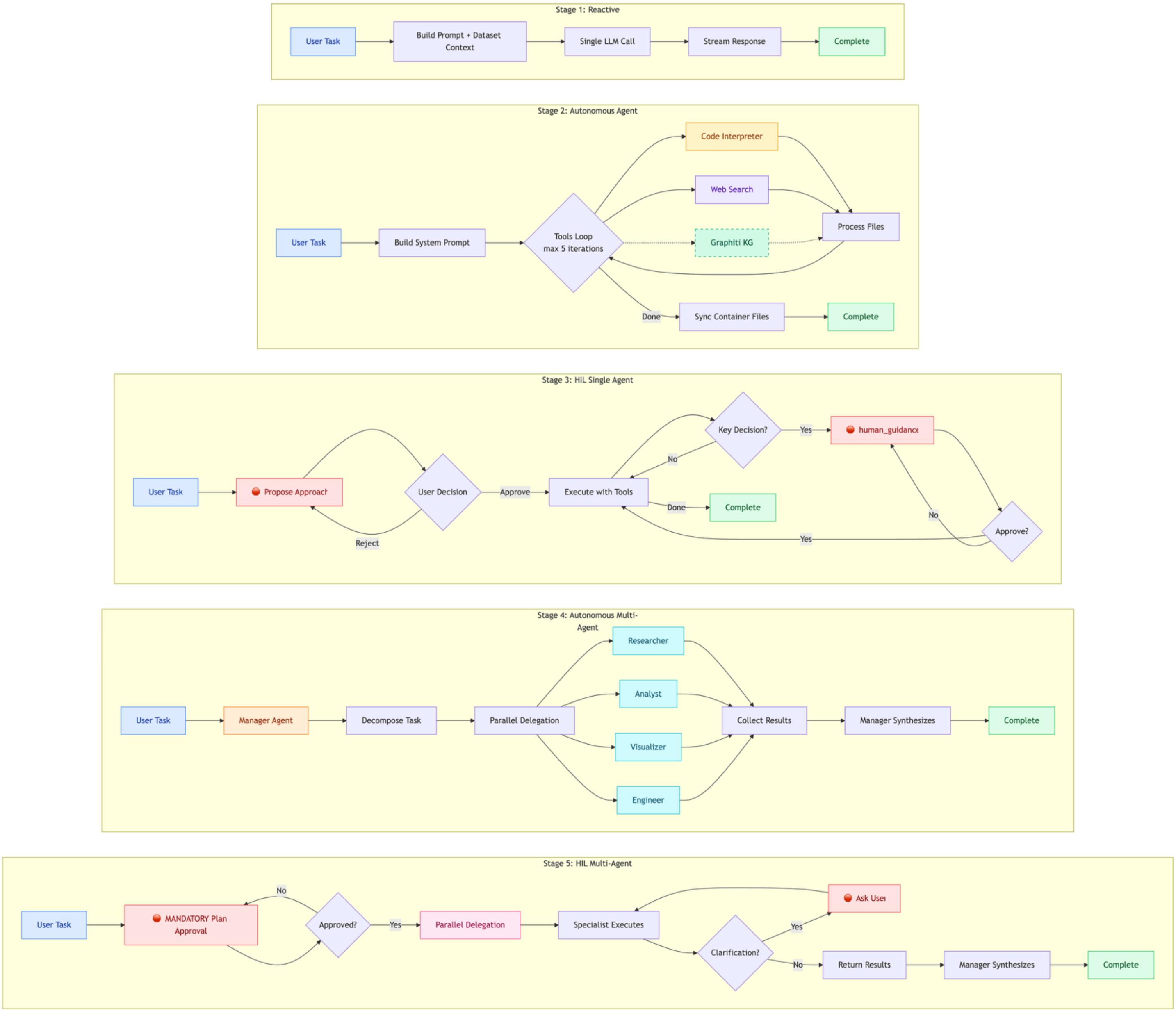

Figure 29 Logical Flowcharts of the 5 Autonomy Stages in the Agentic AI Research Environment

Appendix K

Table 15 Raw Interview Durations by Participant

Participant	*Phase 1 (mm:ss)*	*Phase 2 (mm:ss)*	*Delta (mm:ss)*	*Change (%)*
P1	22:25	45:16	+22:51	+101.9
P2	28:42	33:19	+4:37	+16.1
P3	31:22	42:43	+11:21	+36.2
P4	21:37	22:39	+1:02	+4.8
P5	25:19	36:27	+11:08	+44.0
P6	21:35	38:59	+17:24	+80.6
P7	28:15	27:30	−0:45	−2.7