Bridging Humans and LLMs: Investigating Human-AI Collaboration in Multi-agent Requirements Analysis for Organizational AI Adoption

Malik Abdul Sami[*][iD], Zheying Zhang[*][iD], Muhammad Waseem[iD],
Kai-Kristian Kemell[iD], Zeeshan Rasheed[iD], Tomas Herda[*][iD], Pekka Abrahamsson[iD]

[*]Corresponding authors: `malik.sami@tuni.fi`, `zheying.zhang@tuni.fi`, `herda.tom@gmail.com`

Article info

Abstract

Context: Organizations adopting Artificial Intelligence (AI) face challenges in eliciting and analyzing requirements that align with strategic objectives, especially when human oversight and iterative refinement are needed. Large Language Models (LLMs)-based Multi-agent systems provide a potential solution by supporting structured and collaborative Requirements Engineering (RE) processes for AI adoption planning.

Objective: The objective of this study is to investigate whether a multi-agent system, built on LLMs and supported by human input, can assist in requirements analysis for AI adoption.

Method: We used a mixed-method approach: (i) designed and developed a multi-agent system to support the generation and prioritization of requirements for AI adoption, (ii) conducted multiple case studies with four companies to evaluate the system, and (iii) collected data through post-session questionnaires from nine participants and follow-up interviews, one per company.

Results: Questionnaire and interview findings together indicate that the system may assist in identifying relevant and goal-aligned requirements. Seven participants considered the generated requirements relevant, and six found them aligned with organizational goals. Participants noted that iterative feedback improved completeness and feasibility, often within two feedback rounds. Both data sources show that human input was essential to clarify technical details, ensure contextual accuracy, and validate prioritization results. Participants from all companies also identified usability, transparency, and scalability as areas requiring further refinement for broader organizational use.

Conclusions: LLM-based multi-agent systems can support strategic AI planning by enabling iterative refinement with human experts. Future work will include more interviews with stakeholders and adjustments to system features to improve transparency, usability, and scalability.

1

1. Introduction

Organizations across diverse sectors increasingly integrate Artificial Intelligence (AI) into their strategic planning. This integration aims to achieve goals such as improving operational efficiency, creating new business models, or enhancing customer experience [1, 2]. AI technologies are used in industry, but advances in Generative AI (GenAI) have accelerated organizational adoption efforts due to its reasoning capabilities and its potential for integration into workflows and systems [3, 4].

Despite growing interest, companies face various challenges in trying to adopt GenAI tools in real-world contexts. These include concerns about data privacy and the perceived unpredictability of AI model outputs [5, 6]. In this study, we utilize GenAI's natural language processing capabilities to simulate stakeholders and generate interactive multi-agent dialogues. This enables functions such as summarization, clarity checking, classification, and prioritization within requirements analysis. While high-level AI strategy can guide organizational efforts, adopting concrete tools to realize these strategies requires rigorous planning, with Requirements Engineering (RE) providing structured processes for identifying, analyzing, and prioritizing requirements that support organizational AI strategies [7].

RE plays a key role in the early planning phase of AI adoption by helping companies articulate and prioritize needs, constraints, and concerns [8, 9]. Specifying requirements for AI-based systems includes unique challenges that differ from traditional software development [10]. These include ensuring data availability and quality [11] and validating and interpreting model behavior [12].

Recent advances in LLMs show potential for supporting RE tasks such as elicitation, specification, and prioritization [13, 14]. Prior work has explored multi-agent LLM frameworks for generating user stories and analyzing requirements [14]. Zhang *et al.* [14] showed how LLM agents can collaborate to improve user story quality. Role-based multi-agent frameworks such as the multi-agent collaboration framework for RE (MARE) [15] have been proposed, but their integration into company workflows, particularly for strategic AI planning and Human-in-the-Loop (HITL), remains underexplored [11]. Multi-agent approaches can improve coordination, adaptability, and human oversight in planning tasks. In industrial contexts, agentic systems have supported monitoring, scheduling, and dependency management in manufacturing operations, improving responsiveness and reducing downtime. Similarly, in enterprise sales, agent-based workflows help organize proposal generation and stakeholder communication, supporting human-led validation [16].

This paper extends our earlier work [17], which introduced a LLM-based multi-agent system for user story generation and prioritization. Building on that foundation, the present study extends the approach from software-level user stories to organizational-level requirements. The goal is to explore how multi-agent system can support human–AI collaboration in company-level AI adoption, where requirements represent strategic goals rather than product features. We propose a system that supports strategic AI planning by simulating multiple stakeholder roles and implement it.

In the implemented system, each LLM-based agent is assigned a stakeholder-inspired role such as Product Owner, Compliance Officer, or AI Strategist. The agents collaborate to generate and prioritize high-level requirements for AI adoption through two core activities: (i) generating requirements based on predefined role instructions and (ii) prioritizing these requirements using ranking methods. Stakeholders configure the roles, review the outputs, and refine the requirements to ensure alignment with company strategy.

The research is guided by three questions:

– **RQ1:** How can a multi-agent system be designed to support companies in analyzing requirements for strategic AI adoption planning?
– **RQ2:** To what extent does human involvement influence the quality, relevance, and prioritization of AI-generated requirements for strategic AI adoption?
– **RQ3:** What challenges, limitations and, improvement suggestions do companies identify when using the multi-agent system for AI adoption planning?

To answer these questions, we designed multi agent system and conducted a multiple-case study in four companies. We used a mixed-methods approach, combining post-session questionnaires and semi-structured interviews to capture participant feedback and contextual insights. The core contributions of this research are:

– We propose and evaluate a novel LLM-based multi-agent system for supporting strategic AI adoption planning through requirements generation and prioritization.
– We provide insights and lessons learned on human-AI collaboration, system usability, and practical challenges based on a multi-case study across four companies.

The rest of the paper is structured as follows. Section 2 provides the background and related work. Section 3 presents the research methodology, which consists of two parts: (A) the design and implementation of the proposed multi-agent system and (B) its evaluation through multi-case studies, detailed in Section 4. Section 5 reports the results from the four companies. Section 6 discusses the implications of the findings. Section 7 presents the threats to validity. Finally, Section 8 concludes the paper.

2. Background

This study examines strategic organizational requirements for AI adoption, supported by a multi-agent system that combines generative AI and human-in-the-loop collaboration.

2.1. Organizational AI adoption and challenges

Adopting AI in organizations extends beyond technical implementation. It requires strategic alignment, data governance, and readiness across departments [18–21]. Smit *et al.* [20] emphasize that trust, data quality, and stakeholder coordination are essential for effective AI planning. Russo [6] formalizes these relationships through the Human–AI Collaboration and Adaptation Framework, showing that workflow compatibility is the strongest determinant of GenAI adoption, outweighing traditional factors such as perceived usefulness or social influence. His findings demonstrate that integration success depends on how well AI tools align with existing processes and roles rather than on the technology itself. Ahmad *et al.* [8] similarly note that conventional RE tools remain ill-suited for human-centered AI projects, where reasoning and contextual interpretation are required. Leadership commitment, shared understanding, and collaboration remain critical to sustainable AI integration [22, 23].

Empirical evidence from industrial studies confirms these challenges. Kemell *et al.* [5] identify four recurring barriers in European software companies: data privacy and regulation, unclear value measurement, fragmented tool ecosystems, and limited employee guidance. Vaz Pereira *et al.* [24] report similar issues in a large media company, where developers valued GenAI for automation but expressed concerns about reliability, data security, and skill erosion. Flyckt *et al.* [25] highlight comparable barriers in manufacturing firms, including weak data governance, limited explainability, and lack of process integration.

The evidence demonstrates that adopting AI is not only a technical task but a strategic transformation that affects structures, roles, and practices across the organization. Consequently, requirements analysis in this context must extend beyond defining product features and instead capture organizational changes, governance mechanisms, and activities that enable sustainable AI adoption. Such analysis involves multiple stakeholders in different departments of an organization, and coordination and shared understanding are essential. This calls for tools that support strategic planning at the organizational level rather than only specification at the system level. LLM-based multi-agent systems provide a promising means to address this need. By facilitating multi-level requirements analysis to align with diverse stakeholder inputs, they can help organizations plan AI adoption in a transparent and coordinated manner

2.2. Generative AI and multi-agent collaboration in requirements engineering

Building on the organizational context, GenAI provides mechanisms to support RE by generating, summarizing, and refining textual content for tasks such as elicitation, prioritization, and validation. Large Language Models (LLMs) can generate user stories and related artifacts in zero- and few-shot settings [26, 27]. Ferrari et al. [28] note that while LLMs can support RE tasks, their outputs raise concerns about correctness and trustworthiness. They propose using formal methods to ensure reliable and verifiable results. Abed et al. [29] find that AI-generated user stories improve through human review.

To extend the capabilities of GenAI in RE, multi-agent systems can coordinate autonomous entities with defined roles. In RE, such systems distribute tasks like elicitation, refinement, verification, and prioritization among agents. MARE [15] defines role-based RE agents for negotiation and validation. LDB [30] supports debugging through verification-based reasoning. Hong *et al.* [31] developed a GPT-based system for requirements refinement and code generation. Sanwal *et al.* [32] proposed a pipeline for user story generation and sprint planning with LLM agents. These studies show the potential of multi-agent collaboration but remain limited to simulated or academic contexts.

Human involvement improves reliability and oversight in AI-assisted RE. Hymel *et al.* [33] report that users find it difficult to assess requirement correctness without feedback. Ferrari and Spoletini [34] suggest integrating LLMs with formal methods to improve accountability. Ahmad *et al.* [8] note that current RE tools do not address human-centered aspects in AI projects. This highlights the need for human-in-the-loop (HITL) mechanisms that maintain stakeholder understanding and alignment. Few studies have explored LLM-based multi-agent systems with HITL refinement in real organizations.

Table 1 summarizes recent multi-agent frameworks proposed for requirements analysis and software development. These frameworks differ in scope, coordination method, evaluation type, and the extent of human-in-the-loop (HITL) integration. Most remain conceptual or simulation-based, with no industrial validation. Only a few include partial human feedback or traceability mechanisms. Frameworks such as MARE [15], AutoGen-based systems, and LDB [30] define agent roles for elicitation and collaborative debugging but have been tested only in academic settings. Similarly, ProAgent [35], ALAS [14], Goal2Story [36], MAGIS [37], and RTADev [38] focus on conceptual design or prototype-level demonstrations without industrial application or systematic human validation.

Existing RE approaches provide limited support for the human, technical, and strategic dimensions of AI adoption planning within organizations [8]. Addressing this gap requires methods that combine automation with human oversight to enable multi-level analysis

Table 1: Comparison of multi-agent approaches in requirements engineering

Framework	Scope	Method	Evaluation Type	Industrial Validation	HITL
MARE [15]	Multi-agent RE framework	Role-based coordination and task assignment	Conceptual	No	Not addressed
LDB [30]	Debugging and elicitation	Verification-based reasoning and error tracing	Prototype	No	Partial reasoning trace
ProAgent [35]	Cooperative agent orchestration	Task coordination with limited feedback	Simulation	No	Limited visibility
ALAS [14]	Designed agent roles for automated user story enhancement	Role-based coordination and task assignment	Prototype	No	Not addressed
Goal2Story [36]	Goal-driven RE via Impact Mapping	LLM-based user story generation with adaptive prompts	Prototype	No	Partial HITL traceability
MAGIS [37]	Issue resolution and maintenance planning	Role-specialized agents for planning, coding, and QA review	Prototype	No	Not addressed
RTADev [38]	Software development pipeline	Multi-agent consensus and alignment-checking process	Simulation	No	Not addressed
our study	Requirements generation and prioritization	Role-based multi-agent system with iterative feedback	Empirical multi-case	**Yes**	**HITL multi-agent collaboration**

and informed decision-making. This study empirically evaluates a role-based multi-agent system that supports collaborative requirements analysis and prioritization in strategic AI adoption. Human-in-the-loop feedback is integrated throughout to refine outputs and align them with organizational objectives. Evaluation across four companies demonstrates practical applicability and identifies key challenges in human–AI collaboration within RE. To the best of our knowledge, this is the first empirical evaluation of a multi-agent RE tool in industrial settings.

3. Research methodology

The methodology consists of two sections: (A) the design and implementation of the proposed system and (B) its evaluation through multi-case studies across four organizations. Part A describes the system architecture, implemented features, and technical components. Part B outlines the case study design, including case selection, data collection, and analysis procedures. Figure 1 illustrates the methodology, highlighting the system lifecycle (design and development) and its subsequent evaluation in real-world settings.

3.1. Proposed and implemented system

3.1.1. System design

Human-in-the-loop Workflow: As illustrated in Figure 2, the framework comprises multiple LLM-based agents, each assigned a distinct role for task completion. They collaborate by exchanging intermediate outputs and iteratively refining their responses based on human feedback. Integrating human feedback into the response refinement process fulfills an effective HITL mechanism in improving the quality of task outputs.
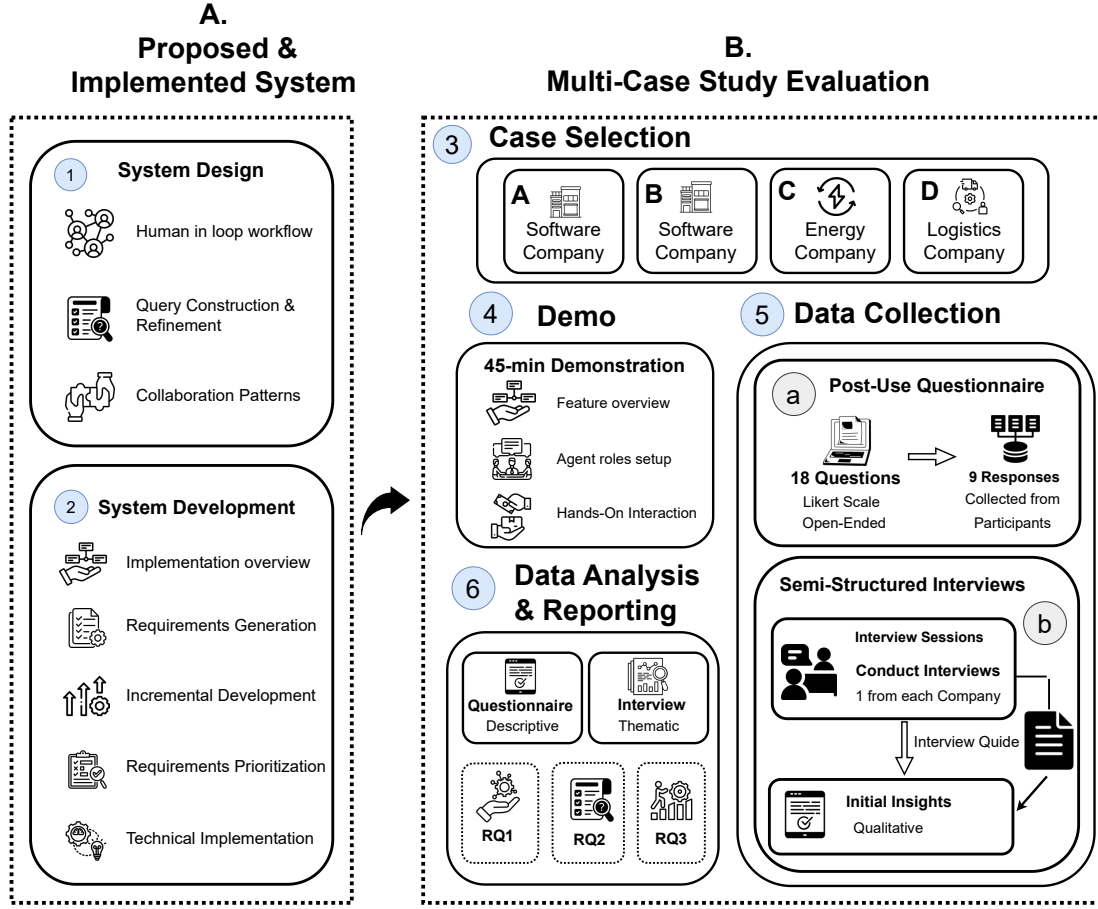
Figure 1: Overview of the research methodology, illustrating (A) system design and implementation and (B) multi-case study evaluation process

The process begins with a task description and associated contextual information, which are transformed into a structured query. This query is then processed by multiple agents, where agents interact according to a predefined interaction pattern to complete the task. The generated outputs are evaluated by human users. If the outputs are accepted, a final requirements are saved to system; otherwise, user feedback is used to revise the query and the context, and the process iterates again.

The framework includes a data store that records and maintains agent profiles, conversation logs, user feedback, and history outputs. The repository supports iterative query refinement by enabling the system to retrieve relevant contextual information and incorporate user feedback effectively. This framework supports dynamic agent interaction and continuous outputs improvement through feedback loops. The following subsections detail how queries are constructed and refined for task completion, and the collaborative patterns agents use to execute these queries.

Query Construction and Refinement: The proposed framework relies on prompt engineering to define and control how agents perform their assigned tasks. Prompts serve as the primary mechanism for guiding LLMs in role interpretation, contextual input processing, and response generation [35]. To balance system-wide consistency with agent-specific customization, the structure of queries sent to the agents comprises two types of prompts, i.e. user prompts and system prompts.
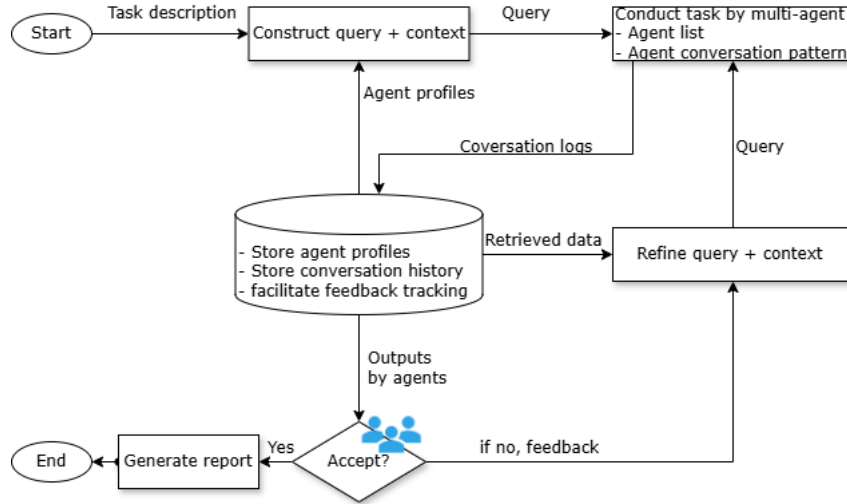
Figure 2: Human intervention with LLM agents for task completion

–   **User prompts** specify agent-specific roles, task descriptions, and contextual information
    relevant to each task.
–   **System prompts** specify shared instructions applicable to all agents involved in a task,
    including the response structure, tone, output constraints, and formatting conventions.
        At runtime, the user and system prompts are merged into a single query that enables
differentiated agent behavior without modifying the underlying model parameters [39].
If the output generated by agents fails to meet user expectations, a query refinement is
triggered. Rather than reconstructing a new query from scratch, the framework reuses the
original query and adds user feedback to improve the resulting response. Therefore, the
refinement query incorporates two additional components, i.e. previous output and user
feedback.
–   **Previous outputs** are responses generated by agents based on the original query.
–   **User feedback** includes clarifications, suggestions, or corrections provided by the user
    to guide the output enhancement.
        The updated prompt may optionally revise the original task, role assignments, or
contextual information based on the feedback. All other components remain consistent
with the initial prompt structure. The iterative human-in-the-loop mechanism is expected
to support the incremental enhancement of agents' outputs so that the framework can
improve the responsiveness to evolving user feedback and enhance the quality and relevance
of task execution.

Multi-agent Collaboration Patterns: The agents in the proposed framework follow
multi-agent design principles and support multiple interaction patterns, each aligned with
a general strategy for coordinating task execution [40, 41]. Upon task initiation, the user
identifies a group of agents based on the nature of the task and the roles required for
task completion. Each agent operates under a predefined role and a corresponding set of
instructions specified within the query. The selection of conversation pattern is decided
by the degree of inter-agent dependency needed for task completion. As illustrated in
Figure 3, we use three multi-agent collaboration patterns from the existing literature
[42]. When three agents execute a task, the system supports three collaboration patterns:
role-delegation, peer coordination, and independent execution. The role-delegation pattern
follows a hierarchical structure. A commander agent (Agent 1) receives the task input,

assigns sub-tasks to subordinate agents (Agent 2 and Agent 3), and synthesizes their outputs into a final result. This pattern fits modular tasks with minimal inter-agent dependency.

The peer coordination pattern enables decentralized collaboration, where all agents communicate directly, exchange intermediate results, and refine responses through dialogue. It is ideal for tasks requiring consensus, comparison, or integration of diverse viewpoints, such as eliciting and analyzing requirements from multiple stakeholders.
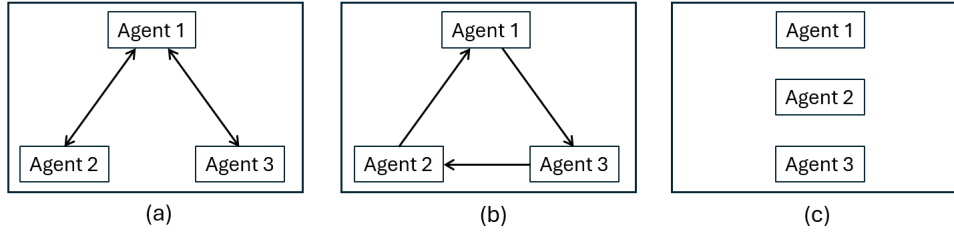

Figure 3: Collaboration patterns between agents: (a) role delegation, (b) peer coordination, and (c) independent interaction

The independent execution pattern involves agents working autonomously. Each receives the same input, performs its task independently, and produces a separate output. This pattern fits tasks requiring parallel analysis. These configurable collaboration patterns provide flexibility in orchestrating agent workflows and enable the system to accommodate a broad range of task structures, from tightly integrated to fully independent processes.

3.1.2. System development

To demonstrate the implementation of the system and evaluate its use in practice, we developed a multi-agent system to support human-in-the-loop strategic AI planning. The system focuses on high-level strategic organizational requirements that support decision-making for strategic AI adoption. As shown in Figure 4, the system addresses two core tasks: (1) requirements generation and (2) requirements prioritization. Requirements are represented as user stories, which facilitate communication of stakeholder expectations and support agile, incremental planning even at strategic levels. This approach bridges high-level organizational goals with concrete implementation details.

Task 1: Requirements Generation. Requirements generation follows a role delegation pattern. Senior stakeholders, such as executives, strategic planners, and IT leaders, initiate this task by providing strategic context through a vision statement and a process-oriented Minimum Viable Product (MVP) [43]. The MVP outlines an iterative roadmap emphasizing processes, stakeholder interactions, governance structures, and scalability pathways rather
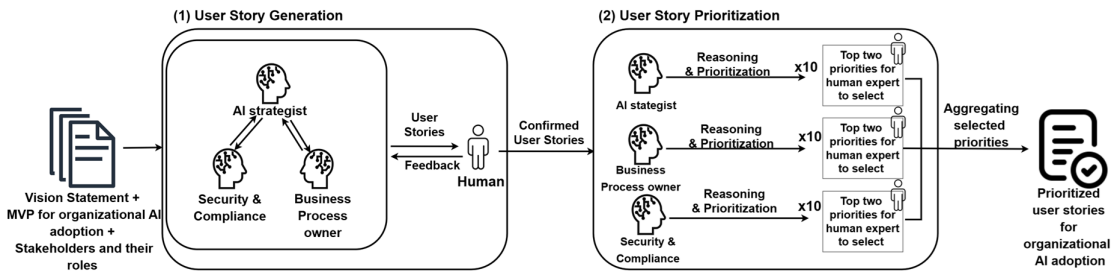

Figure 4: AI-driven multi-agent system for requirements analysis and prioritization

---

Vision and MVP for Stakeholder-Driven Strategic AI Planning

Vision Statement

Our vision is to transform our IT company into a leading AI-enabled organization that leverages cutting-edge artificial intelligence to optimize processes, empower data-driven decision making, and deliver exceptional value to stakeholders. By implementing a stakeholder-driven, LLM-based multi-agent approach to requirements analysis for each AI adoption use case, we ensure that every AI solution addresses real business challenges and evolves with the dynamic market and technological landscape...

MVP

Minimum Viable Plan (MVP) for AI-Enabled Organizational Transformation. 1. Vision Statement: Transform our IT company into a leading AI-enabled organization that leverages cutting-edge artificial intelligence to optimize processes, empower data-driven decision making, and deliver exceptional value to stakeholders. 2. Strategic Objectives: Establish a structured AI adoption framework to guide investments, develop a stakeholder-driven AI requirements process, implement governance and ethical frameworks, and define a roadmap for scalable AI integration across business units. 3. Stakeholders and Their Goals: Executives align initiatives with business strategies, IT Leaders develop scalable infrastructure, Data Scientists optimize AI models, Compliance Officers ensure governance, and Business Units leverage AI for decision intelligence and process automation. 4. MVP Scope (Initial Phase): Deploy a stakeholder-driven, LLM-based multi-agent system for AI use case analysis, identify and prioritize initiatives using structured criteria, and develop governance policies. 5. Success Metrics: Strategic AI Integration Index, AI Governance Readiness Score, Business Impact Metrics, and Scalability Readiness. 6. Iterative Roadmap: Phase 1 involves requirements elicitation and prioritization with LLM agents, while Phase 2 focuses on feasibility assessment and scalable implementation...
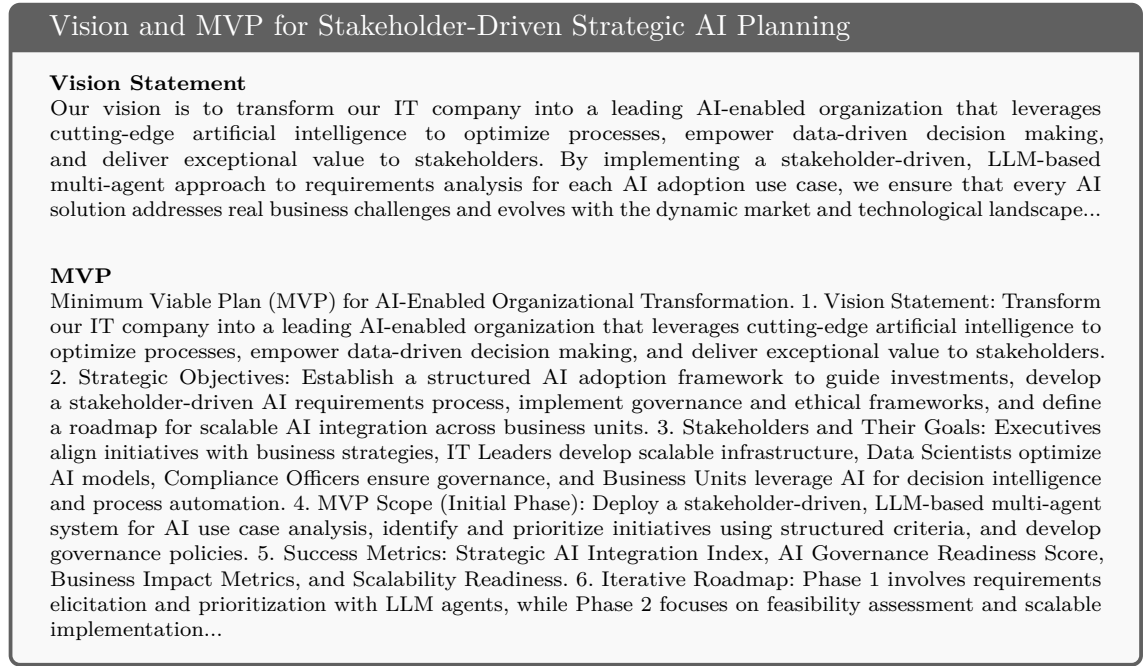
Figure 5: Vision and MVP for stakeholder-driven strategic AI planning

than specific technological solutions. Figure 5 presents excerpts from the vision statement and MVP. Senior stakeholders also define strategic AI planning roles and required expertise. Agents such as AI Strategist, Compliance Officer, and Business Owner are then assigned to generate candidate user stories reflecting business value, regulatory compliance, and technical feasibility. The AI Strategist coordinates and disseminates contextual information to subordinate agents, enabling parallel outputs. Human users review, refine, and confirm these generated requirements to ensure alignment with organizational objectives. Table 2 summarizes the agent roles implemented in the system.

Table 2: LLM-based agent roles and corresponding responsibilities and tasks

Agent	Role	Responsibilities	Tasks
AI Strategist	Lead strategic agent	Identify high-impact AI use cases, align initiatives with strategic goals, and foster cross-departmental collaboration. Continuously reassess and refine AI direction to stay ahead in a rapidly evolving landscape.	Translate business challenges into AI opportunities; generate and prioritize requirements related to strategic alignment and innovation; coordinate agent activities across the system.
Security and Compliance Officer	Regulatory oversight agent	Ensure AI initiatives adhere to legal frameworks and ethical guidelines, mitigating risks related to bias, privacy, security, and transparency.	Identify regulatory constraints; generate compliance-focused requirements; validate system outputs for regulatory alignment and risk mitigation.
Business Owner	Business value agent	Translate operational needs into actionable requirements for AI solutions. Ground initiatives in business processes and practical value. Facilitate collaboration between business and technical stakeholders.	Contribute business-centric requirements; prioritize based on business value; refine based on practical challenges and measurable outcomes.

Task 2: Requirements Prioritization. Requirements prioritization is implemented as an independent interaction pattern. The user selects three agents and chooses a prioritization technique, e.g. the Weighted Scoring Model (WSM), Weighted Shortest Job First (WSJF), or the $100 method [44], along with the number of iterations (three to ten). As shown in Figure 4, each agent independently repeats the prioritization process X times using

the same prompt, generating one ranking list per iteration. To evaluate the stability of these rankings, the system computes Kendall's Tau correlation for every pair of lists across all iterations. After all iterations, it calculates the average Kendall's Tau distance $(1 - \tau)$ of each ranking to the others [45, 46]. The ranking with the lowest average distance is identified as the most consistent, indicating convergence toward a stable prioritization. For each agent, the system identifies the two most stable rankings, i.e., those with the lowest average Tau distances, thereby filtering out random variation and ensuring reproducible prioritization outcomes. The user then reviews these top two rankings per agent and selects the one that best reflects their judgment.

Users may assign weights to the selected rankings to reflect the relative importance of each agent's viewpoint. For example, the AI strategist may prioritize feasibility and alignment with AI objectives, the business process owner may focus on business value, and the security/compliance specialist may emphasize technical and regulatory constraints. Weights are fully configurable, to select a value from a drop-down, such as 40–35–25, and can be adapted to the decision context. The system aggregates the chosen rankings using the provided weights to produce the final prioritized list. This workflow combines statistical stability analysis with expert input, ensuring that the outcome reflects both data-driven consistency and human judgment. Screenshots and source code for the full workflow are available in the public repository[1].

Technical Implementation Details: This system extends our earlier prototype [17]. The current implementation comprises a React frontend and a Python backend built with the Starlette framework [47, 48], using real-time WebSocket communication. The backend manages user sessions, agents profiling, processes OpenAI API calls, and structures the generated outputs. GPT-4o [49] was used for all agent roles through the OpenAI API, selected for its performance and cost efficiency in real-time multi-agent interaction. The temperature parameter was set to 0.1 to reduce generation variability, while other parameters such as top-p, frequency penalty, and presence penalty remained unchanged.

Supabase is used for role-based authentication, and MongoDB stores the application data. The system integrates human-in-the-loop mechanisms that enable iterative feedback during requirements refinement. Agents operate under predefined roles such as AI Strategist, Business Owner, and Compliance Officer.

The multi-agent environment was developed iteratively to support organizational AI adoption. Development progressed through staged integration of agent roles, communication logic, and feedback mechanisms to ensure stable and efficient operation. Its performance was evaluated in earlier work [17], which confirmed technical feasibility and provided the foundation for this study.

4. Multi-case study evaluation

The evaluation aimed to assess how the multi-agent LLM system supports strategic AI adoption planning through human–AI collaboration in the requirements analysis process. Following the guidelines of Runeson and Höst [50], we conducted case studies in four companies. Data collection included post-use questionnaires and semi-structured interviews. Questionnaire responses $(n = 9)$ were analyzed using descriptive statistics, while interview data were analyzed thematically using Braun and Clarke's method [51].

---

[1]https://github.com/GPT-Laboratory/multiagent-story-generation/tree/main/screenshots

The study timeline was coordinated with the participating organizations. The initial evaluation was conducted in Company D during February and March 2025, during which the system was refined based on participants' feedback. The updated version was then used in the remaining three companies for evaluation in April and May 2025. In total, the study spanned four months, with all interviews completed in May 2025. The following subsections describe the demonstration, data collection process, analysis methods, and data reporting for each research question.

4.1. Case selection

The study was conducted in four organizations: two software consultancies, one energy provider, and one logistics company. Following established case study guidelines [50, 52], we applied a convenience sampling strategy [53], selecting organizations that were both accessible and actively engaged in AI adoption initiatives. The cases were chosen to reflect variation in industry sector, company size, and level of AI maturity.

Each organization agreed to participate voluntarily and nominate individuals involved in AI-related planning or decision-making processes. These included IT leaders, product owners, an agile coach, directors, and development managers responsible for planning or implementation tasks.

Company A is a small software consultancy providing development and technology services to clients in the USA, Pakistan, and the UAE. It uses generative AI tools such as Cursor and Copilot to improve productivity in software development and is exploring AI for product planning, marketing, and client engagement. The main challenge is integrating AI into daily workflows to achieve measurable business impact. Motivated by these goals, the company joined the study to evaluate how the tool could support strategic AI planning and requirements analysis. The Director of Technology (P2) and Product Owner (P1) participated, focusing on AI-assisted requirement elicitation and planning.

Company B is a mid-sized software consultancy providing product development and process automation services in the USA and Pakistan. It uses generative AI tools such as ChatGPT, Codium, Cline, and Cursor AI across departments, with most teams using them to accelerate workflows and improve customer satisfaction. The company is also developing an agent-based HR automation system for CV prioritization and AI-supported interviews. The main challenge is aligning AI initiatives with existing systems and governance processes. Motivated by these goals, the company joined the study to explore how the tool could support AI-assisted prioritization and planning. The IT Leader (P3), Business Process Owner (P4), and IT Leader (P5) participated in the study.

Company C operates in the European energy sector, providing electricity, heating, and renewable energy services. It has an internal software unit managing digital systems and data-driven operations. The company is identifying AI use cases to optimize energy management and improve productivity across departments. About half of the employees use tools such as Copilot and ChatGPT to enhance efficiency and reduce routine work. The main challenge is managing change and ensuring smooth technology adoption. Motivated by these goals, the company joined the study to explore how the tool could support strategic AI planning. The Development Director (P6) and Development Manager (P7) participated, focusing on aligning strategic and technical planning.

Company D is an international postal, logistics, and service provider with a strong presence in Central and Eastern Europe. It delivers a broad range of mail, parcel, and freight solutions, supported by dedicated internal software teams responsible for routing,

Table 3: Overview of case companies and participants

Company	Region	Industry	#Employees	Participant (ID, Years of Experience)
A	South Asia	Software Consultancy	10–49	Product Owner (P1, 10+), Director of Technology (P2, 12+)
B	South Asia	Software Consultancy	50–249	IT Leader (P3, 8+), Business Process Owner (P4, 7+), IT Leader (P5, 5+)
C	Europe	Energy Sector	50–249	Development Director (P6, 10+), Development Manager (P7, 10+)
D	Europe	Logistics	499–750	Scrum Master (P8, 5+), AI Strategist (P9, 9+)

operational efficiency, and service analytics. The organization is committed to high-quality standards and continuously adapts its offerings to meet evolving customer needs and is piloting AI to improve operational efficiency, automation, and decision-making while maintaining compliance with security and legal standards. A key challenge lies in identifying AI use cases with measurable business value and clear return on investment. Motivated by these goals, the company joined the study to explore how the tool could support strategic AI planning and prioritization. The Scrum Master (P8) and AI Strategist (P9) participated in the study, focusing on aligning operational and strategic AI initiatives.

Table 3 summarizes the context of the four participating companies, including their region, industry, size, and participant experience. The cases represent diverse sectors and organizational scales, ranging from small consultancies to large enterprises, with participants holding senior roles in AI and technology management.

4.2. Demo

Following system implementation, participants from all four organizations were introduced to the tool through guided demonstration sessions. Each organization took part in a 45-minute session that included an overview of the system's features, configuration of agents and roles, and hands-on interaction. The purpose of these sessions was to familiarize participants with the system and ensure they were prepared for independent use.

During the demonstration, participants explored key capabilities, including assigning agent roles, generating and refining user stories using given vision and MVP statements, and prioritizing them. This phase served both to collect initial user feedback and to support participants in applying the system autonomously within their own planning workflows.

4.3. Data collection

Post-Use questionnaire: In this phase, participants completed a post-session questionnaire comprising 18 questions organized into five sections aligned with the research questions. The first section collected demographic and organizational background information (Q1–Q7). The second and third sections included 16 five-point Likert-scale statements (Q8–Q9, Q12–Q13) examining requirement generation, refinement, prioritization, and strategic alignment. Evaluation followed a subset of requirement quality criteria from ISO/IEC/IEEE 29148:2011 [54]. These sections measured perceptions of relevance, correctness, completeness, feasibility, transparency, and alignment with organizational goals. The fourth section assessed human–AI collaboration and integration through Likert

Table 4: Summary of data collection, participants, sessions, and average durations

Instrument	#Participants	#Sessions / #Responses	Avg. Duration
Demonstration and Q&A	9 (P1–P9)	4 sessions (one per company)	45 min
Post-Session Questionnaire	9 (P1–P9)	9 responses	14:24 min
Semi-Structured Interviews	4 (P2, P3, P7, P9)	4 interviews (one per company)	35 min

and open-ended questions (Q13–Q14). The fifth section contained multiple-selection and open-ended questions (Q15–Q17) capturing perceived benefits, challenges, and improvement suggestions, followed by one numerical rating question on overall recommendation likelihood (Q18). The complete questionnaire and interview are available in the project's GitHub repository[2].

Semi-Structured Interviews: After the post-session questionnaire, we conducted interviews with one participant from each company, invited based on prior questionnaire completion and informed consent. These interviews explored deeper reflections on system usefulness, organizational fit, and suggestions for improvement. Discussions began by establishing organizational context, including current GenAI tool usage, roll out stage, and the participant's role in AI adoption. Participants then shared initial impressions of the system, focusing on ease of use and alignment with strategic planning goals.

The interviews examined the feedback process in detail, including types of revisions to system-generated requirements, number of feedback rounds required, and how human input shaped final outputs. Further, the sessions investigated how well the system fit within organizational workflows, addressing trust in AI-generated logic, team-specific needs, and internal resistance. Finally, participants reflected on the system's overall value and proposed concrete suggestions for improving functionality, usability, and integration.

Table 4 summarizes the data collection process. Each company participated in one demonstration session, followed by a post-session questionnaire completed by nine participants, with an average completion time of approximately 14 min. Four follow-up interviews, one per company, provided deeper insights into system use and organizational fit. Interview durations ranged from 33 to 38 min (P2: 38 min, P3: 33 min, P7: 34 min, P9: 34 min), with an average duration of approximately 35 min.

4.4. Data analysis and reporting

We applied both quantitative and qualitative methods to analyze the collected data, including post-session questionnaire responses and semi-structured interviews. These two sources of data were used for triangulation to strengthen validity and ensure consistency between quantitative and qualitative findings.

We performed descriptive statistical analysis, calculating means and standard deviations to summarize participants' ratings, while frequencies and percentages were used for categorical items. Responses were grouped by research question to examine perceptions of system usefulness, effectiveness of feedback, and integration into organizational workflows.

For qualitative data, we analyzed interview transcripts using thematic analysis following Braun and Clarke's six-phase method [51]. We adopted an inductive, data-driven approach where codes and themes were developed directly from the data. ATLAS.ti[3] was used for coding. Coding decisions were discussed among the research team to ensure consistency.

---

[2]https://github.com/GPT-Laboratory/multiagent-story-generation/tree/main/instrument
[3]https://atlasti.com/

Table 5: Examples of requirements before and after refinement during iterative development

ID	Requirements	Feedback Summary	Refined Requirements
1	As a compliance officer, I aim to establish AI governance practices to ensure ethical and legal AI use in the organization.	Add reference to applicable regulatory frameworks.	As a compliance officer, I aim to establish AI governance practices aligned with the EU AI Act and internal ethical standards to ensure legal and responsible AI use.
2	As a project stakeholder, I aim to validate AI project priorities through structured review sessions.	Clarify evaluation criteria and link to decision process.	As a project stakeholder, I aim to validate AI project priorities through structured evaluation sessions using defined criteria for business value and feasibility.
3	As an executive, I want to establish a structured AI adoption framework to guide organizational planning.	Add measurable outcome or acceptance criteria.	As an executive, I want to establish a structured AI adoption framework that defines key milestones and measurable outcomes for responsible deployment.
4	As a project stakeholder, I need AI governance policies to guide system design.	Avoid duplication and make scope explicit.	As a project stakeholder, I need AI governance policies specifying data handling, accountability, and audit mechanisms to guide AI system design.

Participant quotations were edited for readability by removing filler words and applying minor grammatical refinements without altering meaning.

The analysis identified one high-level theme for RQ1 (first-use impressions and initial support), two themes for RQ2 (feedback-driven refinement and benefits of human-in-the-loop), and four themes for RQ3 (usability barriers, scalability, trust and transparency, and user suggestions). A total of 37 unique codes were applied to 175 coded quotations across four interviews. The number of quotations per code ranged from 1 to 20. The qualitative findings supported RQ2 and RQ3 by capturing feedback on user trust, refinement processes, human-in-the-loop interactions, integration challenges, and design suggestions. The full list of codes available on GitHub[4].

5. Results

In total, nine participants completed the post-session questionnaire, which included Likert-scale and open-text questions across five sections. Four participants (P2, P3, P7, and P9), one from each company, took part in follow-up semi-structured interviews. The interviews followed RQ-aligned prompts focusing on system use, human-in-the-loop interaction, trust, and improvement suggestions.

Nine participants used the system within their organizational planning workflows. For example, in Company D, participant P8 received 21 initial requirements generated by the system. In the feedback rounds, one duplicate was removed, four overlaps were merged, and missing contextual details were added. At the end, the process produced 16 approved user stories. Table 5 presents representative examples of initial and refined user stories, illustrating how user feedback improved the relevance and precision of outputs. Practitioner comments mainly focused on clarifying scope, removing duplicates, and ensuring alignment with compliance and strategic context. The complete set of generated and refined stories with feedback notes is available in the project repository.

---

[4]https://github.com/GPT-Laboratory/multiagent-story-generation

5.1. System support for requirement analysis (RQ1)

The analysis identifies one main theme: *System Support for Requirement Analysis*. It describes how the system assists users in generating, reviewing, and aligning requirements with organizational goals, showing that it enables structured formulation of requirements while allowing users to provide feedback and maintain control during refinement.

Questionnaire responses indicate generally positive perceptions of the generated requirements. Seven participants (78%) agreed or strongly agreed that the requirements were relevant ($M = 4.2, SD = 0.6$), and six participants (67%) agreed that they aligned with organizational goals ($M = 4.1, SD = 0.7$). Five participants (56%) reported that only minor modifications were needed before using the generated requirements. Overall, the quantitative results suggest that the system was perceived as supportive for early requirement formulation and for maintaining alignment with organizational objectives.

The interview data provide further insight into how this support was experienced in practice. Participants emphasized the relevance and usefulness of the generated requirements. As P2 noted, *"When I started exploring it, it was quite impressive… it gives me some promising results as well."* Pariticipant P3 also highlighted the efficiency of the tool: *"It was very fast in giving the requirements results quickly, at least for short projects.".*

Participants also described how the system contributed to organizational goal alignment, as P7 reflected, *"I was curious about how it would bring more knowledge… it shows us how agents perform different roles."* This indicates that the multi-agent design helped broaden requirement perspectives.

Finally, the interviews also support the finding that only minor refinement was needed. For example, P9 explained, *"I did like the human-in-the-loop… it's possible for a human to adjust the results that are not satisfactory. That was really something good."* This illustrates how feedback mechanisms enabled efficient revision rather than extensive rework. Together, the questionnaire and interview data converge: the system helps generate relevant, goal-aligned requirements while allowing users to refine outputs with minimal effort.

5.2. Human-in-the-loop for refinement and prioritization (RQ2)

This question reports how human–AI interaction supports requirement refinement and prioritization. The analysis integrates quantitative and qualitative evidence to understand how human feedback influences the quality and acceptance of generated requirements. Questionnaire data indicate that iterative feedback improves perceived requirement quality and alignment with project goals.

As illustrated in Figure 6a, six of nine participants (67%) approved the requirements after two feedback rounds ($M = 2.1, SD = 0.9$), indicating that two iterations were generally sufficient to reach an acceptable outcome. The interview data provide explanatory context for this pattern. Participants consistently reported needing two to three refinement cycles, with the exact number depending on the clarity of the initial input. As P2 explained, *"Most of the requirements are at a high level… some take two rounds, and for others more than three."* P3 similarly observed, *"For the small project, it was just two to three rounds."* P5 noted that unclear prompts increased iteration effort: *"If the input is vaguely described, the system stays at the same level. The better the preparation, the better the output."* They also observed that longer discussions sometimes led to *"losing track of missing points"* requiring additional iterations.

Table 6: Perceived improvements in requirements before and after human feedback

Requirement Quality Aspect	Before Feedback (M ± SD)	After Feedback (M ± SD)	Change in Mean
Reflects Intent	3.67 ± 0.47	4.11 ± 0.57	+0.44
Updated by Feedback	3.78 ± 0.63	3.89 ± 0.74	+0.11
Clarity and Wording	4.00 ± 0.47	3.67 ± 0.47	-0.33

Moreover, As shown in Figure 6b, missing requirements were reported by 5 participants, unnecessary requirements, completeness, and technical feasibility by 4 participants each, business feasibility by 3 participants, and clarity and wording by 1 participant. No participant mentioned correctness as a concern. These quantitative results suggest that while the system generated technically sound outputs, additional human input was necessary to ensure contextual completeness and feasibility.

Interview data further clarify why such refinements were required. We identified theme "Feedback-driven refinement" where P2 (Company A) explained, *"Some minor improvements will be there, like the technical feasibility should be there, and if I think about the correctness, most of the things are correct, but need to be aligned with the, you know, requirements. And so obviously the human involvement will be there."* Similarly, P3 (Company B) observed, *"Mostly, when the requirements were big enough, then it missed some technical points. So I needed to clarify it to come back to the context. Sometimes it loses focus, so that's what I did."* These insights indicate that participants viewed their role as ensuring contextual and technical alignment, particularly for larger or more complex requirements.

The questionnaire and interview data show that while the system produced accurate and well-structured outputs, users needed to verify and adapt them for organizational and technical relevance. Human review thus complemented system performance by addressing contextual completeness and ensuring feasibility across diverse project scopes.

Building on the previous findings on feedback rounds and quality aspects, the next analysis examines how participants perceived improvement in the generated requirements before and after human feedback. Figure 6c and Table 6 summarize these perceptions. The average score for "Reflects Intent" increased from $M = 3.67$ ($SD = 0.47$) to $M = 4.11$ ($SD = 0.57$), and "Updated by Feedback" rose from $M = 3.78$ ($SD = 0.63$) to $M = 3.89$ ($SD = 0.74$). "Clarity and Wording" decreased slightly from $M = 4.00$ ($SD = 0.47$) to $M = 3.67$ ($SD = 0.47$), indicating mixed perceptions of phrasing simplicity after refinement. Overall, the quantitative results suggest that participants perceived improvement through feedback, especially in terms of intent reflection and iterative updates.

Interview evidence provides further explanation for these patterns. Participants described the feedback process as useful for removing redundant or unclear requirements while improving overall accuracy. As P9 (Company D) stated, *"It can be time-consuming if there are duplicates. When I was generating some requirements, I got two that were almost the same, which required another round of editing. The human should be in the loop. It improved the output, for example by approving selected requirements or completely removing them from the system."* This comment illustrates how iterative human review contributed to improving output quality through selection and refinement.

Figure 6d shows agreement levels on prioritization usefulness and alignment. Five of nine participants agreed that prioritization outcomes aligned with organizational goals, while four remained neutral. Participants familiar with prioritization techniques (e.g., WSJF, Weighted Scoring, $100 allocation) reported better understanding of the rationale behind prioritization outcomes. Interview data further illustrate this point. As P7 (Company C)

(a) Number of feedback rounds before approval

(b) Quality aspects requiring human feedback

(c) Requirement improvements needing human feedback

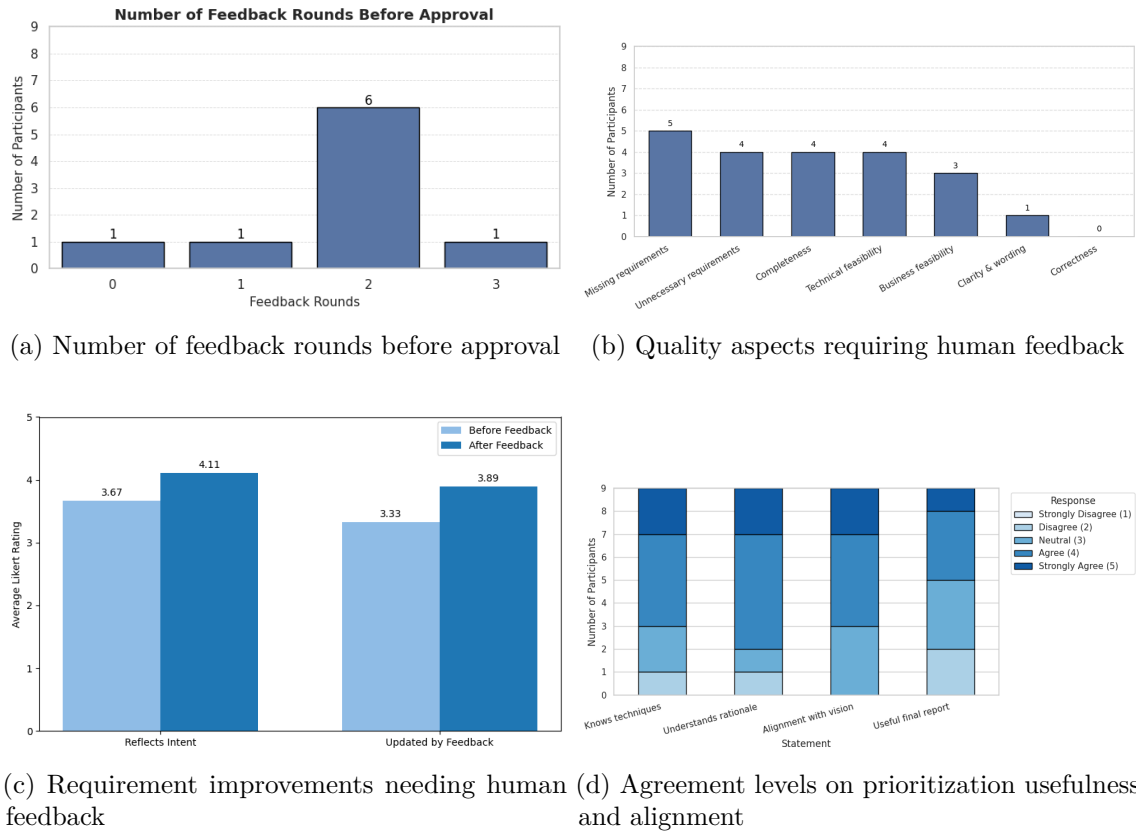(d) Agreement levels on prioritization usefulness and alignment

Figure 6: User-reported feedback cycles and quality aspects: (a) feedback rounds before approval, (b) quality aspects requiring human feedback, (c) requirement improvements, and (d) agreement levels on prioritization outcomes

stated, *"In some cases, it's good to have the human opinion in your decision-making process, of course. But it's sometimes good to have also that... nonhuman opinion, ... think again about your own idea of prioritization and gives another perspective."* This reflects how human-in-the-loop prioritization encouraged reflection and comparative reasoning, allowing participants to validate and reassess their prioritization logic in relation to the system's suggestions.

Overall, the questionnaire and interview data indicate a consistent pattern. Two feedback rounds were sufficient for most participants (6 of 9) to finalize requirements. Perceived quality improved after feedback, particularly in capturing intent and integrating user input. Interview insights confirm that human review addressed omissions and redundancies, enhanced completeness and feasibility, and guided prioritization. Collectively, the findings demonstrate that human–AI collaboration strengthened refinement while maintaining contextual accuracy and user control.

5.3. Challenges and suggestions (RQ3)

The analysis identified challenges and improvement areas for system adoption, as illustrated in Figure 7. Four participants selected *employee resistance to adoption* and *difficulty customizing the system* as the most frequent issues. *Lack of transparency in prioritization*, *trust in model output*, and *AI–human misalignment* were each mentioned by three participants.

17

Two participants reported *difficulty integrating feedback*, and one noted *too many refinement iterations*. These results indicate that challenges were mainly human and organizational rather than technical. The higher frequencies for resistance and customization suggest that usability and workflow alignment were central to adoption concerns.
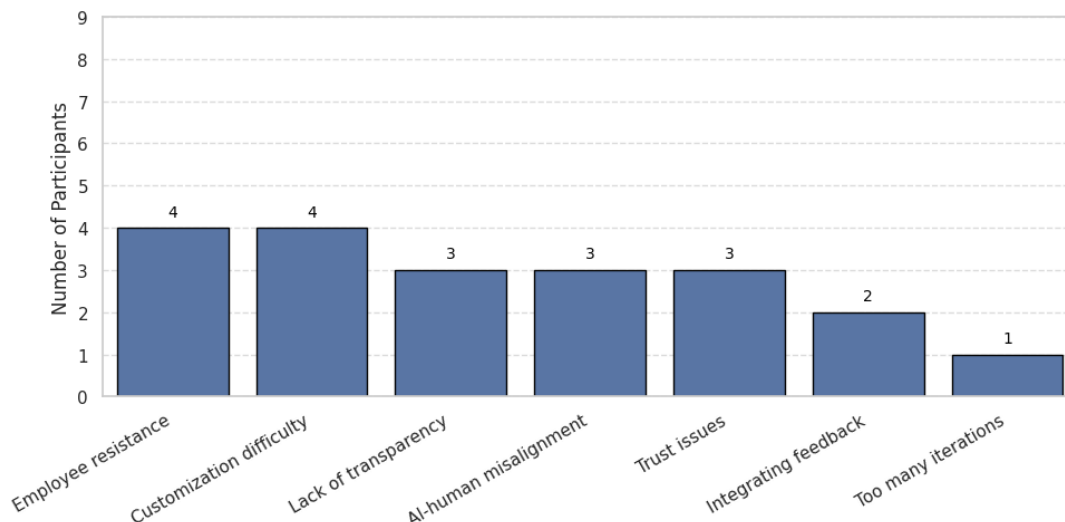

Figure 7: Most commonly reported challenges identified in the questionnaire

Interview findings supported these results and explained why they occurred. Three themes were identified: *usability and interface barriers*, *scalability and adaptability constraints*, and *trust and transparency limitations*. Participants also proposed practical improvements that align with these themes.

Usability and interface barriers. The interview data provide further detail on usability challenges. Participants explained that navigation and layout issues reduced efficiency during requirement review and prioritization. P2 (Company A) explained that the interface caused friction, noting, *"When I tried to assign weightings in the prioritization view, it did not work properly."* P3 (Company B) similarly stated that the system *"needs more development to reach a usable quality level."* These statements clarify that resistance and customization concerns reflected workflow interruptions rather than reluctance toward AI assistance. Participants suggested improving layout consistency, clearer task flow, and visual indicators for new results to enhance speed and reduce friction.

Scalability and adaptability constraints. Participants also discussed performance limitations when handling larger datasets. Although scalability appeared less frequently in questionnaires, it was discussed in every interview. P3 (Company B) noted, *"When I tested it for larger concepts, it lost focus. I guess it's a general issue with LLM."* P9 (Company D) added, *"If you're generating 100+ requirements, grouping or deduplication becomes hard. There needs to be more hierarchy."* These views indicate that the system performed effectively for smaller datasets but lost structure as data volume increased. Participants recommended hierarchy views, filters, and duplicate-detection tools to preserve clarity in large-scale projects.

Trust and transparency limitations. Interview participants raised concerns about transparency of reasoning behind prioritization outcomes. P3 (Company B) pointed out that *"It doesn't show how WSJF scores are calculated. That undermines the trust."* P2

(Company A) similarly stated, *"The team was concerned about how the agent was making decisions, especially in prioritization."* These insights help explain the medium frequency of trust-related issues reported in the questionnaire. Participants stressed the need for visible reasoning to sustain confidence in outputs and suggested adding intermediate calculations, manual adjustment options, and rationale logs to improve traceability and trust.

These interview findings reinforce the quantitative results shown in Figure 7. Participants valued the system's capability to analyze and prioritize requirements but emphasized that adoption required smoother interface design, better scalability, and greater transparency. Addressing these needs through improved layout, hierarchical organization, and clear explanation of reasoning processes was considered essential for broader organizational use.

6. Discussion

This section discusses the key findings by interpreting results in context, identifying implications for practice and research, and summarizing lessons learned. It also reflects on challenges and limitations observed during system use in four companies.

6.1. Interpretation of findings

The results indicate that the proposed multi-agent system that across the four companies, similar patterns were observed in how the system supported requirement generation, refinement, and prioritization. In all cases, participants confirmed that human oversight was necessary to ensure contextual and organizational accuracy. First, participants aligned AI-generated requirements with domain terminology and internal objectives. Second, requirements were adjusted for compliance and feasibility constraints that were not captured by prompts. Third, cross-department alignment required human judgment to maintain consistency with organizational priorities. These refinements indicate that human input focused on contextual interpretation rather than only on error correction. The feedback mechanism was consistently valued for enabling targeted revisions without extensive rework, typically within two feedback rounds. These commonalities suggest that human–AI collaboration was effective in maintaining alignment between generated outputs and organizational goals, regardless of company type or domain.

Differences across cases were linked mainly to organizational context, sector characteristics, and participant roles. The software consultancies (Companies A and B) emphasized efficiency and task-level integration into ongoing projects, reflecting their agile development environments. In contrast, the energy and logistics organizations (Companies C and D) focused on compliance, governance, and strategic alignment due to their regulated and large-scale operational contexts. Participant backgrounds also influenced these differences: technical leaders prioritized feasibility and iteration speed, whereas strategic and managerial roles emphasized explainability, traceability, and long-term value. These contextual factors shaped how participants perceived the system's strengths and areas for improvement.

Overall, it indicates that organizational maturity in AI adoption and participant expertise guided the nature of human feedback. While all companies recognized the system's value in structuring early requirement analysis, their emphasis differed according to internal objectives and operational settings. This variation demonstrates that human–AI collaboration in RE must adapt to sector-specific needs, governance structures, and user expectations to ensure reliable and contextually appropriate results.

6.2. Implications and lessons for practitioners

This study identifies practical aspects for applying LLM-based multi-agent systems in organizational AI planning and requirements analysis. A structured human–AI workflow supports requirement generation, refinement, and prioritization. Participants can approve, modify, delete, or request clarification for individual requirements. This controlled interaction maintains human oversight throughout, ensuring contextual alignment, compliance, and quality assurance while reducing effort for initial drafting.

Role-based configurable agents distribute work between specialized roles such as generation, verification, and prioritization. This configuration supports accountability and allows adaptation to different organizational contexts and stakeholder responsibilities.

Explainability and auditability should be embedded in system design to maintain trust and enable traceability of decisions [55]. Integration with existing planning and documentation environments can reduce adoption effort and support iterative improvement [18, 56]. Scalability and governance mechanisms are needed to preserve clarity and accountability in large projects. Human oversight remains central for contextual judgment, compliance, and verification in agentic workflows [57]. Organizations should define clear human roles that balance automation and accountability, ensuring that agent-generated outputs are reviewed and adjusted before implementation. For practitioners, the key lessons are: (i) multi-agent systems should function under defined human oversight, (ii) iterative feedback cycles should be part of the review process to improve output quality, (iii) transparency mechanisms should make intermediate reasoning visible for review, (iv) workflow compatibility and interface clarity should support organizational adoption, and (v) governance structures should ensure responsible and traceable decision-making.

These implications guide the development and use of transparent, auditable, and context-aware multi-agent systems for requirement analysis in industrial environments.

6.3. Implications for research

This study offers preliminary empirical insights into the operation of LLM-based multi-agent systems within real organizational contexts, extending earlier conceptual and experimental work [15, 58]. While prior studies focused mainly on technical feasibility, our findings illustrate how human-in-the-loop interaction and contextual adaptation may influence system usefulness and user acceptance. These results provide early empirical support for further investigation into AI-supported RE and decision-making in organizational settings.

The findings also complement existing theories of human–AI collaboration [59, 60], suggesting that user feedback can function both as a quality control mechanism and a means of trust calibration. The observed role of transparency is consistent with work on explainable AI [55], indicating that explainability-by-design remains an important research direction for RE tools. Future research should examine systematic approaches for tracing agent reasoning and assessing how visibility of decision logic affects user trust and adoption.

Finally, the study highlights the value of integrating socio-technical perspectives into the design of agentic systems. Role-based configuration and transparent feedback loops may provide a useful foundation for studying distributed reasoning and accountability in human–agent teams. These observations align with emerging industry perspectives that emphasize deliberate collaboration between humans and agents [57]. Further work should expand on these exploratory findings to develop structured evaluation methods and theoretical models connecting AI engineering and organizational research.

6.4. Limitations and future work

This study has several limitations that constrain interpretation. Current large language models remain limited by hallucination, weak domain adaptation, and incomplete reasoning transparency. Although human-in-the-loop validation reduced inconsistencies, generation errors and redundancy persisted, showing that the system cannot yet operate autonomously. Agent coordination was constrained by the absence of shared memory and adaptive communication, limiting collaborative reasoning and continuity across refinement cycles. The system also operated under short context windows, affecting traceability during iterative interaction. We use GPT-4o model for evaluation, restricting comparison across architectures and excluding analysis of model bias and fairness. The participant sample was small (nine questionnaires and four interviews across four organizations), limiting generalizability beyond similar contexts.

Future studies expand the participant base to improve representativeness and assess generalizability across industries. Integrating retrieval-augmented memory or extended-context models could improve reasoning continuity and traceability during multi-round refinement. Embedding explainability features such as reasoning logs, decision-tracing, or rationale visualization can strengthen interpretability and trust [55]. Comparative evaluation using multiple open and proprietary models is needed to assess robustness and reproducibility. Longitudinal and domain-specific studies can examine system adaptation over time and under regulatory constraints. Finally, developing adaptive coordination mechanisms [58] may enhance information sharing and collaborative reasoning among agents, supporting scalability in organizational deployment.

7. Threats to validity

There are a number of threats to the validity of our findings, despite our efforts to mitigate them. In reporting these, we adopt the classification scheme proposed by Runeson and Höst [50], which builds on Yin's work [61] and is widely used in software engineering case study research. The scheme includes four dimensions: construct validity, internal validity, external validity, and reliability. While originally developed for positivist studies, Runeson and Höst note that it can also be applied to flexible design studies. We therefore use it here to structure our validity discussion

7.1. Construct validity

The evaluation focused on a role-based multi-agent system using GPT-4o to support AI planning tasks. The questionnaire was aligned with the research questions and reviewed by multiple authors for coverage and clarity. For requirements-related quality aspects, constructs such as relevance, completeness, and refinement were informed by ISO/IEC/IEEE 29148:2011 guidelines. Triangulation with qualitative interview data supported the interpretation of participant responses. Variation in participant roles and company contexts helped ensure consistent application of the constructs. Future studies will include more companies, a larger and more diverse participant pool, and additional stakeholder roles to evaluate construct stability across different organizational settings.

7.2. Internal validity

The study conducted to evaluate LLM-based multi-agent system in companies for strategic AI planning. Participant responses may have been influenced by expectations, organizational roles, or group dynamics during planning activities. Company-specific factors such as team composition, AI maturity, and strategic priorities may also have shaped perceptions of the system. We used a standardized task structure and fixed agent prompts across all case studies to reduce variation. Data triangulation across questionnaires, and interviews reduce interpretation bias. Similar patterns across companies suggest findings are not context-specific. No detailed interaction logs captured; future work wil incorporate these to support behavioral analysis. Furthermore, researcher bias represents an additional internal validity threat. The authors participated in both system design and evaluation, which may influence the interpretation of participant feedback. Anonymization and analysis help reduce such effects, although complete neutrality cannot be ensured.

7.3. External validity

The study included four case companies from different industrial sectors, and the consistency of their feedback strengthens the external validity and overall rigor. Each company contributed post-session questionnaire responses and one follow-up interview. Although participants held relevant roles in AI planning, the small sample size and early-stage deployment context limit generalizability. The limited number of participants (nine questionnaires and four interviews) further constrains external validity. Future studies with broader representation, including small and medium-sized enterprises, would enhance generalizability and allow deeper assessment of transferability across industrial contexts.

7.4. Reliability

System configuration, agent roles, task prompts, and instruments were consistent across all cases. The first author conducted and thematically analyzed all interviews, while questionnaire data were examined using descriptive statistics. All protocols were designed with the second and other authors, though minor variation in facilitation may have occurred. A shared coding scheme was applied to reduce subjectivity. Future studies could strengthen reliability by using automated instrumentation and expanding data collection and analysis.

7.5. Ethical concerns

All respondents and case companies were anonymized to protect confidentiality and encourage honest feedback. Only essential contextual details were retained. This reduced the risk of biased or overly positive responses. One participating company also signed a Non-Disclosure Agreement (NDA) to ensure proper handling of sensitive data.

8. Conclusions

This study examined a LLM-based multi-agent system for requirements analysis in the context of strategic AI adoption. The system assigns stakeholder-inspired roles that generate, refine, and prioritize requirements in collaboration with human users. Evidence from

a multi-case study involving four companies suggests that the system can support early requirement structuring and facilitate planning discussions. Participants valued the ability to review agent-generated drafts but emphasized that human oversight remained necessary to ensure contextual relevance, compliance alignment, and strategic fit. Reported challenges included limited transparency of prioritization logic, difficulty in adapting outputs to organizational standards, and the lack of integration with existing enterprise tools.

Overall, the results indicate that such a system could assist early planning activities by reducing manual effort and providing structured starting points for discussion, but it does not replace human judgment or domain expertise. Future research investigate transparency mechanisms such as decision logs and traceable reasoning, explore collaborative multi-user operation, and examine integration with enterprise platforms. Broader evaluations across diverse organizations may provide stronger evidence on the system's practical adoption, scalability, and organizational readiness.

CRediT authorship contribution statement

M. A. Sami: system implementation, methodology, data collection and analysis, writing the original draft. Z. Zhang: conceptualization, methodology, review and editing, supervision. M. Waseem: conceptualization, methodology, review and editing, supervision. K-K. Kemmel: conceptualization, methodology, review and editing, supervision. Z. Rasheed: review and editing. T. Herda: conceptualization, review and editing. P. Abrahamsson: conceptualization, methodology, review and editing, supervision.

Declaration of competing interest

One of the authors is affiliated with an industrial partner organization. However, the research was conducted independently, and the company was not involved in the design, data analysis, or interpretation of the findings. The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data and code availability

The source code, experimental scripts, and data used in this study (excluding confidential data) are publicly available on GitHub at
https://github.com/GPT-Laboratory/multiagent-story-generation.

Funding

Declaration of AI assistance in the writing process

ChatGPT and Grammarly were used for language editing and summarization to improve clarity and grammar. No generative AI was used for figures. All AI-assisted content was reviewed and finalized by the authors.

References

[1] A. Dasgupta and S. Wendler, "Ai adoption strategies," Working Paper Series No. 9, Centre for Technology and Global Affairs, University of Oxford, March 2019. [Online]. https://www.politics.ox.ac.uk/sites/default/files/2022-03/201903-CTGA-Dasgupta%20A-Wendler%20S-aiadoptionstrategies.pdf

[2] S. Kaggwa, T.F. Eleogu, F. Okonkwo, O.A. Farayola, P.U. Uwaoma et al., "Ai in decision making: transforming business strategies," *International Journal of Research and Scientific Innovation*, Vol. 10, No. 12, 2024, pp. 423–444.

[3] Gartner, "Gartner survey finds generative ai is now the most frequently deployed ai solution in organizations," https://www.gartner.com/en/newsroom/press-releases/2024-05-07-gartner-survey-finds-generative-ai-is-now-the-most-frequently-deployed-ai-solution-in-organizations, 2024, accessed: 2025-04-06.

[4] Gartner, "Gartner survey finds 79% of corporate strategists see ai and analytics as critical to their success over the next two years," 2023. [Online]. https://www.gartner.com/en/newsroom/press-releases/2023-07-05-gartner-survey-finds-79-percent-of-corporate-strategists-see-ai-and-analytics-as-critical-to-their-success-over-the-next-two-years

[5] K.K. Kemell, M. Saarikallio, A. Nguyen-Duc, and P. Abrahamsson, "Still just personal assistants? a multiple case study of generative ai adoption in software organizations," *Information and Software Technology*, Vol. 186, 2025, p. 107805.

[6] D. Russo, "Navigating the complexity of generative ai adoption in software engineering," *ACM Trans. Softw. Eng. Methodol.*, Vol. 33, No. 5, Jun. 2024. [Online]. https://doi.org/10.1145/3652154

[7] D. Herremans, "aistrom–a roadmap for developing a successful ai strategy," *IEEE access*, Vol. 9, 2021, pp. 155 826–155 838.

[8] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements practices and gaps when engineering human-centered artificial intelligence systems," *Applied Soft Computing*, Vol. 143, 2023, p. 110421.

[9] A. Choudhury and H. Shamszare, "Investigating the impact of user trust on the adoption and use of chatgpt: survey analysis," *Journal of Medical Internet Research*, Vol. 25, 2023, p. e47184.

[10] F. Zambonelli, N.R. Jennings, and M. Wooldridge, "Developing multiagent systems: The gaia methodology," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, Vol. 12, No. 3, 2003, pp. 317–370.

[11] U.e. Habiba, M. Haug, J. Bogner, and S. Wagner, "How mature is requirements engineering for ai-based systems? a systematic mapping study on practices, challenges, and future research directions," *Requirements Engineering*, 2024, pp. 1–34.

[12] X. Franch, A. Jedlitschka, and S. Martínez-Fernández, "A requirements engineering perspective to ai-based systems development: A vision paper," in *international working conference on requirements engineering: foundation for software quality*. Springer, 2023, pp. 223–232.

[13] A. Mehraj, Z. Zhang, and K. Systä, "A tertiary study on ai for requirements engineering," in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2024, pp. 159–177.

[14] Z. Zhang, M. Rayhan, T. Herda, M. Goisauf, and P. Abrahamsson, "Llm-based agents for automating the enhancement of user story quality: An early report," in *International Conference on Agile Software Development*. Springer Nature Switzerland Cham, 2024, pp. 117–126.

[15] D. Jin, Z. Jin, X. Chen, and C. Wang, "Mare: Multi-agents collaboration framework for requirements engineering," *arXiv preprint arXiv:2405.03256*, 2024.

[16] P. Bornet, J. Wirtz, T.H. Davenport, D.D. Cremer, B. Evergreen et al., *Agentic Artificial Intelligence: Harnessing AI Agents to Reinvent Business, Work, and Life*, kindle edition ed. Singapore: World Scientific Publishing, March 2025, available on Amazon Kindle, 572 pages. [Online]. https://www.amazon.com/dp/B0CQYZN7QM

[17] M.A. Sami, Z. Zhang, M. Waseem, K.K. Kemell, Z. Rasheed et al., "A multi-agent llm system for automated requirements analysis: A study on user story generation and prioritization," in *Euromicro Conference on Software Engineering and Advanced Applications*. Springer, 2025, pp. 178–187.

[18] M.C. Lee, H. Scheepers, A.K. Lui, and E.W. Ngai, "The implementation of artificial intelligence in organizations: A systematic literature review," *Information & Management*, Vol. 60, No. 5, 2023, p. 103816.

[19] P. Hamm and M. Klesel, "Success factors for the adoption of artificial intelligence in organizations: A literature review," in *27th Americas Conference on Information Systems, AMCIS 2021: Digital Innovation and Entrepreneurship*. Association for Information Systems, 2021.

[20] D. Smit, S. Eybers, and A. van der Merwe, "Towards human-ai symbiosis: Designing an artificial intelligence adoption framework," *South African Computer Journal*, Vol. 36, No. 1, 2024, pp. 76–104.

[21] A. Tursunbayeva and H.C.B. Gal, "Adoption of artificial intelligence: A top framework-based checklist for digital leaders," *Business Horizons*, Vol. 67, No. 4, 2024, pp. 357–368.

[22] J. Yang, Y. Blount, and A. Amrollahi, "Artificial intelligence adoption in a professional service industry: A multiple case study," *Technological Forecasting and Social Change*, Vol. 201, 2024, p. 123251.

[23] F. Selten and B. Klievink, "Organizing public sector ai adoption: Navigating between separation and integration," *Government Information Quarterly*, Vol. 41, No. 1, 2024, p. 101885.

[24] G. Vaz Pereira, V. Jackson, R. Prikladnicki, A. van der Hoek, L. Fortes et al., "Exploring genai in software development: Insights from a case study in a large brazilian company," in *IEEE International Conference on Software Engineering (ICSE)*, 2025.

[25] J. Flyckt, T. Gorschek, D. Mendez, and N. Lavesson, "Identifying key ai challenges in make-to-order manufacturing organisations: A multiple case study," *Journal of Systems and Software*, Vol. 230, 2025, p. 112559.

[26] Y. Feng, S. Vanam, M. Cherukupally, W. Zheng, M. Qiu et al., "Investigating code generation performance of chat-gpt with crowdsourcing social data," in *Proceedings of the 47th IEEE Computer Software and Applications Conference*, 2023, pp. 1–10.

[27] F.A. Shah, A. Sabir, R. Sharma, and D. Pfahl, "How effectively do llms extract feature-sentiment pairs from app reviews?" in *International Working Conference on Requirements Engineering: Foundation for Software Quality*. Springer, 2025, pp. 123–138.

[28] A. Ferrari and P. Spoletini, "Formal requirements engineering and large language models: A two-way roadmap," *Information and Software Technology*, Vol. 181, 2025, p. 107697. [Online]. https://www.sciencedirect.com/science/article/pii/S0950584925000369

[29] O. Abed, K. Nebe, and A.B. Abdellatif, "Ai-generated user stories supporting human-centred development: an investigation on quality," in *International Conference on Human-Computer Interaction*. Springer, 2024, pp. 3–13.

[30] L. Zhong, Z. Wang, and J. Shang, "Ldb: A large language model debugger via verifying runtime execution step-by-step," *arXiv preprint arXiv:2402.16906*, 2024.

[31] B. Wei, "Requirements are all you need: From requirements to code with llms," in *2024 IEEE 32nd International Requirements Engineering Conference (RE)*. IEEE, 2024, pp. 416–422.

[32] S. Manish, "An autonomous multi-agent llm framework for agile software development," *International Journal of Trend in Scientific Research and Development*, Vol. 8, No. 5, 2024, pp. 892–898.

[33] C. Hymel and H. Johnson, "Analysis of llms vs human experts in requirements engineering," *arXiv preprint arXiv:2501.19297*, 2025.

[34] A. Ferrari and P. Spoletini, "Formal requirements engineering and large language models: A two-way roadmap," *Information and Software Technology*, Vol. 181, 2025, p. 107697.

[35] C. Zhang, K. Yang, S. Hu, Z. Wang, G. Li et al., "Proagent: building proactive cooperative agents with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 16, 2024, pp. 17 591–17 599.

[36] X. Zou, Y. Liu, X. Shi, and C. Yang, "Goal2story: A multi-agent fleet based on privately enabled sllms for impacting mapping on requirements elicitation," *arXiv preprint arXiv:2503.13279*, 2025.

[37] W. Tao, Y. Zhou, Y. Wang, W. Zhang, H. Zhang et al., "Magis: Llm-based multi-agent framework for github issue resolution," *Advances in Neural Information Processing Systems*, Vol. 37, 2024, pp. 51 963–51 993.

[38] J. Liu, G. Wang, R. Yang, J. Zeng, M. Zhao et al., "Rtadev: Intention aligned multi-agent framework for software development," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 1548–1581.

[39] H. Takagi, S. Moriya, T. Sato, M. Nagao, and K. Higuchi, "A framework for efficient development and debugging of role-playing agents with large language models," in *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025, pp. 70–88.

[40] K.T. Tran, D. Dao, M.D. Nguyen, Q.V. Pham, B. O'Sullivan et al., "Multi-agent collaboration mechanisms: A survey of llms," *arXiv preprint arXiv:2501.06322*, 2025.

[41] A. Singh, A. Ehtesham, S. Kumar, and T.T. Khoei, "Enhancing ai systems with agentic workflows patterns in large language model," in *2024 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2024, pp. 527–532.

[42] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li et al., "Autogen: Enabling next-gen llm applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[43] N.A. Parikh, "Managing ai-first products: Roles, skills, challenges, and strategies of ai product managers," *IEEE Engineering Management Review*, 2025, pp. 1–11.

[44] N. Tasneem, H.B. Zulzalil, and S. Hassan, "Enhancing agile software development: A systematic literature review of requirement prioritization and reprioritization techniques," *IEEE Access*, 2025.

[45] F. Essam, H. El, and S.R.H. Ali, "A comparison of the pearson, spearman rank and kendall tau correlation coefficients using quantitative variables," *Asian J. Probab. Stat*, 2022, pp. 36–48.

[46] M.A. Sami, M. Waseem, Z. Zhang, Z. Rasheed, K. Systä et al., "Early results of an ai multiagent system for requirements elicitation and analysis," in *Product-Focused Software Process Improvement*, D. Pfahl, J. Gonzalez Huerta, J. Klünder, and H. Anwar, Eds. Cham: Springer Nature Switzerland, 2025, pp. 307–316.

[47] P. Murley, Z. Ma, J. Mason, M. Bailey, and A. Kharraz, "Websocket adoption and the landscape of the real-time web," in *Proceedings of the Web Conference 2021*, 2021, pp. 1192–1203.

[48] B. Lubanovic, *FastAPI*. " O'Reilly Media, Inc.", 2023.

[49] A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[50] P. Runeson and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Empirical software engineering*, Vol. 14, 2009, pp. 131–164.

[51] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, Vol. 3, No. 2, 2006, pp. 77–101.

[52] R.K. Yin, "Case study research and applications: Design and methods," *SAGE Publications*, 2018.

[53] I. Etikan, S.A. Musa, R.S. Alkassim et al., "Comparison of convenience sampling and purposive sampling," *American journal of theoretical and applied statistics*, Vol. 5, No. 1, 2016, pp. 1–4.

[54] I. Garcia, C. Pacheco, A. León, and J.A. Calvo-Manzano, "A serious game for teaching the fundamentals of iso/iec/ieee 29148 systems and software engineering–lifecycle processes–requirements engineering at undergraduate level," *Computer Standards & Interfaces*, Vol. 67, 2020, p. 103377.

[55] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng et al., "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, Vol. 15, No. 2, 2024, pp. 1–38.

[56] D. Russo, "Navigating the complexity of generative ai adoption in software engineering," *ACM Transactions on Software Engineering and Methodology*, Vol. 33, No. 5, 2024, pp. 1–50.

[57] McKinsey & Company, "One year of agentic ai: Six lessons from the people doing the work," https://www.mckinsey.com/capabilities/quantumblack/our-insights/one-year-of-agentic-ai-six-lessons-from-the-people-doing-the-work, 2025, accessed: 2025-10-13.

[58] J. He, C. Treude, and D. Lo, "Llm-based multi-agent systems for software engineering: Literature review, vision and the road ahead," *ACM Transactions on Software Engineering and Methodology*, 2024.

[59] S. Amershi, A. Begel, C. Bird, R. DeLine, G. Gall et al., "Software engineering for machine learning: A case study," *International Conference on Software Engineering (ICSE)*, 2019, pp. 291–300. [Online]. https://doi.org/10.1109/ICSE.2019.00042

[60] C.O. Retzlaff, S. Das, C. Wayllace, P. Mousavi, M. Afshari et al., "Human-in-the-loop reinforcement learning: A survey and position on requirements, challenges, and opportunities," *Journal of Artificial Intelligence Research*, Vol. 79, 2024, pp. 359–415.

[61] R.K. Yin, *Case Study Research: Design and Methods*, 5th ed. Thousand Oaks, CA: Sage Publications, 2014.

Authors and affiliations

Malik Abdul Sami
e-mail: malik.sami@tuni.fi
ORCID: https://orcid.org/0000-0001-5136-2587
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Muhammad Waseem
e-mail: muhammad.waseem@tuni.fi
ORCID: https://orcid.org/0000-0001-7488-2577
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Zeeshan Rasheed
e-mail: zeeshan.rasheed@tuni.fi
ORCID: https://orcid.org/0000-0001-9655-3096
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Pekka Abrahamsson
e-mail: pekka.abrahamsson@tuni.fi
ORCID: https://orcid.org/0000-0002-4360-2226
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Zheying Zhang
e-mail: zheying.zhang@tuni.fi
ORCID: https://orcid.org/0000-0002-6205-4210
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Kai-Kristian Kemell
e-mail: kai-kristian.kemell@tuni.fi
ORCID: https://orcid.org/0000-0002-0225-4560
Faculty of Information Technology and
Communication Sciences, Tampere University,
Finland

Tomas Herda
e-mail: herda.tom@gmail.com
ORCID: https://orcid.org/0009-0005-2912-380X
AI Center of Excellence, Austrian Post, Vienna,
Austria