



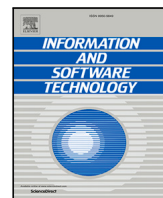
Recommendations for efficient and responsible LLM adoption within industrial software development

Downloaded from: <https://research.chalmers.se>, 2026-06-18 14:42 UTC

Citation for the original published paper (version of record):

Ronanki, K., Cabrero-Daniel, B., Herda, T. et al (2026). Recommendations for efficient and responsible LLM adoption within industrial software development. *Information and Software Technology*, 196. <http://dx.doi.org/10.1016/j.infsof.2026.108171>

N.B. When citing this work, cite the original published paper.



Recommendations for efficient and responsible LLM adoption within industrial software development

Krishna Ronanki ^a ^{*}, Beatriz Cabrero-Daniel ^a , Tomas Herda ^b , Stefan Sitkovich ^b, Jennifer Horkoff ^a, Christian Berger ^a 

^a Chalmers University of Technology | University of Gothenburg, Gothenburg, Sweden

^b Austrian Post, Vienna, Austria

ARTICLE INFO

Keywords:

Large language models
Software engineering
Trustworthy AI

ABSTRACT

Context: Large language models (LLMs) are observed to have a significant positive impact on various software engineering (SE) activities. With improved accessibility, the adoption of powerful LLMs in industry has surged recently. However, there is a lack of actionable best practices for the efficient and responsible adoption of LLMs within industrial software settings.

Objectives: We developed seven actionable recommendations to address this research gap.

Methods: We conducted a multi-case study with three organisations that use LLMs within their SE activities and synthesised seven recommendations through qualitative thematic analysis. We conducted a complementary online survey with software practitioners from various industries to evaluate the perceived relevance of our recommendations.

Results: Our results and recommendations focus on (i) users' preference to use LLMs as AI assistants, (ii) the importance of relevant stakeholders' satisfaction in the LLM-output evaluation, (iii) scoping the applicability of LLMs within SE tasks, (iv) the effect of LLMs on SE workflows, (v) the necessity and directions for developing human oversight mechanisms, and (vi) the necessary skills for practitioners for leveraging LLMs within SE. The online survey indicates a high level of agreement from the participants regarding the perceived relevance of the recommendations.

Conclusion: We outline future research directions, including mapping the seven recommendations to the principles of the EU AI Act (AIA) in order to examine how they relate to the current regulatory compliance frameworks.

1. Introduction

Many state-of-the-art LLMs are demonstrating increasingly impressive capabilities in performing a wide range of tasks [1], including tasks within software engineering (SE), and are observed to increase user productivity [2]. The adoption of LLMs in industrial SE settings is observed to be growing thanks to easy-to-use interfaces that make the LLMs more accessible to practitioners of all backgrounds [3].

Despite their ease of use, there are a few impediments in using LLMs for SE in practice such as confidential data privacy, extrinsic hallucinations [4], and the lack of best practices [5]. Third party LLM-based solutions such as Azure OpenAI services [6], Microsoft Copilot [7], GitHub Copilot [8], ChatGPT enterprise [9] aim to mitigate some of these issues.

However, that still leaves the shortage of best practices for using LLMs in industrial SE as an open challenge. This points towards the

need for more empirical research on LLM-based SE [5]. To address this gap, we define the research problem addressed in this paper as follows:

Users and organisations lack actionable recommendations to leverage LLMs to assist them in their SE activities in an AIA-compliant manner.

Although studies such as Fan et al. [5] and Hou et al. [10] reinforce the benefits and highlight the challenges of leveraging LLMs within SE, the insights presented are based on secondary evidence-based studies [11] or controlled empirical experiments [12] that do not fully capture the nuances of adopting LLM in industrial settings. Hence, we see a clear need for an empirical study that focuses on directly capturing practitioners' experiences regarding the adoption of LLMs

* Corresponding author.

E-mail address: krishna.ronanki@gu.se (K. Ronanki).

<https://doi.org/10.1016/j.infsof.2026.108171>

Received 1 July 2025; Received in revised form 24 April 2026; Accepted 27 April 2026

Available online 28 April 2026

0950-5849/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

within industrial SE environments. With the software industry’s increasing interest in leveraging LLMs in their processes, more empirical evidence collected from real-world environments is required to devise and understand the mechanisms that facilitate efficient and responsible adoption of LLMs in SE [5].

To that extent, we conducted an interview-based multi-case study with three organisations to gather empirical evidence regarding the adoption of LLM-based tools and services to assist users in their SE activities. We conducted a total of fifteen interviews with eighteen different participants across the three organisations. We collected qualitative data regarding their insights on their motivation, intentions and process of integrating LLMs in various SE activities, as well as the advantages and limitations of adopting LLMs within industrial settings. Based on these findings, we developed seven recommendations for the adoption of LLMs for SE within industrial settings.

However, findings of a multi-case study are limited to the context of the case companies under investigation and the recommendations are potentially not directly generalisable to broader contexts of adopting LLMs for SE activities. Therefore, we complemented our multi-case study results with a broader survey with software practitioners to assess their level of agreement with the recommendations that we synthesised from our study. We performed this survey to investigate the perceived relevance of our recommendations, as the software practitioners we surveyed were sampled from various industries, who, self-reportedly, have experienced LLM adoption within various SE tasks under different contexts. We conclude our study with a post hoc mapping of the validated recommendations with the seven key trustworthy AI principles proposed by the European Commission’s High Level Expert Group on AI (AI HLEG). Based on the results of this process, the contributions of this study are as follows:

- Findings surrounding the benefits, concerns, challenges, limitations, and experiential learnings of LLM-assisted SE in practice.
- Seven recommendations for the efficient and responsible adoption of LLMs in industrial SE activities.
- Assessing the perceived relevance of the recommendations beyond the context of the case companies.

The rest of the paper is organised as follows. Section 2 covers the background concepts relevant to our study. Section 3 reviews related works that inform our research motivation and provide a basis for comparing the results of our study. Section 4 describes our research design. Section 5 presents the results and analysis of the multi-case study we conducted. Section 6 presents the seven proposed recommendations. Section 7 presents the practitioner’s perception of the proposed recommendations. We present our post hoc mapping of the validated recommendations with the seven AI HLEG’s trustworthy AI principles in Section 8. We discuss the implications and the validity threats to our results in Section 9. Finally, we conclude our study in Section 10 and present potential areas for further research.

2. Background

LLMs, often built upon deep learning techniques like transformers, can produce useful natural language outputs. This led to them being employed in various language-related tasks such as text generation [13], question answering [14], translation [1], summarisation [15], and sentiment analysis [16]. The application of LLMs for automating or assisting users within certain SE practices is an emerging area of interest. This is indicated by the studies conducted by Fan et al. [5] and Hou et al. [10] who, respectively, presented a survey and systematic literature review on the application of LLMs within SE.

2.1. Retrieval augmented generation

While LLMs have achieved remarkable success, they still encounter some substantial limitations. These are particularly evident in tasks that are specialised or require extensive knowledge [17]. Notably, they tend to generate misleading or false information termed as “hallucinations” [18] when dealing with inquiries that exceed their training data or necessitate up-to-date information.

To address this limitation, recent research has explored the integration of both parametric and non-parametric memory into LLMs. Parametric memory refers to the learned parameters within the model, while non-parametric memory typically involves accessing external knowledge sources such as large text corpora or databases. One approach to combining these types of memory is through a method called Retrieval Augmented Generation (RAG). RAG endows pre-trained parametric-memory generation models, such as transformers, with non-parametric memory by incorporating a dense vector index of external knowledge. This integration is achieved through a general-purpose fine-tuning approach, allowing the model to access extensive knowledge during inference without additional training [19].

2.2. Trustworthy AI guidelines

In order to achieve trustworthy AI in practice, the AI HLEG proposed seven key principles that must be continuously assessed and managed throughout the entire lifecycle of an AI system: *Human Agency and Oversight, Technical Robustness and Safety, Privacy and Data Governance, Transparency, Diversity, Non-discrimination, and Fairness, Societal and Environmental Well-being, and Accountability* [20,21]. These principles are not legally binding under the AIA but are referenced in Recital (27) as voluntary guidance for developing best practices and standards that foster trustworthy, human-centric AI [21].

It is beneficial to ensure all actors involved in LLM-assisted SE tasks or processes comply with the trustworthy AI principles. According to the AIA, “any natural or legal person, including a public authority, agency or other body, using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity”, is referred to as a deployer [21]. Under this definition, employees of software organisations who participate in the design, development, deployment and maintenance of a software system can be categorised as deployers if they use the assistance of an LLM. As such, we hypothesise that adhering to AI HLEG’s seven principles during the adoption and integration of LLMs within SE workflows can aid the deployers in aiming for trustworthiness.

3. Related work

In this section, we provide a brief summary of the state-of-the-art literature relevant to our study, particularly looking at empirical studies applying LLMs in SE contexts.

Pereira et al. [22] conducted an industrial case study with a large media group that has recently begun to adopt OpenAI ChatGPT and GitHub Copilot for software development activities and provide early insights into potential benefits and concerns in the form of eight initial lessons learnt. These eight lessons are centred around aspects such as generative AI’s impact on a user’s learning, improving unit testing, the importance of context and the user’s domain expertise when using LLMs and other potential benefits and challenges surrounding the usage of LLMs within SE.

Similarly, Süße et al. [23] conducted a case study, focusing on how employees in SE perceive the collaboration with AI-powered chatbots such as ChatGPT. They identified fourteen distinct insights into the perceived collaboration with AI-powered chatbots in the case company’s software development context: Developing a general understanding on how AI works, possessing expertise on the task or topic to which the AI agent is assigned, interpreting and evaluating AI agent’s outputs

context-specifically, determining the division of tasks between the AI and oneself, dealing with the AI agent’s outputs in a reflective manner, complying with data protection rules, considering the AI agent as an enabler, considering the AI agent as a sort of a virtual colleague, being able to adapt and be open for change and innovation, feeling confident to work with new and unfamiliar technologies, complying with ethical and moral standards, appreciating the AI agent’s support, engaging oneself in a constant discourse with the AI agent, and expressing oneself comprehensibly towards the AI agent.

Wang et al. [24] conducted a mixed methods field study to incorporate LLMs into the vulnerability remediation process effectively. As part of their study, they design, implement, and empirically validate an LLM-supported collaborative vulnerability remediation process. The lessons learnt from this study are centred around incorporating LLMs into practical processes, facilitating collaboration among all associated stakeholders, reshaping LLMs’ roles according to task complexity, and how to approach the short-term side effects of improved user engagement facilitated by LLMs.

While these studies offer valuable early perspectives drawn from individual organisational contexts, they primarily adopt a descriptive stance, capturing lessons learnt and practitioner perceptions regarding the adoption and integration of LLMs in SE. However, as organisations increasingly explore operational adoption, there remains a need for prescriptive, actionable guidance that extend beyond descriptive observations to support efficient and responsible adoption in practice. Addressing this gap, our study developed and validated a set of recommendations to guide the adoption of LLMs within SE organisations. We also compare and contrast the contributions of our study with the lessons learnt presented by Pereira et al. [22], Süße et al. [23], and Wang et al. [24] in Section 9.

4. Methodology

4.1. Overview

Fig. 1 is a visual representation of this study’s design. We employed a mixed methods approach as part of our research design. We provide an overview of our study design in this section, with additional details regarding the description of the case companies along with the data collection and analysis activities in Sections 4.2, 4.3, 4.4, 4.5, and 4.6 respectively.

We formulated the following research questions (RQs) to obtain our study’s contributions:

RQ1. What are the benefits, challenges and experiential learnings of practitioners concerning LLMs’ assistance in their SE activities?

Justification: The insights from prior studies have mainly been derived from individual case studies conducted in specific organisational contexts. This leaves a need for further empirical work that examines whether similar benefits, challenges, and experiential learnings can be observed across multiple cases. Building on these findings, our RQ1 examines practitioners’ reported benefits, challenges, and experiential learnings from using LLMs in SE activities.

RQ2. What are recommendations for the adoption of LLMs’ assistance for SE activities?

Justification: Existing work mainly reports lessons learnt and practitioner perceptions [22–24]. While these studies improve understanding of LLM use in practice, they provide limited explicit guidance for adoption. Based on the experiences captured in RQ1, our RQ2 focuses on deriving recommendations for adopting LLM assistance in SE activities.

Table 1

Participant roles, LLM-based tools and LLM-assisted use cases of case 1.

Roles	Use cases	LLM tools used
Requirements Engineer, Scrum Master, Product Owner, SE Team Lead	User Stories generation, Stakeholder Analysis, Artefact generation	GitHub Copilot, Microsoft Copilot, RAG POC tool

RQ3. What is the practitioners’ view on the perceived relevance of the proposed recommendations?

Justification: Prior studies suggest that the use of LLMs in SE is shaped by context, practitioner expertise, and task characteristics [23, 24]. For this reason, proposed recommendations should also be examined from the perspective of practitioners. Our RQ3 therefore investigates practitioners’ views on the perceived relevance of the proposed recommendations.

We conducted the interview-based multi-case study following the guidelines for conducting and reporting case study research in SE by Runeson and Höst [25] to answer RQ1. We conducted a total of fifteen interviews with eighteen participants across three cases. Case 1 had five individual interviews and four group interviews. The group interviews were conducted with three people in each. Case 2 had four individual interviews while case 3 had two individual interviews.

We performed reflexive thematic analysis to answer RQ2. We followed the guidelines for conducting thematic synthesis proposed by Clarke and Braun [26] on the data collected from all fifteen interviews, which resulted in six primary themes having 13 codes. Based on the results of the thematic analysis, we synthesised seven recommendations. This was followed by a survey to assess the perceived relevance of these recommendations beyond the contexts of the case companies to answer RQ3. Together, the multi-case study and the survey addressed the overarching goal of synthesising and validating recommendations for the efficient and responsible adoption of LLMs within SE.

After answering the RQs, we conducted a post-hoc mapping of our seven recommendations to the seven key principles for trustworthy AI outlined by the AI HLEG. This mapping was done to assess how the recommendations relate to the EU AI Act (AIA).

The multi-case study provided in-depth findings by focusing on unique aspects of real-world adoption of LLMs. The online survey broadened the scope by involving practitioners from diverse organisational and contextual backgrounds. This approach validated the applicability and relevance of the recommendations across a variety of scenarios. The multi-case study acted as an exploratory phase, uncovering key themes and actionable recommendations grounded in the experiences of the case companies. The survey was conducted as a validation phase, assessing the perceived relevance of the recommendations across a diverse group of practitioners, reinforcing their reliability.

4.2. Case descriptions

4.2.1. Case 1

Case company 1 is an Europe-headquartered international postal and logistics service provider, whose main business activities include shipping and delivering letter mail, advertising mail, print publications and parcels. The case company’s IT division is responsible for the design, development, deployment, maintenance and operation of a suite of software products and services that support the case company’s logistical operations. The roles of the participants, the LLM-based tools used and the LLM-assisted SE use cases within case company 1 are presented in Table 1.

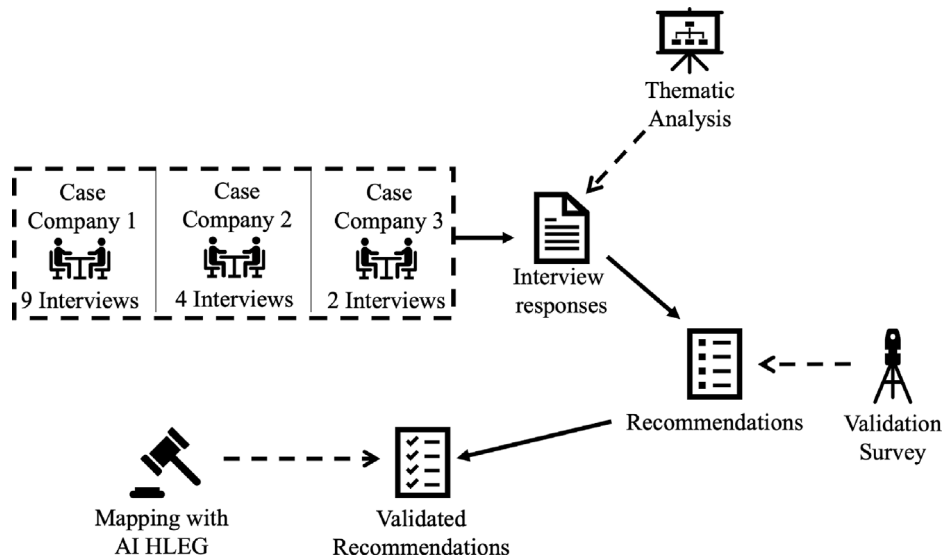


Fig. 1. Methodology. The dotted arrows represent the means of data collection or analysis while the solid arrows point to the outcome of the applied method.

Table 2
Participant roles, LLM-based tools and LLM-assisted use cases of case 2.

Roles	Use cases	LLM tools used
System Manager, Product Owner, Developer and Scrum Master, Customer Product Information (CPI) Architect	Software documentation, coding assistance, information search and retrieval, brainstorming and ideation, scrum activity support	GitHub Copilot, Codeium, Internal LLM tool, ChatGPT

Case company 1 had their own proof-of-concept RAG implementation called **RAG POC tool**, which was also tested within their workshop for generating the stakeholder analysis artefact. The stakeholder analysis artefact (case company specific artefact) serves to identify stakeholders (a person or an organisation who influences a system’s requirements or who is impacted by that system) and assigns them clear responsibilities. The **RAG POC tool** was developed within case company 1 using Microsoft Azure Open AI services and a standard RAG architecture [27].

4.2.2. Case 2

Case company 2 is a Europe-headquartered multinational organisation in telecommunications and networking. It develops and deploys mobile network technologies, and provides infrastructure, software, and services for global connectivity. Case company 2 has an internal LLM tool, which shall remain unnamed within this article for confidentiality reasons. The internal LLM tool was developed in-house and is based on a RAG implementation. This RAG implementation has access to over 2000 company specific documents, and is primarily used for the information search and retrieval use case.

We conducted four individual semi-structured interviews with the participants from this case company. The tools used and the LLM-based use cases at this case company are presented in Table 2.

4.2.3. Case 3

Case company 3 is a Europe-headquartered multinational organisation in transportation and infrastructure. It develops and delivers digital services, focusing on connectivity, data analytics, and fleet management to enhance vehicle performance and operational efficiency.

Table 3
Participant roles, LLM-based tools and LLM-assisted use cases of case 3.

Roles	Use cases	LLM tools used
Developer and Scrum Master, Solution Architect	Coding Assistance, Unit Tests Generation, Cloud Infrastructure Management, Documentation	GitHub Copilot, Amazon Q developer, ChatGPT

We conducted two individual semi-structured interviews with the participants from this case company. The tools used and the LLM-based use cases at this case company are presented in Table 3.

4.3. Data collection

We employed semi-structured interviews [28] as the method of data collection for the multi-case study. All interviews were conducted while following an ethical interview checklist [29]. They were audio-recorded with participant consent and transcribed verbatim using the video conferencing tool’s built-in record and transcribe feature. Due to scarcity of industry practitioners’ time, no pilot interviews were conducted at any of the case companies.

The interview participants were selected using a purposive sampling approach [30], as this method is well-suited for exploratory qualitative studies where participants are chosen to provide relevant insights on the phenomenon being investigated. Our primary inclusion criterion was that participants actively use LLMs in at least one of their regular SE use cases. Our intention behind this purposive approach emphasised that the selected participants had direct, practical experience with integrating LLMs into their SE workflows, aligning with the study’s objective of exploring real-world applications and perceptions of LLM adoption within SE tasks. The exception to this criterion was case 1, where they were still in the pilot phase of adopting LLMs and started the adoption process in parallel with the case study and the participants did not have much experience using LLMs in their use cases.

We used different interview questionnaires for case 1 and case 2 & 3. This is because of the difference in their LLM adoption contexts. The interview participants from case 2 and case 3 have been using LLMs within some of their software development activities for a year while case company 1 had just began to pilot test LLM-based use cases

as of the date of data collection. We used different questionnaires to account for this difference when collecting the data. The two sets of interview questionnaires for the three cases are provided within an online repository.¹

We collected data from case company 1 with two rounds of interviews. The first round of interviews (round-1 interviews) was individual and conducted with five people. The aim of these interviews was to understand participants' motivations for adopting LLMs, their evaluation strategies, expected improvements, and key concerns, including ethical and legal considerations for LLM-assisted RE workflows. The purpose was to gain insights into the feasibility and impact of LLM-based techniques, informing best practices for their adoption and optimisation in RE workflows.

A pilot testing of LLM-assisted RE workflows was conducted during an internal "workshop" hosted and organised by the case company. This internal workshop allowed users to experiment with LLM-assisted RE workflows and gain practical insights on how LLM could be implemented within their RE use cases. However, no data was collected during workshop.

After this workshop, we conducted the second round of interviews (round-2 interviews) with twelve people, five of whom are the same participants from round-1 interviews. This round of interviews was done to assess the feasibility and practical adoption of LLMs in case company 1's RE use cases based on hands-on workshop experiences. It explores necessary tool adaptations, process changes, and lessons learned from using LLMs. They were conducted as four group semi-structured interviews with three participants in each interview. Each interview lasted between 40-50 min.

The interviews conducted within case 2 and case 3 explore the use of LLMs in SE practice, focusing on their impact, benefits, and challenges. We aimed at understanding the participants' roles, the LLM-based tools they use, and their specific use cases, workflow changes, best practices, and strategies for optimising LLM usage while addressing limitations. Findings gathered from these interviews helped in refining approaches for adopting LLMs within SE activities. Each interview lasted between 30-40 min and was conducted virtually via video conferencing.

4.4. Data analysis

We conducted reflexive thematic analysis [26] on the data gathered from the interviews to derive findings. This approach treats coding and theme development as interpretive activities carried out by the researcher, rather than as procedures intended to measure agreement or reliability between coders. The interview transcripts were edited, validated through member checking [31], and anonymised to ensure confidentiality. The analysis was first carried out manually by a single author. The first author thoroughly familiarised themselves with the data from an interview transcript to gain an overall sense of the content. During this phase, notes were taken to document emerging observations and analytic decisions to maintain transparency and traceability in the analysis process.

Then, the author began extracting relevant and meaningful segments of text, i.e., the verbatim source quotes from the interview transcript (which we refer to as findings within this study). An example finding is: "I still reviewed the code line by line to ensure it does what I expect. I'm still involved in the process and treat the LLM as an assistant, not a decision-maker." This process was repeated for all fifteen interview transcripts from all three cases.

Then, similar source quotes (findings) from all fifteen interviews were collated together under a distinct and meaningful code. For example, the quotes "I think AI can't do that (the task) for me, but it can support me in this [task]", "I could imagine an AI serving as a co-pilot, checking for

¹ Zenodo repository containing the supplementary material, i.e., interview and survey questionnaires: <https://doi.org/10.5281/zenodo.15754264>.

Table 4

Thematic analysis codebook.

Code	Meaning
Role	Expectations and perception about the kind of role LLMs should or will play in SE.
Criteria	The criteria under which LLMs can be employed in SE.
Benefit	The benefits of employing LLMs in SE.
Method	The evaluation methods to assess LLMs' outputs.
Metric	The metrics using which LLMs' outputs are evaluated.
Model selection	Insights on how to select the appropriate LLM based on the SE task at hand.
Implementation	Strategies and techniques to effectively use LLMs in SE.
Mandatory requirement	Mandatory practices when employing LLMs in SE.
Autonomy	The need for human autonomy when employing LLMs in SE.
Changing user responsibilities	The effect on user responsibilities when employing LLMs in SE.
Changing workflows	The effect on development workflows when employing LLMs in SE.
Facilitation	How to facilitate training to use LLMs more effectively in SE.
Requirements	The necessary skills users need to have to use LLMs effectively in SE.

inconsistencies in diagrams or other output artefacts", and "In essence, this tool (the LLM) would act like an assistant or co-pilot. It wouldn't automate things" were collated together under the code "Role". The first two quotes were gathered from two different participants from case 1 and the third quote was gathered from a participant from case 3. Codes were developed inductively by comparing quotes across all interviews and refining code definitions as analysis progressed. This process resulted in a total of 13 distinct and meaningful codes. The code descriptions (code book) are presented in Table 4. Table 5 in Section 5.1 provides the traceability of findings to the codes and themes under which they are categorised under. The codebook.xlsx file provided within the online repository² provides a trail of the development of themes and codes to the findings along with the source of the findings as well.

To improve transparency and to check that the development of codes and themes were understandable beyond the primary analyst, a second author became involved in the coding process at this stage. Based on the guidelines for assessing and reporting reliability of coding procedures in qualitative research [32,33], the involvement of the second author was planned for consensus building. The goal was to assess whether the codes and themes could be reasonably interpreted similarly by another researcher and to reduce the risk that the analysis relied on implicit knowledge held only by a single author. No major inconsistencies in the coding and interpretation process were observed between both the researchers. While in some instances both researchers coded the same data extract using different labels, the definitions of these codes were semantically similar. In such cases, the researchers reached a consensus on a single code name, which was then used consistently to code similar data extracts in subsequent analysis as well re-checking the existing annotations to ensure they align with the updated code names and/or definitions. Any such disagreements between the two authors were discussed and resolved.

Once coding was complete, the researchers reviewed and collated codes that address the same aspect into potential themes, considering how different codes combined to capture significant patterns within the data. To reach a consensus in interpretation, we also had regular team discussions and iterative reviews of the developing themes between four of the six involved authors. Finally, the themes were clearly defined and named, with representative quotes from participants selected to illustrate each theme as presented within Tables 5 and 6 in Section 5. This resulted in the creation of six distinct and exclusive primary themes as follows:

² Zenodo repo with codebook.xlsx file: <https://doi.org/10.5281/zenodo.15754264>.

Table 5
Primary themes, the corresponding codes and findings, and the source(s) from which the findings were extracted.

Theme	Code	Findings	Source case
AI assistant	Role	C1F1- I5; C1F2 - I3; C1F3 - I2, C2F1- I12; C3F1- I15	Case 1, Case 2, Case 3
	Criteria	C1F4 - I2; C1F5 - I6; C3F2 - I14	Case 2, Case 3
	Benefit	C1F6 - I7; C2F2 - I12; C3F3 - I14; C3F4 - I15	Case 1, Case 2, Case 3
Evaluation	Method	C1F7 - I1; C1F8 - I4; C1F9 - I5; C3F5 - I15	Case 1, Case 3
	Metrics	C1F10 - I6; C1F11- I5; C2F3 - I11	Case 1, Case 2
Applicability	Model Selection	C1F12- I7; C1F13- I8; C3F6 - I15	Case 1, Case 3
	Implementation	C2F4- I 12; C3F7 - I15; C3F8 - I15	Case 2, Case 3
Human oversight and agency	Mandatory Requirement	C1F14- I1; C1F15- I4; C2F5- I12; C2F6- I12; C3F9- I14; C3F10- I14	Case 1, Case 2, Case 3
	Autonomy	C1F16 - I1	Case 1
LLM effect on workflows	Changing user responsibilities	C1F20- I8; C2F8 - I12	Case 1, Case 2
	Changing Workflows	C1F17- I8; C1F18- I8; C1F19- I9; C2F7- I11; C3F11- I15; C3F12- I14; C3F13 - I14	Case 1, Case 3
User skills	Facilitation	C1F21- I9; C1F22- I8; C3F14- I14; C3F15- I14; C3F16- I15	Case 1, , Case 2, Case 3
	Requirements	C2F9- I12; C2F10- I13; C3F17- I14; C3F18- I15	Case 2, Case 3

1. The “**AI Assistant**” theme encompasses findings about practitioners’ expectations and findings on utilising LLMs as AI assistants within their SE activities rather than automation tools.
2. The “**Evaluation**” theme refers to the methods and metrics for assessing various aspects of the LLM-generated artefacts.
3. The “**Applicability**” theme encapsulates what users expect the capabilities of the LLMs to be and the boundary of the LLMs’ applicability for the selected use cases.
4. The “**Human Oversight and Agency**” theme captures findings related to the role of users in overseeing and guiding LLM-generated outputs. It encompasses practitioners’ perspectives on maintaining control, ensuring accountability, and balancing AI assistance with human expertise in SE activities.
5. The “**LLM Effect on Workflows**” theme refers to the impact of LLM integration on existing SE workflows. It includes observations on how LLMs reshape task execution as well as potential adaptations required for efficient LLM adoption.
6. The “**User Skills**” theme encapsulates findings into the skills and knowledge practitioners need to effectively utilise LLMs in SE.

Our analysis of the multi-case study data revealed several similar findings across the three cases within each primary theme. Each finding extracted from the interviews that contributed to the synthesis of the recommendations was given an identifier in the format of CXFY, where X = [1,3] refers to the case from which the findings was extracted. FY, where F = [1,22], refers to the specific finding within each case. So, C1F1 represents the first finding from case 1 that was found relevant to the synthesis of a recommendation. Based on these common findings, we synthesised seven recommendations to support the adoption of LLMs in SE activities within industrial settings. Every recommendation is supported by multiple findings from different participants from two or more cases. The seven recommendations derived following this process are presented in Section 6.

4.5. Validation survey

We conducted an online survey to assess the perceived relevance of the seven recommendations. The survey questionnaire had a total of ten closed-ended questions. The first question, **Q1**, was a multiple choice single selection question aimed to gather the participant’s demographic

background, i.e., the industry in which they work. The next two questions, **Q2** and **Q3**, were control questions focusing on checking whether the respondents use enterprise versions of the LLMs within their SE use cases and how familiar they are with prompt engineering. Then, in the next seven questions, **Q4** to **Q10**, we presented the recommendations and asked the participants to rate their agreement with the seven recommendations on a scale of 1-5, where 1 is “strongly disagree”, 2 is “somewhat disagree”, 3 is “neutral”, 4 is “somewhat agree”, and 5 is “strongly agree”.

The target population of the survey were software practitioners (e.g. software developers, requirements engineers, product owners, solution architects, and DevOps engineers) from various industries. A pilot study of this survey was administered to three people to ensure the questions in the survey were understandable and no ambiguity was present. The pilot round of survey was active for one week. Based on the feedback from the pilot respondents, we added examples for the recommendations at end of the questions to demonstrate how they can be applied in industrial practice. Four of the six authors of this study contributed to the development, refinement and validation of the survey questionnaire. Once we were satisfied with the changes made, we distributed the survey online to collect the data. The participants were recruited from our professional network, following a purposive sampling approach [30], based on fulfilling this criteria: have experience using any LLM-based tool within their work. The survey was also shared on *LinkedIn* with the criteria highlighted in the post description. We followed the ethical interview checklist [29] to collect the participants’ informed consent for processing for their responses to synthesise the results. The survey was open for seven weeks.

We used the data collected from this survey to strengthen our arguments surrounding the perceived relevance of the recommendations. The survey questionnaire along with the collected data is provided within the supplementary data within an online repository.³

4.6. Mapping to AI HLEG principles

The final phase of our study involves mapping the validated recommendations to the seven Trustworthy AI principles proposed by the AI HLEG. This mapping was conducted after the recommendations had

³ Zenodo repo with survey data: <https://doi.org/10.5281/zenodo.15754264>.

Table 6
Supporting quotes from one or more of the case companies for each recommendation.

Theme	ID	Supporting quotes
AI Assistant	R1	<p>C1F1: “In essence, this tool (the LLM) would act like an assistant or co-pilot. It wouldn’t automate things”</p> <p>C1F2: “I think AI can’t do that (the task) for me, but it can support me in this [task]”</p> <p>C2F1: “I see LLMs as a helper tool”</p> <p>C3F1: “I would love to have something like a coding assistant”.</p>
Evaluation	R2	<p>C1F7: “This process (LLM’s outputs evaluation) would involve not just me, but also expert stakeholders”.</p> <p>C1F8: “ask software testers, not developers, if the solution meets the requirement. Testers usually have a better understanding of how things are connected and what’s needed to reach the goal”.</p> <p>C1F9: “I plan to conduct feedback (for generated output) sessions with different relevant stakeholders to determine its quality and usefulness”</p> <p>C3F5: “I always prefer a domain expert to review it [LLM’s output] during code review”.</p>
Evaluation	R3	<p>C1F10: “The tool doesn’t need to write a perfect user story for me; it just needs to provide me with useful information in an intelligent way”.</p> <p>C1F11: “the tool should be helpful to those who use it”</p> <p>C2F3: “we focus on desirability: does it help end users?”</p>
Applicability	R4	<p>C1F12: “It definitely makes sense to have models specialised for specific tasks in terms of performance”</p> <p>C1F13: “GitHub Copilot wasn’t the right tool for what we tried because it’s more code-based”</p> <p>C3F6: “Amazon Q, it’s better suited for AWS CDK related tasks. Amazon Q helps specifically with the Amazon-related things, because it’s trained on that”.</p> <p>C3F7: “Amazon Q and GitHub Copilot are designed for different things, so I use each specifically for its intended purpose”.</p>
Human oversight and agency	R5	<p>C1F14: “there’s still a need for human oversight and verification, especially when it’s being used on a large scale and by people who may not fully understand how it works”</p> <p>C1F15: “But ultimately, human feedback is crucial”.</p> <p>C2F5: “I’m definitely in favour of keeping humans in the loop”.</p> <p>C2F6: “I still reviewed the code line by line to ensure it does what I expect. I’m still involved in the process and treat the LLM as an assistant, not a decision-maker”.</p> <p>C3F9: “end of the day, you’re still responsible for checking and making sure everything works as expected”.</p> <p>C3F10: “Whatever code an LLM generates, you are still responsible for it”.</p>
LLM effect on workflow	R6	<p>C1F17: “our existing processes would need to be restructured”</p> <p>C1F18: “We need to carve out a space for LLMs in our business process”</p> <p>C1F19: “The integration of such tools into our workflow would likely require a careful evaluation and adjustment of our current processes”.</p> <p>C2F7: “Another challenge is that it changes the workflows, and agreeing on what the new workflows should be is tough”.</p>

(continued on next page)

Table 6 (continued).

Theme	ID	Supporting quotes
		<p>C2F8: “As LLMs become more advanced, I think we’ll see a shift where developers to review the code and ensure things are working properly, but they won’t be doing the groundwork”</p> <p>C3F11: “AI is influencing our planning process as well. We’re seeing a shift in traditional practices due to LLMs, even in the planning stage”.</p> <p>C3F12: “we may not yet be fully considering how it will impact our workflows and which use cases will provide the most benefit”.</p> <p>C3F13: “We’re still figuring out how to design software that’s LLM-friendly. If we were to build something new today, how do we ensure it can be integrated with an LLM down the line? That’s a question we don’t fully have the answer to yet”.</p>
User skills	R7	<p>C1F21: “there’s a need for training on how to interact with these tools, as communicating with them is not the same as conversing with another person”</p> <p>C1F22: “provide proper education about them. If you don’t know how to use it, in most cases, it will be completely useless”</p> <p>C2F9: “I think the key is to understand what you’re doing. LLMs can generate code, but it may become too complex for you to fully comprehend. If that happens, you’re probably already in over your head”.</p> <p>C2F10: “I think the main thing is to have domain knowledge before using LLMs”.</p> <p>C3F14: “one of the biggest challenges is getting people up to speed on how to use LLMs effectively. It takes proper planning, understanding, and the right implementation to really make use of it”.</p> <p>C3F15: “Part of our responsibility is to stay updated on these tools and understand how they can help us. We’re even getting some training on generative AI to see how it could fit into our work”.</p>
User skills	R7	<p>C3F16: “we often share ideas on what works well for specific use cases”</p> <p>C3F17: “knowing which tool to use and understanding what a model is trained on are key”</p> <p>C3F18: “LLMs aren’t just plug-and-play; you need to understand how they work, how to prompt them correctly, and where they can actually add value”.</p>

been developed and validated. The recommendations themselves were derived inductively from participants’ experiences and perspectives.

Although we were aware of the Trustworthy AI principles prior to the study, these principles were deliberately not introduced during the interviews. This decision was made to avoid influencing participants’ responses or framing their experiences in terms of predefined regulatory concepts. As a result, participants did not consistently reference or recall these principles, and the resulting recommendations primarily reflect practical concerns raised in the case studies rather than compliance considerations.

The objective of the mapping step is therefore descriptive and reflective: to assess which of the seven principles are supported by the empirically derived recommendations and to identify which principles are not addressed. Hence, we are open to the degree of coverage of all seven principles by design, and also allow for extending the recommendations to proactively address principles that did not emerge from the data.

We discuss the impact of this methodological decision in Section 8.

5. Multi-case study results

This section presents the results of the thematic analysis performed on interview data gathered from all fifteen interviews from the three cases. We first present the common findings across all three cases within Section 5.1, organised by each of the six common themes, before presenting case specific findings for each of the three cases. The

themes and findings presented in this section are summarised in Table 5. The source of the direct quotes supporting the presented findings are denoted with an interview ID: IX, with X ranging from 1 to 15, e.g., I3 refers to the 3rd interview.

5.1. Cross case analysis

[AI Assistant]: Practitioners view LLMs as tools or assistants that enhance their work, providing guidance and suggestions rather than replacing their decision making. As stated in I5, *“In essence, this tool (the LLM) would act like an assistant or co-pilot. It wouldn’t automate things.”* This highlights the potential for AI to become an interactive and collaborative partner in software development. There is a strong preference for using LLMs in a supervised manner, with concerns arising when they operate without human oversight. Echoing this sentiment, I2 noted, *“I see no concerns and quite some advantages in using them supervised or in a supporting role.”*

[Evaluation]: The effectiveness of LLMs is judged not only by developers but also by testers, customers, and other stakeholders. As noted in I1, *“This process (LLMs’ output evaluation) would involve not just me, but also expert stakeholders.”* Rather than demanding flawless outputs, practitioners emphasised whether LLM suggestions are useful and offer valuable insights. Echoing this view, I6 remarked, *“The [LLM] tool doesn’t need to write a perfect user story for me; it just needs to provide me with useful information in an intelligent way.”* This suggests that LLMs are expected to assist rather than fully automate SE tasks, with usefulness regarded as a more meaningful evaluation criterion than strict accuracy.

[Applicability]: Practitioners recognise that different LLMs are optimised for specific tasks, and their effectiveness varies depending on the domain. A one-size-fits-all approach is not viable, and selecting the right model for a given task is crucial for achieving optimal performance. As noted in I15, *“Amazon Q developer and GitHub Copilot are designed for different things, so I use each specifically for its intended purpose.”* Using LLMs beyond their intended scope can lead to suboptimal results. This concern was highlighted in I7, where it was stated, *“The copilot also expressed dissatisfaction about being misused, so that’s probably not a viable approach moving forward.”* Rather than relying on LLMs for end-to-end solutions, practitioners break down larger goals into smaller steps and use LLMs selectively at each stage. As explained in I15, *“I break down a bigger goal into smaller steps and tackle them one at a time.”* This reflects a structured and controlled approach to leveraging LLM assistance.

[Human Oversight and Agency]: Practitioners stress the importance of human involvement in reviewing and verifying LLM-generated outputs. Despite LLMs’ impressive capabilities, human judgement remains critical, especially in large-scale or complex applications. As emphasised in I14, *“There’s still a need for human oversight and verification, especially when it’s being used on a large scale and by people who may not fully understand how it works.”*

[LLM Effect on Workflows]: The adoption of LLMs is expected to cause significant changes in existing workflows. Organisations need to carefully evaluate and adapt their current processes to incorporate LLMs effectively, which may require restructuring or redesigning workflows to accommodate LLMs’ capabilities. As noted in I8, *“We need to carve out a space for LLMs in our business process . . . our existing processes would need to be restructured.”* There is uncertainty around how the new LLM-assisted workflows should be designed and developed. This challenge was highlighted in I11: *“Another challenge is that it changes the workflows, and agreeing on what the new workflows should be is tough.”* As LLMs advance, the role of developers is expected to shift. According to I3, *“If we have less work with basic or standard activities, we could focus more on our main task, which is to talk to our customers.”* Developers may no longer need to engage in the more repetitive groundwork, instead focusing on validating and ensuring the quality of LLM-generated outputs.

[User Skills]: Practitioners emphasise the need for proper training on how to interact with LLMs. As communication with these tools differs significantly from human interactions, learning how to engage with them effectively is critical for maximising their value. As stated in I8, *“Provide proper education about them. If you don’t know how to use it, in most cases, it will be completely useless.”* A strong understanding of the domain is considered essential for using LLMs effectively. This was highlighted in I15: *“LLMs aren’t just plug-and-play; you need to understand how they work, how to prompt them correctly, and where they can actually add value.”* Practitioners also stress the importance of sharing insights on what works well and what does not. As noted in I15, *“We often share ideas on what works well for specific use cases.”* Effective use of LLMs involves not just technical know-how but also collaboration and knowledge sharing within teams to understand where and how AI can add the most value.

5.2. Case 1

The findings specific to case 1 are based on interviewees’ understanding of LLMs, prompt engineering, and RE artefacts, tasks, and processes at the case company before and after testing LLM-assisted RE use cases in the workshop. These findings focus on (i) expected improvements in their RE process with LLM assistance, (ii) planned evaluation methods for LLMs’ outputs, and (iii) qualities of LLMs prioritised for real-world adoption.

Practitioners believe LLMs can perform analogical reasoning tasks, such as generating high-level project requirements and deriving insights from past projects. They also see potential in using LLMs to identify redundancies in requirements elicitation meetings and system requirements specifications (SRS). Multiple participants expressed interest in using LLMs to check the completeness of RE artefacts like SRS. As stated in I5, *“This tool could ask us questions or we could ask it questions, which could help us remember ideas or aspects of our ecosystem that we might have overlooked. It could also help us learn from past projects, both successful and unsuccessful.”* After using three LLM-based tools during the workshop, participants believed LLMs could aid RE processes by retrieving and analysing data, improving productivity, providing different perspectives, and enhancing coverage of overlooked aspects.

Participants also showed interest in automating tasks such as generating artefacts (e.g., user stories) and checking artefact consistency and dependencies for requirements traceability and change management. As noted in I4, *“When it comes to requirements, having support in creating them can be very helpful. If we do not have to worry about the formal structure, or if the acceptance criteria and affected systems are correctly mentioned, it can save a lot of time.”*

The RAG-POC tool required more detailed input prompts than Microsoft Copilot to convey users’ intentions and sometimes failed to provide answers consistently for the same prompts. Participants attributed these challenges to the lack of sufficient stakeholder analysis artefact data within the RAG database and felt the tool needs greater access to company-specific data and knowledge. This was reflected in I6: *“While the user story use case worked well, the generating stakeholder analysis artefacts use case was a bit more complicated, and we lacked sufficient knowledge about the information already provided and what we could upload [to the RAG database].”* This highlights the challenges faced with the RAG POC tool.

Efficiently setting up the RAG-POC tool and modifying the RAG architecture to include company-specific additions can be a challenging task due to its complexity. The construction and curation of the knowledge base to be used for RAG is currently presenting research challenges [34]. The findings of this case strengthen the arguments focusing on the lack of methodology that guides organisations on how to identify valuable and relevant information for developing useful RAG architectures.

5.3. Case 2

In this case, LLMs were widely valued for their role in learning and cognitive support. Practitioners used them to explore new domains, summarise information, and offload mentally taxing tasks. As noted in I13, “*For me, the main motivation is to learn new things, summarise information, and explore new fields.*” LLMs were also observed to reduce cognitive strain on users. According to I10, “*It eases the cognitive load because I do not have to remember all the command lines when it comes to coding.*”

While LLMs assist in ideation, maintaining control over their creative outputs remains a challenge. As stated in I11, “*The LLMs tend to be too creative, and simple prompting isn’t enough to make them adhere to these constraints.*” However, some saw value in leveraging LLMs for creative exploration: “*I think it would be useful to at least try to perform the [creative] task with the LLM and see if the answer is satisfactory.*” as expressed in I10. This approach, however, requires strong domain knowledge on the user’s part, along with human oversight mechanisms to be advantageous.

Fragmented LLM adoption across teams led to concerns about inefficiency. One participant suggested a more centralised approach: “*Many different teams want to try it out, which leads to multiple LLMs being used across the organisation. Perhaps we shouldn’t have so many and instead focus on having one base that can be applied in different contexts.*” as stated in I11. Others emphasised testing high-value use cases before scaling: “*Instead of evaluating all possible use cases at once, we should focus on one small, high-value use case.*” also from I11. In contrast, some practitioners experimented with LLMs interacting as specialised agents to refine results. One described their approach: “*I’m using one LLM to optimise the input for another. It’s like an agent approach, where each tool works together to improve the end result.*” as mentioned in I12. Another proposed developing domain-specific agents: “*We need one agent for generating standard documentation and another for troubleshooting content.*” from I11. Overall, while diversification of LLM models across organisations adds complexity to the adoption process, utilising multiple LLMs in an agentic approach or similar can contribute to the resilience of LLM-assisted workflows by mitigating the limitations that come with relying on a single LLM model.

The value LLMs brought to SE was seen to be assessed through multiple factors, including efficiency, cost, and business impact. One team developed a structured approach: “*We look at value from three angles: what it brings to the customer, what it brings to the company in terms of profitability, and the trade-offs in development and maintenance costs.*” as stated in I11.

LLMs were also observed to impact the user’s psychological aspects such as increasing confidence when tackling new or complex tasks. As reflected in I13, “*I feel more confident taking on new or unknown tasks*” and “*It also helps with confidence since I have an initial helper before reaching out to colleagues.*”

5.4. Case 3

Similar to case 2, practitioners from case 3 also found LLMs particularly useful for quickly grasping new topics without sifting through extensive online resources. As noted in I15, “*If I don’t know anything about it (a new task), it’s easier to go to the LLMs instead of googling and reading through all the stuff.*” Additionally, the effect of LLMs on cognitive load was similarly reflected: “*I feel a lot less overwhelmed when dealing with unfamiliar tasks.*” as expressed in I15.

However, unlike case 2 where most practitioners preferred the auto complete feature of Copilot for coding assistance, many users from case 3 favoured chat-based interactions over code completion due to greater context awareness. According to I15, “*I use the chat option within Copilot a lot. I don’t use the code completion much because it lacks the context awareness that the chat option provides. For me, the multiple suggestion chat is more useful than the code completion.*”

LLMs were recognised for improving efficiency and reducing workload. As stated in I14, “*I think LLMs can really help cut down on day-to-day work and make things more efficient.*” However, limitations in contextual understanding were also highlighted: “*The AI doesn’t think the way a developer does. It doesn’t have context or real understanding of the system.*” as noted in I14.

Training provided to users regarding LLM usage was limited more to compliance than practical application. As mentioned in I15, “*We received some training on using AI, but it was mostly focused on what not to do, along with legal guidelines.*” Additionally, the abundance of LLM tools available in the market was observed to create confusion around adoption strategy. This concern was expressed in I14: “*With so many models out there, it can be a bit confusing to know which one to choose.*”

Summary of RQ1 results: The results of our multi-case study highlight that practitioners prefer to use LLMs as an assistant, prioritise the usefulness of the outputs over their accuracy, follow a structured approach to leveraging LLMs within a defined scope of applicability, focus on understanding and adapting to effects of LLMs on traditional SE workflows, emphasise the need to develop human oversight mechanisms and the necessary skills to leverage LLMs effectively.

6. Recommendations from the multi-case study

From the themes and findings, we were able to derive seven recommendations. Each recommendation is mapped to one the themes from our case analysis. Some themes were complex enough to warrant more than one recommendation. Each recommendation is supported by findings from at least two cases. The recommendations are as follows:

- R1:** It is beneficial to use the LLMs as assistants with human oversight rather than try and automate the task using the LLMs.
- R2:** When using LLMs as AI assistants, it is better to prioritise the usefulness of the LLMs’ outputs rather than output accuracy.
- R3:** The relevant stakeholders’ satisfaction should be used as part of the LLM-generated outputs evaluation criteria.
- R4:** It is not beneficial to use any task-specific fine-tuned LLMs like GitHub Copilot for tasks they were not designed for.
- R5:** It is crucial to develop and implement human oversight and validation mechanisms while using LLMs’ assistance for any task.
- R6:** There is a need to create room for incorporating LLMs into business processes by restructuring (parts of) them.
- R7:** Knowledge-sharing activities like seminars and specialised training are crucial for the users of the LLMs.

The mapping of the recommendations to the six primary themes and supporting quotes can be found in [Table 6](#).

Summary of RQ2 results: We have synthesised seven recommendations to support efficient and responsible adoption of LLMs within industrial software development. These recommendations are aimed to guide practitioners when deciding how to use LLMs, determining the scope of LLM’s application and criteria for evaluating LLMs’ outputs, emphasising necessary considerations to enhance regulatory compliance, ease LLM integration into business processes, maximise the adoption benefits, and improve overall user experience and ease of use.

7. Survey results

This section presents the results of the online survey conducted to understand the extent to which the recommendations from the case study are applicable to contexts outside of the selected case companies. Section 7.1 presents the demographics of the survey participants. Section 7.2 presents the quantitative results of the survey responses.

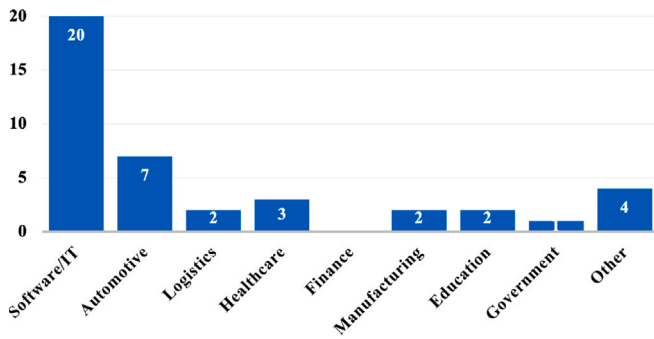


Fig. 2. Distribution of respondents based on the industry they are currently working in.

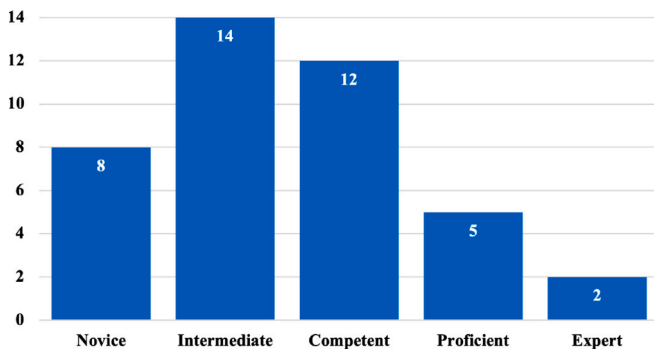


Fig. 3. Distribution of respondents based on their self-identified level of skill within prompt engineering.

7.1. Participants' demographic information

We collected a total of 43 responses over a span of seven weeks, out of which 41 responses were usable; we discarded two responses because they were incomplete. The median response time for answering the survey was 5 min 31 s. The complete breakdown of the demographic data for the survey participants is presented within Figs. 2 and 3.

Fig. 2 represents the demographic breakdown of survey participants by industry. Out of the 41 respondents, 20 indicated working in the Software/IT services sector, which we use to refer to organisations providing IT services or consultancy not dedicated to a specific application domain (e.g. Automotive, Finance, Healthcare). The second highest number of respondents came from the Automotive industry with 7 participants. The Manufacturing and Education sectors each had 3 participants, accounting for 7.3% of responses, while the Logistics, Healthcare, and Government sectors each had 2 participants, representing 4.87% of responses. Notably, although Finance was provided as an option, no respondents selected it. Lastly, the category labelled 'Other' includes 4 participants who indicated domains outside the predefined options. This distribution shows a significant inclination of survey responses from individuals in the Software/IT sector compared to other sectors. 68% of participants reported to have used or using an enterprise version of LLM-based tools such as Microsoft Copilot, Github Copilot, or ChatGPT enterprise. Out of the participants who use an enterprise version of the LLM, 70% of them belong to the software/IT industry.

Fig. 3 represents the self-reported prompt engineering skills of survey participants. The participants rated their skills on a Likert scale from 1 to 5, with 1 being "Novice" and 5 being "Expert". The majority of participants rated themselves as "Intermediate" with a total of fourteen respondents, accounting for 34% of the total participants. This was closely followed by twelve participants who rated themselves as "Competent", making up 29% of the total participants. "Novice" and

"Proficient" categories had eight and five participants, making up 20% and 12% of the total participants respectively. Only two participants rated themselves as "Expert". This distribution indicates a moderate level of prompt engineering skills among the survey respondents, with most participants identifying themselves as having intermediate to competent skills.

7.2. Recommendations' evaluation

The online survey was meant to explore whether software practitioners outside the case companies also saw value in the recommendations. This section reports the level of agreement of the surveyed practitioners with each of the recommendations R1-R7 presented to them within questions Q4-Q10 of the survey. Fig. 4 presents the aggregated Likert scale responses from all 41 participants.

Fig. 4 illustrates the distribution of responses for the remaining survey questions, Q4 to Q10. The dark green horizontal lines indicate a score of 5, light green upward diagonal lines indicate a score of 4, the grey vertical lines indicate a score of 3, the orange downward diagonal lines indicate a score of 2 and the red dotted pattern indicate a score of 1.

Fig. 4 shows that 18 out of 41 respondents (45%) for Q4 strongly agree with the presented recommendation (survey value of 5), 16 out of 41 respondents (38%) showed moderate agreement (survey value for 4), while 6 (15%) remained neutral (survey value of 3). Only 1 respondent (3%) somewhat disagreed (survey value of 2) with the recommendation. If we aggregate both strong and moderate agreement (a response of 4 or 5), 83% of respondents for Q4 agree with the recommendation. Looking at the demographic details of the participants who gave favourable responses (a response of 4 or 5), 16 out of 34 participants (47%) are from the software/IT industry.

Similarly, for Q5, 8 out of 41 respondents (20%) strongly agree, 18 respondents (44%) show moderate agreement while 12 (29%) remained neutral. 2 (5%) somewhat disagree while 1 (2%) strongly disagree (survey value of 1) with the recommendation. 64% of the total respondents agree with the recommendation. 16 of the 26 (62%) participants who gave a favourable response are from the software/IT industry.

For Q6, 9 out of 41 respondents (22%) strongly agree, 18 respondents (37%) show both moderate and neutral agreements. Only 2 respondents (4%) somewhat disagreed with the recommendation. 59% of the total respondents agree with the recommendation. 10 out of the 24 (42%) participants who gave a favourable response are from the software/IT industry.

For Q7, only 5 out of 41 respondents (12%) strongly agree, 8 respondents (20%) moderate agreement while 10 (24%) remained neutral. 17 respondents (42%) somewhat disagreed with the recommendation and only 1 (2%) strongly disagreed. Only 32% of total respondents agree with the recommendation while 44% of them disagreed. 9 out of the 13 (69%) participants who gave a favourable response are from the software/IT industry.

For Q8, 20 out of 41 respondents (50%) strongly agree, 14 respondents (35%) show moderate agreement while 5 (13%) remained neutral. Only 1 respondent (2%) somewhat disagreed with the recommendation. 85% of the total respondents agree with the recommendation. 18 out of 35 (52%) participants who gave a favourable response are from the software/IT industry.

For Q9, 8 out of 41 respondents (20%) strongly agree, 16 respondents (39%) show moderate agreement while 13 (32%) remained neutral. Only 3 respondents (7%) somewhat disagree and 1 respondent (2%) with the recommendation. 59% of the total respondents agree with the recommendation. 14 out of 24 participants (58%) who gave a favourable response are from the software/IT industry.

For Q10, 19 out of 41 respondents (46%) strongly agree, 11 respondents (27%) show moderate agreement while 10 (25%) remained

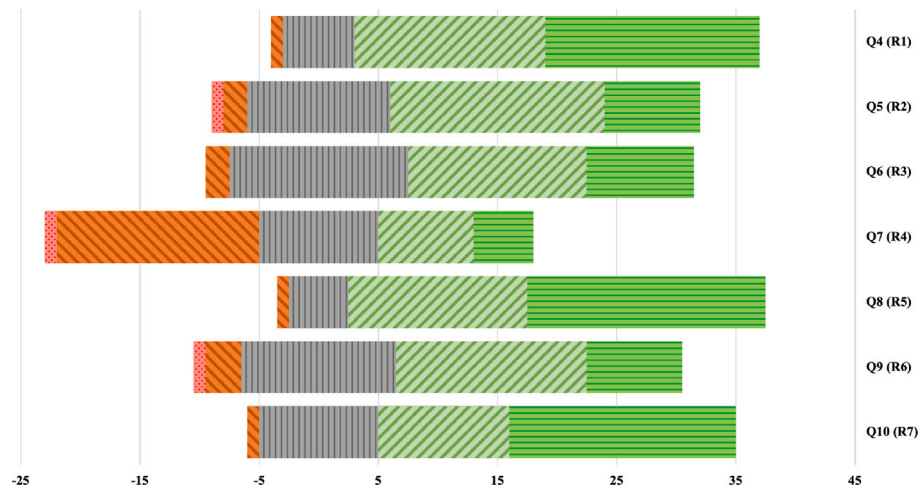


Fig. 4. Aggregated Likert scale responses for Q7 to Q14.

neutral. Only 1 respondent (2%) somewhat disagreed with the recommendation. 73% of the total respondents agree with the recommendation. 16 out of 30 participants (53%) who gave a favourable response are from the software/IT industry.

In addition, we conducted subgroup analyses to explore potential differences in perceptions across participants with varying backgrounds. The analysis examined respondents' agreement with the seven recommendations based on three factors: (1) industry (Q1), (2) prompting proficiency (Q2), and (3) access to enterprise LLMs and RAG enhanced LLM system (Q3). For analytical clarity, responses for Q2 rated 1–3 (novice to competent) were classified under less proficient users, and ratings of 4–5 (proficient to expert) as highly proficient users in the subgroup analysis.

We applied a one-way ANOVA to assess whether differences in these factors influenced respondents' levels of agreement with the recommendations, followed by Welch's t-tests for pairwise and population-level comparisons. Across the three grouping factors, the results revealed no statistically significant differences in agreement levels, except for two cases associated with prompting proficiency. Participants with higher prompting proficiency showed significantly greater agreement with R2 ($p = 0.0319$) and with R3 ($p = 0.0169$). These findings suggest that individuals with higher prompting proficiency agree more with our recommendations concerning the evaluative aspects of LLM-assisted workflows (R2 and R3). This preference might be a consequence of their significant experiences prompting and collaborating with LLMs within SE.

Overall, the strongest agreement was concentrated around recommendations that frame LLMs as assistive rather than autonomous technologies, emphasise human oversight and validation, and encourage knowledge-sharing and training for users. Taken together, this pattern suggests that practitioners see successful LLM adoption in SE less as a matter of simply deploying the technology and more as a matter of integrating it within appropriate human-centric and organisational mechanisms. This is an important finding because it indicates that practitioners associate effective use not only with the technical performance of LLMs, but also with governance, learning, and responsible integration into everyday work.

These findings address RQ3 by showing that the perceived relevance of the recommendations is highest when they address the practical conditions that make LLM-assisted work usable and trustworthy in organisational settings. The results therefore suggest that practitioners place particular importance on recommendations that preserve accountability, support informed human judgement, and help teams develop the skills needed to work effectively with LLMs. At the same time, the subgroup analysis indicates that these perceptions were largely consistent across different respondent backgrounds, which strengthens

the view that these concerns are broadly shared rather than limited to a particular subset of participants.

Summary of RQ3 results: The survey results indicate that most of the recommendations synthesised from the case study are applicable beyond the context of the case company, with six out of the seven recommendations receiving a majority (>50%) agreement from the survey participants. The software/IT industry practitioners make up roughly 48% of the total sampled population, yet they provided 54% of the favourable responses for seven recommendations. Finally, software practitioners with higher prompting proficiency tend to agree more with our recommendations about LLMs' output evaluation more than others.

8. Trustworthy AI compliance

To use LLMs for SE activities in practical settings beyond sandbox environments, software practitioners are recommended to comply with the seven key principles prescribed by the AI HLEG. However, while our findings provide insights relevant to selected principles (e.g., R1 and R5 emphasise human oversight), the interview data reveal a substantial gap between the AI HLEG principles and the day-to-day concerns and practices of software practitioners.

Motivated by this observation, we examine how the seven empirically derived recommendations relate to the Trustworthy AI principles. Specifically, we performed a post hoc mapping of the recommendations to the seven principles with the goal of identifying which principles are supported by the recommendations and, equally importantly, which are not. This mapping is intended to be descriptive rather than prescriptive and serves to highlight gaps between experience-based guidance and policy-level expectations.

Fig. 5 illustrates the relationship between the recommendations and the AI HLEG principles. For example, R1 and R5 emphasise the need for human oversight when using LLMs in SE activities. These recommendations contribute to the principles of *Human Agency and Oversight* and *Accountability*, as responsibility for decisions and outputs remains with the human actor. Similarly, R3, which recommends involving relevant stakeholders in the evaluation of LLM outputs, also aligns with these two principles by reinforcing oversight and shared responsibility.

R2, which guides practitioners on appropriate reliance on LLMs, contributes to the principle of *Technical Robustness and Safety*. Likewise, R4, which limits the use of fine-tuned LLMs to their intended contexts, supports output reliability and thus also aligns with *Technical Robustness and Safety*. R6 highlights the need to restructure existing

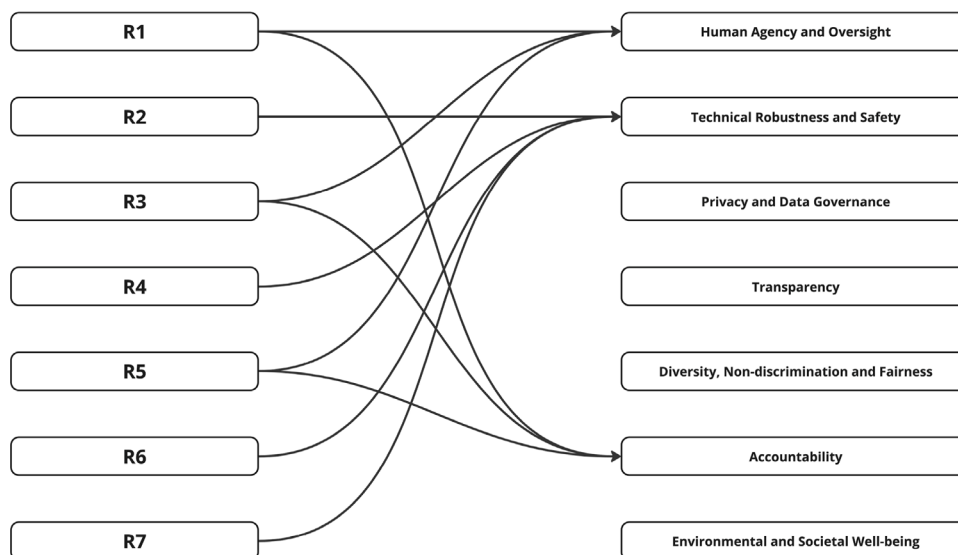


Fig. 5. Mapping of the seven recommendations to trustworthy AI principles.

workflows when integrating LLMs, primarily to ensure appropriate human oversight while maintaining effectiveness and efficiency. Finally, **R7** contributes to *Technical Robustness and Safety* by promoting knowledge sharing and skill development, which helps practitioners use LLMs more reliably in practice.

The mapping shown in Fig. 5 makes clear that the recommendations align with only *three* of the seven Trustworthy AI principles. The remaining four principles, *Privacy and Data Governance*, *Transparency*, *Diversity, Non-discrimination and Fairness*, and *Environmental and Societal Well-being* are not addressed by the recommendations derived from the case studies reported in Section 6.

The post hoc mapping shows that practices and recommendations based mainly on practical and technical experience do not provide broad regulatory coverage. It suggests that practitioners may focus on visible issues such as control, oversight, and technical reliability, while giving less attention to other Trustworthy AI principles. This points to the need for complementary measures to ensure that the full set of principles is addressed.

9. Discussion

This study provides empirical evidence on the adoption of LLM-based tools to support practitioners in their SE activities within industrial contexts. Our findings show that LLM adoption in SE is inherently multidisciplinary, encompassing aspects of (i) technology adoption, (ii) organisational change management, (iii) human–AI collaboration, and (iv) regulatory compliance. Accordingly, we compare and contrast our findings with those presented in the Related Work (Section 3), relating them to relevant literature across the aforementioned multidisciplinary themes, as presented below.

9.1. Technology adoption

One of the notable findings from both Case 2 and Case 3 is the effect of LLMs on users' experiential learning and their willingness to tackle unfamiliar tasks. This aligns with prior work [22,23], which reports that LLMs can boost users' learning and confidence when approaching new challenges. Consequently, users are able to engage in a broader range of SE tasks rather than limiting themselves to a few specific use cases.

Findings from Case 2 further suggest that LLM adoption in SE is gradually moving towards a multi-agent approach. LLM-based multi-agent (LMA) systems are found to be better capable of addressing

real world challenges that often spread across multiple domains and require expertise from different areas compared to singular LLM-based agents [35], including within SE [36].

Overall, LLM adoption appears to expand practitioners' capabilities, enabling them to tackle a wider variety of tasks while elevating the LLM's role from simply providing suggestions to taking on more agentic responsibilities. This shift also increases the user's accountability and reinforces the need to restructure workflows and maintain robust human oversight.

9.2. Organisational change management

Some of our findings underscore that successful LLM adoption in SE extends beyond technical integration and requires deliberate change management efforts. The need to restructure workflows (**R6**) and provide continuous training and knowledge sharing (**R7**) reflects the principles of organisational change theory, which emphasises readiness, communication, and capability building to sustain transformation [37]. The fragmented adoption observed across teams and the call for a more centralised, coordinated strategy (Case 2), along with the confusion caused by the abundance of LLM tools (Case 3) further illustrate the importance of establishing clear vision and leadership. While a grass root level adoption of LLMs is a positive approach that minimises the risk and effect of change resistance, without proper top-down support and leadership strategy, the adoption process will face significant challenges and might not prove to be sustainable. Our findings suggest that LLM adoption should be approached as an organisational change initiative that integrates deliberate strategic planning with the flexibility of an emergent change approach to support long-term success [38].

9.3. Human-AI collaboration

We found that a majority of participants expressed a strong preference for utilising LLMs as AI assistants rather than mere task automation tools. This preference to perform manual validation of the outputs appears to be tied to concerns about trust and reliability. However, as LLMs become more explainable and their outputs more reliable, it is likely that users' trust in LLMs will increase. This will, in turn, increase their willingness to adopt automation-oriented workflows [39].

Practitioners also reflected that the helpfulness of the LLM's outputs is prioritised over their accuracy. However, most of the state-of-the-art literature focuses on the quality of the LLM-generated output and benchmark analysis to support their claims [5]. This approach is better

suited for evaluating LLM's performance for task automation and aligns with the established theory of conventional human-machine interaction (HMI), which emphasises usability and task performance [40]. However, the effectiveness of human-AI collaboration depends not only on system performance but also on the quality of human-AI interaction, trust, and adaptability [41]. These multidimensional aspects are often neglected during evaluations in collaborative workflows, which are a better-suited criteria for task assistance.

Our results are based on practitioners' experiences with LLM-based tools in industrial settings, providing a more relevant context for assessing the feasibility of LLM-based collaborative SE. This reinforces our findings that success from adopting LLM is more closely connected to stakeholder satisfaction with the LLM-assisted workflows than to rigid quantitative benchmarks.

However, as the degree of LLM automation in SE increases, qualities like accuracy and control becomes increasingly critical, as LLM-based autonomous systems must minimise errors and have fall-back mechanisms in place to maintain user trust and operational reliability [42]. We will need trade-off mechanisms to balance the usefulness and accuracy of LLMs' outputs and the level of control user can maintain over the system.

9.4. Regulatory compliance

Our survey results further reinforce the perceived relevance of our recommendations within various industries that design and develop software products and services. Six out of the seven recommendations received more than 50% favourable responses (a response of 4 or 5 on the Likert scale) from all survey respondents. **R4** was the only recommendation that did not.

Recommendation **R4** is about limiting the application of fine tuned LLMs to use cases for which they were originally intended and designed. This recommendation received only 32% favourable responses when evaluated for agreement with software practitioners from various industries. However, 69% of those who did give a favourable response were from the software/IT industry. This indicates that software practitioners from software and IT services industry tend to agree more with **R4** compared to software practitioners from the other sampled industries.

R4 was mapped to the "technical robustness and safety" principles from the AI HLEG precisely because the AIA also aims to prevent the misuse of LLMs by disallowing their application beyond their intended scope. However, the lower agreement rate for **R4** from the sampled practitioners, especially from non-software or IT industries, indicate that either (i) the practitioners are not aware of the impact of the AIA on the adoption of LLMs to assist them in industrial practice, or (ii) they, unwittingly or otherwise, are observed to disregard the regulatory implications of the misuse of LLMs beyond their intended scope. This indicates the need to emphasise the importance and spreading awareness regarding compliance with the AIA and/or any other applicable AI regulations concerning the adoption of LLMs in industrial practice.

An important question raised by our mapping is what it means when empirically grounded, practitioner-driven recommendations align with only a subset of the AI HLEG principles. While one interpretation is that the recommendations simply need to be extended in future work to better cover the remaining principles, another possibility is that this gap could be a reflection of an underlying mismatch between policy-level principles and the realities of day-to-day SE practice. If experienced practitioners do not naturally focus on concerns such as fairness, environmental impact, or transparency when discussing the use of LLMs, in order to be effective, AI HLEG principles must be updated to be sufficiently operationalised, visible, and actionable in real-world development contexts.

Table 7

Comparison of this study's contributions with existing related work.

Recommendations from this study	Pereira et al. [22]	Süße et al. [23]	Wang et al. [24]
Prioritise usefulness over strict accuracy (R2)			
Stakeholder satisfaction as evaluation criterion (R3)			
Avoid task-misaligned fine-tuned LLMs (R4)			
Restructure business processes to integrate LLMs (R6)			✓
Knowledge-sharing activities (seminars, training) (R7)	✓	✓	
Use LLMs as assistants with human oversight (R1)	✓	✓	✓
Implement human oversight and validation mechanisms (R5)	✓	✓	✓

9.5. Novelty compared to state-of-the-art

Our findings align with the lessons learnt reported by Pereira et al. [22], Süße et al. [23], and Wang et al. [24] regarding the necessity of LLM users to possess domain knowledge regarding the task at hand, gaining deeper understanding of LLMs and how they work, limiting the LLM's role to a virtual assistant, and adapting the approach to leveraging LLMs depending on the task.

While prior studies have extensively explored the implications of integrating LLMs into SE contexts, this study introduces several practical, operational-level recommendations that emphasise human oversight, task-specificity, and organisational restructuring. For instance, Pereira et al. [22], Süße et al. [23], Wang et al. [24] primarily focus on users' learning, collaboration dynamics, and ethical considerations when interacting with LLMs. In contrast, this work provides actionable process-level recommendations for software practitioners seeking to embed LLMs into their workflows in an efficient and responsible manner.

Specifically, the contributions of this study differ in three notable ways: (i) placing emphasis on prioritising usefulness over strict accuracy in LLMs' outputs, (ii) advocating for stakeholder satisfaction as an explicit evaluation criterion for LLM-generated outputs, and (iii) ensuring the selection of appropriate LLM-tools for the task at hand. These perspectives complement existing literature by operationalising higher-level lessons into actionable recommendations for practitioners. [Table 7](#) summarises how the study's contributions relate to insights from existing works.

9.6. Threats to validity

We followed Runeson and Höst's guidelines for qualitative research analysis in software engineering [25] to discuss the potential threats to the validity of this research study.

Internal Validity: (1) The selection and grouping of the participants for the "workshop" conducted within case 1 is a potential threat as the research team could only advise on that matter. A scrum master from the case company 1 played a crucial role in the design and facilitation of the workshop based on the round-1 interview participants' responses. To mitigate any potential bias arising out of this, the scrum master was informed beforehand of how this would affect the validity of the study and they did not participate in the data collection processes. (2) Despite the promised anonymity, there is a threat of *social desirability bias* being introduced into the responses of the interview participants to provide answers they think align with the organisation's goals rather than their true opinions or experiences. (3) There is a threat of selection bias for the survey participants, especially for the participants recruited via LinkedIn as the fulfilment of the participation criteria is self-reported by the participants and we have no way of verifying that information. (4)

Only one author was part of first round of thematic coding performed on the case study data. To mitigate subjectivity bias, we had another author perform coding on a subset of the data, ensuring transparency in the coding process, and incorporating member checking to validate interpretations. However, we did not find any major inconsistencies in the coding and interpretation process between both the researchers and any minor changes or discussions brought up by the second coder was incorporated into the coding process.

External Validity: Since the recommendations are synthesised from the findings of the multi-case study, one of the possible threats to the validity of this study is the generalisability of the results. To mitigate this, we conducted a survey with software practitioners from various industries to evaluate the generalisability of the recommendations beyond the context of the case study. (3) Since the data collected from case 1 was based on user experiences of a pilot programme, the short duration of the workshop may have limited participants' ability to fully explore LLM tools. (3) The generalisability of the survey results might not extend to entire population of software practitioners due to potential threat of non-response bias by practitioners who are not proponents of using LLMs.

Construct Validity: (1) There is threat of misunderstanding the constructs presented in the interview and online survey by the participants, which might also affect their responses. Since no pilot interviews were conducted within the multi-case study, the understanding of the constructs by the participants from all the cases is not ensured. To mitigate this threat, we employed semi-structured interviews and informed the participants that they were welcome to ask clarification questions during the interview. (2) For the survey, we conducted a pilot survey and follow-up with the respondents of the pilot to ensure the understandability of the constructs within the survey remained consistent across the participants.

10. Conclusions

In conclusion, our study collected and presented the results of our empirical investigation into the aspects of adopting LLMs for SE activities in industrial settings via a multi-case study. A few of the major findings from the case study focus on aspects like user experiences and perceptions surrounding (i) the preference for LLMs to be used as AI assistants, (ii) the importance of stakeholder satisfaction in the LLM-output evaluation, (iii) scoping the applicability of fine-tuned LLMs, (iv) the effect of LLMs on SE workflows, (v) the necessity and directions for developing human oversight and agency mechanisms, and (vi) necessary skills for practitioners when leveraging LLMs within SE activities. The seven recommendations for the adoption of LLMs within SE activities in practical settings were synthesised based on the findings of our study. To validate the perceived relevance of these seven recommendations, we conducted a survey with software practitioners from various industries. The results of the survey indicate a good level of agreement with most of the presented recommendations from the survey participants. The implications of our post-hoc mapping highlight that assessing AIA compliance only after developing solutions, rather than integrating AIA principles during development, falls short to ensure full compliance. Accordingly, the mapping serves as a contextual "heads-up" that illustrates where the proposed recommendations already contribute to Trustworthy AI principles, while also indicating where further work is needed for broader compliance.

Building upon the findings of this study, future work should therefore aim to extend the scope of the current recommendations to investigate further the remaining Trustworthy AI principles. In line with the objectives and methodology of this study, such extensions should be empirically grounded and informed by practitioners' perspectives, leading to additional guidance that complements existing recommendations rather than retroactively reshaping them. Since recommendation R5, concerned with the need for human oversight and validation

mechanisms, received the highest agreement rate from the survey respondents, researching and developing human oversight mechanisms for employing LLMs responsibly in practical settings is another crucial area of research. Moreover, there is a need for developing further prescriptive artefacts that facilitate the adoption of LLMs into SE activities within industrial settings. As LLMs become more integrated, the evolving role of engineers is an important area of focus. How will their responsibilities shift? What skills and training will be essential? Investigating these questions will be key in guiding the efficient and responsible adoption of LLMs within software organisations.

CRedit authorship contribution statement

Krishna Ronanki: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Beatriz Cabrero-Daniel:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Tomas Herda:** Resources, Project administration. **Stefan Sitkovich:** Project administration. **Jennifer Horkoff:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Christian Berger:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Vinnova project ASPECT [2021-04347] and Wallenberg AI, Autonomous Systems and Software Program (WASP).

Data availability

The data that has been used is confidential.

References

- [1] T. Brown, et al., Language models are few-shot learners, *Adv. Neural Inf. Process. Syst.* 33 (2020) 1877–1901.
- [2] S. Peng, et al., The impact of AI on developer productivity: Evidence from GitHub copilot, 2023, URL: <https://arxiv.org/abs/2302.06590>, arXiv:2302.06590.
- [3] J. White, et al., ChatGPT prompt patterns for improving code quality, refactoring, requirements elicitation, and software design, 2023, arXiv:2303.07839.
- [4] Y. Bang, et al., A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity, 2023, URL: <https://arxiv.org/abs/2302.04023>, arXiv:2302.04023.
- [5] A. Fan, B. Gokkaya, M. Harman, M. Lyubarskiy, S. Sengupta, S. Yoo, J.M. Zhang, Large language models for software engineering: Survey and open problems, in: 2023 IEEE/ACM International Conference on Software Engineering: Future of Software Engineering, ICSE-FoSE, 2023, pp. 31–53, <http://dx.doi.org/10.1109/ICSE-FoSE59343.2023.00008>.
- [6] Microsoft, Azure openai service, 2023, URL: <https://azure.microsoft.com/en-gb/products/ai-services/openai-service/>. (Accessed 23 October 2023).
- [7] Microsoft, Microsoft copilot, 2023, URL: <https://www.microsoft.com/en-us/microsoft-copilot>. (Accessed 18 December 2023).
- [8] GitHub, GitHub copilot, 2023, URL: <https://github.com/features/copilot>. (Accessed 23 October 2023).
- [9] OpenAI, Introducing ChatGPT enterprise, 2023, URL: <https://openai.com/blog/introducing-chatgpt-enterprise>. (Accessed 23 October 2023).
- [10] X. Hou, et al., Large language models for software engineering: A systematic literature review, *ACM Trans. Softw. Eng. Methodol.* 33 (8) (2024) <http://dx.doi.org/10.1145/3695988>.
- [11] B.A. Kitchenham, T. Dyba, M. Jorgensen, Evidence-based software engineering, in: *Proceedings. 26th International Conference on Software Engineering, IEEE, 2004*, pp. 273–281.

- [12] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, *Experimentation in Software Engineering*, Springer Science & Business Media, 2012.
- [13] T. Goyal, J.J. Li, G. Durrett, News summarization and evaluation in the era of GPT-3, 2023, URL: <https://arxiv.org/abs/2209.12356>, arXiv:2209.12356.
- [14] R. Nakano, et al., WebGPT: Browser-assisted question-answering with human feedback, 2022, URL: <https://arxiv.org/abs/2112.09332>, arXiv:2112.09332.
- [15] Q. Xie, et al., A survey on biomedical text summarization with pre-trained language model, 2023, URL: <https://arxiv.org/abs/2304.08763>, arXiv:2304.08763.
- [16] K. Kheiri, H. Karimi, Sentimentgpt: Exploiting GPT for advanced sentiment analysis and its departure from current machine learning, 2023, URL: <https://arxiv.org/abs/2307.10234>, arXiv:2307.10234.
- [17] N. Kandpal, H. Deng, A. Roberts, E. Wallace, C. Raffel, Large language models struggle to learn long-tail knowledge, in: *International Conference on Machine Learning*, PMLR, 2023, pp. 15696–15707.
- [18] Y. Zhang, et al., Siren's song in the AI ocean: A survey on hallucination in large language models, 2025, URL: <https://arxiv.org/abs/2309.01219>, arXiv:2309.01219.
- [19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, 2021, arXiv:2005.11401.
- [20] European Commission, Ethics guidelines for trustworthy artificial intelligence (AI), 2019, URL: www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.
- [21] European Commission, Artificial intelligence act, 2024, URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>. (Accessed 06 March 2025).
- [22] G. Pereira, R. Prikladnicki, V. Jackson, A. van der Hoek, L. Fortes, I. Macaubas, Early results from a study of genai adoption in a large Brazilian company: The case of globo, in: *Generative AI for Effective Software Development*, Springer Nature Switzerland, Cham, 2024, pp. 275–293, http://dx.doi.org/10.1007/978-3-031-55642-5_13.
- [23] T. Süße, M. Kobert, S. Grapenthin, B.-F. Voigt, AI-powered chatbots and the transformation of work: Findings from a case study in software development and software engineering, in: *Collaborative Networks in Digitalization and Society 5.0*, Springer Nature Switzerland, Cham, 2023, pp. 689–705.
- [24] X. Wang, Y. Tian, K. Huang, B. Liang, Practically implementing an LLM-supported collaborative vulnerability remediation process: A team-based approach, *Comput. Secur.* 148 (2025) 104113, <http://dx.doi.org/10.1016/j.cose.2024.104113>.
- [25] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, *Empir. Softw. Eng.* 14 (2009) 131–164.
- [26] V. Clarke, V. Braun, in: T. Teo (Ed.), *Thematic Analysis*, Encyclopedia of Critical Psychology, 2014, pp. 1947–1952, http://dx.doi.org/10.1007/978-1-4614-5583-7_311.
- [27] K. Kumar, Fundamentals of generative AI, in: Y. Vasimalla, S. Kumar (Eds.), *Generative AI for Photonic Sensing*, Springer Nature Singapore, Singapore, 2025, pp. 33–77, http://dx.doi.org/10.1007/978-981-95-1561-5_2.
- [28] C. Robson, *Real World Research: A Resource for Social Scientists and Practitioner-Researchers*, Blackwell, 1993, URL: .
- [29] P.E. Strandberg, Ethical interviews in software engineering, in: 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM, 2019, pp. 1–11, <http://dx.doi.org/10.1109/ESEM.2019.8870192>.
- [30] S. Baltes, P. Ralph, Sampling in software engineering research: A critical review and guidelines, *Empir. Softw. Eng.* 27 (4) (2022) <http://dx.doi.org/10.1007/s10664-021-10072-8>.
- [31] C. Seaman, Qualitative methods in empirical studies of software engineering, *IEEE Trans. Softw. Eng.* 25 (4) (1999) 557–572, <http://dx.doi.org/10.1109/32.799955>.
- [32] M. Lombard, J. Snyder-Duch, C.C. Bracken, Content analysis in mass communication: Assessment and reporting of intercoder reliability, *Hum. Commun. Res.* 28 (4) (2002) 587–604, <http://dx.doi.org/10.1111/j.1468-2958.2002.tb00826.x>.
- [33] N. McDonald, S. Schoenebeck, A. Forte, Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice, 3, (CSCW) 2019, <http://dx.doi.org/10.1145/3359174>.
- [34] V.D. Datta, et al., GREAT AI in medical appropriateness and value-based-care, in: *Big Data and Artificial Intelligence*, Springer Nature Switzerland, 2023, pp. 16–33.
- [35] J. He, C. Treude, D. Lo, LLM-based multi-agent systems for software engineering: Literature review, vision and the road ahead, *ACM Trans. Softw. Eng. Methodol.* (2025) <http://dx.doi.org/10.1145/3712003>.
- [36] S. Hong, et al., MetaGPT: Meta programming for a multi-agent collaborative framework, 2023, URL: <https://arxiv.org/abs/2308.00352>, arXiv:2308.00352.
- [37] J.R. Kotter, *Leading change - why transformation efforts fail*, *Harv. Bus. Rev.* (2007).
- [38] H. Mintzberg, J.A. Waters, Of strategies, deliberate and emergent, *Strat. Manag. J.* 6 (3) (1985) 257–272, <http://dx.doi.org/10.1002/smj.4250060306>.
- [39] M.A. Haque, LLMs: A game-changer for software engineers? *BenchCouncil Trans. Benchmarks*, *Stand. Eval.* 5 (1) (2025) <http://dx.doi.org/10.1016/j.tbench.2025.100204>.
- [40] R. Parasuraman, T. Sheridan, C. Wickens, A model for types and levels of human interaction with automation, *IEEE Trans. Syst. Man, Cybern. - Part A: Syst. Humans* 30 (3) (2000) 286–297, <http://dx.doi.org/10.1109/3468.844354>.
- [41] G. Fragiadakis, C. Diou, G. Kousiouris, M. Nikolaidou, Evaluating human-AI collaboration: A review and methodological framework, 2025, URL: <https://arxiv.org/abs/2407.19098>, arXiv:2407.19098.
- [42] Q. Roy, F. Zhang, D. Vogel, Automation accuracy is good, but high controllability may be better, in: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 1–8, <http://dx.doi.org/10.1145/3290605.3300750>.