Can We Trust AI Agents? A Case Study of an LLM-Based Multi-Agent System for Ethical AI

José Antonio Siqueira de Cerqueira[1,*], Mamia Agbese[2], Rebekah Rousi[3], Nannan Xi[1], Juho Hamari[1] and Pekka Abrahamsson[1]

[1]Tampere University, Tampere, Finland
[2]University of Jyväskylä, Jyväskylä, Finland
[3]University of Vaasa, Vaasa, Finland

Abstract

AI-based systems, including Large Language Models (LLMs), impact millions by supporting diverse tasks but face issues like misinformation, bias, and misuse. AI ethics is crucial as new technologies and concerns emerge, but objective, practical guidance remains debated. This study explores the extent to which trustworthiness-enhancing techniques in LLMs can support the development of ethically aligned AI software. We adopt a single exploratory cycle of Design Science Research (DSR). First, we identify trustworthiness-enhancing techniques for LLMs: multi-agents, distinct roles, structured communication, and multiple rounds of debate. Second, we design a multi-agent prototype LLM-MAS in which agents address real-world AI ethics issues from the AI Incident Database. Finally, we evaluate the prototype across three case scenarios using thematic analysis, hierarchical clustering, a baseline comparison, and code execution. The system generates approximately 2,000 lines of code per case, compared to only 80 lines in baseline trials. Results reveal terms like bias detection, transparency, accountability, user consent, GDPR compliance, fairness evaluation, and EU AI Act compliance, showing this prototype ability to generate extensive source code and documentation addressing often overlooked AI ethics issues. However, practical challenges in source code integration and dependency management may limit its use by practitioners.

Keywords

Ethics, Artificial Intelligence, Trustworthiness, Large Language Models

1. Introduction

Artificial Intelligence (AI) is emerging as a transformative force, reshaping industries, economies, and everyday life. Despite its rapid development and adoption, a number of negative reports on its use highlight the importance of adhering to ethical norms and principles [1]. Several ethical guidelines are available with plenty of ethical principles providing high level and abstract guidance for developers and stakeholders on AI ethics [2]. These are important, but there is a lack of practical guidance for developers to operationalise ethics in AI [2]. As new AI-based technologies emerge, ethics in AI will also become increasingly critical, such as the latest breakthrough: Large Language Models (LLMs).

LLMs are becoming ubiquitous and have significant impact on decision-making processes and human interactions [3, 4]. LLMs are advanced AI algorithms capable of generating, interpreting, and predicting text based on vast amounts of data they have been trained on [5]. Of these LLMs, Generative Pre-trained Transformer (GPT) LLMs, such as GPT-4o, have showcased unprecedented proficiency in the human language, coding, logic, reasoning, and other associated natural language tasks [4]. They excel at guiding complex conversations and are used in countless endeavours such as software engineering (SE) - as in AI for Software Engineering (AI4SE) [6] - and qualitative research [7].

Within the AI4SE field, the capabilities of LLMs are being explored in the software development, maintenance, and evolution [8, 9, 7, 6], namely LLM4SE. Accordingly, they find applications across

8th Conference on Technology Ethics (TETHICS2025), November 11–12, 2025, Vaasa, Finland
*Corresponding author.
jose.siqueiradecerqueira@tuni.fi (J. A. S. d. Cerqueira); mamia.o.agbese@jyu.fi (M. Agbese); rebekah.rousi@uwasa.fi (R. Rousi); nannan.xi@tuni.fi (N. Xi); juho.hamari@tuni.fi (J. Hamari); pekka.abrahamsson@tuni.fi (P. Abrahamsson)
0000-0002-8143-1042 (J. A. S. d. Cerqueira); 0000-0002-5479-7153 (M. Agbese); 0000-0001-5771-3528 (R. Rousi); 0000-0002-9424-8116 (N. Xi); 0000-0002-6573-588X (J. Hamari); 0000-0002-4360-2226 (P. Abrahamsson)

various stages of the software development life cycle, including requirement analysis, software design, code implementation, testing, refactoring, defect detection, and maintenance [9]. Albeit the notable use of LLM in SE, to the best of our knowledge there are no studies that focus on the use of LLM in the ethical AI development. In qualitative research, LLMs are also gaining prominence, particularly as a supplement to tasks traditionally performed by humans, e.g., analysing qualitative data [10, 11, 12]. Specifically, in the coding process - where qualitative data is organised and interpreted by assigning codes or labels to text segments or different data forms - LLMs have demonstrated significant utility [13, 11].

However, several concerns arise, especially related to trustworthiness, as more practitioners are relying on LLMs to perform their task [6]. Several studies are drawing attention to possible problems in adopting LLMs for SE, in particular when syntactically correct but non-functional code is produced, which affects the reliability and effectiveness of LLM-based code generation [14]. Pearce et al. [15] used GitHub Copilot to produce 1,689 programs, and found that 40% of them have security vulnerabilities. Liu et al. [16] systematically analysed ChatGPT code generation reliability identifying quality issues as many of the programs generated provided wrong output or contained compilation or runtime error.

Effective AI4SE should be trustworthy and synergistic with the practitioner's workflow, otherwise "such AI4SE solutions risk becoming obstacles rather than facilitators" [6]. Currently, there is a growing need for context-dependent empirical studies that explore how trust in LLMs affects their adoption in software engineering tasks [6]. Furthermore, there remains a significant gap in research focused on the practical operationalisation of AI ethics, particularly through the application of LLMs. To date, no studies have attempted to explore the development of ethical AI-based systems using LLMs. We aim to explore trustworthiness-enhancing techniques in LLMs in the ethical AI development context.

The following Research Question guides this study:

- RQ: To what extent can trustworthiness-enhancing techniques in LLM-based systems support the development of ethically aligned AI software?

To this end, we employed an adapted single cycle of the Design Science Research (DSR) method [17] to identify trustworthiness-enhancing techniques, build a prototype, evaluate it through three case studies and communicate the results. We identified four trustworthiness-enhancing techniques for LLM-based systems in the literature and implemented them: **(1) multi-agent collaboration, (2) specialised roles, (3) structured communication, and (4) multiple rounds of debate.** Furthermore, we evaluated the prototype through three case studies based on real-world AI incidents in recruitment, deepfake detection, and image classification. Each case is analyzed using (1) LLM for thematic analysis, (2) iRaMuTeQ for hierarchical clustering, (3) a comparative baseline with standard ChatGPT GPT-4o prompts, and (4) source code execution to verify whether the generated code runs, without assessing its runtime behaviour or correctness in depth. **The prototype can generate around 2,000 lines of code (with proper documentation), encompassing themes such as bias detection, transparency, accountability, user consent and GDPR compliance, fairness evaluation and compliance with the EU AI Act.** The hierarchical clustering analysis revealed the ability of the prototype to generate outputs related to legal aspects of ethics in AI through words such as: bias_type, evaluate_bias, bias_score. The baseline study produced around 80 lines of output without source code while covering considerably less AI ethical issues. Therefore, the evaluations conducted suggest that the trustworthiness-enhancing techniques employed can be beneficial in the AI ethical development context. However, practical challenges in source code integration and dependency management may limit its use by practitioners, e.g., deprecated dependencies and modularity. We hope to shed light both on how to improve trustworthiness in LLM and on how to help practitioners to operationalise ethical principles in the development of AI-based systems.

2. Background and Related Work

2.1. Large Language Models

LLMs have been pre-trained on vast amounts of text data, often scraped from the internet, allowing them to learn patterns and structures inherent in human language [18]. One of the key features of LLMs is their ability to generate contextually relevant and apparently coherent text across a wide range of topics [4]. These models can perform tasks such as language translation, text summarization, question answering, and creative writing [18]. They achieve this by leveraging the vast amounts of data they have been trained on to predict the next word or sequence of words in each context [19]. Some well-known examples of LLMs include Generative Pre-trained Transformer (GPT) models developed by OpenAI [20], particularly GPT-4o and o3.

2.2. Trustworthiness in Large Language Models

Recent advancements and use of LLMs have raised concerns about their ethical implications [21, 22]. Issues like information hallucination, biases, and the need for factual correctness are critical in determining the trustworthiness of these systems in real-world applications [23, 24]. Trustworthiness refers to the quality or attribute of being reliable, dependable, and deserving of trust. It encompasses several key components that establish trust in a person, organisation, product, or system [25]. In the context of LLMs, trustworthiness largely refers to their reliability and ethical adherence to social norms [3, 25]. Notably, several studies have formulated taxonomies accompanied by evaluation methodologies, wherein the taxonomy acts as a measure for assessing trustworthiness, with the identified aspects serving as evaluation parameters. Among these efforts, the Holistic Evaluation of Language Models (HELM) devised by Liang et al. [22] introduces a comprehensive taxonomy consisting of seven **metrics** — accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency — and applied this framework to analyze 30 different models. Notably, the authors stress the significance of adopting a multi-metric approach, enabling metrics prioritization beyond mere accuracy.

Wang et al. [21] provide eight trustworthiness **perspectives** of language models: toxicity, stereotype and bias, adversarial robustness, out-of-distribution robustness, privacy, robustness to adversarial demonstrations, machine ethics, lastly, fairness. The authors state that they only provided objective definitions for some of the perspectives coined - due to their intrinsic subjectivity, e.g., fairness toxicity - leaving the subjective exploration of model behaviours based on human understanding as future work. As scholars delve into characterizing trustworthiness within LLMs, a unified terminology for measuring it remains unclear. Furthermore, it is seen that there is no clear consensus on a framework to build aligned LLMs, nor clear assessment guidelines, both posing as unresolved issues [3].

2.3. Trustworthiness enhancing techniques in LLM4SE

Hong et al. [26] propose MetaGPT to simulate a group of agents as software company. Worth noting the use of **specialised roles** (Product Manager, Architect, Project Manager, Engineer, and QA Engineer), **workflow for the agents** (all agents work in a sequential fashion), and a **structured communication interface**. An interesting point regarding the specialised roles, is the ability for agents to have skills, i.e., perform actions, such as being able to run code and search on the web. The authors argue that the incorporation of aforementioned techniques can considerably enhance code generation, as well as mitigating hallucinations risks.

Qian et al. [27] developed ChatDev, a chat-based software development framework using LLMs to enhance communication and collaboration across various roles, aimed at reducing hallucination risks like bugs and missing dependencies. Though, it does generate different outputs for the same inputs and carries risks due to untested code.

Meanwhile, Wu et al. [28] from Microsoft introduced AutoGen, a multi-agent LLM framework enhancing coding productivity by structuring tasks among specialised roles (Commander, Writer, Safeguard) that interact to refine outputs. Through an ablation study they demonstrated that this method

significantly outperforms single-agent systems by breaking down complex tasks into manageable components.

The aforementioned studies run in line with a joint study by MIT and Microsoft, Du et al. [24]. Albeit not using LLM-based multi-agents to directly generate code, the authors argue that using multiple agents along with multiple rounds of debate improve the reasoning and factuality of the generated solution compared to not using them. Hence, multiple agents and rounds of debate can reduce hallucinations and increase the overall trustworthiness of using LLMs. Such techniques can exhibit emergent behaviours that are not evident when considering individual agents in isolation [26, 29].

2.4. AI Ethics

Despite the recent academic and industry interest, AI ethics is a long-standing area of research, in which diverse incidents have recently raised public concerns about its use and development [30]. Consequently, in the past few years a variety of principles and guidelines have emerged, from various sources (academia, industry, civil society) to frame and define what is ethical AI [31]. Ryan and Stahl [32] provide 11 ethical principles along with ethical issues related to AI: 1) Transparency, 2) Justice and Fairness, 3) Non-maleficence, 4) Responsibility, 5) Privacy, 6) Beneficence, 7) Freedom and Autonomy, 8) Trust, 9) Sustainability, 10) Dignity, 11) Solidarity. These ethical principles serve as the basis of our analysis.

The European Parliament is currently putting into force the world's first regulation on AI [33], which highlights the fact that the plethora of guidelines available are in fact soft law, with no legally binding nor real consequences [2]. Nevertheless, it is reminiscent of the abstract principles on how AI ethics is mainly approached [30, 31, 34]. Thus, the obstacles in operationalising ethical principles in AI-based systems stem from the subjectivity necessary to interpret and put into practice highly abstract principles by the practitioners [2]. Despite its importance, it has often been a task commonly taken as an afterthought [1]. As opposed to other efforts available in the literature, we aim to study the use of LLM-based multi-agent systems in the development process of AI-based systems, taking into consideration the operationalization of ethical principles from the first stage of the development process onwards. The use of LLM agents in developing ethically-aligned AI systems introduces significant complexity and novelty, making their ethical implications a compelling area of study. LLMs face unique challenges in producing accurate, unbiased text and source code while adhering to ethical guidelines. Their dynamic and often unpredictable outputs in multi-agent interactions provide a valuable opportunity to study the application of ethical principles in real-world settings, although there is a noted lack of literature on the operationalization of AI ethics through LLM agents.

2.5. LLMs for Qualitative Analysis

Qualitative analysis is another opportune expanding area where LLMs can be applied [35, 36, 12], particularly in qualitative research within the field of software engineering [37, 11]. Thematic analysis is labour-intensive and time-consuming for humans, and are prone to mistakes and bias [38]. Therefore, LLMs present an opportunity to improve the accuracy and efficiency of this method. In thematic analysis, coders require multiple rounds of discussion to achieve consensus and resolve ambiguities for an in-depth understanding of the data [38].

Thus, LLMs can tackle the challenges of traditional qualitative research, such as the time-intensive nature of data analysis, limited generalisability, consistency issues, and personal subjective biases [37, 10]. Hamilton et al. [10] argue that both humans and LLMs have identified unique and overlapping themes from interview data, indicating the potential for recognising patterns and themes in qualitative data. Drápal et al. [36] proposed a framework for collaboration between legal experts and LLM for performing thematic analysis. The LLM used could identify themes that would probably be missed due to human error. Also, they discuss that LLM can perform the initial coding of the data with reasonable quality. De Paoli [39] states that "there is no denying that the model can infer codes and themes and even identify patterns which researchers did not consider relevant and contribute to better qualitative

analysis."

Differently from the aforementioned approaches, our study aims to provide case studies using LLMs automating the whole thematic analysis process, as well as implementing the trustworthiness-enhancing techniques found to propose a prototype that generates source code in the AI ethical development context. For the best of our knowledge, there are no works in the literature that aims to use LLM to operationalise AI ethical principles.

3. Methodology

In this study, we used an adapted DSR approach [17] to explore and enhance the trustworthiness of LLMs for software engineering tasks. We chose this method because it is a problem-solving paradigm that advances knowledge through the building and evaluation of artefacts [17]. In our study, its role is to provide a structured approach in four phases: (1) exploration, (2) prototyping, (3) evaluation through case studies and (4) communication of results. Due to the limited scope of this initial case study, we implemented only a single exploratory cycle. The evaluation phase consists of applying the LLM-MAS prototype to three case scenarios, each derived from real-world AI incidents, and analysing the generated outcomes using thematic analysis, hierarchical clustering, comparative baselines, and source code execution. These case studies serve to examine how the identified trustworthiness-enhancing techniques impact the development of ethically aligned AI-based systems in practice. This Section will cover phases 1 to 3, while phase 4 will be presented in Section 4.

3.1. Exploring

Initially, we explored the existing literature and identified the need to improve trustworthiness in AI4SE, as presented in Section 2.2. This need is most clearly demonstrated in the work of David Lo [6]. Furthermore, we identified recently proposed techniques in the literature to enhance trustworthiness in LLMs, particularly in the context of software engineering (i.e. LLM4SE). In summary, the trustworthiness-enhancing techniques identified and presented in Section 2.3 provide the basis for our conceptual approach, which is visualised in Table 1.

Table 1
Trustworthiness-enhancing techniques adopted

Trustworthiness-enhancing technique	Description
Multi-agent collaboration	Multiple agents interact on a task
Specialised roles	Agents assume roles, e.g., Developer or Ethicist
Structured communication	Predefined message passing structure
Multiple debate rounds	Iterative dialogue to refine results

Next, we will present a detailed overview of the technical implementation of the prototype.

3.2. Prototype: LLM-MAS Technical Implementation

We devised a prototype of an LLM-based Multi-Agent System (henceforth referred to as the LLM-MAS). The LLM-MAS is a Python script consisting of three agents that make use of the GPT-4o-mini model via the OpenAI API. The model was selected on the basis of its cost-efficiency, accessibility, and strong reasoning performance, which rendered it a practical choice for iterative experimentation involving multiple agents. Rather than being equipped with memory or tool access, the agents are LLMs with custom instructions in the form of system prompts that describe their roles and the communication structure. Consequently, they have different specialised roles, and their conversations are structured in the same way. Specifically, there are **two senior Python developers and one AI ethicist**. The **conversation structure** comprises: **reply, reflection, code, and critique**. The agent

prompts are available in the Zenodo: agent1_prompt_PythonDev.txt, agent2_prompt_PythonDev.txt and agent3_prompt_AIEthicist.txt.

The workflow for the agents is sequential. A round is defined as a sequence of conversations between agents 1, 2 and 3. **The conversation history is appended after each agent output, and each agent receives the entire conversation history as an input.** The number of rounds is modifiable. We selected five rounds per project to allow agents enough interaction to develop and refine ideas collaboratively, while keeping the output concise and manageable for analysis. Figure 1 presents an overview of our prototype.

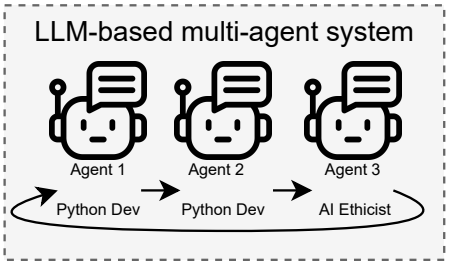

Figure 1: Prototype overview

The agents engage in a structured conversation, through multiple rounds, to discuss and collaborate on a project. That is, as a team, they receive an input – a project description (PD) – that they should work on collaboratively to implement, through structured, iterative discussions.

3.3. Evaluation: Case Studies

We begin with an overview of the evaluation approach, followed by an in-depth explanation of each component.

To evaluate the effectiveness of the prototype, we conducted a case study-based evaluation across three distinct scenarios inspired by real-world AI incidents. **Each case study** involved applying the LLM-MAS prototype to a different project description, enabling us to assess how the system behaved in various ethical and technical contexts. For each case, we employed a multi-faceted evaluation strategy. First, we performed a **thematic analysis** using an LLM to automatically identify and categorise key ethical and technical themes in the output. This helped us to assess the system's ability to address ethical requirements. Secondly, **hierarchical clustering** (via iRaMuTeQ) was used to group related text segments, providing an additional layer of validation by highlighting patterns in the generated content. Thirdly, we conducted a **comparative baseline study**, prompting the same project descriptions directly to GPT-4o in ChatGPT interface (without any multi-agent configuration), to evaluate the added value of the proposed trustworthiness-enhancing techniques. Figure 2 presents an overview of the evaluation methodology approached.

Due to the probabilistic nature of LLMs, running the same input multiple times can yield different outputs. This means that even when using the same system configuration, each run may produce a different conversation, source code, and thematic analysis. Prior work has similarly observed this variability in multi-agent LLM systems [27, 6]. Thus, in order to enrich our analysis, we ran the system three different times to obtain distinct output (yielding the raw output, e.g., $O1_1$, $O1_2$, $O1_3$ for the first project description).

Then, the collected data is analysed through the use of ChatGPT GPT-4o with a specific *custom instruction* acting as qualitative analyst, available in the file Custom_instructions_Qualitative_GPT.txt in Zenodo. The prompt used as a header for the thematic analysis in available in file Prompt_Perform_thematic_analysis.txt. Each of the raw output produces a partial thematic analysis (e.g., $B2_1$, $B2_2$, $B2_3$ for the second project description). Consequently, the partial thematic analysis for each project description is summarised into one thematic analysis through the use of the same GPT-4o, yielding TA, TB and TC. The prompt serving as header for the merging of the thematic analysis is available in file

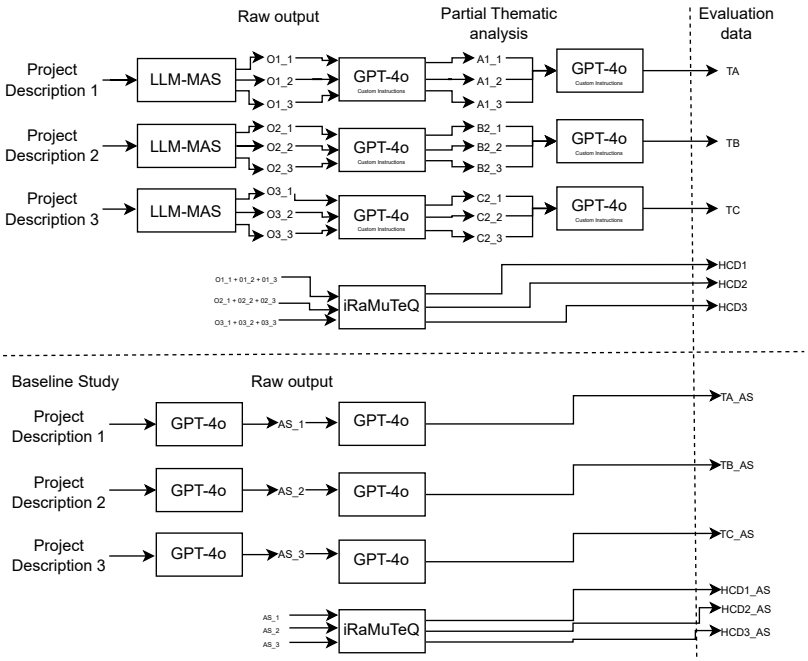

Figure 2: Evaluation overview

Prompt_Merging_Thematic_Analysis.txt, equipped with the same custom instruction.

We used Descending Hierarchical Classification (DHC) from iRaMuTeQ v0.7-alpha running in R v4.2. The input for the generation of each hierarchical clustering is the append of the raw output of each project description (e.g., $O3_1$, $O3_2$, $O3_3$ together generates the hierarchical clustering for the third project description.

With the aim of understanding the benefits of this LLM-MAS approach, a baseline study is also performed, in which a project description is prompted to a regular ChatGPT-4o using OpenAI user interface, without any of the techniques found. For example, for project description 1, we have only one output (AS_1), that serves as input for creation of its thematic analysis also with ChatGPT-4o, and for the iRaMuTeQ analysis, yielding TA_AS and HCD_AS respectively.

Finally, we performed a practical evaluation of the prototype by running the source codes generated by the LLM. This step is for evaluating the system's practical utility in real-world software development scenarios. Our methodology offers an initial framework for assessing the trustworthiness and reliability of LLMs in the software engineering domain. The results and discussions arising from the comparison between the different thematic analysis and hierarchical clustering, coupled with the baseline study and source code execution are presented next.

4. Results

In this Section, we present the results for each of the project descriptions used in the case study. Each project description is derived from real AI incidents available at https://incidentdatabase.ai/. The selected incidents focus on AI applications in recruitment, deepfake detection, and image classification. They were selected due to their high relevance to AI ethics and the need for compliance with regulatory standards. All data generated during the study – including agent prompts, custom instructions, and thematic analyses – are available in Zenodo.

4.1. Project Description 1

The first project description derived from Incident 37: Female Applicants Down-Ranked by Amazon Recruiting Tool:

Table 2
Thematic Analysis PD 1

Theme	Codes
Ethical and Regulatory AI Development	Compliance with the EU AI Act, Ethical considerations in AI development, Bias elimination strategies
Fair and Transparent AI Implementation	Technical development of the AI tool, Integration of bias detection mechanisms, ensuring fairness in AI operations
Iterative Model Training and User-Centered Validation	Scenario-based testing, Continuous model refinement, Incorporating actionable user feedback
Transparency in AI Processes	Documentation of AI processes, Clear communication of updates and improvements, Stakeholder engagement and transparency

> Develop an AI-powered recruitment tool designed to screen resumes impartially, complying with the EU AI Act. The project aims to eliminate biases related to gender and language, ensuring fair evaluation of all applicants. The AI Ethics Specialist will guide the team in addressing ethical concerns and risk levels. The senior Python developers will utilise NLP to process resumes, referencing relevant EU AI Act guidelines.

It is important to note that the project description references the EU AI Act, but the LLM used may not have had access to its full content at the time this study was conducted. In future work, we will introduce this document as a vector to be embedded and accessed by an open-source LLM. Moreover, according to David Lo [6], guaranteeing that AI4SE solution will comply with legislation is one way to enhance trustworthiness in AI4SE, since practitioners need to avoid conflict with laws.

Raw data obtained vary from 1,712 lines to 1,976 lines of discussion plus source code. The thematic analysis in Table 2 was obtained from the merge of the three thematic analyses attained from each output. In Figure 3, 5 classes were devised, however, classes such as 1, 3, 4 and 5 are related to technical implementation. Nevertheless, it is possible to understand that themes 1, 2 and 3 are closely related to the second class, where ethical, EU, act appear. Moreover with class number 3, w.r.t functions as `evaluate_bias` and `bias_score`.

In all cases, the source codes produced were spread around the text. As mentioned, this can be almost 2,000 lines long. This makes the source code difficult to differentiate, and sometimes there are pieces that need to be glued together from different rounds of conversation. After trying to run the source code available in O1_1, it was found that some packages were needed to be installed manually. Furthermore, deprecated packages were being used: `sklearn` PyPI package is deprecated, use `scikit-learn`. This results in extra effort by the practitioners, impacting on their workflow.

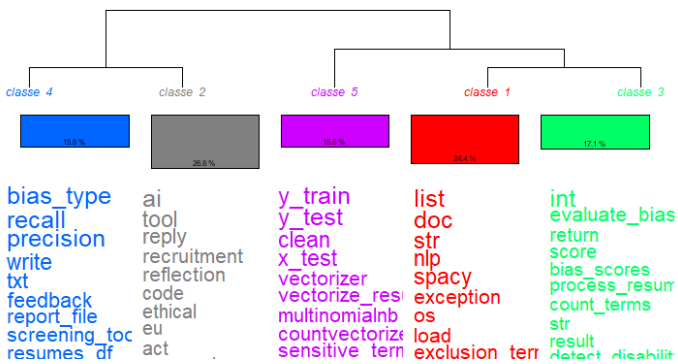

Figure 3: Hierarchical clustering dendogram - PD1

Table 3
Thematic Analysis PD 2

Theme	Codes
System Architecture and Design	Modularity and Scalability, AI Integration and Media Processing, Role-Based Access Control (RBAC)
Ethical Compliance and AI Responsibility	AI Ethics and Bias Detection, Transparency and Accountability, User Consent and GDPR Compliance
Security and Collaboration	User Authentication and Data Security, Role Management and Access Control, Collaboration between Developers and Ethics Specialists
System Reliability and Continuous Improvement	Error Handling and Logging, Feedback Mechanisms, Testing and Maintenance

In regard to the baseline study, only 78 lines of text were obtained, while no code was produced. On the other hand, three themes were devised from the thematic analysis: Ethical AI Development, Technical Implementation and Fairness, and Compliance. It was not possible to generate the hierarchical clustering for any of the baseline studies due to small size of text data available. The quality between both studies is very disparate.

4.2. Project Description 2

The second project description derived from Incident 635: AI-Generated Fake News Targets Black Celebrities on YouTube:

> Create a deepfake content authentication system that uses AI to verify the authenticity of video, image and audio content. This system will distinguish between real and manipulated media, reducing the spread of harmful deepfake content. The AI Ethics Specialist will ensure the tool adheres to ethical standards. The senior Python developers will enhance the system's detection capabilities.

In this PD, there is no reference to EU AI Act directly. However, it is worth pinpointing that it is available in the prompt of the third agent, the AI Ethicist. Raw data obtained vary from 1414 lines to 2178 lines of text data. The discussion here progresses in a similar way to the previous one.

As the relationship between the thematic analysis in Table 3 and the hierarchical clustering produced in Figure 4, it is possible to highlight the class number 1 with theme number 2, Ethical Compliance and AI Responsibility. Moreover, class number 2 is closely related to theme Security and Collaboration, with words such as authentication and logger. Similarly to the previous PD, no source code was produced in the baseline study. This yielded 78 lines of text and themes such as Technological Framework, Ethics and Compliance, User Experience and Support, Strategic Objective. While the study with the proposed prototype provides a more fruitful outcome, yielding almost 2,000 lines of text, the baseline study provides a generic text that can only serve as study material.

In relation to the source code produced, on top of the overhead to developers encountered in the previous PD, in O2_2 it is perceived that our prototype created modules, as in `from your_module import VideoAuthenticator, ImageAuthenticator, AudioAuthenticator`. Again, this poses a problem to practitioners when working with this tool, due to the fact that they will need to go through all the text and manually search for the source code that he/she needs to copy and paste somewhere else, test it, install dependencies, and manage modules.

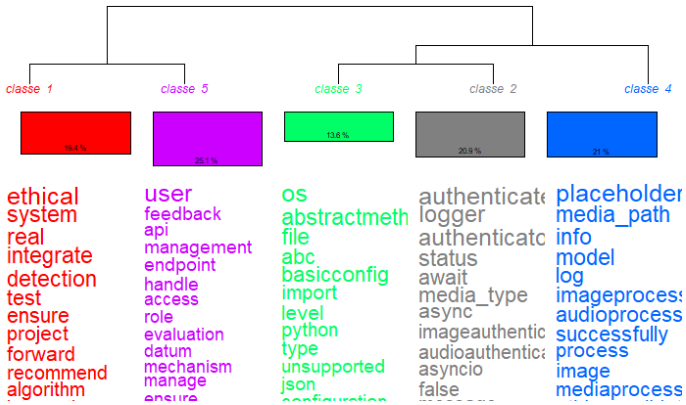

Figure 4: Hierarchical clustering dendogram - PD2

Table 4
Thematic Analysis PD 3

Theme	Codes
Ethical AI Development and Compliance	Ethical AI, Compliance, Bias Detection and Mitigation, Fairness Evaluation, Transparency and Accountability
Technical Implementation, Optimization, and Robustness	Data Management, Model Design and Implementation, Long-term Monitoring and Continuous Improvement, Error Handling and Modular Design
Stakeholder Engagement and Continuous Improvement	Stakeholder Engagement, User Feedback Integration, Collaboration, Transparency in Operations
Data-Driven Evaluation	Statistical Analysis, Confidence Scores, Class-wise Analysis

4.3. Project Description 3

The third project description derived from Incident 16: Images of Black People Labeled as Gorillas:

> Develop an AI-based image processing and classification system that complies with the EU AI Act, ensuring accurate and respectful labelling of images. The system must avoid racial biases and harmful misclassifications. The AI Ethics Specialist will focus on identifying and mitigating psychological harm and harm to social systems. The senior Python developers will implement and refine image classification algorithms.

This discussion follows in a similar fashion to the previous discussions. Raw data obtained vary from 1,676 lines to 2,122 lines of text data. The thematic analysis in Table 4 revolves around Ethical AI Development and Compliance and Technical Implementation, Optimization and Robustness. Class number 1 is related to theme number 1, however, the rest of the classes available contain words that are related to technical implementations, such as `image_path`, `import`, `sklearn`, `sensitive_attribute`, which are related to theme number 2. Interesting to note that `sklearn` is present in a class in Figure 5, even though it had already been recognised to be deprecated. This highlights the challenge of relying on static LLM knowledge in a rapidly evolving software landscape, where tools and dependencies can quickly become outdated.

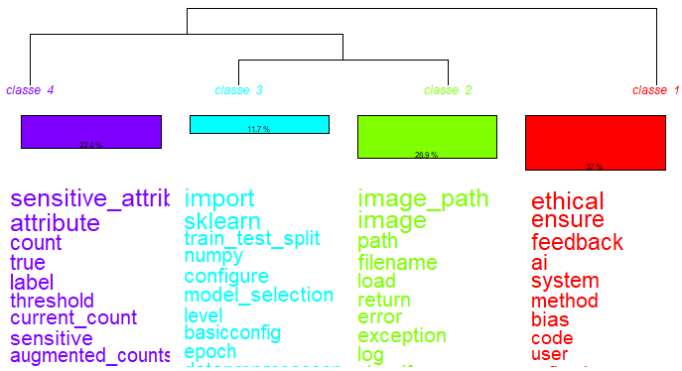

Figure 5: Hierarchical clustering dendogram - PD3

5. Discussion

The results suggest that the trustworthiness-enhancing techniques employed in the prototype, such as multi-agent collaboration, specialised roles, structured communication, and multiple debate rounds, can improve the quality of generated artefacts, as observed in [26, 27, 28]. While LLM agents are increasingly integrated into traditional software engineering tasks (e.g., requirements engineering, code generation, software QA, and maintenance) [40, 41], none systematically study their use in the operationalisation of AI ethics. Thus, our results extend prior work by adding the ethical AI dimension. Specifically, we implement LLM-trustworthiness techniques not only to complete software engineering tasks, but to tackle AI ethical issues, that is, to bring fairness, accountability, transparency, privacy, and legal alignment (EU AI Act, GDPR) into the generated output from the start.

The prototype produced more code and richer ethics-related artefacts (e.g. bias-related functions, consent handling, documentation) compared to the baseline study. This is consistent with evidence that multi-agent debate improves reasoning and factuality [24] and suggests that improving role specialisation in LLM-based agents is a next step [40]. Nonetheless, we observed practical issues that matter for adoption: scattered modules, dependency management, and extra packaging/integration work. These align with calls to improve developer–AI synergy in AI4SE [6]. For practice, this gives developers tools to consider ethical impacts during AI system development. For research, it invites further work to test, compare, and improve these techniques.

6. Threats to Validity

This study has several limitations that may affect the validity and generalisability of its findings. First, the prototype relies on a proprietary LLM by OpenAI, which may evolve over time in undocumented ways. This creates a reproducibility challenge, as future versions of the model may behave differently, even under identical conditions. Secondly, the evaluation relies on an LLM to perform a thematic analysis of the output, which could raise questions regarding its validity and objectivity. Although recent studies (e.g. [10]; [36]) have demonstrated the potential of LLMs to support, and even enhance, qualitative analysis by identifying patterns that human coders might overlook, this approach remains methodologically debated. In our study, we did not conduct a parallel human-led thematic analysis, which limits our ability to triangulate or verify the consistency of the themes generated. This increases the risk that the LLM-based analysis will reflect the model's own limitations or biases. Future work should either involve expert human analysts to validate the AI-generated codes and themes or adopt a hybrid human-in-the-loop approach to enhance analytical rigour. Third, the thematic analysis and hierarchical clustering processes were fully automated and lacked human validation. While these methods offer efficiency and consistency, they may miss subtle, context-sensitive insights that would be more easily detected by domain experts, particularly in the ethical dimensions of software engineering.

Fourth, although the prototype generates source code, our execution tests were limited to confirming

whether the code runs, rather than assessing its functionality, correctness, or ethical alignment. This leaves open questions about the practical reliability and appropriateness of the generated software. Finally, the study was limited to three case scenarios. While these were selected for their ethical relevance, the small sample size restricts the scope of generalisability. Future work should expand on this with broader case coverage, incorporate human-in-the-loop validation, and include a more rigorous assessment of the ethical dimensions of the generated artefacts. It is also important to note that the evaluation was conducted using a case study approach. Our objective was not to produce generalisable results, but rather to explore and gain in-depth insights into how trustworthiness-enhancing techniques in LLM-based systems can support the development of ethically aligned AI. Case studies are particularly valuable for examining complex, real-world phenomena in context, and for uncovering practical challenges and opportunities that may not emerge in more controlled or large-scale experiments [42].

7. Conclusion

This study aimed to explore trustworthiness in AI4SE, specifically in the use of LLM for software engineering (LLM4SE), in the context of ethically aligned AI-based systems development. To the best of our knowledge, there are no studies in the literature that address practical AI ethics using LLM. Our Research Question is: To what extent can trustworthiness-enhancing techniques in LLM-based systems support the development of ethically aligned AI software. In order to provide an answer, we used the Design Science Research method, we identified the motivation and different trustworthiness-enhancing techniques in LLM4SE, proposed a prototype - an LLM-based multi-agent system (LLM-MAS) -, and evaluated it through three case studies. The techniques discovered to improve trustworthiness are (1) multi-agent collaboration, (2) specialised roles, (3) structured communication, and (4) multiple rounds of debate. Each case study had a project description derived from real AI incidents extracted from the AI Incident Database in recruitment, deepfake detection, and image classification. For each project description, LLM-MAS produced about 2,000 lines of text, encompassing source code and documentation to implement the AI-based systems.

Some of the codes produced through thematic analysis include: bias detection, transparency and accountability, user consent and GDPR compliance, fairness evaluation and compliance with the EU AI Act. In relation to hierarchical clustering, some of the words that depict ethics in AI include: ethical, bias, `bias_type`, `evaluate_bias`, `bias_score`. This highlights the ability of LLM-MAS to generate outputs related to legal aspects of ethics in AI (i.e. GDPR and EU AI Act), in addition to theoretical and practical ethical issues, both in the documentation and in the source code, through codes in the thematic analysis (e.g., transparency, fairness, bias) and functions detected in the hierarchical clustering (e.g., `evaluate_bias`), respectively. Regarding the baseline study, for each project description it produced around 80 lines of output without source code, covering considerably less AI ethical issues. Therefore, this study underlines the potential of employing trustworthiness-enhancing techniques in LLMs to the development of more ethically aligned AI-based systems.

Nevertheless, there are some issues that hinder the synergy for practitioners [6]. Worth highlighting is the lack of practicality, from scraping the source code from the text - even more difficult when it comes to modules -, and installing packages and dealing with deprecated dependencies stemming from the LLM training date. While it is possible to improve trustworthiness and quality in the use of LLM4SE, there are still open issues to work on to make it more practical for practitioners. Whilst this experiment was conducted relying only on the internal knowledge of the Open AI LLM, it is important to note that providing the AI ethical agent with legislation (e.g., GDPR and EU AI Act) or AI ethical methods or guidelines (e.g., ECCOLA [1]) is a prospective future work that we intend to pursue. Moreover, in-depth analysis of the source code provided by the prototype is necessary to assess ethical alignment. With this study, we hope to further the understanding of the benefits of trustworthiness-enhancing techniques in LLM for the development of ethical AI-based systems. We aim to open source the prototype source code in further studies.

Acknowledgments

This research was supported by CONVERGENCE of Humans and Machines (220025) and the EVIL-AI "The identification and the mitigation of the negative effects of Artificial Intelligence Agents" (JAES/2024/EVIL-AI) projects by Jane and Aatos Erkko Foundation and the "Multifaceted ripple effects and limitations of human-AI interplay at work, business and society (SYNTHETICA)" project (358714) by Research Council of Finland.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to perform grammar and spelling checks, and to enhance clarity in some portions of the text. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] V. Vakkuri, K. Kemell, M. Jantunen, E. Halme, P. Abrahamsson, ECCOLA - A method for implementing ethically aligned AI systems, J. Syst. Softw. 182 (2021) 111067. doi:10.1016/J.JSS.2021.111067.

[2] J. A. S. de Cerqueira, A. P. D. Azevedo, H. A. T. Leão, E. D. Canedo, Guide for artificial intelligence ethical requirements elicitation - RE4AI ethical guide, in: 55th Hawaii International Conference on System Sciences, HICSS 2022, Virtual Event / Maui, Hawaii, USA, January 4-7, 2022, ScholarSpace, 2022, pp. 1–10. URL: http://hdl.handle.net/10125/80015.

[3] Y. Liu, Y. Yao, J.-F. Ton, X. Zhang, R. G. H. Cheng, Y. Klochkov, M. F. Taufiq, H. Li, Trustworthy llms: a survey and guideline for evaluating large language models' alignment, arXiv preprint arXiv:2308.05374 (2023).

[4] P. P. Ray, Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope, Internet of Things and Cyber-Physical Systems 3 (2023) 121–154. doi:10.1016/j.iotcps.2023.04.003.

[5] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, arXiv preprint arXiv:2307.03109 (2023).

[6] D. Lo, Trustworthy and synergistic artificial intelligence for software engineering: Vision and roadmaps, in: IEEE/ACM International Conference on Software Engineering: Future of Software Engineering, ICSE-FoSE 2023, Melbourne, Australia, May 14-20, 2023, IEEE, 2023, pp. 69–85. doi:10.1109/ICSE-FOSE59343.2023.00010.

[7] B. Ni, M. J. Buehler, Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge, Extreme Mechanics Letters 67 (2024) 102131. doi:10.1016/j.eml.2024.102131.

[8] I. Ozkaya, Application of large language models to software engineering tasks: Opportunities, risks, and implications, IEEE Software 40 (2023) 4–8. doi:10.1109/MS.2023.3248401.

[9] X. Peng, Software development in the age of intelligence: embracing large language models with the right approach, Frontiers of Information Technology & Electronic Engineering 24 (2023) 1513–1519. doi:10.1631/FITEE.2300537.

[10] L. Hamilton, D. Elliott, A. Quick, S. Smith, V. Choplin, Exploring the use of ai in qualitative analysis: A comparative study of guaranteed income data, International Journal of Qualitative Methods 22 (2023). doi:10.1177/16094069231201504.

[11] Z. Rasheed, M. Waseem, A. Ahmad, K.-K. Kemell, W. Xiaofeng, A. N. Duc, P. Abrahamsson, Can large language models serve as data analysts? a multi-agent assisted approach for qualitative data analysis, arXiv preprint arXiv:2402.01386 (2024).

[12] D. L. Morgan, Exploring the use of artificial intelligence for qualitative data analysis: The case of

chatgpt, International journal of qualitative methods 22 (2023) 16094069231211248. doi:10.1177/160940692312112.

[13] L. A. Siiman, M. Rannastu-Avalos, J. Pöysä-Tarhonen, P. Häkkinen, M. Pedaste, Opportunities and challenges for ai-assisted qualitative data analysis: An example from collaborative problem-solving discourse data, in: International Conference on Innovative Technologies and Learning, volume 14099, Springer, 2023, pp. 87–96. doi:10.1007/978-3-031-40113-8_9.

[14] X. Hou, Y. Zhao, Y. Liu, Z. Yang, K. Wang, L. Li, X. Luo, D. Lo, J. Grundy, H. Wang, Large language models for software engineering: A systematic literature review, ACM Transactions on Software Engineering and Methodology 33 (2024) 1–79. URL: http://dx.doi.org/10.1145/3695988. doi:10.1145/3695988.

[15] H. Pearce, B. Ahmad, B. Tan, B. Dolan-Gavitt, R. Karri, Asleep at the keyboard? assessing the security of github copilot's code contributions, in: 43rd IEEE Symposium on Security and Privacy, IEEE, 2022, pp. 754–768. doi:10.1109/SP46214.2022.9833571.

[16] Y. Liu, T. Le-Cong, R. Widyasari, C. Tantithamthavorn, L. Li, X. D. Le, D. Lo, Refining chatgpt-generated code: Characterizing and mitigating code quality issues, ACM Trans. Softw. Eng. Methodol. 33 (2024) 116:1–116:26. doi:10.1145/3643674.

[17] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS Q. 28 (2004) 75–105.

[18] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al., Training compute-optimal large language models, arXiv preprint arXiv:2203.15556 (2022).

[19] L. Floridi, M. Chiriatti, GPT-3: its nature, scope, limits, and consequences, Minds Mach. 30 (2020) 681–694. doi:10.1007/S11023-020-09548-1.

[20] OpenAI, Chatgpt: Optimizing language models for dialogue, https://openai.com/chatgpt, 2023. [Accessed 18 February 2024].

[21] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, Decodingtrust: A comprehensive assessment of trustworthiness in GPT models, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, 2023.

[22] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, arXiv preprint arXiv:2211.09110 (2023).

[23] O. Lemon, Conversational ai for multi-agent communication in natural language: Research directions at the interaction lab, AI Communications 35 (2022) 295–308. doi:10.3233/aic-220147.

[24] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, I. Mordatch, Improving factuality and reasoning in language models through multiagent debate, arXiv preprint arXiv:2305.14325 (2023).

[25] L. Sun, Y. Huang, H. Wang, S. Wu, Q. Zhang, C. Gao, Y. Huang, W. Lyu, Y. Zhang, X. Li, et al., Trustllm: Trustworthiness in large language models, arXiv preprint arXiv:2401.05561 (2024).

[26] S. Hong, X. Zheng, J. P. Chen, Y. Cheng, C. Zhang, Z. Wang, S. K. S. Yau, Z. H. Lin, L. Zhou, C. Ran, L. Xiao, C. Wu, Metagpt: Meta programming for multi-agent collaborative framework, ArXiv abs/2308.00352 (2023).

[27] C. Qian, X. Cong, C. Yang, W. Chen, Y. Su, J. Xu, Z. Liu, M. Sun, Communicative agents for software development, arXiv preprint arXiv:2307.07924 (2023).

[28] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework, arXiv preprint arXiv:2308.08155 (2023).

[29] Y. Talebirad, A. Nadiri, Multi-agent collaboration: Harnessing the power of intelligent llm agents, arXiv preprint arXiv:2306.03314 (2023).

[30] E. Halme, M. Jantunen, V. Vakkuri, K. Kemell, P. Abrahamsson, Making ethics practical: User stories as a way of implementing ethical consideration in software engineering, Inf. Softw. Technol. 167 (2024) 107379. doi:10.1016/J.INFSOF.2023.107379.

[31] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nature Machine

Intelligence 1 (2019) 389–399. doi:10.1038/S42256-019-0088-2.
[32] M. Ryan, B. C. Stahl, Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, Journal of Information, Communication and Ethics in Society 19 (2021) 61–86. doi:10.1108/JICES-12-2019-0138.
[33] E. Commission, EU AI Act: first regulation on artificial intelligence, https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. [Accessed 01 April 2024].
[34] N. K. Corrêa, C. Galvão, J. W. Santos, C. D. Pino, E. P. Pinto, C. Barbosa, D. Massmann, R. Mambrini, L. Galvão, E. Terem, N. de Oliveira, Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance, Patterns 4 (2023) 100857. doi:10.1016/J.PATTER.2023.100857.
[35] S. Dai, A. Xiong, L. Ku, Llm-in-the-loop: Leveraging large language model for thematic analysis, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, 2023, pp. 9993–10001. doi:10.18653/V1/2023.FINDINGS-EMNLP.669.
[36] J. Drápal, H. Westermann, J. Savelka, Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies, IOS Press, 2023. URL: http://dx.doi.org/10.3233/FAIA230965. doi:10.3233/faia230965.
[37] M. Bano, R. Hoda, D. Zowghi, C. Treude, Large language models for qualitative research in software engineering: exploring opportunities and challenges, Autom. Softw. Eng. 31 (2024) 8. doi:10.1007/S10515-023-00407-8.
[38] V. Braun, V. Clarke, Using thematic analysis in psychology, Qualitative Research in Psychology 3 (2006) 77–101.
[39] S. De Paoli, Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach, Social Science Computer Review 42 (2023) 997–1019. URL: http://dx.doi.org/10.1177/08944393231220483. doi:10.1177/08944393231220483.
[40] J. He, C. Treude, D. Lo, Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead, ACM Transactions on Software Engineering and Methodology 34 (2025) 1–30. URL: http://dx.doi.org/10.1145/3712003. doi:10.1145/3712003.
[41] J. Liu, K. Wang, Y. Chen, X. Peng, Z. Chen, L. Zhang, Y. Lou, Large language model-based agents for software engineering: A survey, 2024. URL: https://arxiv.org/abs/2409.02977. arXiv:2409.02977.
[42] P. Runeson, M. Host, A. Rainer, B. Regnell, Case study research in software engineering, John Wiley & Sons, Nashville, TN, 2012.