

CodePori: Large-Scale System for Autonomous Software Development Using Multi-Agent Technology

Zeeshan Rasheed

`zeeshan.rasheed@tuni.fi`

Tampere University

Muhammad Waseem

Tampere University

Kai-Kristian Kemell

Tampere University

Aakash Ahmad

University of Derby

Malik Abdul Sami

Tampere University

Mika Saari

Tampere University

Jussi Rasku

Tampere University

Pekka Abrahamsson

Tampere University

Research Article

Keywords: Generative AI, Artificial intelligence, AI for SE, Large Language Model, AI Adoption, Industry Case Study

Posted Date: July 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9701113/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

CodePori: Large-Scale System for Autonomous Software Development Using Multi-Agent Technology

Zeeshan Rasheed^{1*}, Muhammad Waseem¹, Kai-Kristian Kemell¹,
Aakash Ahmad¹, Malik Abdul Sami¹, Mika Saari¹, Jussi Rasku¹,
Pekka Abrahamsson¹

¹Faculty of Information Technology and Communication Science,
Tampere University, Tampere, Finland.

²Data Science Research Center, School of Computing, University of
Derby, UK, Derby, UK.

*Corresponding author(s). E-mail(s): zeeshan.rasheed@tuni.fi;

Contributing authors: muhammad.waseem@tuni.fi;

kai-kristian.kemell@tuni.fi; a.abbasi@derby.ac.uk; malik.sami@tuni.fi;
mika.saari@tuni.fi; jussi.rasku@tuni.fi; pekka.abrahamsson@tuni.fi;

Abstract

Context: Existing Large Language Model (LLM)-based multi-agent systems are capable of executing tasks and providing data-driven recommendations, thereby enabling automation and decision support that assist practitioners in software development. However, existing studies have tested the outcomes of agents on benchmark datasets, offering only binary pass-or-fail results, which provide limited insight into their practical applicability. There remains a lack of empirical research examining the potential and limitations of LLM-based agents in addressing challenging, real-world tasks such as automated code generation for large-scale systems.

Objective: This study empirically investigates the potential of LLM-based agents in software development by engaging participants to evaluate the agents performance in autonomous software development tasks.

Method: We employed a two-phase approach comprising (i) the development of a multi-agent system, CodePori, to automate code generation, and (ii) participant-based evaluation to assess agent performance and explore their

practical applicability for software development.

Results: The results of this study present participants’ feedback and insights into the use of LLM-based multi-agent systems in real-world software development, including their strengths, challenges, and areas for improvement. We also highlight key aspects missed by code-generation benchmarks, which collectively enhance our understanding of the practical applicability of LLM-based agents.

Conclusions: Based on the results, we conclude that, while LLM-based multi-agent systems show potential for large-scale software development, their successful integration requires the addressing of specific challenges (e.g., short-term memory limitations, hallucinations, and code smells) and incorporating a practitioner-centric perspective. The study highlights the need to move beyond standard benchmarks to evaluate real-world applicability and identifies new opportunities for broader adoption in both industry and academia.

Keywords: Generative AI, Artificial intelligence, AI for SE, Large Language Model, AI Adoption, Industry Case Study

1 Introduction

Software Engineering (SE) as a discipline supports the application of structured principles and practices to the development, implementation, and maintenance of software applications and systems that deliver functional solutions to meet specific user needs and business requirements [1]. Traditional SE processes rely heavily on human effort for critical tasks such as requirements gathering, system design, coding, testing, and quality assurance [2]. As depicted in Figure 1, this human-centric approach has historically been a defining feature of conventional software development practices.

However, the recent integration of Artificial Intelligence (AI) into SE has led to significant advancements and opened up new possibilities for synthesizing software processes and practices with data/intelligence-driven tools and techniques to engineer software systems [3]. These advancements have driven the development of Natural Language Processing (NLP) models such as GPT, BERT, T5, RoBERTa, and LLaMA [4–6], which are now commonly referred to as Large Language Models (LLMs) and serve as the foundation for a wide range of AI applications. Numerous studies have explored the integration of LLMs into the SE process and have demonstrated remarkable capabilities in handling various downstream tasks in SE, such as code generation, testing, and documentation [7]. However, utilizing LLMs in SE raises many challenges, such as hallucinations, a limited context window, and a lack of domain-specific expertise [8].

In recent years, a growing number of studies have explored the role and impact of LLM-based agents in automating SE processes [9]. As shown in Figure 1, modern SE processes follow an agent-centric approach, where agents play an important role in the software development process [10]. For instance, the development of agent-based systems such as MetaGPT [11], ChatDev [12], AgentCoder [13], MapCoder

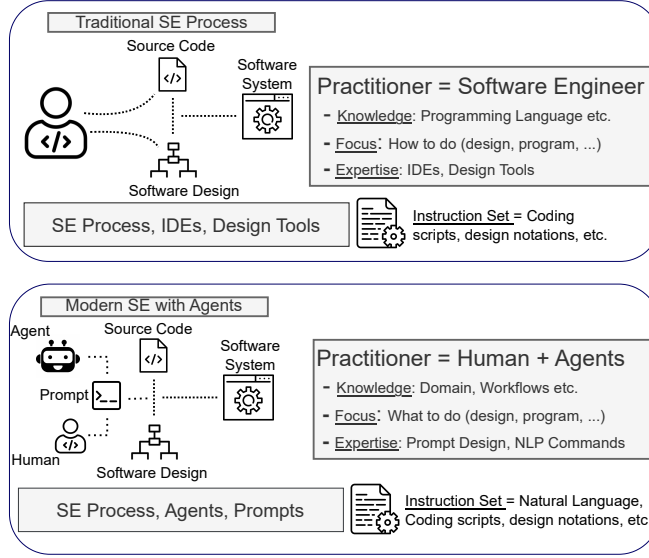


Fig. 1: Traditional (human-centric) vs modern (human-agent collaborative) SE processes.

[14], SWE-Agent [15], and CodeSim [16] represents a step forward in automating the software development process. These systems utilize a collaborative, multi-agent approach where each agent is a prompt-based component with a specific persona and a defined set of instructions [17]. These frameworks allow large-scale tasks to be decomposed into smaller, manageable sub-tasks, thereby potentially mitigating issues such as hallucination and limited memory by distributing the cognitive load and specializing each agent's function [18]. According to Levy *et al.* [19], a large-scale system includes advanced architectural design, a multi-file modular structure, and intricate inter-component communication. Despite the steady growth in published research on LLM-based agents as AI assistants in software development, empirical evidence on their effectiveness in practical development settings remains limited. In particular, most existing studies rely on benchmark-driven or automated evaluations, with little emphasis on practitioner-centric assessments that capture developers' perspectives on the capabilities and limitations of these agents in real-world software development tasks.

Motivation and Contribution: Our study is part of the BF/Amalia-SW project (2023–2025) funded by Business Finland. In collaboration with Finnish companies, this study aims to advance the SE field by utilizing LLM-based agents and evaluating their capabilities in real-world scenarios. This research addresses an industry demand to understand the practical applicability of agents for large-scale software development, including their strengths, challenges, and areas for improvement.

In recent years, agentic frameworks such as AgentCoder [13], MetaGPT [11], and SWE-agent [15] have shown promise in automating key aspects of software development workflows. These frameworks have shown high accuracy on benchmarks like HumanEval [20], MBPP [21], APPS [22], and SWE-bench [23]. Despite promising results on benchmarks, these systems often fail to scale to realistic development scenarios, frequently producing inaccurate or contextually inappropriate code in large-scale systems [8]. This limitation arises from the narrow scope of existing benchmarks, which focus on simplified environments and fail to capture repository-level challenges such as reasoning across multiple files and interdependent functions [24]. The lack of real-world evaluation benchmarks has created a critical research gap, motivating the need to design benchmarks that can assess and advance the practical applicability of LLM-based multi-agent systems in SE [25].

To bridge this gap, we conducted an empirical study to understand the practical applicability of LLM-based agents for large-scale software development. This study involved two main steps, as shown in Figure 2: (i) the development of a multi-agent system, CodePori, that autonomously generates code for large-scale systems from high-level descriptions, and (ii) the evaluation of the performance, strengths, and limitations of the agents through the findings of a survey with 118 participants.

Contribution: Our key contributions can be summarized as follows:

- We have designed and developed a multi-agent system, CodePori, which autonomously generates code for large-scale systems from high-level descriptions.
- We provide empirical insights into multi-agent systems for large-scale code generation based on feedback from 118 participants.
- We present a comprehensive analysis of the multi-agent system’s performance, including its strengths, challenges, and areas for improvement, providing researchers and practitioners with valuable evidence to guide future research on agent-based systems in SE.
- To support replication, validation, and broader exploration of the findings of our study, we have publicly released the survey dataset [26] and the proposed multi-agent system via GitHub [27] to aid further research and implementation.

Structure: The rest of the paper is organized as follows. An overview of the background study and related work is presented in Sections 2 and 3. The study methodology is described in Section 4. The results of the study are presented in Section 5, followed by a discussion in Section 6, and threats to validity are addressed in Section 7. Finally, the study is concluded with recommendations for future work in Section 8.

2 Background

This section presents the background study, outlining the foundational concepts and prior research relevant to this work. Section 2.1 provides an overview of previous studies concerning LLMs and SE.

2.1 Large Language Models in Software Engineering

Over recent years, LLMs have shown significant growth and potential in various SE applications [28]. These models have been trained on large code repositories, enabling them to generate code or entire programs based on an input description [29]. According to Hou *et al.* [30], the LLM and SE are interconnected by applying NLP techniques to different tasks within the software development life cycle. LLMs provide valuable support to SE processes through their language generation capabilities, enabling improvements in tasks such as documentation, code generation, and debugging [31], [32]. Nowadays, LLMs have been applied to different fields of SE, including data analysis [33], text classification [34], code generation [35], code search [36], automated program repair [37], unit test case generation [38], etc. [39].

Some existing review studies have examined SE literature related to LLMs, e.g., [40], [41], [42]. Hou *et al.* [40] analyzed 395 research articles. Their review aims to provide a comprehensive understanding of how LLMs can be used to optimize SE processes and outcomes. The study addresses key research questions by categorizing different LLMs used in SE tasks, analyzing data collection methods, and exploring their applications. Zheng *et al.* [41] investigated the potential of LLMs to improve current SE tasks and categorized these tasks into seven types. For each type, they presented application examples of LLMs, highlighting their strengths and limitations to facilitate researchers in recognizing and addressing potential challenges in applying LLMs to SE tasks. Zheng *et al.* [42] examined the effectiveness and significance of LLMs in the context of SE. They reviewed and evaluated 134 studies focused on code LLMs and provided a detailed assessment of their capabilities across different SE tasks.

3 Related Work

This section presents the most relevant existing research to contextualize the scope and justify the contribution of the proposed study. Sections 3.1 and 3.2 review prior works that have utilized LLMs and LLM-based agents for code generation.

3.1 Code Generation with Large Language Models

Nowadays, LLMs have shown impressive effectiveness across different SE domains. As shown in Table 1, several studies have investigated the potential of LLMs in facilitating automatic programming. For example, AlphaCode (Li *et al.* [37]) achieves human-level performance in accurate code generation during actual programming contests, while Codex (Chen *et al.* [20]) introduced capabilities that later enabled GitHub Copilot to provide instant coding recommendations. Wang *et al.* [51] introduced CodeT5, a cohesive pre-trained encoder-decoder transformer model, using a unified approach to effortlessly facilitate both code comprehension and generation tasks, while enabling multi-task learning. Other open-source code generation models include GPTNeo (Black *et al.* [45]), GPT-J (Wang *et al.* [52]), CodeParrot (Tunstall *et al.* [53]), PolyCoder (Xu *et al.* [44]), CODEGEN (Nijkamp *et al.* [48]), INCODER (Fried *et al.* [49]), and (Rasheed *et al.* [54]). Chen *et al.* [55] leverage the Codex inference API from OpenAI, along with the two robust open-source models, CODEGEN

Table 1: Large pre-trained language models for code automation

S No	Papers	Parameters	Language	Size	Benchmark	Reference
01	Codex	12B	Python	Code: 159GB	HumanEval, APPS	Chen <i>et al.</i> [20]
02	AlphaCode	41B	12 Langs	Code: 715.1GB	HumanEval, APPS Code Contest	Li <i>et al.</i> [37]
03	PaLM-Coder	8B	Multiple	Text: 741 B Tokens Code: 39GB	HumanEval, MBPP TransCoder, DeepFix	Chowdhery <i>et al.</i> [43]
04	PolyCoder	2.7B	12 Langs	Code: 253.6GB	HumanEval	Xu <i>et al.</i> [44]
05	GPT-Neo	1.3B	Multiple	Text: 730GB Code: 96GB	HumanEval	Black <i>et al.</i> [45]
06	GPT-NeoX	20B	Multiple	Text: 730GB Code: 95GB	HumanEval	Black <i>et al.</i> [46]
07	GPT-J	6B	Multiple	Text: 730GB Code: 96GB	HumanEval	Arora <i>et al.</i> [47]
08	CodeGen-Multi	6.1B	6 Langs	Code: 150B Tokens Text: 355B Tokens	HumanEval, MBPP	Nijkamp <i>et al.</i> [48]
09	InCoder	6.1B	28 Langs	Code: 159GB	HumanEval, MBPP CodeXGLUE	Fried <i>et al.</i> [49]
10	CodeGeeX	13B	23 Langs	Code: 15B Tokens	HumanEval, MBPP, XLCOST HumanEval-X, CodeXGLUE	Zheng <i>et al.</i> [50]
11	CodeGen-Mono	6.1B	Python	Code: 150B tokens Text: 355B Tokens	HumanEval, MTPB	Nijkamp <i>et al.</i> [48]

and INCODER, for zero-shot code generation. Zheng *et al.* [50] presented CodeGeeX, a 13-billion-parameter multilingual model designed for code generation. As of June 2022, CodeGeeX has been pre-trained on a massive dataset comprising 850 billion tokens from 23 different programming languages.

3.2 Code Generation with LLM-Based Multi-Agents

In 2023, Qian *et al.* [12] introduced the ChatDev model, which utilizes LLM-based agents throughout the entire software development life cycle. This was the first model to use LLM-based agents for software development. Hong *et al.* [11] introduced MetaGPT, an innovative meta-programming framework that incorporates efficient human workflows into LLM-based multi-agent collaborations. This framework achieved 89% accuracy on HumanEval and 87.7% on the MBPP benchmark [11]. MetaGPT uses an assembly line paradigm to assign diverse roles to various agents, efficiently breaking down complex tasks into sub-tasks. Following MetaGPT, Huang *et al.* [13] introduced AgentCoder, a multi-agent framework designed to improve the quality and efficiency of code generation by simulating a collaborative team of SE. AgentCoder achieved 96.3% on HumanEval and 91.8% on the MBPP benchmark, demonstrating higher accuracy compared to both MetaGPT and ChatDev. In 2024, Islam *et al.* [14] proposed MAP Coder, an agent-based system focused on improving performance by breaking down large problems into smaller, manageable sub-tasks and using a structured communication protocol between agents to ensure a coherent final solution. This framework achieved 93.9% on HumanEval and 83.1% on the MBPP benchmark, showing an improvement over MetaGPT, especially on the HumanEval benchmark. However, AgentCoder’s performance of 96.3% and 91.8% on HumanEval and MBPP, respectively, still shows higher accuracy compared to MetaGPT, ChatDev, and MAP Coder. Most recently, SWE-Agent [15] is an agent-based framework designed to solve software issues from GitHub repositories. It uses a specialized agent that can autonomously execute code, navigate repositories, and apply fixes, demonstrating high accuracy compared to the models mentioned above.

However, as we have observed, while current agent-based systems show high accuracy on benchmarks, their performance on real-world problems is often low [56]. As noted by Yadav *et al.* [57], these benchmarks contain a limited set of projects and do not represent the full range of real-world programming challenges. Similarly, Dai *et al.* [58] pointed out that the programming tasks in the HumanEval dataset are designed for small projects. In particular, benchmark accuracy primarily reflects isolated task performance and does not capture system-level challenges such as agent capabilities, coordination among agents, and robustness across multi-step development workflows. As a result, strong benchmark performance does not necessarily translate into effective system behavior in realistic software development settings [59]. This represents a notable gap in evaluating systems intended to generate code for large-scale systems. To address this gap, our study extends existing work by involving participants to examine the effectiveness and practical limitations of multi-agent systems that are not captured by traditional benchmarks.

4 Research Method

This section presents the methodology used to conduct this study. Section 4.1 states the research questions, Section 4.2 describes the multi-agent workflow and roles, and Section 4.3 outlines the evaluation framework. The overall research method is illustrated in Figure 2.

4.1 Research Questions

Based on the study goal, we formulated the following two Research Questions (RQs). RQ1 addresses autonomous code generation, and RQ2 addresses solution evaluation.

RQ1. *Can the multi-agent system autonomously generate code, and is it capable of handling the generation of large-scale systems?*

Objective. The main goal of RQ1 is to determine if the CodePori system can perform autonomous code generation for large-scale systems without human intervention.

RQ2. *What are the perceived effectiveness and practical limitations of CodePori for large-scale code generation, as experienced by participants?*

Objective. The objective of RQ2 is to examine the effectiveness and practical limitations of the CodePori system in autonomous code generation for large-scale systems, achieved by analyzing participants’ feedback on the system’s performance in real-world scenarios.

Table 2: Main tasks of multi-agent system

No	Agent	Main Task	Prompt Content
1	Manager agent	Task segmentation	Take project description and segment these tasks into smaller and structured chunks
2	Architecture agent	Generate architecture overview	Provide a detailed overview of the project architecture tailored to the specified LANGUAGE
3	Flow structure agent	Generate folder structure	Generate folder structure based on project requirements
4	Dev agent	Initial code generation	Generate initial code based on given description and folder architecture
5	Verification agent	Final verification of code	Focus on identifying and rectifying any flaws in code that the above-mentioned agent might have missed
6	Finalization agent	Review and finalize code quality	Suggest an iterative review process to enhance code quality and finalize code

4.2 Multi-Agent System Design

CodePori is a multi-agent system designed to automate the code generation process based on given project requirements for large-scale software systems, while demonstrating the effectiveness of multi-agent collaboration in software development. In this context, a large-scale system refers to software systems characterized by advanced architectural design, multi-file modular structures, and intricate inter-component communication [19].

In Section 4.2.1, we present the architecture of the multi-agent system that works collaboratively to perform coding tasks autonomously based on the provided project description.

4.2.1 Multi-Agent Architecture

As shown in Figure 2, we designed six LLM-based agents that collaborate to autonomously perform coding tasks based on the project description. We evaluated several alternatives and found that a six-agent system produced better overall results. For instance: (i) a single LLM agent handling both development and refinement produced lower-quality outputs, as errors were more likely to persist without a secondary review; (ii) a three-agent system comprising one agent for development, one for review, and one for finalization was also tested but resulted in slower performance due to an overly compressed refinement process. Ultimately, the six-agent system achieved the best balance between quality and performance by mirroring a real-world software development pipeline that separates responsibilities into various phases, ensuring a structured workflow that mimics human development teams and leads to more accurate and maintainable code generation.

The multi-agent system, as illustrated in Table 2, outlines each agent’s profile, including its name, role, goal, and prompt content. In this study, we utilized instruction-based prompts following established prompting guidelines [5], which involve identifying the role, explaining the task, and defining the output format for each agent. Each agent specializes in specific tasks and interacts with others through APIs, facilitating continuous communication and refinement of the generated code. Additionally, we instructed each agent to iterate with the others at least three times to

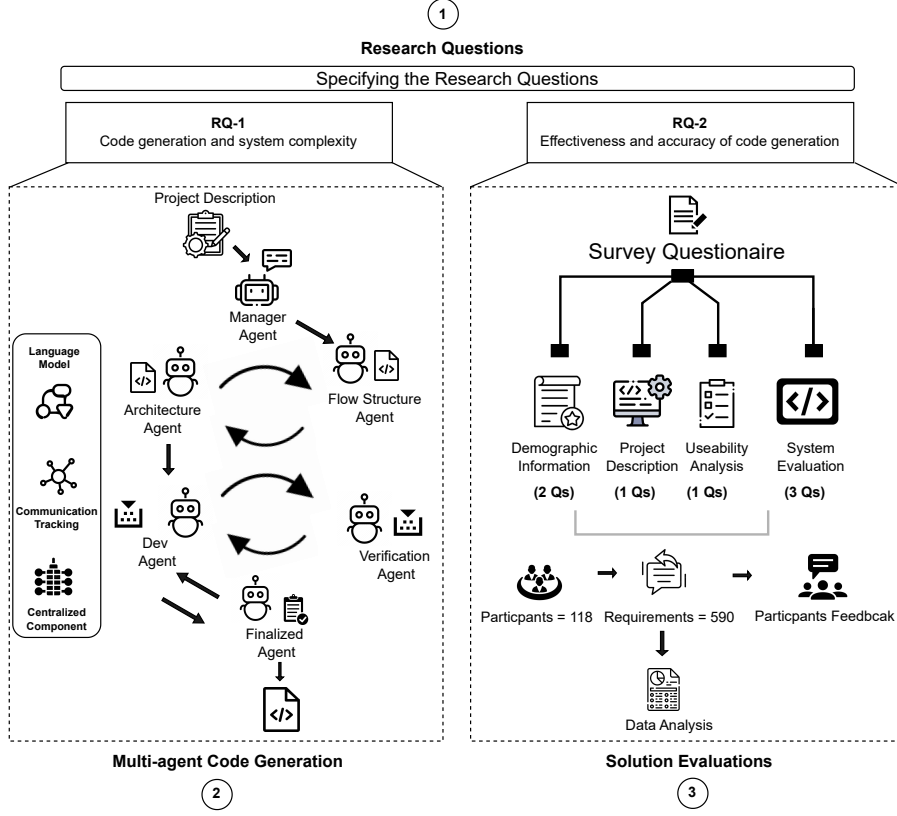


Fig. 2: Multi-agent system and survey workflow.

refine and update the code based on feedback from previous agents, ensuring a thorough and collaborative development process. In the following, we provide a detailed explanation of each agent.

The **Manager agent** plays an important role in the autonomous code generation process by serving as a project manager responsible for analyzing the high-level project description. The *Manager agent* is tasked with breaking down the project into structured components. This is achieved by processing the provided project description and identifying the necessary modules for development. The *Manager agent* ensures a structured decomposition by requiring each module to be described in a standardized format, including attributes such as module name, detailed description, objectives, expected inputs, expected outputs, dependencies, additional notes, and best practices. The *Manager agent* prompt is designed to extract well-defined, JSON-formatted module descriptions, which serve as foundational inputs for further development steps. Appendix 6a provides the prompt for the *Manager agent*.

The **Architecture agent** plays an important role in the autonomous code generation pipeline by transforming modular specifications into a comprehensive architectural overview. Once the *Manager agent* has identified and structured the modules from the high-level project description, the *Architecture agent* then designs the system’s architecture based on the specified programming language. Its primary task is to provide a detailed, standardized architectural overview that captures how different components interact, manage data flow, and connect with the underlying databases. This agent considers language-specific best practices; for instance, separating client and server logic in JavaScript projects or following MVC principles in Python-based Flask applications. Overall, the *Architecture agent* ensures that the resulting architecture is clean, scalable, and aligned with the intended development language. The *Architecture agent* prompt can be found in Appendix 6b.

The **Flow structure agent** transforms high-level architectural planning into a structured file and folder hierarchy aligned with the selected programming language and tech stack. It interprets the project’s modular descriptions and organizes them into logically grouped directories, ensuring that configuration files, dependency managers, and code components are properly structured. The *Flow structure agent* outputs a plain-text layout that includes key elements such as `main.py`, `config.py`, and subdirectories. By converting abstract architecture into a concrete file structure, the *Flow structure agent* enables the other agents to generate and place code systematically. The prompt for the *Flow structure agent* is provided in Appendix 7a.

The **Dev agent** is responsible for generating initial code for each module or file identified by the *Flow structure agent*. The *Dev agent* collaborates with the *Verification agent* to iteratively refine and develop the software components while ensuring clean code. The *Dev agent* ensures that each module is developed without placeholders or TODOs, aligns with previously completed modules, and follows Google-style docstrings for maintainability. Additionally, each module contains at least 150-200 lines of code. To ensure consistency and quality, the *Dev agent* ensures that the code follows predefined coding guidelines and industry best practices, including optimized algorithm selection, structured exception handling, and efficient memory management. The agent also integrates logging mechanisms to facilitate debugging and provide transparency in decision-making, further enhancing the quality and maintainability of the code. The development process is structured as a multi-step iterative loop between the *Dev agent* and the *Verification agent*. This iterative process refines each module and ensures production-level quality before proceeding to finalization. The prompt for *Dev agent* is provided in Appendix 7b.

The **Verification agent** is responsible for conducting a detailed code analysis to ensure that the current module code aligns with the project description and that all modules interact correctly. This agent does not modify the code directly, but instead sends it back to the *Dev agent*, accompanied by clear instructions on the necessary improvements to enhance the system’s functionality. The *Verification agent* plays a strict role in code review, ensuring the highest code quality. Appendix 8a contains the prompt for the *Verification agent*.

The **Finalization agent** specializes in the refinement and enhancement of the current module code to ensure that it is production-ready. The process starts with

receiving the project description, file and folder structure, generated code, and a review report detailing the necessary improvements. The *Finalization agent* first analyzes the existing implementation, understands the required modifications, and then collaborates iteratively with the other agents to enhance the code quality. Each response must include a fully executable update, ensuring that the code can be used directly without further modifications. By enforcing these standards, the *Finalization agent* ensures that the final code maintains high integrity, follows industry best practices, and aligns with the overall architectural requirements of the project. The prompt for the *Finalization agent* is provided in Appendix 8b.

4.2.2 Core Components

The proposed system consists of the following core components:

- **Language Model:** In this project, we used the Claude-3.5 Sonnet model for code generation. We chose this model because it outperforms previous state-of-the-art models on the HumanEval benchmark, achieving 92% accuracy. Additionally, Claude-3.5 Sonnet provides enhanced coding and reasoning capabilities [60]. We utilized an API key to connect to the system, with API requests being made during each iteration of the discussions among the six agents to generate code collaboratively.
- **Agent Communication Tracking:** We developed a structured environment, which includes a specialized logger to track all agents' communication activities. This initial configuration ensures that all operations, from initial project design to final code generation, are performed within a controlled and traceable environment.
- **Centralized Component:** To facilitate coordination and data persistence across agents, a state manager was created. This centralized component maintains critical state information such as the project description, architectural overview, folder structure, all generated code files, and verification reviews. The state manager ensures that each agent has access to the necessary context and accumulated results from previous steps, thereby enabling smooth collaboration between agents and minimizing redundant processing.
- **Agent Orchestration Pipeline:** The pipeline is orchestrated through a central function called `generate_project_stream()`, which manages the end-to-end process. This function first calls the *Architecture agent* and the *Flow structure agent* to obtain the architectural blueprint and folder layout. It then iterates over each file path and prompts the *Dev agent* to generate the file's code. If a verification step raises issues, the *Finalization agent* is invoked to address them. Upon completion of codebase generation, all files are packaged into a downloadable ZIP archive, and a link is returned to the user. This procedural flow ensures that the software project is developed in a modular and verifiable manner, utilizing the strengths of LLM-based agents at each critical stage of the pipeline.

4.3 Survey Design

Concerning the research goals, we decided to conduct a survey as the primary data collection method to understand participants' perceptions of the proposed multi-agent system for code generation. In the evaluation phase of our methodology, we

asked participants to evaluate the effectiveness, practical limitations, and real-world applicability of the proposed system.

4.3.1 Survey Questionnaire

We created the survey in English using Microsoft Forms, consisting of both open- and closed-ended questions. As depicted in Table 3, the first two questions were about the participants’ background, covering their education and experience in software development and LLMs. This information was intended to help us assess the participants’ eligibility for the survey.

In the second part, we designed five questions (Q3 to Q7) aimed at understanding their perspectives on the use of multi-agent systems for large-scale code generation.

Table 3: Survey questions

ID	Questions
SQ1	What is the highest level of education you’ve completed?
SQ2	How much experience do you have in software development and LLM?
SQ3	Write the five projects you provided for code generation.
SQ4	How long does the system take to generate code for a single project description? Please write the time taken for each project.
SQ5	Is it required to modify the code generated by the AI agent system? If yes, what modifications did you make to execute the code? Please write the modification you made for each project.
SQ6	How do you rate the proposed AI agent system for code generation?
SQ7	Are you satisfied with the performance of the proposed system? Please elaborate on its strengths and areas that need improvement.

4.3.2 Participant Recruitment

To evaluate the effectiveness, practical limitations, and real-world applicability of the multi-agent system, we engaged 118 students enrolled in the “Fine-Tuning Large Language Models” course held at Tampere University in 2025. Aside from practical assignments where the students fine-tuned open source LLMs, the course included teaching on other topics related to LLMs and SE, such as LLM-based AI agents.

In the final lecture of this course, we introduced our system to these 118 students. We first provided a detailed introductory explanation of the proposed system, and then demonstrated its functionality in practice for the students. This was meant to ensure that the students had an understanding of the system’s functionality before participating in the evaluation. We then shared the survey link together with the GitHub repository of the proposed system, along with detailed instructions on how to use the system.

We acknowledge that the use of student subjects for the evaluation poses some threats to the validity of our study. We discuss these in detail in Section 7. However, we do note that, having taken this course before participating in the evaluation, these students had verifiable prior experience with LLMs through the course itself.

4.3.3 Analyzing Survey Data

Descriptive statistics [61] were used to perform a quantitative analysis of the closed-ended questions. Moreover, we applied the open-coding method [62] to analyze the open-ended response data. The first author conducted the open coding and then used selective coding to identify the core categories generated by aggregating a set of concepts. To reduce personal bias during data analysis, the second, third, and fourth authors participated in validating the generated codes.

5 Results

This section presents the study results. Section 5.1 describes the implementation and demonstration of CodePori. Section 5.2 reports on the survey findings from the 118 participants who evaluated the system.

5.1 CodePori: LLM-Based Multi-Agent System (RQ1)

The proposed multi-agent system, CodePori, is designed for autonomous large-scale code generation. In this context, autonomous refers to code generation without step-by-step human prompting, rather than fully unattended deployment. In this project, we utilized six specialized agents that collaborate to transform high-level requirements into executable code. Each agent performs a specific role and communicates with the others to complete tasks ranging from requirement analysis to finalized implementation.

First, the *Manager agent* analyzes the given requirements and breaks them into structured components, including module names, detailed descriptions, objectives, expected inputs and outputs, dependencies, additional notes, and recommended best practices. Next, the *Architecture agent* generates a detailed, standardized architectural overview that defines component interactions, data flow management, and connections with underlying databases. The *Flow structure agent* then translates the high-level architecture into a structured file and folder hierarchy aligned with the chosen programming language and technology stack. The *Dev agent* then begins writing the initial code for each module or file defined by the *Flow structure agent*. The *Verification agent* performs a comprehensive analysis of the generated code to ensure that each module aligns with the project description and that inter-module interactions function correctly. Finally, the *Finalization agent* reviews the overall implementation, identifies required refinements, and iteratively collaborates with the other agents to improve code quality and ensure the output is production-ready without major modifications. More details about the agents' workflow can be found in Section 4.2.

Figure 3 illustrates how the six agents collaborate to transform high-level requirements into a fully developed software application. As shown in Figure 3, the detailed project description is first provided as input. The initial agent then interprets these requirements and generates five modules, each accompanied by a detailed explanation of its functionality. Then two agents create a detail architectural overview and establish a structured file and folder hierarchy for the project. Once the architecture

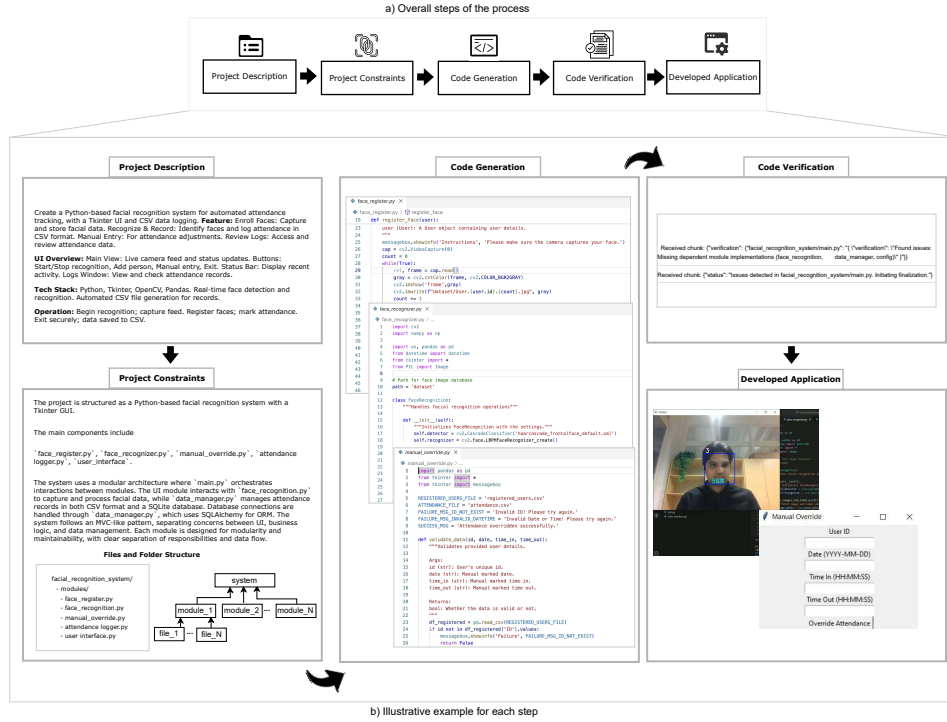


Fig. 3: Detailed representation of the software system developed by CodePori.

is defined, the system proceeds with the initial code generation for the identified modules. The generated code is then verified by the verification agent, which refines and improves it to ensure consistency and correctness. Finally, the developed application is produced as the end result.

In this project, certain manual modifications were necessary to refine the generated code and ensure accurate execution. For instance, it was important to install the required libraries, update file paths for CSV or other necessary files, and, when applicable, load the appropriate APIs or pre-trained models. In the example shown in Figure 3, we created two CSV files within the same project directory and installed the haarcascade frontalface default.xml file from GitHub, which was also stored in the same directory. This file is used for detecting frontal faces during the facial recognition process.

Our initial findings indicate that the proposed multi-agent system is capable of generating code for a large-scale system. However, we believe that although some challenges remain in fully automating SE, the results highlight the practical opportunities and limitations of applying multi-agent approaches beyond isolated function-level generation.

5.2 Survey Results (RQ2)

We collected feedback from 118 participants to evaluate the proposed multi-agent system for autonomous code generation for large-scale systems. In total, 590 project descriptions submitted by the participants were analyzed and classified to assess how the system handled different types of project inputs. The term project description refers to the textual input provided to the system, based on which the proposed system autonomously generates executable code. The analysis also included measuring the average completion times across project categories to assess system efficiency. Furthermore, we evaluated the overall performance of the system by identifying its key strengths, existing challenges, and areas for improvement based on participant feedback, focusing on the effectiveness of multi-agent collaboration in large-scale code generation tasks.

5.2.1 Demography of Participants (SQ1-SQ2)

The demographic details given in Figure 4 provide information on the background of the participants, such as their experience in the specific area and their highest level of education. Below, we outline the demographic information to provide a clearer picture of the participants’ characteristics.

- **Participants’ experience:** As shown in Figure 4, 82 out of 118 (69.49%) of the participants have around 2–5 years of experience in AI-based software development, 31 out of 118 (26.28%) have approximately 2 years of experience, and 5 out of 118 (4.23%) have around 10 years of experience in this domain.
- **Qualifications:** As shown in Figure 4, the majority of participants, i.e., 56 out of 118 (47.46%) held a bachelor’s degree, followed by 33 out of 118 (27.97%) with a master’s degree, 26 out of 118 (22.03%) with a high school diploma, and only 3 out of 118 (2.54%) with a doctoral degree.

5.2.2 Project Description (SQ3)

In this survey, each participant was asked to provide five different project descriptions as input. In total, 590 project descriptions were provided to test the system’s capabilities. We first analyzed and classified the descriptions to understand how the multi-agent system responds to varying levels of detail and project scope. Our analysis also examined how the system performs when given vague project descriptions versus highly detailed and specific requirements. **Project classifications:** Initially, we classified 590 project descriptions. As shown in Table 4, these descriptions were grouped into 11 different categories: Games, AI/ML, Finance, 3D/2D Visualizations, Chatbots, Benchmarking, Mathematical, Image Processing, E-commerce, Medical, and Others.

As shown in Table 4, 149/590 (25.25%) of the project descriptions relate to game development. A further 48/590 (8.14%) focus on AI/ML, 50/590 (8.47%) address mathematical tasks, and 24/590 (4.07%) concern finance. Additionally, 7/590 (1.19%) involve image processing, 7/590 (1.19%) relate to e-commerce, and 26/590 (4.41%) focus on chatbots. Projects covering 2D/3D visualization account for 6/590 (1.02%), medical topics 6/590 (1.02%), and benchmarking 1/590 (0.17%). Finally, 266/590

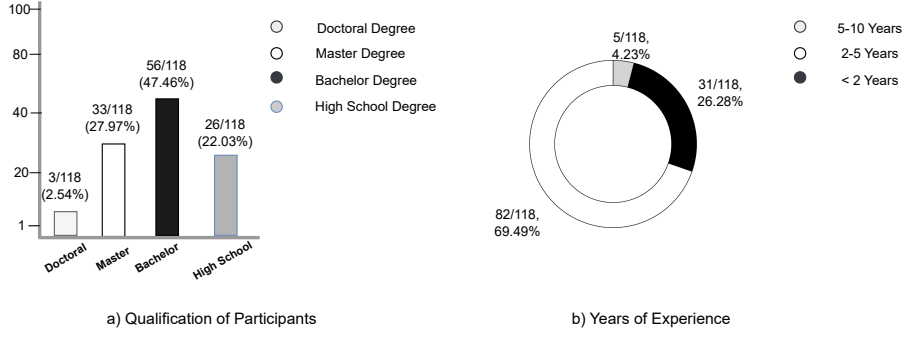


Fig. 4: Demographic data showing participants' qualifications and experience.

(45.08%) fall into the Others category (e.g., weather apps, password generators, food recipe apps, and to-do lists).

Table 4: Classification of project descriptions

Category	# of Projects	Percentage (%)	Brief Description
Games	149	25.25%	Projects involving game development
AI/ML	48	8.14%	Projects focused on AI and specifically on ML/DL
Mathematical	50	8.47%	Projects focused on mathematical computations
Finance	24	4.07%	Projects related to financial calculations
Image Processing	7	1.19%	Projects that analyze or perform operations on digital images
E-commerce	7	1.19%	Projects related to online shopping platforms or features
Chatbots	26	4.41%	Projects involving conversational AI or automated chat systems
2D/3D Visualization	6	1.02%	Projects for creating visual representations in two or three dimensions
Medical	6	1.02%	Projects related to healthcare, medical data, or applications
Benchmarking	1	0.17%	Projects designed to test or compare system performance
Others	266	45.08%	Projects involve a weather app, password generator, URL shortener, food recipe app, and to-do list
Total Projects	590	100%	

5.2.3 Overall Performance (SQ6-SQ7)

Figure 5 presents the overall performance of the multi-agent system based on participant feedback. We define overall performance as the system's speed (time to generate/complete code) and accuracy (functional correctness and absence of defects) in code generation. In this study, 30.51% of participants rated the system poor, 20.33%

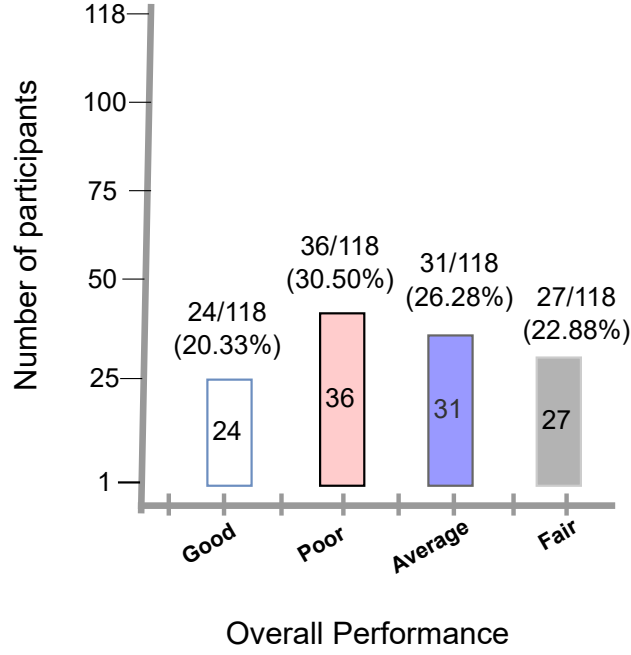


Fig. 5: Overall performance.

rated it good, and the remaining 49.16% rated it average or fair (26.28% and 22.88%, respectively).

Strengths (SQ7)

Participants highlighted several key strengths in the proposed system for autonomous code generation. These strengths are grouped into seven categories: fast code generation, a well-structured file and folder architecture, a clear understanding of user requirements, multi-language support, a robust modular design, and suitability for generating machine learning code. A detailed discussion of these identified strengths is given below.

Fast code generation: To evaluate the speed of the proposed system, we performed a descriptive analysis on SQ4, examining the average completion time across 590 projects. The results indicate that the average completion time with the system was approximately 9 minutes 14 seconds, based on a total of 5,391 minutes across 590 data points. This suggests that, on average, the system enables relatively quick project turnovers. Additionally, 10 out of 118 participants (8.47%) highlighted that

the proposed system reduces development time by handling various aspects of coding. One representative quotation is provided below.

✚ “*This AI agent code generation tool has shown strong capabilities in automating the development process by generating structured and functional code. It can effectively reduce development time by handling various aspects of coding, including routing, database integration and user interface components, even for complex project descriptions.*” (Participant 65):

Better requirement interpretation refers to the system’s ability to accurately understand, process, and translate user-provided requirements into corresponding code structures and functionalities [63]. In this study, 3 out of 118 participants (2.54%) highlighted that the proposed system clearly understands the provided requirements and effectively translates them into well-structured and functional code outputs. An example quotation is included below for reference.

✚ “*I’m very satisfied as personally I wouldn’t have achieved structure and professionalism in code generation without these tools. The strength of the system is that it very clearly understands what is asked and what the user wants to be done.*” (Participant 46):

Code comment and documentation are important aspects of the software development process, as they enhance code readability, maintainability, and collaboration among developers. In this study, 2 out of 118 participants (1.69%) highlighted that the proposed system effectively generates well-structured and concise code comments, as well as supporting documentation. This indicates that the system not only produces functional code but also facilitates understanding and future use by providing explanatory details at both the code block and function levels. The following is a representative example of a participant’s statement.

✚ “*Yes, I am satisfied with the performance of the proposed system. Its first strength is that it gave well-written and concise comments, both the code block comment and the function comment, also the program documentation.*” (Participant 101):

Efficient and ease of use: In this study, 12 out of 118 participants (10.17%) emphasized that the proposed system was efficient in automating setup tasks, managing dependencies, and organizing the project structure. Participant feedback highlights that the system accelerates the development process and improves ease of use by simplifying the initial setup and ensuring a clean, well-structured workflow. Below, we provide a quotation that exemplifies this finding.

👉 “Yeah, it actually works way better than I expected! The way it automatically sets up the whole directory and installs all the required dependencies is super convenient. It saves a lot of time and effort, making it easy to get started without worrying about setup hassles. I also love how it keeps everything well-organized, making the project structure clean and easy to navigate.” (Participant 42):

Multi-layer architecture refers to the structured design of software systems into distinct layers, where each file is responsible for specific functionalities. In this study, 3 out of 118 participants (2.54%) highlighted that the proposed system generates projects with a clear multi-layer architecture, making development easier, less error-prone, and more accessible even to non-coders. One representative quotation is provided below.

👉 “One strength is that the system has a multi-layer architecture, which allows for different perspectives on the project. Both the front and back end of the project can be completed with little to no bugs, which saves much of the human effort needed for development. This allows for even non-coders or people with little technical background to be able to develop applications easily and quickly.” (Participant 52):

Good for ML tasks: In this study, one participant argued that the proposed system is particularly effective for ML tasks, as it generates functional ML code that requires only minor modifications to work correctly. An example quotation is included below for reference.

👉 “The best experience is offered when I try to generate machine learning algorithms instead of applications or mini games, I only need to make minor modifications to the codes and most importantly eventually they work.” (Participant 57):

Good for prototype level: In this study, 6 out of 118 participants (5.08%) mentioned that the proposed system works efficiently for prototype-level or small-scale projects. However, they also noted that the system fails to handle more complex projects, where errors become more frequent and persistent. The following is a representative example of a participant’s statement.

👉 “The system handles simpler projects efficiently, but it gave persistent failure in generating complex project structures. It mostly had finalizer bot error and stopped.” (Participant 24):

Challenges SQ7

As indicated in Figure 5, 36 out of 118 (30.51%) participants reported challenges with the agent-based system for autonomous code generation. Participants highlighted

several challenges such as hallucination, short memory, code smell, poor code quality, etc. Each challenge is explained in detail below.

Hallucination: According to Maleki *et al.* [64], hallucination refers to the model’s generation of content that is factually inaccurate, illogical, or entirely fabricated, despite appearing reasonable and coherent. In this study, 14 out of 118 (11.86%) of the participants pointed out that CodePori hallucinated by generating unrequested code components, introducing unnecessary functionalities, or misinterpreting project requirements by including elements not aligned with the specified description (e.g., adding unwanted dependencies or architectural choices). One representative quotation is provided below.

✎ *“Generated code tend to have more functions than requested, sometimes making code unusable without additions. Like database situation where model hallucinates need of database although application or descriptioned functionality doesn’t need one.” (Participant 46):*

Short memory: According to Wu *et al.* [65], memory in LLM-based agents refers to their capability to store and retrieve information, which is crucial for maintaining consistency and enabling learning across tasks. However, LLM-based agents’ short-term memory has emerged as a significant limitation; for instance, as tasks become more complex, previous information tends to be forgotten [66]. In this study, 7 out of 118 (5.93%) participants highlighted that agents often forget the previously generated class or function in another module when working with large-scale systems, leading to errors such as calling non-existent methods or failing to maintain inter-file references following code refactoring. Below, we provide a quotation that exemplifies this finding.

✎ *“There are couple of issues that makes me wonder whether the AI actually has full understanding of the whole system. For example calling non-existent methods. I believe this could be caused by a phenomena where file x is using file y’s methods but then file y is refactored due to failed tests and the AI forgets that the methods were referred in file x.” (Participant 25):*

Code smell is not a bug, but it is a structural or stylistic characteristic in the source code that indicates a potential deeper problem or suboptimal design, making the code harder to understand, modify, or maintain [67]. In this study, one participant raised the code smell issue with the proposed system’s generated code. Participants highlighted specific instances of poor code quality, such as the hardcoding of constants in multiple locations, extensive duplication of code segments, and the presence of dead or unused code. Below, we provide a quotation that exemplifies this finding.

✎ *“Code quality was also often very poor, with the same constants hardcoded in several places and lots of duplicated code and dead code.” (Participant 118):*

Code inconsistency refers to the phenomenon where the generated code lacks uniform patterns or consistent logical choices within a single output or across multiple

generated modules [68]. In this study, 4 out of 118 (3.39%) participants identified code inconsistency issues in the proposed system’s generated code. Specifically, they highlighted the absence of uniform patterns in function calls and variable definitions across generated files. A sample quote illustrating this point is presented below.

✚ “The system does provide codes that offers an overall framework and all the basic functionalities in related to the project description, but the generated codes never work out of the box, there are a lot of inconsistencies for function calls and variable definitions across generatedpython files, regardless of the complexity of the projects.” (Participant 57):

Poor code quality is defined as code that is difficult to understand, modify, extend, or maintain [69]. In this study, 20 out of 118 (16.95%) participants mentioned that the generated code quality of the proposed system was bad. For instance, the generated code frequently suffered from functional defects and dependency conflicts, which made it unusable without significant debugging and modifications. One representative quotation is provided below.

✚ “No, messy code and hard to debug. Considering the quality vs time spent coding, I still think that coding them by yourself is better.” (Participant 61):

Dependency management issue: According to Pashchenko *et al.* [70], dependency management issues arise when software systems improperly handle, track, or maintain their reliance on external libraries, frameworks, modules, or services. In this study, 2 out of 118 (1.69%) participants raised dependency management issues with the code generated by the proposed system. An example quotation is included below for reference.

✚ “It gave pretty good skeleton for a program, but the programs dont work straight out of the package, there is too much errors and problems there. Dependencies had conflicting versions that had weird errors etc. Then module errors so much.” (Participants 59):

Syntax error occurs when the grammatical rules or structural conventions of a programming language are violated, which prevents the code from being correctly understood or processed by a compiler or interpreter [71]. In this study, 3 out of 118 (2.54%) participants highlighted that the code generated by the proposed system often suffered from syntax errors. Below, we provide a representative quotation.

✚ “Most of the cases It can not generate the full code. When it generates the full code, those section needed to update as well, syntax error and import required libraries or modules are some common problem I faced with the generated code.” (Participant 82):

Orchestration failure: Agent orchestration is the automated coordination of multiple AI agents, to ensure that specialized agents can work together in a cohesive workflow to achieve a complex goal [72]. In this study, 2 out of 118 (1.69%) participants highlighted agent orchestration issues between different files. The quotation below serves as an illustrative example.

✚ “*At least from what I noticed, it doesn’t seem to know how to orchestrate between different files. It has a tendency of creating a requirements-file but not actually adding the requirements (also other excess files).*” (**Participant 49**):

Code complexity: According to Geraci *et al.* [73], code complexity is defined as “the degree to which a system or component has a design or implementation that is difficult to understand and verify, which directly affects the maintainability and defect proneness of code”. In this study, 6 out of 118 (5.08%) participants highlighted the code complexity issue with agent-based generated code. The following is a representative example of a participant’s statement.

✚ “*However, due to the complexity of the project structure and processing logic, there are inevitably some bugs, which require careful analysis and one-by-one processing by humans. The larger the project, the more complex the associations, and the more difficult it will be for the system to perform.*” (**Participant 103**):

A **configuration issue** refers to the process of setting up and customizing software applications, systems, or environments to operate according to specific requirements. It involves defining parameters, settings, and options that control how a program behaves during execution [74].

In this study, 1 out of 118 (0.85%) of the participants identified system integration and configuration as key challenges, specifically highlighting difficulties in getting generated files to work cohesively and ensuring compatibility between different components. One representative quotation is provided below.

✚ “*Importing seems to be very problematic for the system to get right and sometimes it uses non-compatible methods in different files. As I mentioned before, the biggest limitation seems to be to get the generated files to work with each other.*” (**Participant 45**):

Areas of Improvement (SQ7)

Based on real-world testing of the proposed system, participants suggested several areas for improvement. For instance, participants highlighted that a progress tracking system needed to be integrated to see the agent’s real-time progress, the need for improving debugging features, and proper framework selection. Below we discuss each area of improvement in detail.

A **Progress tracking** system provides users with real-time updates and visual indicators regarding the status and progress of an ongoing process or task. This enables

transparency and allows users to monitor efficiency and identify any potential delays or issues within the system’s operation.

In this study, 8 out of 118 (6.78%) participants suggested the addition of a progress tracking system, emphasizing its importance for users to monitor the agent’s progress. Such a feature would enhance user experience by providing critical feedback on long-running or complex operations. One representative quotation is provided below.

✎ *“For the python app in general I would like to see some kind of progress bar to better gauge if the app has gotten stuck during execution or not.” (Participant 50):*

Improve code debugging: Code debugging is the systematic process of identifying, analyzing, and removing errors (bugs) or defects from a computer program. Effective code debugging through AI agents can significantly improve development efficiency, reduce development time, and enhance the overall quality and reliability of software systems [75], [76]. In this study, 14 out of 118 (11.86%) participants argued that the system should provide comprehensive instructions for all types of errors across all relevant files. Furthermore, there should be a separate debugging option that allows users to provide a file containing an error for agents to analyze and resolve. This would empower agents to interact with the actual execution environment to gather precise runtime information, leading to more accurate and efficient bug resolution. Below, we provide a quotation that exemplifies this finding.

✎ *“The idea iss very good and for the most part the code runs well. There should be a place to debug, or give instructions for debugging. That would be a nice add.” (Participant 18):*

Human intervention with agents: Human intervention with agents, especially in complex tasks like code debugging, is crucial for improving performance and overall effectiveness. In this study, one participant argued that the system lacked human intervention when agents were developing code or crashed at some point, meaning that humans were unable to step in to diagnose, guide, or rectify issues, leading to frustration and the inability to resolve problems effectively. One representative quotation is provided below.

✎ *“One problem is that the whole system crashes when the agents face errors, and as a user you cannot intervene at all. This was many times the case and since there is no detailed error-messages, I have no idea how to fix the bug or what went wrong.” (Participant 97):*

Framework selection: In this study, two participants suggested that JavaScript framework detection should be improved to generate vanilla JavaScript by default and only use Node/React when explicitly requesting framework-based projects. A representative quote is given below.

✎ “For javascript it generates code using node and react if I mention that the code is required for js. I have to explicitly mention that I just need it for vanilla javascript then it generates the required code. It should generate code using node or react when I mention that I want a project using javascript’s framework.” (Participant 99):

Library import management: In this study, one participant suggested that the system needs improvement in library import management, particularly for Python projects where relative importing can cause issues. The participant recommended that the system should automatically use absolute imports to avoid known Python compatibility problems, rather than requiring users to explicitly specify this preference. An example quotation is included below for reference.

✎ “The areas that it needs to improve can be importing libraries specifically for python where we use relative importing of components into main file. We should need to specify separately to the model to use absolute imports because it is known that python has this issue and it should generate code accordingly.” (Participant 99):

6 Discussion

In this paper, we utilized an LLM-based agent to automate the code generation process. Our study explores the potential of these agents in software development, with the primary objective of evaluating their capability for large-scale code development in a real-world scenario. Below, we discuss the implications of our results, including the identified challenges and opportunities for future research.

6.1 LLM-Based Agent Code Generation (RQ1)

The implementation of LLM-based agents in software development demonstrated their capacity to autonomously generate functional, modular, and code for large-scale systems. The results show that these agents can manage coding tasks through collaboration and iterative refinement, producing code that aligns with predefined design and performance objectives. Several researchers have explored the use of multi-agent systems for code generation. For instance, LLM-based AI agent systems such as AgentCoder [13], SWE-Agent [15], MetaGPT [11], and similar frameworks utilize multi-agent collaboration to generate code. However, these existing systems primarily focus on generating isolated functions or small code snippets. In contrast, our proposed LLM-based AI agent system focuses on multi-file and project-oriented code generation, rather than only individual function generation. **Implications:** For practitioners, the use of LLM-based agents can reduce development time and assist in software development tasks, allowing them to focus on the more complex and creative aspects of their projects. This can lead to increased productivity and efficiency within development teams. Additionally, the ability to generate code autonomously can result in cost savings, as projects are completed more efficiently and with fewer resources [11]. From an academic and research perspective, this advancement enhances

the understanding of AI-driven software development by improving the capabilities and efficiency of autonomous code generation. Furthermore, LLM-based agents create opportunities for future research, such as examining their impact on the SE workforce, studying the implications of AI-generated code, and developing frameworks to ensure transparency, accountability, and reliability in AI-assisted programming systems.

Additionally, there is a need to extend the capabilities of the system by integrating it with existing Integrated Development Environments, version control systems, and other software development tools to make it a consistent part of the software development ecosystem. Finally, enhancing the system’s learning capabilities based on user input and customizing its outputs according to specific user preferences or project requirements could further improve its usefulness.

6.2 Performance Evaluation of LLM-Based Agents (RQ2)

In this study, we provide empirically grounded insights into the practical applicability of LLM-based agents for large-scale code development, their associated challenges, and key areas for future investigation.

6.2.1 Challenges of LLM-Based Agents:

The study highlights the strong potential of the technology, e.g., fast code generation, improved efficiency, better requirement interpretation, and support for multi-file architectures, while also identifying several limitations, including hallucinations, code smells, inconsistencies, agent orchestration failures, limited memory capacity, etc. (for details, see Section 5.2.3). These limitations are the key empirical findings of this work, as they provide valuable insights into the current challenges faced by multi-agent systems. Below, we discuss these challenges and their implications for practitioners and researchers.

Hallucination is one of the primary challenges with LLM-based agents for software development. In this study, some participants (11.86%) pointed out that agents frequently generated logically inconsistent or factually incorrect code segments. **Implications:** For practitioners, this finding highlights hallucination as a critical challenge in multi-agent systems, indicating the need for careful output verification through appropriate validation and testing practices. We believe that human oversight, continuous evaluation, and structured review processes are crucial to ensuring reliability and preventing the inclusion of incorrect or misleading code in production systems. We also argue that relying solely on the agents’ outputs without verification can be detrimental to future software projects. For researchers, this finding opens up new opportunities to explore how hallucination can be effectively mitigated in LLM-based agents for software development. Future research should investigate mechanisms and frameworks that can detect or mitigate hallucinated outputs, such as integrating verification modules, refining model architectures, or enhancing contextual grounding. Moreover, understanding how developers interact with and correct hallucinated outputs can guide the design of more trustworthy and transparent agent systems.

Participants also highlighted **Code smells and code inconsistency** issues with the proposed system output. For instance, based on participants’ feedback, we found

that when multiple agents collaborate on different modules of a system, without a unified coding style or coordination mechanism, the resulting code often differs in naming conventions, module boundaries, and documentation quality. Although frameworks such as MetaGPT [11] and AgentCoder [13] structure agent roles and workflows, they do not fully address the enforcement of consistent coding standards across agents. These inconsistencies accumulate as technical debt, limiting maintainability in large-scale systems. **Implications:** For practitioners, these findings highlight the importance of establishing clear coding standards and coordination mechanisms when deploying multiple LLM-based agents in a shared development environment. Without unified guidelines for naming conventions, documentation, and modular design, the resulting code can quickly diverge in structure and style, leading to code smells and inconsistencies. In practice, this means treating agents not as independent developers, but as contributors operating under a well-defined organizational and architectural framework. Researchers should investigate how coordination mechanisms, such as shared memory architectures, style alignment models, or communication protocols, can promote consistent and maintainable multi-agent outputs. Moreover, there is a need for empirical studies on how varying coordination strategies influence code quality and team productivity in agent-based development. Addressing these gaps could lead to the creation of frameworks that distribute tasks among agents while also preserving consistency, readability, and the long-term maintainability of the generated code.

Orchestration failure refers to breakdowns or inefficiencies in coordinating multiple agents toward a shared objective. In this study, participants highlighted coordination failures between agents, including missing modules and circular dependencies, arising from poor synchronization. According to Bhatt *et al.* [77], while orchestrating multiple agents can enhance collective performance by leveraging diverse expertise, it also introduces practical risks such as misaligned cost structures, unavailable or unsuitable agents, and overlapping knowledge. These factors can lead to ineffective orchestration, resulting in communication breakdowns, suboptimal task allocation, and redundancy. This finding suggests that orchestration is beneficial only when there are meaningful differences in agent capabilities and clearly defined coordination protocols [77]. **Implications:** Practitioners should ensure that agent interactions are governed by explicit task dependencies, communication standards, and fallback mechanisms to avoid duplication, circular dependencies, or module omissions. Implementing monitoring dashboards and orchestration controllers can help detect coordination breakdowns early. For researchers, there is an opportunity to explore hybrid orchestration approaches that combine symbolic planning with LLM-based reasoning to enable more reliable agent collaboration. Improving agent communication will contribute to building scalable multi-agent systems capable of coordinating complex SE tasks with minimal human intervention, while maintaining robustness and efficiency.

Memory is an important aspect of agent-based technology, enabling learning across tasks [78]. Our findings suggest that larger code outputs can often lead to memory or data management issues, resulting in the omission of essential code sections and incorrect output. To address these memory issues, we segmented the code into multiple modules, which helped to reduce memory constraints and ensured the integrity of

the generated code. However, further research is needed to develop practical solutions to the memory problem. **Implications:** For future research, there is an opportunity to explore memory-optimization techniques, modular generation strategies, and adaptive resource allocation methods to mitigate out-of-memory errors without compromising code completeness.

Finally, researchers have proposed several benchmarks, including HumanEval, MBPP, APPS, and SWE-Bench, to evaluate the capabilities of LLM-based agent systems for code generation. However, the existing benchmarks are designed for one-off code generation tasks. They require the model to produce code directly without the opportunity to ask for additional information. This does not reflect real-world software development, where effective communication is a critical skill [79]. As noted by Zhang *et al.* [80], while current systems perform well on these benchmarks, their performance on real-world problems is low. This indicates a lack of focus on practical code generation scenarios. Additionally, these benchmarks often rely on simple data types, such as lists and numbers. In contrast, real-world problems frequently involve more complex data, including various file types, which are more challenging to handle. For instance, benchmarks such as HumanEval and MBPP primarily assess a model’s functional correctness—whether the code passes a set of unit tests—while largely ignoring critical aspects of code quality, such as readability, maintainability, security, and adherence to coding standards. These benchmarks also contain a limited set of projects and do not represent the full range of real-world programming challenges [57]. In addition, Dai *et al.* [58] mentioned that the programming tasks contained in the HumanEval datasets are designed for small projects. This represents a notable gap in evaluating systems designed to generate code for large-scale systems. Furthermore, although advanced benchmarks such as SWE-bench attempt to address this, they still face challenges such as inconsistent development environments and outdated datasets.

Keeping in mind the above challenges, we tested the capability of the proposed multi-agent system by involving practitioners, to understand its performance in real-world scenarios and to identify practical limitations not captured by traditional benchmarks. However, relying on practitioners to test the system’s capability is not a long-term solution, as it can introduce data bias and is not a scalable or reproducible evaluation method. To this end, we argue that there is a need for a benchmark that moves beyond simple, single-function problems. It should include complex, real-world tasks that require understanding an entire system, managing multiple files, and interacting with external APIs or databases. Additionally, this benchmark should also evaluate the quality of the generated code, focusing on critical aspects such as readability, maintainability, and security.

6.2.2 Improvements of LLM-Based Agents:

This study offers empirically grounded insights into the practical applicability of agents, highlighting crucial areas for future investigation, including enhanced code verification, agent monitoring, human intervention throughout the process, and the selection of an appropriate framework.

Human intervention during agent-driven development is crucial for improving performance and ensuring effective problem resolution. In this study, participants suggested that human interaction with the agent is necessary during the development process, as agents may encounter issues or system crashes, and humans can step in to diagnose, guide, and rectify problems. **Implications:** This highlights a key area for future exploration, the need to propose an adaptive intervention framework in which humans can interact with agents to optimize performance. For instance, when agents encounter uncertainty, conflicts, or execution failures, humans can intervene to provide guidance, enabling agents to make real-time adjustments.

Improve code debugging and progress tracking: In this study, participants recommended that monitoring the real-time progress of the agents would enhance transparency and allow users to evaluate efficiency and identify potential delays or issues within the system’s operation. Additionally, participants suggested that such real-time monitoring could also help improve code debugging. **Implications:** For the academic community, these findings highlight a vital research direction focused on developing monitoring and visualization frameworks that can support real-time tracking of agent behavior and code execution. Researchers could also investigate adaptive monitoring mechanisms that integrate with orchestration frameworks to detect performance bottlenecks or coordination failures automatically. Such work would contribute to building more interpretable, reliable, and efficient agent-driven development environments.

Framework selection and library import management: In this study, participants identified opportunities to improve language-specific handling in LLM-based agents. For instance, our findings highlighted the need to prevent unnecessary framework dependencies and to align generated outputs with user intent better. Similarly, participants emphasized the need to improve library import management, particularly in Python projects where relative imports can cause compatibility issues. The findings highlight that the system should automatically detect when to use absolute imports to avoid known Python import path errors, thus eliminating the need for users to specify this manually. **Implications:** These findings suggest that careful configuration of agent environments is essential for producing reliable and context-appropriate code. Development teams should define explicit coding standards and framework policies within agent workflows to minimize unnecessary dependencies and ensure consistent code generation.

7 Threats to Validity

We followed the guidelines of Runeson *et al.* [81] to discuss threats to validity and found several threats to the validity of the study results.

7.1 Internal Validity

According to Runeson *et al.* [81], internal validity refers to the factors that may influence the data collection and data analysis process. The following factors were identified as potential threats to the internal validity of this study.

Developed System: LLM-based agents introduced two potential threats to internal validity. First, LLM-based systems are non-deterministic, meaning that the same prompts may produce different outputs. To mitigate this, each participant was assigned five projects, enabling us to obtain a more stable and representative picture of the system’s performance. Second, interruptions in agent communication present an additional threat. Since multi-agent communication relies on the API, a stable internet connection is necessary. Temporary network disruptions can interrupt agent communication, potentially leading to incomplete or incorrect outputs. To address this issue, we implemented an agent-level memory mechanism that stores conversation history, allowing agents to resume their tasks once the connection is restored.

Surveys: Two primary threats to the internal validity of the survey were identified. First, recruiting participants with extensive knowledge of LLMs, software development, and agent-based systems is inherently challenging. As a result, the participant pool consisted of individuals who had completed the “Fine-Tuning LLMs” course and had varying backgrounds, experience levels, and technical skills. While this selection ensured a baseline level of familiarity with LLMs and related technologies among participants, it is nonetheless student data.

However, we note that our questions (Table 3) were not ones that would require very specific types of SE knowledge or expertise, and were instead quite general in nature. We thus refer to past literature that argues that the suitability of student subjects largely depends on the task at hand and the expertise required (e.g., [82]), and argue that for this evaluation, students would be reasonably analogous to industry professionals in terms of being able to provide a general evaluation of the system. Nonetheless, industry professionals might have, for example, asked the system to generate more domain-specific systems related to their work domain(s), leading to different results. Our evaluation simply provides some initial indication of the performance of the system.

Second, a potential threat emerges from the limited time available to complete the system evaluation and survey. Participants were required to test the system and then respond within a fixed time frame, which may have affected the quality of their responses. To reduce this risk, we kept the survey concise and focused, limiting it to seven essential questions.

7.2 External Validity

According to Runeson *et al.* [81], external validity refers to the extent to which the results of a study can be generalized to other contexts.

Developed System: The Claude Sonnet 3.5 model was used because it was the most advanced model available when our experiment was conducted. Since then, more capable LLMs have been released, which may demonstrate different performance characteristics. In addition, recently developed agentic frameworks now offer advanced external tool capabilities, such as web search, retrieval, and automated tool integration; however, these frameworks were not available during the development or evaluation of our agent system. These advancements should be considered when assessing the generalizability of our results.

Furthermore, the proposed system relies on a commercial LLM whose behavior may vary across model versions or when deployed under different computational or network conditions. While modular task design may improve applicability, caution is still needed when applying our findings to large-scale industrial software development. Collectively, these factors limit the generalizability of the study’s outcomes to real-world development contexts.

Surveys: Although we recruited a large number of participants, they were all affiliated with a single university, which may limit the generalizability of our findings and may not fully represent the perspectives of all LLM practitioners. Conducting a follow-up study with a more diverse group of participants, including professionals from industry, academia, and other institutions, would help strengthen the external validity of the results.

7.3 Construct Validity

According to Runeson *et al.* [81], construct validity focuses on whether the survey constructs are appropriately defined and interpreted.

Surveys: A potential threat to construct validity arises from how participants evaluated the quality of the generated code. Participants’ interpretations of code quality varied: some viewed successful execution as a sufficient indicator of high performance, while others considered broader aspects such as readability, structural quality, or technical debt. This variation between participants’ interpretations and the broader construct of software quality introduces ambiguity regarding how accurately the evaluation reflects the system’s true capabilities.

7.4 Conclusion Validity

Threats to conclusion validity involve issues, such as inaccurate data, that can prevent researchers from drawing accurate conclusions.

Developed System: To mitigate this threat, the last author validated the performance of the proposed system using multiple independent test cases that represent diverse real-world scenarios in large-scale software development. Any disagreements over evaluation were discussed collaboratively until consensus was reached.

Surveys: To address this threat, the first author conducted the initial analysis of the survey data. In contrast, the second, third, and fourth authors independently reviewed the study to ensure alignment with the intended constructs. Any disagreements were resolved through collaborative discussion and joint interpretation among all authors.

8 Conclusions

In this study, we gained insights into the real-time applicability of LLM-based agents for large-scale software development. The study’s findings are meaningful for both researchers and practitioners, as they highlight the current capabilities, limitations, and practical challenges associated with integrating autonomous LLM-based agents

into real-world software development workflows. In addition, this study provides valuable evidence to guide the future fine-tuning and enhancement of LLM-based agents by addressing the challenges identified above and adopting the proposed areas for improvement. The main findings of this survey are summarized as follows:

- This study introduces CodePori, an LLM-based multi-agent system designed to evaluate the capability of agents for large-scale code generation.
- We engaged 118 participants with advanced knowledge of LLMs and diverse expertise. Of these, 69.49% had 2–5 years of experience in AI-based software development, 26.28% had approximately 2 years, and 4.23% had around 10 years’ experience. In terms of education, 47.46% held a bachelor’s degree, 27.97% a master’s degree, 22.03% a high school diploma, and 2.54% a doctoral degree.
- This study highlights 11 common challenges that practitioners and researchers can address in the future to improve the performance of LLM-based agent systems in software development. The identified difficulties include hallucinations, limitations in short-term memory, code smells in generated code, code inconsistency, poor code quality, dependency management issues, syntax errors, orchestration failures, increased code complexity, and configuration problems.
- This study identified five key areas for improvement, as suggested by participants, to enhance the future performance of agent-based systems: real-time monitoring of agent performance, effective human–agent intervention during development, improved code debugging, careful selection of appropriate frameworks, and correct library imports—particularly in Python.

We also identified several promising research directions to improve the use of LLM-based agents in software development (see the research implications in Section 6). We argue that researchers and practitioners should collaborate to develop dedicated solutions that address the challenges reported in this study.

Author Contributions

Zeeshan Rasheed, as the first author, wrote the manuscript. All authors contributed to improving the manuscript, study design, data analysis, and final approval of the submitted version.

Funding

This project is co-funded by the European Union and Business Finland under the project BF/Amalia-2023/SW.

Data Availability

The source code of our proposed LLM-based multi-agent system is publicly available for research and reproducibility purposes [27]. In addition, we have provided the complete set of participant feedback and analyzed survey data to support further validation of the findings of this study [26].

Declaration of AI Tools

AI tools were used for language editing and correct the grammar to improve clarity and grammar. No generative AI was used for figures.

Declarations

Ethical Approval

This study was conducted in accordance with the research ethics guidelines of Tampere University and relevant data protection legislation and regulations. Participation in the study was voluntary, and all survey responses were collected and processed anonymously. Participants were informed about the purpose of the study before taking part.

The study did not involve any physical intervention, clinical data, vulnerable groups, or exposure to physical risk. It also did not involve a risk of causing mental harm beyond what may be encountered in normal daily life. Therefore, according to the applicable ethical guidelines, formal approval from an ethics committee was not required.

Consent to Participate

Freely given informed consent to participate in the study was obtained from all participants before their survey responses were collected. Participants were informed that their responses would be used for research purposes and could be reported in anonymized form in a scientific publication. No personally identifiable information was collected or reported, and the anonymity of participants was maintained throughout the study.

Consent to Publish

Not applicable, as no identifiable personal data or images are included in this manuscript.

Acknowledgments

This project is co-funded by the European Union and Business Finland under the project BF/Amalia-2023/SW. We gratefully acknowledge their support, which made this research possible.

References

- [1] Bosch, J.: Speed, data, and ecosystems: the future of software engineering. *IEEE Software* **33**(1), 82–88 (2015)
- [2] Liu, J., Wang, K., Chen, Y., Peng, X., Chen, Z., Zhang, L., Lou, Y.: Large language model-based agents for software engineering: A survey. *arXiv preprint arXiv:2409.02977* (2024)

- [3] Treude, C.: Navigating complexity in software engineering: A prototype for comparing gpt-n solutions. In: 2023 IEEE/ACM 5th International Workshop on Bots in Software Engineering (BotSE), pp. 1–5 (2023). IEEE
- [4] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
- [5] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022)
- [6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.*: Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901 (2020)
- [7] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., *et al.*: A survey on large language model based autonomous agents. *Frontiers of Computer Science* **18**(6), 186345 (2024)
- [8] He, J., Treude, C., Lo, D.: Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology* **34**(5), 1–30 (2025)
- [9] Jin, H., Huang, L., Cai, H., Yan, J., Li, B., Chen, H.: From llms to llm-based agents for software engineering: A survey of current, challenges and future. *arXiv preprint arXiv:2408.02479* (2024)
- [10] Hassan, A.E., Li, H., Lin, D., Adams, B., Chen, T.-H., Kashiwa, Y., Qiu, D.: Agentic software engineering: Foundational pillars and a research roadmap. *arXiv preprint arXiv:2509.06216* (2025)
- [11] Hong, S., Zheng, X., Chen, J., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S.K.S., Lin, Z., Zhou, L., *et al.*: Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352* (2023)
- [12] Qian, C., Cong, X., Yang, C., Chen, W., Su, Y., Xu, J., Liu, Z., Sun, M.: Communicative agents for software development. *arXiv preprint arXiv:2307.07924* (2023)
- [13] Huang, D., Zhang, J.M., Luck, M., Bu, Q., Qing, Y., Cui, H.: Agentcoder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010* (2023)
- [14] Islam, M.A., Ali, M.E., Parvez, M.R.: Mapcoder: Multi-agent code generation for

competitive problem solving. arXiv preprint arXiv:2405.11403 (2024)

- [15] Yang, J., Jimenez, C.E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., Press, O.: Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems* **37**, 50528–50652 (2024)
- [16] Islam, M.A., Ali, M.E., Parvez, M.R.: Codesim: Multi-agent code generation and problem solving through simulation-driven planning and debugging. arXiv preprint arXiv:2502.05664 (2025)
- [17] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N.V., Wiest, O., Zhang, X.: Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680 (2024)
- [18] Cheng, Y., Zhang, C., Zhang, Z., Meng, X., Hong, S., Li, W., Wang, Z., Wang, Z., Yin, F., Zhao, J., et al.: Exploring large language model based intelligent agents: Definitions, methods, and prospects. arXiv preprint arXiv:2401.03428 (2024)
- [19] Levy, O., Feitelson, D.G.: Understanding large-scale software systems—structure and flows. *Empirical Software Engineering* **26**(3), 48 (2021)
- [20] Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
- [21] Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al.: Program synthesis with large language models. arXiv preprint arXiv:2108.07732 (2021)
- [22] Hendrycks, D., Basart, S., Kadavath, S., Mazeika, M., Arora, A., Guo, E., Burns, C., Puranik, S., He, H., Song, D., et al.: Measuring coding challenge competence with apps. arXiv preprint arXiv:2105.09938 (2021)
- [23] Jimenez, C.E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K.: Swe-bench: Can language models resolve real-world github issues? arXiv preprint arXiv:2310.06770 (2023)
- [24] Tang, X., Liu, Y., Cai, Z., Shao, Y., Lu, J., Zhang, Y., Deng, Z., Hu, H., An, K., Huang, R., et al.: ML-bench: Evaluating large language models and agents for machine learning tasks on repository-level code. arXiv preprint arXiv:2311.09835 (2023)
- [25] Wang, K., Li, T., Zhang, X., Wang, C., Sun, W., Liu, Y., Shi, B.: Software development life cycle perspective: A survey of benchmarks for code large language models and agents. arXiv preprint arXiv:2505.05283 (2025)

- [26] Anonymous: CodePori: Large-Scale System for Autonomous Software Development Using Multi-Agent Technology. <https://doi.org/10.5281/zenodo.17619419>. Dataset (2025)
- [27] Anonymous: CodePori_v02: Large-Scale Multi-Agent System for Autonomous Software Development. https://github.com/GPT-Laboratory/CodePori_v02. Accessed: 2025-11-15 (2025)
- [28] Feng, Y., Vanam, S., Cherukupally, M., Zheng, W., Qiu, M., Chen, H.: Investigating code generation performance of chat-gpt with crowdsourcing social data. In: Proceedings of the 47th IEEE Computer Software and Applications Conference, pp. 1–10 (2023)
- [29] Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* **30**, 681–694 (2020)
- [30] Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H.: Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology* **33**(8), 1–79 (2024)
- [31] Thiergart, J., Huber, S., Übellacker, T.: Understanding emails and drafting responses—an approach using gpt-3. arXiv preprint arXiv:2102.03062 (2021)
- [32] Hörnemalm, A.: ChatGPT as a Software Development Tool: The Future of Development (2023)
- [33] Weng, L., Wang, X., Lu, J., Feng, Y., Liu, Y., Feng, H., Huang, D., Chen, W.: Insightlens: Augmenting llm-powered data analysis with interactive insight management and navigation. *IEEE Transactions on Visualization and Computer Graphics* (2025)
- [34] Chae, Y., Davidson, T.: Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation (2023)
- [35] Rasheed, Z., Waseem, M., Kemell, K.-K., Xiaofeng, W., Duc, A.N., Systä, K., Abrahamsson, P.: Autonomous agents in software development: A vision paper. arXiv preprint arXiv:2311.18440 (2023)
- [36] Crupi, G., Tufano, R., Velasco, A., Mastropaolo, A., Poshyvanyk, D., Bavota, G.: On the effectiveness of llm-as-a-judge for code generation and summarization. *IEEE Transactions on Software Engineering* (2025)
- [37] Li, Y., Choi, D., Chung, J., Kushman, N., Schrittwieser, J., Leblond, R., Eccles, T., Keeling, J., Gimeno, F., Dal Lago, A., *et al.*: Competition-level code generation with alphacode. *Science* **378**(6624), 1092–1097 (2022)

- [38] Tufano, M., Drain, D., Svyatkovskiy, A., Deng, S.K., Sundaresan, N.: Unit test case generation with transformers and focal context. arXiv preprint arXiv:2009.05617 (2020)
- [39] Nascimento, N., Alencar, P., Cowan, D.: Comparing software developers with chatgpt: An empirical investigation. arXiv preprint arXiv:2305.11837 (2023)
- [40] Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., Luo, X., Lo, D., Grundy, J., Wang, H.: Large language models for software engineering: A systematic literature review. arXiv preprint arXiv:2308.10620 (2023)
- [41] Zheng, Z., Ning, K., Chen, J., Wang, Y., Chen, W., Guo, L., Wang, W.: Towards an understanding of large language models in software engineering tasks. arXiv preprint arXiv:2308.11396 (2023)
- [42] Zheng, Z., Ning, K., Wang, Y., Zhang, J., Zheng, D., Ye, M., Chen, J.: A survey of large language models for code: Evolution, benchmarking, and future trends. arXiv preprint arXiv:2311.10372 (2023)
- [43] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., *et al.*: Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research* **24**(240), 1–113 (2023)
- [44] Xu, F.F., Alon, U., Neubig, G., Hellendoorn, V.J.: A systematic evaluation of large language models of code. In: *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pp. 1–10 (2022)
- [45] Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow **58** (2021)
- [46] Black, S., Biderman, S., Hallahan, E., Anthony, Q., Gao, L., Golding, L., He, H., Leahy, C., McDonell, K., Phang, J., *et al.*: Gpt-neox-20b: An open-source autoregressive language model. arXiv preprint arXiv:2204.06745 (2022)
- [47] Arora, S., Narayan, A., Chen, M.F., Orr, L., Guha, N., Bhatia, K., Chami, I., Re, C.: Ask me anything: A simple strategy for prompting language models. In: *The Eleventh International Conference on Learning Representations* (2022)
- [48] Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., Savarese, S., Xiong, C.: Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:2203.13474 (2022)
- [49] Fried, D., Aghajanyan, A., Lin, J., Wang, S., Wallace, E., Shi, F., Zhong, R., Yih, W.-t., Zettlemoyer, L., Lewis, M.: InCoder: A generative model for code infilling and synthesis. arXiv preprint arXiv:2204.05999 (2022)

- [50] Zheng, Q., Xia, X., Zou, X., Dong, Y., Wang, S., Xue, Y., Wang, Z., Shen, L., Wang, A., Li, Y., et al.: Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. arXiv preprint arXiv:2303.17568 (2023)
- [51] Wang, Y., Wang, W., Joty, S., Hoi, S.C.: Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. arXiv preprint arXiv:2109.00859 (2021)
- [52] Wang, B., Komatsuzaki, A.: GPT-J-6B: A 6 billion parameter autoregressive language model (2021)
- [53] Tunstall, L., Von Werra, L., Wolf, T.: Natural Language Processing with Transformers. " O'Reilly Media, Inc.", ??? (2022)
- [54] Rasheed, Z., Waseem, M., Sami, M.A., Kemell, K.-K., Ahmad, A., Duc, A.N., Systä, K., Abrahamsson, P.: Autonomous agents in software development: A vision paper. In: International Conference on Agile Software Development, pp. 15–23 (2024). Springer Nature Switzerland Cham
- [55] Chen, B., Zhang, F., Nguyen, A., Zan, D., Lin, Z., Lou, J.-G., Chen, W.: Codet: Code generation with generated tests. arXiv preprint arXiv:2207.10397 (2022)
- [56] Paul, D.G., Zhu, H., Bayley, I.: Benchmarks and metrics for evaluations of code generation: A critical review. In: 2024 IEEE International Conference on Artificial Intelligence Testing (AITest), pp. 87–94 (2024). IEEE
- [57] Yadav, A., Singh, M.: Boldly going where no benchmark has gone before: Exposing bias and shortcomings in code generation evaluation. arXiv preprint arXiv:2401.03855 (2024)
- [58] Dai, J., Lu, J., Feng, Y., Ruan, R., Cheng, M., Tan, H., Guo, Z.: Mhpp: Exploring the capabilities and limitations of language models beyond basic code generation. arXiv preprint arXiv:2405.11430 (2024)
- [59] Mohammadi, M., Li, Y., Lo, J., Yip, W.: Evaluation and benchmarking of llm agents: A survey. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 6129–6139 (2025)
- [60] Sobo, A., Mubarak, A., Baimagambetov, A., Polatidis, N.: Evaluating llms for code generation in hri: A comparative study of chatgpt, gemini, and claude. Applied Artificial Intelligence **39**(1), 2439610 (2025)
- [61] Wohlin, C., Höst, M., Henningsson, K.: Empirical research methods in web and software engineering. Web Engineering, 409–430 (2006)
- [62] Blair, E.: A reflexive exploration of two qualitative data coding techniques. Journal of Methods and Measurement in the Social Sciences **6**(1), 14–29 (2015)

- [63] Wei, B.: Requirements are all you need: From requirements to code with llms. In: 2024 IEEE 32nd International Requirements Engineering Conference (RE), pp. 416–422 (2024). IEEE
- [64] Maleki, N., Padmanabhan, B., Dutta, K.: Ai hallucinations: a misnomer worth clarifying. In: 2024 IEEE Conference on Artificial Intelligence (CAI), pp. 133–138 (2024). IEEE
- [65] Wu, Y., Liang, S., Zhang, C., Wang, Y., Zhang, Y., Guo, H., Tang, R., Liu, Y.: From human memory to ai memory: A survey on memory mechanisms in the era of llms. arXiv preprint arXiv:2504.15965 (2025)
- [66] Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., Wen, J.-R.: A survey on the memory mechanism of large language model-based agents. ACM Transactions on Information Systems **43**(6), 1–47 (2025)
- [67] Guggulothu, T., Moiz, S.A.: Code smell detection using multi-label classification approach. Software Quality Journal **28**(3), 1063–1086 (2020)
- [68] Wang, Y., Jiang, T., Liu, M., Chen, J., Mao, M., Liu, X., Ma, Y., Zheng, Z.: Beyond functional correctness: Investigating coding style inconsistencies in large language models. Proceedings of the ACM on Software Engineering **2**(FSE), 690–712 (2025)
- [69] Börstler, J., Bennin, K.E., Hooshangi, S., Jeuring, J., Keuning, H., Kleiner, C., MacKellar, B., Duran, R., Störrle, H., Toll, D., *et al.*: Developers talking about code quality. Empirical Software Engineering **28**(6), 128 (2023)
- [70] Pashchenko, I., Vu, D.-L., Massacci, F.: A qualitative study of dependency management and its security implications. In: Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, pp. 1513–1531 (2020)
- [71] Alfred, V.A., Monica, S.L., Jeffrey, D.U.: Compilers Principles, Techniques & Tools. Pearson Education, ??? (2007)
- [72] Xiong, Y., Wang, J., Li, B., Zhu, Y., Zhao, Y.: Self-organizing agent network for llm-based workflow automation. arXiv preprint arXiv:2508.13732 (2025)
- [73] Geraci, A.: IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries. IEEE Press, ??? (1991)
- [74] Pereira, J.A., Acher, M., Martin, H., Jézéquel, J.-M., Botterweck, G., Ventresque, A.: Learning software configuration spaces: A systematic literature review. Journal of Systems and Software **182**, 111044 (2021)
- [75] Ashrafi, N., Bouktif, S., Mediani, M.: Enhancing llm code generation: A systematic evaluation of multi-agent collaboration and runtime debugging for improved

- accuracy, reliability, and latency. arXiv preprint arXiv:2505.02133 (2025)
- [76] Padurean, V.-A., Denny, P., Singla, A.: Bugspotter: Automated generation of code debugging exercises. In: Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1, pp. 896–902 (2025)
 - [77] Bhatt, U., Kapoor, S., Upadhyay, M., Sucholutsky, I., Quinzan, F., Collins, K.M., Weller, A., Wilson, A.G., Zafar, M.B.: When should we orchestrate multiple agents? arXiv preprint arXiv:2503.13577 (2025)
 - [78] Zhang, Z., Dai, Q., Bo, X., Ma, C., Li, R., Chen, X., Zhu, J., Dong, Z., Wen, J.-R.: A survey on the memory mechanism of large language model based agents. *ACM Transactions on Information Systems* (2024)
 - [79] Wu, J.J., Fard, F.H.: Humanevalcomm: Benchmarking the communication competence of code generation for llms and llm agent. arXiv preprint arXiv:2406.00215 (2024)
 - [80] Zhang, S., Zhao, H., Liu, X., Zheng, Q., Qi, Z., Gu, X., Zhang, X., Dong, Y., Tang, J.: Naturalcodebench: Examining coding performance mismatch on humaneval and natural user prompts. arXiv preprint arXiv:2405.04520 (2024)
 - [81] Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering* **14**, 131–164 (2009)
 - [82] Höst, M., Regnell, B., Wohlin, C.: Using students as subjects—a comparative study of students and professionals in lead-time impact assessment. *Empirical Software Engineering* **5**(3), 201–214 (2000)

9 Appendix

Below, we present the complete prompts for all six agents. Each prompt is shown in full, with brief context on its role in the overall workflow. Finally, we provide the algorithm at the end to tie these components together and explain how the agents interact step-by-step.

<p>Agent Name: <i>[Manager_Bot]</i></p> <p>You are ManagerBot.</p> <p>Your goal is to read and analyze the following project description:</p> <p>\$PROJECT_DESCRIPTION</p> <p>From this description, break down the system into modules, each of which will handle a specific part of the functionality.]</p> <p>**IMPORTANT**:</p> <ol style="list-style-type: none"> 1. You must output "only" valid JSON. 2. Do "not" include any commentary, explanation, or Markdown formatting. 3. The JSON must be a "strict array" of module objects. 4. Each module object should have the following keys: <ul style="list-style-type: none"> "name" (string): The name of the module. "description" (string): What the module does. "dependencies" (array of strings): The names of other modules that this module depends on. <p>## Example JSON Format</p> <pre>[{ "name": "User Management", </pre>	<pre> "description": "Handles user registration, login, profile management, and authentication.", "dependencies": [] }, { "name": "Product Catalogue", "description": "Manages the list of products, their details, and categories.", "dependencies": [] }, { "name": "Shopping Cart", "description": "Manages the shopping cart functionality for users.", "dependencies": ["User Management", "Product Catalogue"] }] </pre> <p>### Final Instructions:</p> <ul style="list-style-type: none"> • If you generate a different set of modules, that's fine; just ensure the JSON array structure is correct. • "No" extra keys besides "name", "description", and "dependencies" • "No" text or lines outside the JSON array. • Now read the project description and generate your JSON array.
---	---

(a) Manager agent

<p>Agent Name: <i>[Architecture_Bot]</i></p> <p>Objective: [Return a valid JSON object with exactly one key: - "architecture_overview" (string): A thorough textual explanation of the overall project architecture design, including how components interact, data flow, and database connections...]</p> <p>Instructions:</p> <ul style="list-style-type: none"> • Content Requirements: [Provide a detailed overview of the project architecture tailored to the specified LANGUAGE] • Exclude: [any information about test files or log files. • - Focus solely on the main code files and their interactions. • - Limit the description to "5-10 main code files" relevant to the project's functionality. • - Do not include unnecessary files such as "readme.md", ".env", ".gitignore", etc.] <p>Language-Specific Architecture:</p> <p>*JavaScript*</p> <ul style="list-style-type: none"> • Separate the project into "client" (frontend) and "server" (backend) components. • The client (React) communicates with the server (Node.js and Express) via RESTful APIs. • Create "package.json" for both "client" and "server", detailing dependencies. • Explain how MongoDB is used for database operations and how API endpoints handle data flow between client and server. • Describe how the server manages database connections, including any ORM or ODM used (e.g., Mongoose). <p>*Python*</p> <ul style="list-style-type: none"> • Design the project with a Flask app that includes both backend and frontend components, if applicable. • Detail the MVC architecture or other relevant architectural patterns. • Explain interactions between different modules, data management, and any frontend integration (e.g., HTML templates or a frontend framework). • Specify the contents and purpose of "requirements.txt" if necessary. • Describe how the Flask app connects to the database (e.g., SQLAlchemy for ORM). 	<p>*Java*</p> <ul style="list-style-type: none"> • Focus on a backend-only architecture with a command-line interface (CLI). • Explain how different Java classes and packages interact. • Describe any design patterns employed (e.g., Singleton, Factory). • Detail how the application connects to the database, including JDBC configurations or ORM frameworks like Hibernate. <p>*Other Languages*</p> <ul style="list-style-type: none"> • Adapt the architectural overview based on the standard practices and frameworks associated with the specified LANGUAGE. • Follow the same separation of concerns and modularity principles tailored to the language. • Include details on how the application interfaces with the database, relevant to the chosen language and framework. <p>*Best Practices*</p> <ul style="list-style-type: none"> • Ensure clarity and coherence in the architectural explanation. • Use technical terminology appropriate to the LANGUAGE and its ecosystem. • Avoid jargon that may not be universally understood. • Maintain objectivity and focus on the structural aspects of the project. • Provide detailed descriptions of what each file/module does and how they interconnect, including database interactions. <p>*Output Format*</p> <ul style="list-style-type: none"> • Return only the raw JSON object without any additional text, explanations, or formatting. • Do not use triple backticks or any other markdown formatting. <p>*Example Output*</p> <pre>{ "architecture_overview": "Here is a description..." }</pre> <p>*Imports*</p> <ul style="list-style-type: none"> • Do not include any extra commentary or explanations outside of the JSON object. • Ensure that the JSON is properly formatted and valid.
---	--

(b) Architecture agent

Fig. 6: Prompts for manager agent and architecture agent

<p>Agent Name: <i>[Flow_Structure_Bot]</i></p> <p>You are FlowStructureBot.</p> <p>Given the project description: PROJECT_DESCRIPTION: \$PROJECT_DESCRIPTION PROJECT_LANGUAGE: \$PROJECT_LANGUAGE</p> <p>Coding language and stack-specific guidelines:</p> <ul style="list-style-type: none"> Select the coding language and tech stack based on the provided project description and language. <p>**JavaScript-based projects**:</p> <ul style="list-style-type: none"> Use the MERN stack or any other appropriate stack as needed. Separate the structure into 'client/' and 'server/' folders for front-end and back-end code, respectively. Include configuration files such as 'package.json' inside 'client/' and 'server/' if relevant. Ensure database configuration files (e.g., '.env', 'config.json') are placed appropriately within the 'server/' directory.. <p>**Java-based projects**:</p> <ul style="list-style-type: none"> Do not include HTML or CSS files. - Focus only on Java files and backend logic, structured appropriately. - Organize code into packages reflecting different modules or functionalities. - Include configuration files for database connections if necessary. <p>**Python-based projects**</p> <ul style="list-style-type: none"> Include folders like 'templates/' and 'static/' for Flask or Django applications if needed. Do not include unnecessary files such as '.env', or 'README.md', unless explicitly required. Ensure configuration files like 'config.py' are placed in a dedicated 'config/' directory if applicable. Include 'requirements.txt' for managing dependencies. 	<p>**Other Languages**</p> <ul style="list-style-type: none"> Adapt the file structure based on standard practices for the selected language or stack. Include necessary configuration and setup files as per the language's conventions. Ensure database configuration files are included if required. <p>**Output Requirements**</p> <ul style="list-style-type: none"> Return a "plain-text tree structure" with the folder and file hierarchy. "No JSON or additional commentary". Include a maximum of "5-10 relevant code files" tailored to the project description. "Exclude unnecessary files" like requirements, readme, .env, .gitignore, etc <p>**Output format**: (Not to include three backticks in response)</p> <ul style="list-style-type: none"> project_name/ - app.extension - modules/ - file1.extension - file2.extension - file3.extension <p>**Additional Notes**:</p> <ul style="list-style-type: none"> Ensure all file and folder names follow conventions for the chosen language or framework. Dynamically adapt to the stack and project-specific requirements, including naming conventions and folder organization. For frameworks like Flask/Django, include necessary subdirectories (e.g., 'templates/', 'static/'). For MERN stack, ensure clear separation between 'client' and 'server' directories. Include database configuration files within the appropriate directories based on the tech stack. Ensure that the folder structure supports absolute imports to facilitate seamless module integration.
---	--

(a) Flow structure agent

<p>Agent Name: <i>[Dev_Bot]</i></p> <p>Context:</p> <ul style="list-style-type: none"> -ARCHITECTURE_OVERVIEW: \$ARCHITECTURE_OVERVIEW -FLOW_STRUCTURE: \$FLOW_STRUCTURE -MODULE_DESCRIPTION: \$MODULE_DESCRIPTION -ACCUMULATED_CODE: \$ACCUMULATED_CODE -PROJECT_DESCRIPTION: \$PROJECT_DESCRIPTION -LANGUAGE: \$LANGUAGE <p>Goal: Produce "a single code file" for the path in MODULE_DESCRIPTION.</p> <p>**Instructions**:</p> <p>1. **Language-Specific Handling**:</p> <p>-**JavaScript**:</p> <ul style="list-style-type: none"> Utilize the "MERN" stack (MongoDB, Express, React, Node.js). Generate both "front-end" and "back-end" code as required. Include necessary configuration files such as 'config.json', etc. Separate 'client/' (frontend) and 'server/' (backend) components properly. Ensure 'package.json' files for both 'client/' and 'server/' are generated and accurately configured. Ensure proper integration between front-end and back-end components. Use absolute imports based on the file/folder structure to avoid relative import issues. <p>**Python**</p> <ul style="list-style-type: none"> Use appropriate frameworks (e.g., Django, Flask) based on the project description. Include necessary configuration files like 'config.py', 'requirements.txt' and 'package.json' (only if specified). Generate both backend (API or logic) and frontend (HTML templates or JS) components if needed. Ensure module imports use absolute paths reflecting the file/folder structure. Generate a proper 'requirements.txt' listing all necessary libraries used. <p>**Java**:</p> <ul style="list-style-type: none"> Design the project with a Flask app that includes both backend and frontend components, if applicable. Detail the MVC architecture or other relevant architectural patterns. Explain interactions between different modules, data management, and any frontend integration (e.g., HTML templates or a frontend framework). Specify the contents and purpose of 'requirements.txt' if necessary. Describe how the Flask app connects to the database (e.g., SQLAlchemy for ORM). <p>**Other Languages**</p> <ul style="list-style-type: none"> Adapt the technology stack based on the selected language (e.g., Spring Boot for Java, Ruby on Rails for Ruby). Include relevant configuration and setup files as per standard practices. Ensure proper import statements and dependency management as per the language's conventions. 	<p>2. **Module Type Determination**:</p> <p>**Tests**:</p> <ul style="list-style-type: none"> If 'MODULE_DESCRIPTION' indicates tests (e.g., path includes "tests/test_..."), generate unit tests referencing 'ACCUMULATED_CODE'. Use testing frameworks appropriate to the LANGUAGE (e.g., Jest for JavaScript, unittest or pytest for Python).s and how they interconnect, including database interactions <p>3. **Code Generation Guidelines**:</p> <p>**Formatting**:</p> <ul style="list-style-type: none"> Wrap the final code in triple backticks with the language identifier, e.g.: <ul style="list-style-type: none"> '''\$LANGUAGE // code here ''' <p>**File Management**</p> <ul style="list-style-type: none"> Do not recreate or overwrite existing files. Only produce new code needed for the specified 'MODULE_DESCRIPTION'. Use the 'rel_path' specified in 'MODULE_DESCRIPTION' for the generated file path, ensuring the code integrates seamlessly. <p>**Best Practices**</p> <ul style="list-style-type: none"> Ensure that the code adheres to the best practices and coding standards of the specified LANGUAGE. Maintain seamless integration with the existing codebase provided in 'ACCUMULATED_CODE'. Focus primarily on backend logic and functionalities to maintain project efficiency and scalability. For HTML files, ensure that CSS files are correctly linked and that all input fields are properly defined and functional. <p>4. **Additional Instructions**:</p> <p>**Exclude Unnecessary Files**:</p> <ul style="list-style-type: none"> Do not create files such as 'readme.md', '.env', '.gitignore', unless explicitly required by the LANGUAGE-specific framework. <p>**Configuration Files**:</p> <ul style="list-style-type: none"> For frameworks or stacks that require configuration files (e.g., 'config.json' for MERN), ensure they are properly generated and integrated. <p>**Dynamic Adaptation**:</p> <ul style="list-style-type: none"> Adjust the code generation process dynamically based on the specified LANGUAGE and its associated technology stack. <p>**File Paths**:</p> <ul style="list-style-type: none"> Use the 'rel_path' specified in 'MODULE_DESCRIPTION' for the generated file path, ensuring the code integrates seamlessly.
--	---

(b) Dev agent

Fig. 7: Prompts for flow structure agent and dev agent

<p>Agent Name: <i>[Verification_Bot]</i></p> <p>[You are VerificationBot. You receive: ACCUMULATED_CODE: \$ACCUMULATED_CODE PROJECT_DESCRIPTION: \$PROJECT_DESCRIPTION MODULE_CODE: \$MODULE_CODE MODULE_NAME: \$MODULE_NAME</p> <p>Your task: Provide a short JSON snippet with your verification or any issues.</p> <p>Example: { "verification": "All good" }</p> <p>OR { "verification": "Found issues: missing X, Y..." }</p>	<p>*Importants*</p> <ul style="list-style-type: none"> • Ensure that the verification process accounts for the specific tech stack and coding standards. • Do not include unnecessary files or code elements. <p>Output only this JSON. No extra commentary..</p>
---	--

(a) Verification agent

<p>Agent Name: <i>[Finalizer_Bot]</i></p> <p>You are FinalizerBot1.</p> <p>PROJECT_DESCRIPTION: \$PROJECT_DESCRIPTION</p> <p>FILE_FOLDER_STRUCTURE: \$FILE_FOLDER_STRUCTURE</p> <p>ACCUMULATED_CODE: \$ACCUMULATED_CODE</p> <p>REVIEW: \$REVIEW</p> <p>Instructions:</p> <p>1. **Combine and Correct**:</p> <ul style="list-style-type: none"> • Use the provided 'ACCUMULATED_CODE' and 'REVIEW' to address any corrections or improvements for the files. • Ensure all corrections suggested in 'REVIEW' are applied and the final code adheres to best practices for the specified project language or stack. • For missing files mentioned in 'FILE_FOLDER_STRUCTURE', generate the required code based on the project description. <p>2. **Language-Specific Notes**:</p> <p>**JavaScript**:</p> <ul style="list-style-type: none"> • Ensure correct integration between front-end and back-end components (e.g., React and Express). • Properly handle HTML, CSS, and JavaScript files, including dependency imports/exports. • Ensure absolute imports are used based on the file/folder structure to avoid relative import issues. • Verify that 'package.json' files for both 'client' and 'server' include all necessary dependencies and scripts. <p>**Python**</p> <ul style="list-style-type: none"> • Ensure alignment with frameworks like Flask or Django, including templates and static files if required. • Address module imports and ensure any configuration files like 'config.py' are functional. • Generate a proper 'requirements.txt' listing all necessary libraries used. • Confirm that ORM configurations (e.g., SQLAlchemy) are correctly implemented and connected. <p>**Java**</p> <ul style="list-style-type: none"> • Focus on backend logic and command-line utilities. • Use appropriate design patterns (e.g., Singleton, Factory) as required. • Ensure that imports are absolute and reflect the file/folder structure accurately. • Verify that database connections using JDBC or ORM frameworks like Hibernate are properly configured. <p>*Other Languages*</p> <ul style="list-style-type: none"> • Adapt corrections based on standard practices and frameworks of the specified language. • Ensure proper import statements and dependency management as per the language's conventions. • Validate that database connections and configurations are correctly implemented. 	<p>3. **Error Handling**:</p> <ul style="list-style-type: none"> • If issues in 'REVIEW' are unclear, make logical improvements based on the provided project description and accumulated code. • Do not leave any file unaddressed. • Ensure that all database connections are secure, efficient, and correctly integrated. • Validate that all modules and components interact seamlessly as per the 'ARCHITECTURE_OVERVIEW'. <p>4. **Output Format**:</p> <ul style="list-style-type: none"> • - For each file in the structure, return an object with: • "rel_path": Path to the file (from 'FILE_FOLDER_STRUCTURE'). • "updated_code": The finalized code for that file after corrections. • Ensure the JSON contains all files from 'FILE_FOLDER_STRUCTURE', even if no changes are required. <p>**Additional Steps**:</p> <ul style="list-style-type: none"> • After applying corrections, verify that all files are correctly linked and that database connections are functional. • Ensure that frontend and backend components interact as described in the 'ARCHITECTURE_OVERVIEW'. • Confirm that all configuration files (e.g., 'config.json', 'config.py', 'package.json') are correctly set up and integrated. • Return the response in the following JSON format: <pre>{ "final_codes": [{ "rel_path": "path/to/file1.js", "updated_code": "finalized code for file1.js" }, { "rel_path": "path/to/file2.py", "updated_code": "finalized code for file2.py" }, ...] }</pre> <p>**Additional Instructions**:</p> <ul style="list-style-type: none"> • Ensure that all imports within the code use absolute paths based on the file/folder structure to prevent relative import issues, especially in Java and Python. • For HTML files, ensure that CSS files are correctly linked and that all input fields are properly defined and functional. • Focus primarily on backend logic and functionalities to maintain project efficiency and scalability. • Provide detailed descriptions in the architecture to clarify the role and interaction of each file. • Avoid creating unnecessary or redundant files unless explicitly required by the project specifications.
---	--

(b) Finalization agent

Fig. 8: Prompts for verification agent and finalization agent

```

1  # Require: Project description, preferred programming language
2  # Ensure: Structured project code, reviewed and finalized
3
4  # === Initialization ===
5  LOAD OpenRouter API key from environment
6  INITIALIZE OpenAI client with base URL
7  CONFIGURE logging for both file and console output
8
9  # === Define abstract agent structure ===
10 DEFINE class Agent with methods:
11     - load_prompt(prompt_file)
12     - update_prompt(replacements)
13     - communicate(user_message)
14     - reset_conversation()
15
16 # === Specialized Agent Bots ===
17 DEFINE Manager agent:
18     USE prompt template to GENERATE module_descriptions JSON
19     INITIALIZE Manager agent with 'manager agent.txt'
20     CALL Manager agent with project description
21     RETURN module_descriptions as JSON
22
23
24
25
26 DEFINE Architecture agent:
27     USE prompt template to GENERATE architecture_overview JSON
28     INITIALIZE Architecture agent with 'architecture agent.txt'
29     CALL Architecture agent with project description and language
30     RETURN architecture_overview as JSON
31
32 DEFINE FlowStructure agent:
33     GENERATE folder/file hierarchy in plain text
34     INITIALIZE FlowStructure agent with 'flow_structure agent.txt'
35     ,
36     CALL FlowStructure agent with architecture and language
37     RETURN project folder structure as plain text
38
39 DEFINE Dev agent:
40 FOR each file in folder structure DO
41     CREATE prompt with architecture, flow structure, and partial
42     code
43     CALL Dev agent to generate code snippet
44     STORE code in accumulated_code[file_path]
45 END FOR
46
47 DEFINE Verification agent:

```

```

48     INITIALIZE Verification agent with 'verification agent.txt'
49     FOR each code file DO
50         CALL Verification agent with project description, code,
           and context
51         STORE feedback/review
52     END FOR
53
54 DEFINE Finalization agent:
55     INITIALIZE Finalization agent with 'finalizer_agent.txt'
56     FOR each review in reviews DO
57         IF review indicates issues THEN
58             CALL Finalization agent with accumulated code,
               reviews, and language
59             EXTRACT finalized code and overwrite original
60         END IF
61     END FOR
62
63 # === State Tracking ===
64 INITIALIZE StateManager:
65     - project_description, architecture, flow_structure
66     - accumulated_code, reviews
67
68 # === Folder Structure Parsing ===
69 PARSE flow_structure text INTO list of file paths
70 HANDLE indentation to determine directory nesting
71
72 # === File Management Utilities ===
73 DEFINE create_directories_and_save_file():
74     CREATE directories recursively
75     CLEAN code blocks
76     SAVE to disk
77
78 # === Main Project Generation ===
79 FUNCTION generate_project_stream(project_description, language):
80     CALL Architecture agent → GET architecture_overview
81     CALL FlowStructure agent → GET folder hierarchy
82     PARSE folder structure to file_paths
83
84     FOR each file_path in file_paths:
85         CALL Dev agent → GENERATE initial file code
86         CALL Verification agent → REVIEW generated code
87
88         IF review CONTAINS issues:
89             CALL Finalization agent → REFINER and FIX code
90             SAVE Finalized code
91         ELSE:
92             SAVE original code
93
94     PACKAGE all files for download
95     RETURN zip download link

```

Listing 1: **Algorithm 1: Agent-Based Software Development Pipeline**