

The Impact of Generative AI on Collaborative Open-Source Software Development: Evidence from GitHub Copilot

Fangchen Song

University of Texas at Austin
2110 Speedway
Austin, Texas
United States
78712

fangchen.song@mccombs.utexas.edu

Ashish Agarwal

University of Texas at Austin
2110 Speedway
Austin, Texas
United States
78712

ashish.agarwal@mccombs.utexas.edu

Wen Wen

University of Texas at Austin
2110 Speedway
Austin, Texas
United States
78712

wen.wen@mccombs.utexas.edu

Abstract

Generative artificial intelligence (AI) enables automated content production, including coding in software development, which can significantly influence developer participation and performance. To explore its impact on collaborative open-source software (OSS) development, we investigate the role of GitHub Copilot, a generative AI pair programmer, in OSS development where multiple distributed developers voluntarily collaborate. Using GitHub's proprietary Copilot usage data, combined with public OSS repository data obtained from GitHub, we find that Copilot use increases project-level code contributions by 5.9%. This gain is driven by a 2.1% increase in individual code contributions and a 3.4% rise in developer coding participation. However, these benefits come at a cost as coordination time for code integration increases by 8% due to more code discussions enabled by AI pair programmers. This reveals an important tradeoff: While AI expands who can contribute and how much they contribute, it slows coordination in collective development efforts. Despite this tension, the combined effect of these two competing forces remains positive, indicating a net gain in overall project-level productivity from using AI pair programmers. Interestingly, we also find the effects differ across developer roles. Peripheral developers show relatively smaller gains in project-level code contributions and face a higher increase in coordination time than core developers, likely due to the difference in their project familiarity. In summary, our study underscores the dual role of AI pair programmers in affecting project-level code contributions and coordination time in OSS development. Our findings on the differential effects between core and peripheral developers also provide important implications for the structure of OSS communities in the long run.

Keywords: Generative AI, AI Pair Programmer, Open-source Software Development, Project-level Code Contributions, Coordination Time, Core Developers, Peripheral Developers

1. Introduction

The continuous advancements in generative artificial intelligence (AI) are transforming content production across a wide range of domains. Cutting-edge generative AI tools can not only automate mundane tasks but also enhance original content. In the context of software development, generative AI-powered pair programmers (i.e., AI pair programmers) like GitHub Copilot, Amazon Q Developer, Google Gemini, and Chat GPT can swiftly generate code based on developers' prompts, parameters, and descriptions (Dakhel et al. 2023). By reducing common coding errors and repetitive coding needs, these AI pair programmers hold great potential to shape the software development process. A growing body of literature has investigated various impacts of generative AI on well-defined and discrete tasks on individuals, such as writing and customer service tasks (Noy and Zhang 2023, Brynjolfsson et al. 2025). However, there is limited understanding on the role of generative AI in complex tasks that involve team collaboration and require rich contextual knowledge, such as software development.

As a popular form of software development, open-source software (OSS) development relies on the voluntary participation of geographically dispersed developers without formal hierarchies or standardized workflows (Levine and Prietula 2014, Lindberg et al. 2016). Prior work shows that generative AI can enhance individual developer productivity (Peng et al. 2023, Cui et al. 2024). However, it remains unclear how such improvements at the individual level translate into project-level code contributions, which are a function of both individual contribution intensity and the extent of developer participation. While AI pair programmers accelerate individual coding tasks, they may also affect developers' decisions to participate. Research has documented generative AI has led to a significant reduction in participation in online discussions on platforms like Stack Overflow (Burtch et al. 2024). However, given the distinctions between OSS development and those information exchange communities, it is not clear how generative AI affects participation in OSS settings. On one hand, developers may be encouraged to participate because generating code becomes easier. On the other hand, this may reduce their incentive to contribute to OSS as a means of skill development or reputation building.

Beyond project-level code contributions, effective coordination is critical to the success of OSS development (Koushik and Mookerjee 1995, Reagans et al. 2016, Mawdsley et al. 2022). However, coordination is often informal and decentralized in OSS settings, shaped by fluid team composition and voluntary participation. In our study, we focus on one key aspect of coordination efficiency, namely, the coordination time needed to integrate individual code contributions into the codebase of an OSS project (Howison and Crowston 2014, Lindberg et al. 2016, Medappa and Srivastava 2019, Shaikh and Vaast 2023). Shorter coordination time for code integration enables faster software development. Despite its importance, it remains unclear how AI pair programmers affect coordination time in OSS projects. Since AI pair programmers can make it easier to generate and interpret code, developers may be able to allocate more effort to integration, potentially reducing coordination time. However, if developers also participate more in the review process, this could increase the amount of communication required, thereby raising coordination time. Moreover, given the inherent tradeoff between contribution volume and the coordination effort required to manage them (Bakos and Brynjolfsson 1993), it is not yet clear how AI pair programmers shape the overall project-level productivity of OSS development.

Moreover, OSS development typically involves two types of developers: core developers and peripheral developers (Setia et al. 2012). Core developers design the overall architecture, write code, and maintain control over the codebase, whereas peripheral developers contribute on an irregular basis and focus on fixing bugs and adding incremental features (Setia et al. 2012, Gousios et al. 2014, Medappa and Srivastava 2019). Thus, one important distinction between core and peripheral developers lies in their level of contextual knowledge about a focal project, or project familiarity, which refers to understanding of the project's architecture, design, and norms. While existing literature has focused on the heterogenous impacts of generative AI based on workers' skill levels (e.g., Brynjolfsson et al. 2025), the conclusions cannot be easily generalized to understand how generative AI affects core versus peripheral developers, due to the important theoretical distinctions between skills needed to complete individual tasks and project familiarity required for complex software development.

Motivated by these observations, in this study we aim to answer the following research questions.

First, how do AI pair programmers affect project-level code contributions in OSS development? Second, how do AI pair programmers affect coordination time for code integration in OSS development? Third, how do AI pair programmers affect code contributions and coordination time for core versus peripheral developers differently? Answers to questions could provide important implications on how to leverage generative AI tools to support collaborative and distributed software development in the OSS community.

We argue that AI pair programmers can increase project-level code contributions in OSS development, because AI pair programmers not only boost individual code contributions but also encourage greater developer participation. However, because AI pair programmers also encourage more participation in discussions on OSS projects, they could lead to longer coordination time for code integration. Moreover, we argue that because of peripheral developers' lower project familiarity and the limited ability of AI pair programmers to learn the project's full context, the increase in project-level code contributions from peripheral developers may be smaller than that from core developers. At the same time, code contributed by peripheral developers may require more coordination time to integrate than code from core developers. Overall, the productivity gains from AI pair programmers may be smaller for peripheral developers than for core developers.

To empirically test these hypotheses, we examine how the AI pair programmer, GitHub Copilot, influences project-level code contributions and coordination time of OSS projects on GitHub, as well as how its effect varies among core versus peripheral developers. GitHub is one of the largest code-hosting repositories based on the Git version control system (Dabbish et al. 2012). In this setting, a repository is a fundamental unit that typically contains the source code and resource files for a software project, along with information related to the project's evolution history, high-level features, and developer details (Zhang et al. 2017). Such repositories are often used to investigate collaborative development practices (Dabbish et al. 2012). Our unit of analysis is at the repository-month-level, with the sample period from January 2021 to December 2022. To examine our research questions, we use a combination of publicly available data on GitHub repositories and proprietary data on Copilot use provided by GitHub organization. Our treatment

group consists of repositories where Copilot was both supported by local coding environments and used by developers to code. Thus, the post-treatment period includes the months during which Copilot was supported and used in a focal repository, and the pre-treatment period includes all other months. The control group includes repositories where Copilot was not used throughout the sample period. We estimate our model using the Generalized Synthetic Control Method (GSCM) and validate the results through alternative matching techniques and a comprehensive set of robustness checks.

Our empirical results show that the adoption of GitHub Copilot is associated with a 5.9% increase in the number of project-level code contributions but also an 8% increase in coordination time for code integration. These findings indicate a tradeoff between contribution gains and coordination time in the OSS development following the adoption of Copilot. Further analysis of the underlying mechanism suggests that the observed increase in project-level code contributions is accompanied by a significant increase in both individual code contributions and developer participation. At the same time, the increase in coordination time is driven by a higher volume of discussions surrounding code contributions, a broader set of developers participating in these discussions, and greater discussion intensity per developer. Importantly, the combined effect of these two competing forces still yields an overall positive effect on the project-level productivity, measured by the total code contributions with timely integration into the codebase.

Furthermore, we find that compared to core developers, AI pair programmers lead to a smaller increase in project-level code contributions made by peripheral developers; following the adoption of AI pair programmers, there is also a larger increase in coordination time for integrating code contributed by peripheral developers. These results are consistent with our hypothesis that due to the different levels of project familiarity held by core versus peripheral developers and the limitations of generative AI tools, peripheral developers may realize less productivity gain from AI pair programmers than core developers.

Our study provides several contributions to the literature. First, it contributes to the literature on generative AI in software development (Imai 2022, Barke et al. 2023, Peng et al. 2023, Cui et al. 2024). While prior research has shown that generative AI improves individual developer productivity (e.g., Peng et al. 2023; Cui et al. 2024), less is known about its impact on voluntary participation in collaborative

software development. Existing studies suggest that generative AI might reduce voluntary participation in Q&A communities by substituting for information exchange (Xu et al. 2023, Burch et al. 2024). However, OSS communities are fundamentally different in that they involve not only information sharing but also complex problem solving and team collaboration. To our knowledge, we are among the first to show generative AI encourages more participation in OSS development, including both coding and non-coding participation (i.e., code discussions).

Second, our study contributes to the literature on generative AI in team-based collaboration (Li et al. 2024, Dell'Acqua et al. 2025). While prior work has examined the impact of generative AI within traditional teams characterized by fixed size and formal coordination processes for performing a common task, we extend this work to open collaborative environments, where team composition is fluid, participation is voluntary, and individuals perform distinct tasks that must be integrated. Our study is among the first to uncover some unexpected impacts of generative AI tools—because these AI tools encourage developers' participation in non-coding activities such as discussions, they could lead to longer coordination time in order to reconcile different ideas and perspectives among developers.

Third, our study adds to the literature that explores the heterogeneity in the roles of generative AI among individuals (Dell'Acqua et al. 2023, Demirci et al. 2025). Prior studies have focused on how individual skills play a role in shaping the effect of generative AI tools on completing discrete tasks and they found that individuals with lower skills usually obtain greater productivity gain from these tools than highly skilled individuals (e.g., Peng et al. 2023, Cui et al. 2024). Our results draw a sharp contrast with these findings, as we demonstrate that peripheral developers obtain less productivity gain from AI pair programmers than core developers in OSS settings, potentially because the former do not possess important and necessary contextual knowledge about an OSS project to effectively use the AI tools.

2. Literature Review

2.1 Generative AI in Software Development

A growing body of literature has started to examine the impact of generative AI on software development. Most studies have focused on the implications of generative AI for individual productivity on specific tasks.

For example, Imai (2022) finds that GitHub Copilot produces more lines of code than a human pair programmer when completing a task in Python. Peng et al. (2023) find that GitHub Copilot enables individual developers to implement an HTTP server 55.8% faster than those not using the tool. Hoffmann et al. (2024) show that GitHub Copilot causes individual developers to shift focus towards coding tasks and away from project management.

Despite these insights, research on how generative AI influences project-level outcomes for complex tasks involving multiple developers remains limited. Yeverechyahu et al. (2024) investigate the innovation capabilities of generative AI, particularly its role in extrapolative versus interpolative thinking, and compare its influence on innovation in Python versus R. Different from the existing literature and built upon the OSS literature, our hypotheses are motivated by the unique characteristics of OSS, namely, the software development process in an open and collaborative environment. Because participation is often voluntary and coordination does not follow formal centralized processes in this environment, it remains unclear how generative AI influences open participation in both coding and non-coding activities, as well as team coordination, all of which could have important implications for project-level software development productivity.

In addition, existing research that explored heterogeneity in the roles of generative AI among individuals has mostly focused on understanding the differential effects between workers with high skills against those with low skills for discrete tasks (e.g., Cui et al. 2024, Brynjolfsson et al. 2025). However, it remains unclear whether the results hold in settings with complex tasks that require not only skills but also contextual knowledge and team collaboration. In the context of OSS development, the distinction between core and peripheral developers lies not in their programming skills, but in their roles, responsibilities, and the resulting level of contextual knowledge about a focal project (Crowston et al. 2006, Setia et al. 2012). Therefore, the prediction from prior studies on how individuals with different levels of skills benefit from generative AI provides limited insights on how it influences core and peripheral developers within OSS communities. We seek to bridge this important gap in the literature by investigating the differential impact of generative AI on core versus peripheral developers.

2.2 Generative AI and Voluntary Participation

Voluntary participation is essential to online communities. Prior research suggests that individuals contribute to online communities not only to exchange information, but also to derive social benefits (Zhang and Zhu 2011). With generative AI tools capable of producing content autonomously, there have been growing concerns that AI tools may crowd out human contributions. For example, studies on platforms such as Stack Overflow find that generative AI reduces participation by replacing straightforward information exchange (Xu et al. 2023, Burtch et al. 2024). Similarly, research on freelancer platforms shows a decline in job postings for simple coding tasks (Demirci et al. 2025).

However, these online communities differ from collaborative environments such as OSS development, so prior studies on the implications of AI for participation in online communities may not hold in the OSS setting. More specifically, beyond information exchange, developers participate in open-source projects to solve complex problems, with the goals of improving skills, enhancing software quality, and building professional reputation (Shah 2006, Kononenko et al. 2018). On the one hand, generative AI may reduce the need for deep coding expertise, potentially diminishing developers' incentives to contribute if they no longer view OSS as a valuable avenue for skill development or reputation building. On the other hand, by helping developers overcome technical challenges more efficiently, generative AI could lower barriers to contribution and support the broader goals of improving code quality and sustaining collaborative progress. As a result, it is unclear *ex ante* the overall impact of generative AI on participation in OSS development.

2.3 Generative AI in Team Coordination

While team collaboration can improve problem-solving and lead to higher-quality outcomes by leveraging diverse perspectives, it also introduces coordination challenges that can hinder team-level productivity and performance (Janardhanan et al. 2020, Mattarelli et al. 2022). In software development, collaboration is a core activity as developers jointly design and implement software development. The extent to which developers can effectively coordinate their activities has a direct impact on software team performance (Koushik and Mookerjee 1995, Reagans et al. 2016, Mawdsley et al. 2022). In the OSS setting, team

members are often globally distributed, participation is voluntary, and team composition is fluid. Without centralized task allocation or hierarchical oversight, developers must self-organize and manage interdependent tasks (Lindberg et al. 2024). This fluid and decentralized structure makes efficient coordination challenging in the OSS development environment.

Although some research has examined coordination in open-source projects (Medappa and Srivastava 2019), little is known about how emerging technologies, such as generative AI, shape coordination in open collaboration. Li et al. (2024) demonstrate that the use of generative AI does not enhance coordination within teams, largely due to the increased volume and diversity of ideas it introduces, which can hinder convergence. Conversely, Dell'Acqua et al. (2025) find that generative AI facilitates coordination during idea generation in cross-functional teams by supporting alignment across disciplinary boundaries. In both studies, team size is held constant, and coordination is operationalized as the collective development of a single, shared solution. As such, the observed coordination primarily reflects convergence toward a common outcome, rather than the integration of heterogeneous, individual contributions. Moreover, these studies do not examine how differences in participants' familiarity with the task may influence coordination performance. To the best of our knowledge, we are among the first to examine the impact of generative AI on team coordination in an open collaboration setting, where team size and composition are fluid and individuals are performing distinct coding tasks in a self-organizing manner. We focus specifically on coordination time for code integration, a key indicator of coordination efficiency that is critical for rapid iteration and timely delivery of software.

3. Theoretical Motivation

3.1 AI Pair Programmers and Project-level Code Contributions

Open-source software (OSS) development is characterized by a fully decentralized and open environment, where a group of voluntary software developers work together to develop and refine code, subsequently making it accessible to fellow developers and the broader community (Levine and Prietula 2014).

AI pair programmers, guided by explicit prompts from developers, offer immediate code suggestions and corrections. These tools support software developers by fixing bugs, proposing

enhancements, and facilitating the transfer of coding knowledge across various domains.¹ AI pair programmers use extensive databases and advanced machine learning algorithms to provide suggestions that follow best practices, reducing search time. This helps lower the time developers spend writing code and waiting for peer support, thereby accelerating the code production. Recent research has documented the beneficial effects of AI pair programmers on individual developers' productivity (Imai 2022).

Besides the positive effect on individual productivity as documented in the prior literature, we argue that AI pair programmers can also lead to more participation in the development of a focal OSS project (i.e., more OSS code contributors), due to the following reasons. According to the expectancy value models, developers evaluate both the expected outcomes and the value of participation when deciding whether to contribute to a project (Atkinson 1957, Hertel et al. 2003, Setia et al. 2012). Given their limited time, developers would select projects that are both meaningful and rewarding, balancing the costs of participation against the potential benefits (Wen et al. 2013). Traditionally, developers needed specialized knowledge to effectively contribute to OSS projects, creating significant entry barriers. AI pair programmers help reduce these barriers by shortening the learning curve and lowering the time and effort needed to contribute. This allows developers to contribute more easily to different projects. By lowering entry barriers and reducing the cost of participation, AI pair programmers could encourage more developers to participate in a focal OSS project.

Taken together, because AI pair programmers not only enable greater intensity of individual code contributions but also encourage more developers to participate in the project, we expect AI pair programmers to result in an increase in total code contributions to the project.

Hypothesis 1: AI pair programmers lead to an increase in project-level code contributions in OSS development.

3.2 AI Pair Programmers and Coordination Time

¹ The GitHub website discusses the use cases of Copilot: <https://github.com/features/copilot>. <https://github.blog/2022-09-14-8-things-you-didnt-know-you-could-do-with-github-copilot/>.

Unlike traditional software teams, OSS projects involve distributed developers and operate without formal hierarchies or centralized oversight, making coordination more essential for maintaining coherence and progress. As OSS teams grow in size and diversity, the complexity of integrating code contributions increases (Weber 2006, Feri et al. 2010, Yu et al. 2015, Roels and Corbett 2024). Differences in development styles, coding conventions, design philosophies and architectural decisions can exacerbate integration challenges and, if not adequately managed, may undermine team effectiveness (Ancona and Caldwell 1992, Janardhanan et al. 2020, Mattarelli et al. 2022). While many open-source projects adopt modular designs and parallel development strategies to reduce interdependencies (Howison and Crowston 2014), evidence suggests that significant code and developer interdependencies remain (Lindberg et al. 2016). These interdependencies underscore the need for effective coordination to align efforts, resolve conflicts, and ensure that individual contributions integrate effectively into the existing codebase.

Therefore, besides making code contributions to the codebase, developers also spend a significant amount of time and efforts managing project-level dependencies and ensuring cohesive and aligned actions among team members. Coordination theory highlights communication as a central mechanism (Malone and Crowston 1994, Pikkarainen et al. 2008), as communication enables teams to align objectives, share progress updates, resolve interdependencies, and establish a common ground (Okhuysen and Bechky 2009). This is especially critical in interdependent task environments, where coordination requires continuous adjustments and information exchange (Srikanth and Puranam 2014, Oliveira and Lumineau 2017, Im and Ahuja 2023). In OSS settings, online discussions serve as a primary channel for communication (Roberts et al. 2006), helping developers clarify project goals, receive feedback on contributions, and resolve integration conflicts (Harbring 2006). Motivated by these observations, in this study, we focus on coordination time for code integration, which is often spent on discussing code and resolving conflicts and different ideas among developers (Howison and Crowston 2014, Lindberg et al. 2016, Medappa and Srivastava 2019, Shaikh and Vaast 2023).

Besides encouraging coding activities as discussed above, AI pair programmers may also encourage more participation in discussions on the focal project, which could in turn lead to longer

coordination time for code integration, for reasons as follows. AI pair programmers can assist developers in understanding algorithms and exploring alternative solutions (Barke et al. 2023). This improved code comprehensibility increases individual developers' capacity to contribute comments and participate in discussions. It also lowers the barrier to participation for those who might otherwise be discouraged by unfamiliar or intricate code. In addition, by automating repetitive coding tasks, AI pair programmers enable developers to have more time to provide feedback and participate in discussions about implementation and design decisions. As more developers participate in code discussions with each developer contributing more thoughts and ideas, it could potentially become more difficult to align perspectives and reach consensus. As a result, given the same amount of code contributions to an OSS project, AI pair programmers could potentially lead to longer coordination time for the code to be integrated into the codebase of the project.

Hypothesis 2: AI pair programmers lead to an increase in coordination time for integrating code contributions in OSS development.

3.3 Core and Peripheral Developers

The OSS development process involves two categories of developers: core and peripheral (Setia et al. 2012). Core developers, who often include the projects' administrators and key maintainers, are responsible for defining the overarching objectives and ensuring the delivery of the final code product. As shown in Figure 1, software development usually begins with core developers designing the primary codebase and hosting it on a platform for open collaboration (Singh and Phelps 2013). With the write access and full control over the projects, core developers can submit code and either directly integrate it into the primary codebase or have it reviewed by other core developers for integrity.

On the other hand, peripheral developers introduce additional perspectives shaped by different domains, use cases, and user needs. Peripheral developers typically participate on a more limited basis and often contribute across multiple projects (Howison and Crowston 2014, Krishnamurthy et al. 2016). They mainly focus on enhancing the existing codebase such as fixing bugs and adding new features (Setia et al. 2012). Meanwhile, without the write access to codebases, peripheral developers' code submissions need to

be reviewed by core developers, who decide whether to accept the code changes, request additional modifications, or reject them (Gousios et al. 2014, Medappa and Srivastava 2019).

Because of the different roles and responsibilities assumed by core versus peripheral developers, a key distinction between core developers and peripheral developers lies in their level of contextual knowledge about the OSS project, which refers to understanding of a software project on aspects such as its overall design, architecture, coding norms and practices (labeled as “project familiarity” for simplicity). In particular, due to their intensive involvement with the project, core developers typically possess higher familiarity with the project’s architecture, decision history, and development norms. On the other hand, because peripheral developers often participate on an irregular basis and complete specific tasks such as fixing bugs and adding patches, they tend to have more limited familiarity with the project.

This difference in project familiarity becomes particularly relevant when considering the use of AI pair programmers. While AI pair programmers can facilitate code generation, they may not be able to automatically fully incorporate a software project’s specific context, such as broader architectural structure, interdependencies, performance considerations, into the code they generated (Feldman et al. 2023, Liguori et al. 2024, Piñeiro-Martín et al. 2025). Human developers remain essential for interpreting high-level design intent, validating architectural coherence, and ensuring that code generated by the AI tools is compatible with the overall OSS project vision. In this context, the developer’s familiarity with the project becomes a critical resource in maximizing the effective use of AI pair programmers.

Given the difference in project familiarity between core developers and peripheral developers and the limitations of AI pair programmers, we expect these two types of developers may realize different levels of productivity gain from AI pair programmers. In the following sub-sections, we first discuss how the impact of AI pair programmers on project-level code contributions may vary between those from core and peripheral developers, followed by discussions on how the impact on coordination time for code integration could differ for code submitted by these two types of developers.

3.3.1. AI Pair Programmers and Code Contributions for Core vs. Peripheral Developers

As noted above, core developers play a central role in defining the project's architecture and guiding its long-term development (Crowston et al. 2006). Given their familiarity with the project's structure, coding conventions, and historical evolution, core developers may be able to craft more effective prompts and critically evaluate AI-generated code within the context of project requirements. Moreover, due to the productivity gain enabled by AI pair programmers, core developers may be more willing to contribute; they may also be more motivated to reallocate their time and efforts towards more coding activities from project management activities (Hoffmann et al. 2024). As a result, the individual code contributions made by a core developer to a focal project could be significantly boosted by AI pair programmers.

If AI pair programmers can effectively reduce the time and effort required for core developers to code, the cost of participating in coding activities could also be lowered. Accordingly, a focal OSS project may attract more core developers aided by these AI tools to participate. Thus, we expect AI pair programmers could lead to a significant increase in total project-level code contributions made by core developers. In contrast, because peripheral developers tend to have limited contextual knowledge about a focal OSS project, they may not be able to realize as much productivity gain from AI pair programmers as core developers, for reasons as follows. First, they may struggle to frame contextually appropriate prompts to generate code. Since AI-generated code is highly dependent on the quality and specificity of the prompts (Feldman et al. 2023, Liguori et al. 2024, Piñeiro-Martín et al. 2025), peripheral developers may generate less relevant or misaligned code. Second, because peripheral developers tend to lack comprehensive knowledge of its architectural decisions, development history, and coding norm, they may not recognize when AI outputs violate project-specific conventions. As a result, their code generated through AI tools is more likely to require revision, delay integration, or be rejected (Zhang et al. 2022).

Therefore, while AI pair programmers could benefit peripheral developers by facilitating their code generation and reducing their cost of participation in an OSS project, the benefit could be constrained by their limited ability in leveraging the AI tools due to their lack of familiarity with a focal project. This may not only lead to a smaller increase in individual code contributions but also a smaller increase in their incentive to participate, compared to core developers. Overall, we expect AI pair programmers lead to a

smaller increase in total project-level code contributions made by peripheral developers than those from core developers.

Hypothesis 3a: AI pair programmers lead to a smaller increase in project-level code contributions made by peripheral developers compared to core developers.

3.3.2. AI Pair Programmers and Coordination Time for Core vs. Peripheral Developers

We have argued above that AI pair programmers may encourage developers to participate in code discussions more, which could lead to longer coordination time for code integration. We further expect that the extent of such an increase in coordination time could also depend on the code being submitted. Because of the difference in project familiarity between peripheral developers and core developers, we expect the increase in coordination time could be even bigger for code contributions made by peripheral developers than those made by core developers, for reasons as follows.

When AI pair programmers assist peripheral developers, the limited contextual knowledge held by peripheral developers suggests that they may generate code that is semantically unclear or misaligned with the project, as generative AI relies on rich contextual information to perform effectively (Feldman et al. 2023, Liguori et al. 2024, Piñeiro-Martín et al. 2025). This is likely because the AI tools generate code based on generalized training data rather than project-specific patterns, and peripheral developers often fail to provide complete context in their prompts. Although AI pair programmers can help peripheral developers produce code quickly, the output may potentially be inconsistent with the implicit design principles, create integration challenges, or require substantial clarification. These issues could lead to many comments from other developers to request the focal peripheral developer to clarify design intent, align code with norms, and resolve ambiguities (Gousios et al. 2014). Consequently, their code is more likely to prompt extended discussions, leading to relatively long coordination time.

In contrast, core developers' deep project familiarity allows them to better leverage AI pair programmers to generate code that can be integrated well into the existing system. As a result, for core developers, their code contributions facilitated by AI pair programmers may require fewer revisions or

clarifications. This suggests less coordination time would be needed for integrating their code contributions into the codebase.

Hypothesis 3b: AI pair programmers lead to a larger increase in coordination time for integrating code contributed by peripheral developers compared to the code contributed by core developers.

4. Data

We use GitHub data for our empirical analysis where a repository serves as the basic unit for collaborative software development (Dabbish et al. 2012). We investigate how the introduction of AI pair programmer GitHub Copilot impacts project-level code contributions and coordination time for GitHub repositories. Copilot's introduction occurred in stages, commencing with limited availability in June 2021 and a formal public launch in June 2022². During the one-year period between June 2021 and June 2022, Copilot underwent a technical preview phase. Based on this timeline, we collect panel data of GitHub repositories from GitHub Archive Dataset, spanning a two-year timeframe, from January 2021 to December 2022³.

To ensure the generalizability of our analysis, we follow established procedures in the literature (Kalliamvakou et al. 2014, AlMarzouq et al. 2020) to identify active OSS repositories during our panel period. We select repositories with non-zero size, at least one specified programming language and license, a description, and no mirror or personal store designation. To exclude ghost or abandoned projects, we require at least one code submission every six months from 2021 to 2022 and at least one additional activity, such as a release or creation, each year. As we are interested in evaluating project-level code contributions and coordination time involving multiple developers, we focus on repositories with at least three developers contributing each month. Additionally, as discussed in greater detail below, we use IDE⁴ information to

² GitHub launched the technical preview of Copilot in June 2021: <https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/>. It then announced the formal launch and public availability of Copilot in June 2022: <https://github.blog/2022-06-21-github-copilot-is-generally-available-to-all-developers/>.

³ We choose this endpoint to ensure our results are not influenced by the rise in popularity of Chat GPT, which began in early 2023.

⁴ IDE stands for integrated development environment, which is a software application that provides local environments for coding, testing, and debugging.

identify repositories in the treatment group versus those in the control groups, so we further restrict our sample to those that disclose IDE information. These criteria result in a sample of 9,244 repositories.

To identify repositories where Copilot was used by developers (i.e., the treatment group), we collaborated with GitHub organization, which provided proprietary aggregated Copilot usage data at the repository level. This dataset indicates the monthly proportion of developers contributing code to a given repository who also used Copilot.⁵ Since Copilot requires a compatible IDE, we also consider IDE usage. During our analysis period, only a limited number of IDEs supported Copilot: Visual Studio Code, the JetBrains suite of IDEs, Neovim, and Visual Studio.⁶ To determine IDE usage, we examine the webpages of the repositories in our sample to gather information on the IDE usage by contributing developers. We categorize repositories with developers who used Copilot and one of the supported IDEs as our treatment group, and those using unsupported IDEs and not using Copilot as our control group. In total, our sample includes 5,687 repositories in the treatment group and 3,557 repositories in the control group. We designate the first month when Copilot was supported by IDEs and used by a non-zero proportion of developers as the treatment start time for each repository. The months before this time are defined as the pre-treatment period and the months after this time as the post-treatment period for each repository, covering the two-year span from 2021 to 2022.

To further improve comparability, we restrict the sample to repository-month observations with at least one code contribution. This focus allows us to estimate Copilot's effect on collaborative development in actively maintained projects. Moreover, because the analysis of coordination time requires at least one code contribution to compute acceptance time, this criterion allows us to use the same sample to investigate H1 and H2. The final sample includes 7,637 repositories, with 4,491 in the treatment group and 3,146 in the control group. Table 1 provides descriptions and summary statistics for repository-month-level variables used in the main analysis. We check the robustness of our results by removing the restriction on the number

⁵ We do not know the identity of individual Copilot users because of privacy concerns. Although such Copilot usage is measured at the GitHub platform level, developers are likely to integrate Copilot into their workflows extensively and thus would use it for all repositories where it is feasible (Marangunić and Granić 2015).

⁶ GitHub lists the supported IDEs for Copilot: <https://github.com/features/copilot>.

of code contribution. Additionally, we test the robustness using a larger sample that includes repositories with as few as two developers.

[insert Table 1 here]

5. Empirical Analyses

5.1 Measures

To test H1 and H2, the objective in our empirical analyses is to determine how Copilot influences project-level code contributions and coordination time for code integration. To measure project-level code contributions, we use the total number of pull requests (PRs) submitted to each repository that were eventually merged⁷. PRs represent code changes submitted by developers and need further evaluation by core developers. This evaluation results in either approval, leading to merged PRs, or rejection, leading to closed but unmerged PRs. Thus, merged PRs reflect successful code changes that have been eventually incorporated into the development of repositories and are commonly used in the literature to assess meaningful contributions by developers (Gousios et al. 2014, Tsay et al. 2014, Kononenko et al. 2018).

To evaluate coordination time for code integration, we follow prior work in both economics (Simcoe 2012) and software engineering (Espinosa et al. 2007, Yu et al. 2015, El Mezouar et al. 2019) by measuring the duration between the submission and acceptance of each PR. Because a repository may receive many PRs in a month, we then compute the average duration (i.e., average time) to merge a PR for a repository in a month.

5.2 Model and Estimation

In our setup, repositories adopt Copilot at different times after it is available. The variation in treatment timing and the larger number of treated repositories compared to untreated ones complicate the application of traditional matching techniques in this context. Given these complexities, we adopt the Generalized Synthetic Control Method (GSCM) (Xu 2017), which allows for improved estimation of treatment effects by generating weighted synthetic controls that closely resemble the treated units prior to treatment. This

⁷ More specifically, we consider all PRs initiated in a month that get merged anytime between that month and 12/2023 as code contributions for that month and call these as merged pull requests.

approach combines the idea of a synthetic control (Abadie et al. 2010) with interactive fixed effects (Bai 2009). This allows us to address multiple treated units with staggered treatment times while accounting for time-varying unobservables. The GSCM shares core assumptions with the standard synthetic control method, effectively managing unobserved time-variant confounders by giving more weight to control units that mirror the pre-treatment trends of the treatment group (Xu 2017). This approach offers improved pre-treatment control and more credible parallel trends relative to traditional difference-in-differences (DID) and matching methods and has been shown to exhibit lower bias (Xu 2017, Wang et al. 2021).

We use the following linear factor model (Bai 2009, Xu 2017) with the unit of analysis at the repository-month level:

$$Y_{it} = \delta_{it} D_{it} + X'_{it}\beta + \lambda'_i f_t + \varepsilon_{it} \quad (1)$$

where Y_{it} represents the outcome variable, which is either the project-level code contributions measured by total number of merged PRs of repository i in month t , or coordination time for code integration measured by the average time taken to merge a PR of repository i in month t . We take log transformation to reduce the skewness of both variables. D_{it} is the treatment indicator which equals one if repository i has been developed with Copilot by month t and zero otherwise. The parameter of primary interest is δ_{it} , which signifies the dynamic impact of Copilot on log number of merged PRs and log average time taken to merge a PR. δ_{it} is heterogeneous treatment effect and its subscripts i and t indicate that the estimates vary across repositories and months.

In the above equation, $f_t = [f_{1t}, \dots, f_{rt}]'$ is an $(r \times 1)$ vector of unobserved common factors associated with factor loadings $\lambda'_i = [\lambda_{i1}, \dots, \lambda_{ir}]$ and ε_{it} is the error term with a mean of zero. The factor component $\lambda'_i f_t$ can be expressed as $\lambda'_i f_t = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \dots + \lambda_{ir} f_{rt}$. The factor component nests a range of unobserved heterogeneities including additive unit and time fixed effects and unit-specific linear and quadratic time trends. Note that a two-way fixed effects specification is a special case of the factor component, where $r = 2$, $f_{1t} = 1$, and $\lambda_{i2} = 1$, so that $\lambda'_i f_t = \lambda_{i1} + f_{2t}$. Here, λ_{i1} represents the

repository fixed effects, while f_{2t} is the month fixed effects. We specify the two-way fixed effects to account for heterogeneity across repositories and time, while considering other unobserved latent factors.

We estimate the optimal number of latent factors using a cross-validation procedure (Xu 2017). Briefly, this involves first estimating the parameters of model using the control group data only and employing a cross-validation procedure to determine the number of latent factors r . Next, the optimal number of factor loadings for each treated unit is estimated by minimizing the mean squared errors of the predicted treated outcomes in the pre-treatment periods. We provide a detailed explanation of the model and the estimation including the determination of the optimal latent factors in Online Appendix A. After accounting for time and repository fixed effects, the optimal number of unobserved factors determined by the cross-validation technique is zero. This suggests that the fixed effects setting has effectively accounted for any unobserved time-varying characteristics (Xu 2017).

The GSCM estimator predicts the counterfactuals for treated units in the post-treatment periods using the parameter estimates obtained in the previous two steps. The causal effect of the treatment is calculated as the average treatment effect on the treated (ATT), based on the differences between the observed outcome of a treated unit $Y_{it}(1)$ and its constructed counterfactual $\hat{Y}_{it}(0)$:

$$ATT_t = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} [Y_{it}(1) - \hat{Y}_{it}(0)] = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \delta_{it} \quad (2)$$

where \mathcal{T} denotes the set of treated units and $|\mathcal{T}|$ represents the number of units in \mathcal{T} .

Identification: Repository fixed effects account for static unobservable differences across repositories such as the characteristics of the projects or the developers that can influence code contribution to these repositories and the required coordination time. Additionally, time fixed effects account for the common time trends that could influence the outcomes. There may be dynamic unobservables which could influence the outcomes. GSCM synthesizes a weighted control unit that closely mirrors the data pattern of the log number of merged PRs and the log average time taken to merge a single PR during the pre-treatment period for the treated unit. The outcome variable of this synthetic control unit during the post-treatment period serves as the counterfactual prediction for the treated unit. By modeling the trend of the outcome

variables, the GSCM can naturally accommodate the influence of unobservable confounders that evolve over time. Furthermore, it allows each treatment unit to have a different treatment period and can efficiently construct synthetic control units from a relatively small control sample.

To assess the validity of results, we conduct equivalence tests to examine the presence of any pre-treatment trends. The results of these tests are reported in Online Appendix B and suggest that there are no significant pre-treatment differences in the outcomes between treated and corresponding synthetic control repositories. Additionally, we perform placebo tests and visualize the time-varying treatment effects across pre-treatment and post-treatment periods to ensure the robustness of our findings. Detailed results of these analyses are presented in Online Appendix C. We also validate our results using different samples, matching technique, and alternative estimation as described in the following sections.

6. Results

6.1 Main Results

Table 2 presents the estimated overall effects of Copilot adoption on project-level code contributions and coordination time for code integration. We find that, following Copilot adoption, the number of merged PRs of active repositories increased by 5.9%, while the average time required to merge a single PR increased by 8%. These results indicate that, following Copilot adoption, project-level code contributions significantly increased, but coordination time for code integration also became longer.

The increase in project-level code contributions may be driven by two key mechanisms. First, Copilot might increase individual code contributions by providing useful code suggestions and auto-completions, thereby streamlining the coding processes and accelerating code submissions. Second, Copilot may reduce the participation barriers and costs for developers, enabling more developers to contribute to repositories and thus leading to an increase in project-level code contributions.

Meanwhile, the observed increase in coordination time could be attributable to increased code discussions. As discussed earlier, generative AI tools such as Copilot may improve code comprehension, enabling developers to propose and discuss more suggestions. Additionally, these AI tools may free up time

for developers to participate more actively in code discussions, leading to longer coordination time. We explore these potential drivers in detailed mechanism analyses in Section 6.3.

[insert Table 2 here]

6.2 Robustness Checks

To further investigate the robustness of our findings, we employ a range of empirical methods to examine the relationship between Copilot use and the project-level code contributions and coordination time for code integration. A summary of these checks is provided in Table D.1 of Online Appendix D.

Within-IDE Analysis: In our baseline identification, we use IDEs that supported Copilot during our sample period as part of the criteria to identify repositories in the treatment group and consider repositories developed with IDEs that did not support Copilot for the control group. To mitigate concerns regarding potential unobservable differences in IDE usage, we conduct the analysis by only focusing on repositories using Copilot-supported IDEs and further restricting the treatment versus control group comparison within the same Copilot-supported IDE. This approach addresses three key sources of potential bias. First, it eliminates concerns about differences between repositories using supported versus unsupported IDEs, the latter of which may be associated with lower popularity. Second, it controls for variation across different supported IDEs, which may attract distinct types of projects or developer teams. Third, it accounts for the possibility that developers who choose IDEs compatible with Copilot may differ from those who do not. By focusing on comparing repositories developed within the same Copilot-supported IDE, we ensure that both treatment and control groups share the same development environment and that contributing developers are likely to have similar preferences and characteristics. Therefore, the only variation across groups lies in the timing of Copilot adoption.

Specifically, our first step is to identify the sample of repositories using Copilot-supported IDEs. Then, in order to identify treatment and control groups, we use repositories that adopted Copilot in the first half of the period from July 2021 to December 2022 when Copilot was available, i.e., July 2021 - March 2022, as the treatment group. Accordingly, we conduct this analysis using only the period from January 2021 to March 2022, so repositories that adopted Copilot after March 2022 can be considered as the control

group. We choose March 2022 as the endpoint so that there are enough repositories with Copilot usage during this time window (i.e., a total of nine months from July 2021 to March 2022). Our second step is to match treatment and control repositories within the same IDE. After matching, we use the pooled sample to run the GSCM model. This analysis includes 3,215 repositories in the treatment group and 1,247 repositories in the control group, a total of 4,462 repositories. The treatment turn-on time for each repository in the treatment group is decided in the same manner as in our baseline analysis above. The results based on estimation using GSCM in column (1) of Tables 3 and 4 show that Copilot increased the number of merged PRs and the average time taken to merge a single PR, which are consistent with our main findings.

Refined Sample: As developers work on multiple repositories, it is possible that some may be working on both treatment and control repositories. Approximately 3.5% of the selected repositories in our main analysis have developers involved in both groups. Although these developers might not be using Copilot for the control repositories due to IDE restrictions, there is potential for knowledge transfer, which could bias our results. To eliminate this bias, we refine our sample to exclusively assess the impact of Copilot on repositories where no developer is involved in both groups. Corresponding results are shown in column (2) of Tables 3 and 4, indicating that Copilot increased the number of merged PRs and the average time taken to merge a single PR, which are qualitatively similar to our main results.

There may be spillover effects wherein repositories not using Copilot could still benefit if their developers have prior experience with Copilot from other projects. In such cases, our estimates would represent a lower bound on the true impact of Copilot on code contributions and coordination time for code integration. The observed increase in code contributions would likely be greater if developers in the control repositories had no prior exposure to Copilot. Similarly, the increase in coordination time would be more pronounced when compared to repositories unaffected by such spillover effects.

Outlier Removal: To rule out the possibility that extreme values are disproportionately influencing the results, we conduct an analysis that excludes repositories exhibiting unusually high or low values in terms of the number of merged PRs and the average time taken to merge a single PR. Such extreme observations may arise due to idiosyncratic project dynamics. By removing these outliers, we ensure that

our results reflect generalizable trends rather than being driven by a few atypical cases. The results, reported in column (3) of Tables 3 and 4, remain consistent with our main findings.

Non-AI Topic Projects: Another potential concern is that the observed effects may be driven by project-specific enthusiasm associated with certain topics, particularly those related to artificial intelligence. AI-related repositories may attract different groups of developers compared to repositories focused on other topics, which could confound the estimated effect of Copilot adoption. To address this concern, we conduct a robustness check by excluding repositories associated with AI-related topics and re-estimate the models using the remaining sample. This approach allows us to test whether the impact of Copilot is generalizable across a broader range of repository topics and not limited to those inherently aligned with AI development. As shown in column (4) of Tables 3 and 4, the results remain consistent with our main findings.

PSM and DID: We further validate our results using an alternative matching and estimation approach. Specifically, we use the difference-in-differences (DID) estimation combined with the propensity score matching (PSM). In this analysis, we use June 2021, the first availability date of Copilot, as the single treatment turn-on time for all repositories classified into the treatment group in the baseline analysis. Hence, this approach may alleviate concerns about non-random treatment turn-on time for Copilot usage in the treatment group in the main analysis. However, we acknowledge that because developers may sign up gradually after the initial availability date, this approach may consider the period when Copilot was actually not used as the post-treatment period, leading to an underestimation of the positive effect.

We calculate each repository's propensity score, defined as the probability of adopting Copilot, using a logit regression model. To construct the covariates for matching, we calculate the monthly values of several repository characteristics over the pre-treatment period (January to June 2021), average them across the six-month window, and then apply a log transformation. Specifically, we use the log number of merged PRs, the log average time taken to merge a single PR, the log number of unique developers whose PRs have been merged, the log number of push events, the log number of release events, the log number of opened issues, and the log number of closed issues.

Next, we employ the nearest-neighbor matching algorithm without replacement and set a caliper of 0.05 to match each treated repository with a control unit. Following this matching process between repositories where Copilot was used and those that Copilot was not used, we arrive at a dataset comprising 2,122 repositories, consisting of 1,061 treatments and 1,061 controls. We report the details of the balance checks of our matched sample in Online Appendix E. Overall, there is no statistically significant differences between the treatment and control groups across these pre-treatment covariates after matching.

Based on the matched sample, we analyze the impact of Copilot on the number of merged PRs and the average time taken to merge a single PR using the following regression specification:

$$Y_{it} = \beta_0 + \beta_1 \text{Copilot}_i \times \text{Post}_t + \alpha_i + \mu_t + \varepsilon_{it} \quad (3)$$

where the dependent variable Y_{it} represents the log number of merged PRs and the log average time taken to merge a PR of repository i in month t . Copilot_i is a binary indicator that is denoted with “1” for repository i where Copilot was used and “0” otherwise. Post_t is a binary indicator that is set to “1” in months after June 2021 and “0” otherwise. To control for time-invariant heterogeneity across repositories and time trends, we include repository-level fixed-effects α_i and month-level fixed-effects μ_t .

The main coefficient of interest is β_1 , as it indicates the change in the number of merged PRs or the change in the average time taken to merge a PR of repositories before and after the availability of Copilot, relative to the changes of repositories where Copilot was not used at all throughout the sample period. Additionally, we confirm that the parallel trend assumption holds for the DID model (details of the parallel pre-trend test conducted using the relative time model are provided in Online Appendix F). We report the results in column (5) of Tables 3 and 4. The results show that Copilot significantly increased the number of merged PRs and the average time taken to merge a single PR, which are qualitatively consistent with our findings from the main analysis.

[insert Table 3 here]

[insert Table 4 here]

Alternative Measures: In the main analysis, we measure project-level code contributions using merged PRs and coordination time using hours between PR submission and acceptance. To ensure our results are not driven by specific measures, we examine alternative measures, including submitted PRs and commits for project-level code contributions, and time intervals in days and minutes for coordination time. Results in Tables G.1 and G.2 of Online Appendix G remain consistent, indicating our findings are robust to outcome variable definitions.

Effect of Workload: One important alternative explanation for the observed increase in coordination time for merging a PR following Copilot adoption is that there were simply more PRs that needed to be reviewed, which could lead to some delay. To address this concern, we include the log number of cumulative submitted PRs till the focal month as a control variable to account for potential increases in overall workload when analyzing coordination time. The results, reported in Tables G.3 of Online Appendix G, are qualitatively similar to our main results for coordination time.

Alternative Estimation: We also validate our results using a two-way fixed effects model applied to the full sample. The findings, presented in Table G.4 of Online Appendix G, are qualitatively consistent with our main analysis.

Expanded Sample: Because our main sample includes repositories with three or more developers, we also examine the robustness of our results using an expanded sample that includes repositories with two developers. Results are reported in Table G.5 of Online Appendix G and are qualitatively similar to our main analysis.

Additional Alternative Sample: In our main analysis, we restrict the sample to repository-month observations in which at least one PR was eventually merged. To assess the robustness of our findings, we replicate the analysis by including all repository-month observations, regardless of whether any PRs were merged. As shown in Table G.6 of Online Appendix G, the results remain consistent with those from our main analysis. However, it is important to note that we cannot extend this analysis to coordination time as measuring coordination time requires data on PR acceptance time, which is only available when PR is merged.

6.3 Mechanism Analyses

6.3.1 Project-level Code Contributions

In our hypothesis development for H1, we argue that AI pair programmers could influence project-level code contributions by increasing individual code contributions and encouraging more developers to participate in coding. To shed light on these potential mechanisms, we conduct the following analyses.

Individual Code Contributions: The increase in project-level code contributions could be driven by a higher level of individual code contributions, as AI pair programmers support software development activities of individual developers. To explore this possibility, we replicate our main analysis using the average number of merged PRs per developer as the dependent variable (Subramaniam et al. 2009, Medappa and Srivastava 2019). As shown in column (1) of Table 5, Copilot use is associated with a 2.1% increase in average number of merged PRs per developer in active repositories. At the same time, as shown in column (1) of Table H.1 of Online Appendix H, the volume of individual code contributions is positively correlated with project-level code contributions. These results collectively support our argument that AI pair programmers could lead to an increase in project-level code contributions by increasing individual code contributions.

Developer Coding Participation: If AI pair programmers encourage more developers to contribute to repositories, this could also increase project-level code contributions. To test this potential mechanism, we measure developer's coding participation by counting the number of unique developers whose PRs have been merged into the primary codebase (Subramaniam et al. 2009, Medappa and Srivastava 2019). We conduct the analysis using the same GSCM model as discussed above with developer's coding participation as the dependent variable. The results, shown in column (2) of Table 5, indicate that Copilot increased the number of developers whose PRs have been merged by 3.4%. Meanwhile, as shown in column (2) of Table H.1 of Online Appendix H, developer coding participation is positively correlated with project-level code contributions. These results are consistent with our argument that AI pair programmers could increase developer coding participation by reducing the participation costs developers face when deciding whether

to contribute (Atkinson 1957, Hertel et al. 2003, Setia et al. 2012, Wen et al. 2013). As a result, project-level code contributions increase.

[insert Table 5 here]

6.3.2 Coordination Time

In our hypothesis development for H2, we suggest that AI pair programmers may promote more participation in discussions on the focal OSS project, which could lead to longer coordination time for code integration. To probe this potential mechanism, we analyze how AI pair programmers influence code discussions for each merged PR. Moreover, because more code discussions for each merged PR could be driven by more participation by developers into code discussion, higher code discussion intensity per developer, or both, we further examine how AI pair programmers influence each of these factors.

Code Discussion Volume: To provide some direct evidence on this mechanism, we re-estimate our baseline model using code discussion volume as the outcome variable. We measure code discussion volume using the average number of total comments per merged PR (Cataldo et al. 2006, Oh and Jeon 2007, Gousios et al. 2014). As shown in column (1) of Table 6, Copilot adoption is associated with a 6.5% increase in code discussion volume per merged PR. Moreover, as shown in column (1) of Table H.2 in Online Appendix H, we observe a positive correlation between code discussion volume and coordination time. These findings support our argument that AI pair programmers may promote more code discussions, which in turn leads to longer coordination time.

Code Discussion Participation: Increased participation by developers in code discussion can introduce more perspectives, thereby extending the time required to reach consensus (Weber 2006, Feri et al. 2010, Yu et al. 2015, Roels and Corbett 2024). To explore whether Copilot impacts code discussion participation, we measure code discussion participation using the average number of unique developers who left comments per merged PR (Oh and Jeon 2007). Then we employ the same GSCM model using code discussion participation as the dependent variable. Column (2) of Table 6 presents the results, showing a significant 1.7% increase in code discussion participation. Meanwhile, column (2) of Table H.2 in Online Appendix H suggests a positive correlation between code discussion participation and coordination time.

These results suggest that AI pair programmers could encourage broader participation in code discussions, which leads to longer coordination time.

Code Discussion Intensity: AI pair programmers may also affect the intensity of code discussion made by each developer on a merged PR, which could contribute to more code discussion volume and coordination time for code integration. To test this, we measure code discussion intensity using the average number of comments per developer per merged PR (Oh and Jeon 2007) and use it as the dependent variable based on the same GSCM model as our baseline model. As shown in column (3) of Table 6, repositories using Copilot show a 5.9% increase in code discussion intensity per developer. Column (3) of Table H.2 in Online Appendix H also shows a positive correlation between code discussion intensity and coordination time. These results indicate that AI pair programmers could encourage each individual developer to participate more in code discussion, resulting in longer coordination time.

[insert Table 6 here]

We further examine how Copilot affects the content of code discussions among developers. Specifically, we conduct a supplementary text analysis of topic diversity using Latent Dirichlet Allocation (LDA) to extract discussion topics from 5,404,735 merged PR comments of our selected repositories in the main analysis from 2021 to 2022. Details are provided in Online Appendix I. As shown in Table I.1, Copilot increased the diversity of code discussion topics per merged PR, suggesting that Copilot also expanded the range of topics being discussed, further increasing coordination time.

Alternative Explanation: One alternative explanation for the increased coordination time is that core developers, who are responsible for reviewing and merging the code, may be allocating less effort to these activities after adopting Copilot, as they become more engaged in contributing code themselves. To investigate this possibility, we examine three measures of core developer review activity on merged pull requests (PRs): (1) the total number of reviews finished by core developers, which reflects the overall volume of review work; (2) the number of unique core developers who reviewed code submissions for merged PRs, which indicates the breadth of participation; and (3) the average number of reviews per core developer, which captures review productivity of each core developer. As reported in Table 7, we find that

Copilot adoption is associated with an increase in all three measures. Although Copilot leads to greater code contribution volume, it also appears to enhance review productivity and encourage broader participation in review among core developers. These findings suggest that, rather than reducing their review activity, developers are completing more reviews per month following Copilot adoption. Therefore, the increased coordination time for code integration does not seem to be driven by less effort spent on reviewing the code.

[insert Table 7 here]

6.4 Combined Effects of Project-level Code Contributions and Coordination Time

The results in the previous sections indicate that AI pair programmers may increase both project-level code contributions and coordination time for code integration. Thus, the combined effects on project-level productivity, defined as timely merge of code contributions into the codebase of a project, remain unclear *ex ante*. To provide some empirical evidence on this, we consider the number of PRs that were merged within a particular time window as a measure of project-level productivity. For robustness, we use three time windows based on the distribution of merge times prior to Copilot's introduction: (1) the 25th percentile (1 day), (2) the median (3 days), and (3) the 75th percentile (10 days). Then, we calculate the number of PRs merged within each of the three time windows.

The results, presented in Table 8, indicate that Copilot adoption is associated with a 3.5% increase in PRs merged within one day, a 4.1% increase in PRs merged within three days, and a 5.1% increase in PRs merged within ten days. As reported in Table H.3 of Online Appendix H, we further confirm a positive correlation between project-level code contributions and project-level productivity, and a negative correlation between coordination time and project-level productivity. These findings suggest that, despite the increased coordination time, the effect of Copilot on overall project-level productivity is positive.

[insert Table 8 here]

Impact on Code Quality: While the use of Copilot increased productivity at the project level, it is equally important to understand its impact on code quality. On one hand, Copilot may reduce errors and improve the quality of code written by individual developers. On the other hand, it could potentially lower

overall quality if it encourages increased participation from less-skilled developers. Thus, ex-ante, the effect of Copilot on quality is theoretically ambiguous.

To examine whether and how Copilot affects code quality, we include four additional dependent variables related to issue reports. First, we measure the total number of issues reported in each repository. Second, since issues can vary in nature and severity, we focus specifically on bug-related issues, which are more severe and widely used in the software engineering literature as a proxy for quality (Vasilescu et al. 2015). Third, we construct two normalized measures: the average number of total issues per merged PR and the average number of total bug-related issues per merged PR. These normalized metrics allow us to assess quality relative to the volume of contributions. As reported in Table 9, we find that although Copilot adoption is associated with an increase in the total number of issues and bugs, the normalized measures remained statistically unchanged. These findings suggest that the increased issues and bug reports are due to a higher volume of code contributions, rather than a reduction in the quality of individual contributions.⁸

[insert Table 9 here]

6.5 Core and Peripheral Developers

Our H3a and H3b highlight some differential effects of AI pair programmers between peripheral and core developers. To assess the effect of AI pair programmers on the relative project-level code contributions made by peripheral developers compared to core developers (H3a), we compute the proportion of merged PRs from peripheral developers to the total number of merged PRs in each repository.⁹ A negative treatment effect estimate on this outcome variable indicates that after using Copilot, the proportion of code generated by peripheral developers among all code merged in a repository became lower, i.e. the adoption of AI pair programmers led to a relative decline in contribution share from peripheral developers.

⁸ We also examine whether Copilot's effects differ by project complexity and find no significant differences between simple and complex projects. Detailed discussions of the results are in Online Appendix J.

⁹ Note that there are only two types of developers in a repository: core or peripheral developers. We use this measure instead of the ratio of peripheral developers' merged PRs to core developers' merged PRs because the latter measure does not allow us to incorporate cases where core developers had zero merged PR in a repository in a month.

To assess the relative change in coordination time for integrating code contributed by peripheral developers compared to the code from core developers (H3b), we calculate the ratio of the average time to merge peripheral developers' PRs to the overall average code-merging time in each repository. A positive estimate on this outcome would indicate that PRs from peripheral developers faced relatively longer coordination time after Copilot adoption.

We next examine the difference in productivity gain from AI pair programmers between peripheral developers and core developers. Specifically, we compute the proportion of PRs from peripheral developers that were merged within one, three, or ten days, and use each of these as the dependent variable. Negative estimates based on these outcome variables suggest that AI pair programmers led to less productivity gain for peripheral developers than for core developers.

As shown in column (1) in Table 10, the proportion of merged PRs from peripheral developers significantly declined by 0.019 or 3.7%¹⁰. Additionally, as shown in column (2), peripheral developers experienced an increase of 0.085 or 5.4%¹¹ in their relative average merge time. Collectively, these findings suggest that Copilot led to relatively fewer project-level code contributions from peripheral developers and longer coordination time for them. The results, as well as the ones in columns (3) to (5) that use the proportion of timely merging of peripheral developers' code as the dependent variables, suggest peripheral developers realized less productivity gain from Copilot than core developers.

[insert Table 10 here]

Table 11 next reports the results on the mechanisms through which Copilot shapes project-level code contributions and coordination time for peripheral versus core developers differently. We find that following the adoption of Copilot, peripheral developers had relatively lower individual code contributions, as shown in column (1), and relatively lower coding participation, as shown in column (2). These could

¹⁰ The mean proportion of merged PRs by peripheral developers to the total number of merged PRs, as reported in Table 1, is 0.52. This proportion decreased by 0.019, representing a 3.7% decrease.

¹¹ The mean ratio of the average time taken to merge a PR submitted by peripheral developers to the average time taken to merge a PR submitted by all types of developers, as reported in Table 1, is 1.56. This ratio increases by 0.085, representing an 5.4% increase.

potentially explain the effect of Copilot on reducing the proportion of project-level code contributions made by peripheral developers, supporting our arguments for H3a.

At the same time, as shown in columns (3) to (5) of Table 11, peripheral developers' code contributions received relatively higher code discussion volume, more developers participating in discussing their code, and greater code discussion intensity per developer on their code. These results could collectively explain why peripheral developers experienced a relatively longer coordination time for merging their code following the adoption of Copilot and are consistent with our arguments for H3b.

[insert Table 11 here]

It is worth noting that the observed differential effects in Tables 10 and 11 are either due to lower increases or reductions in activities of peripheral developers. To provide further support to H3a (which suggests the project-level code contributions from both peripheral and core developers could increase and the former may have a smaller increase than the latter), we construct a range of metrics to capture absolute changes in project-level code contributions for both core and peripheral developers. Similarly, to provide further support to H3b (which suggests the coordination time for both peripheral and core developers could increase, and the former may have a larger increase than the latter), we also use a range of metrics to capture absolute changes in coordination time for both core and peripheral developers. The detailed results are reported in the Online Appendix K. Note that as GSCM generates different synthetic controls for these measures across samples, we cannot directly compare the magnitudes across tables. However, these findings corroborate the results in Tables 10 and 11 and suggest that observed lower relative share of project-level code contributions and greater coordination time for peripheral developers are not driven by absolute declines in their activity levels following Copilot adoption. Rather, they reflect differential returns to Copilot shaped by developer roles.

In our hypothesis development, we argue that one key distinction between peripheral developers and core developers is that the former may lack the level of project familiarity needed to overcome AI pair programmers' limitations and use them effectively. This could be one potential reason why AI pair programmers benefit peripheral developers less than core developers, as hypothesized in H3a and H3b. To

provide some preliminary evidence on this, in Online Appendix L, we show that peripheral developers indeed have less project familiarity compared to core developers, measured by their tenure with the project. We also show project familiarity seemed to influence the benefits from using Copilot.

Alternative Explanations: The heterogeneous effects of Copilot on peripheral and core developers may result from other differences between peripheral and core developers, such as programming language expertise or Copilot usage, instead of project familiarity. We conduct additional analyses to examine these alternative explanations and report the detailed results in Online Appendix M. Overall, there are no significant differences among core and peripheral developers in terms of their programming language expertise. Furthermore, the Copilot usage seems even higher among peripheral developers as compared to core developers. These results suggest that following Copilot adoption, the lower project-level code contributions and longer coordination time of peripheral developers as compared to core developers are not likely to be driven by lower programming skills or lower Copilot usage by peripheral developers.

Another possible explanation is that the heterogenous effects of Copilot on core and peripheral developers could be driven by changes in task allocations between the two groups following the adoption of Copilot. However, prior OSS literature highlights the self-organizing nature of these communities (Lindberg et al. 2024). In the absence of a formal hierarchical structure, developers are generally not assigned with tasks. Instead, they voluntarily select areas in which they wish to contribute. Given that open-source projects often adopt modular architectures and support parallel development (Howison and Crowston 2014, Medappa and Srivastava 2019), core and peripheral developers may contribute simultaneously to similar features, effectively working on the same tasks. As such, task reallocation from peripheral to core developers may not be a likely explanation for the observed pattern.

7. Discussion and Conclusions

The rise of AI pair programmers has garnered increasing attention due to their potential to transform software development. In this study, we explore the effects of AI pair programmers within the context of collaborative OSS development, with a focus on their impact on project-level code contributions and coordination time for code integration. Using repository-level proprietary data from GitHub organization,

we evaluate the impact of GitHub Copilot, an AI pair programmer, on the development outcomes of open-source repositories, along with its underlying mechanisms.

Our findings indicate that Copilot adoption significantly increases project-level code contribution alongside longer coordination time, highlighting a tradeoff between facilitated code generation and heightened coordination cost in collaborative software development. Mechanism analyses reveal that the increased project-level code contributions are driven by more developers participating in coding and greater individual code contributions. In contrast, the increased coordination time seems driven by greater code discussions, including higher code discussion volumes, more developers participating in code discussions, and greater discussion intensity per developer. We extend prior research on generative AI's productivity benefits for individual developers (e.g., Peng et al. 2023; Cui et al. 2024) by examining its impact on collaborative, voluntary participation in open-source software (OSS) projects. We also contribute to the literature on the role of generative AI in team collaboration. Unlike earlier work focused on traditional teams with fixed roles and structured coordination (Li et al. 2024, Dell'Acqua et al. 2025), we explore fluid, volunteer-driven OSS environments. We reveal that AI tools, by encouraging engagement in discussions, may increase coordination time due to the need to align diverse perspectives.

We also examine the effects of Copilot on core and peripheral developers within the OSS community, who differ in their levels of familiarity with the focal projects. Our analysis shows that, following Copilot adoption, peripheral developers experienced a relatively smaller increase in project-level code contributions but a larger increase in coordination time, when compared to core developers. In addition, we find that peripheral developers contribute a smaller proportion of timely integrated code, suggesting that the productivity gain from AI pair programmers could be less for peripheral developers than for core developers. Prior studies (e.g., Dell'Acqua et al. 2023; Peng et al. 2023) show that less-skilled users often benefit more from AI tools. In contrast, we find that in OSS settings where developers handle complex software development projects, core developers gain more from AI pair programming than peripheral ones, likely due to their greater contextual knowledge of the project.

7.1 Implications

Our paper has several implications for both practice and the broader OSS communities. First, our findings suggest that organizations should invest in AI pair programmers, as they increase project level code contributions by increasing individual code contributions and developer participation. However, these benefits come with increased coordination time. Organizations should adopt complementary strategies to manage integration challenges, such as aligning coding practices, clarifying project goals, and improving communication workflows, to fully capitalize on the benefits of AI-assisted software development.

Second, our results raise concerns about potential shifts in team structure within OSS communities. While AI pair programmers bring benefits, the gains favor core developers, who possess greater familiarity with project structure and context, which underscores the importance of integrating human intelligence with the automation and augmentation capabilities of AI pair programmers. In contrast, peripheral developers face relatively longer coordination time that reduces their overall project productivity. Over time, this imbalance could lead to more concentrated development teams, where core developers dominate and peripheral contributions decline. Such a trend risks undermining the open-source ethos of diversity of perspectives, potentially making open collaboration more closed.

In addition, AI pair programmers introduce new challenges to traditional norms of code ownership. Historically, contributors retain copyright over their code while licensing it under open-source terms that allow others to freely use, modify, and distribute the work. However, AI-generated code blurs the boundaries of code ownership, as these tools are trained on public repositories without attributing credit to the original contributors. Because it is often unclear whether a piece of code was written by a human or generated by AI, the transparency of contributions may diminish. Over time, this lack of attribution and clarity may erode trust in the origin and quality of code within OSS communities.

7.2 Limitations and Future Work

Our study has several limitations. First, due to privacy constraints, we lack individual-level Copilot use data. Future research could leverage more granular AI usage data to validate our findings and explore role transitions in the OSS development and potential spillover effects enabled by AI pair programmers.

Second, our work represents the first step in examining how generative AI affects core and peripheral developers differently. We acknowledge that beyond different levels of project familiarity, developers may differ in motivation, ability, and preference for the type of work they undertake. Due to the observational nature of our data, we are not able to measure these metrics. Future work could explore these nuanced behavioral dynamics, potentially through surveys, interviews, or mixed-method approaches.

Third, although our study emphasizes the quantitative tradeoff between project-level code contributions and coordination time, it does not capture the subjective aspects of working with AI tools. Future research could examine how developers perceive and adapt to AI pair programmers, including their satisfaction, trust, and perceived control over their contributions.

Finally, the potential risks associated with AI pair programmers warrant further investigation. One important concern is the possibility of skill erosion. Developers may become overly reliant on AI-generated suggestions, potentially diminishing their coding proficiency over time. In addition, AI systems trained on large-scale public repositories may perpetuate existing coding biases, potentially reinforcing suboptimal practices. Furthermore, AI-generated code may inadvertently introduce security vulnerabilities, especially if it lacks the contextual awareness required for sensitive or security-critical components. Future studies should propose frameworks to ensure safe and ethical AI usage in software development.

References

- Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association* 105(490):493-505.
- AlMarzouq M, AlZaidan A, AlDallal J (2020) Mining GitHub for research and education: challenges and opportunities. *International Journal of Web Information Systems* 16(4):451-473.
- Ancona DG, Caldwell DF (1992) Demography and design: Predictors of new product team performance. *Organization science* 3(3):321-341.
- Atkinson JW (1957) Motivational determinants of risk-taking behavior. *Psychological review* 64(6p1):359.
- Bai J (2009) Panel data models with interactive fixed effects. *Econometrica* 77(4):1229-1279.
- Bakos JY, Brynjolfsson E (1993) Information technology, incentives, and the optimal number of suppliers. *Journal of Management Information Systems* 10(2):37-53.
- Barke S, James MB, Polikarpova N (2023) Grounded copilot: How programmers interact with code-generating models. *Proceedings of the ACM on Programming Languages* 7(OOPSLA1):85-111.
- Brynjolfsson E, Li D, Raymond L (2025) Generative AI at work. *The Quarterly Journal of Economics*:qjae044.
- Burtch G, Lee D, Chen Z (2024) The consequences of generative AI for online knowledge communities. *Scientific Reports* 14(1):10413.

- Cataldo M, Wagstrom PA, Herbsleb JD, Carley KM (2006) Identification of coordination requirements: Implications for the design of collaboration and awareness tools. *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, 353-362.
- Crowston K, Wei K, Li Q, Howison J (2006) Core and periphery in free/libre and open source software team communications. *Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06)* (IEEE), 118a-118a.
- Cui ZK, Demirer M, Jaffe S, Musolff L, Peng S, Salz T (2024) The effects of generative ai on high skilled work: Evidence from three field experiments with software developers. Available at SSRN 4945566.
- Dabbish L, Stuart C, Tsay J, Herbsleb J (2012) Social coding in GitHub: transparency and collaboration in an open software repository. *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 1277-1286.
- Dakhel AM, Majdinasab V, Nikanjam A, Khomh F, Desmarais MC, Jiang ZMJ (2023) Github copilot ai pair programmer: Asset or liability? *Journal of Systems and Software* 203:111734.
- Dell'Acqua F, Kogut B, Perkowski P (2023) Super mario meets ai: Experimental effects of automation and skills on team performance and coordination. *Review of Economics and Statistics*:1-47.
- Dell'Acqua F, Ayoubi C, Lifshitz-Assaf H, Sadun R, Mollick ER, Mollick L, Han Y, Goldman J, Nair H, Taub S (2025) The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise.
- Demirci O, Hannane J, Zhu X (2025) Who is AI replacing? The impact of generative AI on online freelancing platforms. *Management Science*.
- El Mezouar M, Zhang F, Zou Y (2019) An empirical study on the teams structures in social coding using GitHub projects. *Empirical Software Engineering* 24(6):3790-3823.
- Espinosa JA, Slaughter SA, Kraut RE, Herbsleb JD (2007) Familiarity, complexity, and team performance in geographically distributed software development. *Organization science* 18(4):613-630.
- Feldman P, Foulds JR, Pan S (2023) Trapping LLM hallucinations using tagged context prompts. *arXiv preprint arXiv:2306.06085*.
- Feri F, Irlenbusch B, Sutter M (2010) Efficiency gains from team-based coordination—large-scale experimental evidence. *American Economic Review* 100(4):1892-1912.
- Gousios G, Pinzger M, Deursen Av (2014) An exploratory study of the pull-based software development model. *Proceedings of the 36th international conference on software engineering*, 345-355.
- Harbring C (2006) The effect of communication in incentive systems—an experimental study. *Managerial and Decision Economics* 27(5):333-353.
- Hertel G, Niedner S, Herrmann S (2003) Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel. *Research policy* 32(7):1159-1177.
- Hoffmann M, Boysel S, Nagle F, Peng S, Xu K (2024) Generative AI and the Nature of Work. Report.
- Howison J, Crowston K (2014) Collaboration through open superposition: A theory of the open source way. *Mis Quarterly* 38(1):29-50.
- Im GP, Ahuja M (2023) The Interdependence of Coordination and Cooperation in Information Technology Outsourcing. *Information Systems Research* 34(4):1791-1806.
- Imai S (2022) Is github copilot a substitute for human pair-programming? an empirical study. *Proceedings of the ACM/IEEE 44th International Conference on Software Engineering: Companion Proceedings*, 319-321.
- Janardhanan NS, Lewis K, Reger RK, Stevens CK (2020) Getting to know you: motivating cross-understanding for improved team and individual performance. *Organization Science* 31(1):103-118.
- Kalliamvakou E, Gousios G, Blincoe K, Singer L, German DM, Damian D (2014) The promises and perils of mining github. *Proceedings of the 11th working conference on mining software repositories*, 92-101.
- Kononenko O, Rose T, Baysal O, Godfrey M, Theisen D, De Water B (2018) Studying pull request merges: A case study of shopify's active merchant. *Proceedings of the 40th international conference on software engineering: software engineering in practice*, 124-133.
- Koushik MV, Mookerjee VS (1995) Modeling coordination in software construction: An analytical approach. *Information Systems Research* 6(3):220-254.

- Krishnamurthy R, Jacob V, Radhakrishnan S, Dogan K (2016) Peripheral developer participation in open source projects: an empirical analysis. *ACM Transactions on Management Information Systems (TMIS)* 6(4):1-31.
- Levine SS, Prietula MJ (2014) Open collaboration for innovation: Principles and performance. *Organization Science* 25(5):1414-1433.
- Li N, Zhou H, Mikel-Hong K (2024) Generative AI Enhances Team Performance and Reduces Need for Traditional Teams. *arXiv preprint arXiv:2405.17924*.
- Liguori P, Improta C, Natella R, Cukic B, Cotroneo D (2024) Enhancing AI-based Generation of Software Exploits with Contextual Information. *2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)* (IEEE), 180-191.
- Lindberg A, Berente N, Gaskin J, Lyytinen K (2016) Coordinating interdependencies in online communities: A study of an open source software project. *Information Systems Research* 27(4):751-772.
- Lindberg A, Schecter A, Berente N, Hennel P, Lyytinen K (2024) THE ENTRAINMENT OF TASK ALLOCATION AND RELEASE CYCLES IN OPEN SOURCE SOFTWARE DEVELOPMENT. *MIS Quarterly* 48(1).
- Malone TW, Crowston K (1994) The interdisciplinary study of coordination. *ACM Computing Surveys (CSUR)* 26(1):87-119.
- Marangunić N, Granić A (2015) Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14:81-95.
- Mattarelli E, Bertolotti F, Prencipe A, Gupta A (2022) The effect of role-based product representations on individual and team coordination practices: A field study of a globally distributed new product development team. *Organization Science* 33(4):1423-1451.
- Mawdsley JK, Meyer-Doyle P, Chatain O (2022) Client-related factors and collaboration between human assets. *Organization Science* 33(2):518-540.
- Medappa PK, Srivastava SC (2019) Does superposition influence the success of FLOSS projects? An examination of open-source software development by organizations and individuals. *Information Systems Research* 30(3):764-786.
- Noy S, Zhang W (2023) Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654):187-192.
- Oh W, Jeon S (2007) Membership herding and network stability in the open source community: The Ising perspective. *Management science* 53(7):1086-1101.
- Okhuysen GA, Bechky BA (2009) 10 coordination in organizations: An integrative perspective. *Academy of Management annals* 3(1):463-502.
- Oliveira N, Lumineau F (2017) How coordination trajectories influence the performance of interorganizational project networks. *Organization Science* 28(6):1029-1060.
- Peng S, Kalliamvakou E, Cihon P, Demirer M (2023) The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.
- Pikkainen M, Haikara J, Salo O, Abrahamsson P, Still J (2008) The impact of agile practices on communication in software development. *Empirical Software Engineering* 13:303-337.
- Piñeiro-Martín A, Santos-Criado F-J, García-Mateo C, Docío-Fernández L, López-Pérez MdC (2025) Context Is King: Large Language Models' Interpretability in Divergent Knowledge Scenarios. *Applied Sciences* 15(3):1192.
- Reagans R, Miron-Spektor E, Argote L (2016) Knowledge utilization, coordination, and team performance. *Organization Science* 27(5):1108-1124.
- Roberts JA, Hann I-H, Slaughter SA (2006) Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management science* 52(7):984-999.
- Roels G, Corbett CJ (2024) Too many meetings? Scheduling rules for team coordination. *Management Science* 70(12):8647-8667.
- Setia P, Rajagopalan B, Sambamurthy V, Calantone R (2012) How peripheral developers contribute to open-source software development. *Information Systems Research* 23(1):144-163.

- Shah SK (2006) Motivation, governance, and the viability of hybrid forms in open source software development. *Management science* 52(7):1000-1014.
- Shaikh M, Vaast E (2023) Algorithmic interactions in open source work. *Information Systems Research* 34(2):744-765.
- Simcoe T (2012) Standard setting committees: Consensus governance for shared technology platforms. *American Economic Review* 102(1):305-336.
- Singh PV, Phelps C (2013) Networks, social influence, and the choice among competing innovations: Insights from open source software licenses. *Information Systems Research* 24(3):539-560.
- Srikanth K, Puranam P (2014) The firm as a coordination system: Evidence from software services offshoring. *Organization Science* 25(4):1253-1271.
- Subramaniam C, Sen R, Nelson ML (2009) Determinants of open source software project success: A longitudinal study. *Decision Support Systems* 46(2):576-585.
- Tsay J, Dabbish L, Herbsleb J (2014) Influence of social and technical factors for evaluating contribution in GitHub. *Proceedings of the 36th international conference on Software engineering*, 356-366.
- Vasilescu B, Yu Y, Wang H, Devanbu P, Filkov V (2015) Quality and productivity outcomes relating to continuous integration in GitHub. *Proceedings of the 2015 10th joint meeting on foundations of software engineering*, 805-816.
- Wang Y, Ramachandran V, Liu Sheng OR (2021) Do fit opinions matter? The impact of fit context on online product returns. *Information Systems Research* 32(1):268-289.
- Weber RA (2006) Managing growth to achieve efficient coordination in large groups. *American Economic Review* 96(1):114-126.
- Wen W, Forman C, Graham SJ (2013) Research note—The impact of intellectual property rights enforcement on open source software project success. *Information Systems Research* 24(4):1131-1146.
- Xu B, Nguyen T-D, Le-Cong T, Hoang T, Liu J, Kim K, Gong C, Niu C, Wang C, Le B (2023) Are we ready to embrace generative ai for software q&a? *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (IEEE), 1713-1717.
- Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1):57-76.
- Yeverechyan, D., Mayya, R., & Oestreicher-Singer, G. (2024). The impact of large language models on open-source innovation: Evidence from GitHub Copilot. arXiv preprint arXiv:2409.08379.
- Yu Y, Wang H, Filkov V, Devanbu P, Vasilescu B (2015) Wait for it: Determinants of pull request evaluation latency on github. *2015 IEEE/ACM 12th working conference on mining software repositories* (IEEE), 367-371.
- Zhang X, Zhu F (2011) Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *American Economic Review* 101(4):1601-1615.
- Zhang X, Yu Y, Gousios G, Rastogi A (2022) Pull request decisions explained: An empirical overview. *IEEE Transactions on Software Engineering* 49(2):849-871.
- Zhang Y, Lo D, Kochhar PS, Xia X, Li Q, Sun J (2017) Detecting similar repositories on GitHub. *2017 IEEE 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)* (IEEE), 13-23.

Figures and Tables:

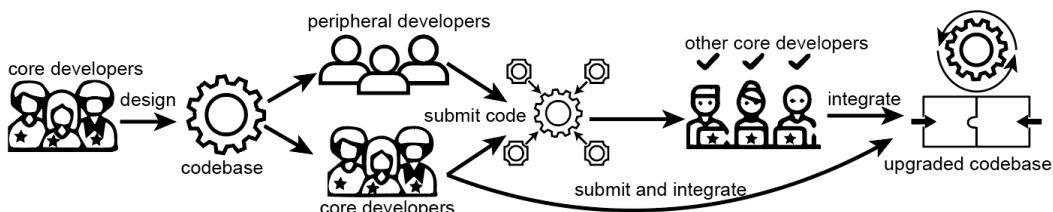


Figure 1: Open-Source Software Development Process

Table 1: Definitions and Summary Statistics, Repository-Month-Level

Variables	Description	Obs	Mean	Std.	Min	Max
Merged_PR	The number of code contributions (i.e., merged PRs).	139,329	23.65	70.59	1	3956
Merge_time	The average time taken (in hours) to accept a code submission.	139,329	355.63	956.50	0	23810
MergedPR_per_dev	The average number of code contributions per developer.	139,329	3.19	6.63	1	613
Num_devs_with_mergedPR	The number of distinct developers with code contributions.	139,329	6.30	13.77	1	575
Comments_per_mergedPR	The average number of comments per code contribution.	139,329	220.44	646.57	0	17301
Comments_per_dev_per_mergedPR	The average number of comments per developer per code contribution.	139,329	118.19	340.95	0	9757
Num_devs_with_comments_per_mergedPR	The average number of developers who provided comments per code contribution.	139,329	1	0.95	0	7
Reviews	The number of total reviews of core developers.	139,329	24.7	85.09	1	4104
Reviews_per_core	The average number of reviews per core developer.	139,329	3.58	7.67	1	810
Num_core_with_reviews	The number of core developers who reviewed code submissions.	139,329	3.01	7.92	1	186
MergedPR_1D	The number of code submissions accepted within one day.	139,329	14.98	47.95	0	2569
MergedPR_3D	The number of code submissions accepted within three days.	139,329	17.56	55.08	0	2953
MergedPR_10D	The number of code submissions accepted within ten days.	139,329	20.77	63.37	0	3431
Total_issues	The number of total issues.	139,329	35.36	114.59	0	3741
Total_bugs	The number of total bug-related issues.	139,329	3.84	18.51	0	1049
Total_issues_per_PR	The average number of total issues per code contribution.	139,329	1.42	2.36	0	201
Total_bugs_per_PR	The average number of total bug-related issues per code contribution.	139,329	0.14	0.59	0	57
Peri_merged_PR_share	The share of code contributions from peripheral developers.	77,888	0.52	0.36	0.0005	1
Peri_merge_time_ratio	The ratio of the average time taken (in hours) to accept a code submission from peripheral developers to the average time taken to accept a code submission from all developers.	77,888	1.56	2.56	0	254.1
Peri_mergedPR_1D_share	The share of code submissions from peripheral developers and accepted within one day.	66,005	0.41	0.38	0	1
Peri_mergedPR_3D_share	The share of code submissions from peripheral developers and accepted within three days.	68,419	0.43	0.37	0	1
Peri_mergedPR_10D_share	The share of code submissions from peripheral developers and accepted within ten days.	71,447	0.46	0.37	0	1
Peri_mergedPR_per_dev_ratio	The ratio of the average code contributions per peripheral developer to the average code contributions per developer of both types (peripheral and core).	77,888	0.76	0.36	0.008	7.5
Peri_dev_with_mergedPR_share	The share of peripheral developers with code contributions to all developers with code contributions.	77,888	0.63	0.29	0.006	1
Peri_comment_per_mergedPR_ratio	The ratio of average number of comments per peripheral developers' merged PR to average number of comments per merged PR.	77,888	0.55	0.60	0	13.09
Peri_comment_dev_per_mergedPR_ratio	The ratio of average number of developers who provided comments per peripheral developers' merged PR to average number of developers who provided comments per merged PR.	77,888	0.55	0.52	0	2.86
Peri_comment_per_dev_per_mergedPR_ratio	The ratio of average number of comments per developer per peripheral developers' merged PR to the average number of comments per developer per merged PR.	77,888	0.24	0.46	0	8.62

Table 2: The Impact of Copilot on Project-Level Code Contributions and Coordination Time

	(1) Project-level Code Contributions Log (Merged PR)	(2) Coordination Time Log (Merge time)
ATT of Copilot	0.059*** (0.007)	0.080*** (0.021)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	7,637	7,637
Observations	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: The Impact of Copilot on Project-Level Code Contributions (Robustness Checks)

	(1) GSCM Within-IDE Analysis	(2) GSCM Refined Sample	(3) GSCM Outlier Removal	(4) GSCM Non-AI Topic	(5) PSM and DID Single Treatment Turn-on Time
	Log (Merged PR)	Log (Merged PR)	Log (Merged PR)	Log (Merged PR)	Log (Merged PR)
ATT of Copilot	0.031*** (0.012)	0.062*** (0.007)	0.060*** (0.008)	0.057*** (0.008)	0.172*** (0.027)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,462	7,369	7,557	6,668	2,122
Observations	60,659	135,731	138,123	121,366	42,251

Notes: Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: The Impact of Copilot on Project-Level Coordination Time (Robustness Checks)

	(1) GSCM Within-IDE Analysis	(2) GSCM Refined Sample	(3) GSCM Outlier Removal	(4) GSCM Non-AI Topic	(5) PSM and DID Single Treatment Turn-on Time
	Log (Merge time)	Log (Merge time)	Log (Merge time)	Log (Merge time)	Log (Merge time)
ATT of Copilot	0.055** (0.027)	0.080*** (0.022)	0.074*** (0.022)	0.088*** (0.021)	0.148*** (0.043)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,462	7,369	7,557	6,668	2,122
Observations	60,659	135,731	138,123	121,366	42,251

Notes: Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 5: The Impact of Copilot on Project-Level Code Contributions (Mechanism)

	(1) Individual Code Contributions Log (MergedPR_per dev)	(2) Developer Coding Participation Log (Num devs with mergedPR)
ATT of Copilot	0.021*** (0.005)	0.034*** (0.004)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	7,637	7,637
Observations	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: The Impact of Copilot on Project-Level Coordination Time (Mechanism)

	(1) Discussion Volume Log (Comments_per mergedPR)	(2) Developer Discussion Participation Log (Num_devs_with_comments_per mergedPR)	(3) Individual Discussion Intensity Log (Comments_per_dev_per mergedPR)
ATT of Copilot	0.065*** (0.017)	0.017*** (0.005)	0.059*** (0.014)
Repository FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637
Observations	139,329	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 7: The Impact of Copilot on Core Developers' Review Activities

	(1) Total Reviews Log (Reviews)	(2) Individual Review Productivity Log (Reviews_per_core)	(3) Developer Review Participation Log (Num core with reviews)
ATT of Copilot	0.059*** (0.007)	0.016** (0.007)	0.036*** (0.003)
Repository FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637
Observations	139,329	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 8: The Impact of Copilot on Project Productivity

	(1) Timely Merged PRs (1 Day) Log (MergedPR_1D)	(2) Timely Merged PRs (3 Days) Log (MergedPR_3D)	(3) Timely Merged PRs (10 Days) Log (MergedPR_10D)
ATT of Copilot	0.035*** (0.009)	0.041*** (0.009)	0.051*** (0.008)
Repository FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637
Observations	139,329	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 9: The Impact of Copilot on Project-Level Code Quality

	(1) Log (Total issues)	(2) Log (Total bugs)	(3) Log (Total issues per PR)	(4) Log (Total bugs per PR)
ATT of Copilot	0.070*** (0.007)	0.052* (0.030)	0.006 (0.004)	0.002 (0.002)
Repository FE	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637	7,637
Observations	139,329	139,329	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 10: Heterogeneous Effect of Copilot on Core versus Peripheral Developers

	(1) Project-level Code Contributions	(2) Coordination Time	(3) Timely Merged PRs (1 Day)	(4) Timely Merged PRs (3 Days)	(5) Timely Merged PRs (10 Days)
	Peri_merged_PR share	Peri_merge_time_ratio	Peri_mergedPR_1D_share	Peri_mergedPR_3D_share	Peri_mergedPR_10D_share
ATT of Copilot	-0.019*** (0.004)	0.085** (0.036)	-0.021*** (0.006)	-0.024*** (0.006)	-0.020*** (0.005)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,154	4,154	3,350	3,528	3,729
Observations	77,888	77,888	66,005	68,419	71,447

Notes: This table is based on a smaller sample (i.e., a total of 4,154 repositories) than the main sample because this analysis requires at least one merged PR from peripheral developers (in order to compute merge time). The sample becomes smaller in columns (3) to (5) due to the exclusion of observations with zero PR merged in that particular time window, which would lead to a value of zero for the denominator. All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

Table 11: Heterogeneous Effect of Copilot on Core versus Peripheral Developers (Mechanism)

	(1) Individual Code Contributions	(2) Developer Coding Participation	(3) Discussion Volume	(4) Discussion Participation	(5) Individual Discussion Intensity
	Peri_mergedPR_per_dev_ratio	Peri_dev_with_mergedPR_share	Peri_comment_per_mergedPR_ratio	Peri_comment_dev_per_mergedPR_ratio	Peri_comment_per_dev_per_mergedPR_ratio
ATT of Copilot	-0.008* (0.004)	-0.017*** (0.003)	0.029*** (0.008)	0.026*** (0.006)	0.013** (0.006)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,154	4,154	4,154	4,154	4,154
Observations	77,888	77,888	77,888	77,888	77,888

Notes: This table is based on the same analysis sample as the one used for Table 10 to maintain consistency. All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. Please refer to Table 1 for detailed definitions of the dependent variables used in this table. *** p<0.01, ** p<0.05, * p<0.1.

ONLINE APPENDIX

Appendix A: Details of the Main Estimation

Our main causal inference uses the generalized synthetic control method (Xu 2017). Below, we provide a brief overview of its estimation process and relate it to our study.

Model: The GSCM combines the Interactive Fixed Effects (IFE) model (Bai 2009), which considers latent factors with effects that can vary across time and units, with the Synthetic Control (SC) method (Abadie et al. 2015), which creates synthetic control units to act as counterfactuals for the treated units. This combination enables GSCM to (1) relax the assumption of pre-treatment parallel trends, and (2) consider potential unobserved factors that vary over time at the unit level. Furthermore, GSCM offers several advantages over other estimation methods. Unlike the SC method, which is limited to a single treated unit, GSCM is designed for multiple treated units, eliminating the need to construct synthetic control units for each treatment unit one by one. Additionally, unlike the typical difference-in-differences model combined with a specific matching technique, which is most effective for a single treatment turn-on time and a large control group, GSCM allows each treatment unit to have a different treatment period and can efficiently construct synthetic control units from a relatively small control sample. These advantages make GSCM particularly well-suited to our data, as repositories adopted Copilot at different points in time and we have relatively fewer repositories in the control sample than in the treatment sample.

First, the GSCM adopts a linear IFE framework to model the latent factors. In our study, the outcome Y_{it} , the log number of merged pull requests (PR) and the log average time taken to merge a single PR of repository i in month t , can be expressed as:

$$Y_{it} = \delta_{it} D_{it} + X'_{it}\beta + \lambda'_{it} f_t + \varepsilon_{it} \quad (1)$$

where D_{it} is the treatment indicator which equals one if repository i has been developed with Copilot by month t and zero otherwise; the parameter of primary interest is δ_{it} , which signifies the dynamic impact of Copilot on the log number of merged PRs and the log average time taken to merge a single PR. δ_{it} is the heterogeneous treatment effect and its subscripts i and t indicate that the estimates vary across units and

time. Let r be the number of latent factors, then f_t represents the $(r \times 1)$ vector of unobserved common factors, and λ'_i is the $(1 \times r)$ vector of unknown factor loadings. The factor component $\lambda'_i f_t$ can be expressed as $\lambda'_i f_t = \lambda_{i1} f_{1t} + \lambda_{i2} f_{2t} + \dots + \lambda_{ir} f_{rt}$. A two-way fixed effects specification is a special case of the factor component, where $r = 2$, $f_{1t} = 1$, and $\lambda_{i2} = 1$, so that $\lambda'_i f_t = \lambda_{i1} + f_{2t}$. Here, λ_{i1} represents the repository fixed effects, while f_{2t} is the month fixed effects. We specify the two-way fixed effects to control for heterogeneity across repositories and time, while considering other unobserved latent factors. In general, $\lambda'_i f_t$ is estimated by an iterative factor analysis of the residuals from the model. If optimal number of unobserved latent factors determined by the cross-validation technique is zero, it means that the fixed effects setting has effectively accounted for any unobserved time-varying characteristics (Xu 2017). Lastly, ε_{it} is the error term with a mean of zero.

Estimation of Latent Factors: An essential task when utilizing the IFE framework is to identify the number of latent factors, denoted as r . Xu (2017) suggests a predictive analytics method for this purpose. For clarity, we divide the total number of repositories into two groups: Tr for treated repositories and Co for control repositories. The latent-factor selection algorithm first applies Equation (1) to data from control units only, covering both pre-treatment and post-treatment time periods, using the following equation:

$$Y_{it} = X'_{it} \beta + \lambda'_i f_t + \varepsilon_{it}, \forall i \in Co \quad (2)$$

Since the control repositories never received treatment during the data-collection period, $\delta_{it} D_{it}$ is excluded from Equation (2). The value of r can vary within a range of candidate values specified by the researchers. For each specified r , the algorithm runs Equation (2) and derives the estimates $\hat{\beta}$ and \hat{f}_t . Subsequently, a cross-validation procedure is carried out for all treated repositories. Specifically, let s be the index of a pre-treatment period, ranging from one (the first pre-treatment period) to t^0 (the last pre-treatment period). Equation (3) is then applied to each pre-treatment period, starting with $s = 1$, for all repositories in the treatment group by leaving one period out and using the remaining periods to estimate the factor loadings.

$$\hat{\lambda}_{i,-s} = (\hat{f}_{-s}^{0'} \hat{f}_{-s}^0)^{-1} \hat{f}_{-s}^{0'} (Y_{i,-s}^0 - X_{i,-s}^{0'} \hat{\beta}), i \in Tr, s = 1, \dots, t^0 \quad (3)$$

In Equation (3), \hat{f}_{-s}^0 and $\hat{\beta}$ are the estimates obtained from Equation (2). The superscript 0 indicates periods before the introduction of the treatment, and the subscript $-s$ denotes all periods except for s . Based on these estimates, the algorithm predicts the outcome for treated unit i in period s using $\hat{Y}_{is}(0) = X'_{is}\hat{\beta} + \lambda'_{i,-s}\hat{f}_s$ and records the out-of-sample error $e_{is} = Y_{is}(0) - \hat{Y}_{is}(0)$ for all $i \in Tr$. In the final step, the algorithm calculates the Mean Squared Error (MSE) of the prediction, summed over all pre-treatment periods s , for each candidate value of r . The value of r that minimizes this prediction error is selected, and the corresponding set of latent factors will be used in the causal inference process.

Estimation of Average Treatment Effects: The GSCM algorithm uses the estimated function to predict the counterfactuals for treated units, denoted as $\hat{Y}_{it}(0)$, representing the outcome of treated repositories had they not received treatment in the post-treatment period. The causal effect of the treatment is quantified as the Average Treatment effect on the Treated unit (ATT), calculated mathematically as:

$$ATT_t = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} [Y_{it}(1) - \hat{Y}_{it}(0)] \text{ for } t > t^0$$

where \mathcal{T} denotes the set of treated units and $|\mathcal{T}|$ represents the number of units in \mathcal{T} . $Y_{it}(1)$ is the observed outcome for treated repository i at time t . The essence of GSCM lies in the fact that if the predicted outcomes for the treated repositories during the pre-treatment periods are accurate, the algorithm can produce a valid counterfactual for each treated unit during the post-treatment periods. Finally, the GSCM uses bootstrapping to estimate confidence intervals and standard errors (Xu 2017).

Unlike typical econometric modeling, the latent-factor selection algorithm in GSCM prioritizes models with the lowest out-of-sample prediction error to accurately predict counterfactuals, rather than those with the best fit based on information criterion. As a result, models with more latent factors may not be selected, even if they account for more confounding factors. Moreover, the GSCM without any latent factor still yields a more valid estimate than a difference-in-differences (DID) model, except under two conditions: (1) the treatment is randomly assigned, satisfying the parallel trend assumption; and (2) no treatment heterogeneity exists. If either of these conditions is not met, the DID estimate is likely to be invalid.

Appendix B: Equivalence Test of Pre-trend in GSCM

We apply the equivalence test to examine the presence of a pre-treatment trend in GSCM (Pan and Qiu 2022, Egami and Yamauchi 2023, Wang et al. 2023). This is a standard way to test the performance of GSCM (Liu et al. 2024), as the equivalence test better incorporates substantive considerations of what constitutes good balance on covariates and placebo outcomes compared to traditional tests (Hartman and Hidalgo 2018). Specifically, we use the two-one-sided t (TOST) test. The test includes an equivalence range within which differences are deemed inconsequential (Hartman and Hidalgo 2018, Lakens et al. 2018). The test is considered passed if the average prediction error for any pre-treatment period is within the equivalence range (Liu et al. 2024).

The result of the equivalence test for the main outcome variables, including project-level code contributions, coordination time and project productivity which combines the effect of these two competing forces, are shown in Figures B.1, through B.5. We observe that the average prediction error with 90% confidence intervals (the gray dotted line) is within the equivalence range (the red dotted line). Thus, we can reject the null of inequivalence (equivalence test p-value is 0) and conclude that there exists no pre-treatment trend (Hartman and Hidalgo 2018, Pan and Qiu 2022, Egami and Yamauchi 2023, Wang et al. 2023, Liu et al. 2024). This indicates that a sufficient set of confounders have been controlled to address the endogeneity concerns and that GSCM provides a good control group.

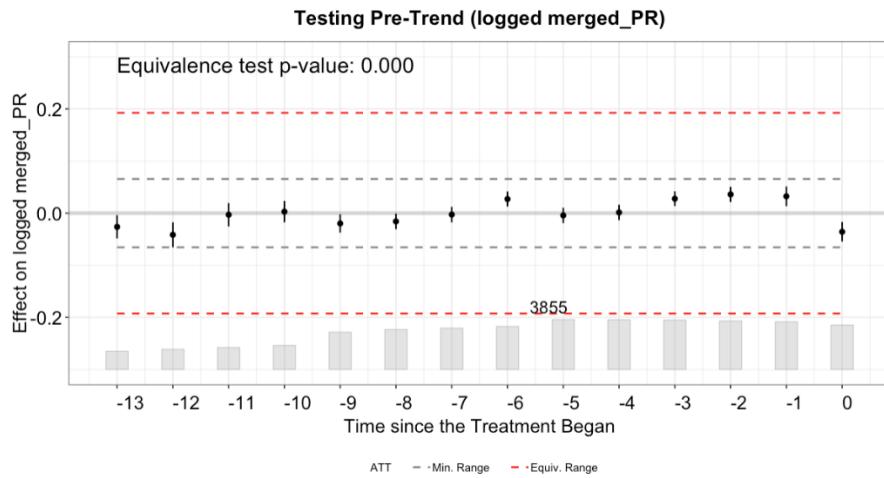


Figure B.1: Pre-trend test (TOST) of logged merged_PR

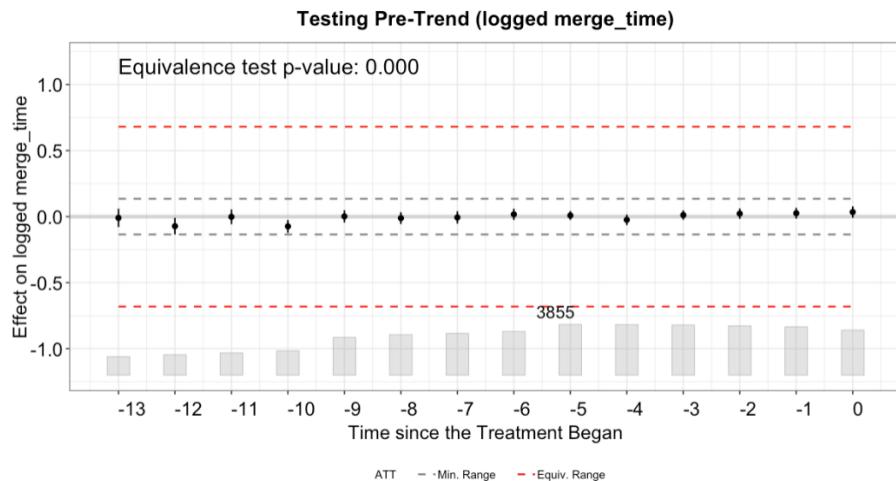


Figure B.2: Pre-trend test (TOST) of logged merge_time

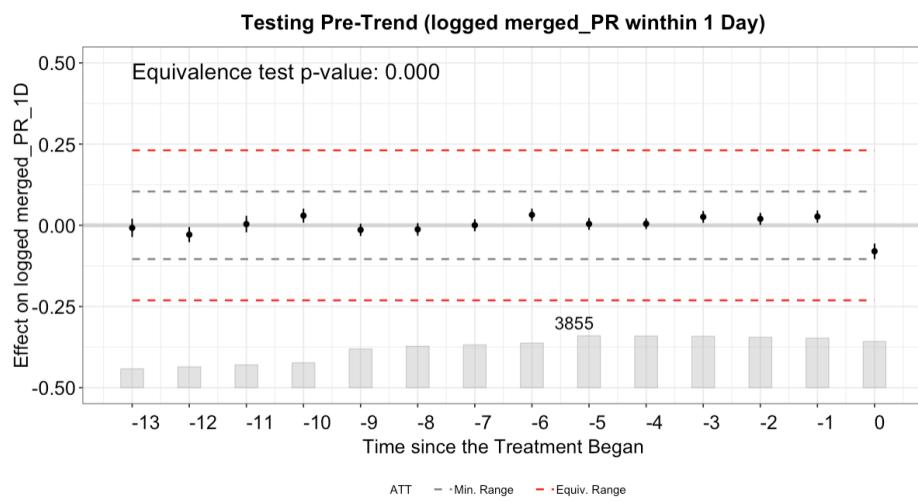


Figure B.3: Pre-trend test (TOST) of logged merged_PR_1D

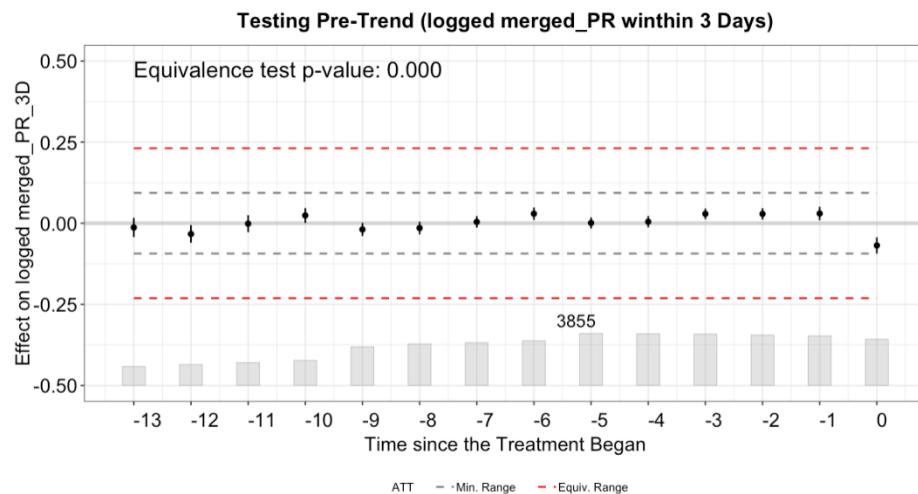


Figure B.4: Pre-trend test (TOST) of logged merged_PR_3D

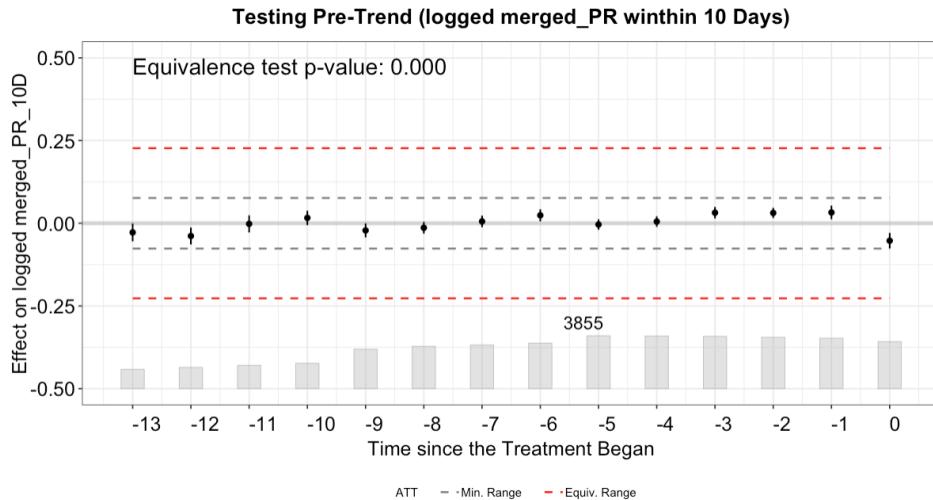


Figure B.5: Pre-trend test (TOST) of logged merged_PR_10D

Appendix C: Placebo Test in GSCM

To further validate our causal identification strategy based on GSCM, we conduct placebo tests by artificially shifting the actual Copilot adoption time earlier. Specifically, we move the observed Copilot adoption date three months earlier than its true occurrence and re-estimate the GSCM model under these placebo scenarios. These falsification tests are essential for verifying the strict exogeneity assumption, as they enable us to detect potential violations or model misspecifications. If our model specification is valid and the observed effects genuinely result from Copilot adoption, these placebo tests should yield no significant effects.

To statistically evaluate the results from our placebo analyses, we apply the two-one-sided t (TOST) equivalence test, which assesses whether the estimated placebo effects are practically equivalent to zero (Liu et al. 2024). Similar to pre-trend assessments, the TOST approach involves specifying an equivalence interval around zero and testing the null hypothesis that there is a meaningful placebo effect.

The results from these tests for the main outcome variables, including project-level code contributions, coordination time and project productivity, are shown in Figures C.1, through C.5. We observe that the placebo effect estimates during the artificial pre-treatment period are at or below zero, while those from the artificial post-treatment period are substantially above zero. The equivalence tests indicate that the placebo effects remain within the equivalence range defined around zero (equivalence test p-value is 0). Thus, we can reject the null hypothesis of inequivalence, thereby supporting the validity and robustness of our causal identification strategy.

It is worth noting that there is a drop at the onset of treatment (i.e., time = 0) in contribution-related measures, such as the number of merged PRs and the number of timely merged PRs. This drop likely reflects an initial adjustment period as developers adapt to the AI pair programmer or integrate it into their workflow. Importantly, the equivalence holds in the pre-treatment window, supporting the credibility of the synthetic control construction.

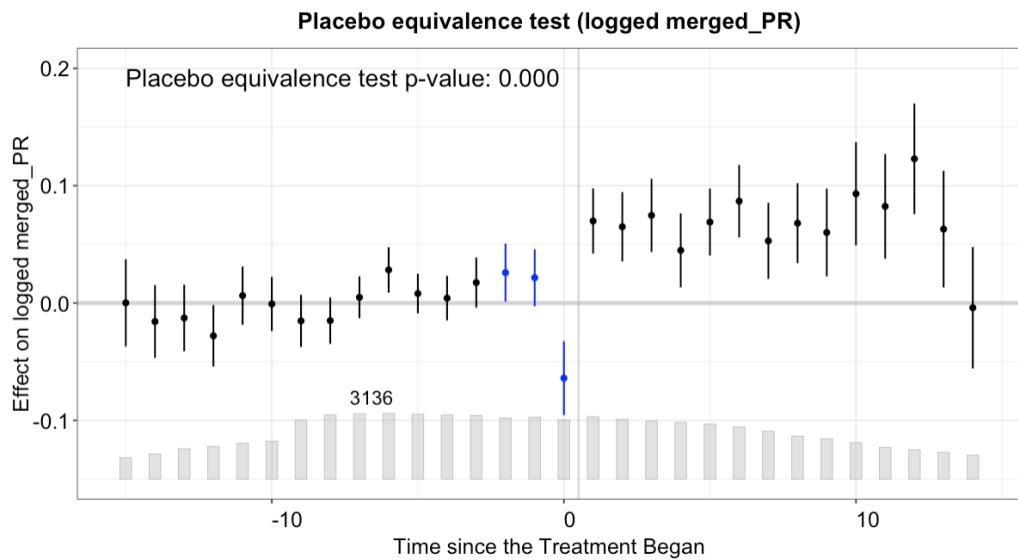


Figure C.1 Placebo Equivalence Test of logged merged_PR

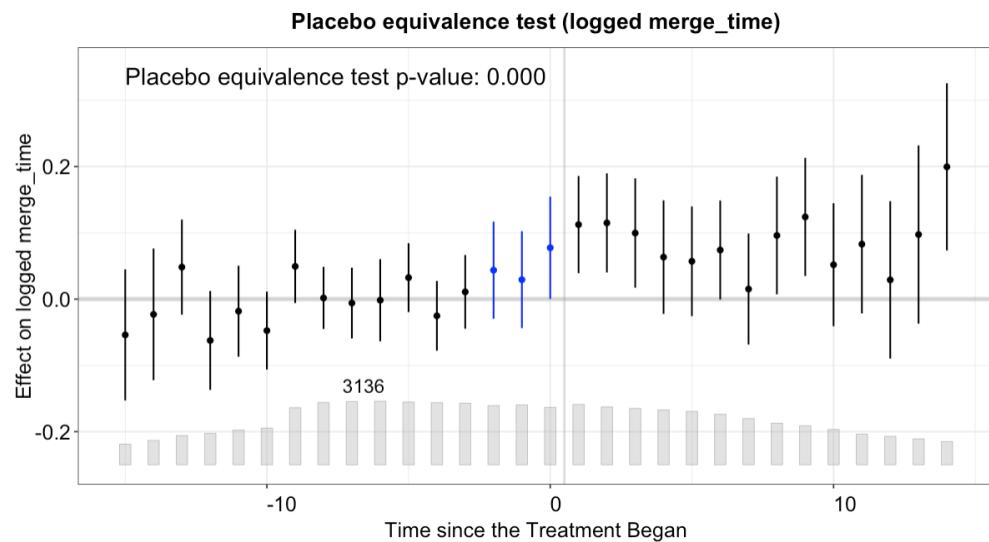


Figure C.2: Placebo Equivalence Test of logged merge_time

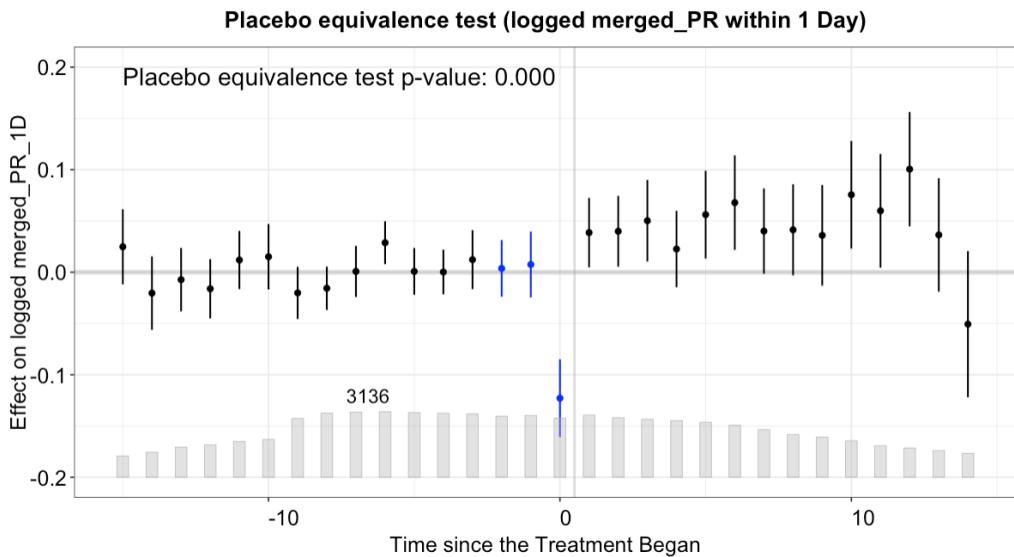


Figure C.3: Placebo Equivalence Test of logged merged_PR_1D

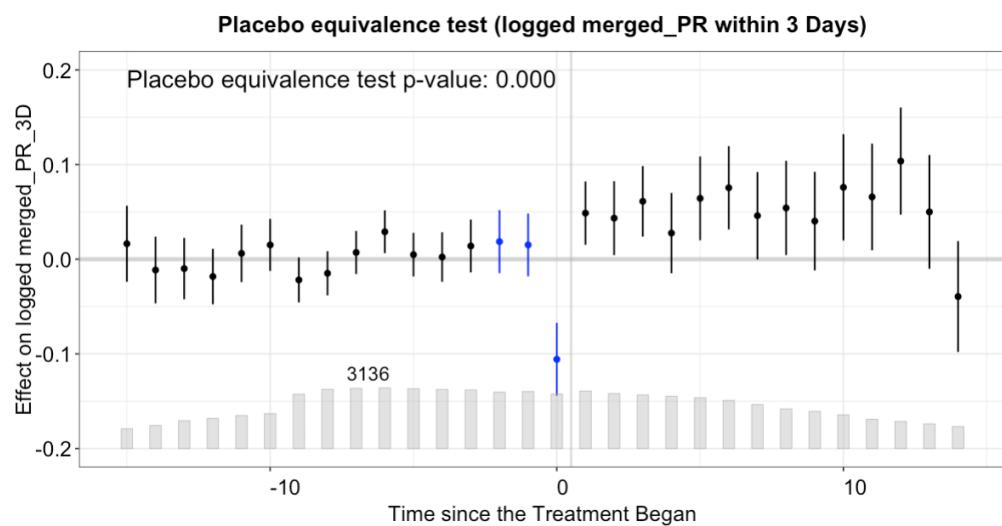


Figure C.4: Placebo Equivalence Test of logged merged_PR_3D

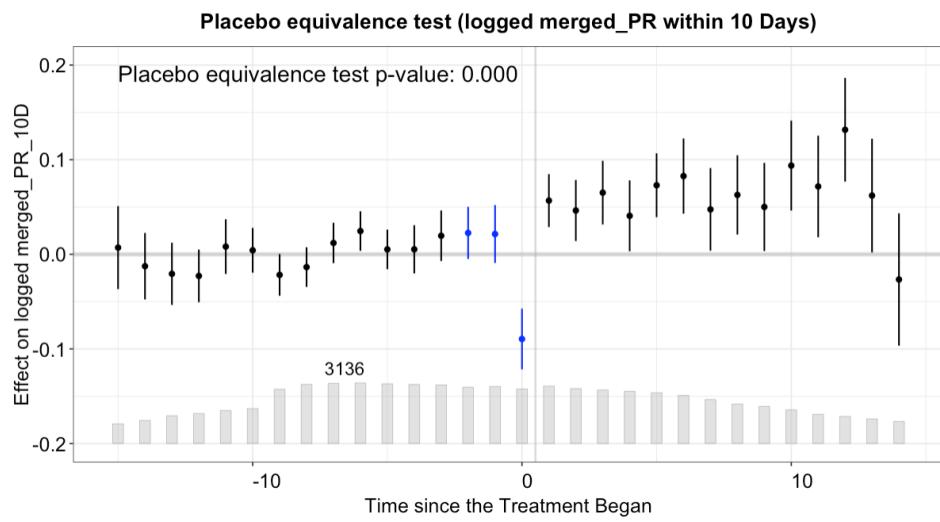


Figure C.5: Placebo Equivalence Test of logged merged_PR_10D

Appendix D: Summary of Empirical Methods

Table D.1: Overview of Empirical Methods

Empirical methods	Description	Key takeaways
Generalized Synthetic Control Method in the Main Analysis	We use the GSCM to construct a weighted control unit that matches the outcome variable of the treated unit during the pre-treatment period.	Because the GSCM models the trend of the outcome variable, our results are robust to the effects of unobservable confounders that change over time.
Generalized Synthetic Control Method in a Within-IDE Analysis	We focus on repositories using Copilot-supported IDEs and further restrict the treatment versus control group comparison within the same Copilot-supported IDE.	Our results remain robust in this within-IDE analysis, which addresses concerns regarding unobserved differences between repositories using supported versus unsupported IDEs, as well as potential differences across different IDEs that support Copilot.
Generalized Synthetic Control Method with Refined Sample	We exclude repositories associated with developers who participate in both treatment and control repositories.	Our results are robust with this refined sample and address the concern that the knowledge transfer of developers might bias the results.
Generalized Synthetic Control Method and Outlier Removal	We exclude repositories with extreme values in project-level outcomes to test robustness against the remaining sample.	Our results are consistent after outlier removal, indicating results are not driven by extreme cases.
Generalized Synthetic Control Method and non-AI Topic Projects	We remove repositories focused on AI topics to test whether the observed effects are driven by AI-intensive projects.	Our results remain stable, suggesting that the observed effects are generalizable beyond AI-focused repositories.
Propensity Score Matching and Difference-in-Differences	We use the PSM to construct a comparable control group with a single treatment turn-on time and analyze the effect using the DID estimation.	Our results are consistent after addressing the concerns about non-random treatment turn-on time for Copilot usage in the baseline analysis.
Generalized Synthetic Control Method and Alternative Measures	We use alternative measures to test if the observed effects are sensitive to the choice of dependent variables.	Our results show consistency across alternative outcome variables, ruling out bias from specific definitions of outcomes.
Generalized Synthetic Control Method and Workload Effect	We add the cumulative submitted code as the control in coordination time analysis.	The finding that Copilot increases coordination time remains robust after controlling for submitted code, suggesting that the increase in coordination time is not solely driven by higher levels of code submission.
Two-Way Fixed Effects with Full Sample	We apply two-way fixed effects to test the direct effect of Copilot on project-level code contributions and coordination time.	Our results are consistent with TWFE and the full sample, suggesting that our results are robust to alternative estimation strategies.
Generalized Synthetic Control Method with Expanded Sample	We validate our results using an expanded sample that includes small repositories with only two developers.	Our results are consistent with the expanded sample, suggesting that the observed patterns are not specific to active repositories with large developer teams.
Generalized Synthetic Control Method and Additional Validation	We include months without any merging activity in the analysis.	The results are consistent, suggesting that our results are robust to different sample selection.
Equivalence Pre-trend Tests and Placebo Tests	We conduct equivalence tests to evaluate the performance of GSCM and eliminate concerns of pre-trend and model misspecification.	Our results support the validity of the GSCM specification, confirming no significant pre-trend effects and indicating that observed effects are attributable to Copilot adoption.
Relative Time Model (RTM)	We use the RTM to verify whether the parallel trend assumption holds in the analysis of PSM and DID.	Our results pass the parallel pre-trend test, indicating that we have constructed a comparable control group in the analysis of PSM and DID.

Appendix E: Balance Checks in PSM

To construct a matched sample, we employ a one-to-one nearest neighbor matching approach without replacement, pairing each treated repository that adopted Copilot with a single control repository that did not use Copilot. Matching without replacement ensures that each control repository is used only once, preventing any single control project from disproportionately influencing the results. While this may reduce the total number of matches, it improves the integrity of the comparison by limiting reuse of controls. Matching is performed based on a set of repository characteristics measured during the six-month pre-treatment period (January to June 2021). For each characteristic, we calculate the monthly values, average them across the six-month window, and then apply a log transformation. These variables include the number of merged PRs, the average time to merge a PR, the number of distinct developers with merged PRs, and the number of push events, release events, opened issues, and closed issues.

Table E.1 reports the balance check results for the matched sample, indicating that there are no statistically significant differences between the treatment and control groups across these pre-treatment covariates after matching.

Table E.1: Balance Check of Matched Sample

Variable	Treatment Mean	Control Mean	Difference	P-value
Log (Merged_PR)	2.10	2.05	0.05	0.20
Log (Merge_time)	4.12	4.09	0.03	0.60
Log (Num_devs_with_mergedPR)	1.32	1.29	0.03	0.13
Log (Push)	2.96	2.91	0.05	0.33
Log (Release)	0.28	0.27	0.01	0.79
Log (Open_issue)	1.13	1.11	0.02	0.68
Log (Closed_issue)	1.05	1.04	0.01	0.79

Appendix F: Relative Time Model (RTM)

We employ the RTM to examine whether there is any pre-treatment trend in the analysis of PSM and DID, i.e., whether the outcome variables of repositories in the treatment group are different from that in the control group even before the treatment (i.e., the availability of Copilot) happens. The RTM has been used in the economics and information systems literature to validate the pre-treatment parallel trend assumption (Lu et al. 2019, Alyakoob and Rahman 2022). More specifically, we use the following RTM (Autor 2003).

$$Y_{it} = \beta_0 + \sum_{\eta=1}^q \beta_{1,\eta} (Copilot_i \times Post_{t-\eta}) + \sum_{\eta=0}^p \beta_{2,\eta} (Copilot_i \times Post_{t+\eta}) + \alpha_i + \mu_t + \varepsilon_{it}$$

In this model, the sum on the right-hand side represents q lead indicators (anticipatory effects) and p lagged indicators (posttreatment effects) where Y_{it} denotes the log number of merged PRs or the log average time taken to merge a single PR for repository i in month t . For our analysis of PSM and DID, we set $q = 4$ and $p = 18$. When the pre-treatment trend exists, the lead indicators would be able to predict changes in dependent variables, the log number of merged PRs and the log average time taken to merge a PR. Conversely, if the pre-treatment trend is unlikely to be a concern, the lead indicators would not predict dependent variables (Autor 2003, Angrist and Pischke 2009). The estimation results of the RTM are shown in Table F.1. We observe that the coefficients for the four lead indicators are statistically insignificant, indicating that there does not exist a pre-treatment trend.

Table F.1: Pre-trend Test for PSM and DID Analysis

	(1) PSM and DID Log (Merged PR)	(2) PSM and DID Log (Merge time)
$Copilot_i \times Pre4m$	-0.033 (0.034)	0.037 (0.088)
$Copilot_i \times Pre3m$	-0.020 (0.038)	0.094 (0.090)
$Copilot_i \times Pre2m$	-0.050 (0.041)	0.089 (0.094)
$Copilot_i \times Pre1m$	-0.031 (0.041)	0.100 (0.094)
$Copilot_i \times Post0m$	0.036 (0.044)	0.165* (0.098)
$Copilot_i \times Post1m$	0.068 (0.046)	0.100 (0.101)
$Copilot_i \times Post2m$	0.097** (0.047)	0.090 (0.100)
$Copilot_i \times Post3m$	0.042 (0.046)	0.237** (0.116)
$Copilot_i \times Post4m$	0.159*** (0.047)	0.325*** (0.100)
$Copilot_i \times Post5m+$	0.175*** (0.042)	0.257*** (0.075)

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix G: Additional Robustness Checks

Table G.1: The Impact of Copilot on Project-Level Code Contributions
(Alternative Measures of Code Contributions)

	(1) GSCM Main Sample Log (Submitted PRs)	(2) GSCM Main Sample Log (Commits)
ATT of Copilot	0.075*** (0.009)	0.107*** (0.014)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	7,637	7,637
Observations	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table G.2: The Impact of Copilot on Project-Level Coordination Time
(Alternative Measures of Coordination Time)

	(1) GSCM Main Sample Log (Merge time in days)	(2) GSCM Main Sample Log (Merge time in minutes)
ATT of Copilot	0.049*** (0.016)	0.101*** (0.032)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	7,637	7,637
Observations	139,329	139,329

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table G.3: The Impact of Copilot on Project-Level Coordination Time with Code Submission

	(1) GSCM Main Analysis Log (Merge time)	(2) GSCM Within-IDE Analysis Log (Merge time)	(3) GSCM Refined Sample Log (Merge time)	(4) GSCM Outlier Removal Log (Merge time)	(5) GSCM Non-AI Topic Log (Merge time)	(6) PSM and DID Single Treatment Turn-on Time Log (Merge time)
ATT of Copilot	0.079*** (0.023)	0.055** (0.028)	0.077*** (0.021)	0.072*** (0.023)	0.087*** (0.021)	0.147*** (0.044)
Log (Open_PR)	0.031 (0.029)	0.041 (0.041)	0.054* (0.031)	0.021 (0.032)	0.028 (0.032)	0.057 (0.037)
Repository FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
# of repositories	7,637	4,462	7,369	7,557	6,668	2,122
Observations	139,329	60,659	135,731	138,123	121,366	42,251

Notes: *Open_PR* refers to the total number of cumulative submitted PRs till the focal month and is added to the regressions to control for the effect of workload. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table G.4: The Impact of Copilot on Project-Level Code Contributions and Coordination Time
(Two-way Fixed Effects with Full Sample)

	(1) TWFE Project-level Code Contributions	(2) TWFE Coordination Time
	Log (Merged PR)	Log (Merge time)
ATT of Copilot	0.029*** (0.008)	0.036** (0.018)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	9,244	9,244
Observations	140,084	140,084

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table G.5: The Impact of Copilot on Project-Level Code Contributions and Coordination Time
(Expanded Sample that Includes Repositories with Two Developers)

	(1) GSCM Expanded Sample Project-level Code Contributions	(2) GSCM Expanded Sample Coordination Time
	Log (Merged PR)	Log (Merge time)
ATT of Copilot	0.013** (0.006)	0.083*** (0.015)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	12,512	12,512
Observations	240,433	240,433

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table G.6: The Impact of Copilot on Project-Level Code Contributions
(Expanded Sample that Includes Repository-month Observations with Zero Merged PR)

	(1) Project-level Code Contributions	(2) Individual Code Contributions	(3) Developer Coding Participation
	Log (Merged PR)	Log (MergedPR_per dev)	Log (Num devs with mergedPR)
ATT of Copilot	0.066*** (0.006)	0.058*** (0.005)	0.050*** (0.004)
Repository FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
# of repositories	8,965	8,965	8,965
Observations	206,897	206,897	206,897

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix H: Additional Evidence on the Mechanisms for H1 and H2

Table H.1: How Individual Code Contributions and Developer Coding Participation are Correlated with Project-Level Code Contributions

	(1) Log (Merged PR)	(2) Log (Merged PR)
Log (MergedPR_per_dev)	1.173*** (0.006)	
Log (Num_devs_with_mergedPR)		1.213*** (0.006)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	7,637	7,637
Observations	139,329	139,329

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table H.2: How Code Discussion is Correlated with Project-Level Coordination Time

	(1) Log (Merge time)	(2) Log (Merge time)	(3) Log (Merge time)
Log (Comment_per_mergedPR)	0.059*** (0.004)		
Log (Num_devs_with_comments_per_mergedPR)		0.259*** (0.016)	
Log (Comments_per_dev_per_mergedPR)			0.063*** (0.004)
Repository FE	Yes	Yes	Yes
Month FE	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637
Observations	139,329	139,329	139,329

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table H.3: How Project-Level Code Contributions and Coordination Time are Correlated with Project-Level Productivity

	(1) Log (MergedPR 1D)	(2) Log (MergedPR 1D)	(3) Log (MergedPR 3D)	(4) Log (MergedPR 3D)	(5) Log (MergedPR 10D)	(6) Log (MergedPR 10D)
Log (Merge_time)	-0.152*** (0.002)		-0.144*** (0.002)		-0.119*** (0.002)	
Log (Merged_PR)		1.027*** (0.003)		1.058*** (0.002)		1.069*** (0.002)
Repository FE	Yes	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes	Yes
# of repositories	7,637	7,637	7,637	7,637	7,637	7,637
Observations	139,329	139,329	139,329	139,329	139,329	139,329

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix I: Text Analysis of Comment Content

Prior research in team coordination emphasizes that the content of discussions also plays a role in shaping coordination time, as greater diversity in discussion topics can increase the complexity and difficulty of coordination (Hansen et al. 2018). Building on this insight, we argue that the same logic may apply to software development, where developer code discussions serve as a primary mechanism for coordination. In this context, a broader range of topics discussed may similarly introduce additional coordination challenges.

To examine the effect of GitHub Copilot on the diversity of code discussion topics within merged PRs, we apply a Latent Dirichlet Allocation (LDA) topic modeling approach. Specifically, we measure both the number of topics discussed per merged PR and the entropy of topic distribution. We adopt the LDA model for three key reasons. First, LDA addresses the data sparsity challenges inherent in traditional approaches such as TF-IDF. Second, the topics and associated keywords generated by LDA are human-interpretable, in contrast to the latent dimensions produced by models like doc2vec. Third, LDA has been widely validated and adopted in prior literature across various research contexts, including scientific publications (Griffiths and Steyvers 2004, Wang and Blei 2011), social media content (Ramage et al. 2010, Weng et al. 2010, Lee et al. 2016, Shin et al. 2020), business descriptions (Shi et al. 2016), and mobile App description (Lee et al. 2020). The key advantage of using LDA lies in its ability to identify latent themes from the data by grouping semantically related keywords into coherent topics, rather than treating each word as an independent feature. We exclude comments that are not in English or are too short to convey meaningful information. After this step, we estimate the topic model using a corpus of 5,404,735 merged PR comments collected from selected repositories between 2021 and 2022. This modeling approach provides a multinomial distribution over the inferred topics for each merged PR, enabling us to assess both the number of distinct topics present and the evenness of their distribution.

A critical hyperparameter in LDA modeling is the number of topics to extract. Several methods have been proposed to determine the optimal number, including metrics such as perplexity, topic coherence, and topic diversity. Given our objective of categorizing advice-related content into distinct yet coherent

themes, we rely on both coherence and diversity scores to guide model selection. Based on these criteria, we identify 10 as the optimal number of topics, balancing topic distinctiveness and coherence.

The results, presented in Table I.1, indicate that following the adoption of Copilot use, there was an increase in both the number of code discussion topics and the entropy score. This suggests that Copilot introduces diverse code discussion content, leading to increased coordination time.

Table I.1: The Impact of Copilot on Project-Level Code Discussion Content

	(1) Entropy_score	(2) Topics_count
ATT of Copilot	0.017** (0.007)	0.029** (0.014)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	4,009	4,009
Observations	78,636	78,636

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix J: The Impact of Copilot and Project Complexity

In OSS communities, complex projects typically contain more features and code modules. Prior work suggests that developers are more likely to contribute to complex projects, as they offer greater status and visibility benefits (Chengalur-Smith et al. 2010, Hann et al. 2013, Medappa and Srivastava 2019). However, these projects also tend to impose greater demands for knowledge sharing, communication, and mutual understanding (Williams 2011, Zimmerman et al. 2018). While generative AI tools enhance developers' capacity to produce code, they do not reduce the inherent complexity of a project. There are two plausible, yet opposing, mechanisms through which project complexity might influence the observed effects of Copilot.

On one hand, complex projects may attract more contributors due to their higher status potential, which can lead to greater project-level code contributions. However, the increased number of contributors also brings a wider range of perspectives and development styles, making it more difficult to reach consensus and resolve conflicting views, thereby increasing coordination time. On the other hand, these projects may present higher entry barriers, which could limit participation in both coding and non-coding activities (i.e., code discussion) and reduce coordination time. In either case, it is important to determine whether the effects we observe are simply artifacts of project complexity rather than the result of Copilot adoption. To address this concern, we conduct the analysis by splitting the sample into complex versus simple projects based on the number of lines of code in each repository prior to Copilot adoption. We use a median split, classifying repositories above the median as complex and those below the median as simple.

The results in Tables J.1 and J.2 show that both simple and complex projects exhibit similar patterns. Copilot adoption is associated with increased project-level code contributions and longer coordination time in both cases. To further examine this, we conduct PSM and DID analyses using an interaction term for project complexity. We define a binary variable “Complexity” that equals 1 if a project's lines of code exceed the median value before the Copilot use and 0 otherwise. As shown in Table J.3, the interaction term is not statistically significant, indicating no difference in Copilot's impact between simple and complex projects.

Table J.1: The Impact of Copilot on Simple Projects

	(1) Project-level Code Contributions	(2) Coordination Time	(3) Timely Project Outputs (1 Day)	(4) Timely Project Outputs (3 Days)	(5) Timely Project Outputs (10 Days)
	Log (Merged PR)	Log (Merge time)	Log (MergedPR_1D)	Log (MergedPR_3D)	Log (MergedPR_10D)
ATT of Copilot	0.046*** (0.011)	0.082** (0.032)	0.025** (0.012)	0.029*** (0.011)	0.039*** (0.011)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	3,591	3,591	3,591	3,591	3,591
Observations	54,404	54,404	54,404	54,404	54,404

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table J.2: The Impact of Copilot on Complex Projects

	(1) Project-level Code Contributions	(2) Coordination Time	(3) Timely Project Outputs (1 Day)	(4) Timely Project Outputs (3 Days)	(5) Timely Project Outputs (10 Days)
	Log (Merged PR)	Log (Merge time)	Log (MergedPR_1D)	Log (MergedPR_3D)	Log (MergedPR_10D)
ATT of Copilot	0.086*** (0.012)	0.068** (0.029)	0.065*** (0.014)	0.071*** (0.013)	0.079*** (0.014)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,046	4,046	4,046	4,046	4,046
Observations	78,706	78,706	78,706	78,706	78,706

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table J.3: The Impact of Copilot on Project Complexity (PSM and DID)

	(1) Project-level Code Contributions		(2) Coordination Time
	Log (Merged PR)	Log (Merge time)	
Copilot	0.142*** (0.034)		0.150*** (0.057)
Copilot*Complexity	0.059 (0.039)		-0.003 (0.063)
Repository FE	Yes		Yes
Month FE	Yes		Yes
# of repositories	2,122		2,122
Observations	42,251		42,251

Notes: Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Appendix K: Additional Analyses on the Differential Effects of Copilot among Core and Peripheral Developers

We construct 20 metrics to capture absolute changes in project-level outcomes for both core and peripheral developers. These metrics span four key dimensions: project-level code contributions, coordination time, project-level productivity, and the underlying mechanisms including individual code contributions, developer coding participation, code discussion volume, code discussion participation, and code discussion intensity per developer.

Specifically, for each repository-month, we compute: (1) the number of merged PRs submitted by core and peripheral developers; (2) the average time taken to merge a single PR submitted by each group; (3) the number of PRs submitted by core and peripheral developers that were merged within one, three, and ten days; (4) the average number of merged PRs per core and peripheral developer; (5) the number of distinct core and peripheral developers submitting PRs which were eventually merged; (6) the average number of comments per merged PR submitted by each group; (7) the average number of unique developers participating in code discussions per merged PR submitted by each group; and (8) the average number of comments per developer per merged PR submitted by each group.

We report the results for core developers' project-level code contributions, coordination time, and project-level productivity in Table K.1 and for peripheral developers in Table K.2. Overall, the findings indicate that both core and peripheral developers benefited from the adoption of Copilot, though the magnitude and nature of these benefits varied. For core developers, Copilot led to a 6.1% increase in project-level code contributions, a 6.7% increase in coordination time, and increases in timely integrated code contributions within one, three, and ten days by 4.4%, 4.8%, and 5.8%, respectively. In contrast, peripheral developers experienced a 5.4% increase in project-level code contributions and a 5.3% increase in coordination time. Their timely integrated code contributions increased by 2.3%, 2.4%, and 3.6% within one, three, and ten days, respectively.

The results for the underlying mechanisms, such as individual code contributions, developer coding participation, and code discussion-related metrics, are presented in Table K.3 for core developers and Table

K.4 for peripheral developers. As shown in Table K.3, mechanism analyses further reveal that core developers experienced a 5.6% increase in individual code contributions, a 5.5% increase in developer coding participation, and increases of 4.1%, 1.6%, and 3.6% in code discussion volume, code discussion participation, and code discussion intensity per developer, respectively.

For peripheral developers, as indicated in Table K.4, they experienced a 2.2% increase in individual code contributions, a 2.8% increase in developer coding participation, and increases of 12.6%, 2.3%, and 11.4% in code discussion volume, code discussion participation, and code discussion intensity per developer, respectively.

Table K.1: The Effect of Copilot on Core Developers (Absolute Changes)

	(1) Project-level Code Contributions	(2) Coordination Time	(3) Timely Merged PRs (1 Day)	(4) Timely Merged PRs (3 Days)	(5) Timely Merged PRs (10 Days)
	Log (Core_merged PR)	Log (Core_merge time)	Log (Core_mergedPR_1D)	Log (Core_mergedPR_3D)	Log (Core_mergedPR_10D)
ATT of Copilot	0.061*** (0.008)	0.067*** (0.023)	0.044*** (0.011)	0.048*** (0.010)	0.058*** (0.010)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	6,423	6,423	6,423	6,423	6,423
Observations	118,853	118,853	118,853	118,853	118,853

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table K.2: The Effect of Copilot on Peripheral Developers (Absolute Changes)

	(1) Project-level Code Contributions	(2) Coordination Time	(3) Timely Merged PRs (1 Day)	(4) Timely Merged PRs (3 Days)	(5) Timely Merged PRs (10 Days)
	Log (Peri_merged PR)	Log (Peri_merge time)	Log (Peri_mergedPR_1D)	Log (Peri_mergedPR_3D)	Log (Peri_mergedPR_10D)
ATT of Copilot	0.054*** (0.008)	0.053* (0.028)	0.023** (0.010)	0.024** (0.012)	0.036*** (0.010)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,154	4,154	4,154	4,154	4,154
Observations	77,888	77,888	77,888	77,888	77,888

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table K.3: The Effect of Copilot on Core Developers (Absolute Changes, Mechanism Analyses)

	(1) Individual Code Contributions	(2) Coding Participation	(3) Discussion Volume	(4) Discussion Participation	(5) Individual Discussion Intensity
	Log (MergedPR_per core)	Log (Num_core_with mergedPR)	Log (Comment_per mergedPR)	Log (Num_dev_with_ comments_per mergedPR)	Log (Comments per dev_per mergedPR)
ATT of Copilot	0.056*** (0.011)	0.055*** (0.007)	0.041** (0.019)	0.016*** (0.005)	0.036** (0.017)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	6,423	6,423	6,423	6,423	6,423
Observations	118,853	118,853	118,853	118,853	118,853

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05,
* p<0.1.

Table K.4: The Effect of Copilot on Peripheral Developers (Absolute Changes, Mechanism Analyses)

	(1) Individual Code Contributions	(2) Coding Participation	(3) Discussion Volume	(4) Discussion Participation	(5) Individual Discussion Intensity
	Log (MergedPR_per peri)	Log (Num_peri_with mergedPR)	Log (Comment_per_ mergedPR)	Log (Num_dev_with_ comments_per mergedPR)	Log (Comments per dev_per mergedPR)
ATT of Copilot	0.022*** (0.006)	0.028*** (0.005)	0.126*** (0.033)	0.023*** (0.007)	0.114*** (0.026)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	4,154	4,154	4,154	4,154	4,154
Observations	77,888	77,888	77,888	77,888	77,888

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05,
* p<0.1.

Appendix L: Project Familiarity

We argue that one potential reason that explains the difference in how core and peripheral developers benefit from Copilot could lie in their different levels of project familiarity. In the OSS community, core developers are more embedded in focal projects and maintain sustained involvement (Setia et al. 2012), while peripheral developers contribute more sporadically (Howison and Crowston 2014, Krishnamurthy et al. 2016).

To explore this potential explanation, we first compare the level of project familiarity for core versus peripheral developers. Based on the sample of repositories with continuous activity from 2018 to 2021, we measure a developer’s project familiarity based on their activity during this period related to a focal repository. More specifically, a developer’s project familiarity with a focal repository is measured by the developer’s tenure with the repository, i.e., the number of months a developer was active in the repository before June 2021 (i.e., the introduction of Copilot to GitHub). As reported in Table L.1, the t-test results show that on average, core developers have significantly longer tenure than peripheral developers, consistent with our argument that core developers have greater project familiarity.

Table L.1: Comparison between Core and Peripheral Developers on their Project Familiarity,
Measured by Project Tenure

Variable	Core Developers	Peripheral Developers	Mean diff	p-value
Avg tenure	6.98	1.64	5.34	0.00

Notes: The table is based on the same sample as the one used in our main analyses in the paper.

We next explore whether the differential effects of Copilot on project-level code contributions and coordination time among core versus peripheral developers can be partly explained by the different levels of project familiarity. To do so, our main idea is to examine whether Copilot had differential effects on various outcome variables for developers with different levels of familiarity. Developers are classified into high versus low project familiarity based on whether their tenure exceeds the median value of tenure for that repository before the adoption of Copilot.

Using the same empirical approach as Section 6.5, we compute 1) the proportion of project-level code contributions made by developers with high project familiarity to all contributions, labeled as *Ratio*

of merged PR, and 2) the ratio of coordination time for merging code contributed by high-familiarity developers to the average coordination time, labeled as *Ratio of merge time*. Using each of these two variables as the dependent variable, the results are shown in Tables L.2. Consistent with our arguments, Copilot led to an increased proportion of code contributions from high-familiarity developers; it also led to a relative decrease in coordination time for merging code contributed by high-familiarity developers compared to the average coordination time.

Table L.3 reports the results related to the mechanism analyses. It indicates Copilot led to higher individual code contributions (column [1]) and developer coding participation (column [2]) by high-familiarity developers. Meanwhile, as shown in columns (3) to (5), following the adoption of Copilot, there was a relative decline in code discussion volume, discussion participation, and discussion intensity on code contributed by high-familiarity developers, compared to the average code discussion level. Collectively, these results support our argument that the differential effects of Copilot on project-level code contributions and coordination time among core versus peripheral developers might be partly explained by their different levels of project familiarity.

Table L.2: Heterogeneous Effect of Copilot on Developers with High vs. Low Project Familiarity

	(1)	(2)
	Project-level Code Contributions	Coordination Time
	Ratio of merged PR	Ratio of merge time
ATT of Copilot	0.024*** (0.004)	-0.052*** (0.009)
Repository FE	Yes	Yes
Month FE	Yes	Yes
# of repositories	6,171	6,171
Observations	87,048	87,048

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table L.3: Heterogeneous Effect of Copilot on Developers with High vs. Low Project Familiarity
(Mechanism)

	(1) Individual Code Contributions	(2) Developer Coding Participation	(3) Discussion Volume	(4) Discussion Participation	(5) Individual Discussion Intensity
	mergedPR_per_dev_ratio	dev_with_mergedPR_share	comment_per_mergedPR_ratio	comment_dev_per_mergedPR_ratio	comment_per_dev_per_mergedPR_ratio
ATT of Copilot	0.174*** (0.008)	0.007* (0.004)	-0.106*** (0.005)	-0.099*** (0.005)	-0.103*** (0.005)
Repository FE	Yes	Yes	Yes	Yes	Yes
Month FE	Yes	Yes	Yes	Yes	Yes
# of repositories	6,171	6,171	6,171	6,171	6,171
Observations	87,048	87,048	87,048	87,048	87,048

Notes: All estimations are based on the GSCM method. Robust standard errors clustered at repository level in parentheses. *** p<0.01, ** p<0.05,
* p<0.1.

Appendix M: Alternative Explanations for the Differential Impacts between Core and Peripheral Developers

To address the concern that the heterogeneous effects of Copilot on core versus peripheral developers may be driven by differences such as programming language expertise or Copilot usage, we conduct the following additional analyses.

First, it is worth noting that the theoretical distinction between core and peripheral developers does not necessarily reflect differences in programming language expertise. As suggested by the literature, peripheral developers include core developers from other projects and end users (Setia et al. 2012), which indicates that peripheral developers may also possess substantial expertise in the relevant programming languages. Core developers may be new to a domain and rely on peripheral contributors with specialized knowledge to support project development. Alternatively, core developers may be domain experts, while peripheral developers participate in the project to learn and gain experience.

Given the plausibility of both scenarios, we compare the fraction of expert developers in core developers against the fraction of expert developers in peripheral developers. Specifically, we classify developers as programming language experts if they used that language in at least 75% of their prior projects. We then compare each developer's expertise to the primary language used in the focal project to determine whether they met the expert criteria.

As reported in the first row in Table M.1, the t-test result shows no significant difference between core and peripheral developers in terms of the percentage of expert developers. This suggests that differences in language expertise may not be the primary reason for the observed heterogeneity between core and peripheral contributors.

Second, we consider the possibility that core developers benefit more simply because they are more likely to use Copilot or use it more intensively. To explore this possibility, we leverage the proprietary data provided by GitHub organization that indicates for each repository, the percentage of Copilot users among core developers and the percentage of Copilot users among peripheral developers. Due to privacy

constraints, we do not have access to individual-level Copilot usage data for specific core or peripheral developers.

As reported in the second row in Table M.1, interestingly, we find that peripheral developers had a significantly higher Copilot adoption rate than core developers. Despite this, our analysis in Section 6.5 of the main paper shows that core developers experienced relatively greater project-level code contributions and lower coordination time from Copilot usage. This pattern further supports our argument that the observed differences in outcomes are not driven by differential Copilot adoption rates but are more plausibly attributed to differences in project familiarity.

Table M.1: Comparison between Core and Peripheral Developers

Variable	Core Developers	Peripheral Developers	Mean diff	p-value
% of expert developers	0.68	0.67	0.01	0.15
% of Copilot users	0.20	0.30	-0.10	0.00

Notes: The table is based on the same sample as the one used in our main analyses in the paper.

References

- Abadie A, Diamond A, Hainmueller J (2015) Comparative politics and the synthetic control method. *American Journal of Political Science* 59(2):495-510.
- Alyakoo M, Rahman MS (2022) Shared prosperity (or lack thereof) in the sharing economy. *Information Systems Research* 33(2):638-658.
- Angrist JD, Pischke J-S (2009) *Mostly harmless econometrics: An empiricist's companion* (Princeton university press).
- Autor DH (2003) Outsourcing at will: The contribution of unjust dismissal doctrine to the growth of employment outsourcing. *Journal of labor economics* 21(1):1-42.
- Bai J (2009) Panel data models with interactive fixed effects. *Econometrica* 77(4):1229-1279.
- Chengalur-Smith I, Sidorova A, Daniel SL (2010) Sustainability of free/libre open source projects: A longitudinal study. *Journal of the Association for Information Systems* 11(11):5.
- Egami N, Yamauchi S (2023) Using multiple pretreatment periods to improve difference-in-differences and staggered adoption designs. *Political Analysis* 31(2):195-212.
- Griffiths TL, Steyvers M (2004) Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl_1):5228-5235.
- Hann I-H, Roberts JA, Slaughter SA (2013) All are not equal: An examination of the economic returns to different forms of participation in open source software communities. *Information Systems Research* 24(3):520-538.
- Hansen S, McMahon M, Prat A (2018) Transparency and deliberation within the FOMC: A computational linguistics approach. *The Quarterly Journal of Economics* 133(2):801-870.
- Hartman E, Hidalgo FD (2018) An equivalence approach to balance and placebo tests. *American Journal of Political Science* 62(4):1000-1013.
- Howison J, Crowston K (2014) Collaboration through open superposition: A theory of the open source way. *Mis Quarterly* 38(1):29-50.
- Krishnamurthy R, Jacob V, Radhakrishnan S, Dogan K (2016) Peripheral developer participation in open source projects: an empirical analysis. *ACM Transactions on Management Information Systems (TMIS)* 6(4):1-31.
- Lakens D, Scheel AM, Isager PM (2018) Equivalence testing for psychological research: A tutorial. *Advances in methods and practices in psychological science* 1(2):259-269.
- Lee GM, Qiu L, Whinston AB (2016) A friend like me: Modeling network formation in a location-based social network. *Journal of Management Information Systems* 33(4):1008-1033.
- Lee GM, He S, Lee J, Whinston AB (2020) Matching mobile applications for cross-promotion. *Information Systems Research* 31(3):865-891.
- Liu L, Wang Y, Xu Y (2024) A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science* 68(1):160-176.
- Lu Y, Gupta A, Ketter W, Van Heck E (2019) Information transparency in business-to-business auction markets: The role of winner identity disclosure. *Management Science* 65(9):4261-4279.
- Medappa PK, Srivastava SC (2019) Does superposition influence the success of FLOSS projects? An examination of open-source software development by organizations and individuals. *Information Systems Research* 30(3):764-786.
- Pan Y, Qiu L (2022) How ride-sharing is shaping public transit system: A counterfactual estimator approach. *Production and Operations Management* 31(3):906-927.
- Ramage D, Dumais S, Liebling D (2010) Characterizing microblogs with topic models. *Proceedings of the international AAAI conference on web and social media*, 130-137.
- Setia P, Rajagopalan B, Sambamurthy V, Calantone R (2012) How peripheral developers contribute to open-source software development. *Information Systems Research* 23(1):144-163.
- Shi Z, Lee GM, Whinston AB (2016) Toward a better measure of business proximity. *MIS quarterly* 40(4):1035-1056.

- Shin D, He S, Lee GM, Whinston AB, Cetintas S, Lee K-C (2020) *Enhancing social media analysis with visual data analytics: A deep learning approach* (SSRN Amsterdam, The Netherlands).
- Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 448-456.
- Wang Q, Qiu L, Xu W (2023) Informal Payments and Doctor Engagement in an Online Health Community: An Empirical Investigation Using Generalized Synthetic Control. *Information Systems Research*.
- Weng J, Lim E-P, Jiang J, He Q (2010) Twitterrank: finding topic-sensitive influential twitterers. *Proceedings of the third ACM international conference on Web search and data mining*, 261-270.
- Williams C (2011) Client–vendor knowledge transfer in IS offshore outsourcing: insights from a survey of Indian software engineers. *Information Systems Journal* 21(4):335-356.
- Xu Y (2017) Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1):57-76.
- Zimmermann A, Oshri I, Lioliou E, Gerbasi A (2018) Sourcing in or out: Implications for social capital and knowledge sharing. *The Journal of Strategic Information Systems* 27(1):82-100.