Barcelona School of Economics

Overflow Under-Flowed

ChatGPT's Impact on Stack Overflow Question Patterns

Blanca Jimenez (161331)
Maria Simakova (172708)
Moritz Peist (254017)

Abstract

This study investigates how ChatGPT's release has transformed question patterns on Stack Overflow, combining causal inference with text mining to measure both quantitative and qualitative impacts. Applying the Technology Acceptance Model framework, we analyze how ChatGPT's perceived usefulness and ease of use have reshaped developers' information-seeking behavior. Using a Synthetic Difference-in-Differences approach with data spanning January 2021 to March 2024, we establish that ChatGPT caused a significant 39.5% reduction in scripting language questions (JavaScript, Python, R, and PHP). Beyond this volumetric decline, we demonstrate a statistically significant increase in question complexity following ChatGPT's introduction. Our TF-IDF analysis reveals meaningful linguistic shifts: terms related to troubleshooting and technical infrastructure increased in importance, while basic programming concepts declined significantly. These findings align with recent research suggesting developers strategically allocate questions between platforms based on perceived usefulness for specific query types. Our research provides empirical evidence of how large language models reshape knowledge-sharing dynamics in technical communities, pointing to a complementary relationship between AI tools and human-moderated forums.

March 25, 2025

Contents

1 Introduction **1**

1.1 Research Questions and Approach 1

2 Literature Review **2**

2.1 General Technology Adaptation Literature 2

2.2 Empirical Literature on AI Coding Assistants 2

3 Causal Impact Analysis **4**

3.1 Synthetic Difference-in-Differences Methodology 4

3.1.1 Model Specification 5

3.2 Causal Impact Results 5

3.2.1 Base DiD Estimates 6

3.2.2 Synthetic DiD Results 6

3.2.3 Event Study Analysis 6

3.3 Robustness and Potential Confounders 7

3.4 Implications for Text Analysis 7

4 Natural Language Processing Methodology **8**

4.1 Complexity Analysis 8

4.2 Term Importance Analysis 9

5 Results **9**

5.1 Question Complexity Impact 9

5.2 Analysis of Question Content Changes 10

5.3 Interpretation 11

References **12**

A Difference-in-Difference Question Counts **14**

B Processing Pipeline **17**

C Difference-in-Difference Complexity Scores **18**

1 Introduction

The advent of Large Language Models (LLMs) has triggered a paradigm shift in how individuals seek and obtain technical information. Stack Overflow, as the premier programming question-and-answer platform, has long been the go-to resource for developers facing coding challenges. However, with the public release of ChatGPT in November 2022, developers gained access to an AI assistant capable of providing immediate, contextual programming guidance—potentially disrupting established knowledge-seeking patterns on specialized forums.

This paper investigates how ChatGPT's introduction has altered the landscape of programming questions on Stack Overflow, employing natural language processing (NLP) techniques to identify and quantify changes in question content, complexity, and topical focus. We first establish the causal impact of ChatGPT on Stack Overflow question volumes using difference-in-differences methods, then leverage this foundation to conduct a comprehensive NLP analysis of the changing patterns in developer queries.

Our dataset spans January 2021 to March 2024, encompassing over 1.3 million questions from Stack Overflow. The initial dataset size was around 100 GB, and still 2.5 GB of data, or roughly 4 million rows, for our selected time frame, including all questions. Thus, we focused on scripting languages (i.e., JavaScript, Python, R, and PHP) (Stack Overflow, 2025).

The focus on scripting languages has also several other advantages, i.e.: (1) We effectively reduce the dataset size and thus also pre-processing times significantly. (2) The selected languages are also the top-contributing programming languages on the platform itself and thus representative of the overall platform trends. (3) The chosen languages are those which, assumedly, saw the largest impact as they are high-level languages that earlier ChatGPT versions were especially trained on and sound in. For the causal analysis component, we incorporate data from four non-programming Stack Exchange forums (Mathematics, Physics, Superuser, and AskUbuntu) as control units (approx. 0.5 million questions) without any exclusions (Internet Archive, 2024).

1.1 Research Questions and Approach

Thus, our primary research questions ask:

1. To what extent has ChatGPT's introduction causally affected question volume on Stack Overflow?
2. How has the nature and complexity of questions changed post-ChatGPT?

We approach these questions through a two-stage methodology. First, we establish causality through a Synthetic Difference-in-Differences (SDID) framework, quantifying the volumetric impact while controlling for temporal trends following Arkhangelsky, Dmitry et al. (2021) and Clarke, Damian et al. (2023). This causal foundation motivates our core NLP analysis, considering changes in term frequencies and re-employing the SDID framework upon a question complexity score to analyze Stack Overflow questions before and after ChatGPT's release.

Therefore, our study applies concepts from the Technology Acceptance Model to understand how ChatGPT's high perceived usefulness and ease of use may have shifted developers' question-asking behavior. We provide empirical evidence of how LLMs reshape technical knowledge sharing dynamics by examining quantitative changes (question volume) and qualitative changes (complexity and content).

2 Literature Review

This question can be approached through the Technology Adoption field, which is the conceptual area concerned with user acceptance processes of new technologies, especially those related to Management Information Systems.

2.1 General Technology Adaptation Literature

The current longest-standing and most hegemonic framework is Roger Davis' Technology Acceptance Model (TAM) which defined two main predictors of end-user attitudes towards a new technology: Perceived Usefulness (PU) and Perceived Ease of Use (PEOU). Through the later empirical literature, PU has been found to have the strongest relationship with positive attitudes towards technology. PEOU's predictive ability, on the other hand, has always been found to be weaker, and also, been found to be smaller and smaller as new studies have been made (Davis, Fred D., 1985; Davis, Fred D., 1989; Kelly, Sage et al., 2023).

In 2003, a formal expansion of the TAM, the Unified Theory of Acceptance and Use of Technology (UTAUT), was suggested by Venkatesh et al. (2003). It suggested four predictors instead of two: performance expectancy, social influence, effort expectancy, and facilitating conditions. It aimed to address the limitations of TAM by incorporating external influences and structural factors that affect technology adoption. It has since been widely applied, with researchers often modifying or extending it to better fit specific technological contexts (Venkatesh et al., 2003; Hasan Emon, Md Mehedi, 2023).

Throughout the literature, many authors have applied extended versions (including additional traits) of either the TAM or the UTAUT. In an extensive revision of the literature, Kelly, Sage et al. (2023) find "trust" and "attitudes" to be the most widely applied in empirical studies. "Trust" refers to both trust in the technology itself and trust in the provider and can be both a predictor of PU or a direct predictor of acceptance. "Attitudes" refers to the subjective sentiment towards the technology and can both influence and be influenced by PU and PEOU and, therefore, act as a direct predictor of acceptance in case of the latter.

Many authors have also included personal and cultural characteristics as predictors of attitude, including age, religion, and social context, and, in relation to this, the cognitive process has emerged as a widely considered factor. The element of cognitive process includes a combination of social characteristics and past experiences with new or similar technologies that contribute to the individual's perception of themselves and of the adoption process, thus influencing intention of use.

2.2 Empirical Literature on AI Coding Assistants

Across multiple empirical studies, developers report that AI programming assistants offer significant utility in their workflow, indicating high perceived usefulness. Bird, Christian et al. (2023) observed that early Copilot users found the tool helpful for a range of coding tasks beyond simple autocompletion (ibid.). Use cases observed include delegating tedious tasks such as generating unit tests, boilerplate code, and code comments as well as providing essential assistance in the use of languages or tasks which the developer is unfamiliar with (Bird, Christian et al., 2023; Sergeyuk, Agnia et al., 2025). This had key implications on respondents' perceived productivity.

Moreover, their support is multipurpose and versatile, offering assistance in practically the whole spectrum of potential issues within the same conversation. Not only does it generate code but it also engages in dialogue to explain concepts or debug errors. Its conversational

ability allows developers to obtain detailed explanations and alternative solutions, thereby enhancing its role as an effective answering agent (Kabir, Samia et al., 2023). The tool's extensive knowledge base and real-time responsiveness enable users to address both simple and complex queries without delay that can be immediately followed up with clarifications or further questions in a single session, therefore maximizing the possibility of obtaining an effective answer within a short period of time.

AI coding assistants also significantly reduce the effort required to obtain coding assistance, giving them a very high Perceived Ease of Use. Unlike traditional Q&A forums, which often demand well-structured inquiries and involve waiting periods for responses, ChatGPT provides immediate, natural-language answers, thereby minimizing help-seeking friction. The conversational format streamlines the process and eliminates the social risks associated with public forums, as users can ask questions privately without fear of judgment or negative feedback (ibid). Furthermore, the low learning curve related to these tools enables even novice developers to use them effectively immediately (Bird, Christian et al., 2023). Overall, the high perceived ease of use and the efficiency of obtaining timely answers reinforce the adoption of AI assistants as integral components of modern development workflows.

Despite these strengths, limitations remain that temper the usefulness of AI assistants. Both Copilot and ChatGPT are known to produce outputs that may be incorrect or misleading. Kabir, Samia et al. (2023) report that a 52% of answers contained incorrect information while 78% were inconsistent with human-generated answers. Compatibility problems, internal errors, and context misunderstandings were the most reported problems (Zhou, Xiyu et al., 2025). Nonetheless, ease of use and answer presentation still cause developers to use AI in most cases.

This caveat leads us to two different observations. Firstly, it shows how perceived usefulness is a far more important indicator of use than actual usefulness. This idea can be reinforced by the fact that the initial acceptance rate of suggestions, regardless of whether such suggestions make it to the final version of the code, is the most significant predictor of self-reported productivity by programmers. Secondly, this might be the main mechanism through which the change of use of Stack Overflow is affected, with developers now assigning each tool a specific type of question or changing the presentation of their questions.

Based on TAM and UTAUT frameworks, we can predict that ChatGPT's significant perceived usefulness (ability to provide immediate, contextual programming assistance) and ease of use (conversational interface with low friction) would lead to widespread adoption for certain types of programming questions. This adoption would likely cause a redistribution of question types across platforms, with developers strategically choosing between AI assistants and human-moderated forums based on question characteristics and expected value of responses. Following the technology adoption literature, we would expect this redistribution to reflect not just preference for convenience but optimization of outcomes - using each tool for what it does best.

Thus, our thesis is that, through our language analysis on Stack Overflow questions, we will be able to see an increase in the complexity of questions asked as well as a shift in the topics, with a significant reduction in more basic and generalized tasks and an increase in more specific questions and in questions that solve the most prominent issues arising from ChatGPT answers (compatibility and context-related problems). This aligns with predictions made by some authors as well as with early observations (Kabir, Samia et al., 2023; Sergeyuk, Agnia et al., 2025; Zhou, Xiyu et al., 2025), but these are mainly based on qualitative analysis based on personal interviews. Our work expects to answer this question based on quantitative analysis and large-scale direct analysis of Stack Overflow questions.

3 Causal Impact Analysis

Before conducting our primary text mining analysis, we first establish the causal impact of ChatGPT on Stack Overflow question volumes. This section outlines our causal inference methodology and findings, which provide a critical context for interpreting the subsequent textual analysis results.

3.1 Synthetic Difference-in-Differences Methodology

To identify the causal impact of ChatGPT on Stack Overflow question volumes, we employ a Synthetic Difference-in-Differences (SDID) approach (Arkhangelsky, Dmitry et al. (2021)). This methodology combines the strengths of traditional difference-in-differences and synthetic control methods, allowing us to construct a credible counterfactual for Stack Overflow in the absence of ChatGPT.

The selection of Mathematics, Physics, Superuser, and AskUbuntu as control units was strategically motivated by several considerations. First, these Stack Exchange forums represent technical knowledge domains with structured question patterns similar to Stack Overflow. Yet, they address distinct subject matters less effectively handled by early ChatGPT versions. While ChatGPT demonstrated strong capabilities in programming tasks from its initial release, it exhibited notable limitations in advanced mathematics, physics reasoning, and system-specific troubleshooting—areas central to our control forums. Second, these forums maintain sufficient question volumes to provide statistical power while exhibiting pre-treatment correlation with Stack Overflow question patterns (cf. Figure 4 in Appendix A), suggesting similar responsiveness to seasonal trends and external factors affecting forum usage.

A critical assumption for traditional difference-in-differences analysis is the parallel trends assumption, which requires that treatment and control groups would follow similar trajectories in the absence of treatment. While examining raw counts shows substantial scale differences between forums, for visualization purposes, we thus use an indexed count (cf. Figure 1). However, in the following regressions, we use log transformations for interpretability.

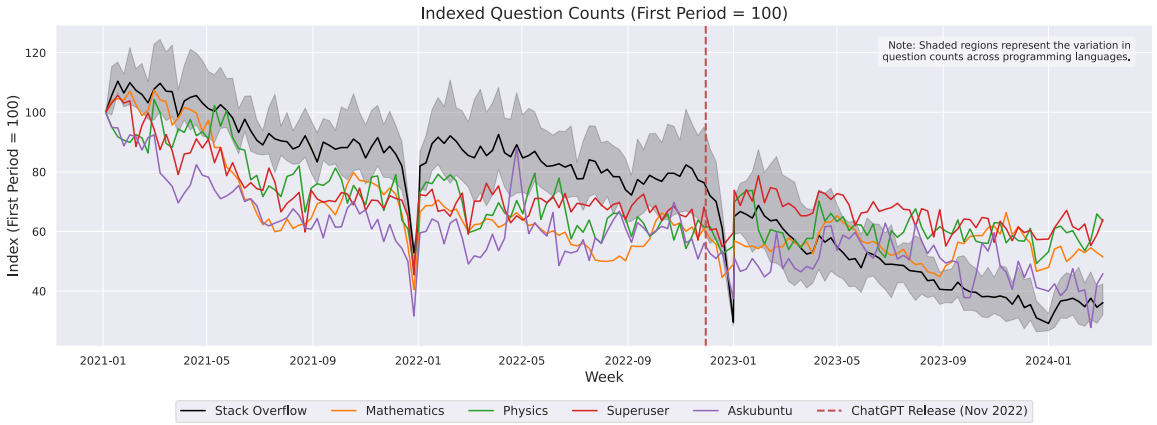


Figure 1: Parallel trends: Weekly indexed question counts

Despite the similar pre-treatment trends, i.e., all Stack Exchanges saw slightly decreasing question counts, concerns remain about external shocks that might differentially affect Stack Overflow and control forums. Our methodological approach addresses this in two complementary ways. First, our baseline DiD implementation using Stata's `xtdidregress` automatically adjusts for both panel (forum) effects and time effects in calculating the treatment effect.

We verified the robustness of these results by explicitly testing specifications with additional time-fixed effects at various granularities (weekly, monthly, quarterly), which yielded identical treatment effect estimates.

Second, our synthetic DiD approach further strengthens causal identification by constructing a weighted combination of control units that better approximates the counterfactual for Stack Overflow. This method implicitly accounts for time-varying factors to the extent they similarly affect both treatment and control units, creating a more credible counterfactual than standard DiD approaches relying solely on parallel trend assumptions. Together, these approaches provide robust evidence that our findings reflect the causal impact of ChatGPT rather than coincidental time-specific shocks.

The same methodological approach is applied to both question volume and complexity score analyses, allowing for consistent causal inference across both dimensions.

3.1.1 Model Specification

Our base difference-in-differences (DiD) model can be expressed as:

$$Y_{it} = \beta_1(Treatment_i \times Post_t) + \gamma_i + \lambda_t + \varepsilon_{it} \tag{1}$$

where Y_{it} represents either the log-transformed question count $\ln(Q_{it})$ for volume analysis or the standardized complexity score for complexity analysis (cf. 4.1) for forum i at time t. Treatment_i is an indicator for Stack Overflow (being 1 in the case of Stack Overflow and 0 otherwise), Post_t is an indicator for periods after ChatGPT's release (November 30, 2022, being 1 after and 0 otherwise). Furthermore, γ_i represents forum fixed effects, λ_t represents time fixed effects, and ε_{it} is the error term. Lastly, the coefficient β_1 captures the average treatment effect on the treated (ATET) - the causal impact of ChatGPT on Stack Overflow question volume.

For our synthetic DiD approach, we follow Arkhangelsky, Dmitry et al. (2021), where the estimator can be expressed as:

$$\hat{\tau}_{\text{SDID}} = \sum_{t=T_0+1}^{T} \lambda_t \left(Y_{1t} - \sum_{j=2}^{J} \omega_j Y_{jt} \right) - \sum_{t=1}^{T_0} \lambda_t \left(Y_{1t} - \sum_{j=2}^{J} \omega_j Y_{jt} \right) \tag{2}$$

where Y_{jt} represents the log-transformed question count $\ln(Q_{it})$ for volume analysis or the standardized complexity score for complexity analysis (cf. 4.1) for forum i at time t, ω_j are unit weights, λ_t are time weights, T_0 is the last pre-treatment period. Finally, unit $j = 1$ represents Stack Overflow. We implemented this methodology using Clarke, Damian et al. (2023) and Ciccia, Diego (2024)'s Stata implementation to ensure the robustness of our findings and conducted both static SDID analysis and dynamic event study specifications.

3.2 Causal Impact Results

The following section provides an overview of our motivating Difference-in-Difference approach to post ChatGPT question volumes on Stack Overflow.

3.2.1 Base DiD Estimates

We begin with standard DiD estimates for both all Stack Overflow questions and specifically for scripting language questions (JavaScript, Python, R, and PHP). Table 1 presents the base DiD results. These results indicate statistically significant negative effects across all model specifications. The standard DiD model (Table 1) indicates a 21.4% decrease in question volume for all Stack Overflow questions. In contrast, for scripting language questions, we observe a much more significant decline of 35.8% in the standard DiD model and 39.5% in the synthetic DiD model. This differential impact suggests that ChatGPT has been particularly effective at addressing programming questions related to these popular scripting languages.

Table 1: Basic DiD Estimates of ChatGPT's Impact on Stack Overflow Question Volume

	Dependent variable: Log Question Count	
	All Questions	Scripting Languages
Treatment Effect	-0.241^{***}	-0.443^{***}
	(0.034)	(0.054)
Time fixed effects	Yes	Yes
Group fixed effects	Yes	Yes
Percent Change	-21.4%	-35.8%
Observations	830	1,328
Number of groups	5	8
Pre-treatment periods	101	101
Post-treatment periods	65	65

Notes: Standard errors in parentheses, clustered by forum/group. $^{***}p < 0.001$. The dependent variable is log question count. Treatment is defined as the period after ChatGPT's release (November 30, 2022).

3.2.2 Synthetic DiD Results

To address potential violations of the parallel trends assumption and create a more credible counterfactual, we employ the SDID approach, which yields larger treatment effect estimates compared to standard DiD, suggesting that the traditional DiD may underestimate the impact. The SDID results indicate a 26.7% reduction in overall question volume and a substantial 39.5% reduction in scripting language questions (covariate-adjusted results are 26.5% and 39.2%, respectively).

With additional details in the Appendix A, Figure 5a visualizes the results for all Stack Overflow questions, while Figure 5b focuses on scripting language questions, and Table 3 presents the formal SDID estimates.

3.2.3 Event Study Analysis

We conduct a synthetic event study analysis to explore how the treatment effect evolved over time, following Ciccia, Diego (2024). Unlike standard event study approaches, the synthetic difference-in-differences (SDID) event study estimator can be expressed as:

$$\hat{\tau}_{\ell}^{sdid} = \sum_{a \in A_{\ell}} \frac{N_{tr}^{a}}{N_{tr}^{\ell}} \hat{\tau}_{a,\ell}^{sdid} \tag{3}$$

where $\hat{\tau}_{a,\ell}^{sdid}$ represents the dynamic treatment effect ℓ periods after treatment for cohort a:

$$\hat{\tau}_{a,\ell}^{sdid} = \frac{1}{N_{tr}^{a}} \sum_{i \in I^a} Y_{i,a-1+\ell} - \sum_{i=1}^{N_{co}} \omega_i Y_{i,a-1+\ell} - \sum_{t=1}^{a-1} \left(\frac{1}{N_{tr}^{a}} \sum_{i \in I^a} \lambda_t Y_{i,t} - \sum_{i=1}^{N_{co}} \omega_i \lambda_t Y_{i,t} \right) \tag{4}$$

In our application, Y_{it} is the log-transformed question count for forum i at time t, with the different programming languages on Stack Overflow as the treated units (corresponding to I^a). The weights ω_i and λ_t are optimally chosen to create a synthetic control that best approximates Stack Overflow's pre-treatment outcome trajectory. For each relative time period ℓ, the estimator compares the difference between actual and synthetic outcomes to their pre-treatment average difference.

This approach allows us to examine treatment effects at specific time points relative to ChatGPT's release (November 30, 2022), with $\ell < 0$ for pre-treatment periods and $\ell > 0$ for post-treatment periods. By implementing this within the synthetic DiD framework, each relative time coefficient is estimated using the synthetic control method, comparing Stack Overflow to a weighted combination of control forums optimized for that specific time period. Table 4 presents the event study estimates across different time periods, while Figure 6 in the Appendix A displays the results for scripting language questions.

The event study reveals several important patterns: (1) An immediate and substantial drop in question volume following ChatGPT's release. (2) Persistence of the effect throughout the post-treatment period. (3) Intensification of the effect over time, with the most recent period showing the strongest effect, suggesting continued adoption of ChatGPT for programming assistance.

3.3 Robustness and Potential Confounders

The stability of our estimates across different model specifications provides strong evidence of robustness. Adding monthly covariates yields nearly identical treatment effects for both all questions (-0.308 vs. -0.311) and scripting languages (-0.497 vs. -0.502). While our methodology addresses many potential confounders, some limitations remain: (1) concurrent AI tool releases may have contributed to the observed effects, such as the introduction and evolution of GitHub Copilot; (2) despite controlling for time-invariant forum characteristics and common shocks, forum-specific trends could still influence results; and (3) potential spillover effects may exist if users reduced activity across multiple Stack Exchange forums after discovering ChatGPT. Despite these considerations, the magnitude, immediacy, and persistence of effects—particularly for scripting languages—strongly suggest a causal relationship between ChatGPT's introduction and the decline in Stack Overflow question volume.

3.4 Implications for Text Analysis

These findings establish a causal impact of ChatGPT on Stack Overflow question volumes, particularly for scripting language questions. The differential impact on scripting languages (approximately 39% reduction compared to 27% overall) suggests that ChatGPT has been particularly effective at addressing common programming queries.

This causal foundation motivates our core research question: How has the nature of the remaining questions changed? The dramatic reduction in volume indicates a fundamental shift in how developers seek programming assistance. Still, it raises important questions about the characteristics of questions that continue to be asked on Stack Overflow despite

the availability of ChatGPT. Our subsequent text mining analysis will identify and quantify these changes in question content, complexity, and topical focus.

4 Natural Language Processing Methodology

Building on the established causal impact, we now turn to our primary contribution: a comprehensive NLP analysis of how question content, complexity, and focus have evolved in response to ChatGPT's introduction. This section outlines our NLP methodology for detecting and characterizing these changes. Figure 7 in Appendix B provides a sketched overview of our entire processing pipeline including file names in a flowchart.

4.1 Complexity Analysis

Based on the pre-processed data, we construct a parsimonious complexity score for forum posts which is composed of 4 key elements: (1) title length, (2) body length, (3) number of tags and (4) length of technical expressions (i.e. code, or equation length for the respective forums)[1]:

$$\begin{aligned}\text{Complexity Score}_{i,t} = \\ \frac{1}{4}\left(\frac{\text{TagCount}_{i,t} - \mu_{\text{TagCount}}}{\sigma_{\text{TagCount}}} + \frac{\text{TechExprLength}_{i,t} - \mu_{\text{TechExprLength}}}{\sigma_{\text{TechExprLength}}}\right. \\ \left. + \frac{\text{BodyLength}_{i,t} - \mu_{\text{BodyLength}}}{\sigma_{\text{BodyLength}}} + \frac{\text{TitleLength}_{i,t} - \mu_{\text{TitleLength}}}{\sigma_{\text{TitleLength}}}\right) \quad (5)\end{aligned}$$

Eq. 5 thus shows the complexity score of forum i in time (week) t. The reason for choosing this relatively simple score is that questions often use snippets of code or equations. Thus, more sophisticated and established equation or code complexity algorithms become unusable. Table 2 presents the results over time:

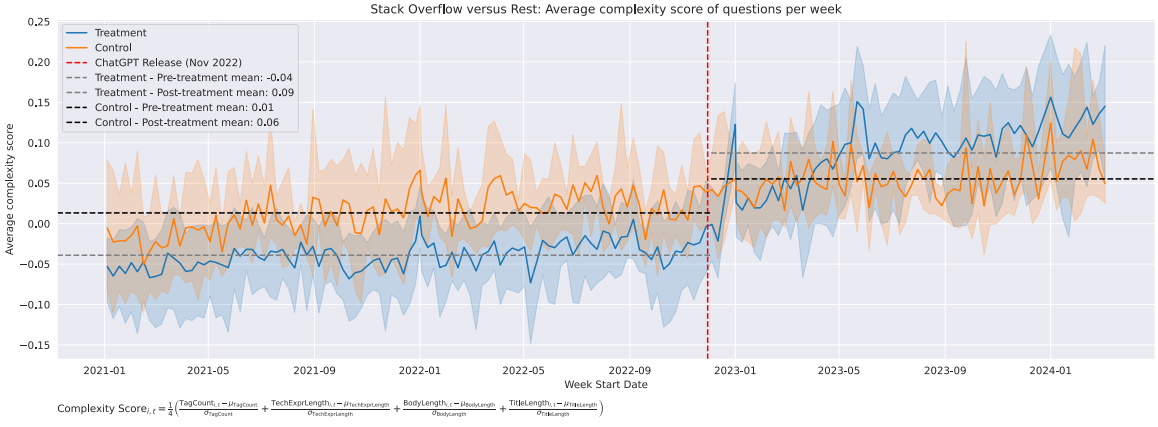


Figure 2: Treatment vs. Control: Question complexity over time

[1]We used different metrics appropriate to each forum's content: code length was measured for Stack Overflow, Superuser, and Askubuntu, while equation length was used for Mathematics and Physics. This distinction reflects the content patterns in our data, as Physics and Mathematics questions contained code in fewer than 5% of posts, while the technology-focused forums rarely contained equations.

4.2 Term Importance Analysis

To understand the nature of the complexity changes identified in our Synthetic DiD analysis, we examine shifts in term importance using TF-IDF (Term Frequency-Inverse Document Frequency) analysis. Unlike raw term frequency, TF-IDF weights terms based on their relative importance within documents and across the corpus, highlighting discriminative terms while down-weighting common ones. Figure 3 shows the significant changes in term importance before and after ChatGPT's introduction. To ensure observed term changes represent genuine shifts rather than random variation, we applied statistical testing using bootstrapped confidence intervals (100 replications) and permutation tests ($\alpha = 0.05$). This approach allowed us to identify statistically significant changes in term importance that correlate with the introduction of ChatGPT.

5 Results

While our NLP analysis remains in progress, our causal findings suggest substantial changes in Stack Overflow usage patterns. Beyond the volume reduction, we observed a fundamental shift in the correlation structure between Stack Exchange forums after ChatGPT's release, as visualized in Figure 4 in the Appendix A.

This dramatic weakening of correlations (from 0.76-0.87 to 0.18-0.65) suggests that ChatGPT has reduced question volume and potentially altered the relationship between programming questions and those in other knowledge domains. This finding motivates our hypothesis that the content and nature of Stack Overflow questions have fundamentally changed in the post-ChatGPT era – a hypothesis we further strengthened by the following results of complexity scores and term frequencies.

5.1 Question Complexity Impact

The average treatment effect on the treated (ATT) indicates a significant increase in our standardized complexity measure of 0.059 standard deviations (cf. Table 2), as shown in Figure 9 in the Appendix C. This effect remains robust when including time-fixed effects and various covariates (ATT = 0.073, SE = 0.010).

Our complexity scores (cf. Eq. 5), capture multiple dimensions of question sophistication, providing a comprehensive measure of question complexity at the individual level. The traditional DiD regression also confirms this effect (cf. Table 5 in Appendix C), with consistent findings across various model specifications. Figure 10 in Appendix C presents the event study results, demonstrating both the immediate impact following ChatGPT's introduction and the persistence of this effect throughout the post-treatment period.

The event study in Figure 10 in Appendix C reveals that while there was an initial positive but non-significant effect in the first twelve weeks after ChatGPT's release, this effect became statistically significant and stronger from week 13 onward. The impact has persisted and strengthened over time, with the largest effect observed in the most recent period (weeks 61-65), suggesting a fundamental shift in how developers utilize Stack Overflow rather than a temporary adjustment.

These findings align with our theoretical predictions derived from the Technology Acceptance Model. The substantial decrease in question volume (39.5% for scripting languages) suggests the high perceived usefulness of ChatGPT for certain question types. In contrast, the shift toward more complex questions on Stack Overflow indicates that developers are optimizing

Table 2: Impact of ChatGPT on Stack Overflow Question Complexity

	Dependent variable: Complexity Score		
	(1)	(2)	(3)
Treatment Effect	0.084***	0.059***	0.073***
	(0.011)	(0.014)	(0.010)
Model	Basic DiD	Synthetic DiD	Synthetic DiD
Time fixed effects	Yes	Yes	Yes
Group fixed effects	Yes	Yes	Yes
Month covariates	No	No	Yes
Observations	1,328	1,328	1,328
Number of groups	8	8	8
Pre-treatment periods	101	101	101
Post-treatment periods	65	65	65

Notes: Standard errors in parentheses, clustered at the group level. * p<0.1, ** p<0.05, *** p<0.01. The dependent variable is the standardized complexity score calculated at the individual question level. Each question's complexity is measured as the average of four z-standardized components: tag count, code length, body length, and title length. Treatment is defined as the period after ChatGPT's release (November 30, 2022). Models 1-2 use traditional DiD specifications, while Model 3 uses synthetic control methods with month covariates.

their information-seeking strategy as predicted by the literature (Kelly, Sage et al., 2023). The reduction in basic programming questions on Stack Overflow mirrors the delegation patterns observed by Bird, Christian et al. (2023) in Copilot users, who reported offloading routine coding tasks to AI assistance.

5.2 Analysis of Question Content Changes

The TF-IDF analysis reveals patterns consistent with our complexity findings (cf. Figure 3). Terms related to troubleshooting and debugging (e.g., "error", "issue", "expect"), as well as technical infrastructure terms ("version", "package", "library") showed significant increases in importance, while terms associated with basic programming concepts ("array", "loop", "list") decreased significantly. These shifts in term importance indicate that ChatGPT has likely absorbed simpler programming questions, leaving Stack Overflow to serve more complex, specific troubleshooting needs.

Notably, conversational terms ("like", "want", etc.) also decreased in importance, suggesting a shift toward more technical, problem-specific language in post-ChatGPT questions. This aligns with our hypothesis that questions remaining on Stack Overflow have become more specialized and technical in nature.

The linguistic changes revealed through our TF-IDF analysis provide qualitative context for the quantitative complexity increases observed in our Synthetic DiD model. The notable shift toward troubleshooting terminology coupled with decreased prevalence of basic programming terms suggests that ChatGPT has fundamentally altered the nature of questions on Stack Overflow, not just their volume or overall complexity.

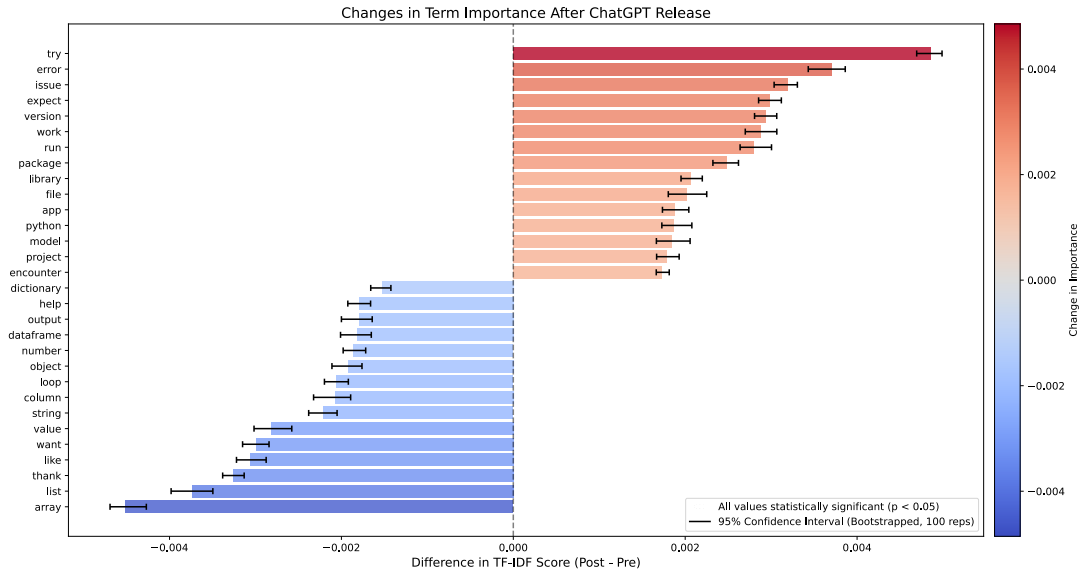


Figure 3: Top 10 increases and decreases in TF-IDF scores after ChatGPT's introduction

5.3 Interpretation

These findings support our hypothesis that ChatGPT has altered information-seeking behavior in programming communities. Developers now appear to reserve simpler questions for ChatGPT while turning to Stack Overflow for more complex programming challenges that require human expertise. The magnitude of this effect—approximately 0.059 standard deviations increase in question complexity—represents a modest but statistically significant shift in the types of questions users bring to Stack Overflow.

To put this effect size in context, it represents a meaningful change in question complexity, particularly given that the complexity measure was calculated at the individual question level using standardized metrics across all questions in our dataset. The gradual increase in effect size over time further suggests that this is not merely a temporary adjustment but rather reflects an evolving shift in how programmers allocate their questions between AI tools and human-moderated forums.

Our findings provide quantitative support for predictions from the Technology Acceptance literature. The substantial decrease in question volume coupled with increased complexity suggests that developers are making strategic choices between platforms based on perceived usefulness - precisely as the TAM framework would predict. The differential impact across question types also aligns with Kabir, Samia et al. (2023) observation that ChatGPT performs better on common programming tasks than on compatibility and context-dependent issues. This pattern of adoption reflects what Venkatesh et al. (2003) described as performance expectancy driving technology acceptance, with users quickly adapting their behavior to leverage ChatGPT's strengths while continuing to rely on human expertise for more challenging problems.

This empirical evidence points to a complementary relationship between AI-powered assistants and human-moderated Q&A forums, with each platform serving distinct informational needs within the programming community. Stack Overflow appears to be evolving toward a repository for more complex programming questions, while more straightforward queries may be increasingly handled through interaction with large language models like ChatGPT.

References

Arkhangelsky, Dmitry et al. (Dec. 2021). "Synthetic Difference-in-Differences". en. In: *American Economic Review* 111.12, pp. 4088–4118. ISSN: 0002-8282. DOI: 10.1257/aer.20190159. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20190159 (visited on 03/07/2025).

Bird, Christian et al. (June 2023). "Taking Flight with Copilot". en. In: *Communications of the ACM* 66.6, pp. 56–62. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/3589996. URL: https://dl.acm.org/doi/10.1145/3589996 (visited on 03/23/2025).

Ciccia, Diego (Nov. 2024). *A Short Note on Event-Study Synthetic Difference-in-Differences Estimators*. arXiv:2407.09565 [econ]. DOI: 10.48550/arXiv.2407.09565. URL: http://arxiv.org/abs/2407.09565 (visited on 03/08/2025).

Clarke, Damian et al. (Feb. 2023). *Synthetic Difference-in-Differences Estimation*. en. SSRN Scholarly Paper. Rochester, NY. URL: https://papers.ssrn.com/abstract=4346540 (visited on 03/08/2025).

Davis, Fred D. (1985). "A technology acceptance model for empirically testing new end-user information systems : theory and results". eng. Thesis. Massachusetts Institute of Technology. URL: https://dspace.mit.edu/handle/1721.1/15192 (visited on 03/09/2025).

— (Sept. 1989). "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology". In: *MIS Quarterly* 13.3, p. 319. ISSN: 02767783. DOI: 10.2307/249008. URL: https://www.jstor.org/stable/249008?origin=crossref (visited on 03/09/2025).

Hasan Emon, Md Mehedi (2023). "INSIGHTS INTO TECHNOLOGY ADOPTION: A SYSTEMATIC REVIEW OF FRAMEWORK, VARIABLES AND ITEMS". In: *Information Management and Computer Science* 6.2, pp. 55–61. ISSN: 26165961. DOI: 10.26480/imcs.02.2023.55.61. URL: https://www.theimcs.org/archives2023/issue2/2imcs2023-55-61.pdf (visited on 03/25/2025).

Internet Archive (July 2024). *Stackexchange directory listing*. en. Archive. URL: https://archive.org/download/stackexchange (visited on 03/08/2025).

Kabir, Samia et al. (2023). "Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions". In: DOI: 10.48550/ARXIV.2308.02312. URL: https://arxiv.org/abs/2308.02312 (visited on 03/09/2025).

Kelly, Sage, Sherrie-Anne Kaye, and Oscar Oviedo-Trespalacios (Feb. 2023). "What factors contribute to the acceptance of artificial intelligence? A systematic review". en. In: *Telematics and Informatics* 77, p. 101925. ISSN: 07365853. DOI: 10.1016/j.tele.2022.101925. URL: https://linkinghub.elsevier.com/retrieve/pii/S0736585322001587 (visited on 03/09/2025).

Sergeyuk, Agnia et al. (Feb. 2025). "Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward". en. In: *Information and Software Technology* 178, p. 107610. ISSN: 09505849. DOI: 10.1016/j.infsof.2024.107610. URL: https://linkinghub.elsevier.com/retrieve/pii/S0950584924002155 (visited on 03/09/2025).

Stack Overflow (2025). *Tags*. en. URL: https://stackoverflow.com/tags (visited on 03/08/2025).

Venkatesh et al. (2003). "User Acceptance of Information Technology: Toward a Unified View". In: *MIS Quarterly* 27.3, p. 425. ISSN: 02767783. DOI: 10.2307/30036540. URL: https://www.jstor.org/stable/10.2307/30036540 (visited on 03/09/2025).

Zhou, Xiyu et al. (Jan. 2025). "Exploring the problems, their causes and solutions of AI pair programming: A study on GitHub and Stack Overflow". en. In: *Journal of Systems and Software* 219, p. 112204. ISSN: 01641212. DOI: 10.1016/j.jss.2024.112204. URL:

https://linkinghub.elsevier.com/retrieve/pii/S0164121224002486 (visited on 03/09/2025).

A Difference-in-Difference Question Counts

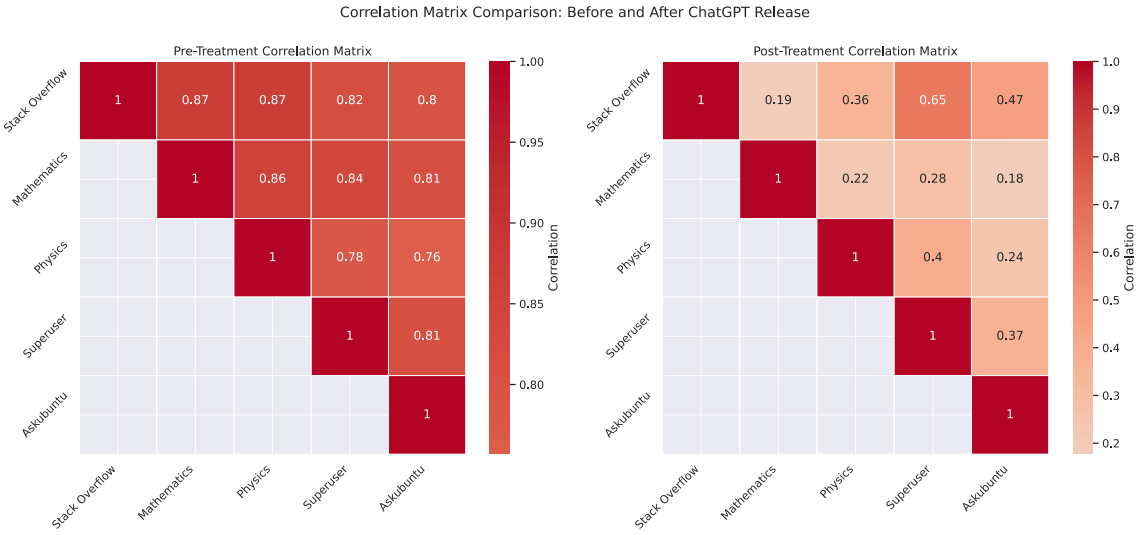
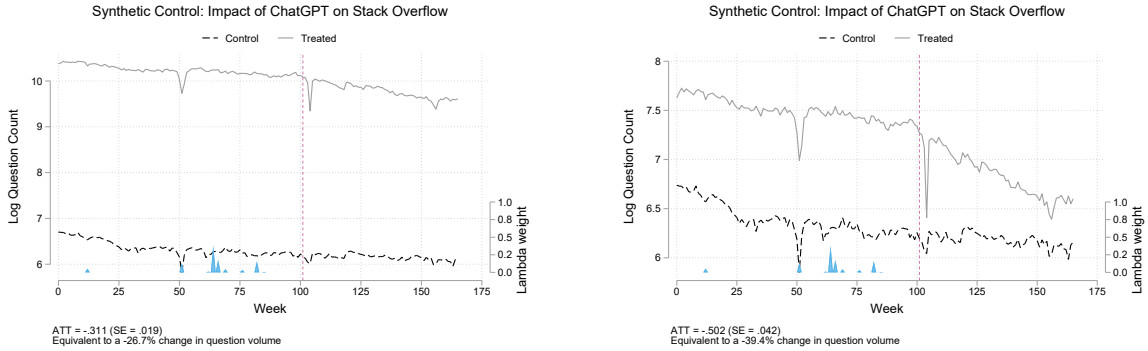


Figure 4: Correlation Matrices Before and After ChatGPT Release



(a) Impact on All Stack Overflow Questions

(b) Impact on Scripting Language Questions

Figure 5: Synthetic difference-in-difference plots

Table 3: Synthetic DiD Estimates of ChatGPT's Impact on Stack Overflow Question Volume

	All Questions		Scripting Languages	
	(1)	(2)	(3)	(4)
Treatment Effect	−0.311***	−0.308***	−0.502***	−0.497***
	(0.019)	(0.000)	(0.042)	(0.039)
Month covariates	No	Yes	No	Yes
Percent Change	−26.7%	−26.5%	−39.5%	−39.2%
Observations	830	830	1,328	1,328
Number of groups	5	5	8	8
Pre-treatment periods	101	101	101	101
Post-treatment periods	65	65	65	65

Notes: Standard errors in parentheses based on placebo/bootstrap replications (100 repetitions). $^{***}p < 0.001$. The dependent variable is log question count. All models include time and group fixed effects. Treatment is defined as the period after ChatGPT's release (November 30, 2022).

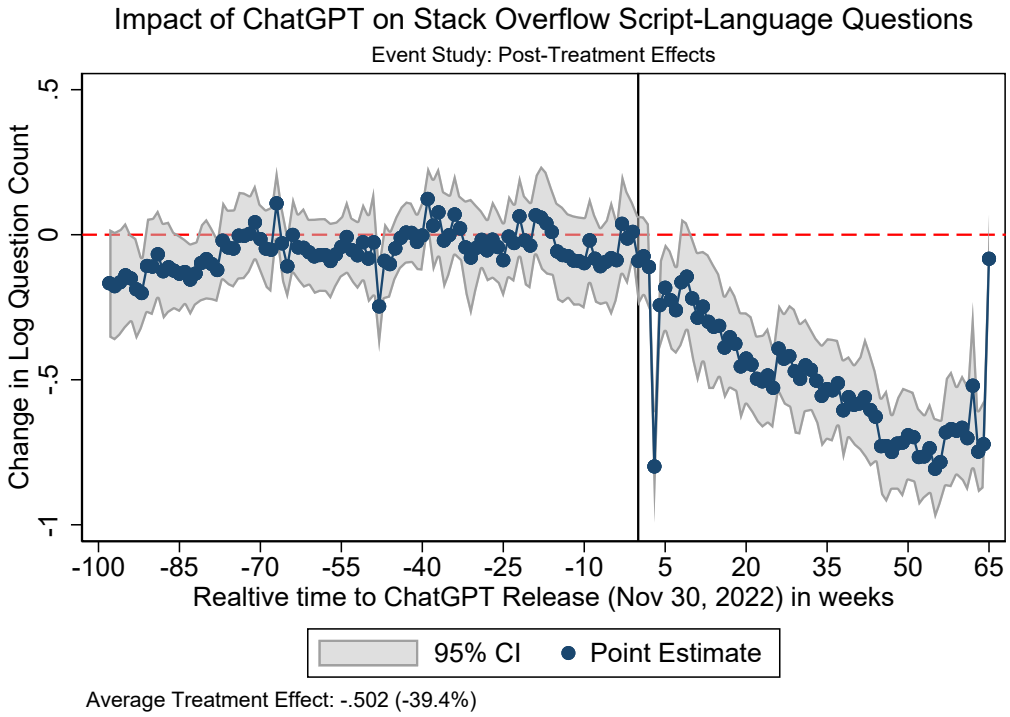


Figure 6: Event Study Analysis for Scripting Language Questions

Table 4: Event Study: Post-Treatment Effects Over Time for Scripting Languages

Time Period	Estimate	SE	95% CI
Week 1-4 after treatment	-0.335	0.079	[-0.490, -0.180]
Week 5-12 after treatment	-0.267	0.084	[-0.432, -0.102]
Week 13-24 after treatment	-0.482	0.080	[-0.639, -0.325]
Week 25-36 after treatment	-0.577	0.082	[-0.738, -0.416]
Week 37-48 after treatment	-0.644	0.077	[-0.795, -0.493]
Week 49-60 after treatment	-0.718	0.081	[-0.877, -0.559]
Week 61-65 after treatment	-0.739	0.077	[-0.890, -0.588]
Overall treatment effect	-0.502	0.073	[-0.646, -0.358]

Notes: Results from synthetic DiD event study analysis with bootstrapped standard errors (100 repetitions). Estimates show the change in log question count for Stack Overflow scripting language questions relative to the synthetic control group across different post-treatment time periods. All effects are statistically significant at the 0.1% level.

B Processing Pipeline

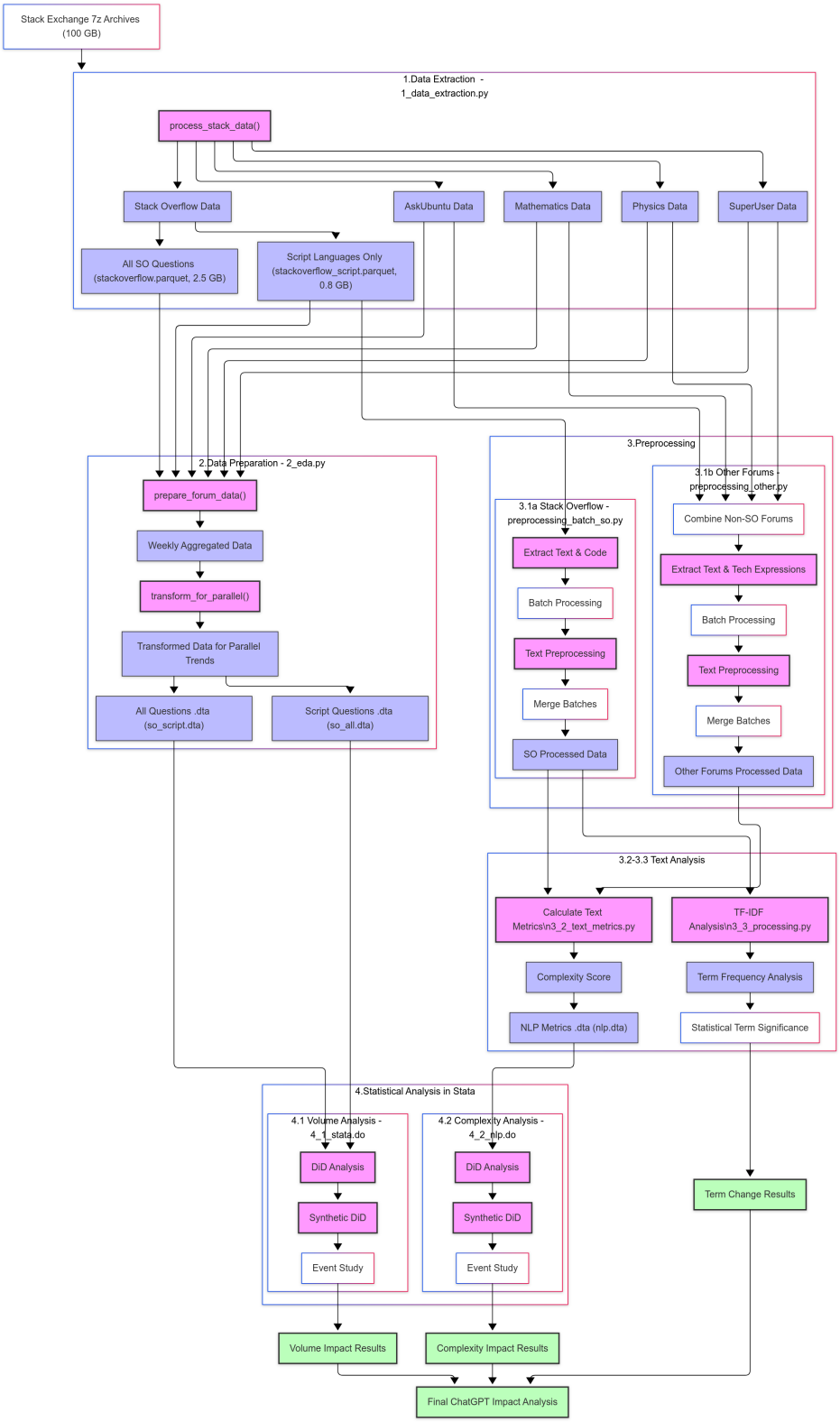
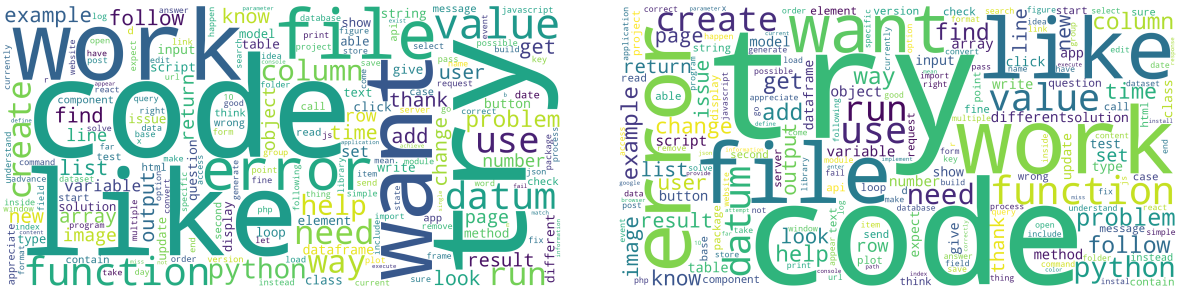


Figure 7: Sketched process and file flow



(a) Pre-treatment wordcloud **(b)** Post-treatment wordcloud

Pre-processing results of wordclouds show no immediate clear trend for Stack Overflow.

C Difference-in-Difference Complexity Scores

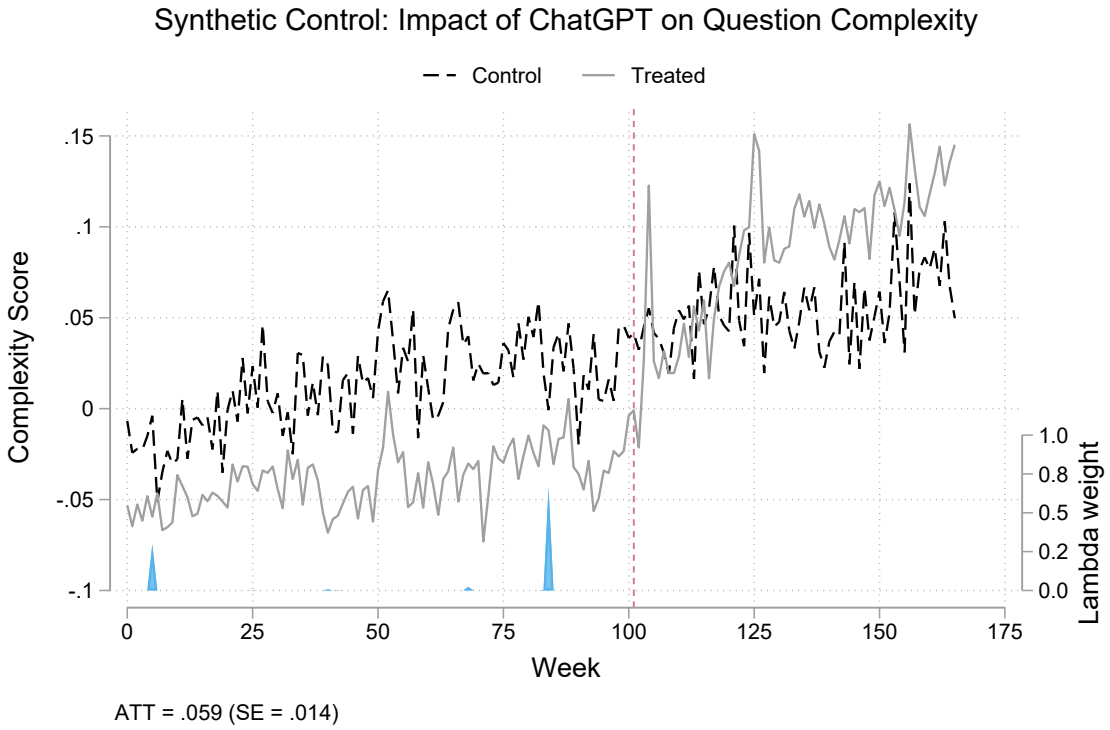


Figure 9: Synthetic DiD: Question complexity

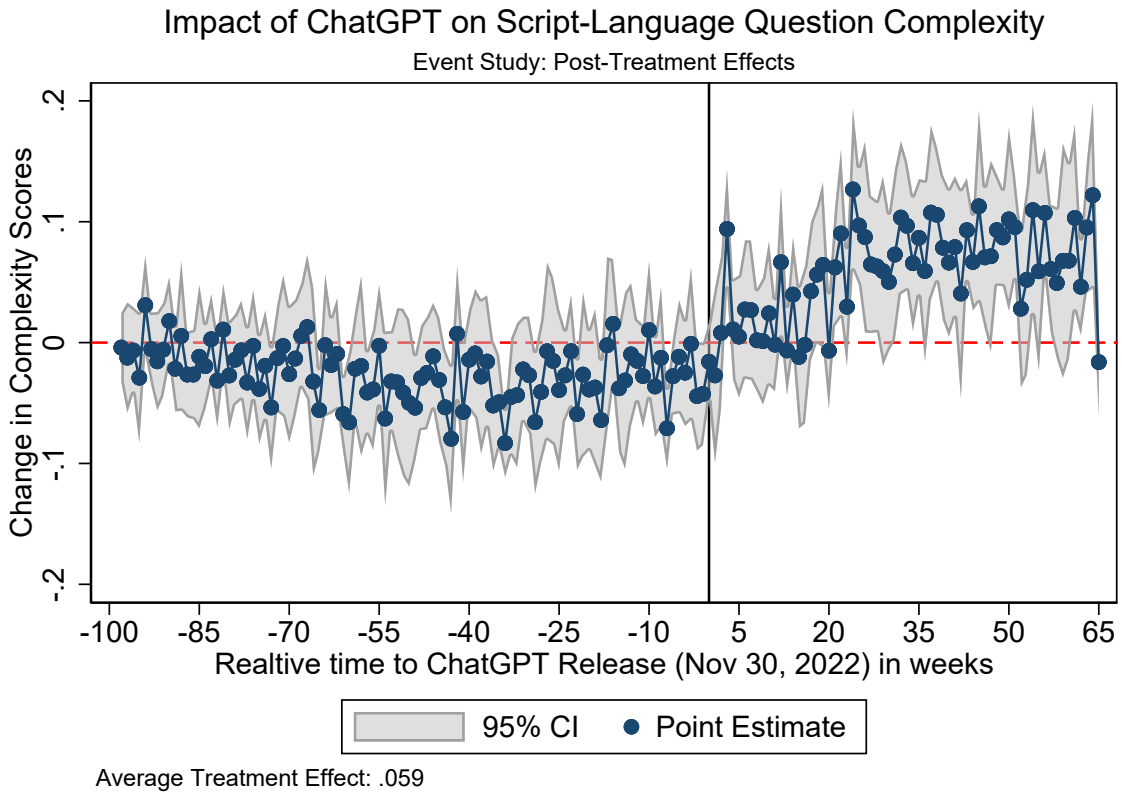


Figure 10: Event study complexity score development

Table 5: Event Study: Post-Treatment Effects Over Time

Time Period	Estimate	SE	95% CI
Week 1-4 after treatment	0.016	0.025	[-0.033, 0.065]
Week 5-12 after treatment	0.028	0.022	[-0.015, 0.071]
Week 13-24 after treatment	0.064	0.018	[0.029, 0.099]
Week 25-36 after treatment	0.078	0.015	[0.049, 0.107]
Week 37-48 after treatment	0.083	0.017	[0.050, 0.116]
Week 49-60 after treatment	0.081	0.019	[0.044, 0.118]
Week 61-65 after treatment	0.092	0.020	[0.053, 0.131]
Overall treatment effect	0.059	0.014	[0.032, 0.086]

Notes: Results from synthetic DiD event study analysis. Estimates show the change in complexity score for Stack Overflow questions relative to the synthetic control group across different post-treatment time periods. Weeks 13-65 effects are statistically significant at the 5% level, while the initial 1-12 week periods show positive but statistically insignificant effects.