Chatbot Overflow - Assessing the Impact of Chat GPT Introduction on Q&A Platforms

Master Thesis in Business Analytics & Management

Rotterdam School of Management

Erasmus University Rotterdam

15th of June 2023

Piotr Andrzej Piwnik (638623)

Academic Year 2022/2023

Thesis Trajectory under supervision of Dr. Dominik Gutt

Co-Reader: Dr. Olga Slivko

Preface

Acknowledgements

I would like to thank my thesis coach, Dr. Dominik Gutt, who inspired me to find a relevant and exceptionally interesting topic, helped me with choosing the right methods for my research, and provided me with constructive feedback throughout the whole thesis trajectory.

Furthermore, I would also like to thank my co-reader, Dr. Olga Slivko, who helped me master my literature review and integrate the existing research on generative AI and Q&A platforms into my study.

Last, but not least, I am exceptionally grateful to my family and friends for their support, which made the research process much easier.

Executive summary

This study investigates the impact of the Chat GPT-3.5 release on the activity on the Stack Exchange forums. A quasi-experimental approach based on the difference-in-differences method is used to determine the causal effect of the introduction of Chat GPT-3.5, treated as an intervention. The treatment group in the study consists of 42 weeks of observations in the years 2022-2023, while the control group consists of observational data from the same weeks in the years 2021-2022. The number of posts, answers, post, and answer scores, as well as post view count, are analysed. Natural Language Processing methods are used for calculating the measures of contemporary content novelty, and readability. Data are aggregated based on the combination of the first two tags attached to a post and on a weekly basis. Results of the Panel Ordinary Least Squares regressions, accounting for fixed group and fixed time effects, suggest that the average treatment effect of Chat GPT introduction is highly negative but not yet significant for the number of posts and answers, post scores and post view count. The results indicate that content posted is significantly more novel and that posts are significantly less readable on average, as a result of Chat GPT release. Implications of the study concern the future of knowledge exchange, claiming that the widespread use of Generative Artificial Intelligence might currently increase the efficiency in problem-solving, however in a longer period of time it is anticipated that usage of those tools will decrease the amount of high-quality, domain-specific knowledge.

Table of Contents

Preface ... **ii**
Acknowledgements ... **iii**
Executive summary ... **iv**
List of Tables ... **vii**
List of Figures ... **viii**
1. Introduction ... **1**
1.1. Problem background and hypothesis formulation ... 2
1.2. Managerial relevance ... 6
1.3. Academic relevance ... 8
1.4. Research approach and structure ... 9
2. Background information ... **10**
2.1. Generative Artificial Intelligence ... 10
2.1.1. Definition of the Generative Artificial Intelligence ... 10
2.1.2. History and development of Generative Artificial Intelligence ... 11
2.2. Role of the Q&A Platforms in knowledge sharing ... 13
3. Related literature ... **14**
3.1. Literature on Generative Artificial Intelligence ... 15
3.2. Literature on online Q&A platforms ... 16
4. Data ... **18**
4.1. Stack Exchange forums ... 18
4.2. Data collection ... 20
4.3. Data transformations ... 21
4.3.1. Data quality checks and text pre-processing ... 21
4.3.2. Feature engineering and data aggregation ... 22
4.4. Exploratory data analysis ... 24
5. Methodology ... **27**
5.1. Difference-in-differences estimation ... 28
5.1.1. Assumptions of the difference-in-differences model ... 29
5.1.2. Econometric specification ... 31
5.2. Natural Language Processing methods ... 33
5.2.1. Content novelty measures ... 33
5.2.2. Content readability measures ... 35

6. Results 37
6.1. Parallel Trends Test 37
6.1.1. Parallel Trends Test results 37
6.1.2. Robustness check 43
6.2. The difference in differences estimation results 44
6.2.1. Results of Ordinary Least Squares regression 44
6.2.2. Heterogeneity analysis 47
6.2.3. Robustness checks 48
7. Discussion 51
7.1. Conclusion 52
7.2. Managerial implications 55
7.3. Academic contributions 57
7.4. Limitations and recommendations 58
8. Appendices 60
9. References 82

List of Tables

Table 1: Counts of unique values of tags, tag combinations, forums present and a number of observations. 24
Table 2: Descriptive statistics of the dependent variables (1/2) 24
Table 3: Descriptive statistics of the dependent variables (2/2), title- and body length 25
Table 4: Formulas used for calculation of the chosen readability metrics and their interpretation 36
Table 5: Results of OLS regressions without the control variables 45
Table 6: Results of the regressions with cosine similarity values 46
Table 7: Values of the Variance Inflation Factor (VIF) 49
Table 8: Results of Poisson regressions 49

List of Figures

Figure 1: Percentage of observations kept in the data set for different percentage values of the most common tag combinations included 23
Figure 2: Total post count share of the five most popular forums in the reduced data set 26
Figure 3: 20 most popular tag combinations in the non-aggregated, reduced data set 27
Figure 4: Weekly post counts in the control and treatment groups 38
Figure 5: Weekly sum of post answers in the control and treatment groups 38
Figure 6: Average post readability in the control and treatment groups 39
Figure 7: Results of Parallel Trends Test for weekly post count 41
Figure 8: Results of Parallel Trends Test for a weekly sum of post answer count 41
Figure 9: Results of Parallel Trends Test for average readability of posts 42

1. Introduction

Over different periods of history, people have constantly sought knowledge in various ways to solve their problems and find answers to their questions. This quest has been extremely difficult throughout most of known history, with only a few well-educated and wealthy individuals possessing almost all the valuable information of any kind. The first books printed by Jan Gutenberg marked the initial turning point, as information sharing got much easier and faster, widening the career opportunities among the clerisy (Burke, 2000, pp. 22–23). Around 300 years later, Industrial Revolution started a process of gradual democratization of knowledge, as society has been equipped with better tools for gathering and sharing knowledge. After the invention of the transistor and the first computers in the mid-20th Century, information technology has become a ubiquitous element of our society and economy, with the aforementioned events marking the beginning of the information age.

Currently, we are in the era of datafication, which can be defined as '*process by which the world is processed, quantified, stored digitally and converted into binary code*' (Posada et al., 2021, p. 1). Entering this era has brought information closer to users than ever before. First and foremost, the rapid digitisation of existing sources of knowledge, such as books, articles and research papers, and then making them freely available online, has made access to these sources easier and more equal than in previous decades. For instance, archive.org (2023), a site dedicated to cataloguing information published on the web, offers free access to 38 million books, while the average US library had only 81,000 books in 2016 (National Centre for Education Statistics, 2016). Moreover, new, purely online knowledge databases developed by communities of users – wikis – have contributed to documenting many areas of science, culture and everyday life, with a notable example of Wikipedia with more than 58 million articles in more than 300 languages (Wikimedia Foundation, 2023). A different sort of Internet community – Question & Answer forums have become centres of crowd-based problem solving, enabling access to domain-specific and expert knowledge, with platforms such as Stack Exchange becoming a source widely used by leading companies and academia. Furthermore, a variety of easily accessible and interactive platforms offering online courses for expanding professional or academical knowledge have been created, fulfilling the self-development needs of millions of people every day and playing a role of a freely accessible educational tool.

Simultaneously with the changes in the knowledge-sharing landscape, more efficient computing capabilities have enabled the deployment of Artificial Intelligence (AI) models trained on big data sets. These solutions can be used for creating, searching, and sharing information on a mass scale and might be another breaking point in the evolution of the knowledge-sharing process in our society.

1.1. Problem background and hypothesis formulation

Even though AI solutions have been developed since the 1950s, when Alan Turing defined the concept of machines using information for problem-solving just as humans (Turing, 1950), it was not until recently that AI solutions became ubiquitous in everyday life. The process of AI development over time is often measured by presenting milestones, such as Deep Blue computer beating Garri Kasparow in 1997 (Anyoha, 2017) or IBM Watson defeating two *Jeopardy!* champions in 2011 (IBM, 2023). Although these events gained some public attention, none of them started such a fierce debate about the future of AI as the release of the Chat GPT-3.5. On 30th November 2022, the prototype of this multi-purpose chatbot has been publicly shared, gaining users extremely quickly and attracting a lot of attention. Based on a reinforced and self-supervised Large Language Model (LLM) and being available for free for all users, Open AI's model is epitomized as the peak of the recent development in the field of general-purpose AI.

Chat GPT has attracted more than 100 million users in record time, who discovered many useful, interesting, and surprising uses for this chatbot. With capabilities such as authoring sophisticated essays, marketing copies, and children's books, creating weight-loss plans, and coding on demand, it has become a commonly used tool. The main advantage of this model, compared to its predecessors are the highly developed problem-solving skills, which enable using it in many tasks that could only be done by humans so far. Many voices in academia and business stated that these capabilities are the beginning of an important transition. '*The tipping point of the AI*' – that is what Chat GPT was called in an article published in Harvard Business Review on 14th December 2022, claiming that now AI can be implemented not only in areas where failure has got severe consequences, but also in areas where occasional failures are acceptable (Mollick, 2022). LLMs are presented as tools that automate various monotonic tasks, create new disruptive business models, and enhance the learning process on every level.

The latest achievements of LLM's creators bring advanced knowledge available on demand for everyone with access to the Internet. They seem to draw a very promising, maybe even utopic vision of the future, where one can have every sophisticated topic explained in simple terms. In this vision, Einstein's quote on understanding "*If you can't explain it to a six year old, you don't understand it yourself.*" is undoubtedly still valid, however with an addition of an AI tool that can explain '*it*' to help one understand. On the contrary, many experts in the area of computational linguistics have been signalising that the reasoning of Large Language Models is fundamentally different from human reasoning, and although the LLMs are excellent at learning statistical patterns, they are not fully aware of their answers. Moreover, the creativity of those models is limited to answers that can be created from parts of natural language in the training data, choosing the wording of an answer based on the occurrence of a given token in the training data. Thus, LLMs are often accused of 'parroting' and generating highly generic answers. Many experts in the fields related to Natural Language Processing emphasize that this drawback cannot be ever overcome by GPT and similar models, because "*meaning cannot be learned from form alone*", as stated in the results of a study by Bender & Koller (2020). These limitations lead to an important and often neglected question: could the mass use of tools such as Chat GPT really be beneficial for knowledge-sharing processes in our society?

Humans of the datafication era get information from various sources, with a growing share of people using the Internet daily to gain knowledge and develop new skills. Sometimes the information gathered from a search engine or website is not enough, especially when it comes to solving domain-specific problems. Users who want to solve such a problem or expand their knowledge often visit Community Question Answering (CQA) platforms where they ask a specific question, which, with a great probability, will get a precise answer verified by other users, usually free of charge. Those online communities have become an important part of the digital economy, playing the role of a knowledge exchange base, a place where users can collaborate, and an inspiring space encouraging individuals to try new ideas (Chen et al., 2018).

The foundation of the CQA existence is the constant flow of User Generated Content (UGC). Maintaining the contributions of the users has been a vital challenge for most of the online communities, as the time and effort of the authors do not necessarily grant them any benefits from being active in the forum. In response to that problem, online community platforms have implemented motivational schemes, such as gamification, recognition (awarding badges) or status

incentives (ranks). However, these measures and actions were taken in an environment where most of the users were not able to use GAI tools, which can affect the multiplicity, diversity, and quality of content on CQA platforms.

Chat-GPT has been labelled as a multi-purpose tool that can increase efficiency and empower users in the process of searching for information and solving problems, but is using it beneficial to the whole knowledge exchange process? Generative AI may solve sophisticated problems faster, have a wider range of capabilities and even be more accurate than humans, however, it does not document the knowledge. This is dangerous in multiple ways, with reducing the amount of well-described and explained knowledge in the first place. Also, the generative models have to be trained on the past content, and if people will stop producing new sources of knowledge, then the data needed for updating these models will not be available. It is not hard to imagine that with this course of events, the Generative Models will become outdated and not adjusted to current standards. On top of that, using GAI for tasks that previously required human input, such as writing essays, can reduce the creativity of authors, as they will be rather validating generic output than creating their own piece of work. The risk of subsidizing human-based knowledge exchange with a chatbot prompt can be one of the most significant risks related to the mass implementation of AI in the future. Despite having rather implicit than explicit character, perhaps it can be more harmful than AI-related reduction of the workforce size or unethical behaviour of AI algorithms. In order to assess whether these are just hypothetical concerns or whether the process of reducing well-documented knowledge on CQA has already begun, the overarching problem of this research has been defined as follows:

What was the impact of the introduction of Chat GPT-3.5 on content posted on Q&A platforms?

In order to systematically measure the influence of Chat GPT on CQA users' behaviour, four key hypothetical mechanisms were defined. Based on these mechanisms, a very select set of dependent variables was chosen and the hypothetical directions of changes in values of those variables were defined, resulting in the formulation of four research hypotheses.

The first mechanism of Chat GPT's influence on the CQA users is substituting the questions posted on the forums with using the prompt of the chatbot. It is assumed that the demand for answers on the platform – and therefore the propensity of the users to create a post is on average lower, as it takes a couple of hours to get an answer on a CQA in an optimistic scenario, whereas Chat GPT

provides an answer within seconds. This mechanism is also expected to decrease the interest of users in browsing the forum and lower the post scores, as users are less engaged in validating answers on the forum. The dependent variables used for measuring the effects of this mechanism are the number of posts created on the Q&A platforms, the number of views of those posts and the total scores received from the users. The hypothetical association between the presence of Chat GPT and the number of new posts, post views, and post scores is expected to be negative. Based on that, Hypothesis 1 reflecting the changes in demand for answers on CQA forums is defined as follows:

H1: The introduction of Chat GPT-3.5 significantly decreased the post count, post view count and the total score of posts on Q&A platforms.

Another important mechanism related to the use of Chat GPT by members of the online Q&A communities is how likely they are to answer new posts published on the platform. In the case of the supply of answers on the platform, there are two contradictory mechanisms influencing the number of answers. Following the reasoning regarding the change in the number of posts, the users may switch partly or completely to Chat GPT and no longer visit nor be active on the forum. On the other hand, Chat GPT provides an opportunity to produce answers to a question posted on the platform which can encourage users to answer more often. It is assumed that the first mechanism is stronger, as posting answers authored by GAI creates a risk of the user's account being removed as a result of violating the rules of the forum (it is forbidden to post content generated by GAI on some of the leading platforms). Therefore, the direction of the expected association between the number of answers and the presence of Chat GPT is negative, and it is expected that the score of the answers also decreased, because the users are less engaged in validating them, and answers created by GAI have lower quality on average. Hypothesis 2 defining the change in the supply of the answers is stated as follows:

H2: The introduction of Chat GPT-3.5 significantly decreased the number of answers and the total scores of answers on Q&A platforms.

The quality of content and its innovativeness have been the key aspects of the CQA platforms that enabled them to be a place of professional knowledge exchange. One of the dimensions in which these characteristics can be measured is the novelty of published content. It is hypothesized that the novelty of content has decreased as a result of the introduction of Chat GPT. This is believed

to be a consequence of users frequently obtaining generic responses via command prompts, which address many of their issues. Such a trend may disincentivize users from initiating discussions about novel problems on the forum. Beyond that, for users commonly utilizing the capabilities of GAI, it may be more comfortable to stick to areas in which GAI can provide guidelines, which may result in a lower likelihood of posting threads dedicated to areas that were not mentioned on the forum before. Additionally, questions posted on the forum might be based to some extent on the content generated by GAI, which tends to be generic and less novel than human-generated content. Thus, the expected association between the presence of Chat GPT and the content novelty is negative, which is reflected in Hypothesis 3, defined below:

H3: The introduction of Chat GPT-3.5 significantly decreased the novelty of content posted on Q&A platforms.

CQA platforms are often a place where users look for answers to sophisticated questions and problems that they cannot solve on their own. As Chat GPT has got the capability of solving challenging problems from various domains, it is hypothesized that the use of Chat GPT may help solve some of the less complicated questions, and that posts created by the users would be more sophisticated on average than before. Because of that, it is assumed that there might be a negative association between the presence of Chat GPT and the readability of posts. Following this reasoning, Hypothesis 4 tackling the change in the readability of posts is stated beneath:

H4: The introduction of Chat GPT-3.5 significantly decreased the readability of posts on Q&A platforms.

In Chapter 1.4, a specific research approach will be defined to designate the methods that can be used to obtain the information needed to assess the validity of the hypotheses.

1.2. Managerial relevance

The dynamically changing landscape of the knowledge exchange affects all stakeholders involved in this process, which is crucial for the functioning of the modern digital economy. This research aims to critically investigate what are the consequences of GAI popularization in areas relevant for the decision-makers in every field that benefits from online knowledge exchange. Promises for further democratization of knowledge go hand in hand with the risks that can easily outweigh the benefits of problem-solving via a chatbot prompt.

The capabilities of GAI promise a new era of rapid development. Not only can those models explain a given domain in simple terms, but they can also solve a variety of specific, even niche problems. Various use cases such as drug design, material composition, chip manufacturing and creating synthetic data are exposed to be partially realized by the use of Generative AI (Wiles, 2023). However, a substantial risk outlined in this paper is the lack of documentation of this process. Community Q&A platforms consist of User Generated Content which is a public good (Chen et al., 2018), while the output generated by GAI models is not shared by default. Moreover, the users of CQA communities constantly validate the content shared on its premises, which is limited to reinforced learning and red-teaming in the case of GAI such as Chat GPT. This urges for investigating the impact of Chat GPT introduction on the knowledge-sharing process in the whole society, which is one of the expected outcomes of this study. This paper critically assesses the effects of Chat GPT introduction to help academics, policymakers, and content creators to understand if this tool is a milestone on the timeline of knowledge democratisation, or a machine that repeats what has been once written.

The CQA platforms themselves are essential stakeholders of the knowledge exchange process. Their survival depends, to a great extent, on the user activity. One of the key results of this study is determining the changes in the number of posts and answers created by users, which can help the management of the platforms retain user engagement in the times of ubiquitous generative AI.

Creators of GAI tools based on Large Language Models are also an important party in the new, reshaped knowledge exchange process. Although leading technology companies switched their focus to LLM-based solutions, there are many unknowns related to the further development of the GAI. Reducing the open-source knowledge bases affects the training of future models. Without new and verified information, which is sourced from the CQA communities, they will not be able to provide out-to-date output.

Users of the CQA communities are also directly affected by the ongoing changes in those communities. From the perspective of readers, the falling number of people willing to help in solving sophisticated problems on demand reduces their chances of finding a good answer. The most active users also keep the older threads updated, when for example a new version of the software is released and the guidelines on the programming posts need to be adjusted. From the perspective of users solving questions, as they do not receive monetary compensation, the most important motivation is recognition in an online community, as proven in research papers

(Chen et al., 2018; Goes et al., 2016). A decrease in traffic on a platform, quality and innovativeness of posts may discourage users from visiting the sites, ending the era of keeping relevant knowledge and experiences in a public space.

1.3. Academic relevance

Q&A platforms are a relatively new field of research with a limited, but growing number of publications relevant to the area of this study. Existing literature focuses on mechanisms shaping user engagement, maintaining the creation of novel user-generated content (UGC), and the role of hierarchy and gamification in online communities. By combining the contributions from the studies on knowledge exchange with the machine learning literature tackling the recent developments in the generative models, this paper addresses ongoing behavioural changes in one of the most developed knowledge sources of modern society.

The area of this study is considered highly relevant for the development of academic literature, as assessing how the GAI tools affect user contribution will help to understand what are the implications of the public availability of the generative models and how the modern knowledge exchange changes. Despite the common positive sentiment associated with the release of the Chat GPT, this research aims to assess what risks are connected with its presence, namely a reduction of user engagement, and therefore the quantity and quality of posts on domain-specific forums. It is considered an extremely important and relevant area for the research, as those Q&A platforms have become a part of knowledge sources also in the academic community, with an example of more than 90 scientific articles having references to the Math Overflow forum dedicated to solving novel questions in mathematics (Y. R. Tausczik, 2016). It poses a considerable challenge to academia when users choose to consult a chatbot - that often delivers confident responses, irrespective of the accuracy of the information on a given topic - over forums that are verified by thousands of users daily. In response to that shift, this paper aims to examine what changes have already occurred in user activity on Q&A forums.

Beyond the contributions related to the research domain, this paper takes an innovative approach to analysing the traffic on Q&A forums, that complements the existing literature in this field. In comparison to the study conducted by Goes et al. (2016) which investigates the impact of hierarchical incentives and gamification on user effort, this study focuses not only on the number of answers, but also on the number of questions posted. What is more, the set of dependent

variables chosen to assess the change in activity on Stack Exchange platforms also differs from the one used by Chen et al. (2018) in a study using panel estimation to determine the effectiveness of various types of incentives on user engagement, based on the motivational state of the users. In comparison to that research model, methods used in this paper also include Natural Language Processing techniques implemented for post content analysis. In relation to textual analysis, this dissertation analyses not only the relative content novelty of posts, as measured in a recent study on the effectiveness of peer awards and recognition of new users authored by Burtch et al. (2022), but also the post readability, represented by a normalized average of different measures denoting ease of reading.

In conclusion, this study significantly extends the academic literature by capturing a relevant research gap at the intersection of online communities and artificial intelligence, and by applying an innovative approach focused on observing changes in key dimensions of content posted on Q&A platforms. By building upon knowledge from various streams of literature—such as machine learning (with a focus on natural language processing), information management, social computing, and data ethics—this study provides insights about changes in the knowledge-sharing process. It also addresses the existing gap in understanding the impact of Generative AI solutions on Q&A platforms.

1.4. Research approach and structure

This study has been conducted based on observational data collected from the 36 most popular Stack Exchange forums that are dedicated to various domains. To assess the impact of Chat GPT introduction on those online communities, a quasi-experimental method of difference-in-differences estimation was chosen. As it was not possible to determine a control group consisting of individuals who have been not affected by the availability of the GAI tools and have been using one of the leading Q&A platforms, this study uses a relatively new approach similar to one implemented by Eichenbaum et al. (2020) to assess the economic activity during the pandemic waves of SARS-CoV-2. The introduction date of Chat GPT-3.5 (30th November 2022) is treated as an intervention date separating the pre-and post-intervention periods of equal length (21 weeks). Observational data on different dimensions of user activity on CQA forums from these periods represent the treatment group and similar data from respective periods in the preceding year represent the control group. Taking such an approach enables assessing the causal effect of the

rapid popularisation of the GAI tools without limiting the scope of the research to an experimental setting. In the analysis conducted in this study, a large data set of observational data can be used, which enables a high-level and accurate assessment of changes in the quantity and quality of content posted on Q&A forums.

The remaining part of this paper consists of six chapters. In the second chapter, the background related to Generative Artificial Intelligence and Q&A platforms is presented. The third chapter consists of a synthesized review of the literature on GAI and Q&A platforms. In the fourth chapter, the important characteristics of Stack Exchange forums are presented and the descriptive statistics of the data as well as insights from exploratory data analysis are provided. The fifth chapter contains the methodology chosen for this study and focuses on the difference-in-differences estimator and its assumptions, as well as on Natural Language Processing methods related to the content novelty and readability of posts. In the sixth chapter, the results of the estimations are presented and discussed. The final, seventh chapter is a critical discussion of the results obtained and their implications for business and academia. Furthermore, a description of the limitations of the study and recommendations for future research is provided in this chapter.

2. Background information

This chapter describes the history, development and roles played by Generative Artificial Intelligence and Q&A Platforms, which states the essential context for this study. On top of that, the influence of Generative Artificial Intelligence on knowledge sharing is discussed.

2.1. Generative Artificial Intelligence

2.1.1. Definition of the Generative Artificial Intelligence

According to the comprehensive survey of AI Generated Content by Cao et al. (2023), *'Generative AI (GAI) techniques belong to the category of Artificial Intelligence Generated Content (AIGC), which involves the creation of digital content, such as images, music, and natural language, through AI models*'. This definition illustrates the fundamental difference between GAI, and AI methods most commonly applied for regression and classification problems throughout the last decades, which are commonly defined as conventional AI or analytical AI. While the conventional methods have a discriminative approach that limits their role only to analysing existing or simulated data, GAI can create new content (Zhang et al., 2023). Van der Zant takes a broader and

more philosophical approach, pointing out the difference in the way those two groups of methods function. Compared to the conventional AI methods that are optimized to some end-point and whose outcome is predictable, the functioning of GAI is strongly dependent on internal dynamics and the interactions with its environment (Van Der Zant et al., 2013). One such example can be the bias-variance trade-off, which plays an essential role in optimizing analytical AI methods, but it is not that important for the Generative models, such as Large Language Models.

The difference in nature of the analytical AI and GAI does not mean that the first group does not generate content, which takes place for example when labels are generated for images recognized by a supervised algorithm. Although the content is generated, it is low dimensional, while GAI provides a high-dimensional output that can be '*used as synthetic data for alleviating the need for more data in deep learning*' (Zhang et al., 2023).

2.1.2. History and development of Generative Artificial Intelligence

The history of generative models started in the early days of Artificial Intelligence when Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) emerged in the 1950s. Those algorithms were used for creating sequential data, with examples such as speech or time-series data. From different types of the early generative models, the closest predecessors of today's GAI such as Chat GPT were Natural Language Processing (NLP) models, which were trained to search for the most similar word sequences from the word distribution of the training data. That process called n-gram language modelling was highly ineffective in creating long sentences, (Cao et al., 2023, pp. 4–5) because those models suffered from the 'curse of dimensionality', as the training data have got way fewer instances than the possible test data. Solutions for those problems came in the 2000s when Recurrent Neural Networks (RNNs) were introduced. Implementing RNNs along with changes such as training on the distributed representation of words, that have taken into account vast numbers of sentences similar to a chosen one, resulted in significant improvements compared to the n-gram models, as presented in the study by Bengio et al. (2003). The usage of Long-Short Term Memory (LSTM) and Gated Recurred Unit (GRU) that enhanced memory control during model training were additional advancements that happened at that time (Cao et al., 2023, p. 4).

A milestone in the development of the GAI was the implementation of the deep learning techniques used for the model training and introduction of the transformer architecture as presented in

(Vaswani et al., 2017), which has become the backbone of the Large Language Models, such as Chat GPT or BERT, replacing the RNNs (Zhang et al., 2023, p. 9) for the language tasks. Deep learning, as defined in the aforementioned paper is a '*data-driven method that optimizes the model parameters with a stochastic gradient*' and has been used so widely, that it started to be commonly associated with the term Artificial Intelligence in general (Zhang et al., 2023, p. 9). After the popularization of deep learning and transformer architecture, the development of GAI has gained enormous momentum, switching the aim of AI from analysing data to creating content.

Chat GPT-3.5 developed by Open AI which was released on 30th November 2022 has utilized most of the developments from recent years and proven that as a Large Language Model, it can fulfil various tasks, such as successfully taking official exams, summarizing text, translating content, generating articles and developing code (Hughes, 2023). It became the consumer application that acquired a 100 million user base in a record time of around two months after launch. (Hu, 2023)

The latest improvement in the development of the GAI is the introduction of Chat GPT-4, which took place on the 14th of March 2023. The new version is capable of processing not only text but also images and has got significantly '*broader general knowledge and problem-solving abilities'* according to its creators from Open AI (*Open AI*, 2023). These claims have been confirmed in the first preliminary studies, with a paper by Katz et al. (2023) serving as an example of the difference between GPT-4 and its predecessor. GPT-4 passed the Uniformed Bar Exam (UBE), an exam that has been viewed as an '*insurmountable summit for even domain-specific models.*' (Katz et al., 2023, p. 10) with a significant margin. It achieved scores better than the average student scores, in opposition to GPT-3.5 which has achieved a score lower by 26%, placing it below the average level of current human takers.

Currently, GAI tools are often named one of the most disruptive technologies shaping the future of business. According to McKinsey's article from 20th December 2022, the reasons why Generative AI will be a disrupting technology are its abilities to create next-generation assistive technology, enable access to new capabilities for non-technical users as well as reduce the application development time (Chui et al., 2022). In the article, an extensive list of business use cases in many areas such as marketing, operations, risk, law, engineering, and R&D has been presented with many Generative AI use cases that '*could create early impact*'. Therefore, knowledge about the impact of technologies such as Chat GPT can be useful for many business stakeholders, especially in fields related to IT and business information management.

What is more, the potential for Generative AI goes far more than optimizing repetitive business tasks, enhancing communication and code development. In the article published by Gartner on 26th January 2023, various use cases such as drug design, material composition, chip manufacturing and creating synthetic data are exposed to be partially realized by the use of Generative AI (Wiles, 2023).

The introduction of Chat GPT has also started a great debate in the education sector and academia since its release at the end of November 2022. These sectors have also recognized the relevance of this new tool and Chat GPT has quickly become commonly used in the scientific community. One of the pieces of evidence of its popularity is a recent survey conducted by the Nature Journal. According to the results published on the journal's website on the 20th of February 2023, " *Of 672 readers (*of the journal – added by author*) who responded to an online questionnaire, around 80% have used Chat GPT or a similar AI tool at least once* " (Owens, 2023), which indicates that this tool has become commonly used in a scientific community. Moreover, recent developments in using Chat GPT in education systems have been observed, as in February 2023 the Singaporean government officially announced that Generative AI tools will be used " (…) *appropriately, these tools (*Generative AI – added by author*) can support students in their learning when students have mastered basic concepts and thinking skills*" (Ministry of Education, Singapore, 2023).

2.2. Role of the Q&A Platforms in knowledge sharing

Between the years 2005 and 2022, the number of individuals using the Internet has increased from 16% to 66% according to the latest ITU report (International Telecommunication Union, 2022). With more than 5.3 billion users Internet has become a ubiquitous part of almost all domains of everyday life of many people, including knowledge exchange and education. One of the key components of the modern knowledge exchange process are the Q&A platforms, which evolved from domain-specific mailing lists (Vasilescu et al., 2014), enabling one to ask various kinds of questions in the community for free or for a fee.

A study quite distant from today's perspective by (Harper et al., 2008) recognized three main types of Q&A forums that have emerged: '*Digital reference services*', '*Ask expert services*' and '*Community Q&A sites*'. This study will focus on Community Q&A sites, which are the most modern and popular type of Q&A forums. The following type of Q&A forum is characterized by a role-based structure with the community asking and answering the questions and the moderators,

who often have been the most active users for a longer time, setting the rules and providing features of an online platform (Harper et al., 2008, p. 3).

Over the last two decades, the Community Q&A forums have gained a lot of popularity, which is reflected in the website traffic rankings such as *semrush.com*, according to which the highest ranked Q&A forums were *reddit.com* with approximately 5.41 billion visits in March 2023 (9th position among all websites) (*Semrush.Com*, 2023a), and *quora.com* with approximately 1.43 billion visits in March 2023 (54th position among all websites) (*Semrush.Com*, 2023b).

For many individuals, those forums have become an important source of information, especially in a technical domain such as programming (Geigle et al., 2019). Online communities have been functioning following the validation scheme based on a number of opinions or scores left by different users, which can be defined as the 'wisdom of the crowd'(Y. Tausczik & Boons, 2018).

In order to analyse how CQA platforms are changing, it is important to acknowledge that Chat GPT indeed caused disruption, but still, it has got many shortcomings and it is not a perfect tool, which is reflected in the public discourse. Chat GPT is not yet able to answer many questions and has got a tendency to provide fraudulent answers and to refer to sources that do not exist. Moreover, recent studies conducted on the previous version of GPT-3 indicated that this model requires additional safeguards to ensure that radical and insulting answers do not appear (McGuffie & Newhouse, 2020). With tens of thousands of threads created each month on forums such as Stack Overflow, assessing the impact of Chat GPT's introduction on CQA is considered relevant for the future of knowledge exchange in many domains, which leads to the core problem of this research paper.

No studies on the impact of Generative AI on Q&A online communities have been found, which is related to the novelty of the topic. The landscape of knowledge sharing is changing dynamically, with emerging technologies such as GAI enhancing the learning process and automating repetitive tasks, but simultaneously increasing the risk of spreading false or wrongful information, which can also affect the Q&A online communities.

3. Related literature

This chapter provides a synthesis of the existing literature on Generative Artificial Intelligence and Q&A Platforms. The literature review aims to define the effects and mechanisms influencing the process of knowledge sharing that were previously observed in the literature and are relevant to

this research. Moreover, the areas in which this study contributes to various streams of scientific literature are outlined.

3.1. Literature on Generative Artificial Intelligence

Many recent events illustrate that Generative AI tools are changing processes of sharing knowledge at a very quick pace. The introduction of Chat GPT-3.5 to the general public enabled every individual with access to the Internet to use this tool for creating new content or reworking existing pieces of information. Given the exponential growth of the number of Chat GPT users, this phenomenon has hypothetically reshaped existing knowledge exchange processes.

The debate about Generative AI often revolves around the consequences of sharing AI-generated content in the future, however, numerous GAI use cases have already been implemented. One of the most substantial examples is news generation by leading newspapers such as the *Associated Press*, *Forbes*, the *New York Times*, the *Washington Post*, and *ProPublica*. In those companies articles related to business, politics, sports and foreign affairs are generated automatically, which can raise concerns about the accuracy of this content, especially since currently there are no strict requirements on reporting the involvement of AI in content generation (Longoni et al., 2022, pp. 1–2). An experimental study conducted by Longoni et al. (2022, pp.1) on representative U.S. samples shows that news headlines generated by AI were "(the respondents were – added by author) *more likely to incorrectly rate news headlines written by AI (vs. a human) as inaccurate when they were actually true, and more likely to correctly rate them as inaccurate when they were indeed false*.". This conclusion points out that the process of sharing relevant facts may change as the involvement of GAI in content creation influences its accuracy in the eyes of readers. The need for regulation of the AI sector and disclosure of AI usage are means often proposed in scientific literature, as there are proven examples of harmful and unethical content generated by AI, such as societal biases – for example associating Muslims with violence by GPT-3 (Abid et al., 2021).

The use of Generative AI may enhance learning on various levels and take part in democratizing knowledge. Due to the availability on demand and high response speed of tools such as Chat GPT, interactive and adaptive learning is possible for a larger number of people than ever before. Teachers can also benefit from the introduction of those tools by automating grading and enhanced tutoring (Baidoo-Anu & Owusu Ansah, 2023, pp. 1–8). The study conducted by Yue et al. (2023)

shows that Chat GPT-3.5 can explain financial terms to individuals with little or no financial knowledge, which has got potential to increase access to responsible investments for these individuals. The use of GAI tools for robo-advice also decreases costs compared to traditional investment advice. Besides that, GAI tools are also reshaping research, and enhancing the development of scientific papers, with proofreading and summarizing being the key functions. GPT-3 has been proven to help researchers save significant time in editing and revising research manuscripts while maintaining sufficient quality (Pividori & Greene, 2023). All things considered, GAI tools are found to be highly disruptive to the existing knowledge-sharing processes, however, they have significant drawbacks that require critical assessment by humans. Besides the risks of societal bias mentioned in Chapter 2, they tend to generate 'hallucinations' – wrongfully interpreting facts and creating nonexistent references, which can be particularly harmful to research applications (Baidoo-Anu & Owusu Ansah, 2023).

Although the impact of GAI on knowledge sharing in academia and teaching was a topic of interest for researchers in recent months, Q&A forums were not found to be mentioned in recent publications. Previous research related to how chatbots influence the Q&A platforms has been conducted before the introduction of Chat GPT-3.5 on November 30th, 2022 and there was no option to assess the impact of Generative AI tools on Q&A platforms, as none of these tools were so widely used at that time. Due to this limitation, the studies in this area conducted before this date had experimental character and mostly focused on one aspect of the platform. A relevant (in the context of this research) example of such a study is a paper published by the University of Antwerp and the University of California, analysing the interactions between users and bots on Stack Overflow (Murgia et al., 2016). It provides valuable insights about the user interactions with the content generated by bots, such as lower ratings for machine-written content, but the impact of Generative AI on the traffic on this website is not analysed. Therefore, analysing the impact of the popularisation of Generative AI tools on Q&A forums states a gap, which can be fulfilled by this study.

3.2. Literature on online Q&A platforms

Online Q&A platforms have been an object of interest for many researchers, focusing mainly on the factors influencing user behaviour, such as various forms of incentives (status in a hierarchy, user score, peer awards) and how they influence the quantity and quality of the user-generated

content (UGC). A variety of studies on maintaining high levels of user contribution is dictated by the fact that this factor is crucial for the Q&A platforms to function and develop, as currently most of them do not offer financial incentives for individuals responding to questions (Guan et al., 2018). Therefore, motivating users to engage in discussions is also the main challenge for the administration of Q&A forums, which makes the implications of research on this topic even more important.

One of the relevant studies aiming to determine the effectiveness of methods increasing user engagement is a study conducted by Chen et al. (2018) which demonstrates that the effect of different incentive categories differs depending on the state of user motivation. Peer recognition is found to be the most effective measure, increasing user engagement in all states of motivation determined in the study. The work of Goes et al. (2016) questions the effectiveness of hierarchy-related incentives and gamification in inducing contributions of the users. It shows that reaching a higher rank in the hierarchy of an online community indeed increases user contribution, however, this effect is temporary and increases are smaller for higher ranks awarded to the users.

A recent study by Burtch et al. (2022) analyses the effect of peer awards on the creativity of users, measured by semantic content novelty (in relation to the content posted on a given subforum) of the UGC on Reddit, with a particular focus on contributions of new users. This experimental study shows that assigning a peer award enhances the creativity of the new users, resulting in the creation of a bigger amount of novel UGC.

Another stream of literature that has been developed over the last years concentrates on the value generated by users of the Q&A forums, and assessing which users create significantly higher value than others. A study by Anderson et al. (2012) concludes that including different kinds of community features (such as measures of reputation, activity, and quality) increases the accuracy of answers on the platform. Potential experts that provide answers more frequently and more accurately are the object of research conducted by Pal et al. (2011). The results of detecting such users by using an algorithm based on user behaviour shows that early identification of those users is crucial for the Q&A platforms to retain engaged users and maintain high accuracy of answers.

Some studies also investigate the accuracy of content verification on community Q&A forums. These platforms operate in a scheme defined as '*crowd problem-solving*' (Y.Tausczik & Boons, 2018) that despite many advantages such as the ability to broadcast difficult

problems to crowds and enhance the learning process of the online discussion participants (Baran & Keleş, 2011, p. 58) has got many significant drawbacks. One of the serious flaws is investigated in the study conducted by Tausczik & Boons (2018), which has proven that a crowd with relevant information distributed among the individuals fails to use diverse and specific knowledge of the participants, as the majority tends to stick to the most common facts and neglect the importance of information held by only some of the participants. This study also illustrates another failure of such communities, as the way in which participants share facts is biased and the decision process is unstable (Y. Tausczik & Boons, 2018, pp. 1–4). A different observation regarding the subjectivity of the Q&A forums is made in the study conducted by (Liu et al., 2013) – individuals visiting more elaborate Q&A forums are more likely to engage in a discussion and to learn more, than individuals who only present basic knowledge.

This study contributes to the literature on online Q&A communities, analysing the presence of a new-generation GAI tool on different dimensions of the content posted. Conclusions presented in the last chapter will illustrate how the user contribution on Q&A forums change, which will state a foundation for future research on mechanisms motivating users to use GAI for increasing their contributions on online platforms.

4. Data

This chapter is dedicated to the introduction of the Stack Exchange forums – the object of this research, and to outlining relevant characteristics of the data gathered on those forums. The latter consists of a description of the data collection and pre-processing as well as a presentation of the insights from the exploratory data analysis.

4.1. Stack Exchange forums

The existence of the Stack Exchange forums began in 2008, when Stack Overflow was founded to solve programming challenges by programmers willing to contribute on a voluntary basis. In 2009 other forums were added, creating the Stack Exchange community, which has become one of the most well-known and most visited CQA on the Internet. Despite how much the landscape of online Q&A forums has changed over the years, Stack Exchange forums are still a vibrant online community. Website traffic statistics from *semrush.com* place *stackoverflow.com* (with more than

24 million threads available) domain in the 170th place (along all domains) with around 512.9 million visits in March 2023 (*Semrush.Com*, 2023c).

Stack Exchange forums are characterized by a structure similar to a domain-specific knowledge base, where users are focused on answering questions accurately and methodologically. The concise and professional character of the forums is presented to the new users just after they finish the registration process: *'This site is all about getting answers. It's not a discussion forum. There's no chit-chat. Just questions... ...and answers.'*. The interface of the forum enhances clarity and transparency – correct answers are voted up and are displayed at the top of the page, to be shown first. The first answer that works for the author of a question is marked as 'accepted', indicating that the thread has been resolved, however, users can still try to provide a better answer or comment on the existing ones to suggest how to optimize them.

The process of choosing the correct answer is the core of the Stack Exchange gamification process. It is worth noting that the forum prioritizes getting a working solution first (and marking it as accepted), even if it is not the best answer. For correct answers users gather points that result in additional privileges such as the ability to comment, give up- or down-votes to other threads and in the case of highly engaged users – to become a moderator. This incentivises users to answer quickly to new questions as well as try to find a better solution for questions that were already answered.

The professional and result-oriented character of forums made them become a source of technical and domain-specific knowledge for many individuals, with Stack Overflow being referred to as a '*computer science department where people go to learn*' in a study conducted by (Dondio & Shaheen, 2019, p. 1). Although Stack Overflow is the most developed and recognizable forum by a great margin, Stack Exchange communities cover a wide range of topics, with more than 170 community-powered Q&A sites. The community allows submitting ideas for creating new forums at *area51.meta.stackexchange.com*. If a proposed forum gets a sufficient amount of recognition tokens, it should become a beta version of a new Stack Exchange forum.

Given the knowledge exchange taking place on the forums, a significant number of studies on the Stack Exchange forums have been conducted (mostly on Stack Overflow), with the most notable themes related to enhancing learning abilities (Dondio & Shaheen, 2019), evaluating the expertise of the Stack Exchange community members (Huang et al., 2018), analysing comments and

community dynamics (Sengupta & Haythornthwaite, 2020), as well as assessing the quality of questions and answers posted on the forums (Hodgins, 2016).

4.2. Data collection

Stack Exchange is an open-source organisation that shares all its data, which is available online via Stack Exchange Data Explorer based on Microsoft SQL Server (Stack Exchange, 2023). Data for this research was collected by executing SQL queries on this database for the 36 most popular Stack Exchange forums (names and the number of posts are presented in Appendix B).

Given that there are around 180 Stack Exchange forums in total, this study includes the upper 20% of all forums that account for the great majority of all questions asked on Stack Exchange. The forums included in the study differ significantly in terms of the total number of posts and answers, from Stack Overflow, the most popular forum with around 24 million questions and 35 million answers to Ethereum with roughly 51 thousand questions and 57 thousand answers. The motivation behind including only the most active forums was to capture variability while aggregating on the weekly level with a reasonable margin of error. Forums with around 50,000 data points (defined as the number of posts) yield around 100 data points per week on average (assuming that those forums exist for 10 years), which gives a sufficient sample size for a panel regression.

For all forums combined, more than 3 million data points representing posts from the periods from 01/07/2021 to 30/04/2022 and from 01/07/2022 – 30/04/2023 were collected. The choice of this time period was dictated by the main method of the study – difference-in-differences regression. Chat GPT-3.5 was released on the 30th of November 2023, and this date is treated as an intervention date in this study. Therefore, data were collected for five months preceding the release (July – November) and for five months following the release (December – April). At the time of the last data update (beginning of May 2023) that was the configuration that could capture the longest post-intervention period while maintaining the same length of pre-intervention and post-intervention periods. Data for the treatment group were collected for the aforementioned period of years 2022-2023. Data from the respective period from a year before, so from years 2021-2022 was collected to represent the control group.

The data in both years starts in week 26 and ends in week 17 next year, however, the scope of the analysis was limited to the period from week 27 to week 16 next year, to include only weeks

containing observations for all seven days of the week. With that configuration, the pre- and post-intervention periods include 21 complete weeks each. All values presented in the next chapters refer to those periods of time.

Data collected by executing queries had 11 different columns, consisting of 9 columns containing values of numerical dependent variables (post score, post view count, answer count, answer scores), values used for aggregation (post creation date, post tags), text data (post body, post title, answer body) and additional information about the users (post author, answer author). It is worth noting that both posts with and without a valid answer were collected, to ensure that the analysis does not omit unanswered questions, whose number and share might have increased after the introduction of Chat GPT.

4.3. Data transformations

4.3.1. Data quality checks and text pre-processing

The data sets downloaded from the Stack Exchange Data Explorer were provided in a structured form, but various data transformations were required to obtain a data set enabling effective implementation of the research approach. Firstly, the data downloaded in a csv format was merged into one file, creating a data set in a long format, suitable for the panel estimations. After that, a set of standard procedures ensuring different dimensions of data quality following the approach presented by Watson (2004) was applied. The accuracy of the data was checked by exploring the values of the numerical variables, completeness by checking the missing values count by column, and timeliness by ensuring the correct merging of the files and the values of different variables over time. Moreover, besides the high-level analysis and exploration, random checks of the database were conducted to validate the consistency of the data. No significant flaws or inconsistencies in the data were found in this process, resulting in keeping all observations.

The next step applied was the pre-processing of textual data, namely the text body of posts. Text data have been prepared for vectorisation needed for the analysis by removing HTML tags, snippets of code, hyperlinks and website addresses. The second part of the textual pre-processing was automated language detection and excluding all posts in other languages than English, which excluded around 4.51% of all observations for the content novelty and readability analysis. In this form, the text of the posts was saved for the readability analysis, as those tools use original natural

language to calculate the readability scores. For the content novelty analysis, the textual data were processed further - a list of stop words was removed from the text. It is a standard procedure to remove stop words (such as “and”, ”the” etc.) as they do not contain much value and they appear frequently in the texts (Balakrishnan & Ethel, 2014). A standard list of stop words available in the NLTK Python library was used for this operation (*NLTK Documentation*, 2023). The final step in preparation for the vectorisation of the post text was applying the lemmatisation of the words. Lemmatisation is a process of reducing the words to the lemma, defined also as a vocabulary form. The inflected parts of a word are combined by an algorithm to create a single lemma. For example, the lemma of the words “studied” “studying” and “studies” is “study”. Contrary to stemming, which is another technique commonly used to reduce the number of words, lemmatising takes into account other information such as the person, singular and plural forms as well as the gerund and infinitive forms of the verb (Khyani & B S, 2021). After applying this transformation, the body text of the Stack Exchanges posts, saved in a lemmatized form, was ready for calculating the content novelty measures.

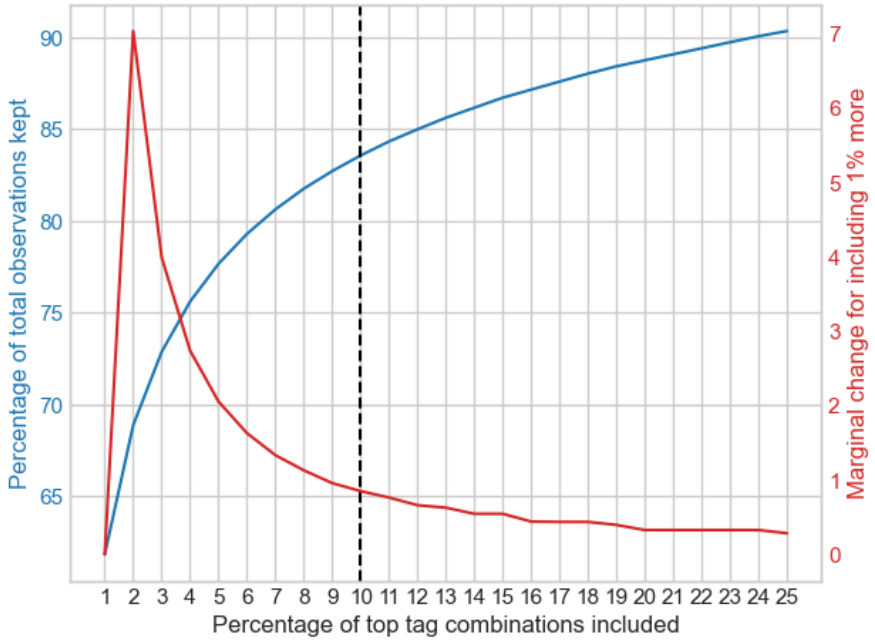
4.3.2. Feature engineering and data aggregation

To investigate the changes in values of the dependent variables defined in Chapter 1, a select set of variables had to be created using the downloaded data. The original (not processed) text was used to calculate the dependent variables representing the post body length and title length. Using the text that was posted in English cleaned from code chunks, html tags and other objects not being words, six readability metrics were calculated. A set of readability metrics consists of Automated Readability Index, Flesch Reading Ease, Flesch-Kincaid Grade, Gunning Fog Index, SMOG Index and Coleman-Liau Index. Values of those features were normalised accordingly to their scale (which will be presented in detail in Chapter 5) and the average readability metric was created. Variables indicating the week of the year and identifying the number of weeks were also added to the data set. The meaning of all variables with their level of aggregation is presented in Appendix A.

The period (T) and treatment group (D) indicators were added to the data set to represent the pre-intervention and post-intervention periods, and assignment to control and treatment groups for difference-in-differences estimations. Tags attached to the posts were separated from a string form, and saved in separate columns – every post can have up to three main tags. Combinations of the

first two tags (coded without repetition – a combination of tag x and tag y has got the same meaning as a combination of tag x and tag y) were saved in a separate column. Due to limitations regarding computing power, the decision about reducing the number of tag combinations (which are treated as entities in the panel regression conducted in this research) was made. To determine what proportion of the tag combinations should be left in the data set, an exploratory analysis of relative and marginal change in a number of observations for different proportions was conducted. The results are presented in Figure 1 below:

Figure 1: Percentage of observations kept in the data set for different percentage values of the most common tag combinations included



Based on the analysis, the optimal cut-off value would limit the analysis to only 2% of the tag combinations that appear the most often, as they are present in almost 70% of all posts in the data set. Given that removing 30% of all observations would be still a significant loss of information, the scope of research was limited to the 10% of most common tag combinations, which are present in 84% of all posts in the data set. Counts of unique tags and numbers of observations before and after the reduction of the data set are presented in Table 1 below:

Table 1: Counts of unique values of tags, tag combinations, forums present and a number of observations.

	Before reduction	After reduction
Number of unique tag1 values	24,140	4,348
Number of unique tag2 values	44,621	9,047
Number of unique tag2 values	53,293	42,570
Number of unique tag1-tag2 combinations	342,992	34,299
Number Forums present	36	36
Number of observations	3,107,968	2,593,153
No. of observations per tag combination	9.06	75.60

The reduction removed 514,815 observations which constituted 16.56% of all observations in the full data set. After the reduction the data were aggregated on a weekly basis per each tag combination, resulting in the creation of a data set with 781,821 rows.

4.4. Exploratory data analysis

The descriptive statistics of the pre-processed, reduced, and aggregated data set are presented in Table 2 and Table 3 below:

Table 2: Descriptive statistics of the dependent variables (1/2)

	Post_Count	Post_Score	Post_Answer_Count	Response_Scores
mean	3.32	1.74	3.19	3.54
std	15.79	8.72	18.30	20.57
min	1.00	-647.00	0.00	-14.00
25%	1.00	0.00	1.00	0.00
50%	1.00	0.00	1.00	1.00
75%	2.00	2.00	2.00	2.00
max	790.00	1,537.00	1,010.00	4,108.00

Note: N = 781,821

Table 3: Descriptive statistics of the dependent variables (2/2), title- and body length

	Post_View_Count	Title_Length	Body_Length	Avg_Readability
mean	1,164.42	61.24	1,667.52	0.062956
std	8,226.38	21.43	1,932.68	0.018554
min	2.00	15.00	38.00	0.002593
25%	68.00	47.00	699.67	0.055185
50%	185.00	58.50	1,187.00	0.065243
75%	590.00	72.00	1,958.00	0.074083
max	1,690,099.00	150.00	108,971.00	0.728542

Note: N = 781,821

Descriptive statistics show that values of all dependent count variables have a positively skewed distribution, with a mean greater than the median. It can be noted that these variables have a very broad range of values, with the most popular tag combinations occurring a couple hundred or thousand times more often than the mean value. Average and median values of title and body length are also positively skewed, however for the title length skewness is smaller, presumably due to the 150-character limit. The values of average- and median post body length suggest that Stack Exchange posts are mostly short. Assuming that an average word in English contains 5 characters (Miller et al., 1958, p. 382), the average length of posts was roughly 335 words, and the median length of posts was roughly 240 words.

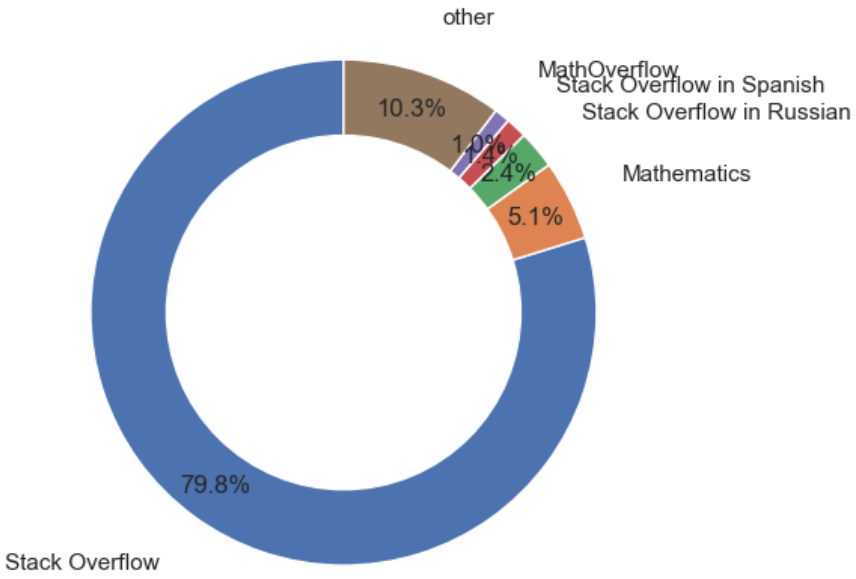
Descriptive statistics of the average readability suggest a negatively skewed distribution, however, it should be noted that the higher the value of this metric, the lower the readability of the text. This interpretation is based on the way of calculating the metrics that mostly reflect an equivalent of a US educational grade needed to understand the text. Those metrics and their interpretation will be discussed in Chapter 5.2.2 in detail. Moreover, the average readability metric is a normalized average so it cannot be interpreted directly, therefore it is better to refer to the descriptive statistics of the different readability metrics presented in Appendix C. As the readability metrics could only be calculated for posts written in English, the values of average readability (calculated on the aggregated level) also represent only the posts written in English.

The average Automated Readability Index (ARI) score was found to be 10.14, which suggests that understanding the post content (in terms of language) requires an educational level similar to the 10^{th} US grade level. The Flesch-Kincaid Grade, Gunning Fog Index, SMOG Index and Coleman Liau Index use similar scale and have lower mean values, equal to 7.8, 9.63, 8.19 and 7.8 US grade equivalents respectively. Flesch Reading Ease (FRE) uses a different scale, where the higher

values denote more readable texts. A mean score of 70.69 is higher than the commonly assumed target value of 60, which is the level at which the text is accessible for most the readers. In conclusion, the values of readability metrics show that Stack Exchange posts are easily accessible and readable, with content that requires language knowledge on the level of 8^{th} to 10^{th} US grade on average. According to all metrics, 75% of posts require knowledge equivalent to 12^{th} grade at most.

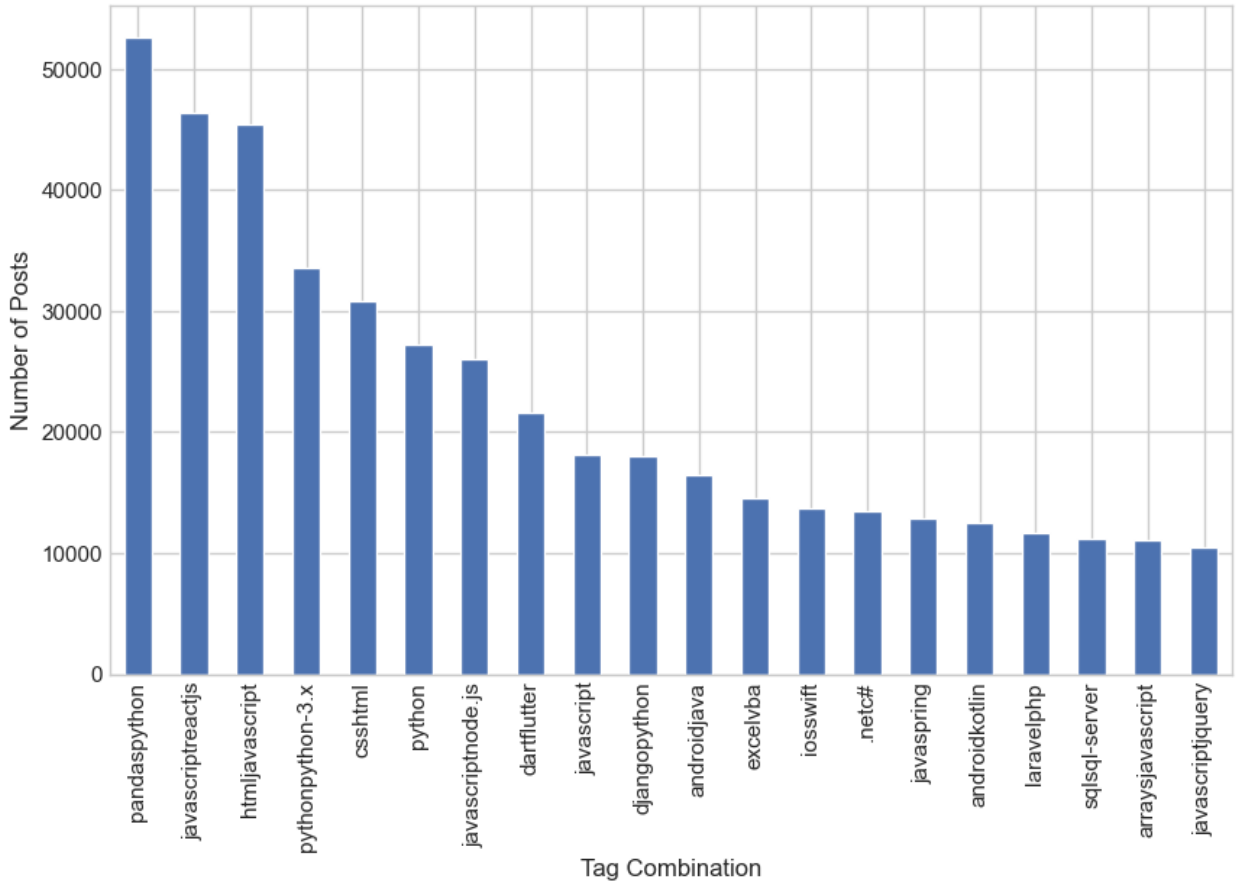
Analysed Stack Exchange forums differ significantly not only in qualitative characteristics such as the domain of content discussed by their users but also in the quantity of content published and level of user activity. Figure 2 below presents the overview of the value share of the 5 forums having the largest post number in the data set:

Figure 2: Total post count share of the five most popular forums in the reduced data set



The oldest and largest forum - Stack Overflow constitutes almost 80% of all observations in the data set. Such a great share of this one forum in the set of 36 most popular forums was the motivation to conduct the analysis on the tag combination level, as much heterogeneity between the forums is observed. All 5 most popular forums that account for nearly 90% of all observations are related to either programming or mathematics, which shows the dominant characteristics of content posted on the most popular Stack Exchange forums. It is also reflected in the most 20 popular tag combinations presented in Figure 3:

Figure 3: 20 most popular tag combinations in the non-aggregated, reduced data set



All 20 most popular tag combinations are related to programming, with numbers reaching 50,000 posts out of 2.6 million observations (values regard to the not aggregated data set). It clearly reflects the popularity of Stack Overflow and the high activity of this forum's community.

A breakdown of post counts of all forums can be found in the table placed in Appendix B. Appendix D contains the descriptive statistics of the dependent variables' body and post length in the reduced, not aggregated data set. In Appendix E the descriptive statistics of the readability measures in the reduced, not aggregated data set can be found.

5. Methodology

The following part contains information about the methodology applied in this study. The relevant characteristics of difference-in-differences estimation are outlined in the first section, with emphasis on the parallel trends assumption. This part also contains the specification of the

regression model. In the second part, the Natural Language Processing (NLP) techniques chosen for calculating content novelty and evaluating the readability of posts are described.

5.1. Difference-in-differences estimation

Difference-in-differences has been one of the most popular quasi-experimental methods used to assess the causal effect of interventions (Callaway & Sant'Anna, 2021), with a history dating back to Snow's analysis of the cholera epidemic in London (Snow, 1855). Using the DiD estimator requires specifying the control group (not receiving treatment) and the treatment group (receiving treatment), as well as (at least) two time periods: pre-treatment (or pre-intervention; a period before the treatment group receives the treatment) and post-treatment (or post intervention; after the treatment group receives the treatment) (Albouy, 2020). The outcome of the difference-in-differences model in a canonical form can be specified as follows (Greene, 2018, p. 168):

$$y_{it} = \beta_1 + \beta_2 T_t + \beta_3 D_i + \delta(T_t \cdot D_i) + \varepsilon_i \quad (1)$$

$$T \in \{0,1\}, D \in \{0,1\}$$

In Equation 1 presented above, variable T indicates the period (0 – pre-intervention, 1 – post-intervention), and variable D is a binary indicator of assignment to the treatment group (with 0 indicating assignment to the control group, and 1 - to the treatment group). Thus, y_{it} is the outcome variable for section i in period t. The difference-in-differences effect is equal to the difference between the change of the outcome variable y for the treatment group and the control group, as presented in Equation 2 below:

$$(\beta_1 - \beta_2 - \beta_3 + \delta) - (\beta_1 + \beta_3) - ((\beta_1 - \beta_2) - \beta_1) = \beta_2 + \delta - \beta_2 = \delta \quad (2)$$

The value of difference-in-differences is equal to the coefficient of the interaction term between the period indicator T and the treatment group indicator D from the canonical Equation 1.

Given the nature of the problem and the available data on Stack Exchange Q&A platforms, finding a control group that has not been affected by the introduction of Chat GPT was not feasible for this research. Therefore, it was decided that a relatively new approach to DiD estimation will be implemented, which assumes that data from the previous year can be treated as a control group. This approach was used by Eichenbaum et al. (2020) when comparing consumer spending between different waves of the SARS-CoV-2 epidemic in Portugal.

Another key assumption for the implementation of DiD estimation in this paper is treating the date of Chat GPT-3.5 introduction (30th of November 2022) as an intervention date. This decision is motivated by the rapid growth of Chat GPT's popularity from the first days after its release, which made it the "*fastest-growing consumer application in history (...)"* (Hu, 2023). Although supposedly there have been significant differences between the usage of this tool among different users of the Stack Exchange platform, Chat GPT reached 1 million users in 5 days, and around 100 million users in January 2023. Additionally, topics related to Generative AI became a part of public discourse, including online communities, and therefore it is assumed that most of the users of CQA forums have been aware of how to use Chat GPT. This theory is supported by the results of a survey conducted by YouGov in January 2023 on a sample of 1000 respondents, 50% of respondents declared that they either used Chat GPT to generate text themselves or have seen someone generate text in front of them (YouGov, 2023).

The period from 1st July 2022 to 30th November 2022 was defined as the pre-intervention period for the treatment group and the period from 1st December 2022 to 30th of April 2023 as the post-intervention period for the treatment group. The exact same periods from previous years: 1st July 2021 to 30th November 2021, and 1st December 2021 to 30th of April 2022 were defined as a control group for pre-intervention and post-intervention periods respectively. For the estimations, the beginning and end of both control and treatment periods fall on week 27 and week 16 next year respectively, as described in Chapter 4.

5.1.1. Assumptions of the difference-in-differences model

As there is no analytics without theory, there are several underlying assumptions that have to be fulfilled in the case of the DiD model specified in the preceding section. The first key assumption for the planned study is the Stable Unit Treatment Value Assumption (SUTVA), which '*implies that the treatments are completely represented and, in particular, that there are no relevant interactions between the members of the population'* according to Lechner (2010). This assumption can be expressed in the mathematical notation as presented in Equation 3:

$$y_t = D * y_t^1 + (1 - D) * y_t^0 \quad (3)$$

$$t \in \{1,2\}$$

In the case of SUTVA violation, neither of the two potential outcomes can be observed. An example of such violation in this particular study could hypothetically be the usage of Generative

AI tools by individuals assigned to the control group (Stack Exchange users from the previous year, as described in Chapter 5.1.), meaning that (a part of) the control group has received the treatment. Because checking whether the post was created with the help of a GAI tool is not feasible for the collected data set, robustness checks on numerical dependent variables will be conducted by adding data for an additional control group in respective periods as defined in Chapter 5.1 from years 2018-2019. Given the fact that the unrestricted access to the API of the GPT-3 model (the predecessor of Chat GPT-3.5) was given to users on the 18th of November 2021, it is assumed in years 2018 and 2019 advanced generative models were not available to the broader public of community online forums.

The second critical assumption for the DiD estimation is the Parallel Trends Assumption (PTA) which implies that the trends in outcome values are parallel to each other before the intervention and that it is reasonable to assume that those trends would be maintained after the intervention if it did not take place (Dimick & Ryan, 2014). This assumption can be examined by graphical analysis of the outcome values over time or by conducting the Parallel Trends Test (PTT) checking the significance of the interaction terms between period indicators (for example dummy week variables) and the treatment group indicator (D). To ensure the robustness of the results, both methods will be used to evaluate if the Parallel Trends Assumption is maintained.

Ordinary Least Squares Regression will be used to calculate the values of the difference-in-differences estimator, and according to the Gauss-Markov theorem (Gauss, 1823) in order to obtain the best linear unbiased estimator, four critical assumptions have to be fulfilled. Equation 4 below presents a general version of the linear regression model:

$$Y = X * \beta + \varepsilon \quad (4)$$

In this equation, Y is the n*1 vector of dependent variable values, X is the n*p matrix of the independent variable values and ε represents the error term of the model. The OLS assumptions can be defined as follows (Greene, 2018, p. 246):

A.1 Linearity - this assumption requires a linear relationship specified between a dependent variable and an independent variable.

A.2 Full rank - this assumption is based on the condition that there are no exact relationships between independent variables, i.e. the vectors of independent variables cannot be specified as linear combinations of each other.

A.3 Exogeneity of independent variables (mean independence) - this assumption states that the expected value of the disturbance term in the model is not a function of the independent variables, which is referred to as mean independence.

A.4 Homoskedasticity - this assumption requires that each error term ε_i is uncorrelated with any other error term conditional on X.

Possible violations of those assumptions in the estimated OLS models will be investigated and if needed mitigated by applying appropriate measures. Poisson models with the same specifications will be estimated for count dependent variables to check the robustness of the results.

5.1.2. Econometric specification

A formal assessment of the validity of the parallel trends assumption following the approach from (Autor, 2003; Burtch et al., 2016) can be done by estimating a regression in which the treatment group indicator D_i is interacted with time dummies representing each period. In case of this study, data are aggregated on a tag combination and weekly level, and the equation of the parallel trends test can be specified as in Equation 5:

$$y_{it} = \alpha_i + \lambda_t + \sum \beta_t D_{it} + \varepsilon_{it} \qquad (5)$$
$$t \in \{-21, 20\} \setminus \{0\}$$

In this equation, the dependent variable y_{it} is regressed on the vector of group fixed effects α_i, the vector of time-fixed effects λ_t, and interaction terms between period indicators β_t and treatment group indicator D_{it}. In the context of this research, fixed group effects are the fixed effects of a given tag combination, and fixed time effects are the fixed effects of a given week of the year. ε_{it} represents the error term of the model. The period indicators β_t are binary variables which take the value 1 for a week of the year t and 0 for any other week of the year. One of the one-hot encoded period indicators has to be removed to avoid a dummy variable trap and therefore it serves the role of a baseline. If all period indicators would be included, the sum of their values would be equal to one for every observation, resulting in perfect multicollinearity (Greene, 2018, p. 157). Hence, it would violate the full rank assumption of the OLS estimator. Although any period indicator could be dropped, the period indicator representing the 48th week of the year (the first weeks of the post-intervention period) when the Chat GPT was introduced (in which $t = 0$) was chosen as a baseline, following the approach implemented by (Burtch et al., 2016). With this specification, the Parallel Trends Test will be conducted for all dependent variables excluding

content novelty measured by average cosine similarity of posts, due to the fact that only pairwise values of this measure for two data points can be obtained. Characteristics of this measure will be provided in Chapter 5.2.1.

After checking the PTA assumption, the difference-in-differences model will be estimated. Equation 6 below presents the econometric specification of the model:

$$y_{it} = \alpha_i + \lambda_t + \beta_2 T_t + \beta_3 D_i + \delta(T_t \cdot D_i) + \varepsilon_{it} \quad (6)$$

$$T \in \{0,1\}, D \in \{0,1\}$$

Similar to the PTT specification, y_{it} is the dependent variable, α_i denotes the vector of tag combination fixed effects, and λ_t represents the vector of weekly fixed effects. Because the fixed effects are included, no constant term β_2 is present in the model specification. Dependent variables of post count, post view count, post scores, post answer count, response scores and average readability are represented by y_{it}. The standard errors of estimates will be clustered by tag combination.

For content novelty model with a simple difference between averages of pairwise cosine similarity values (pairs containing one observation from pre- and post-intervention periods for every tag combination) will be calculated as outlined in Equation 7 below:

$$y = \beta_1 + \beta_3 D_i + \varepsilon_i \quad (7)$$

$$D_i \in \{0,1\}$$

In this specification, y is the dependent variable – average cosine similarity, the coefficient β_3 interacted with treatment group indicator D_i will be treated similarly to the DiD estimator - the interaction term in Equation 6. β_1 is the constant term of the regression. No fixed effects are included in the regression for content novelty, as the values of the pairwise cosine similarity are calculated for observations chosen randomly from pre- and post-intervention periods, therefore the time-fixed effects cannot be taken into account. The fixed effects for the tag combinations are also not included, as the pool of 50 pairs of questions that are randomly drawn (explained in detail in Chapter 5.2.1) does not represent equally the differences in means between tag combinations with a number of questions much higher than the minimum of 50 questions per each period, and tag combinations with a number of questions close to this minimum.

5.2. Natural Language Processing methods

To obtain data suitable for the difference-in-differences analysis for the dependent variables related to content novelty and content readability, textual data needs to be transformed into a numerical form. This section describes Natural Language Processing methods used for this process.

5.2.1. Content novelty measures

To assess the validity of the third hypothesis, Natural Language Processing methods will be used to calculate the change in the relative novelty of content posted by the Stack Exchange users. The textual novelty of content can be defined in multiple ways, with three types of novelty outlined in the study on the relationship between novelty and popularity of user-generated content by Carmel et al. (2010): contemporaneous novelty, self-novelty, and discussion novelty. The first type of novelty refers to the novelty of a given post compared to the contents of the other posts at the forum in a similar time period. The second term captures the relative novelty of a post authored by a particular user compared to the content created by this user in the past. The third definition of novelty compares the novelty of a given post to the novelty of the comments present under this post.

The approach implemented by (Burtch et al., 2022) to calculate the self-novelty of user-generated content by chosen Reddit users was adjusted to the Stack Exchange data set used in this study. Calculating the self-novelty of content was not possible in the case of this research, as a marginal number of users contributed at least twice in all four periods. Instead of focusing on the contribution by the particular user, it was decided that the main focus of the study would be on comparing the change in the contemporaneous content novelty between the pre- and post-intervention periods. In the further part of this study, the definition of contemporaneous content novelty will be used as a definition of content novelty.

A method commonly applied in previous research to calculate the semantic distance between two pieces of text is calculating the cosine similarity. This operation consists of several steps, beginning with transforming the textual data to a numeric representation. Such transformation requires choosing a technique used to create a vector representation of data. In this study, three common vectorization techniques will be used to construct embedding spaces in which the posts' content will be represented in a numerical form.

The first method is TF-IDF (Term Frequency-Inverse Document Frequency) which weights the words by their occurrence in a given document, normalizing the weights by the length of the document and scaling them by the inverse of word frequencies across documents in a chosen word corpus. Therefore words existing only in a single document would have a high TF-IDF score, while words occurring often in the corpus of documents would be scaled down (Burtch et al., 2022; Wang et al., 2019).

An alternative method is LSI (Latent Semantic Indexing), in which the occurrences of a given word with groups of other words in the analysed corpus of documents are used to calculate the vector representation of this word. The main difference compared to TF-IDF is that for LSI the meaning of the words is more important than their frequency of occurrence in the documents. LSI creates a vectorized version of a document by reducing its dimensionality (Burtch et al., 2022; Wang et al., 2019).

Another alternative method is an unsupervised method of Doc2Vec, which uses an unlabelled corpus of text to create a vectorized representation of a given document with reduced dimensions. This method utilizes neural networks to vectorize documents taking into account both the ordering of words and their semantic meaning, which is not taken into account in the bag-of-words methods. Another advantage of this method is that it can be trained on a larger external corpus before implementation on a particular data set, which can improve the tuning of the model (Burtch et al., 2022; Le & Mikolov, 2014).

After obtaining the vectorized form of the post content using a separate word corpus for each tag combination, cosines between the vectors are calculated. Cosines represent the distance between the vectors and can take values from -1 (vectors pointing in opposite directions) to 1 (identical vectors) (Hass, 2017, p. 238). The value of 0 means that the vectors are orthogonal and therefore not related. In the text mining context, the negative values are not observed, and the value of cosine between two non-negative vectors represents their similarity on a scale from 0, for non-related (novel) vectors, to 1 for identical text vectors (Al-Anzi & AbuZeina, 2017).

Given that the data contains more than 3.1 million observations, calculating a matrix with pairwise cosine similarity values was not feasible. To overcome this constraint, an approach based on bootstrapping was implemented on the tag combination level. Data were reduced to contain only tag combinations that appear at least 50 times in all four periods (pre- and post-intervention periods

respectively for the control and treatment groups). For each tag combination, 50 observations from the post-intervention period will be selected and for each of those 50 observations, 50 random observations will be drawn from the pool of the observations with the same tag combination in the pre-intervention period. The values will be saved separately, resulting in 50 average values, averaging out of a total of 2500 values of pairwise cosine similarity for each tag combination present in this part of the analysis. Finally, a simple difference-in-differences regression will be conducted, as specified in Equation 7.

5.2.2. Content readability measures

Validating the fourth Hypothesis requires calculating the change in the readability of posts. This dimension of content published on Stack Exchange will be assessed by calculating six different readability measures and taking an arithmetic average of their normalized values. As most of the commonly used readability metrics are defined similarly, it is expected that they will be highly correlated. Taking the average value will be done to mitigate possible bias that can apply only to one of the metrics. To ensure the robustness of this approach, regressions with each of the readability metrics will be run separately.

The average readability index metric will be based on Automated Readability Index, Flesch Reading Ease, Flesch-Kincaid Grade, Gunning Fog Index, SMOG Index and Coleman-Liau Index. These readability measures were used not only in the studies related to Stack Exchange platforms, such as the study on the quality of the Stack Overflow questions by Hodgins (2016) but also in studies on social media platforms, detecting financial fraud and political bias (Loughran & McDonald, 2020). Table 4 below presents how the chosen readability metrics are calculated:

Table 4: Formulas used for calculation of the chosen readability metrics and their interpretation

Readability Metric	Formula	Interpretation
Automated Readability Index	$4.71 * \frac{characters}{\text{words}} + 0.5 * \frac{words}{sentences} - 21.43$	A higher score means a higher US grade level required to understand the text
Flesch Reading Ease	$206.835 * (1.015 * \frac{words}{sentences}) (84.6 * \frac{\text{syllables}}{words})$	A higher score means a text is easier to understand
Flesch-Kincaid Grade	$(0.39 * \frac{words}{sentences} + (11.8 * \frac{\text{syllables}}{words}) - 15.59$	A higher score means a higher US grade level required to understand the text
Gunning Fog Index	$0.4 * (\frac{\text{words}}{sentences} + \frac{\text{polysyllables}}{words})$	A higher score means a higher US grade level required to understand the text
SMOG Index	$\sqrt{\text{polysyllables} * \frac{30}{sentences} + 3}$	A higher score means a higher number of years of education required to understand the text
Coleman-Liau Index	$(0.0588 * \frac{characters}{words}) - (0.296 * \frac{\text{words}}{sentences}) - 15.8$	A higher score means a higher US grade level required to understand the text

The scales used to assign the grade levels to Flesch Reading Ease are presented in Appendix F of this dissertation.

The average readability will be calculated as presented in Equation 8:

$$average\ readability = \frac{ARI_n + (1 - FRE_n) + FKG_n + GFG_n + SMOG_n + CLL_n}{6} \quad (8)$$

For each data point (representing a post published on Stack Exchange) the measures will be normalized, which is represented in the equation by the superscript *n*. Min-max normalization on a scale of 0 to 1 will be applied, which can be expressed as in Equation 9 below:

$$X' = (\frac{X - min(X)}{max(X) - min(X)}) * (new\ max(X) - new\ min(X)) + new\ min(X) \quad (9)$$

Where X' is the value of a normalized variable, X is the original value of a variable, $min(X)$ and $max(X)$ are the minimal and maximum values of a variable X in the data, and $new\ min(X)$ and $new\ max(X)$ are the desired boundaries of the new range of the variable X values. The two latter values are set to 0 and 1 respectively. The averaged readability measure has no direct interpretation (that could be linked to years of education needed to understand the text), but it indicates the relative readability of posts over time that can be interpreted. An increase in the values of the measure is linked to an increase in the average of underlying scores, meaning that in such cases the text requires more years of education to understand and is less readable on average.

6. Results

The first part of this chapter consists of the results of the parallel trends test conducted for the six dependent variables. In the following part, the results of difference-in-differences regressions with control variables are presented. Both parts also contain a description of additional analyses that were conducted to check the robustness of the results.

6.1. Parallel Trends Test

6.1.1. Parallel Trends Test results

Figures 4 – 6 present the visualisation of the weekly post count, weekly post answer count and average readability of posts for each of the 42 weeks of the year included in the analysed data set. The yellow dashed line represents week 48 in which Chat GPT-3.5 was released. This also marks the first week of the post-intervention period. The red dashed line was added as an auxiliary measure that marks the launch of Chat GPT-4 in week 11, which is faster, more accurate and has got superior capabilities than its predecessor, as discussed in Chapter 2.

Figure 4: Weekly post counts in the control and treatment groups

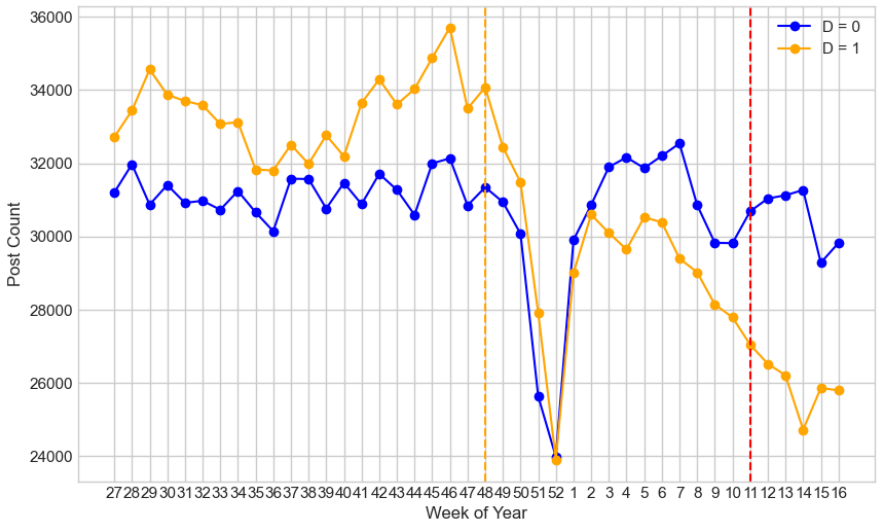
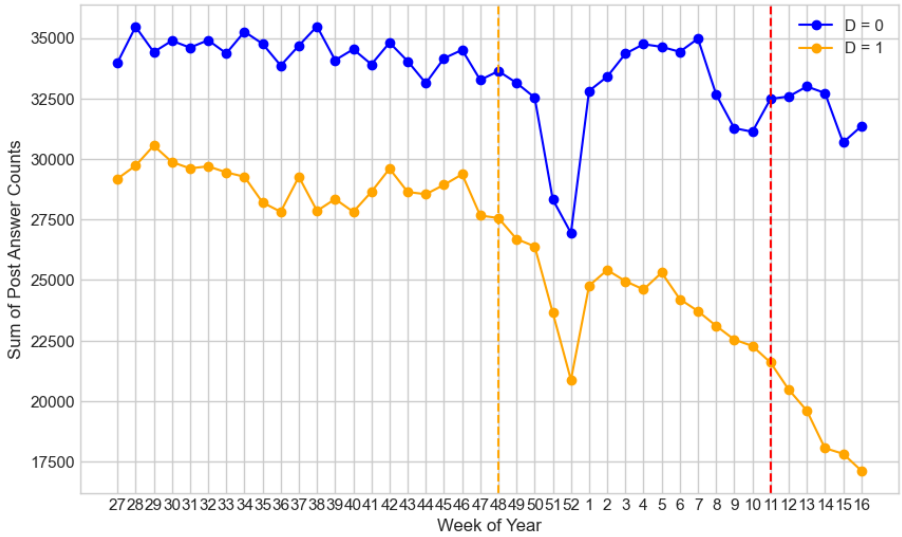


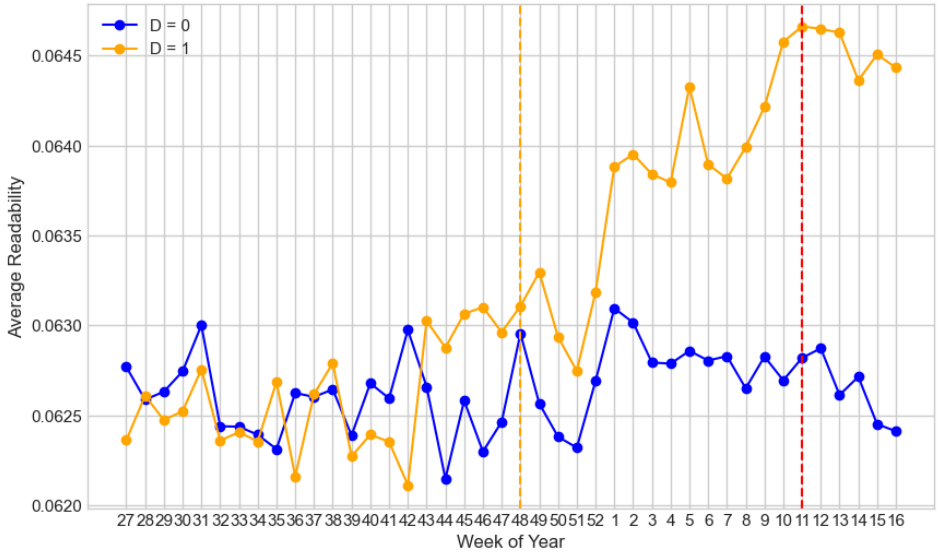
Figure 4 illustrates that the trends of weekly posts count were parallel till week 48 marking the intervention date. In the pre-intervention period, the number of posts in the control group was higher in the treatment group. After the decrease in the number of posts in the last weeks of the year, in weeks 1-6 the weekly count of posts in the treatment group gets smaller, than in the control group, and decreases sharply from week 6 onwards, leading to a difference of around 4000 posts weekly in the last weeks captured in the data set.

Figure 5: Weekly sum of post answers in the control and treatment groups



Trends in counts of post answers presented in Figure 5 are parallel till the intervention date and, similarly to weekly post counts, a decrease in the weekly number of answers can be observed at the end of the year in both groups. After that point, from the first weeks of the following year, the gap between the control and treatment groups is becoming gradually wider, with the most rapid decrease in the period after the introduction of Chat GPT-4.

Figure 6: Average post readability in the control and treatment groups



The trends in the average readability of posts presented in Figure 6 are hard to assess graphicly before week 42, as their mean values fall relatively close to each other. The mean values stay in the range from 0.062 to 0.063. In weeks 43-52, the average readability measure takes higher values for the treatment group, which indicates that posts in that group are harder to understand on average. From the beginning of the following year, the values of the average readability measure in the treatment group are rising to values close to 0.0645, meaning that the posts getting less readable compared to the control group. In contrast to the post count and post answer count, the momentum of decrease in the average readability measure is lower after week 9, however, the difference between the groups still increases.

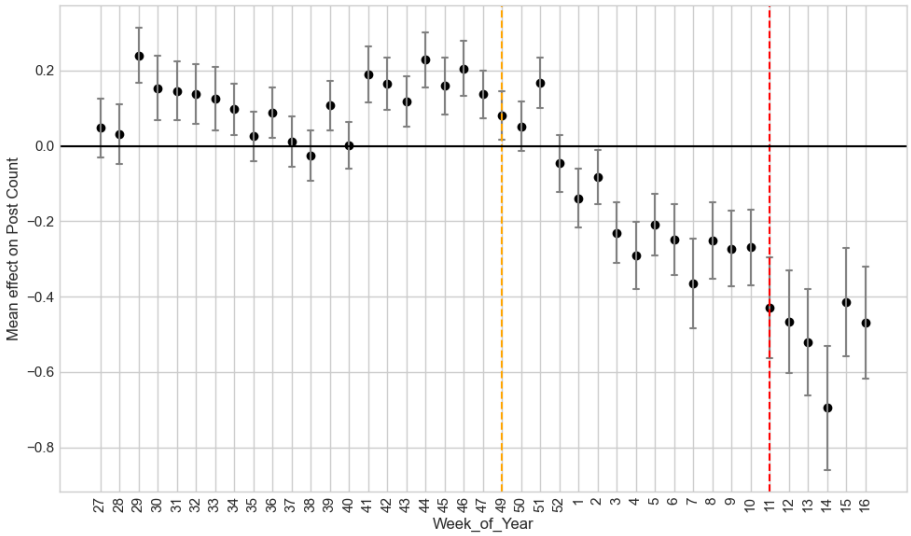
Graphical analysis of the weekly sum of post view count shows that the view count was much higher in the control group. The trends are parallel till the end of the year; from week 1 of the next year the difference between the groups increases. The visualisation of the weekly sum of post scores leads to a conclusion that the trends in the control and treatment groups are indeed parallel, however, the difference between the control and the treatment groups increases after week 13,

which may be connected to the rising popularity of the GPT-4 model. Similarly to other dependent variables, the values of post scores are higher for the control group. Regarding the weekly sum of response scores, the scores are also higher for the control group. Parallel trends are observed from week 27 till week 11 of the following year. Only after week 12 onwards a change in the difference between the control and the treatment group can be observed. The figures presenting the graphic assessment of parallel trends as well as visualisation of the parallel trends test for post view count, post scores and response scores are available in Appendix G.

The conclusions drawn by the graphical analysis are confirmed by the results of the Parallel Trend Test (that can be found in Appendix H) for the weeks in the post-intervention period, as the week indicators in those periods are significant at a 1% level in the case of all dependent variables, except for a few coefficients. However, this pattern is also present in many weeks in the pre-intervention period, in which the coefficients of the period indicators are also significant at a 1% confidence level for all dependent variables, excluding the average readability measure. For the latter, only the dummy variables of weeks directly before the intervention are significant, and the other dummy variables in the pre-intervention period are not significant.

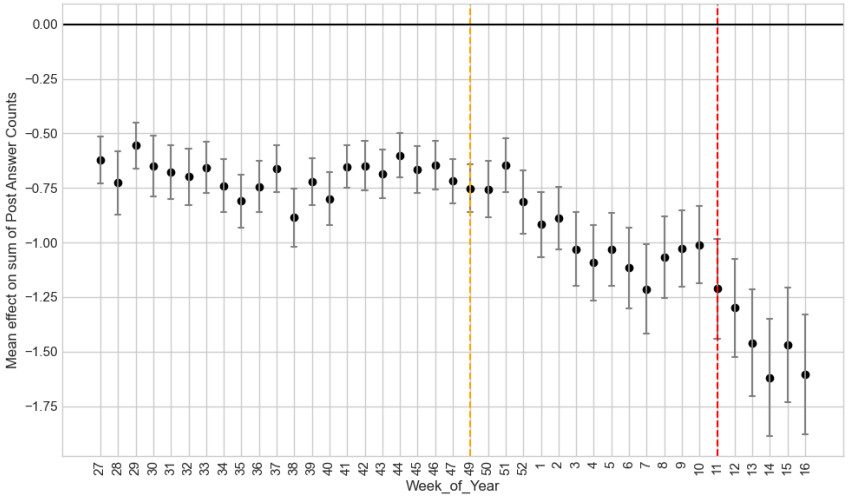
To visualise the results of the PTT for the three variables with the most notable differences between trends, in Figures 7-9 the values of estimated effects with 95% confidence intervals are shown. The interaction term between the treatment group indicator and the time dummy for week 48 (the first week of the post-intervention period) was omitted for the PTT to avoid the dummy variable trap. Therefore, this week serves as a baseline for the test. The values of the constant term suggest that week 48 was characterized by a mean post count close to the average value for the whole data set, a mean post view count higher by around 50% than the average, a mean post score higher by around 30% than the average, mean post answer count higher by around 15% than the average, mean response score higher by 25% than the average, and average readability close to mean value of this metric.

Figure 7: Results of Parallel Trends Test for weekly post count



Based on the PTT results visualised in Figure 7, it can be observed that the effects on weekly post count per tag combination have changed from positive in the pre-intervention period to negative in the post-intervention period, with values decreasing gradually. It is worth noting that the standard errors of the estimates were higher at the end of the analysed periods. The change in the number of posts in the last weeks (compared to the baseline of week 48) was equal to around -0.5 posts per tag combination.

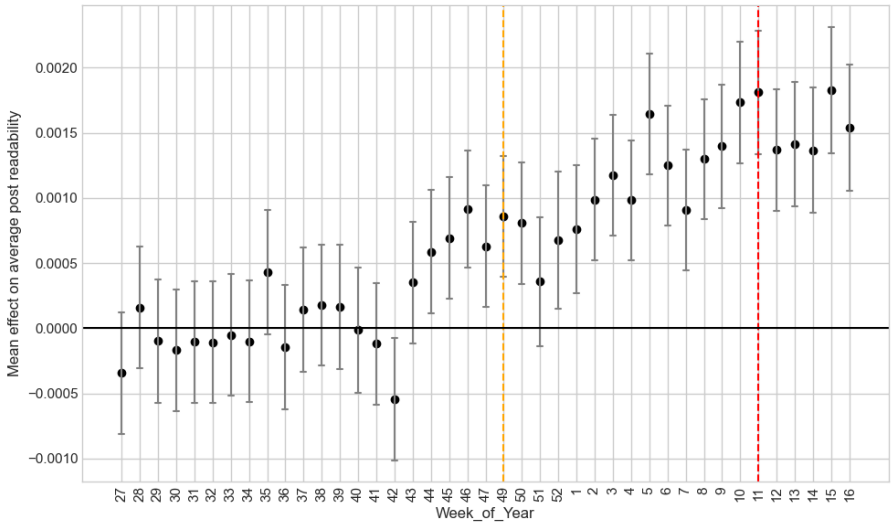
Figure 8: Results of Parallel Trends Test for a weekly sum of post answer count



As shown in Figure 8, changes between the years in the weekly sum of post answer counts are less visible than in the case of the weekly post count, however, a gradually increasing negative tendency after the intervention date can be observed. Similarly to post counts, the confidence

intervals are becoming wider in the last part of the sample, and the effect on the mean number of total post answers per tag combination is close to -1.5 compared to week 48.

Figure 9: Results of Parallel Trends Test for average readability of posts



Results of PTT for average readability plotted in Figure 9 strongly suggest a change in the trends between the years. It is worth underlying that this change started around week 44, 4 weeks before the intervention date. After that, the values of the average readability measure started to rise, denoting the posts becoming harder to understand on average. Compared to the post counts and post answer counts, the standard errors did not increase so visibly towards the end of the sample.

In conclusion, the results of the PTT suggest that the Parallel Trends Assumption is not formally fulfilled for all dependent variables, with the exception of the average readability. For the former five variables, the interaction terms between the week indicators and treatment group indicator are significant in both pre- and post- intervention periods, suggesting that the trends differ significantly throughout the whole analysed period.

On the contrary, the graphic examination of the trends shows that there is a notable change in the slope of the trends for post count, post answer count, and average readability measure that happens around the time of Chat GPT-3.5 release and gains momentum after the release of Chat GPT-4. Those conclusions are supported by the visualisation of coefficients estimated in the PTT for those variables, which show gradually increasing negative change in effects compared to the baseline week 48 for post count (for which some of the pre-intervention coefficients' confidence intervals contain zero), post answer count, and gradually increasing positive change in the average readability measure (meaning that the readability of posts decreases). Similar, but weaker negative

changes are observed for post scores. Post-view count and response scores both decrease in the post-intervention period compared to the pre-intervention period, however, they are not characterized by a strong change in the trend slope. The plotted coefficients estimated for PTT suggest that the decline is sharper for the former variable.

Given the results, the PTA assumption is not fulfilled for all tested variables except average readability, which creates a risk of possible bias in the difference-in-differences regression results. As a robustness check of the results, the significance of the linear trend will be assessed.

6.1.2. Robustness check

Another approach that can provide conclusions about the differences between two trends is conducting a regression, where the difference-in-differences estimator is interacted with a variable representing a linear trend. With the weekly level of data aggregation, the linear trend coefficient has been defined as 'Week Counter', taking value 1 for the first week of the year in the data – week 27, and then increasing incrementally by 1 for every next week, till reaching the value of 42 for the week 16 the following year. This method of analysis is based on the example presented by (Greene, 2018, pp. 173–174), modified accordingly for this study. The results of the regressions are presented in Appendix I. The interaction terms are significant on a 1% level for all dependent variables except response scores. Taking into account the results of the PTT and the graphic analysis suggesting that the trends of weekly response scores do not differ visibly in the post-intervention period except in the last four weeks, it is presumed that the parallel trends assumption is not fulfilled for this variable. Trends of other outcome variables seem to differ significantly between the treatment and control groups, which is confirmed by the significance of the interaction term between the DiD estimator and the linear trend. This result is consistent with the results of the Parallel Trends Test, as it confirms that there is a significant difference between the two trends. Nevertheless, due to the significance of the interaction terms in PTT for the pre-intervention period, the Parallel Trends Assumption is not deemed fulfilled for all dependent variables analysed in this test, except average readability.

At this point, it is worth emphasizing that although the PTA assumption has been proven not valid in the analysed period for five out of six dependent variables, there is a notable change in the trends' slopes after the release of Chat GPT-3.5 in trends for all variables, with the exception of response scores. Moreover, the discontinuity between trends for the control and treatment group

is clearly visible for all five variables in the last weeks of the analysed period, suggesting that the traffic on Stack Exchange decreased sharply after the introduction of a much more capable version of Chat GPT – GPT-4. Those observations are considered highly relevant for this study, as they complement recently published traffic analysis. A report published on *similarweb.com* shows that between the years 2022 and 2023, the number of visits on *stackoverflow.com* dropped by around 10% year-over-year in months from December to February, and by nearly 14% in March (Carr, 2023). CEO update for 2022 published by the Chief Executive Officer of *stackoverflow.com* revealed that the number of posts in the AI-related tags has declined by 20% between the years 2022 and 2021. In addition, an analysis conducted by one of the Stack Exchange users demonstrated that the decrease in the new question activity on *stackoverflow.com* from January onwards has been in the sharpest decline from 2018 (Meta Stack Exchange, 2023). Similarly to the trends presented in Figures 4-6, in January 2023 no rebound in user activity was noted, although it happened in the year 2022 and the previous years (captured in the referred analysis). Thus, based on the results from the analysis of parallel trends of this study and the findings of external reports, it is assumed that despite the trends not parallel in the analysed setting, conducting the difference-in-difference regressions will provide valuable insights for understanding the changing landscape of Q&A platforms.

6.2. The difference in differences estimation results

6.2.1. Results of Ordinary Least Squares regression

Results of the OLS regressions without controlling for post and body length of the posts are presented in Table 5 below:

Table 5: Results of OLS regressions without the control variables

	Post Count	Post View Count	Post Score	Post Answer Count	Response Scores	Avg Readability
DiD	-0.3606*** (0.0666)	-261.0986*** (26.6708)	-0.1491*** (0.0299)	-0.4008*** (0.0587)	-0.026 (0.0658)	0.0011*** (0.0001)
D	0.1162*** (0.0281)	-1283.9493*** (76.6694)	-0.9927*** (0.0518)	-0.6919*** (0.0439)	-1.7519*** (0.1018)	0.0001** (0.0001)
T	0.1626*** (0.0309)	136.7127*** (30.0127)	0.1589*** (0.0534)	0.2484*** (0.0394)	0.3314** (0.1313)	-0.0 (0.0003)
F-statistic	252.8743***	3631.9173***	2149.4353***	2353.5147***	1018.6076***	174.0752***
R-squared	0.0010	0.0144	0.0086	0.0094	0.0041	0.0007
Breusch-Pagan test p-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0014
Tag combination fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

*Legend: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$*

The coefficients of the difference-in-differences estimator have been found significant at 1% for all dependent variables except the response scores. For all variables the effect is negative, suggesting that the activity of Stack Exchange users in creating, answering, and verifying content on the platform has significantly decreased due to the release of Chat GPT. The view count of posts has also fallen, as well as their readability – higher values of the average readability metric denote a higher level of education required to understand the posts on average.

Based on the results, the average treatment effect of Chat GPT introduction resulted in a decrease of 0.36 posts, or 10.87% of post count mean per tag combination and week, on average. The respective decline for the post views count variable was equal to around 261.1 views, or 22.42% of its mean per tag combination and week, on average. The mean post score dropped by 0.15 points, or 8.57% of its mean value per tag combination and week, on average. The mean values of post answer count decreased by 0.40 answers, or 12.57% of its mean value per tag combination and week, on average. The average readability has decreased by 0.0011 or 1.75% of its average per tag combination and week, on average. It should be noted that this composite measure does not have a particular interpretation, hence values of each of the six measures included will be

regressed and analysed separately in Chapter 6.2.3. The presented interpretations of the effects of the OLS regression are valid under the assumption that the other coefficients of the regression are constant.

The estimated models include vectors of group-fixed effects for each tag combination and time-fixed effects for each week of the year. The constant term of the models is not included, as the fixed effects capture the characteristics that are constant over time for different tag combinations and the time characteristics of each week of the year. Robust standard errors clustered by each tag combination are used to counter the heteroskedasticity, which is indicated by the results of the Breusch-Pagan tests for all variables. Low values of R^2 suggest that the variability of the dependent variables is explained by other factors than the difference-in-differences estimator, period indicator and the treatment group assignment indicator. Control variables representing the length of post body and length of title turned out to be significant on 1% for all dependent variables excluding post count and response scores (for the latter title length was significant at the 5% level), however, their effects were relatively close to zero for all variables, supposedly due to low correlation with the depended variables (illustrated in the correlation matrix in Appendix J). Those two variables are not included in the further analysis. The results of the estimations including these control variables are presented in Appendix K.

Results of the regressions with the cosine similarity scores from pre- and post-intervention periods are presented in Table 6 below:

Table 6: Results of the regressions with cosine similarity values

	TF-IDF	**LSI**	**Doc2Vec**
Constant	0.0561***	0.1043***	0.5556***
	(0.0001)	(0.0002)	(0.0007)
D	-0.0005***	-0.0007**	-0.0015
	(0.0001)	(0.0003)	(0.001)
F-statistic	16.2041***	5.5838**	2.0995
R-squared	0	0	0
White's robust SE	Yes	Yes	Yes
Note: N = 781,821			

*Legend: *** - p < 0.01, ** - p < 0.05, * - p < 0.1*

The significance of the coefficients of treatment group indicator D on 1% and 5% for the bootstrapped average cosine similarity calculated for the contents of the post vectorized by using TF-IDF and LSI indicates that the introduction of Chat GPT increased the contemporary novelty

of content on Stack Exchange forums, as values of cosine similarity closer to 0 denote lower similarity between the vector representation of the text. It should be noted that the estimated effects for cosine similarity - a decrease of 0.0005 for TF-IDF and 0.0007 for LSI, have small magnitude, as in relative terms they constitute -0.9% and -0.67% of respective means.

The coefficient of the most advanced method taking into account the context of parts of the text - Doc2Vec is not significant on any commonly assumed level, which further undermines the validity of Hypothesis 3 saying that the Stack Exchange posts are becoming closer to each other in terms of semantic meaning in effect of the Chat GPT release. On the contrary, the Doc2Vec is the most resource-demanding method in terms of computing, which in the case of this study has severely limited the number of parameters that could be included in the model. The descriptive statistics of the cosine similarity values calculated and the overview of the parameters used for the computations is presented in Appendix L.

6.2.2. Heterogeneity analysis

Exploratory analysis of different dimensions of the Stack Exchange data used in this research has shown that content posted on different forums differs to a great extent. To overcome the constraints related to this heterogeneity on the forum level, an approach based on grouping by tag combination has been implemented. Taking into consideration the motives behind the chosen approach, it is expected that the difference-in-difference effects that were estimated differ greatly for different tag combinations. To check the heterogeneity of the results, effects on the group level were estimated for all tags included in the regressions. Values of DiD effects for the coefficients with the 15 highest negative and positive values are presented in Appendix M.

On the lists of the tag combinations with the highest negative effects, almost only programming-related tags can be observed, with the most popular programming languages such as Python, SQL, Java, and C++ appearing in the top 15. The highest negative effects reach values close to 1000 for post count, answer count and response scores, 500 for post score, and more than 100 thousand views for the post view count. A different pattern can be observed for the tag combinations with the highest positive effects – although most of the tag combinations are related to programming, tag combinations from other domains, such as '*expressionsidioms*', '*analytic-geometryvectors*' and '*woodwoodworking*' also appear on the list. Another difference is that the effects are much smaller, with an increase in post and answer count equal to around 20, an increase in post and

response scores of around 150 and an increase in view count of around 10 thousand views. Similar differences can be observed in the effects for average readability (the highest positive effects are presented along with the highest negative effects for the other variables, respectively for the highest negative values), but the values of this measure cannot be directly interpreted. Interestingly, the categories of tag combinations with the highest positive and negative effects are substantially different than the other tags – they refer to a greater extent to topics related to other technical domains than programming. An exploratory analysis of the occurrence of these tag combinations over time has been conducted, as it was hypothesized that those combinations have a higher increase in readability because these tags might be new on the forum and therefore, the threads may be more complicated. This hypothetical assumption was not confirmed by the analysis, showing that counts of the tag combinations did not differ substantially between the control and treatment groups. A possible explanation for the highest effects of this tag-combination might be related to their characteristics, as mathematical or engineering problems might on average require longer and more sophisticated descriptions than programming problems.

6.2.3. Robustness checks

To ensure that the results of this study are valid and accurate, the main concerns were defined and tested by appropriate statistical methods. The first main concern outlined was the validity of the assumptions of the OLS regression used for estimating DiD coefficients because it determines whether the OLS estimator is the best linear unbiased estimator. A possible violation of the assumption of homoskedasticity of the error term was found after conducting the Breusch-Pagan test, which resulted in rejecting the null hypothesis of this test saying that the error variances are constant in favour of the alternative hypothesis indicating heteroskedasticity. To mitigate this concern, robust standard errors were used for all estimations. Application of those errors, which are calculated in a more conservative way did not result in a change in the significance of the regressions.

A second possible violation of the OLS assumptions was the violation of the full rank assumption by a multicollinearity problem. To assess that, the Variance Inflation Factor (VIF) has been calculated and its values are presented in Table 7 below:

Table 7: Values of the Variance Inflation Factor (VIF)

Coefficient	VIF Value
DiD	2.9653
D	1.9392
T	2.0567
Body Length	1.0062
Title Length	1.0072

As the VIF scores are not exceeding the commonly assumed threshold of 5 suggesting a case of multicollinearity problem, the full rank assumption of the OLS method in this case is deemed to be fulfilled.

The second concern about the robustness of the results obtained in the study was the usage of the appropriate method. Ordinary Least Squares regression was chosen as the most suitable method that could be applied to all dependent variables and be interpreted easily, however, given that the three dependent variables are count variables, Poisson regression is a viable alternative. Therefore, three similar regressions were estimated, accounting for fixed effects and clustering the standard errors in the same way as presented in Chapter 6.2.1. The results are presented below in Table 8:

Table 8: Results of Poisson regressions

	Post Count	Post View Count	Post Answer Count
D	0.0359*** (0.0063)	-1.0342*** (0.0128)	-0.1949*** (0.0091)
T	0.0356*** (0.01)	0.2042*** (0.0438)	0.0977*** (0.0174)
DiD	-0.1114*** (-0.0098)	-0.7916*** (0.0193)	-0.1615*** (0.0136)
Log-Likelihood:	-1,121,029	-372,840,055	-1,217,772
Adjusted Pseudo R-squared	0.7383	0.8034	0.7514
Tag combination fixed effects	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes
SE clustered by tag combination	Yes	Yes	Yes

*Legend: *** - p < 0.01, ** - p < 0.05, * - p < 0.1*

Results of the Poisson regression confirm the results of the OLS regression, as the DiD estimators are also significant at 1% and they all yield negative average treatment effects. Estimated rate ratios are equal to 0.8904 for post count, 0.4531 for post view count and 0.8509 for post answer count. Based on the means of those variables, this is equivalent to a decrease of 0.35 posts per tag combination per week on average, 636.8 post views per tag combination per week on average and 0.48 post answers per tag combination per week, on average. Compared to the results of the OLS regression, the greatest difference can be observed for post view count, as the relative decrease is equal to 54.69% of the mean, compared to 22.42% of the mean estimated for OLS. For the post answer count the effect of Chat GPT estimated in the Poisson regression is slightly higher – 14.91% of mean compared to 12.57% for OLS, and the effect of the intervention on post count is nearly identical - 10.54% compared to 10.87% for OLS. These observations lead to the conclusion that the results of Poisson regression confirm the OLS results presented in Chapter 6.2.1.

Another concern regarded the choice of the year of the control group. Choosing years 2021-2022 affected by the SARS-CoV-2 pandemic might not be representative, as the activity of the Stack Exchange users might be influenced by many factors interlinked with the public health crisis. Also, this choice of control group entailed the risk of SUTVA violation, as the GAI tools were already available to some users, and despite their lower capabilities could already be used by those individuals to generate content posted on the Stack Exchange forums. To account for possible bias, the DiD regressions were conducted for an additional control group reflecting the Stack Exchange activity in the years 2018-2019. The same weeks as for the control group from years 2021-2022 were taken and the estimations with count and response variables (no readability metrics or content novelty) were conducted with similar settings. The results presented in Appendix N confirm the conclusions from the regressions with the control group from years 2021-2022, as coefficients of difference-in-difference of all dependent variables are significant at 1% level, including the DiD coefficient for response scores, which was not significant for the original control group. The estimated coefficients are negative, similar to the results presented in Chapter 6.2.1, and the estimated average treatment effects have a much bigger magnitude. The ratio between the effects presented in Table 5 and the results for the additional control group is equal to 5.7 for post count, 5.2 for post view count, 8.3 for post score and 6.2 for post answer count. For the response scores the effect is much higher, but this variable was not significant in the case of the main control group, so it cannot be compared. At this point, it is worth underlining that the mean values of those

variables are higher for the data set created by combining this additional control group with the treatment group from years 2022/2023, as the Stack Exchange forums had higher traffic in the years 2018-2019. Nevertheless, the results prove that the effects of Chat GPT introduction are significant for a control group consisting of years preceding the SARS-CoV-2 pandemic, which proves that the DiD results presented in Chapter 6.1.2 are robust in this context.

The final concern regards the significance of the average readability measure which was created from an average of normalized readability scores. Similar DiD estimations as for the average readability were conducted for each of the measures, taking their original (not normalized) values – the results are presented in Appendix O. The DiD estimator was significant for all six measures, however in the case of Flesch Reading Ease, Flesch-Kinkaid Grade and Gunning-Fog Index the DiD estimator was only significant at 5% level, and in case of Coleman Liau Index – at 10% level. The results are positive for all measures, suggesting a higher educational level is required to understand the text, except the Flesch Reading Ease, in case of which higher score results in higher readability.

For Automated Reading Index and Coleman Liu Index, the estimated average treatment effects are positive, meaning that a higher educational level is required for the comprehensive understanding of the texts, but they are also close to 0, and therefore – negligible. For the Flesch Reading Ease, a positive effect (which, considering the scale of this measure indicates an increase in readability) of 0.053 or 0.19% of the mean of this measure was estimated. The treatment effects for Flesch-Kinkaid Grade, Gunning Fog Index, and SMOG index are equal to 0.065, 0.037 and 0.19 US grade equivalent more needed on average to understand the Stack Exchange posts. Those values represent a 0.15%, 0.27% and 3.15% increase compared to the respective means of these measures. In conclusion, these results suggest that indeed the readability of posts has decreased given the negative effects for five out of six different readability measures, however, the changes have relatively larger error values compared to the results of regression with the average metrics and heterogeneity in the effects between the different measures is observed.

7. Discussion

The final chapter summarizes the findings of this study and assesses them critically, taking into account the limitations of the chosen research approach and possible bias of the results. After that,

the key managerial implications, and academic contributions are provided. The limitations of the study are presented alongside the future research recommendations.

7.1. Conclusion

The purpose of this study was to investigate what is the impact of the Generative Artificial Intelligence of the new generation, such as Chat GPT on the process of knowledge sharing on online Q&A platforms. The first chapters have provided arguments for why this question is highly relevant for various stakeholders and how it may influence not only the future of the platforms themselves but also education, business, academia, and the future development of AI. Based on the overarching research problem four hypotheses were defined, tackling the impact of Chat GPT's release on different dimensions of content posted on the Q&A platforms: number of posts, number of their views, and the sum of their scores, number of answers and the sum of their scores, semantic content novelty and content readability. A replicable modelling approach using open-source tools and a data set downloaded from a public repository has been developed. The code used for this research has been shared via the public GitHub repository (link placed in the Appendix) in order to enable replicating and expanding the results of this study in the future.

The first part of the analysis focuses on testing the validity of the Parallel Trends Assumption. For all dependent variables except content novelty measures the assumption was tested graphically, and by performing the Parallel Trends Test. According to the results of the test, the assumption was fulfilled only for average readability. Interaction terms of the other tested variables, except response scores were significant not only in the pre-intervention but also in the post-intervention period. The graphical analysis of the aggregated data and the results of the PTT indicated that close to the intervention week there is a notable change date in trends for post count and post answer count, and the difference in slopes is increasing towards the last weeks of the analysed data, which is the period when Chat GPT-4 became available. Similar, but weaker changes were observed for post view count, post scores and answer scores. The results of the PTT were validated by assessing the significance of the linear trend coefficient – week counter, which confirmed the significant results of the Parallel Trend Test.

The results of the study based on OLS difference-in-differences estimation show the introduction of Chat GPT has got a significant treatment effect on the number of posts and their scores, the number of answers, and the number of post views. The estimated effects for these variables were

significant at a 1% level. No significant treatment effect on response scores has been found. Estimated mean magnitudes of the effects were all negative and were equal to 10.87% for post count, 22.42% for post view count, 8.57% for post score and 12.57% for post answer count, in relation to the respective mean values of those variables.

DiD regressions estimated for the cosine similarity indicated that the cosine similarity of posts decreased by 0.9% of the mean for the TF-IDF method and by 0.67% of the mean for the LSI method. Those results were significant at 1% and 5% levels respectively and they indicate that the novelty of content posted increased as a result of Chat GPT-3.5 introduction, as lower average cosine similarity denotes more novel content in terms of contemporary novelty. The difference in the averaged cosine similarity values was not significant on any commonly assumed level for the text vectorized by using the Doc2Vec method, which weakens the conclusions drawn, as TF-IDF and LSI tend to be less accurate on average.

The analysis of the readability demonstrated that the release of Chat GPT 3.5 had significantly decreased the readability of the posts' text published on Stack Exchange forums. The estimated average treatment effect was significant on a 1% level and equal to an increase of 1.75% of the mean value of average readability, indicating that a higher educational level is required to understand comprehensively the description of the posts. Separate estimations have shown that from the six readability measures taken into account, five are significant on the 5% level, and two on the 1% level. Some heterogeneity between the measures was observed, with the SMOG index having the highest partial increase of 3.15% which was significant at the 1% level, other measures having effects smaller than 0.3%, and Flesch Reading Ease having a different direction of effect as expected. As described in the introductory chapter of this study, it is assumed that decreasing readability of posts indicates their higher sophistication, as users tend to solve easier problems by using GAI tools and limit their activity on the Q&A platform to more sophisticated questions.

Heterogeneity analysis of the treatment effects by tag combination has shown that the strongest DiD negative effects on a group level are at least 5 times higher than for the tag combinations with the strongest positive effects for all dependent variables. Furthermore, the tag combinations with the top negative effects seem to consist only of tag combinations related to the most popular programming languages, while in the top positive group effects, other tag combinations are present. Additionally, it was noted that the top tags for both positive and negative effects are substantially different for average readability, as they include tags from a variety of different

domains and forums. The results are in line with the research context presented in Chapter 2, namely that Chat GPT is especially accurate and reliable in coding, compared to other domains.

The results of DiD regressions have shown robustness for checks performed, including Poisson regression for the count variables, and estimating the OLS DiD regressions with the control group from the years 2018-2019. In both cases, the effects were significant at the same level as in the main analysis, with the addition of a significant effect for response scores in the case of the control group from the years 2018-2019. The treatment effects estimated by conducting Poisson regression for count variables were notably higher for post view count, slightly higher for post answer count and comparable to OLS regression results for post count.

In light of the outcomes derived from the performed analyses, Hypothesis 1 and Hypothesis 2 are both rejected, due to the invalidity of the Parallel Trends Assumption. The significant difference in the trends in the pre-treatment period does not allow to separate the causal effect of Chat GPT-3.5 release on both the demand and supply side of user Q&A activity analysed in the study, as other factors that are not included in the model may cause the difference in trends. Nevertheless, the results obtained prove that all dependent variables included (except response scores) in those hypotheses have significantly declined and that there is a notable change in the trends of those variables. The significance of DiD estimates is maintained also for the Poisson model, and for a different control group from years 2018-2019. Given the behaviour of the trends confirmed by external statistics (presented in Chapter 6), especially after the release of Chat GPT-4, it is expected that the causal effect of Chat GPT introduction can be identified by a similar study in the future using longer pre- and post-intervention periods.

Given the results for the average cosine similarity measure, Hypothesis 3 saying that the release of Chat GPT-3.5 caused a significant decrease in the novelty of textual content posted on Q&A forums is also rejected. The results are significant on a 5% level for two out of three approaches used to obtain an embedding space of words present in the analysed content of posts, however, the estimated effects suggest a slight increase in content novelty.

Hypothesis 4 claiming that the Chat GPT-3.5 introduction decreased the readability of posts on the Q&A platforms is deemed valid. This decision is based on the validity of the parallel trends assumption and results of the DiD regression for the composite average readability score, created as a normalized average of six commonly used readability metrics. The estimated average treatment effect is significant at 1% for the OLS regression and yields a positive effect, indicating

a higher educational level needed to understand the posts, thus lower readability. A possible concern undermining the validity of this hypothesis is the heterogeneity of effects for different readability metrics, however except for the Flesch Reading Ease and Coleman-Liau index, effects of those metrics are significant on a 5% level. Moreover, the significant differences in the magnitude of the effects presented in the heterogeneity analysis suggest that the causal effect might not be present for all groups included in the research. The difference in the tag combinations with the highest and negative effects makes it difficult to systematically identify the groups with high and low effects, however, the negative (in terms of readability) effects are much more common and have higher values than the positive effects, which only partially weakens the rationale under accepting the Hypothesis 4.

7.2. Managerial implications

Understanding the ongoing changes in the knowledge exchange process is extremely important for all stakeholders involved in this process. The study shows that even though the release of GPT 3.5 has not yet significantly influenced the number of questions, post views, answers, and question ratings on the most popular Stack Exchange forums, all of these quantities have been decreasing sharply from the last weeks of 2022. This tendency suggests that the knowledge exchange might become dependent on rapidly developing Large Language machine learning models. What is more, the decrease in the readability of posts in the effect of the Chat GPT introduction, which has been proven significant in this study, shows that assumingly GAI is a better choice for solving standard and repetitive problems. Although choosing GAI tools may enhance the efficiency of solving simple and everyday problems that previously required more time and effort, it decreases the contribution to online communities, which are a reusable source of knowledge.

Results of the analysis conducted in this study show that the user activity decreases from both: the supply site – the number of answers, and the demand site – the number of questions. In addition, the view count of questions posted and the rating of questions is also in decline, which suggests that the role of Q&A forums as a knowledge source is becoming less important than ever before. Although the effect of Chat GPT introduction on those changes cannot be yet deemed significant, it is expected this trend will continue in the future. This may severely impair the development of the Q&A community and perhaps stop it altogether, as a growing share of users will use GAI tools for solving their problems. Lower engagement in both asking and responding to questions posted

on platforms such as Stack Overflow might lower the inclination of the users previously engaged in CQA communities for the purpose of self-development. Crowd problem-solving on such forums engages the responding party to overcome the given challenge and the asking party to verify the answer received, and in comparison, interacting with tools such as Chat GPT does not incentivize the users for any kind of verification. Moreover, problem-solving with the use of GAI tools does not leave a sharable and accessible resource, as opposed to a discussion on a Q&A platform.

This states the strategic implication of this study – substituting the process of knowledge exchange done and verified by individuals to using advanced generative tools might decrease the creation of new learning resources and diminish the educational role of the Q&A platforms, both from the supply and demand side. This twofold reduction may result in a scarcity of verified and accessible sources of knowledge in the future. This implication is extremely relevant to all decision-makers dealing with the need for domain-specific information on demand, which refers to the majority of the sectors in today's digital economy.

Diminishing the supply of knowledge on the Q&A forums is a problem that requires a prompt response from decision-makers in business, otherwise, vast resources of online knowledge might never be produced in the future. Currently, many leading companies have started adapting their business models to respond to the new opportunities given by GAI and threats posed by these solutions, with a focus on developing in-house GAI tools and training employees in using them. Given the results of this study, this can be called a 'bottom-up' approach, as the clue is to understand how tools such as Chat GPT disrupt the knowledge-sharing process and what might be the consequences of this phenomenon for different organisations. This relates also to the technology companies developing the models - GPT-4 using more than 100 billion parameters might be the best-performing LLM in history, however, if the set of data used for training, which contained books, news articles and Stack Exchange posts will rapidly decrease, the new Open AI's model will not have access to novel information. This refers to the change in the quantity of data available, however, the quality is an even more important dimension. How will the new LLM models be capable of generating creative responses, when their training data are not up-to-date, and it is not documented well?

7.3. Academic contributions

This paper fills a highly relevant gap in the academical literature, providing insights on how modern knowledge exchange changes due to recent developments in Generative Artificial Intelligence. It expands the existing work on Q&A forums, which have played an important role in the scientific community, providing a transparent, free, and accessible space for solving advanced and abstract problems. The conclusion of this dissertation goes beyond the literature published so far, providing insights about especially relevant concern on the disruption of one of the most important channels of knowledge exchange in the modern world.

First of all, this study contributes to the academic community by responding in a timely manner to the research on the effects of the popularization of GAI tools on society. By indicating how various dimensions the user activity on one of the leading Q&A platforms has been changing, valuable insights are provided for the literature on online communities, machine learning and Natural Language Processing. This area of research is highly relevant in academia, which is reflected in the number of working papers regarding various effects of widespread GAI that have already been published in recent months. Analysing the effects of Chat GPT introduction on the activity of the most popular Stack Exchange platforms complements these findings and provides relevant insights and recommendations for future studies, especially those using causal inference methods. It also provides valuable results for Q&A forums, which are the key stakeholder in the knowledge exchange process and need the information on how the human quest for information is changing.

Furthermore, the research approach used to analyse the Stack Exchange data are novel and extends the existing quantitative research conducted on observational data. The scope of the analysis includes both question- and answer- related measures, which differs from the approach of a study by (Goes et al., 2016), where only the answers were analysed. Combining the numerical and textual analysis of data gives a wider overview of the characteristics of the forums and expands the recent work of (Burtch et al., 2022) by adding the readability component to the scope of research. Moreover, the specification of panel regressions in this study creates an alternative approach to (Chen et al., 2018), focusing on data aggregated on tag combination, instead of aggregating on user level. This approach allows to better account for the heterogeneity of the content posted on the platforms, than aggregating data on user, forum, or subforum level. Using the data from a previous year as a control group for the difference-in-difference regressions extends the work of (Eichenbaum et al., 2020) by applying this novel approach in a new context. In addition, the

resources shared by the publication of the study are suitable for open-source software and freely accessible, which can contribute to future studies in areas similar to this dissertation.

7.4.Limitations and recommendations

Within the quasi-experimental research approach implemented in this study, a series of assumptions limiting the outcome of the study were made. First and foremost, the difference-in-differences method treats the release date (week) of Chat GPT as an intervention date, however, the usage of this tool among Stack Exchange community members has increased gradually over time, reaching the level of 100 million users in January 2023. As a recommendation for future studies, using a dynamic DiD model could be a viable option, however, it would also require a broader range of data. Reasons for including a longer period of data will be outlined in the next paragraph.

The second significant limitation of the study was the range of data that could be collected, which is related to the Parallel Trends Test. As the results have shown, the last weeks in the data set are characterized by the strongest negative change, which is significant for most of the dependent variables. Conducting studies with a longer time horizon in the future could differentiate the trends better and draw more precise conclusions. As outlined in the preceding sections, it is anticipated that future similar studies will be able to prove that there is a negative causal effect of Chat GPT introduction on activity on Q&A forums, as the decrease in activity will continue, but it is only stated as an expectation. An additional limitation related to the range of data are related to the possibility that the last months are the peak of Chat GPT's popularity and that the enormous expectations about this tool encourage people to use it more often. In this scenario, observing a causal effect might not be possible, but future studies in this area will have more choices regarding the collection of the data set, and it is advisable to compare trends of the outcome variables to observe what periods should be analysed.

Another limitation of the study was the computing power available for NLP pre-processing and text vectorization, especially for the Doc2Vec method. Due to a large number of observations exceeding 2.4 million rows, those operations had to be limited to extracting the most important features, and also in the case of Doc2Vec, the training data had to be reduced. More computing power, especially operating memory, would provide capabilities to test a wider range of parameters of the vectorizers. With access to such resources, the Doc2Vec model could be trained on a larger

external corpus, which could improve the quality of text prediction. Furthermore, the bootstrapping part in content novelty calculations could be replaced with calculations of all values, which could potentially increase the reliability of estimates.

One limitation related to data collection is also considered relevant to the results, namely the difference in the share of deleted posts between the treatment and control groups. Default queries submitted to the Stack Exchange Data Explorer do not return information about the number of deleted questions, however, the information about their count can be extracted. It is assumed that the share of deleted questions is higher for the control group, as users tend to remove questions over time for different reasons. Including this quantity in similar studies can improve the reliability of the results.

The final limitation is related to the chosen level of group aggregation – the tag combination. Although this approach is effective in dealing with the content heterogeneity on different Stack Exchange forums on one hand and generating fewer groups than by grouping by tag 2 or tag 3 on the other hand, it makes the heterogeneity analysis much more difficult, as the number of groups is very high. Moreover, the occurrence of such high effects makes isolating the causal effect difficult when using a difference-in-difference estimator. As a recommendation for future research, it is suggested that a different level of aggregation might be sought, for example, based on content classification by NLP analysis of the post's contents.

8. Appendices

Below the appendices are provided in the order in which they appear in the main text.

Additionally, Python and R code used to obtain the results is available in the public GitHub repository: https://github.com/peterthebest444/Chatbot-Overflow---Piotr-Piwnik

Appendix A: Description of the variables in the data aggregated on a weekly basis

Variable name	**Post Creation Date**	**Post Count**	**Post Answer Count**	**Post Score**	**Response Scores**	**Post View Count**	**Readability metrics***	**Average readability**
Meaning	Date and time when a post was created	Number of posts in a given week	Number of answers under a given post	Score that the post acquired from the users	Sum of response scores of all answers	Number of views of the post from the date of creation	Values of chosen readability metrics	Average value of normalised and adjusted readability metrics
Data Type	datetime	numerical	numerical	numerical	numerical	numerical	numerical	numerical, values in the range (0,1)
Type of aggregation	-	Weekly count of rows in the data	Weekly sum	Weekly sum	Weekly sum	Weekly sum	Weekly mean	Mean of aggregated readability metrics
Number of observations	2,593,153	2,593,153	2,593,153	2,593,153	2,593,153	2,593,153	2,476,321	2,476,321

*Description of the readability metrics is presented in Chapter 5.2. and their descriptive statistics are presented in Appendices C and E

Appendix B: Number of observations with distinction by Stack Exchange forum in the not aggregated data set

Forum Name	N	Forum Name	N
Stack Overflow	2,377,476	Home Improvement	11,738
Mathematics	160,185	Ask Different	11,622
Stack Overflow in Russian	71,733	Mathematica	10,134
Stack Overflow in Spanish	39,813	Database Administrators	9,720
Ask Ubuntu	36,361	English	7,031
Super User	35,043	Magento	6,999
MathOverflow	33,192	WordPress	6,914
Physics	31,446	Stack Overflow in Portuguese	6,773
Statistical Analysis	28,514	Science Fiction and Fantasy	6,529
Unix and Linux	26,914	Code Review	5,007
TeX - LaTeX	26,909	IT Security	4,561
Electrical Engineering	25,294	SharePoint	4,167
Blender	22,618	Arqade	4,021
GIS	18,565	Android Enthusiasts	3,831
Server Fault	18,473	Game Development	3,545
Salesforce	15,468	Meta Stack Exchange	3,371
Ethereum	14,630	Drupal Answers	2,791
English Language Learners	13,912	Software Engineering	2,668
		Sum of all forums	**3,107,968**

Appendix C: Descriptive statistics of the chosen readability measures in the aggregated data set

	ARI	FRE	FK Grade	Gunning Fog	SMOG Index	Coleman Liau Index
count	781,821	781,821	781,821	781,821	781,821	781,821
mean	10.142723	70.695801	7.808621	9.634146	8.198190	7.801096
std	6.873068	28.996033	5.449808	3.821009	3.912263	6.883033
min	-11.600000	-2,371.076667	-15.700000	0.000000	0.000000	-33.810000
25%	7.000000	60.140000	6.200000	7.730000	6.690625	6.233333
50%	9.200000	68.963333	7.900000	9.337500	9.050000	7.830000
75%	12.000000	76.945000	9.800000	11.150000	10.700000	9.630000
max	1,670.30	206.84	346.83	260.97	43.10	2,044.39

Appendix D: Descriptive Statistics of the dependent variables, Title- and Body Length in the reduced, not aggregated data set

	Post_View_Count	Post_Score	Post_Answer_Count	Response_Scores	Title_Length	Body_Length
count	2,593,153	2,593,153	2,593,153	2,593,153	2,593,153	2,593,153
mean	351.07	0.52	0.96	1.07	61.43	1,823.96
std	2,247.84	2.58	0.91	6.89	24.30	2,359.49
min	2.00	-654.00	0.00	-14.00	15.00	38.00
25%	45.00	0.00	0.00	0.00	44.00	668.00
50%	94.00	0.00	1.00	0.00	57.00	1,173.00
75%	266.00	1.00	1.00	1.00	74.00	2,085.00
max	1,539,162.00	1,442.00	217.00	4,076.00	150.00	115,919.00

Appendix E: Descriptive Statistics of the readability measures in the reduced, not aggregated data set

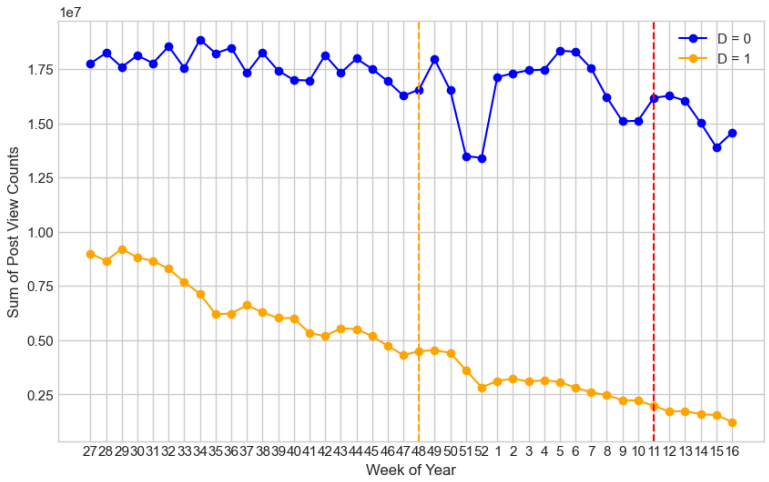
	ARI	FRE	FK Grade	Gunning Fog	SMOG_Index	Coleman Liau Index
count	2,481,784	2,481,784	2,481,784	2,481,784	2,481,784	2,481,784
mean	9.853448	73.283445	7.510565	9.563541	7.696757	7.169151
std	8.089382	36.140468	6.870696	4.800160	4.525574	8.095224
min	-11.600000	-7,238.990000	-15.700000	0.000000	0.000000	-33.810000
25%	6.100000	60.240000	5.700000	7.200000	6.400000	5.680000
50%	8.700000	70.330000	7.600000	9.100000	8.800000	7.460000
75%	12.000000	79.400000	9.900000	11.330000	10.700000	9.390000
max	3,200.70	206.84	1,023.20	760.40	63.80	3,742.00

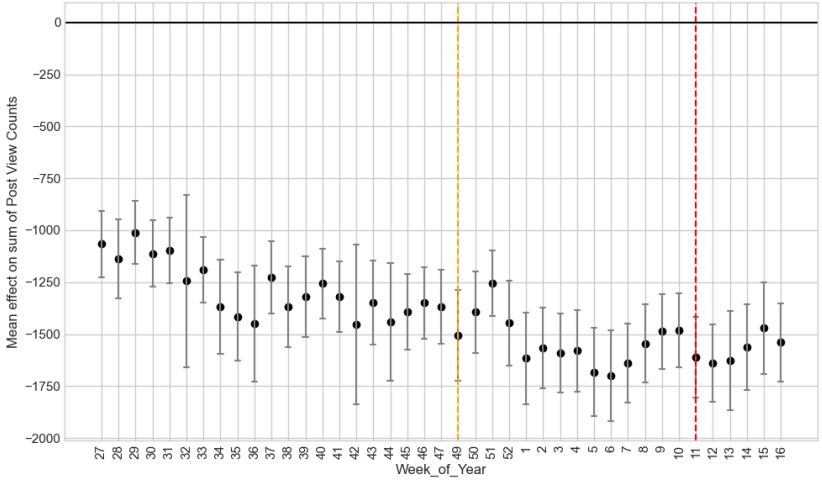
Appendix F: Scale for the school levels (US grade equivalents) used to interpret the Flesch–Kincaid Reading Ease (Wydick & Flesch, 1980)

Score Range	School Level (US)	Notes
100.00–90.00	5th grade	Very easy to read. Easily understood by an average 11-year-old student.
90.0–80.0	6th grade	Easy to read. Conversational English for consumers.
80.0–70.0	7th grade	Fairly easy to read.
70.0–60.0	8th & 9th grade	Plain English. Easily understood by 13- to 15-year-old students.
60.0–50.0	10th to 12th grade	Fairly difficult to read.
50.0–30.0	College	Difficult to read.
30.0–10.0	College Graduate	Very difficult to read. Best understood by university graduates.
10.0–0.0	Professional	Extremely difficult to read. Best understood by university graduates.

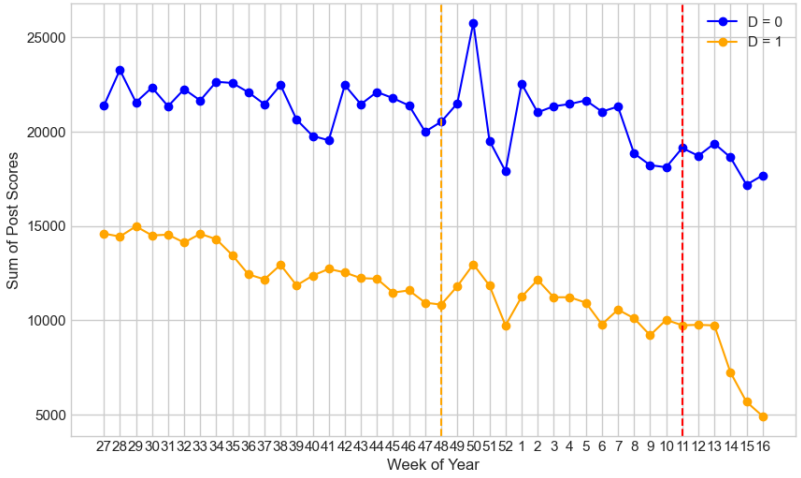
Appendix G: Results of the Parallel Trends Tests (1/2)

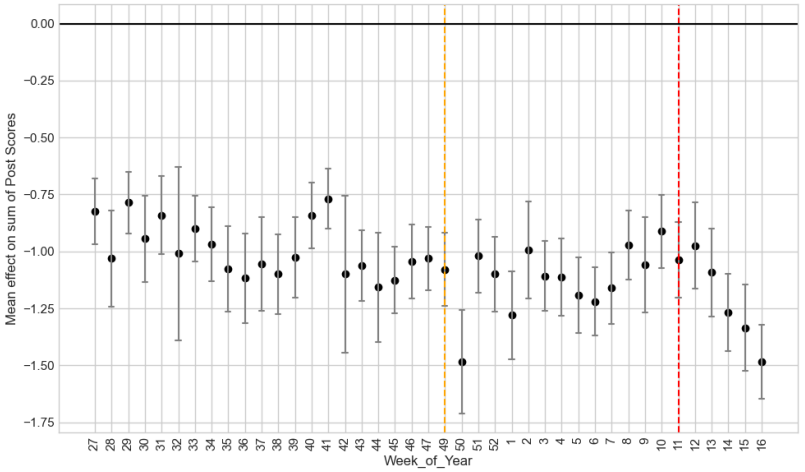
Post View Count





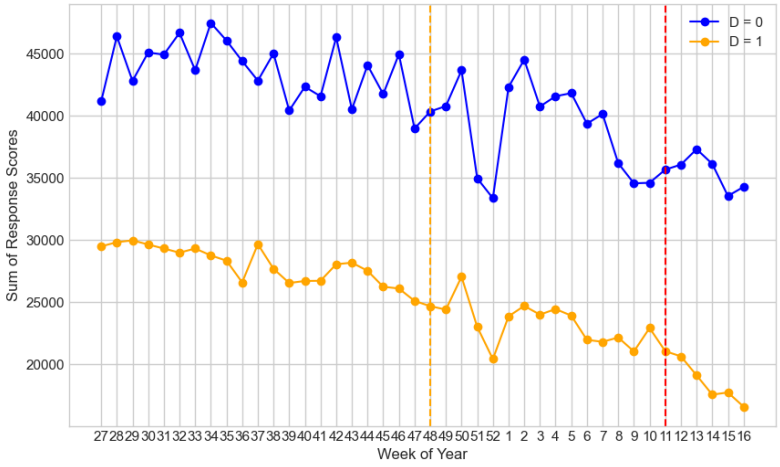
Post Score

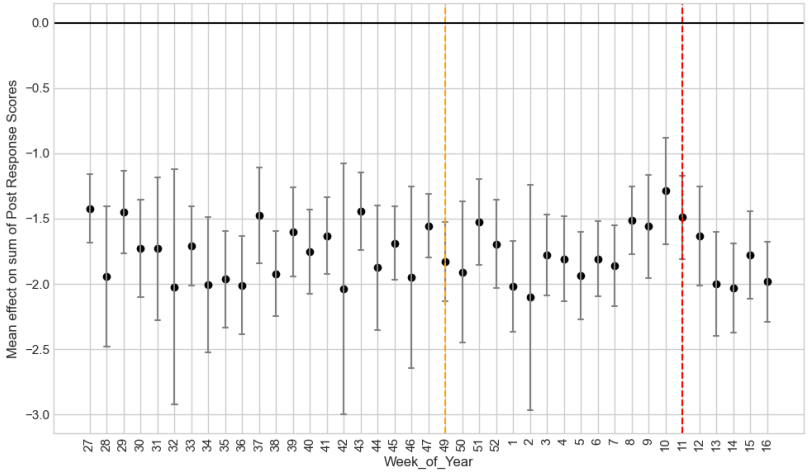




Appendix G: Results of the Parallel Trends Tests (2/2)

Response Scores





Appendix H: Results of the Parallel Trends Test

	Post Count	Post View Count	Post Score	Post Answer Count	Response Scores	Avg Readability
const	3.3466***	1871.1266***	2.2743***	3.6335***	4.431***	0.0626***
Week_of_Year_27_D	0.0477	-1065.6795***	-0.8235***	-0.6223***	-1.4228***	-0.0003
Week_of_Year_28_D	0.0319	-1137.6325***	-1.0315***	-0.7263***	-1.9424***	0.0002
Week_of_Year_29_D	0.2405***	-1010.86***	-0.786***	-0.556***	-1.4484***	-0.0001
Week_of_Year_30_D	0.1537***	-1112.533***	-0.9448***	-0.6482***	-1.7271***	-0.0002
Week_of_Year_31_D	0.1459***	-1096.4072***	-0.8413***	-0.677***	-1.73***	-0.0001
Week_of_Year_32_D	0.1385***	-1243.6643***	-1.0087***	-0.6976***	-2.0223***	-0.0001
Week_of_Year_33_D	0.1264***	-1191.911***	-0.9004***	-0.6568***	-1.7091***	-0.0001
Week_of_Year_34_D	0.0972***	-1369.6965***	-0.9693***	-0.7397***	-2.0074***	-0.0001
Week_of_Year_35_D	0.0262	-1415.5293***	-1.0759***	-0.8095***	-1.9635***	0.0004*
Week_of_Year_36_D	0.0882***	-1450.0613***	-1.1178***	-0.7432***	-2.011***	-0.0001
Week_of_Year_37_D	0.0106	-1227.5383***	-1.0545***	-0.6611***	-1.4735***	0.0001
Week_of_Year_38_D	-0.0257	-1368.6914***	-1.0998***	-0.886***	-1.9225***	0.0002
Week_of_Year_39_D	0.1069***	-1317.9301***	-1.0256***	-0.7209***	-1.601***	0.0002
Week_of_Year_40_D	0.0023	-1255.7568***	-0.8424***	-0.7988***	-1.7539***	-0.0
Week_of_Year_41_D	0.1898***	-1320.9583***	-0.769***	-0.6522***	-1.6305***	-0.0001
Week_of_Year_42_D	0.1652***	-1453.6467***	-1.1***	-0.6483***	-2.0363***	-0.0005**
Week_of_Year_43_D	0.1173***	-1347.3928***	-1.0612***	-0.685***	-1.4419***	0.0004
Week_of_Year_44_D	0.2286***	-1441.7517***	-1.1575***	-0.6004***	-1.876***	0.0006**
Week_of_Year_45_D	0.159***	-1391.5991***	-1.1261***	-0.6651***	-1.6881***	0.0007***
Week_of_Year_46_D	0.2057***	-1348.6968***	-1.0435***	-0.6459***	-1.9492***	0.0009***
Week_of_Year_47_D	0.137***	-1369.1626***	-1.031***	-0.7175***	-1.5549***	0.0006***
Week_of_Year_49_D	0.0802**	-1504.5885***	-1.0789***	-0.7517***	-1.8279***	0.0009***
Week_of_Year_50_D	0.0517	-1393.1505***	-1.4835***	-0.755***	-1.9085***	0.0008***
Week_of_Year_51_D	0.168***	-1256.0709***	-1.0203***	-0.6444***	-1.5263***	0.0004
Week_of_Year_52_D	-0.0464	-1446.4865***	-1.1001***	-0.8145***	-1.6939***	0.0007**

Week_of_Year_1_D	-0.1385***	-1616.336***	-1.2798***	-0.9171***	-2.0178***	0.0008***
Week_of_Year_2_D	-0.0824**	-1566.5741***	-0.9928***	-0.8896***	-2.1037***	0.001***
Week_of_Year_3_D	-0.2306***	-1589.4506***	-1.1075***	-1.0296***	-1.779***	0.0012***
Week_of_Year_4_D	-0.2907***	-1580.2576***	-1.1124***	-1.0919***	-1.8076***	0.001***
Week_of_Year_5_D	-0.208***	-1682.4256***	-1.1926***	-1.0317***	-1.9375***	0.0016***
Week_of_Year_6_D	-0.2483***	-1700.107***	-1.2198***	-1.1169***	-1.8077***	0.0012***
Week_of_Year_7_D	-0.3644***	-1640.9558***	-1.1613***	-1.2127***	-1.8596***	0.0009***
Week_of_Year_8_D	-0.2508***	-1545.119***	-0.9736***	-1.0682***	-1.5124***	0.0013***
Week_of_Year_9_D	-0.2727***	-1486.7471***	-1.0586***	-1.0267***	-1.5599***	0.0014***
Week_of_Year_10_D	-0.2695***	-1482.7664***	-0.9113***	-1.0104***	-1.2875***	0.0017***
Week_of_Year_11_D	-0.4297***	-1610.793***	-1.0358***	-1.2113***	-1.4903***	0.0018***
Week_of_Year_12_D	-0.4663***	-1639.4486***	-0.9744***	-1.2993***	-1.6343***	0.0014***
Week_of_Year_13_D	-0.5216***	-1627.6615***	-1.0922***	-1.4594***	-1.9992***	0.0014***
Week_of_Year_14_D	-0.6949***	-1561.5186***	-1.2689***	-1.6183***	-2.0324***	0.0014***
Week_of_Year_15_D	-0.4146***	-1470.8359***	-1.3353***	-1.4696***	-1.7767***	0.0018***
Week_of_Year_16_D	-0.47***	-1539.1***	-1.4844***	-1.6040***	-1.9839***	0.0015***
F-statistic	31.5804***	261.5434***	155.9101***	177.5284***	73.9078***	15.2748***
R-squared	0.0017	0.0141	0.0085	0.0096	0.0040	0.0008

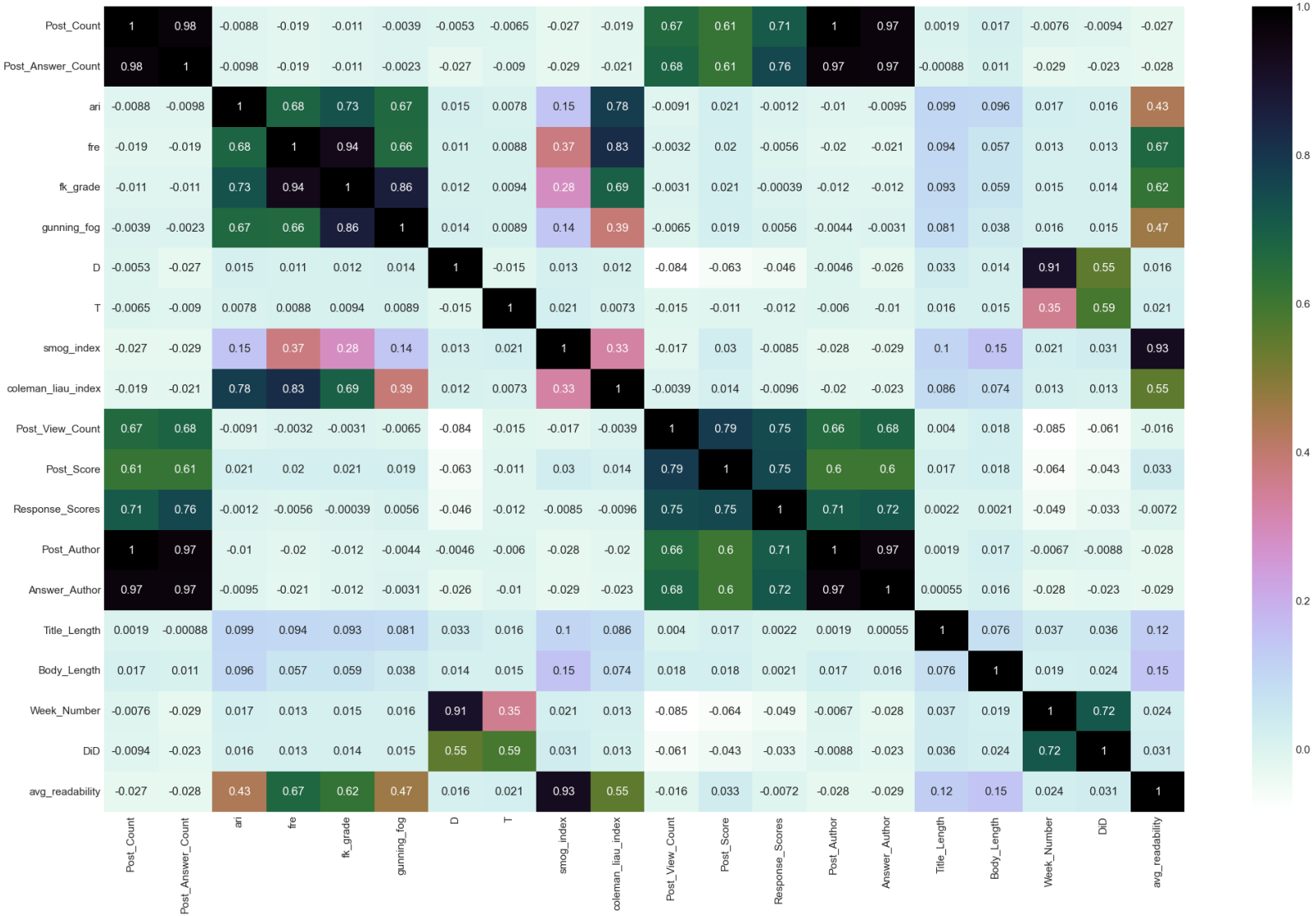
*Legend: *** - p < 0.01, ** - p < 0.05, * - p < 0.1*

Appendix I: Results of evaluation of the Parallel Trend Assumption by regression with a linear trend indicator (week counter)

	Post Count	**Post View Count**	**Post Score**	**Post Answer Count**	**Response Scores**	**Avg Readability**
D	0.1376*** (0.0094)	-1285.4453*** (18.532)	-0.9945*** (0.0182)	-0.6636*** (0.0148)	-1.7549*** (0.0437)	0.0001** (0.0001)
T	0.1332*** (0.0514)	101.466 (101.0748)	0.138 (0.0991)	0.2194*** (0.0805)	0.3255 (0.2381)	0.0001 (0.0003)
DiD * Week Counter	-0.0126*** (0.0004)	-8.0149*** (0.8128)	-0.0045*** (0.0008)	-0.0143*** (0.0006)	-0.0006 (0.0019)	0.0*** (0.0)
F-statistic	332.5134***	3633.2602***	2149.602***	2400.5321***	1018.5866***	185.1632***
R-squared	0.0013	0.0144	0.0086	0.0095	0.0041	0.0007
tag combination fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes

*Legend: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$*

Appendix J: Correlation Matrix of all features used in the data set



Appendix K: Results of OLS regression with control variables – Body Length and Title Length

	Post Count	**Post View Count**	**Post Score**	**Post Answer Count**	**Response Scores**	**Avg Readability**
DiD	-0.3607*** (0.0666)	-264.0461*** (26.6451)	-0.1555*** (0.0299)	-0.3996*** (0.0586)	-0.0267 (0.0659)	0.0009*** (0.0001)
D	0.1161*** (0.0281)	-1285.2122*** (76.6565)	-0.9949*** (0.0518)	-0.6915*** (0.0439)	-1.7529*** (0.1018)	0.0001 (0.0001)
T	0.1627*** (0.0309)	138.6119*** (30.0045)	0.1631*** (0.0534)	0.2476*** (0.0394)	0.3318** (0.1313)	0.0001 (0.0003)
Body_Length	-0.0 (0.0)	0.017*** (0.0014)	0.0*** (0.0)	-0.0*** (0.0)	-0.0 (0.0)	0.0*** (0.0)
Title_Length	0.0001 (0.0001)	1.2364*** (0.1338)	0.002*** (0.0002)	-0.0004*** (0.0001)	0.0011** (0.0006)	0.0001*** (0.0)
F-statistic	151.7876***	2186.9149***	1331.0218***	1414.2819***	611.6883***	5369.1364***
R-squared	0.0010	0.0144	0.0088	0.0094	0.0041	0.0347
Number of Weeks	42	42	42	42	42	42
tag combination fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
SE clustered by tag combination	Yes	Yes	Yes	Yes	Yes	Yes

*Legend: *** - p < 0.01, ** - p < 0.05, * - p < 0.1*

Appendix L (1/2): Descriptive statistics of cosine similarity values for the techniques used to construct the embedding spaces

	TF-IDF	**LSI**	**Doc2Vec**
count	118,800	118,800	118,800
mean	0.055851	0.103989	0.554882
std	0.022429	0.049683	0.180365
min	0.000000	0.000000	-0.674654
25%	0.041999	0.071479	0.489815
50%	0.056113	0.099889	0.595396
75%	0.070502	0.133112	0.671736
max	0.158286	0.349725	0.888278

Appendix L (2/2): Parameters used for the post text vectorisation in the process of calculating the average cosine similarity (gensim, 2023b, 2023a; sklearn, 2023)

Parameter from a given Python library	Python library	Description	Value		
			TF-IDF	LSI	Doc2Vec
max_features	sklearn (TF-IDF), gensim (LSI)	Number of the most frequent terms acrosss the corpus that are included to build the dictionary	1000	1000	-
n_gram_range	sklearn	Upper boundary of n-grams included in the model	2	-	-
vector_size	gensim	Dimensionality of the feature vectors	-	-	10
window	gensim	Maximum distance between the current and predicted word within a sentence	-	-	5
min_count	gensim	Ignores all words with total frequency lower than this	-	-	5
workers	gensim	Number of threads used for computing	-	-	8
epochs	gensim	Number of iterations (epochs) over the corpus	-	-	10

Appendix M: Results of heterogeneity analysis – 15 tag combinations with the highest negative effects (1/2)

Post Count		Post View Count		Post Score	
pandaspython	-893.2533	javascriptreactjs	-643783.5720	javascriptreactjs	-478.4663
javascriptreactjs	-740.7256	pandaspython	-312777.5107	pandaspython	-397.3266
htmljavascript	-668.9153	javascriptnode.js	-297352.9079	discussionstack-exchange	-251.4167
pythonpython-3.x	-508.6339	dartflutter	-260363.4058	dartflutter	-227.7975
python	-503.4314	htmljavascript	-259274.0049	javascriptnode.js	-224.4630
csshtml	-423.3255	pythonpython-3.x	-243107.9370	htmljavascript	-208.8040
javascriptnode.js	-401.6447	csshtml	-155678.4896	pythonpython-3.x	-191.6834
dartflutter	-315.7113	reactjstypescript	-148365.5491	androidkotlin	-180.6186
javascript	-314.8282	androidkotlin	-143532.5604	iosswift	-172.3302
djangopython	-273.7442	javaspring	-139754.2171	computability-theory	-166.0000
androidjava	-235.7877	djangopython	-135632.8891	csshtml	-133.5175
iosswift	-208.3690	python	-122977.5892	djangopython	-128.2485
arraysjavascript	-202.7780	javaspring-boot	-120940.4325	reactjstypescript	-123.3222
excelvba	-189.6938	androidjava	-119100.3467	.netc#	-119.7228
laravelphp	-170.2608	iosswift	-116954.9536	discussionprofile-page	-111.0000

Appendix M: Results of heterogeneity analysis – 15 tag combinations with the highest negative effects (2/2)

Post Answer Count		Response Scores		Avg Readability	
pandaspython	-1176.6004	pandaspython	-1033.5275	customizationphp	0.2253
javascriptreactjs	-906.8552	javascriptreactjs	-888.9696	authorization	0.1900
htmljavascript	-839.4820	pythonpython-3.x	-633.7985	circuit-analysistransfer-function	0.1733
python	-734.9848	designdiscussion	-567.5000	homotopy-type-theorytype-theory	0.1718
pythonpython-3.x	-646.3487	python	-548.4814	derivativespower-series	0.1680
csshtml	-611.7551	htmljavascript	-510.8068	ap.analysis-of-pdeshyperbolic-pde	0.1670
javascript	-442.1119	discussionprofile-page	-508.5000	fermionsquantum-mechanics	0.1647
dartflutter	-437.9001	discussionstack-exchange	-499.4167	electromagnetismfield-theory	0.1607
javascriptnode.js	-428.2747	dartflutter	-408.4937	apache-2.4httpd.conf	0.1595
arraysjavascript	-357.0928	csshtml	-369.1897	oa.operator-algebrasrt.representation-theory	0.1556
djangopython	-286.1396	javascriptnode.js	-363.5679	goproxy	0.1552
listpython	-277.8747	javascript	-315.8124	analysismanifolds	0.1525
androidjava	-225.5338	computability-theory	-292.0000	cachingx86	0.1511
r	-216.6475	arraysjavascript	-254.4218	autofac	0.1505
sqlsql-server	-213.2188	androidkotlin	-252.0059	estimationmathematical-statistics	0.1503

Appendix M: Results of heterogeneity analysis – 15 tag combinations with the highest positive effects (1/2)

Post Count		Post View Count		Post Score	
pythonselenium-webdriver	27.9194	amsmath	13224.925	designdiscussion	134.333
javaselenium-webdriver	6.1582	serverssh	9202.2128	python	62.6714
analytic-geometryvectors	4.0000	consoleterminal	7908.0000	discussionvote-to-close	53.0000
pythonpython-polars	2.5536	discussionvote-to-close	6811.0000	nt.number-theorypr.probability	44.5000
c#maui	2.2883	securitysql	5771.1667	securitysql	44.3333
quadratic-forms	2.0000	nlptext	5116.8000	answerssupport	40.1667
predicate-logicquantifiers	2.0000	american-englishsingle-word-requests	5008.5000	gn.general-topologygt.geometric-topology	32.0833
allocatorc++	2.0000	gitwindows	4706.4580	sqlsql-server	30.1191
tdengine	1.9146	goparsing	4661.3333	expressionsidioms	28.2222
maui	1.7512	tolkiens-legendarium	4107.3333	discussionreputation	22.8905
memgraphdb	1.6826	bolts	3940.8333	java	22.7868
erlangfunctional-programming	1.6667	designdiscussion	3711.6667	nt.number-theoryprime-numbers	20.7173
angularangularjs	1.3412	sql-server-2019	3324.0000	american-englishsingle-word-requests	18.5000
voltage	1.3000	c++synchronization	3206.3333	bountiesfeature-request	18.3333
python-3.xselenium-webdriver	1.2957	c#vb.net	3179.6698	c++synchronization	17.5000

Appendix M: Results of heterogeneity analysis – 15 tag combinations with the highest positive effects (2/2)

Post Answer Count		Response Scores		Avg. Readability	
pythonselenium-webdriver	22.5536	discussionreputation	186.9857	jspmysql	-0.0581
discussionreputation	11.5571	securitysql	147.6667	javascriptwhatsapp	-0.0434
bolts	7.3333	bolts	99.3333	java	-0.0363
gr.group-theoryra.rings-and-algebras	6.8095	gr.group-theoryra.rings-and-algebras	85.4762	c#filter	-0.0347
journals	6.0000	discussionvote-to-close	77.0000	phptxt	-0.0341
javaselenium-webdriver	5.8220	doctor-who	74.4167	inputlaravel	-0.0338
javascriptstack	5.5000	american-englishsingle-word-requests	73.3333	querysql	-0.0329
ho.history-overviewreference-request	5.0000	expressionsidioms	65.7333	journals	-0.0328
predicate-logicquantifiers	5.0000	c++synchronization	63.1667	datatablessql	-0.0325
capacitorsurface-mount	4.5000	tolkiens-legendarium	59.5000	php7	-0.0323
woodwoodworking	4.1000	bootsecurity	40.6667	sqlite3	-0.0318
erlangfunctional-programming	4.0000	fpgamicrocontroller	40.5000	dll	-0.0317
securitysql	3.8333	grammaticalityverbs	39.7500	asyncc#	-0.0303
cputemperature	3.5000	sql-injection	39.6000	arraylistpython-3.x	-0.0295
american-englishidioms	3.3333	journals	38.0000	google-chromeselenium-webdriver	-0.0291

Appendix N: Results of OLS difference-in-difference regressions with additional control group from years 2018-2019

	Post Count	**Post View Count**	**Post Score**	**Post Answer Count**	**Response Scores**
DiD	-2.0669*** (0.179)	-1372.9101*** (155.1841)	-1.2353*** (0.1319)	-2.4778*** (0.2328)	-2.618*** (0.279)
D	1.1685*** (0.1606)	-4304.2505*** (202.4708)	-2.7395*** (0.1203)	-0.1797 (0.1537)	-4.4529*** (0.2289)
T	0.799*** (0.0798)	280.3535** (130.3756)	0.316*** (0.1074)	0.9338*** (0.1007)	0.7603*** (0.2523)
F-statistic	1240.1843***	4434.4875***	3617.888***	2085.0856***	2537.5657***
R-squared	0.0050	0.0178	0.0145	0.0084	0.0102
Observations	777,623	777,623	777,623	777,623	777,623
Number of Weeks	42	42	42	42	42
tag combination fixed effects	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes
SE clustered by tag combination	Yes	Yes	Yes	Yes	Yes

*Legend: *** - p < 0.01, ** - p < 0.05, * - p < 0.1*

Appendix O: Results of separate regressions with the chosen readability metrics

	ARI	FRE	FK_Grade	Gunning_Fog	SMOG_index	Coleman_Liau_Index
DiD	0.0*** (0.0)	0.0001** (0.0)	0.0001** (0.0001)	0.0001** (0.0001)	0.006*** (0.0004)	0.0* (0.0)
D	0.0001*** (0.0)	0.0002*** (0.0)	0.0003*** (0.0)	0.0004*** (0.0)	-0.0002 (0.0003)	0.0*** (0.0)
T	0.0001 (0.0001)	0.0002 (0.0002)	0.0003 (0.0002)	0.0003 (0.0003)	-0.001 (0.0016)	0.0 (0.0)
F-statistic	87.793***	32.1892***	49.9073***	73.0082***	146.4136***	27.0276***
R-squared	0.0004	0.0001	0.0002	0.0003	0.0006	0.0001
Number of Weeks	42	42	42	42	42	42
tag combination fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
SE clustered by tag combination	Yes	Yes	Yes	Yes	Yes	Yes

*Legend: *** - $p < 0.01$, ** - $p < 0.05$, * - $p < 0.1$*

9. References

Abid, A., Farooqi, M., & Zou, J. (2021). *Persistent Anti-Muslim Bias in Large Language Models*. https://doi.org/10.48550/ARXIV.2101.05783

Al-Anzi, F. S., & AbuZeina, D. (2017). Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. *Journal of King Saud University - Computer and Information Sciences*, *29*(2), 189–195. https://doi.org/10.1016/j.jksuci.2016.04.001

Albouy, D. (2020). *Program Evaluation and the Difference in Difference Estimator*. University of California, Berkeley:

Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Discovering value from community activity on focused question answering sites: A case study of stack overflow. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 850–858. https://doi.org/10.1145/2339530.2339665

Anyoha, R. (2017, August 28). The History of Artificial Intelligence. *BLOG, SPECIAL EDITION ON ARTIFICIAL INTELLIGENCE*. https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

Archive.org. (2023, June 11). *Archive.org*. https://archive.org/details/texts

Autor, D. H. (2003). Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, *21*(1), 1–42. https://doi.org/10.1086/344122

Baidoo-Anu, D., & Owusu Ansah, L. (2023). Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4337484

Balakrishnan, V., & Ethel, L.-Y. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. *Lecture Notes on Software Engineering*, *2*(3), 262–267. https://doi.org/10.7763/LNSE.2014.V2.134

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. *Département d'Informatique et Recherche Opérationnelle Centre de Recherche Mathématiques Université de Montréal, Montréal, Québec, Canada*, 19.

Burke, P. (2000). *A social history of knowledge: From Gutenberg to Diderot, based on the first series of Vonhoff Lectures given at the University of Groningen (Netherlands)*. Polity Press ; Blackwell Publishers.

Burtch, G., Carnahan, S., & Greenwood, B. N. (2016). Can You Gig it? An Empirical Examination of the Gig-Economy and Entrepreneurial Activity. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2744352

Burtch, G., He, Q., Hong, Y., & Lee, D. (2022). How Do Peer Awards Motivate Creative Content? Experimental Evidence from Reddit. *Management Science*, *68*(5), 3488–3506. https://doi.org/10.1287/mnsc.2021.4040

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics*, *225*(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*. https://doi.org/10.48550/ARXIV.2303.04226

Carmel, D., Roitman, H., & Yom-Tov, E. (2010). On the relationship between novelty and popularity of user-generated content. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1509–1512. https://doi.org/10.1145/1871437.1871659

Carr, D. (2023, April 27). *Stack Overflow is ChatGPT Casualty: Traffic Down 14% in March*. https://www.similarweb.com/blog/insights/ai-news/stack-overflow-chatgpt/

Chen, W., Wei, X., & Zhu, K. X. (2018). Engaging Voluntary Contributions in Online Communities: A Hidden Markov Model. *MIS Quarterly*, *42*(1), 83–100. https://doi.org/10.25300/MISQ/2018/14196

Chui, M., Roberts, R., & Yee, L. (2022, December 20). Generative AI is here: How tools like ChatGPT could change your business. *McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/generative-ai-is-here-how-tools-like-chatgpt-could-change-your-business#/

Dimick, J. B., & Ryan, A. M. (2014). Methods for Evaluating Changes in Health Care Policy: The Difference-in-Differences Approach. *JAMA*, *312*(22), 2401. https://doi.org/10.1001/jama.2014.16153

Dondio, P., & Shaheen, S. (2019). Is StackOverflow an Effective Complement to Gaining Practical Knowledge Compared to Traditional Computer Science Learning? *Proceedings of the 2019 11th International Conference on Education Technology and Computers*, 132–138. https://doi.org/10.1145/3369255.3369258

Eichenbaum, M. S., Godinho de Matos, M., Lima, F., Rebelo, S., & Trabandt, M. (2020). *Expectations, Infections, and Economic Activity*.

Gauss, C.-F. (1823). *Theoria combinationis observationum erroribus minimis obnoxiae*. Henricus Dieterich.

Geigle, C., Dev, H., Sundaram, H., & Zhai, C. (2019). A Generative Model for Discovering Action-Based Roles and Community Role Compositions on Community Question Answering Platforms. *Proceedings of the International AAAI Conference on Web and Social Media*, *13*, 181–192. https://doi.org/10.1609/icwsm.v13i01.3220

gensim. (2023a). *Gensim Doc2Vec Documentation.* https://radimrehurek.com/gensim/models/doc2vec.html

gensim. (2023b). *Gensim LSI Documentation*. https://radimrehurek.com/gensim/models/lsimodel.html

Goes, P. B., Guo, C., & Lin, M. (2016). Do Incentive Hierarchies Induce User Effort? Evidence from an Online Knowledge Exchange. *Information Systems Research*, *27*(3), 497–516. https://doi.org/10.1287/isre.2016.0635

GPT-4 by Open AI. (2023). https://openai.com/product/gpt-4

Greene, W. H. (2018). *Econometric analysis* (Eighth edition). Pearson.

Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008). Predictors of answer quality in online Q&A sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 865–874. https://doi.org/10.1145/1357054.1357191

Hass, R. W. (2017). Tracking the dynamics of divergent thinking via semantic distance: Analytic methods and theoretical implications. *Memory & Cognition*, *45*(2), 233–244. https://doi.org/10.3758/s13421-016-0659-y

Hodgins, G. (2016). Classifying the Quality of Questions and Answers From Stack Overflow. *University of Dublin, Trinity College*.

Hu, K. (2023, February 2). *ChatGPT sets record for fastest-growing user base—Analyst note*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Huang, C., Yao, L., Wang, X., Benatallah, B., & Zhang, X. (2018). *Software Expert Discovery via Knowledge Domain Embeddings in a Collaborative Network*.

Hughes, A. (2023, April 3). ChatGPT: Everything you need to know about OpenAI's GPT-4 tool. *Science Focus*. https://www.sciencefocus.com/future-technology/gpt-3/

IBM. (2023). *A Computer Called Watson*. https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/

International Telecommunication Union. (2022). *Measuring digital development: Facts and figures 2022*. https://www.itu.int/hub/publication/d-ind-ict_mdd-2022/

Katz, D. M., Bommarito, M. J., Gao, S., & Arredondo, P. (2023). GPT-4 Passes the Bar Exam. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4389233

Khyani, D., & B S, S. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. *Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology*, *22*, 350–357.

Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*. https://doi.org/10.48550/ARXIV.1405.4053

Lechner, M. (2010). The Estimation of Causal Effects by Difference-in-Difference MethodsEstimation of Spatial Panels. *Foundations and Trends® in Econometrics*, *4*(3), 165–224. https://doi.org/10.1561/0800000014

Longoni, C., Fradkin, A., Cian, L., & Pennycook, G. (2022). News from Generative Artificial Intelligence Is Believed Less. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 97–106. https://doi.org/10.1145/3531146.3533077

Loughran, T., & McDonald, B. (2020). Textual Analysis in Finance. *Annual Review of Financial Economics*, *12*(1), 357–375. https://doi.org/10.1146/annurev-financial-012820-032249

McGuffie, K., & Newhouse, A. (2020). *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*. https://doi.org/10.48550/ARXIV.2009.06807

Meta Stack Exchange. (2023, May 24). *Did Stack Exchange's traffic go down since ChatGPT?* https://meta.stackexchange.com/questions/387278/did-stack-exchanges-traffic-go-down-since-chatgpt

Miller, G. A., Newman, E. B., & Friedman, E. A. (1958). *Length-Frequency Statistics for Written English*. *Harvard University*, 370–389.

Ministry of Education, Singapore. (2023, February 7). *Managing the use of artificial intelligence (AI) bots such as ChatGPT in schools*. https://www.moe.gov.sg/news/parliamentary-replies/20230207-managing-the-use-of-artificial-intelligence-bots-such-as-chatgpt-in-schools

Mollick, E. (2022, December 14). ChatGPT Is a Tipping Point for AI. *2023.* https://hbr.org/2022/12/chatgpt-is-a-tipping-point-for-ai

Murgia, A., Janssens, D., Demeyer, S., & Vasilescu, B. (2016). Among the Machines: Human-Bot Interaction on Social Q&A Websites. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1272–1279. https://doi.org/10.1145/2851581.2892311

National Centre for Education Statistics. (2016). *Number of public libraries, number of books and serial volumes, and per capita usage of selected library services per year, by state: Fiscal years 2015 and 2016*. https://nces.ed.gov/programs/digest/d17/tables/dt17_701.60.asp

NLTK Documentation. (2023). https://www.nltk.org/

Owens, B. (2023, February 20). *How Nature readers are using ChatGPT*. https://www.nature.com/articles/d41586-023-00500-8

Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011). Early Detection of Potential Experts in Question Answering Communities. In J. A. Konstan, R. Conejo, J. L. Marzo, & N. Oliver (Eds.), *User Modeling, Adaption and Personalization* (Vol. 6787, pp. 231–242). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22362-4_20

Pividori, M., & Greene, C. S. (2023). *A publishing infrastructure for AI-assisted academic authoring* [Preprint]. Scientific Communication and Education. https://doi.org/10.1101/2023.01.21.525030

Posada, J., Weller, N., & Wong, W. H. (2021). We Haven't Gone Paperless Yet: Why the Printing Press Can Help Us Understand Data and AI. *Proceedings of the 2021*

AAAI/ACM Conference on AI, Ethics, and Society, 864–872. https://doi.org/10.1145/3461702.3462604

Semrush.com. (2023a, April 22). Semrush.Com Website Traffic Statistics - Reddit.Com. https://www.semrush.com/website/reddit.com/overview/

Semrush.com. (2023b, April 22). Semrush.Com Website Traffic Statistics - Quora. https://www.semrush.com/website/quora.com/overview/

Semrush.com. (2023c, April 22). Semrush.Com Website Traffic Statistics - Stackoverflow.Com. https://www.semrush.com/website/stackoverflow.com/overview/

Sengupta, S., & Haythornthwaite, C. (2020). *Learning with Comments: An Analysis of Comments and Community on Stack Overflow*. Hawaii International Conference on System Sciences. https://doi.org/10.24251/HICSS.2020.354

sklearn. (2023). *Sklearn TF-IDF documenation*. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Snow, J. (1855). *On the Mode of Communication of Cholera (J. Churchill, London)*.

Stack Exchange. (2023, May 5). *Stack Exchange Data Explorer*. https://data.stackexchange.com/

Tausczik, Y., & Boons, M. (2018). Distributed Knowledge in Crowds: Crowd Performance on Hidden Profile Tasks. *Proceedings of the International AAAI Conference on Web and Social Media*, *12*(1). https://doi.org/10.1609/icwsm.v12i1.15005

Tausczik, Y. R. (2016). Citation and Attribution in Open Science: A Case Study. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1524–1534. https://doi.org/10.1145/2818048.2820070

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series, Vol. 59, No. 236 (Oct., 1950), Oxford University Press*. https://www.jstor.org/stable/2251299?origin=JSTOR-pdf

Van Der Zant, T., Kouw, M., & Schomaker, L. (2013). Generative Artificial Intelligence. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (Vol. 5, pp. 107–120). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31674-6_8

Vasilescu, B., Serebrenik, A., Devanbu, P., & Filkov, V. (2014). How social Q&A sites are changing knowledge sharing in open source software communities. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 342–354. https://doi.org/10.1145/2531602.2531659

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. https://doi.org/10.48550/ARXIV.1706.03762

Wang, K., Dong, B., & Ma, J. (2019). *Towards Computational Assessment of Idea Novelty*.

Watson, R. T. (2004). *Data management: Databases and organizations* (4th ed). J. Wiley.

Wikimedia Foundation. (2023). *Explore Wikipedia's new look*. https://wikimediafoundation.org/wikipedia-desktop/

Wiles, J. (2023, January 26). *Beyond ChatGPT: The Future of Generative AI for Enterprises*. https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises

Wydick, R. C., & Flesch, R. (1980). Lawyers' Writing. *Michigan Law Review, 78*(5), 711. https://doi.org/10.2307/1288066

YouGov. (2023). *What Americans think about ChatGPT and AI-generated text* [PDF]. https://docs.cdn.yougov.com/p3j2eqjz5c/tabs_ChatGPT_20230124.pdf

Yue, T., Au, D., Au, C. C., & Iu, K. Y. (2023). Democratizing Financial Knowledge with ChatGPT by OpenAI: Unleashing the Power of Technology. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4346152

Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., Dam, S. K., Thwal, C. M., Tun, Y. L., Huy, L. L., kim, D., Bae, S.-H., Lee, L.-H., Yang, Y., Shen, H. T., Kweon, I. S., & Hong, C. S. (2023). *A Complete Survey on Generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 All You Need?* https://doi.org/10.48550/ARXIV.2303.11717