

---

# **ANALYSIS AND COMPARISON OF GENERATIVE AI CHATBOT APPLICATIONS**

**MAGHAV AHUJA**

**UNITEC INSTITUTE OF TECHNOLOGY**  
MASTERS OF APPLIED TECHNOLOGIES

**THESIS SUPERVISOR: PROFESSOR HAMID SHARIFZADEH**

**CO-SUPERVISOR: DR. LEI SONG**

*A thesis submitted in partial fulfilment of the requirements for the Master of Applied Technologies degree in the School of Computing, Electrical & Applied Technology.*

## **ABSTRACT**

This research proposes the analysis to systematically evaluate and compare ChatGPT API applications operating within two categories: AI chatbots and project management tools. The primary aim is to analyse the generative AI chatbots to assess these applications based on domain-specific criteria, aligning with the intended functionalities of their AI feature using various text mining techniques.

The evaluation methodology involves tailored strategies for each category. Text mining techniques are applied for chatbots to assess coherence, relevance, and completeness within the generated text, providing a quantitative basis for comparison. Concurrently, the project management tools undergo text mining analysis with some AI integration feature comparison, including the data utilisation analysis, which they offer to evaluate documentation, summarisation, sentiment analysis, and other types of scores.

The chatbots include CopyAI, HuggingFace chatbot, ChatSonic, YouChat and Rytr.me, and project management tools include ClickUp, Notion and Jira as they support the “Write with AI” feature.

This research seeks to evaluate GPT-powered applications by leveraging text-mining techniques. It intends to provide insights into the strengths and weaknesses of these applications, aiding users in making informed decisions based on specific requirements within each category.

## TABLE OF CONTENTS

Abstract.....	1
List of Tables .....	5
List of Figures .....	6
CHAPTER 1: Introduction .....	7
1.1 Generative AI .....	7
1.1.1 History of Generative AI.....	7
1.1.2 What is Generative AI? .....	9
1.1.3 Key Application Areas of Generative AI .....	10
1.2 Research Objectives.....	12
1.3 Thesis Structure .....	13
CHAPTER 2: Literature Review .....	15
2.1 Anthropic’s Claude .....	15
2.2 Meta’s Llama.....	16
2.3 Google’s Bard (Now Gemini).....	17
2.4 OpenAI’s ChatGPT .....	18
2.4.1 APIs and their Applications .....	19
2.4.2 Why ChatGPT Applications? .....	19
2.4.3 Need of Evaluating the Selected Applications .....	22
2.5 Existing Comparisons of AI Chatbots .....	23
2.6 Research Gap .....	28
CHAPTER 3: Proposed Methodology .....	29
3.1 Dataset Collection .....	29
3.1.1 Prompt Engineering .....	29
3.1.2 Response Generation.....	31

---

3.2 Text Mining Techniques .....	32
3.3 Implementation of Text Mining Techniques.....	33
3.3.1 Combining Techniques for Comprehensive Analysis .....	34
3.2.2 Determining Scores.....	35
3.4.1 Data Utilisation .....	41
3.4.2 Integration Features .....	42
CHAPTER 4: Results & Discussions.....	43
4.1 Lexical Diversity Score for AI Chatbots .....	45
4.2 Lexical Diversity Score for Project Management Tools.....	47
4.3 Sentiment Score for AI Chatbots.....	49
4.4 Sentiment Score for Project Management Tools.....	52
4.5 Subjectivity Score for AI Chatbots.....	53
4.6 Subjectivity Score for Project Management Tools.....	55
4.7 Pronoun Ratio for the AI Chatbots.....	57
4.8 Pronoun ratio for Project Management tools .....	59
4.9 Average Word Length for AI Chatbots.....	61
4.10 Average Word Length for Project Management Tools.....	63
4.11 Readability Scores (Flesch - Kincaid Grade Level) for AI Chatbots.....	65
4.12 Readability Scores (Flesch - Kincaid Grade Level) for Project Management Tools.....	67
4.13 Readability Score (Coleman Liau Index) for AI Chatbots .....	69
4.14 Readability Score (Coleman Liau Index) for Project Management Tools .....	71
4.15 Data Utilisation in AI Features of Project Management Tools .....	73
4.16 Integrated Features in Project Management Tools .....	74
4.17 Discussion.....	77
4.17.1 Limitations.....	78
4.17.2 Future Research & Development.....	78

4.17.3 Ethical Considerations.....	79
CHAPTER 5: Conclusion.....	80
References .....	82
Appendix 1 - AI Chatbots .....	94
Appendix 2 - Project Management Tools .....	99

## LIST OF TABLES

Table 4.1: Lexical Diversity Score for AI chatbots .....	45
Table 4.2: Lexical Diversity Scores for Project Management Tools .....	47
Table 4.3: Sentiment Scores for AI Chatbots .....	49
Table 4.4: Sentiment Scores for Project Management Tools .....	52
Table 4.5: Subjectivity Scores for AI Chatbots .....	53
Table 4.6: Subjectivity Scores for Project Management Tools .....	55
Table 4.7: Pronoun Ratio for AI Chatbots .....	57
Table 4.8: Pronoun Ratio for Project Management Tools .....	59
Table 4.9: AWL for AI Chatbots.....	61
Table 4.10: AWL for Project Management Tools.....	63
Table 4.11: Flesch Kincaid Grade Level for AI Chatbots.....	65
Table 4.12: Flesch Kincaid Grade Level for Project Management Tools.....	67
Table 4.13: Coleman Liau Index Score for AI Chatbots.....	69
Table 4.14: Coleman Liau Index for Project Management Tools.....	71
Table 4.15: Project Management Tool’s Data Utilisation Factors from the Already Created Projects .....	74
Table 4.16: Features Integrated in AI-enhanced Project Management Tools.....	75

## LIST OF FIGURES

Figure 2.1: List of AI Chatbots considered in the thesis.....	20
Figure 2.2: List of AI-enhanced project management tools considered in the thesis .....	21
Figure 4.1: Box plot showing lexical diversity scores for AI Chatbots category.....	45
Figure 4.2: Box Plot showing Lexical Diversity Scores for the Project Management Tools .....	47
Figure 4.3: Box Plot showing Sentiment Scores for AI Chatbots .....	49
Figure 4.4: Box Plot showing Sentiment Scores for Project Management Tools .....	52
Figure 4.5: Box Plot showing Subjectivity Scores for Project Management Tools .....	54
Figure 4.6: Box Plot showing Subjectivity Scores for Project Management Tools .....	56
Figure 4.7: Box Plot showing Pronoun Ratio for AI Chatbots .....	58
Figure 4.8: Box Plot showing Pronoun Ratio for Project Management Tools .....	60
Figure 4.9: Box Plot showing AWL for AI Chatbots.....	62
Figure 4.10: Box Plot showing AWL for Project Management Tools .....	64
Figure 4.11: Box Plot showing Flesch Kincaid Grade Level for AI Chatbots.....	66
Figure 4.12: Box Plot showing Flesch Kincaid Grade Level for Project Management Tools.....	68
Figure 4.13: Box Plot showing Coleman Liau Index results for readability on AI Chatbots.....	70
Figure 4.14: Box Plot showing Coleman Liau Score for Project Management Tools.....	72

## **CHAPTER 1: INTRODUCTION**

Generative AI has seen advancements leading to the launch of various chatbots by companies. These developments have prompted comparisons online, evaluating factors like conversational quality and user experience. This research explores these developments and comparisons, examining the evolution of chatbots and their impact. It is motivated by the increasing importance of AI-driven interactions in applications, aiming to determine which chatbots perform reliably. This chapter provides the background of generative AI, discusses research questions and objectives, and outlines the thesis structure.

The combination of artificial intelligence (AI) and natural language processing (NLP) technologies as generative artificial intelligence (AI) has significantly transformed several application sectors in recent years [1]. Of these, the incorporation of ChatGPT API into various apps has generated significant attention, providing improved functionality and user engagement. This study seeks to examine and analyse apps in two distinct categories - AI Chatbots and project management tools—that use ChatGPT API by text mining techniques and have an active “Write with AI” feature.

The primary objective is to use text-mining [2] To systematically analyse and compare these applications, focusing on their efficacy in fulfilling pre-established criteria tailored to each category's unique requirements. Text mining techniques can be used to examine the responses provided by the ChatGPT API within the chatbot domain. This derives insights by evaluating predefined score criteria, such as sentiment scores, subjectivity scores, average word lengths, integration features in the project management tools, etc.

This research aims to explain the strengths and limits of apps in their respective categories via a complete comparative analysis, which can be determined using the possible applicable text-mining techniques. The acquired insights assist the end-users in making well-informed choices and provide vital guidance for developers aiming to enhance their application's functions and user experiences.

### **1.1 GENERATIVE AI**

#### **1.1.1 HISTORY OF GENERATIVE AI**

Generative models have a lengthy lineage in the field of artificial intelligence, originating in the 1950s with the creation of Hidden Markov Models (HMMs) [3] and Gaussian Mixture Models (GMMs) [4]. These



models produced sequential data, such as speech and time series. Yet, it was only with the emergence of deep learning that generative models saw substantial performance improvements. There is often little overlap across diverse domains in the nascent stages of deep generative models. In the field of natural language processing (NLP), a conventional approach for generating sentences involves learning the distribution of words by N-gram language modelling [5] and then identifying the optimal sequence. Nevertheless, this approach is not proficient at accommodating lengthy phrases. Recurrent neural networks (RNNs) [6] were subsequently designed to address this issue, enabling the modelling of longer dependencies in language problems. Subsequently, Long Short-Term Memory (LSTM) [6] and Gated Recurrent Unit (GRU) [7] were developed, using the gating mechanism to regulate memory during training. These techniques can process around 200 tokens in a sample [8], which is a substantial improvement compared to N-gram language models.

Before the emergence of deep learning-based approaches, standard picture production algorithms in computer vision (CV) included techniques such as texture synthesis [9] and texture mapping [10]. These algorithms relied on manually crafted characteristics and had limitations in producing intricate and varied visuals. The year 2014 saw the introduction of Generative Adversarial Networks (GANs) [11], a notable breakthrough in this field owing to its remarkable outcomes across diverse applications. Variational Autoencoders (VAEs) [12] and other techniques, such as diffusion generative models [13], have been created to enhance the level of precision in the picture creation process and to enable the production of high-quality images. The development of generative models in different fields has taken many routes, but ultimately, the point of convergence has been the emergence of the transformer architecture [14]. The Transformer model, first proposed by Vaswani et al. in 2017 for Natural Language Processing (NLP) tasks, has now been adapted for Computer Vision (CV) applications and has emerged as the primary architecture for several generative models across different domains [15] [16]. In natural language processing (NLP), well-known and influential language models, such as BERT and GPT, use the transformer architecture as their main component. This choice of architecture provides certain benefits compared to earlier components, namely LSTM and GRU. In computer vision, the Vision Transformer (ViT) [17] and Swin Transformer [18] build upon this idea by integrating the transformer architecture with visual elements, enabling its application to image-based tasks. In addition to enhancing certain modalities, integrating transformers at this junction facilitated combining models from other domains to perform multimodal activities. An instance of a multimodal model is CLIP [19]. CLIP is a collaborative vision-language model that merges the transformer architecture with visual elements, enabling it to be trained on a vast volume

of text and picture data. Due to its integration of visual and linguistic information in pre-training, it may also serve as an image encoder in multimodal prompting for generating purposes. Overall, transformer-based models have brought about a revolutionary change in AI generation and made large-scale training feasible. Recently, researchers have also started using novel strategies derived from these models. In Natural Language Processing (NLP), some individuals choose a few-shot prompting technique [20] rather than fine-tuning. This approach involves including a small number of samples from the dataset in the prompt to enhance the model's comprehension of the task's demands. Researchers often integrate modality-specific models with self-supervised contrastive learning goals in visual language to enhance the robustness of representations. As the significance of AIGC grows, an array of technologies is progressively implemented, improving the vitality of this field.

### **1.1.2 WHAT IS GENERATIVE AI?**

Generative artificial intelligence (AI) has emerged as one of the most promising areas of AI research and development. Generative models learn to synthesise new data instances similar to their training data, going beyond just classification and prediction [11]. The capacity to generate realistic text, images, audio, video, and 3D content from scratch has accelerated progress in this field.

Several techniques at the forefront of generative AI include generative adversarial networks (GANs) [11], variational autoencoders (VAEs) [12], autoregressive models, diffusion models [21], normalising flows, and energy-based models. Combined with large datasets and models, these approaches have achieved remarkable results in text, image, speech, music, and drug design generation.

Large language models (LLMs) are exciting - generative models scaled up to billions of parameters focused on text generation [22]. LLMs like GPT-3, Llama, Gemini and Claude are trained on vast text corpora to predict upcoming tokens given the previous context. Their foundation is the transformer architecture, which effectively models long-range dependencies in text. Autoregressive decoding allows recursive sampling from LLMs to generate coherent text [23].

LLMs have shown impressive fluency, semantic coherence, knowledge retention, and few-shot adaptation abilities at scale. Scaling model size, computing, and datasets yield further breakthroughs. Active research directions for LLMs include controllable generation, personalisation to different writing styles,

maintaining topic coherence over long horizons, minimising harmful bias in outputs, and multimodal modelling.

For images, GANs represent a seminal breakthrough in generative modelling [11]. GANs involve adversarial training of a generator network against a discriminator. This competitive framework enables realistic image synthesis across architectures. Critical innovations in GAN research include conditional GANs [24], StyleGANs [25], CycleGANs [26], and [27].

GANs have delivered unprecedented advances in high-resolution photorealistic image generation and editing. Applications span art, entertainment, content creation, facial editing, data augmentation, domain adaptation, image inpainting, and super-resolution. Recent research has focused on training stability, interpretability, controllability, and personalisation.

VAEs learn compressed latent data representations for generative sampling using reconstruction and regularisation losses [12]. The encoder network compresses data into the latent space while the decoder reconstructs from latent points. VAEs have shown promise in image, video, speech, and drug generation by manipulating the latent space. Innovations in VAE research include conditional VAEs, hierarchical VAEs, and VAEs for sequential data.

Diffusion models are an increasingly popular technique for high-fidelity image generation [21]. They are trained to denoise data by reversing additive noise corruption over repeated steps. DALL-E 2 and Imagen demonstrate cutting-edge text-to-image capabilities using diffusion models. Current research focuses on larger models, video generation, sample efficiency, controllable generation, and multimodality.

### 1.1.3 KEY APPLICATION AREAS OF GENERATIVE AI

Generative AI has revolutionised various fields by creating new content, assisting in complex tasks, and providing innovative solutions. This section explores the primary areas where generative AI is making significant impacts, showcasing its versatility and potential across different domains.

**Content Creation:** Generative adversarial networks (GANs) and diffusion models have shown promising results in assisting and amplifying human creativity across design, gaming, architecture, and entertainment [28]. For example, GANs can generate innovative fashion designs, game assets, or

architectural renderings. Recent advancements in GANs, such as StyleGAN3, have further improved the quality and controllability of generated content.

**Text Generation:** GPT-3 and other Large Language Models (LLMs) demonstrate strong performance on text completion, translation, dialogue agents, and summarisation. LLMs can produce human-like text for customer service bots or summarise lengthy reports into concise overviews [29]. The emergence of few-shot learning capabilities in these models has expanded their applicability to various domains with minimal task-specific training [22].

**Image Generation:** Generative models for images, including GANs, Variational Autoencoders (VAEs), and diffusion models, have made significant progress in unconditional and class-conditional image generation. Critical applications include creating profile pictures based on descriptions, product visualisation, and data augmentation [30]. Recent breakthroughs in text-to-image generation, such as DALL-E 2 and Stable Diffusion, have pushed the boundaries of what's possible in AI-generated imagery [31].

**Voice Synthesis:** Models like WaveNet and MelNet can synthesise human-like speech from text by learning robust acoustic models from data. Uses include personal assistants, audiobook narration, and accessibility tools for vision-impaired users [32].

**Drug Discovery:** Generative models can learn meaningful latent representations of molecules, guiding molecular graph generation towards desired pharmacological properties. This assists researchers in designing candidate drug compounds [33]. The integration of reinforcement learning with generative models has shown promising results in optimising generated molecules for specific properties.

**Data Augmentation:** Generative models provide a powerful approach for synthetically expanding datasets during training. By generating additional varied, realistic examples, overfitting can be reduced [11]. Advanced techniques like MixUp and CutMix have shown effectiveness in improving model robustness and generalisation [34].

**Recommendation Systems:** Collaborative filtering methods based on generative models of user preferences have been effective for personalised suggestions. Applications include e-commerce, digital content platforms, and social networks [35]. The integration of graph neural networks with generative models has enhanced the ability to capture complex user-item interactions [36].

**Video Generation:** Generative models can forecast future video frames by modelling scene dynamics and motion trajectories. This assists computer vision research and graphics applications like video codecs.

Recent advances in transformer-based architectures have shown promising results in long-term video prediction and generation tasks [37].

LLMs are AI systems that work with language and form the text-generating part of generative AI. They are trained on many articles, Wikipedia entries, books, internet-based resources, and other inputs to produce human-like responses to natural language queries. LLMs are machine-learning neural networks trained through data input/output sets, frequently using self-supervised or semi-supervised learning methodology [38]. The scale and capabilities of LLMs have been rapidly expanding, with models like GPT- demonstrating impressive performance across a wide range of language tasks. Recent research has focused on improving the efficiency and environmental impact of these large models through techniques like sparse attention and model distillation [39].

OpenAI launched ChatGPT for the public in November 2022. It was the first time any LLM's API was used as a chatbot. In competition, Anthropic, Google and Meta launched their LLMs and developed their chatbot based on their models. When GPT-3.5 Turbo was made public, Anthropic launched an improved version of Claude 2 in early 2023. Meta and Google launched their LLaMA and Bard in February 2023 as their chatbots. However, as of June 2024, all these chatbots were updated to their much more powerful and revised versions, as Bard is now called Gemini, and Claude 2 is now Claude 3 with variants like Opus, Sonnet and Haiku. Meta also launched its LLaMA - 3 version this year. At the time of writing this thesis in May 2024, OpenAI had recently launched another model called GPT-4o [40]. My proposal was approved in mid-2023, and now it is mid-2024. My scope for this study remains limited to the older versions. This hasn't impacted the applications of ChatGPT, which I have used in my comparison framework. AI Chatbots like You.com and Huggingface have added a functionality that allows users to change among different API models, but the use of OpenAI API for ChatGPT is still active.

## **1.2 RESEARCH OBJECTIVES**

This research evaluates the text generation capabilities of generative AI chatbots and AI-enhanced project management tools. By employing quantitative text analysis metrics such as lexical diversity, sentiment, subjectivity, readability, and pronoun usage, we will compare the performance of selected AI chatbots when responding to standardised prompts. This comparative analysis will identify each system's strengths, weaknesses, and unique characteristics.

Additionally, this study will assess how effectively AI features within project management tools utilise and integrate project-specific data to enhance their AI-generated responses and functionalities. I've addressed a few research questions mentioned below.

### *1.2.1 RESEARCH QUESTIONS*

To achieve the objectives outlined above, the following research questions have been formulated:

**RQ1:** How can AI chatbots responses be analysed using text mining techniques?

**RQ2:** How can AI-enhanced project management tool responses be analysed using text mining techniques?

**RQ3:** How well are these AI-enhanced project management tools integrated with their AI feature?

Through this research, we seek to contribute to a deeper understanding of AI language generation technology's current state and practical applications. The findings will provide valuable insights for developers, users, and decision-makers in selecting and improving AI-powered tools, particularly in general-purpose chatbots and project management systems.

## **1.3 THESIS STRUCTURE**

The thesis is organised into six main chapters, each serving a distinct purpose in exploring the applications and comparisons of generative AI chatbots and project management tools.

The abstract concisely summarises the entire thesis, highlighting the research objectives, methodology, key findings, and conclusions. The introduction chapter introduces the concept of generative AI, tracing its history and defining what it entails. It discusses the key application areas of generative AI and sets out the research objectives. The chapter concludes with an overview of the thesis structure, guiding the reader on what to expect in the subsequent chapters.

The literature review surveys existing knowledge on various generative AI models such as Anthropic's Claude, Meta's Llama, Google's Bard (now Gemini), and OpenAI's ChatGPT. It also examines the applications of AI chatbots like You Chat, ChatSonic, Copy.ai, Huggingface, and Rytr.me and project management tools like Notion, ClickUp, and Jira. This chapter identifies gaps in the existing research that this thesis aims to address. The proposed methodology chapter outlines the methodology used for the research. It details the dataset collection process, including prompt engineering and response generation. The chapter also describes the text mining techniques employed, their implementation, and how they

were combined for a comprehensive analysis. It includes the metrics used to analyse project management tools, focusing on data utilisation and integration features.

The results and discussions chapter presents the findings of the research. It includes detailed analyses of lexical diversity, sentiment scores, subjectivity scores, pronoun ratios, average word length, and readability scores for AI chatbots and project management tools. The chapter also examines data utilisation and the integrated features of project management tools, providing a thorough discussion of the results.

The conclusion chapter summarises the key findings of the research, highlighting the performance differences among the evaluated AI chatbots and project management tools. It reflects on the implications of these findings and suggests areas for future research. The references chapter lists all the sources cited throughout the thesis, providing full bibliographic details to ensure proper attribution and allow readers to locate the original sources.

The Appendix 1 provides additional information and data on the AI chatbots evaluated in the study. Similar to Appendix 1, Appendix 2 offers supplementary details on the project management tools included in the research.

This structure ensures a logical flow, from introducing the topic and reviewing existing literature to detailing the methodology, presenting the results, and concluding the study. Each chapter builds on the previous one, comprehensively examining generative AI chatbots and project management tools.

## CHAPTER 2: LITERATURE REVIEW

This chapter provides a comprehensive overview of large language models (LLMs) and their applications, with a particular focus on chatbots and project management tools. It begins by examining key players in the generative AI landscape, including Anthropic's Claude, Meta's LLaMA, Google's Gemini, and OpenAI's ChatGPT. Each model's development, key features, and recent advancements are explored. The chapter then delves into the application of LLMs, with a specific focus on ChatGPT-based tools due to their widespread adoption. These applications are categorised into AI chatbots and project management tools, enabling a detailed comparison of their implementation. Building on existing comparisons of AI chatbots, the chapter highlights research gaps and sets the stage for this thesis' unique contribution to the field.

### 2.1 ANTHROPIC'S CLAUDE

Anthropic, a San Francisco-based AI startup founded in 2021, has been focused on developing safe and helpful artificial intelligence. In early 2022, Anthropic unveiled its first AI conversational agent, constitutional AI, which was later improved and released as Claude AI.

As of 2024, Claude has evolved into the Claude 3 model family, which includes Claude 3 Haiku, Claude 3 Opus, and Claude 3.5 Sonnet. These models offer varying levels of intelligence and specialisation, with Claude 3.5 Sonnet being the most advanced in terms of general intelligence. Despite strategic partnerships, notably with Amazon, Claude's adoption in the developer community has been slower compared to some competitors.

In early 2022, Anthropic unveiled its first AI conversational agent, constitutional AI. This system was trained to have harmless, honest, and helpful dialogues through Constitutional AI, Anthropic's proprietary technique based on self-supervision. While effective, constitutional AI had some limitations in its conversational abilities. So, later that year, Anthropic iterated on their research and released Claude AI [41]. Claude built upon the everyday foundations of constitutional AI but was trained with increased data using Constitutional AI v2.0. This improved Claude's natural language processing and ability to have more coherent, human-like dialogues.

Despite the availability of Anthropic's Claude chatbots API with transparent pricing, the model has yet to gain widespread recognition as a flagship application of Generative AI within the developer community. While LabLab.ai and similar platforms boast listings of over 100 applications purportedly leveraging Claude's API, a closer examination reveals a scarcity of active implementations.



In LLM, Anthropic's Claude has garnered substantial validation through strategic partnerships, notably from Amazon. These factors substantiate Claude's status as one of the leading models in the domain. However, a noticeable trend is a comparative lack of fervour and acknowledgment, especially when juxtaposed with the fervent enthusiasm surrounding the OpenAI API.

This analysis is pivotal when contextualising the landscape of API applications stemming from Language Learning Models like ChatGPT, Meta, Claude, and Bard. It provides a nuanced understanding of the varying degrees of adoption, visibility, and industry enthusiasm for different AI models, illuminating potential factors contributing to their differential trajectories in adoption and acclaim.

## **2.2 META'S LLAMA**

Meta, formerly known as Facebook, has directed significant resources toward advancing Generative Artificial Intelligence research, specifically focusing on developing robust Natural Language Processing (NLP) systems using Large Language Models (LLMs). This evolution has been showcased through milestone advancements in their chatbot development journey.

In 2020, Meta introduced Blender, a chatbot built upon the foundation of the open-source GPT-2 model. While pioneering in its inception, Blender revealed limitations in its conversational abilities. Building upon this groundwork, Meta unveiled Opt in 2021, a substantially larger LLM chatbot trained on Meta's extensive dataset of public social media conversations. Opt demonstrated notable improvements in conversational aptitude compared to its predecessor, Blender.

The subsequent leap in Meta's chatbot evolution arrived in 2022 with the advent of Galactica. This advanced chatbot was developed using Reinforcement Learning from Human Feedback, a rewards-based learning approach that significantly augmented Opt's conversational capabilities, fostering more human-like dialogues.

Continuing this innovation trajectory, Meta revealed LLaMA in late 2022 [42], a powerful LLM equipped with an impressive 137 billion parameters. LLaMA showcased exceptional language prowess and knowledge capabilities, marking a significant milestone in Meta's generative AI research. Building on this, in 2023, Meta launched LLaMA-2, a monumental LLM with a staggering 200 billion parameters. This

cutting-edge model, fine-tuned using Meta's conversational data, represents the pinnacle of their current generative AI research endeavours.

In 2024, Meta further expanded its LLM portfolio with the release of Llama 3. This latest iteration builds upon the strengths of its predecessors while introducing new features and capabilities. Llama 3 continues Meta's commitment to open-source AI development, allowing researchers and developers to access and build upon the model. This open-source approach has fostered a vibrant ecosystem of applications and improvements, distinguishing Meta's strategy from some of its competitors.

However, while Meta offers access to its API exclusively through their models, the pricing structure hinges on the dataset size on which the model has been trained. Meta has not released any standalone official chatbots for direct comparison with competitors. Despite partnerships with DELL Technologies for Generative AI in their laptops and IBM for Watsonx, official collaborations for other applications remain undisclosed. Now, Meta has integrated the LLaMA feature into the Facebook and WhatsApp bot to access the functionalities of the public.

This lack of official chatbots or partnerships beyond specific sectors limits the direct assessment and comparison of applications leveraging Meta's latest Llama models. As a result, analysing the practical applications and effectiveness of Meta's latest generative AI models remains challenging without publicly accessible and diversified implementations. However, the open-source nature of Llama 3 may lead to increased adoption and a broader range of applications in the near future.

### **2.3 GOOGLE'S BARD (NOW GEMINI)**

In 2018, Google introduced BERT, which pioneered the pre-training and fine-tuning approach for NLP. This allowed models to learn general language representations before specialising in downstream tasks [43]. Building on BERT, in 2020, Google released T5 and T5v1.1, which were pre-trained on massive text corpora using self-supervised learning. This significantly improved the models' natural language understanding and generation skills [44].

Later, in 2020, Google unveiled the Pathways Language Model or PaLM with 540 billion parameters. Trained on data centres' worth of text, PaLM demonstrated powerful conversational ability and knowledge. Most recently, in 2023, Google announced Bard, an experimental casual AI service powered

by PaLM [45]. While an initial demo of Bard provided an incorrect answer, Bard represents Google's continued progress in developing LLMs through self-supervised learning.

As of June 2024, Google has fully transitioned from Bard to Gemini. Launched in late 2023, Gemini offers API access for developers and is available in multiple versions, including a free public version and a paid Pro version that competes with GPT-4 and other advanced models from Anthropic and Meta. This shift represents Google's increased confidence in its LLM capabilities and its commitment to competing in the rapidly evolving AI landscape.

## **2.4 OPENAI'S CHATGPT**

OpenAI's Generative Pre-Trained Transformer models have rapidly advanced the capabilities of large language models over a short span of years. In 2020, the release of GPT-3 represented a pioneering breakthrough, leveraging its 175 billion parameters to achieve impressive performance on a wide range of natural language processing tasks requiring generation, comprehension, and reasoning. However, GPT-3's massive model size limited feasibility for real-world deployment [22].

Building on GPT-3, OpenAI unveiled GPT 3.5 Turbo in 2021 using model compression techniques to optimise GPT-3's capabilities to fit into as little as 12GB. This improved the ability to deploy such a complex generative model. In 2023, OpenAI revealed GPT-4, which learns from GPT-3 and GPT 3.5 Turbo to scale up to about 200 billion parameters and establish new state-of-the-art standards in common sense reasoning, factual knowledge, and memory capabilities.

The rapid evolution from GPT-3 to GPT-3.5 Turbo, and now to GPT-4 and its variants like GPT-4 Turbo, demonstrates OpenAI's pioneering work in advancing scalable pre-trained language models. As of 2024, OpenAI continues to refine and expand its offerings, with ChatGPT remaining one of the most widely used and recognised AI chatbots globally.

### **2.4.1 APIs AND THEIR APPLICATIONS**

OpenAI has been working on its Generative Pre-Trained Transformer (GPT) series since 2018. GPT-1 was the first model in the series. It introduced the concept of pre-training transformers on a large corpus of text data and demonstrated the generation of coherent and contextually relevant text [46]. However, it had limitations in handling long-range dependencies and sometimes generated outputs that lacked factual accuracy [46]. GPT-2 marked a significant advancement over GPT-1 in scale and performance. It was designed with 1.5 billion parameters, making it one of the most considerable language models at its release. GPT-3 is the most potent model in the series, featuring a staggering 175 billion parameters [22]. It demonstrated remarkable language understanding, context retention, and the ability to perform a wide range of natural language processing tasks without task-specific training data [22]. GPT-3 has been widely used in various applications, including chatbots, code generation, language translation, and creative writing [22]. GPT-3.5 Turbo is an advanced iteration of GPT-3, designed for improved performance and efficiency. It retains the capabilities of GPT-3 but is optimised for various applications, allowing developers to harness its powerful language generation abilities more effectively.

In late 2023, OpenAI introduced GPT-4 Turbo, an improved version of GPT-4. This model features an expanded knowledge cutoff (up to April 2023), increased token limits for longer conversations, and improved performance across various tasks. It also offers more steerable behaviour, allowing developers to precisely customise the model's responses.

The latest iteration in the series, as of 2024, is GPT-4 Omni. This version builds upon the strengths of its predecessors while introducing novel features and capabilities. GPT-4 Omni is expected to offer even more advanced natural language understanding and generation, potentially incorporating improved multimodal abilities and enhanced real-time information processing. While specific details about GPT-4 Omni are still emerging, it is anticipated to push the boundaries of what's possible with language models even further, empowering developers to create increasingly sophisticated and capable AI applications.

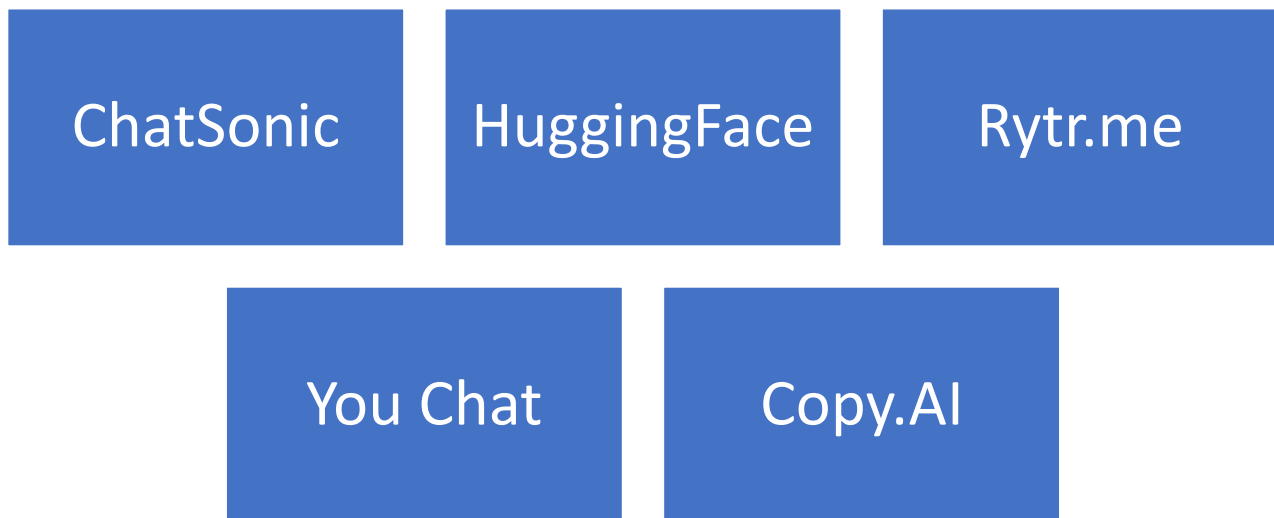
These continuous advancements in the GPT series reflect OpenAI's commitment to advancing the field of artificial intelligence and natural language processing, providing increasingly powerful tools for a wide range of applications and research endeavours.

### **2.4.2 WHY CHATGPT APPLICATIONS?**

OpenAI released access to its GPT APIs much earlier than other chatbot companies. Many researchers have concluded that ChatGPT's performance is better than that of Bard and Claude, Llama. According to

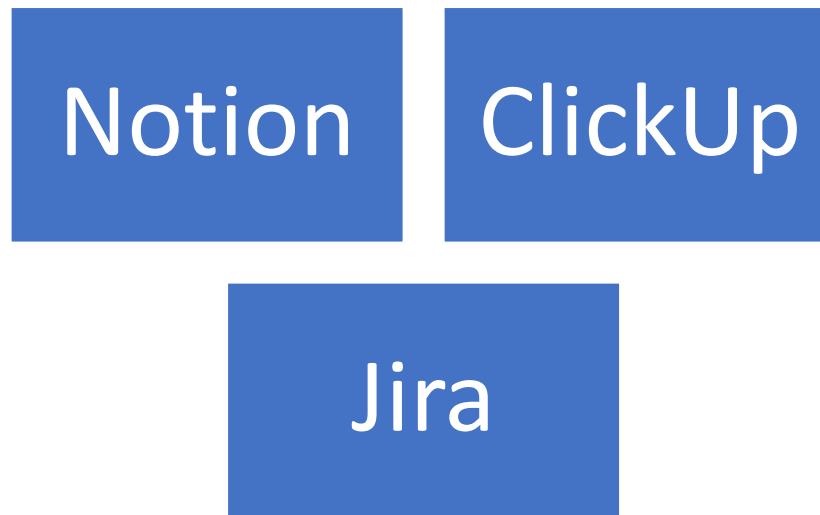
[47], ChatGPT gained much higher hype than any of the technical launches before. This is also a reason why the development of the applications only uses OpenAI's GPT API and not others.

Specifically, as chatbots were extensions to ChatGPT, many chatbots were developed. This leads to the importance of categorising all these applications and comparing them. Dividing the applications into two categories, AI Chatbots and project management tools, can provide a better way to get insights. Table 1 lists a few names of the applications that use ChatGPT API for integrating AI features in their applications. The selection of the listed depends on the popularity, their available premium versions, accessibility, userbase, similar features, and integration. All of them offered similar scope because considering them for comparison may be best to provide the expected results.



*Figure 2.1: List of AI Chatbots considered in the thesis*

You Chat is an AI chatbot that responds to multiple LLM models from its interface. It also uses the internet when generating responses [48]. ChatSonic uses custom-tuned GPT 3.5 Turbo and the internet to generate responses [49]. Copy.AI uses fine-tuned GPT3.5 Turbo and the internet and generates responses in its interface [50]. Rytr.me is an OpenAI-based chatbot for writing long text (fine-tuned for generating long texts) [51]. HuggingFace offers free access to the open-source LLM models in its chat interface [52].



*Figure 2.2: List of AI-enhanced project management tools considered in the thesis*

Notion is a note-making web application for project management [53]. ClickUp is a project management tool that offers AI integration when writing anything on its platform. Jira, similar to ClickUp [54] It is also a project management tool that offers writing up anything with a feature called “Write with AI.” [55] which helps users to use AI wherever they want to write.

The reason for sticking into two categories - Other categories like trip planning, recipe making, and coding extensions for VS Code, which worked using the ChatGPT API, were also considered. But unfortunately, trip planning and recipe maker websites weren’t reliable. There were just a few GitHub repositories [56] [57] in which the developers considered the development of recipe generators and trip planners using ChatGPT API as personal projects. While working with websites like Trip.com, MakeMyTrip.com, and Booking.com, features that generated trips using ChatGPT API were integrated but only in the beta mode. Working with project management tools like Asana and Monday.com (one of the most popular project management tools), I found that Monday.com still has its AI feature under the beta testing mode as of the end of 2023 (when the proposal was approved). While working on Asana, it only offered its AI feature for summarising tasks. There are no prompts and no random generation of text, but only specifics. So, unfortunately, I had to ignore Asana and Monday.com and stay restricted to 3 important project management tools, perhaps the most popular ones on the internet. When working with the coding extensions, I began with prompt engineering. Still, the moment GitHub Copilot, based on OpenAI’s API

[58], was available for users, it ended the use of all of the extensions that were just being used as shortcuts to the chatbots within the VSCode. Soon, they became obsolete and no longer required to be used.

### **2.4.3 NEED OF EVALUATING THE SELECTED APPLICATIONS**

The promotion and widespread use of the ChatGPT API have catalysed the development of various AI chatbots and the integration of AI features into popular project management tools. This rapid advancement highlights several critical areas that warrant thorough evaluation and analysis.

#### **UNDERSTANDING INTEGRATION MECHANISMS**

**Integration Techniques:** Understanding the methodologies and techniques used to integrate AI into project management tools is crucial. This includes analysing the underlying algorithms, data handling processes, and the overall architecture that supports AI functionalities.

**User Interface and Experience:** How AI features are embedded within the user interfaces of project management tools significantly impacts usability and efficiency. Evaluating the design, accessibility, and intuitiveness of these integrations provides insights into their practical utility.

#### **ASSESSING THE DEVELOPMENT OF AI CHATBOTS**

**Innovation and Differentiation:** Various companies have developed AI chatbots, often marketing them as superior alternatives to ChatGPT. An in-depth evaluation of these claims is necessary to understand the unique features, innovations, and technological advancements that set these chatbots apart.

**Performance Metrics:** Analysing the performance of different AI chatbots in terms of response quality, accuracy, and user engagement helps in determining their effectiveness and reliability.

#### **MARKET STRATEGIES AND ECONOMIC IMPLICATIONS**

**Marketing and Positioning:** Companies have employed diverse marketing strategies to promote their AI chatbots. Evaluating these strategies sheds light on how market perceptions are shaped and the impact of marketing on user adoption.

**Pricing Models:** The introduction of AI features has led to changes in pricing models for project management tools. These tools often charge an additional fee for AI capabilities, which raises questions about these features' value proposition and cost-effectiveness.

#### **USER EXPERIENCE AND SATISFACTION**

**Impact on Workflow:** The integration of AI features is expected to streamline workflows and enhance productivity. Evaluating how these features impact day-to-day operations provides a comprehensive understanding of their practical benefits.

**User Feedback:** Collecting and analysing user feedback regarding AI integrations and AI chatbots offers valuable insights into user satisfaction, preferences, and areas for improvement.

#### COMPARATIVE ANALYSIS

**Feature Comparison:** Conducting a comparative analysis of AI features across different project management tools highlights their strengths and weaknesses, helping users make informed decisions based on their specific needs.

**Effectiveness of AI Chatbots:** Comparing the performance and features of various AI chatbots provides a clear picture of their capabilities and helps identify the best options available in the market.

#### IMPLICATIONS FOR FUTURE DEVELOPMENT

**Technological Advancements:** Understanding the current state of AI integration and chatbot development guides future innovations and technological advancements in this field.

**Best Practices:** Identifying best practices in AI integration and chatbot development benefits both software developers and end-users, leading to more effective and user-friendly solutions.

By conducting a thorough evaluation of AI integration in project management tools and the development of AI chatbots, this research aims to provide valuable insights that benefit both academic research and practical applications. This evaluation not only justifies the additional costs associated with AI features but also contributes to the broader understanding of how AI technologies can enhance productivity, efficiency, and user satisfaction in various domains.

## 2.5 EXISTING COMPARISONS OF AI CHATBOTS

Large Language Models have already been compared based on response-generating speed, accuracy, Chi-Squared Tests, and response length. Specifically, on comparing ChatGPT with Bard (now Gemini), The responses from OpenAI's ChatGPT and Google's Gemini were compared based on response generating speed, accuracy, Chi-Squared Tests, and response length according to [59]. ChatGPT displayed superior radiology knowledge compared to Bard. While both chatbots display reasonable radiology knowledge, they should be used with a conscious understanding of their limitations and fallibility. Both chatbots



provided incorrect or illogical answer explanations and did not always address the educational content of the question [59]. In [60], the capabilities of LLMs, including Gemini Ultra 1.0, Claude -3 Opus, and GPT-4, have been compared.

A comparative analysis was conducted in [61] to evaluate ChatGPT-3.5, ChatGPT-4.0, and Google Bard (now Gemini). The study followed a methodology that included thirty-one commonly asked questions about myopia care. These questions were categorised into six domains: pathogenesis, risk factors, clinical presentation, diagnosis, treatment and prevention, and prognosis. The LLMs were asked [about] each topic, and their replies were independently evaluated by three consultant-level paediatric ophthalmologists using a three-point accuracy scale (bad, borderline, good). The final ranking for each answer was determined using a majority consensus technique. The replies that received a rating of 'Good' were then assessed for comprehensiveness using a five-point scale. In contrast, replies classified as 'poor' were given additional prompts for self-correction and then re-evaluated for correctness. Similarly, other benchmarks like RACE for reading comprehension [62], FEVER for fact-checking [63], math [64], coding [65], computer science tasks [66], translation [67] [20], reasoning [68], or bias [69].

Benchmarks like BIG-bench [70] and [71] incorporate diverse tasks [72] conducted an extensive technical evaluation of ChatGPT using 23 data sets. Guo et al. [73] collected tens of thousands of comparison responses (the HC3 dataset) from human experts and ChatGPT, with questions ranging from open-domain, financial, medical, legal, and psychological areas. They assessed ChatGPT's performance on multiple-choice datasets. [74] provides a commonsense reasoning benchmark. Some works have examined the toxicity and ethics of LLMs (e.g., [75] [76]). [74] offers a detailed evaluation of the capabilities of both the GPT-3 and GPT-3.5 series models. Some works have extensively studied the capacities of a specific model (GPT-4 [77]). Some studies have assessed LLMs qualitatively, such as [78] [79] [80]]. These benchmarks are subjective and informal, and they may not satisfy the rigorous standards of scientific evaluation. We follow their approach but try to make it quantitative and thorough. To our knowledge, no benchmark has yet compared the modern LLMs quantitatively and exhaustively by carefully examining their responses. Instead of using multiple-choice questions, conducted in almost all NLP benchmarks, we analyse open-form answers of the LLMs in comparison to human answers, which allows for a more granular examination of the systems' performance and can uncover cases where the system chooses the right solution for the wrong reason, and vice versa.

In [81], the author compares ChatGPT 3.5 Turbo with GPT-4, Bard, and Bing Chat based on 15 questions and grades them according to the US grading system. The responses from the chatbots were treated the same as if human students were given mathematical problems. All literature reviews talk about how the chatbots developed by Microsoft, Google, Meta, or Amazon can be compared among themselves. A comprehensive comparison between Google's Gemini and OpenAI's ChatGPT has been conducted in a recent study titled "Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation" [82]. This research analyses these two leading generative AI models, examining their applications, performance metrics, underlying architectures, capabilities, and implementation strategies. The study highlights the strengths and weaknesses of each model across various domains, including natural language processing, code generation, and multimodal tasks. It offers insights into how Gemini and ChatGPT compare in terms of response accuracy, processing speed, and the ability to handle complex queries. Additionally, the research explores the architectural differences between the two models, shedding light on how these variations impact their respective performances and use cases. This comparative analysis contributes valuable insights to the ongoing discourse on the state-of-the-art large language models and their practical applications in various fields.

A recent study published in Graefe's Archive for Clinical and Experimental Ophthalmology compares the performance of two prominent large language models, ChatGPT and Google Gemini, in the context of glaucoma surgical cases. The research conducted by [83] evaluates these AI models' ability to provide assistance in complex medical scenarios. The study utilised a set of 31 glaucoma surgical case descriptions, assessing the models' responses for accuracy, completeness, and clinical relevance. The researchers found that both ChatGPT and Google Gemini demonstrated impressive capabilities in understanding and responding to complex medical queries. However, there were notable differences in their performance across various aspects of glaucoma surgery. This comparative analysis not only highlights the potential of AI in assisting medical professionals but also underscores the importance of continued evaluation and refinement of these models for specialised medical applications. The study provides valuable insights into the current state and future potential of AI-assisted medical decision-making, particularly in the field of ophthalmology. A recent study in [83] compares four leading large language models (LLMs) in the context of uveitis, a complex ophthalmological condition. The research, available as a preprint, evaluates the performance of ChatGPT-3.5, ChatGPT-4.0, Google Gemini, and Anthropic Claude3 in handling uveitis-related queries. The study employed a set of carefully crafted questions covering various aspects of uveitis, including diagnosis, treatment, and management. The responses from each LLM were assessed for accuracy, completeness, and clinical relevance by a panel of expert ophthalmologists. The results revealed

significant variations in performance among the models, with newer versions generally outperforming their predecessors. This comparative analysis not only highlights the rapid advancements in AI capabilities within specialised medical domains but also emphasises the need for continuous evaluation and refinement of these models for clinical applications. The study provides valuable insights into the potential and limitations of using LLMs as assistive tools in ophthalmology, particularly in managing complex conditions like uveitis.

A recent study by Yoo et al. [84] evaluates the diagnostic capabilities of three advanced large language models (LLMs) in complex medical scenarios. The research, available as a preprint, compares GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro using "Diagnosis Please" cases from the New England Journal of Medicine. These cases are known for their diagnostic challenges, providing a rigorous test for the LLMs' medical reasoning abilities. The study assessed the models' performance in generating differential diagnoses, suggesting appropriate diagnostic tests, and providing final diagnoses. The results revealed significant variations in these models' diagnostic accuracies and reasoning capabilities, offering insights into their potential applications and limitations in medical diagnostics. This comparative analysis contributes to our understanding of how advanced AI models perform in specialised medical tasks and highlights areas for future improvement in AI-assisted medical diagnosis.

Zhang et al. [85] introduce CS-Bench, a comprehensive benchmark designed to evaluate the proficiency of large language models in computer science. This study presents a multifaceted approach to assessing LLMs' capabilities across various computer science domains, including programming, algorithms, data structures, and theoretical concepts. The benchmark consists of diverse tasks and problems, ranging from basic coding exercises to complex algorithmic challenges. By applying CS-Bench to several state-of-the-art LLMs, the researchers provide valuable insights into the strengths and weaknesses of these models in handling computer science-related tasks. The results showcase the current capabilities of LLMs in this field and identify areas where further development is needed. This benchmark serves as a crucial tool for researchers and developers in assessing and improving the performance of AI models in computer science applications.

A study by Zhang et al. [86] explores the public perception and sentiment towards ChatGPT in China through a comprehensive analysis of social media content. The research, published in *Technology in Society*, employs advanced text mining techniques to analyse a large corpus of social media posts related

to ChatGPT on Chinese platforms. The study investigates various aspects of public opinion, including general attitudes, perceived benefits and risks, and potential societal impacts of AI technology. By examining themes and sentiment patterns in the social media discourse, the researchers provide valuable insights into how ChatGPT is perceived in the context of Chinese society and culture. The findings highlight both enthusiasm for AI advancements and concerns about potential challenges, offering a nuanced view of the public's response to emerging AI technologies in China. This research contributes to our understanding of the social acceptance and potential cultural implications of large language models in different global contexts.

Due to the lack of development of the applications using the APIs provided by these chatbots, a major gap can be identified. In fact, there is also a lack of techniques used to compare all the chatbots. Let's discuss it in more detail.

## 2.6 RESEARCH GAP

After completing a thorough literature review in 2023, I identified the lack of use of text mining techniques for comparing generative AI chatbots. Specifically, there has never been any comparison between the model's applications.

Few studies compare ChatGPT with Gemini, Claude, and LLaMA, but it's hard to find any existing research that has compared the applications of these models. However, the API of these chatbots is available.

I couldn't find any research paper that compares the responses generated from the applications of ChatGPT API like You.com, ChatSonic, HuggingChat, Copy.ai and Rytr.me based on text mining techniques. These AI Chatbots are developed using the existing LLM models developed by these big companies. Even for project management tools like Jira, Notion, and ClickUp.com, which have active AI integration for customer text generation, no research paper compares the responses generated from them based on available text mining techniques.

*This research actively focuses on this gap identified, and based on this gap, a proposed methodology has been implemented, which is discussed in the following chapter.*

## CHAPTER 3: PROPOSED METHODOLOGY

This chapter outlines the comprehensive methodology employed in our research to evaluate and compare AI chatbots and project management tools. It is structured into three main sections: dataset collection, text mining techniques, and the implementation of these techniques. The chapter begins by detailing the process of prompt engineering and response generation, which form the foundation of our dataset. It then explores the various text mining techniques utilised in our analysis, including sentiment analysis, lexical diversity measurement, named entity recognition, and pronoun ratio calculation. Finally, the chapter delves into the practical implementation of these techniques using Python libraries such as Pandas, NLTK, and TextBlob. Throughout the chapter, we provide in-depth explanations of how each technique contributes to our comparative analysis, offering insights into the quantitative measures used to evaluate the performance and characteristics of AI-generated responses.

### 3.1 DATASET COLLECTION

The dataset in our research means the collection of the responses from the AI chatbots and the project management tools. So, as the first step, I need to devise a set of prompts that will be used to generate responses from the applications.

#### 3.1.1 PROMPT ENGINEERING

Prompt engineering plays a crucial role in the effective utilisation of AI chatbots and language models. The art and science of crafting input queries elicit desired responses from these AI systems. The importance of prompt engineering in AI chatbots cannot be overstated, as it directly influences the generated responses' quality, relevance, and usefulness. Effective, prompt engineering can enhance the accuracy and relevance of AI-generated content, guide the AI towards producing responses in a specific format or style, help in extracting specific types of information from the AI's knowledge base, mitigate biases and inconsistencies in AI responses, and optimise the AI's performance for specific tasks or domains.

In our research, we developed a comprehensive set of prompts through a systematic approach. Our process involved several key steps, including identifying key themes, crafting diverse prompts, iterative refinement, and standardisation. We began by outlining the primary themes we wanted to explore in our analysis, chosen based on their relevance to evaluating AI capabilities and their potential to generate diverse linguistic features for our text-mining analyses. Within each theme, we created a variety of prompts designed to test different aspects of the AI's capabilities. This diversity allows for a more

comprehensive evaluation of the AI systems. We conducted initial tests with a subset of prompts and refined them based on the quality and nature of responses received. This iterative process helped in optimising the prompts for our research objectives. To ensure consistency across different AI systems, we standardised the format and structure of our prompts while maintaining their diverse content.

The main themes for our prompts included knowledge-based queries (e.g., New Zealand history), industry-specific topics (e.g., automation in the New Zealand IT sector), creative writing tasks, problem-solving and ideation prompts, analytical and comparative queries, emotional and sentiment-oriented prompts, and project management-specific tasks (for project management tools). The prompts and their order are strategically designed to optimise the effectiveness of subsequent text-mining techniques. This approach involves crafting prompts that are likely to elicit responses rich in the linguistic features our analysis aims to examine. For instance, prompts are formulated to potentially generate responses with varied sentiment, diverse vocabulary, and different levels of complexity, thus providing a comprehensive dataset for our text mining analyses.

It is important to note that while the prompt structure is consistent across all chatbots, the interpretation and response to these prompts may vary depending on each chatbot's specific capabilities and training. This variance in interpretation does not compromise the validity of our comparison, as it reflects the inherent differences in the chatbots' performance - a key aspect of our study. In our research design, we developed two distinct sets of prompts: 50 for AI Chatbots and 35 for the AI features of Project Management Tools. For AI chatbots, we crafted a diverse range of prompts covering various topics and designed to elicit responses suitable for our text analysis metrics. For project management tools, our prompts were tailored to evaluate their capacity to generate specific types of content commonly required in project management contexts.

While we employed similar text mining techniques for analysing the responses from project management tools, our assessment extended beyond mere textual analysis. We also evaluated these tools based on their ability to effectively integrate AI-generated content into project management workflows, their user interface for AI interactions, and the relevance and practicality of the generated content in real-world project scenarios. This multifaceted approach allows for a more comprehensive evaluation of how these AI features enhance project management processes beyond just text generation. The complete list of

prompts and their corresponding responses are documented in the appendices, providing a transparent and detailed record of our proposed research methodology.

### 3.1.2 RESPONSE GENERATION

After defining the set of prompts, these prompts were sent to all the chatbots and project management tools individually, and all the responses generated from the chatbots were stored as the dataset in an Excel file [87]. Almost every chatbot or tool has its limitations in its free versions. Copy.ai allowed only 2000 words of generation in its free trial [88], Rytr.me offered 10k words every month [89], whereas You.com [90] and HuggingFace offered unlimited responses as they offered free usage of the GPT-3.5 Turbo model, but Chatsonic offered only 50 responses every day. Even in the project management tools, when trying their AI feature, ClickUp offered 25 responses/month for free, and Jira doesn't offer any AI feature when on the free plan. However, both companies' free trials helped continue the research endlessly. Notion offers only the first 20 responses in its AI feature trial, which can be bypassed by purchasing its subscription for 8. USD When reaching their limitations, I bought subscriptions for Copy.ai, Rytr.me, and Notion which were funded by Unitec. Chatsonic offered 50 response generations daily for free, which was enough to create its dataset. You.com offered a free trial for one month to explore other features, which was a great advantage to understanding how chatbot models other than OpenAI were integrated with You.com and HuggingChat, which offered all-time free access to all chatbot models, but I stuck to GPT-3.5 Turbo for all our comparisons as all the tools and the chatbots preferred to use GPT-3.5 Turbo over other models being cheaper for them. Copy.ai's one-day trial costs around 9 USD, and Rytr.me costs 6.3 USD.

When working on these chatbots with no limitations, I stored all the responses in an Excel file so that I could use Python's Pandas [91] to work with the dataset. Pandas is an open-source data analysis and manipulation tool built on Python. It is designed to be fast, powerful, flexible, and easy to use, making it a go-to tool for data scientists and analysts. Implementing the text mining techniques became much easier when pandas library's functions helped pick up the responses one by one from the Excel file [92]. Now, I could loop on every response, and implementation of text mining became easy. In the following section, I discuss the text-mining techniques that have been employed to extract quantitative values from the responses.



## 3.2 TEXT MINING TECHNIQUES

In our analysis of AI-generated responses, we employed several text-mining techniques to extract quantitative measures for comparison. These techniques allow us to objectively assess various aspects of the text, from emotional content to linguistic complexity. The following subsections detail each technique used in our study:

- **Sentiment Analysis:** Sentiment analysis involves determining a text's emotional tone or opinion. Researchers can gauge the responses' overall positivity, negativity, or neutrality by assigning sentiment scores to each response. This quantitative measure allows for a comparative emotional content analysis across different responses [93]. Sentiment analysis typically employs machine learning algorithms or lexicon-based approaches. In our implementation, we used a lexicon-based method where each word in the text is compared against a pre-defined dictionary of words with associated sentiment scores. The overall sentiment score for a piece of text is then calculated as the sum or average of the individual word scores, often normalised to a range between -1 (extremely negative) and 1 (extremely positive).
- **Lexical Diversity measure:** Lexical diversity measures how many unique words are used compared to the total number of words [94]. The library utilised the Type-Token Ratio (TTR) for this measure. TTR is calculated by dividing the number of unique words (types) by the total number of words (tokens) in a text:  $TTR = (\text{Number of unique words}) / (\text{Total number of words})$  [95]. A high score (closer to 1) indicates a rich vocabulary and varied expression, often seen in academic writing or works by skilled authors. It shows the ability to choose the most precise words to engage the reader. Moderate diversity (around 0.6-0.8) is a good balance for essays or everyday communication. Low diversity (below 0.5) suggests limited vocabulary and repetition, which is common in informal writing.
- **Named Entity Recognition (NER):** NER is a technique that identifies and extracts named entities, such as person names, organisations, locations, or dates, from the text. By quantifying the frequency and distribution of named entities in the responses, researchers can gain insights into the key actors, places, or events mentioned and compare their prominence across different responses [96].

- **Pronouns Ratio:** High pronoun ratios (many pronouns compared to total words) can indicate a more conversational or informal writing style. Conversely, lower ratios suggest a more formal style focusing on nouns and descriptive phrases. Pronoun ratios can give clues about the focus of the text. A high ratio of "I" and "you" pronouns might indicate a personal narrative, while a prevalence of "he," "she," and "they" could suggest a focus on specific people or groups. In some cases, pronoun ratios can hint at objectivity. Formal writing, aiming for neutrality, might use fewer first-person pronouns ("I," "we") and rely more on nouns and descriptive phrases. However, this is not a definitive measure. So, what I did is I collected all the pronouns from [97], stored all of them into an array and then calculated their ratio by the length of all the words in the paragraph.

### 3.3 IMPLEMENTATION OF TEXT MINING TECHNIQUES

Pandas [91], a powerful data manipulation library in Python, has significantly simplified working with Excel files. Mining techniques could be employed to analyse and compare the responses effectively. Pandas function *read\_csv()* [98] was used to convert my Excel file into [99]. Two different data frames were created as there were two different sheets of the dataset (one was AI Chatbot, and the other was the project management tools). To implement any text mining technique for comparison, in our case, it was required to apply it to each and every response. So, selecting each response using the Data frame's available functions became useful.

The pronoun ratio is a measure of how frequently pronouns are used in a text relative to the total word count. This technique involves several steps: a) Tokenization: The text is broken down into individual words or tokens. This process involves splitting the text into whitespace and handling punctuation. b) Part-of-Speech (POS) Tagging: Each token is then assigned a part-of-speech tag. This involves analysing each word in context and determining its grammatical category (noun, verb, pronoun, etc.). c) Pronoun Identification: Using the POS tags, pronouns are identified. This typically includes personal pronouns (I, you, he, she, it, we, they), possessive pronouns (my, your, his, her, its, our, their), and reflexive pronouns (myself, yourself, himself, herself, itself, ourselves, themselves). d) Ratio Calculation: The number of identified pronouns is divided by the total number of tokens to get the pronoun ratio. Lexical diversity assesses the variety of words used in a text. The Type-Token Ratio (TTR) is a common method for this: a) Tokenization: As before, the text is split into individual words. b) Normalization: Often, the tokens are normalised by converting to lowercase and possibly stemming or lemmatising to reduce inflected words to their base form. c) Type Counting: The number of unique words (types) is counted. d) TTR Calculation:

The number of types is divided by the total number of tokens to get the TTR. More sophisticated measures like Moving-Average TTR (MATTR) or Measure of Textual Lexical Diversity (MTLD) can be used to address TTR's sensitivity to text length. Sentiment analysis determines the emotional tone of a text. A basic lexicon-based approach involves a) Tokenization and normalisation of the text. b) Maintaining a pre-defined lexicon of words with associated sentiment scores. c) Identifying sentiment-bearing words in the text. d) Calculating an overall sentiment score by aggregating individual word scores, often considering negations and intensifiers. More advanced methods use machine learning: a) Feature extraction: Converting text into numerical features (e.g., using bag-of-words or TF-IDF). b) Training a classifier (e.g., Naive Bayes, SVM) on a labelled dataset of texts with known sentiments. c) Using the trained model to predict sentiment on new texts. Subjectivity analysis determines how much of a text reflects opinion versus factual information. A basic approach might involve: a) Maintaining separate lexicons for subjective and objective words. b) Tokenizing and normalising the text. c) Counting occurrences of subjective and objective words. d) Calculating a ratio of subjective to total words as the subjectivity score. More sophisticated approaches use machine learning, similar to sentiment analysis: a) Creating a labelled dataset with texts marked as subjective or objective. b) Extracting features from the text (e.g., n-grams, POS tags, syntactic structures). c) Training a classifier to distinguish between subjective and objective text. d) Using the trained model to assign subjectivity scores to new texts.

Now, when I was able to loop in a series of responses, I began applying the text mining techniques using available functions in libraries like Natural Language Toolkit [100], Textblob [101] and Scikitlearn [102]. Using them, I calculated all the possible scores and matrices. The selection of appropriate text-mining techniques relies on two key factors: the expected outcomes and the capabilities of existing text-mining methods. It was evident that quantitative values were necessary for two main reasons - Firstly, all the responses needed statistical analysis to determine which was superior. Quantitative values allow for a more objective and data-driven approach to evaluating the responses, enabling researchers to draw meaningful conclusions based on numerical comparisons. Secondly, quantitative values facilitate further statistical analysis and modelling.

### **3.3.1 COMBINING TECHNIQUES FOR COMPREHENSIVE ANALYSIS**

Combining multiple text mining techniques is often beneficial to comprehensively understand the responses. For example, sentiment analysis can be used with topic modelling to determine the emotional tone associated with specific topics [103]. Similarly, TF-IDF can be employed to identify the most

important words within each topic, providing a more granular analysis of the responses [104]. So, ultimately, I figured out the possible scores that can be derived by applying combined text mining techniques, and these scores can be used to compare the chatbots. Assessing the polarity scores and subjectivity scores from the sentiment analysis processes in the library helps understand interactions' emotional and opinionated nature by knowing how neutral positive or negative responses are. Similarly, in the upcoming section, more details of how all the scores have been determined using the text mining techniques.

### **3.2.2 DETERMINING SCORES**

In comparing project management tools and AI chatbots, I found that quantitative analysis through text mining techniques can provide valuable insights into the performance, capabilities, and characteristics of these systems. By determining various scores, we can objectively evaluate and compare different tools and chatbots based on their responses to project management-related queries. This section presents a structured approach to calculating and interpreting multiple scores derived from text mining techniques. These scores offer a multi-faceted view of the responses generated by project management tools and AI chatbots, allowing for a comprehensive comparison. The scores we'll examine include sentiment scores, subjectivity scores, pronoun ratio, average word lengths, lexical diversity scores, and readability scores (based on the Flesch Kincaid and Coleman Liau index). Each subsection details the calculation method, the significance of the score in the context of project management communication, and how it contributes to our overall comparison. By analysing these scores collectively, we can gain insights into various aspects such as: - The emotional tone of responses (based on sentiment score), the balance between factual and opinion-based information (subjectivity score), the level of personalisation in communication (pronoun ratio), the complexity and sophistication of language used (average word length and lexical diversity), the clarity and accessibility of the information provided (readability scores) these metrics help in evaluating how well each tool or chatbot communicates project management concepts, adapts to different query types, and provides clear, relevant information. This quantitative approach complements qualitative assessments, offering a data-driven foundation for comparing the effectiveness of various project management tools and AI chatbots in addressing user needs and facilitating project-related communication.

### 3.3.2.1 SENTIMENT SCORE CALCULATION:

Sentiment analysis is a crucial aspect of evaluating the responses generated by chatbots. This section explains how sentiment scores were calculated to determine the overall emotional tone of the chatbot and project management tools' responses. First, I discussed the sentiment score calculation, methodology, and weighted average calculation. Then, I have described the implementation of the sentiment analysis in the code.

Sentiment score is determined using a lexical-based approach from [101], which estimates a text snippet's sentiment based on its constituent words. This method relies on sentiment lexicons, which are pre-compiled dictionaries where words are assigned sentiment scores, typically ranging from -1 (extremely negative) to +1 (extremely positive), with 0 representing neutral sentiment. Lexicons are comprehensive lists of words or phrases in a language, often including additional information about each entry. In the context of sentiment analysis, these lexicons specifically include sentiment scores for each word.

The lexical-based approach operates by first tokenising the text into individual words or tokens. Each token is then looked up in the sentiment lexicon, and if found, its pre-assigned sentiment score is retrieved. These individual word scores are subsequently aggregated to produce an overall sentiment score for the text. Aggregation methods may include simple averaging of all word scores, weighted averaging, giving more importance to certain words (e.g., adjectives), or consideration of negations or intensifiers that might modify sentiment. The final score is often normalised to a specific range, such as -1 to +1 [101].

Technical considerations in this approach include handling negations that can invert the sentiment of following words, dealing with intensifiers that can amplify sentiment, addressing context sensitivity where the same word might have different sentiments in different contexts, and managing out-of-vocabulary words not found in the lexicon. Sentiment lexicons can be general-purpose or domain-specific, manually curated by experts, or automatically generated using machine learning techniques.

TextBlob library maintains its own internal sentiment lexicons, but we can also integrate custom lexicons tailored to the specific domain or needs. Each word from the response is compared against the available sentiment lexicons. The word's corresponding sentiment score is retrieved if a match is found. Words not found in the lexicon are typically assigned a neutral score (0). TextBlob library calculates a weighted average of the sentiment scores for all words in the text snippet. This is where it goes beyond a simple

word count. Weights are assigned to each word's sentiment score, considering factors like words expressing stronger sentiment (e.g., "awful" vs. "bad") might have higher weights to reflect their greater emotional impact. Negation words like "not" or "no" likely reverse the polarity of the following word. TextBlob library factors this in, adjusting the sentiment score accordingly [105].

In some cases, the surrounding context might influence the weight assigned to a word. For example, "sick" can be positive ("That concert was sick!") or negative ("I feel sick"). TextBlob library might consider nearby words to determine the intended meaning and adjust weights. The specific weighting scheme used by TextBlob library is not publicly documented, making it somewhat of a "black box." However, the general principles of intensity, negation handling, and potential contextual influence are likely employed [106].

In my code, after the weighted average calculation, TextBlob library stores the resulting score as a floating-point value in the *sentiment.polarityproperty*. In my code, *analyzer = SentimentIntensityAnalyzer()*: This line creates an instance of the *SentimentIntensityAnalyzer* class from the TextBlob library. This analyzer is specifically designed to identify the emotional tone of text data. *sentiment\_scores = analyzer.polarity\_scores(paragraph)*. This line is called the *polarity\_scores* method for the analyser object. It passes the paragraph variable containing the response from the chatbot in the loop as an argument. The *polarity\_scores* method analyses the sentiment of the paragraph and returns a dictionary containing sentiment scores. *sentiment\_scores\_list.append(sentiment\_scores['compound'])*.

It extracts the 'compound' score from the *sentiment\_scores* dictionary and appends it to the *sentiment\_scores\_list*. The 'compound' score is a single numerical value between -1 (most negative) and 1 (most positive), representing the overall sentiment of the response.

### 3.3.2.2 SUBJECTIVITY SCORE CALCULATION:

The subjectivity score measures the presence of opinions and beliefs versus factual information. This can reveal whether chatbots provided objective summaries or leaned towards subjective interpretations. Similar to sentiment analysis, [101] uses pre-built lexicons calculated in the same way it gets calculated by tokenisation, but these lexicons focus on identifying subjective words and phrases (e.g., "I believe," "most people think"). It maintains its own internal subjectivity lexicons, and you can also integrate custom lexicons for domain-specific subjectivity cues.

Each word in the input text is compared against the available subjectivity lexicons. If a match is found, a score of 1 (highly subjective) is typically assigned to that word. Words not found in the lexicon are likely assigned a 0 (objective) score. While [101] doesn't explicitly mention weighting subjectivity scores, they might use a simple average or a more nuanced weighting scheme to account for this. Words expressing stronger subjectivity (e.g., "absolutely love" vs. "kind of like") might have higher weights. In some cases, the surrounding context might influence the subjectivity score. For example, "The president is a strong leader" is more subjective than "The president gave a speech today." TextBlob might consider context to adjust weights (though this is not confirmed). After calculating (potentially a weighted average), [101] stores the resulting score as a floating-point value in the *sentiment.subjectivityproperty*. This typically ranges from 0 (completely objective) to 1 (highly subjective).

In my code, *subjectivity\_score = TextBlob(paragraph).sentiment.subjectivity*, [101] employs a lexicon-based approach to estimate the subjectivity of a text snippet. Subjectivity refers to the degree to which the text expresses personal opinions or beliefs instead of factual information [106]. *TextBlob(paragraph)* creates an object from the paragraph variable [101]. This object represents the text you want to analyse. *.sentiment*: Accesses the sentiment property of the TextBlob object [101]. This property contains various sentiment analysis functionalities. *.subjectivity*: Extracts the subjectivity score from the sentiment property. This score represents the degree to which the text expresses personal opinions, beliefs, or feelings (vs. factual information) [107]. Then, I stored the subjectivity scores in the array to save my results.

### **3.3.2.3 AVERAGE WORD LENGTH CALCULATION:**

Average Word Length shows the average number of letters per word. This can offer insights into the writing style of the chatbot responses. Generally, higher AWL (Average Word Length) indicates a more complex and formal style, while lower AWL suggests a simpler and more informal approach. AWL can be loosely correlated with readability [108]. Longer words can make the text less readable for audiences with a lower reading level. In some cases, AWL can provide clues about the content and conciseness of the responses. A chatbot that consistently produces short responses with low AWL might be struggling to elaborate on ideas. It is calculated by dividing the total number of letters in the paragraph by the total number of words.

### **3.3.2.4 READABILITY SCORES (FLESCH-KINCAID GRADE LEVEL & COLEMAN-LIAU INDEX):**

To assess the complexity and accessibility of the chatbots' responses, we employed two widely used readability metrics: the Flesch-Kincaid Grade Level and the Coleman-Liau Index. These metrics provide an

estimate of the US school grade level required to understand a given text, offering insights into how effectively the chatbots communicate with users of varying educational backgrounds. The Flesch-Kincaid Grade Level score typically ranges from 0 to 18, with lower scores indicating easier readability. For instance, a score of 8.0 suggests that the text is understandable by an eighth-grade student. The Coleman-Liau Index also estimates the U.S. grade level needed to understand the text, but it uses a different calculation method that may yield slightly different results [109]. These readability scores are particularly important in our context as they help evaluate how well the chatbots adapt their language to suit the target audience. A chatbot that consistently produces responses with high readability scores might be less effective for general users, while one that maintains lower scores could be more accessible [110].

Traditionally, the Flesch-Kincaid Grade Level is calculated using the formula:  $0.39 * (\text{total words} / \text{total sentences}) + 11.8 * (\text{total syllables} / \text{total words}) - 15.59$  [111]. Similarly, the Coleman-Liau Index is computed using this formula:  $0.0588L - 0.296S - 15.8$  [112], where L is the average number of characters per 100 words, and S is the average number of sentences per 100 words.

In our implementation, we utilised the *textstats* library in Python to calculate these scores. The library abstracts the complexity of syllable counting and other intricate calculations: `readability_score_fk = textstats.flesch_reading_ease(paragraph)` and `readability_score_cl = textstats.coleman_liau_index(paragraph)`. These functions take a text paragraph as input and return the respective readability scores. The *textstats* library likely implements the standard formulas internally, automatically handling word, sentence, syllable, and character counting. This approach allowed us to efficiently process multiple responses and store the resulting scores in arrays for further analysis and comparison across different chatbots and project management tools.

### 3.3.2.5 PRONOUN RATIO:

It provides a basic indicator of writing style (formal vs. informal) and focus (personal narrative vs. specific entities) [113]. The code I wrote is designed to analyse a list of words (`'words'`) and determine the proportion of words that are classified as pronouns based on a predefined list. Let's break down how this works step by step. Firstly, the `'pronoun_count'` variable is calculated using a generator expression inside the `'sum()'` function. This expression iterates through each word in the `'words'` list. For each word, it converts the word to lowercase (`'word.lower()'`) to ensure case insensitivity when comparing against the list of pronouns. The condition `'if word.lower() in [...]'` checks if the lowercase version of the current word exists in the list of pronouns. The list of pronouns contains various forms such as "you", "he", "she", "it",



"they", "who", "what", and others, encompassing singular and plural forms as well as possessive and reflexive pronouns.

When a word from `words` matches any pronoun in this list, the generator expression yields `1`, indicating that the word is a pronoun. The `sum()` function then calculates the total count of such occurrences, giving us `pronoun_count`. Next, the `pronoun_ratio` is computed by dividing `pronoun_count` by the total number of words in the `words` list (`len(words)`). This ratio represents the proportion of words in `words` that are classified as pronouns according to the predefined pronoun list. Finally, assuming there's a list `pronoun_ratios` where these ratios are being stored for different sets of `words`, `pronoun_ratio` is appended to `pronoun_ratios` after each calculation. This allows for tracking and comparison of pronoun usage across different texts or contexts. I collected the whole list of pronouns from [97]. The ratio I got indicates how well the text connects ideas. How is it calculated? It divides the number of pronouns counted by the number of words in the response. This gives a value between 0 and 1, representing the percentage of words that are pronouns. A higher pronoun ratio might suggest a more cohesive text, as pronouns help connect ideas. However, it's not the only factor to consider [114]. - A high ratio could indicate repetitive use of pronouns or a focus on referring to other entities instead of providing new information. - A low ratio might suggest a text that lacks flow or clear connections between sentences [115]. This approach provides a quantitative measure to compare the usage of pronouns across different chatbot responses. Still, it's important to use it alongside other metrics and consider the context of the responses for a more comprehensive evaluation.

#### **3.3.2.6 LEXICAL DIVERSITY SCORE (TYPE-TOKEN RATIO):**

In my code,  $lexical\_diversity\_score = len(unique\_words) / len(words)$ : This line calculates the lexical diversity score. It divides the number of unique words ( $len(unique\_words)$ ) by the total number of words ( $len(words)$ ) [116]. A higher score indicates greater lexical diversity, meaning the chatbot response uses a wider variety of words and avoids repetition [117]. What does the lexical diversity score mean? Lexical diversity refers to the variety of unique words used in a text. A higher lexical diversity score indicates that a text uses a wider range of words, while a lower score suggests more repetition of the same words. The use of this score can help us know the clarity, precision and engagement of the response [117]. By analysing these scores together, I can gain a comprehensive picture of the strengths and weaknesses of each chatbot response. Here are some examples: A high sentiment score, readability score, and lexical diversity score might indicate a clear, informative, and positive response. A low subjectivity score and a high key phrase relevance score could suggest a factual, focused response that directly addresses the

prompt. A high bias score and low pronoun ratio might highlight a potentially biased and formal response that needs further refinement.

### **3.4 PROJECT MANAGEMENT TOOLS ANALYSIS METRICS**

In addition to the text mining techniques, the Project Management Tools, including ClickUp.com, Jira, and Notion, can be compared on the basis of the features they offer in their AI feature. I have divided these features into two parameters for the comparison metrics – Data utilisation and integration features. In this section, these three categories have determined the comparison of the project management tool's AI feature other than just the text mining techniques.

#### **3.4.1 DATA UTILISATION**

To measure data utilisation, assess how well the tools incorporate existing project data into generated responses. For the evaluation, the primary objective focuses on critical data points such as names, dates, tasks, and other project-specific details to determine how much each tool leverages this data to produce contextually relevant and specific responses.

First, we identify the essential data points commonly used in project management responses. These data points include team members' names, important dates (such as deadlines and milestones), specific tasks, and other relevant project details. Including these data points in generated responses is crucial for ensuring that the responses are tailored to the project's specific context and provide meaningful information. Next, we evaluate the responses generated by each tool against these identified data points. For instance, we examine whether the responses from ClickUp include the names of the sender and receiver, dates from the project timeline, and task details directly from the existing projects. The thoroughness and accuracy of these inclusions are crucial for effective data utilisation.

The assessment criteria for data utilisation include the following factors: the inclusion of names, the incorporation of dates, the presence of specific task details, and the integration of other project-specific details. For example, I analysed whether the responses incorporate specific names from the project, including names of team members, stakeholders, and any other relevant personnel involved in the project. The presence of specific names indicates a higher level of data integration. Also, I checked if the responses included relevant dates such as project start dates, deadlines, and milestones. Accurate inclusion of dates helps provide a clear timeline and context for the responses. I evaluated whether the responses contained specific task details from the project. This involves checking for the inclusion of task names, descriptions, and statuses. Detailed task information in the responses signifies a deeper integration of project data.

Then, I looked for any additional project-specific details that were included in the responses. This could be information such as project objectives, key deliverables, or any other relevant data points that provide a comprehensive view of the project context.

By systematically evaluating the responses based on these criteria, we can measure the effectiveness of each tool in utilising project data. A tool that consistently incorporates these data points in its responses demonstrates more data utilisation, providing more contextually relevant and specific information. This thorough assessment helps understand each tool's strengths and weaknesses in terms of data integration and utilisation in project management.

### **3.4.2 INTEGRATION FEATURES**

ClickUp offers three features “Continue Writing” from where we left off, write with a prompt, and “Write StandUp” [118]. Jira only offers a writing assistant that asks for prompts, which it uses to write. [119]. Notion, being a note-making company, offers various features, the ClickUp’s “Continue Writing”, Find action items, Translate (which it does in English, Korean, Chinese, Japanese, Spanish, Russian, French, Portuguese, German, Italian, Dutch, Indonesian, Filipino, Vietnamese), “Explain this” feature (which is generally used when we need any paragraph, or text to be explained in simpler language), Fix spelling and grammar of any language, Make shorter, Make longer, Change tone, Simplify language, Ask AI to edit..., Brainstorm Ideas (Random), Blog post, Outline, Social Media post, Press release, Creative story, Essay, Poem, To-do list, Meeting agenda, Pros and Cons list, Job description, Sales email, and Recruiting email in Notion AI [53].

As a user, based on the results we got from the prompts of 35 responses, we just focussed on the common feature (writing assistant which asks for prompts) on which we have implemented the text mining techniques. We need to understand the other features as well. In the following chapter, discussions on the results based on text mining techniques and these other integration features with how accurately we can find these tools based on the data utilisation have been discussed.

## CHAPTER 4: RESULTS & DISCUSSIONS

The results I got after implementing the techniques discussed in the previous chapter were stored in a Google Sheets dataset that can be accessed via [87].

I got the results from the code for each and every response, but later, I realised I needed to find the overall scores of the whole chatbot. There is a need to perform statistical analysis to determine the overall performance of the chatbot and answer a few important questions, such as which chatbot is more readable. Or which project management tool is sentimentally positive? So, I calculated the scores' mean, median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile and interquartile range. Let us elucidate the significance of these statistical measures and their practical utility.

**Mean:** The mean or average score for each type of score can provide a general notion of the central tendency or typical value across all the responses. This can help identify which scores tend to be higher or lower on average. Formula:  $Mean = (\text{sum of scores in the list}) / \text{number of observations (questions)}$  [120].

**Median:** The median score represents the middle value when the scores are sorted in ascending or descending order. It can provide a more robust measure of central tendency, especially if the data has outliers or extreme values that might skew the mean. Formula: AI Chatbot case,  $Median = [(n / 2)\text{th term} + ((n / 2) + 1)\text{th term}] / 2$  (where  $n$  is the total number of observations (questions)). In project management tools, the formula is  $Median = [(n + 1) / 2]\text{th term}$  [120].

**Mode:** The mode is the most frequently occurring score for each type of score. It can help identify the most common or typical value that appears across the responses. Formula:  $Mode = L + h * [(fm - f_1) / (fm - f_1 - f_2)]$ . Where details are omitted for brevity [120].

**Standard Deviation:** The standard deviation measures the spread or dispersion of the scores around the mean. A higher standard deviation indicates a wider spread, while a lower one suggests scores clustered closer to the mean. Formula:  $SD = \sqrt{\sum(\bar{x} - x)^2 / (n - 1)}$ . Where details omitted [121].

Skewness: Skewness measures the asymmetry of the distribution of scores. A positive skew indicates a longer tail towards higher scores, while a negative skew suggests a longer tail towards lower scores.

Formula:  $Skewness = \frac{\sum(\bar{x} - x)^3}{(n * SD^3)}$  [121].

First Quartile ( $Q1$ ), Third Quartile ( $Q3$ ), Interquartile Range ( $IQR$ ):  $Q1$  and  $Q3$  represent the values below which 25% and 75% of scores fall, respectively.  $IQR = Q3 - Q1$ , measuring the spread of the middle 50% of scores [122].

Kurtosis: Kurtosis measures the "peakiness" or "flatness" relative to a normal distribution. A higher value indicates a more peaked distribution, while a lower value suggests a flatter distribution. Formula:  $\frac{\sum(\bar{x} - x)^4}{(n * SD^4)} - 3$  [121].

By analysing these statistical measures for each type of score, I gained insights into the score distributions' central tendency, dispersion, skewness, and overall shape. This information can help better understand the patterns, identify potential outliers or deviations, and make more informed interpretations and decisions for the users of the applications based on the text mining results across the different types of scores. I applied these measures using the Python statistics library [123] and SciPy stats library [124] on the scores obtained from the text mining techniques. The following subsections present the results in the tables and box plots.

## 4.1 LEXICAL DIVERSITY SCORE FOR AI CHATBOTS

Table 4.1: Lexical Diversity Score for AI chatbots

Measures	HuggingChat	Chatsonic	You Chat	Copy.ai	Rytr.me
Average(mean)	0.649	0.585	0.577	0.594	0.614
Median	0.667	0.558	0.551	0.584	0.592
Mode	0.768	0.703	0.571	0.702	0.784
Standard Deviation	0.094	0.088	0.104	0.064	0.084
Skewness	-0.28	0.98	0.915	0.55	0.726
Kurtosis	0.028	0.33	0.373	-0.385	-0.579
First Quartile	0.61	0.52	0.494	0.543	0.548
Third Quartile	0.69	0.63	0.62	0.635	0.672
Interquartile Range	0.067	0.11	0.13	0.92	0.124

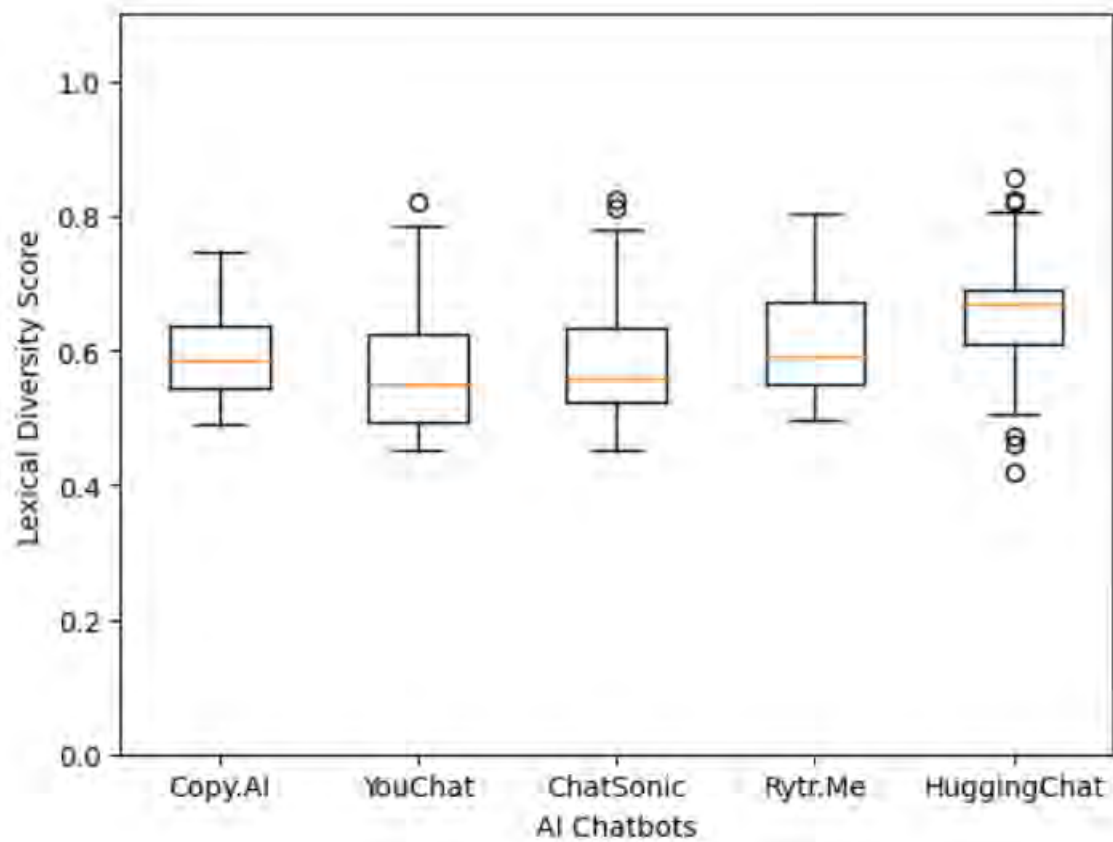


Figure 4.1: Box plot showing lexical diversity scores for AI Chatbots category

The table (Table 4.1) presents various statistical measures for the lexical diversity scores of five AI chatbots: HuggingChat, Chatsonic, YouChat, Copy.ai, and Rytr.me. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that HuggingChat has the highest mean score of 0.649, followed by Rytr.me with 0.614, Copy.ai with 0.594, Chatsonic with 0.585, and YouChat with the lowest mean of 0.577. The median values align closely with the mean values, with HuggingChat again having the highest median score of 0.667. The mode values indicate the most frequently occurring scores, with Rytr.me and HuggingChat showing higher modes (0.784 and 0.768, respectively). In terms of variability, Copy.ai demonstrates the lowest standard deviation (0.064), while YouChat shows the highest (0.104). Skewness values indicate that HuggingChat has a slight left skew (-0.28), while Chatsonic and YouChat show significant right skewness (0.98 and 0.915, respectively). Kurtosis values reveal that HuggingChat has a nearly normal distribution (0.028), whereas Rytr.me has the lowest kurtosis value (-0.579), indicating a flatter distribution with lighter tails.

The box plot in Figure 4.1 represents the distribution of lexical diversity scores for the five AI chatbots, providing a clear comparison of their score distributions. The boxes represent the interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that HuggingChat has the highest median and relatively low variability, as shown by its narrow IQR and short whiskers. In contrast, YouChat exhibits the highest variability, evidenced by its wider IQR and more extended whiskers. Additionally, the presence of outliers in several chatbots, such as HuggingChat and Chatsonic, indicates scores deviate significantly from their distributions' central tendency.

## 4.2 LEXICAL DIVERSITY SCORE FOR PROJECT MANAGEMENT TOOLS

Table 4.2: Lexical Diversity Scores for Project Management Tools

Measures	ClickUp.com	Notion.so	Jira
Average(mean)	0.617	0.602	0.789
Median	0.605	0.598	0.788
Mode	0.59	0.540	0.752
Standard Deviation	0.06	0.100	0.065
Skewness	0.332	1.69	-0.231
Kurtosis	-0.617	5.538	0.153
First Quartile	0.571	0.549	0.752
Third Quartile	0.658	0.636	0.82
Interquartile Range	0.087	0.087	0.067

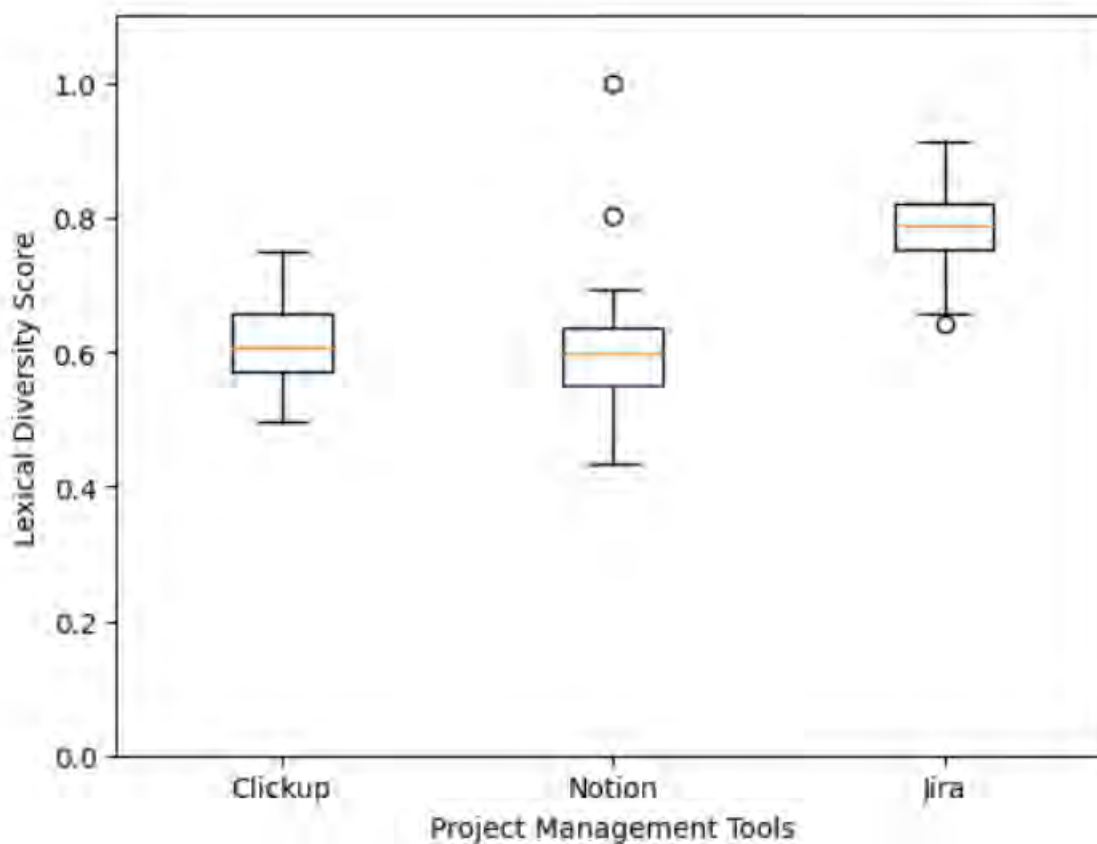


Figure 4.2: Box Plot showing Lexical Diversity Scores for the Project Management Tools



The table (Table 4.2) presents various statistical measures for the lexical diversity scores of three project management tools: ClickUp.com, Notion.so, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Jira has the highest mean score of 0.789, followed by ClickUp.com with 0.617, and Notion.so with the lowest mean of 0.602. The median values align closely with the mean values, with Jira again having the highest median score of 0.788. The mode values indicate the most frequently occurring scores, with Jira showing the highest mode of 0.752. In terms of variability, ClickUp.com demonstrates the lowest standard deviation (0.06), while Notion.so shows the highest (0.100). Skewness values indicate that Notion.so has a significant right skewness (1.69), while Jira shows a slight left skew (-0.231), and ClickUp.com has a moderate right skewness (0.332). Kurtosis values reveal that Notion.so has a highly peaked distribution (5.538), whereas ClickUp.com has the lowest kurtosis value (-0.617), indicating a flatter distribution with lighter tails.

The box plot in Figure 4.2 represents the distribution of lexical diversity scores for the three project management tools, providing a clear comparison of their score distributions. The boxes represent the interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that Jira has the highest median and relatively low variability, as shown by its narrow IQR and short whiskers. In contrast, Notion.so exhibits the highest variability, evidenced by its wider IQR and more extended whiskers. Additionally, the presence of outliers in several tools, such as ClickUp.com and Notion.so, indicates scores deviate significantly from their distributions' central tendency.

### 4.3 SENTIMENT SCORE FOR AI CHATBOTS

Table 4.3: Sentiment Scores for AI Chatbots

Measures	ChatSonic	HuggingChat	You	Rytrme	CopyAI
Average(mean)	0.604	0.574	0.599	0.626	0.533
Median	0.982	0.981	0.968	0.979	0.98
Mode	-0.296	0.996	0.996	0.852	0.994
Standard Deviation	0.733	0.768	0.7164	0.712	0.802
Skewness	-1.623	-1.478	-1.646	-1.763	-1.20
Kurtosis	0.795	0.264	0.937	1.23	-0.27
First Quartile	6.273	5.99	6.054	6.06	0.543
Third Quartile	6.896	6.60	6.435	6.69	0.635
Interquartile Range	0.623	0.061	0.381	0.063	0.092

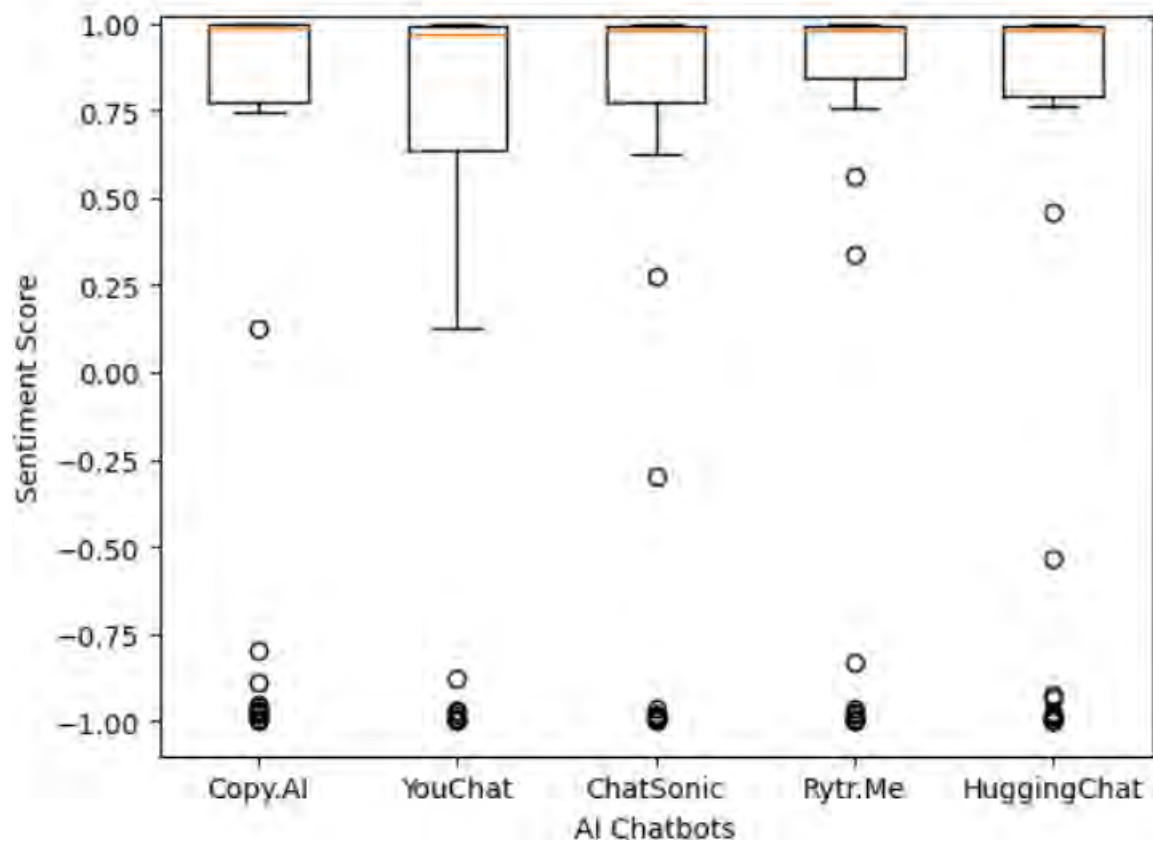


Figure 4.3: Box Plot showing Sentiment Scores for AI Chatbots

ChatSonic shows strong sentiment scores with an average (mean) of 0.604482 and a median of 0.98255, indicating overall positive sentiment. However, its mode is negative (-0.296), suggesting some instances of negative sentiment are common. The standard deviation (0.73) is the highest, reflecting significant variability in sentiment scores. ChatSonic has a negative skewness (-1.62), indicating a bias towards higher positive scores, and a kurtosis (0.79) close to zero, suggesting a normal distribution with some outliers. The first quartile (6.27) and third quartile (6.89) are high, reflecting a broad spread of sentiment scores. Overall, ChatSonic performs well in sentiment but shows high variability and some negative sentiment occurrences. HuggingChat has a moderate average sentiment score (mean) of 0.57 and a high median of 0.9817, indicating a generally positive sentiment. The mode (0.99) is nearly 1, showing frequent high positive sentiment. Its standard deviation (0.77) is quite high, indicating variability. HuggingChat's skewness (-1.47) is negative, showing a bias towards higher positive scores, and its kurtosis (0.26) is close to zero, indicating a relatively normal distribution. The first quartile (5.999529308) and third quartile (6.602) indicate a wide range of sentiment scores. Overall, HuggingChat has strong positive sentiment but with considerable variability. YouChat shows a solid sentiment performance with an average (mean) of 0.59 and a high median of 0.96805, reflecting generally positive sentiment. The mode (0.9961) is high, indicating frequent positive sentiment. The standard deviation (0.71) suggests variability in sentiment scores. You have a negative skewness (-1.64), showing a bias towards higher positive scores, and a kurtosis (0.937) close to one, suggesting some outliers. The first quartile (6.05) and third quartile (6.43) show a broad spread of sentiment scores. Overall, You Chat demonstrate strong positive sentiment with some variability. Rytr.me has the highest average sentiment score (mean) of 0.62 and a high median of 0.979, indicating positive sentiment. The mode (0.852) is high, showing frequent positive sentiment. Its standard deviation (0.71) is moderate, indicating variability. Rytr.me's skewness (-1.76) is highly negative, indicating a strong bias towards higher positive scores, and its kurtosis (1.23) is the highest, suggesting more outliers. The first quartile (6.07) and third quartile (6.697) indicate a broad range of sentiment scores. Overall, Rytr.me shows the highest positive sentiment but with notable variability and outliers. Copy.ai has the lowest average sentiment score (mean) of 0.53 but a high median of 0.9853, indicating generally positive sentiment, with some lower scores pulling the average down. The mode (0.99) is high, indicating frequent positive sentiment. Copy.ai's standard deviation (0.80) is the highest, indicating significant variability. Its skewness (-1.29) is negative, showing a bias towards higher positive scores, and its kurtosis (-0.27) is close to zero, indicating a relatively normal distribution. The first quartile (5.54) and third quartile (6.63) indicate a broad spread of sentiment scores. Copy.ai shows positive sentiment but with high variability and a lower average score.

The box plots in Figure 4.3 illustrate differences in sentiment scores among the AI chatbots despite similar median values. ChatSonic and HuggingChat, for instance, have high medians close to 1.0, indicating a generally positive sentiment. However, ChatSonic's broader interquartile range (IQR) and longer whiskers suggest more variability and a wider distribution of sentiment scores compared to HuggingChat. Additionally, YouChat, despite having a high median, shows a larger spread and more outliers, indicating significant variability and occasional negative sentiment scores. Rytr.me and Copy.ai, while also exhibiting high median scores, show distinct differences in their spread and the presence of outliers, highlighting variability in their sentiment consistency.

#### 4.4 SENTIMENT SCORE FOR PROJECT MANAGEMENT TOOLS

Table 4.4: Sentiment Scores for Project Management Tools

Measures	ClickUp	Notion	Jira
Average(mean)	0.732	0.778	0.794
Median	0.977	0.970	0.940
Mode	0.926	0.318	0.950
Standard Deviation	0.559	0.45	0.437
Skewness	-2.27	-2.63	-3.22
Kurtosis	3.71	6.36	9.436
First Quartile	6.08	5.315	6.37
Third Quartile	6.732	6.164	7.07
Interquartile Range	0.65	0.84	0.067

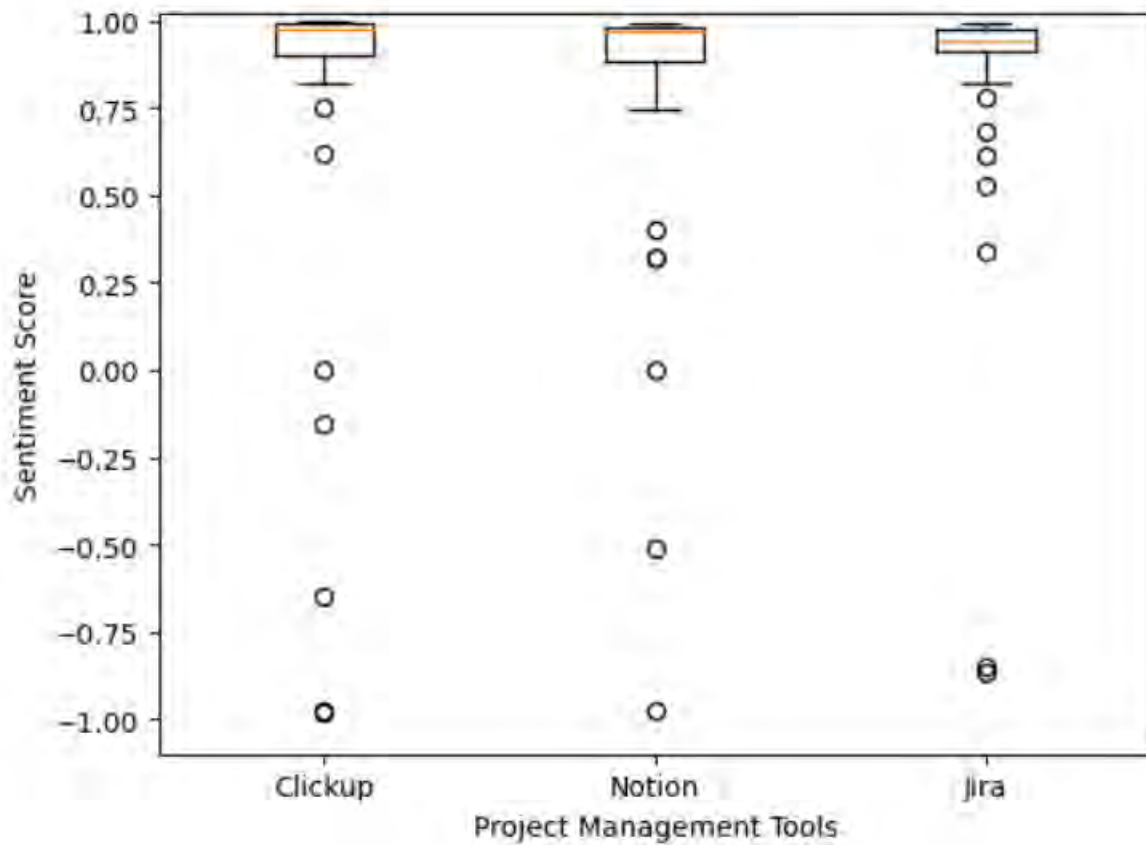


Figure 4.4: Box Plot showing Sentiment Scores for Project Management Tools

The table (Table 4.4) presents various statistical measures for the sentiment scores of three project management tools: Clickup, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Jira has the highest mean score of 0.794, followed closely by Notion with 0.778 and Clickup with the lowest mean of 0.732. The median values are similar to the mean values, with Notion having the highest median score of 0.970. The mode values indicate the most frequently occurring scores, with Jira showing the highest mode at 0.950. In terms of variability, Jira demonstrates the lowest standard deviation (0.437), while Clickup shows the highest (0.559). Skewness values indicate that all three tools have negative skewness, with Jira being the most negatively skewed at -3.22. Kurtosis values reveal that Jira has the highest kurtosis value (9.456), indicating a more peaked distribution with heavier tails compared to Clickup and Notion.

The box plot in Figure 4.4 represents the distribution of sentiment scores for the three project management tools, clearly comparing their score distributions. The box plot highlights that Jira has the highest median and the narrowest interquartile range (IQR), indicating lower variability in its sentiment scores compared to Clickup and Notion. Clickup and Notion exhibit similar median values, but Clickup shows a wider IQR and more extended whiskers, suggesting higher variability in its sentiment scores. Additionally, the presence of outliers in all three tools, particularly in the lower range, indicates instances of scores that deviate significantly from the central tendency of their distributions.

## 4.5 SUBJECTIVITY SCORE FOR AI CHATBOTS

**Table 4.5: Subjectivity Scores for AI Chatbots**

Measures	ChatSonic	HuggingChat	You	Rytrme	CopyAI
Average(mean)	0.430	0.462	0.437	0.437	0.474
Median	0.428	0.476	0.439	0.450	0.485
Mode	0.359	0.311	0.393	0.418	0.427
Standard Deviation	0.094	0.080	0.101	0.108	0.107
Skewness	-0.121	-0.180	0.106	-0.430	0.180
Kurtosis	-0.292	-0.733	-0.76	-0.381	-0.021
First Quartile	0.360	0.4	0.355	0.373	0.420
Third Quartile	0.496	0.51	0.507	0.508	0.541
Interquartile Range	0.135	0.11	0.152	0.135	0.121

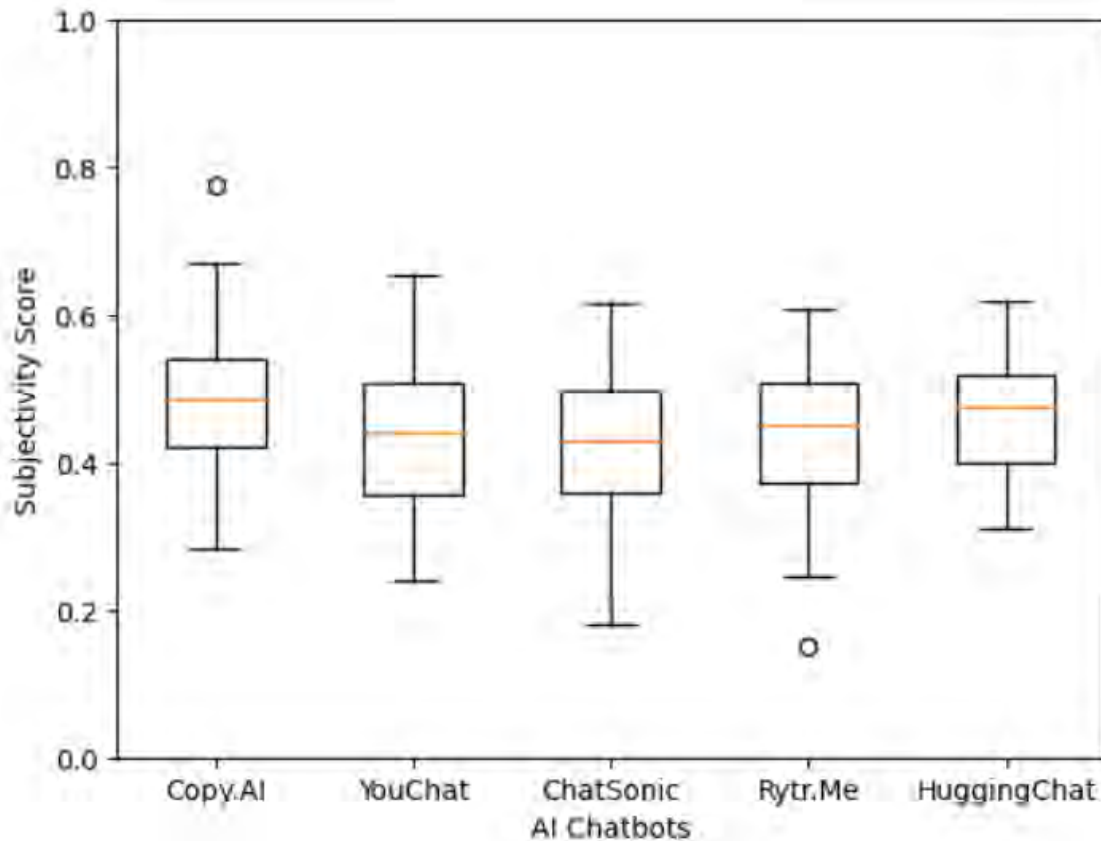


Figure 4.5: Box Plot showing Subjectivity Scores for Project Management Tools

The table (Table 4.5) presents various statistical measures for the subjectivity scores of five AI chatbots: ChatSonic, HuggingChat, YouChat, Rytr.me, and Copy.ai. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Copy.ai has the highest mean score of 0.474, followed by Rytr.me with 0.437, ChatSonic with 0.430, HuggingChat with 0.462, and YouChat with the lowest mean of 0.437. The median values are similar to the mean values, with Rytr.me having the highest median score of 0.450. The mode values indicate the most frequently occurring scores, with HuggingChat showing the highest mode at 0.311. In terms of variability, YouChat demonstrates the lowest standard deviation (0.101), while HuggingChat shows the highest (0.080). Skewness values indicate that Rytr.me has the highest positive skewness (0.430), while ChatSonic shows the most negative skewness (-0.121). Kurtosis values reveal that ChatSonic has the lowest kurtosis value (-0.292), indicating a flatter distribution with lighter tails, whereas Copy.ai has the highest kurtosis value (-0.021).

The box plot in Figure 4.5 represents the distribution of subjectivity scores for the five AI chatbots, clearly comparing their score distributions. The box plot highlights that HuggingChat and ChatSonic have the highest median scores and relatively similar variability, as shown by their comparable interquartile ranges (IQRs) and whisker lengths. In contrast, YouChat exhibits the lowest median and smallest IQR, indicating lower variability in subjectivity scores. Additionally, the presence of an outlier in Copy.ai's distribution suggests an instance of a score that deviates significantly from the central tendency of its distribution.

## 4.6 SUBJECTIVITY SCORE FOR PROJECT MANAGEMENT TOOLS

**Table 4.6: Subjectivity Scores for Project Management Tools**

Measures	ClickUp	Notion	Jira
Average(mean)	0.49	0.49	0.5
Median	0.49	0.50	0.58
Mode	0.41	0.50	0.58
Standard Deviation	0.10	0.14	0.14
Skewness	0.32	-1.28	-0.42
Kurtosis	-0.34	2.82	2.60
First Quartile	0.42	0.43	0.48
Third Quartile	0.55	0.56	0.64
Interquartile Range	0.13	0.13	0.16



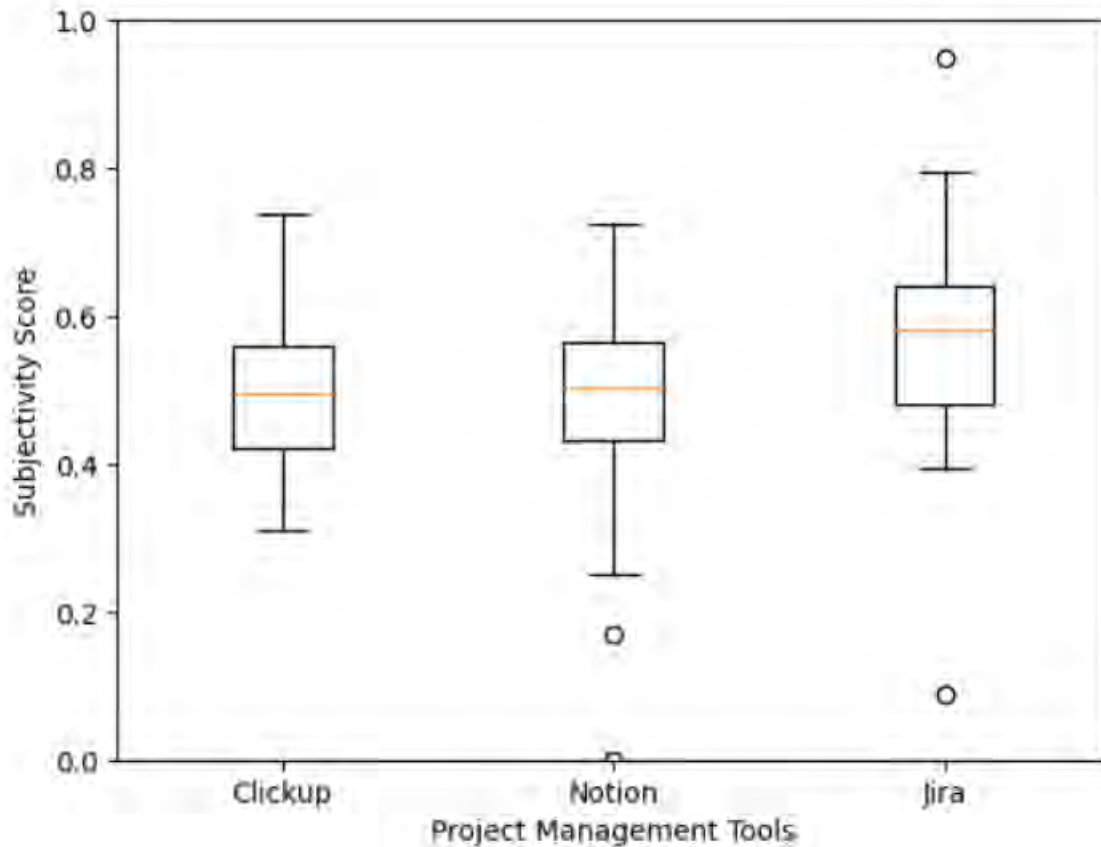


Figure 4.6: Box Plot showing Subjectivity Scores for Project Management Tools

The table (Table 4.6) presents various statistical measures for the subjectivity scores of three project management tools: Clickup, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Notion has the highest mean score of 0.49, followed closely by Jira with 0.5 and Clickup with the lowest mean of 0.49. The median values are identical to the mean values for all three tools. The mode values indicate the most frequently occurring scores, with Notion and Jira sharing the same mode of 0.50, while Clickup has a mode of 0.41. In terms of variability, Jira demonstrates the lowest standard deviation (0.14), while Clickup and Notion have slightly higher values of 0.10 and 0.14, respectively. Skewness values indicate that Jira is the most negatively skewed (-0.42), while Clickup has a positive skewness of 0.32. Kurtosis values are not provided in the table.

The box plot in Figure 4.6 represents the distribution of subjectivity scores for the three project management tools, providing a clear comparison of their score distributions. The box plot highlights that Notion and Jira have similar median scores and interquartile ranges (IQRs), indicating comparable central

tendencies and variability in their subjective scores. Clickup, on the other hand, exhibits a slightly lower median score and a narrower IQR, suggesting lower variability in its subjectivity scores. The presence of an outlier in Jira's distribution suggests an instance of a score that deviates significantly from the central tendency of its distribution.

#### 4.7 PRONOUN RATIO FOR THE AI CHATBOTS

Table 4.7: Pronoun Ratio for AI Chatbots

Measures	ChatSonic	HuggingChat	You	Rytrme	CopyAI
Average(mean)	0.010	0.019	0.01	0.013	0.01
Median	0.008	0.015	0.008	0.010	0.017
Mode	0	0.016	0	0	0
Standard Deviation	0.009	0.013	0.011	0.009	0.011
Skewness	1.60	1.19	3.11	0.71	0.349
Kurtosis	2.35	0.939	13.166	-0.278	-0.87
First Quartile	0.004	0.09	0.004	0.006	0.010
Third Quartile	0.013	0.025	0.014	0.018	0.027
Interquartile Range	0.009	0.015	0.009	0.012	0.016

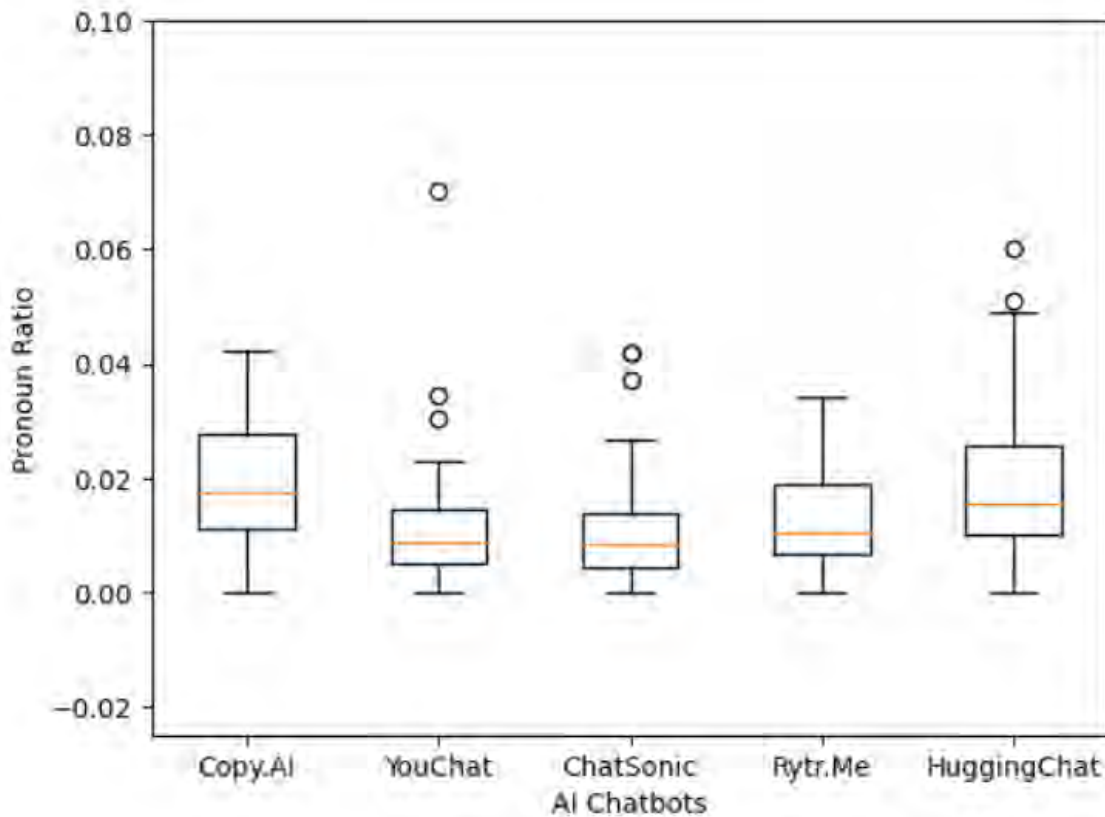


Figure 4.7: Box Plot showing Pronoun Ratio for AI Chatbots

The table (Table 4.7) presents various statistical measures for the pronoun ratio of five AI chatbots: ChatSonic, HuggingChat, YouChat, Rytr.me, and Copy.ai. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Rytr.me has the highest mean pronoun ratio of 0.013, followed by HuggingChat and Copy.ai with 0.019 and 0.01, respectively. ChatSonic and YouChat have the lowest mean pronoun ratios of 0.010 and 0.01, respectively. The median values are similar to the mean values, with HuggingChat having the highest median pronoun ratio of 0.015. The mode values are not provided in the table. In terms of variability, YouChat demonstrates the highest standard deviation (0.011), while Copy.ai and ChatSonic have the lowest (0.009). Skewness values indicate that YouChat is the most positively skewed (3.11), while HuggingChat has the lowest skewness (1.19). Kurtosis values reveal that YouChat has an exceptionally high kurtosis value (13.166), indicating a highly peaked distribution with heavy tails, whereas HuggingChat has the lowest kurtosis value (0.939).

The box plot in Figure 4.7 represents the distribution of pronoun ratios for the five AI chatbots, providing a clear comparison of their ratio distributions. The box plot highlights that the median pronoun ratios for all chatbots are relatively similar, with HuggingChat having a slightly higher median compared to the others. However, the interquartile ranges (IQRs) and the presence of outliers vary among the chatbots. YouChat and ChatSonic exhibit wider IQRs, indicating higher variability in their pronoun ratios, while Copy.ai, Rytr.me, and HuggingChat have narrower IQRs, suggesting lower variability. The presence of outliers in YouChat, ChatSonic, and Rytr.me suggests instances of pronoun ratios that deviate significantly from the central tendency of their respective distributions.

#### 4.8 PRONOUN RATIO FOR PROJECT MANAGEMENT TOOLS

**Table 4.8: Pronoun Ratio for Project Management Tools**

Measures	ClickUp	Notion	Jira
Average(mean)	0.010	0.012	0.016
Median	0.007	0.011	0.008
Mode	0	0	0
Standard Deviation	0.008	0.010	0.019
Skewness	0.523	1.116	1.01
Kurtosis	-0.845	1.053	-0.057
First Quartile	0.003	0.004	0
Third Quartile	0.017	0.019	0.02
Interquartile Range	0.013	0.014	0.028

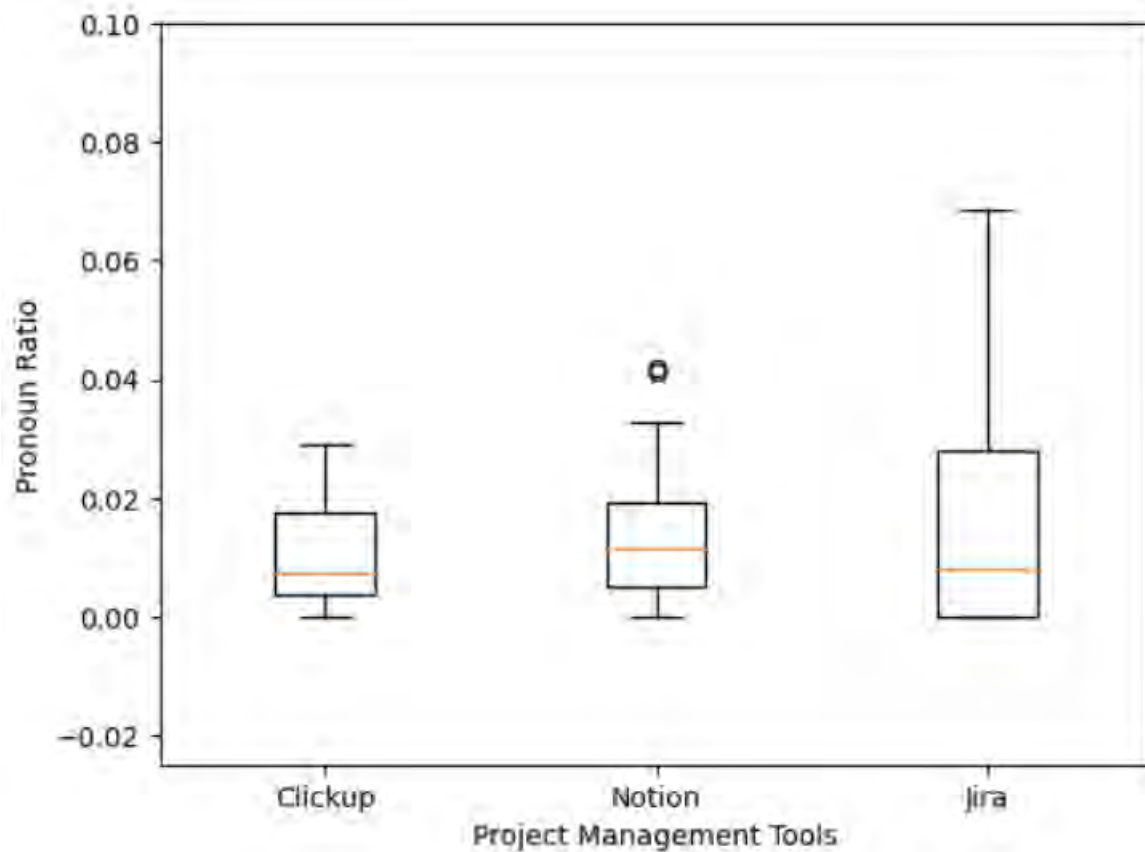


Figure 4.8: Box Pot showing Pronoun Ratio for Project Management Tools

The table (Table 4.8) presents various statistical measures for the pronoun ratio scores of three project management tools: Clickup, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Jira has the highest mean score of 0.016, followed by Notion with 0.012, and Clickup with the lowest mean of 0.010. The median values differ from the mean values, with Notion having the highest median score of 0.011363636. All tools share the same mode of 0. In terms of variability, Jira demonstrates the highest standard deviation (0.019), while Clickup shows the lowest (0.008). Skewness values indicate that all tools have positive skewness, with Notion showing the highest (1.116). Kurtosis values reveal that Notion has a positive kurtosis (1.053), while Clickup and Jira have negative kurtosis values.

The box plot in Figure 4.8 represents the distribution of pronoun ratio scores for the three project management tools, providing a clear comparison of their score distributions. The boxes represent the

interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that Jira has the widest range of scores and the largest IQR, indicating higher variability. Notion shows a slightly smaller range than Jira, with its box positioned higher on the y-axis. Clickup exhibits the smallest range and lowest positioned box. All tools display positive skewness, evidenced by longer upper whiskers. Outliers are visible for Notion and Jira, appearing as points above their respective upper whiskers.

#### 4.9 AVERAGE WORD LENGTH FOR AI CHATBOTS

Table 4.9: AWL for AI Chatbots

Measures	ChatSonic	HuggingChat	You	Rytrme	CopyAI
Average(mean)	6.56	6.28	6.22	6.34	6.92
Median	6.65	6.4	6.26	6.34	6.43
Mode	5.87	5.90	5.5	5.45	5.3
Standard Deviation	0.51	0.48	0.47	0.51	1.62
Skewness	-0.44	-0.76	-0.13	-0.12	2.34
Kurtosis	-0.32	0.88	0.30	-0.21	4.97
First Quartile	6.27	5.99	6.05	6.06	6.21
Third Quartile	6.89	6.60	6.43	6.7	6.84
Interquartile Range	0.62	0.60	0.38	0.62	0.62

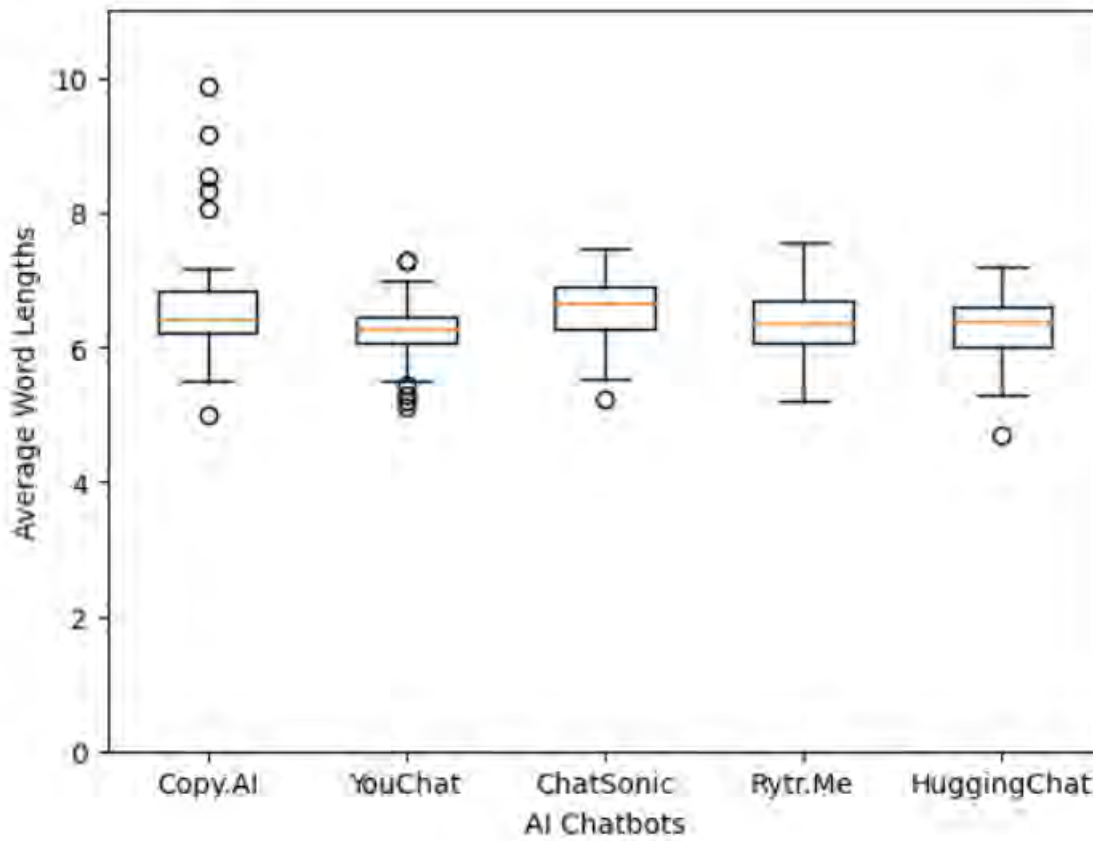


Figure 4.9: Box Plot showing AWL for AI Chatbots

The table (Table 4.9) presents various statistical measures for the average word length (AWL) of five AI chatbots: ChatSonic, HuggingChat, You, Rytrme, and CopyAI. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that CopyAI has the highest mean AWL of 6.92, followed by ChatSonic with 6.56, Rytrme with 6.34, HuggingChat with 6.28, and You with the lowest mean of 6.22. The median values align closely with the mean values, with ChatSonic having the highest median score of 6.65. The mode values indicate the most frequently occurring AWL, with ChatSonic showing the highest mode (5.87). In terms of variability, CopyAI demonstrates the highest standard deviation (1.62), while You shows the lowest (0.47). Skewness values indicate that HuggingChat has the highest negative skew (-0.76), while CopyAI shows significant positive skewness (2.34). Kurtosis values reveal that CopyAI has the highest kurtosis (4.97), indicating a distribution with heavier tails, while ChatSonic has the lowest kurtosis value (-0.32), suggesting a flatter distribution.

The box plot in Figure 4.9 represents the distribution of average word lengths for the five AI chatbots, clearly comparing their AWL distributions. The boxes represent the interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that Copy.AI has the highest median and the largest variability, as shown by its tall box and long whiskers. In contrast, You exhibits the lowest variability, evidenced by its narrow IQR and shorter whiskers. Additionally, the presence of outliers in several chatbots, particularly noticeable for Copy.AI and YouChat, indicates instances of AWL that deviate significantly from the central tendency of their distributions.

#### 4.10 AVERAGE WORD LENGTH FOR PROJECT MANAGEMENT TOOLS

**Table 4.10: AWL for Project Management Tools**

Measures	ClickUp	Notion	Jira
Average(mean)	6.362	6.422	6.652
Median	6.469	5.741	6.76
Mode	6.469	5.315	6.806
Standard Deviation	0.541	4.326	0.574
Skewness	-0.259	5.454	-0.355
Kurtosis	-0.449	28.54	-0.221
First Quartile	6.083	5.315	6.377
Third Quartile	6.732	6.164	7.070
Interquartile Range	0.649	0.849	0.692



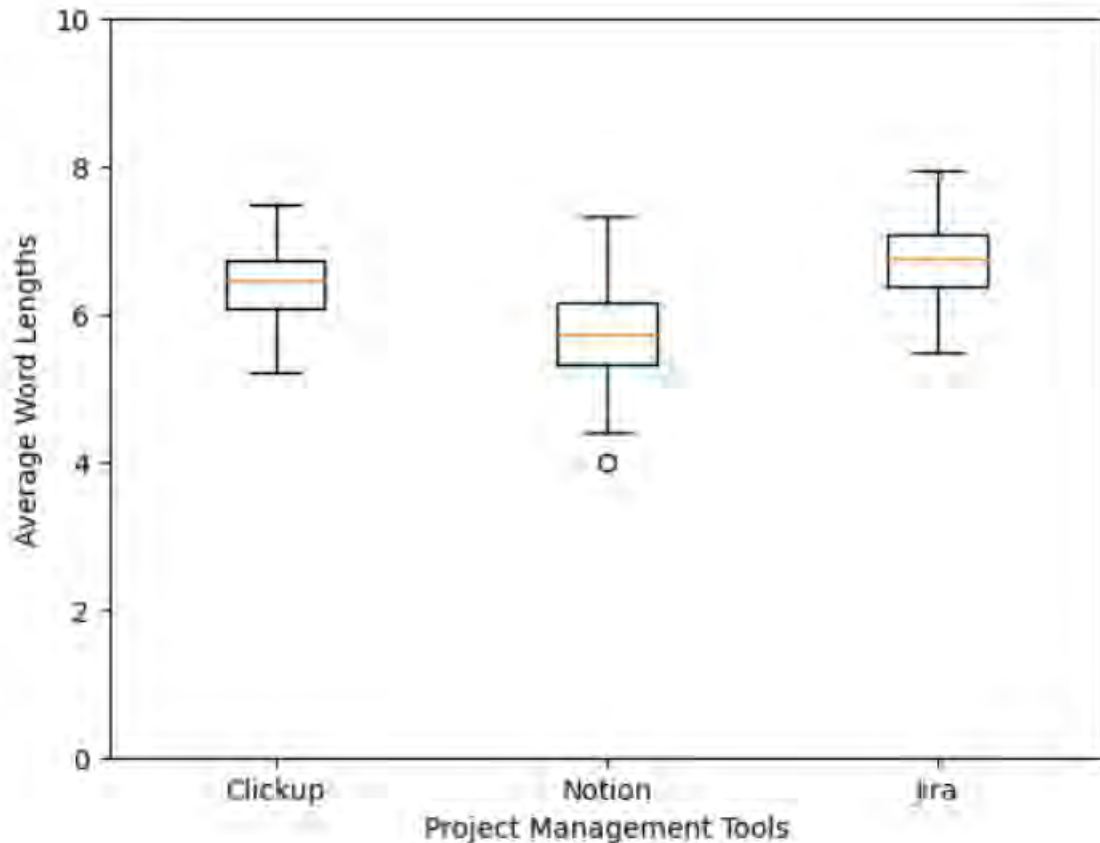


Figure 4.10: Box Plot showing AWL for Project Management Tools

The table (Table 4.10) presents various statistical measures for the average word length (AWL) of three project management tools: ClickUp, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Jira has the highest mean AWL of 6.652, followed by ClickUp with 6.362, and Notion with the lowest mean of 6.422. The median values differ from the mean, with Jira having the highest median score of 6.76. The mode values indicate that Jira has the highest, most frequently occurring AWL (6.806). In terms of variability, Jira demonstrates the highest standard deviation (0.574), while Notion shows the lowest (4.326). Skewness values indicate that Notion has a significant positive skew (5.454), while ClickUp and Jira show negative skewness. Kurtosis values reveal that Notion has an extremely high kurtosis (28.54), indicating a distribution with heavy tails, while ClickUp and Jira have negative kurtosis values, suggesting flatter distributions.

The box plot in Figure 4.10 represents the distribution of average word lengths for the three project management tools, providing a clear comparison of their AWL distributions. The boxes represent the

interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that Jira has the highest median and a relatively large variability, as shown by its tall box and long whiskers. ClickUp shows a more compact distribution with shorter whiskers. Notion exhibits the widest IQR, indicating high variability within the middle 50% of its data, and also displays an outlier below its lower whisker, suggesting the presence of unusually low AWL values in its distribution.

#### 4.11 READABILITY SCORES (FLESCH - KINCAID GRADE LEVEL) FOR AI CHATBOTS

Table 4.11: Flesch Kincaid Grade Level for AI Chatbots

Measures	ChatSonic	HuggingChat	You	Rytrme	CopyAI
Average(mean)	2.121	20.85	16.69	13.704	14.300
Median	-1.64	19.685	14.035	13.075	17.69
Mode	37.5	10.91	9.42	55.74	17.64
Standard Deviation	18.39	11.83	16.23	18.37	23.15
Skewness	0.86	0.33	0.47	0.20	-1.30
Kurtosis	-0.03	-0.29	-0.09	-0.14	3.33
First Quartile	-12.23	10.91	4.7475	0.495	6.41
Third Quartile	13.18	28.09	27.20	26.02	25.7
Interquartile Range	25.41	17.18	22.45	25.52	19.28

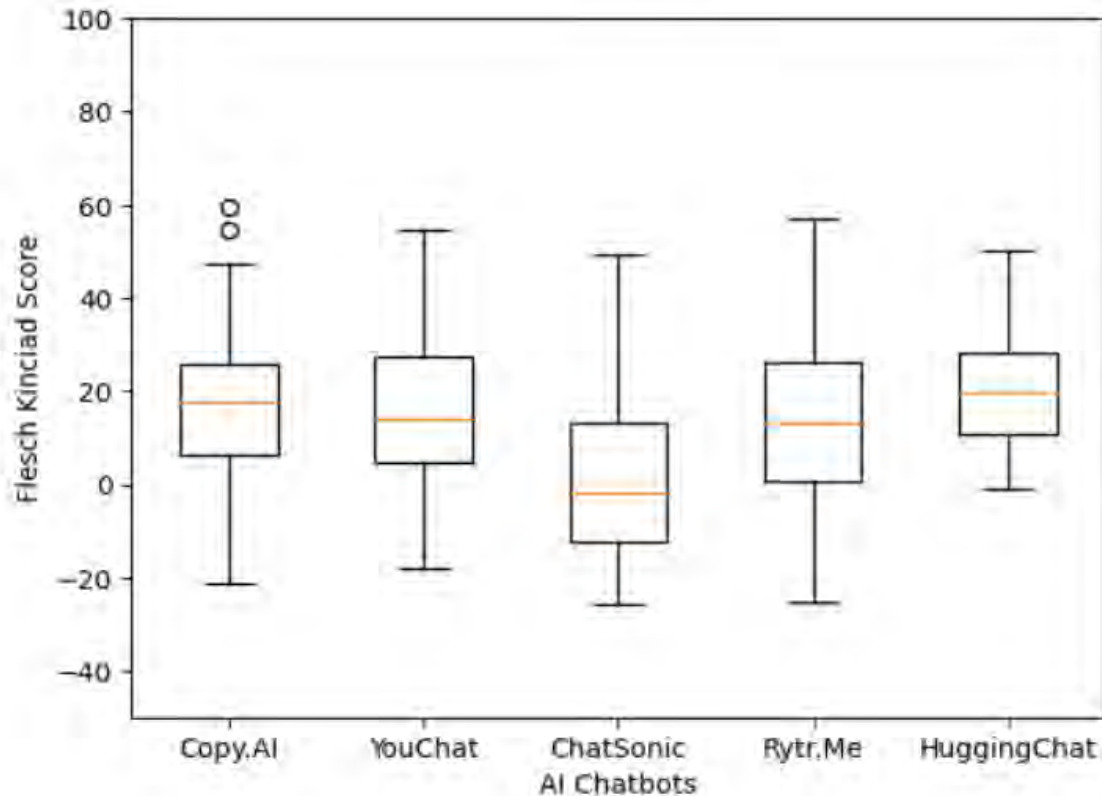


Figure 4.11: Box Plot showing Flesch Kincaid Grade Level for AI Chatbots

The table (Table 4.11) presents various statistical measures for the Flesch-Kincaid Grade Level readability scores of five AI chatbots: ChatSonic, HuggingChat, You, Rytrme, and CopyAI. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that HuggingChat has the highest mean score of 20.85, followed by ChatSonic with 2.1216, You with 16.69, CopyAI with 14.58, and Rytrme with the lowest mean of 13.7048. The median values differ significantly from the means, with HuggingChat having the highest median score of 19.685 and ChatSonic having the lowest at -1.64. The mode values vary widely, with Rytrme showing the highest mode (55.74) and You the lowest (9.42). In terms of variability, CopyAI demonstrates the highest standard deviation (23.15), while HuggingChat shows the lowest (11.83). Skewness values indicate that ChatSonic has the highest positive skew (0.86), while CopyAI shows significant negative skewness (-1.30). Kurtosis values reveal that CopyAI has the highest kurtosis (8.33),

indicating a distribution with heavier tails, while ChatSonic has the lowest kurtosis value (-0.03), suggesting a nearly normal distribution.

The box plot in Figure 4.11 represents the distribution of Flesch-Kincaid Grade Level scores for the five AI chatbots, clearly comparing their score distributions. The boxes represent the interquartile range (IQR), containing the middle 50% of the data, with the bottom and top edges representing the first quartile (Q1) and the third quartile (Q3), respectively. The line inside each box denotes the median (second quartile, Q2). The whiskers extend to 1.5 times the IQR from Q1 and Q3, representing the range of the data excluding outliers, which are displayed as individual points outside the whiskers. The box plot highlights that HuggingChat has the highest median and a relatively compact distribution. ChatSonic shows the widest IQR, indicating high variability within the middle 50% of its data. Copy.AI and YouChat display similar distributions with moderate variability. Rytr.me exhibits a large range with its box positioned lower than the others. Outliers are visible for several chatbots, particularly for Copy.AI and Rytr.me, indicating scores that deviate significantly from the central tendency of their distributions.

## 4.12 READABILITY SCORES (FLESCH - KINCAID GRADE LEVEL) FOR PROJECT MANAGEMENT

### TOOLS

**Table 4.12: Flesch Kincaid Grade Level for Project Management Tools**

Measures	ClickUp	Notion	Jira
Average(mean)	23.31	26.01	22.61
Median	25.76	41.16	22.61
Mode	21.23	31.28	13.34
Standard Deviation	31.79	76.86	16.66
Skewness	-2.49	-4.68	-0.31
Kurtosis	6.57	22.27	0.69
First Quartile	18.1	32.05	13.34
Third Quartile	42.12	51.51	31.01

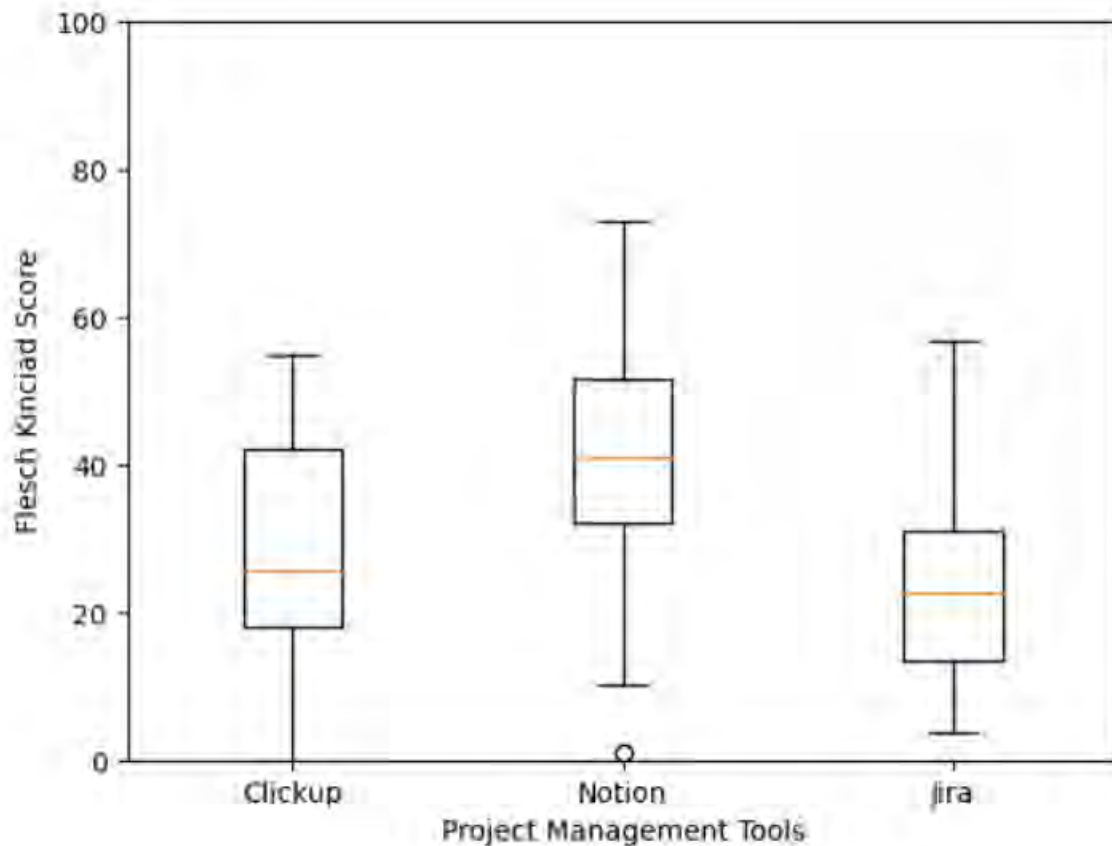


Figure 4.12: Box Plot showing Flesch Kincaid Grade Level for Project Management Tools

The table (Table 4.12) presents various statistical measures for the Flesch-Kincaid Grade Level scores of three project management tools: Clickup, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, and third quartile. The average scores show that Notion has the highest mean score of 26.01, followed by Clickup with 23.31, and Jira with the lowest mean of 22.61. The median values differ significantly from the means, with Notion having the highest median score of 41.16. The mode values indicate the most frequently occurring scores, with Notion showing the highest mode (31.28). In terms of variability, Notion demonstrates the highest standard deviation (76.86), while Jira shows the lowest (16.66). Skewness values indicate that all tools have negative skewness, with Notion having the most extreme value (-4.68). Kurtosis values reveal that Notion has a high kurtosis (22.27), indicating a distribution with heavy tails.

The box plot in Figure 4.12 represents the distribution of Flesch-Kincaid Grade Level scores for the three project management tools, providing a clear comparison of their score distributions. The plot shows that

Notion has the widest spread and highest median, while Jira has the narrowest spread and lowest median. Clickup falls between these two in terms of both spread and median scores. Notion also displays outliers below its lower whisker, indicating some extremely low scores in its distribution.

#### 4.13 READABILITY SCORE (COLEMAN LIAU INDEX) FOR AI CHATBOTS

**Table 4.13: Coleman Liau Index Score for AI Chatbots**

Measures	ChatSonic	HuggingChat	You	Rytr.me	CopyAI
Average(mean)	20.117	18.121	18.027	18.473	20.729
Median	20.57	18.525	18.135	18.45	18.79
Mode	21.42	16.13	18.63	18.68	18.27
Standard Deviation	3.01	2.42	2.61	2.86	7.36
Skewness	-0.55	-0.46	-0.35	-0.12	2.32
Kurtosis	-0.36	-0.32	0.22	-0.27	5.13
First Quartile	18.7	16.67	17.13	17.01	17.73
Third Quartile	22.23	19.84	19.58	20.73	20.39
Interquartile Range	3.53	3.16	2.44	3.71	2.66

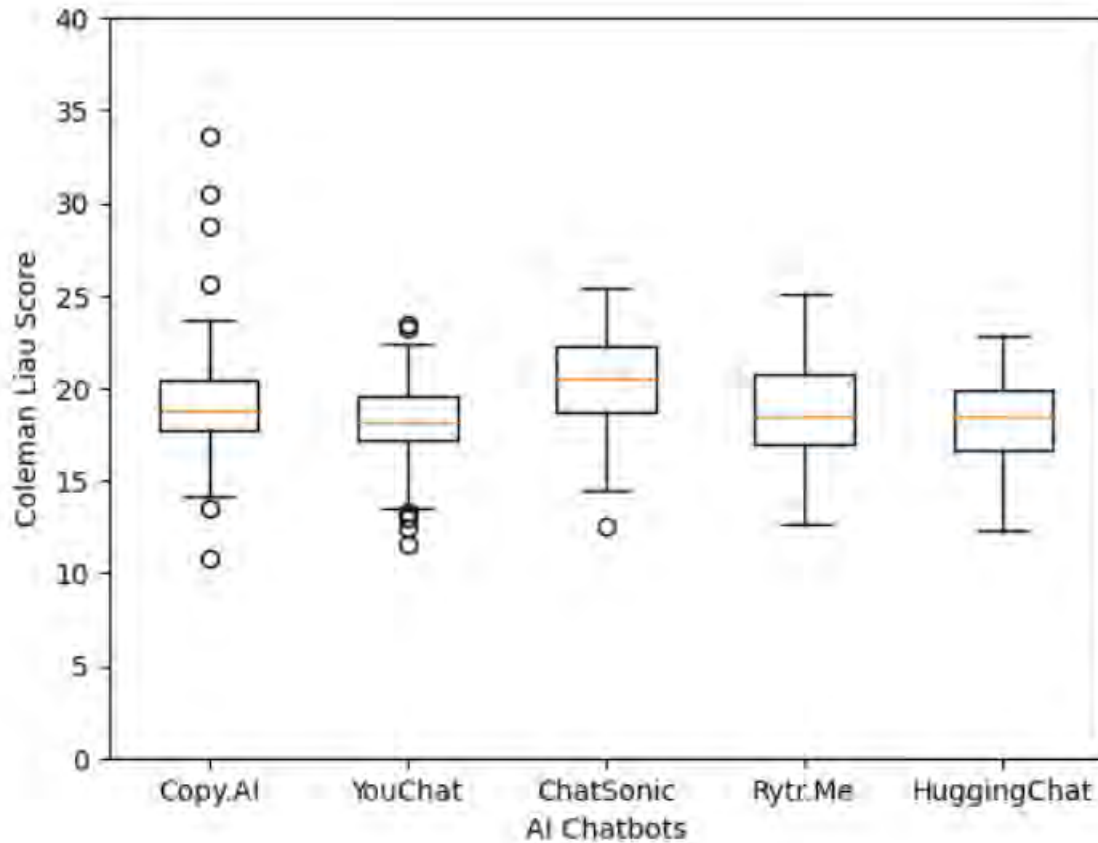


Figure 4.13: Box Plot showing Coleman Liau Index results for readability on AI Chatbots

The table (Table 4.13) presents various statistical measures for the Coleman Liau Index scores of five AI chatbots: ChatSonic, HuggingChat, You, Rytrme, and CopyAI. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that ChatSonic has the highest mean score of 20.1172, followed by CopyAI with 20.729, You with 18.0272, Rytrme with 18.4734, and HuggingChat with the lowest mean of 18.1216. The median values align closely with the mean values, with ChatSonic again having the highest median score of 20.57. The mode values indicate the most frequently occurring scores, with ChatSonic showing the highest mode (21.42). In terms of variability, CopyAI demonstrates the highest standard deviation (7.36), while HuggingChat shows the lowest (2.42). Skewness values indicate that CopyAI has a significant right skew (2.32), while the other chatbots have slight left skews. Kurtosis values reveal that CopyAI has the highest kurtosis (5.13), indicating a distribution with heavier tails.

The box plot in Figure 4.13 represents the distribution of Coleman Liau Index scores for the five AI chatbots, providing a clear comparison of their score distributions. The plot shows that ChatSonic has the highest median and a relatively narrow interquartile range. CopyAI exhibits the widest spread and several outliers above its upper whisker, indicating some extremely high scores in its distribution. HuggingChat, You, and Rytrme have similar median scores and distributions, with Rytrme showing a slightly larger spread than the other two. The presence of outliers in CopyAI and YouChat indicates instances of scores that deviate significantly from the central tendency of their distributions.

#### 4.14 READABILITY SCORE (COLEMAN LIAU INDEX) FOR PROJECT MANAGEMENT TOOLS

**Table 4.14: Coleman Liau Index for Project Management Tools**

Measures	ClickUp	Notion	Jira
Average(mean)	18.70	17.33028571	20.06828571
Median	19.32	15.25	20.34
Mode	20.11	12.98	21.75
Standard Deviation	3.34	13.71	3.193236866
Skewness	-0.085	5.15	-0.228431727
Kurtosis	0.25	26.31	0.146985385
First Quartile	16.87	13.19	18.405
Third Quartile	20.72	16.40	21.83
Interquartile Range	3.85	3.21	3.425



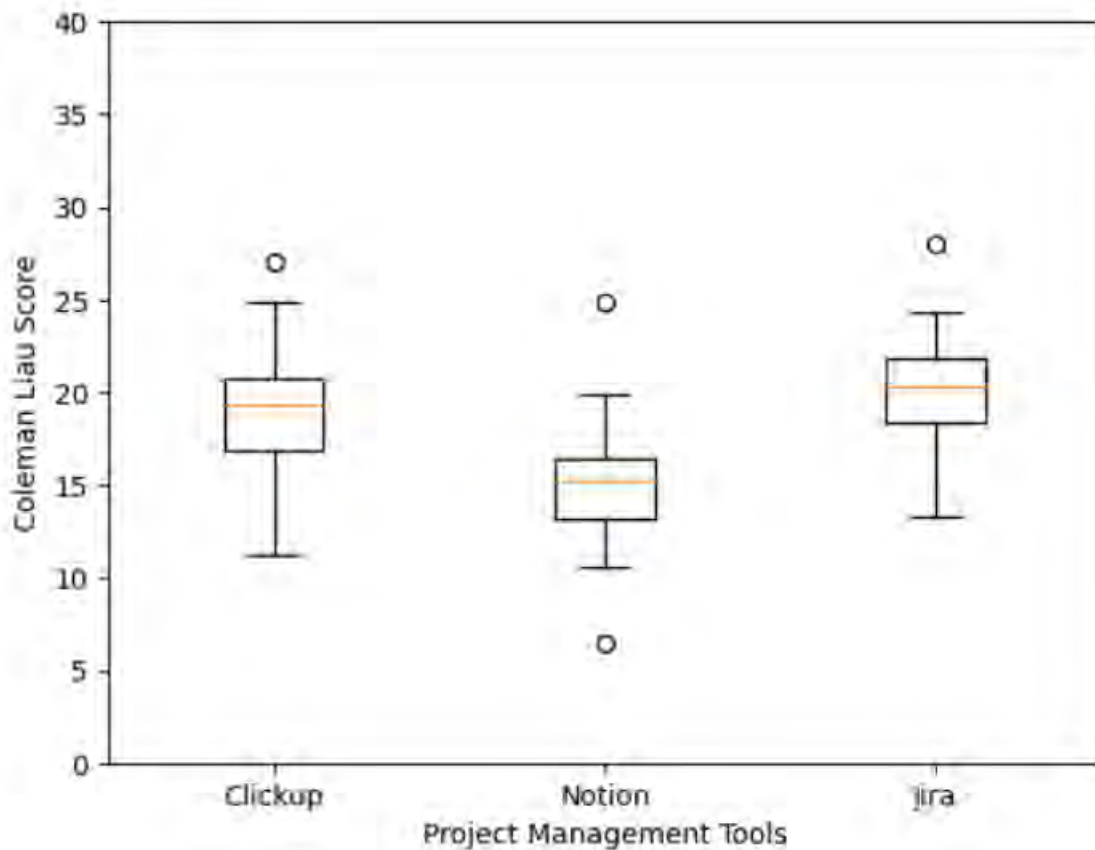


Figure 4.14: Box Plot showing Coleman Liau Score for Project Management Tools

The table (Table 4.13) presents various statistical measures for the Coleman Liau Index scores of three project management tools: ClickUp, Notion, and Jira. The measures include the average (mean), median, mode, standard deviation, skewness, kurtosis, first quartile, third quartile, and interquartile range. The average scores show that Jira has the highest mean score of 20.06828571, followed by ClickUp with 18.70, and Notion with the lowest mean of 17.33028571. The median values align somewhat with the mean values, with Jira having the highest median score of 20.34. The mode values indicate the most frequently occurring scores, with Jira showing the highest mode (21.75). In terms of variability, Notion demonstrates the highest standard deviation (13.71), while ClickUp shows the lowest (3.34). Skewness values indicate that Notion has a significant right skew (5.15), while ClickUp has a slight left skew (-0.085). Kurtosis values reveal that Notion has the highest kurtosis (26.31), indicating a distribution with much heavier tails than the others.

The box plot in Figure 4.13 represents the distribution of Coleman Liau Index scores for the three project management tools, providing a clear comparison of their score distributions. The plot shows that Jira has the highest median and a relatively narrow interquartile range. ClickUp exhibits a wider spread than Jira but a similar median. Notion has the lowest median and shows the widest spread of the three, with outliers both above and below its whiskers, indicating some extreme scores in its distribution. The presence of these outliers in Notion's data suggests instances of scores that deviate significantly from the central tendency of its distribution.

#### **4.15 DATA UTILISATION IN AI FEATURES OF PROJECT MANAGEMENT TOOLS**

The evaluation of ClickUp, Jira, and Notion based on data utilisation reveals distinct differences in how each tool integrates project data into generated responses. By examining the inclusion of names, dates, specific task details, and other project-specific information, we can see varying levels of effectiveness in data utilisation.

ClickUp demonstrates a high level of data utilisation. The responses generated by ClickUp consistently include specific names of project team members and stakeholders, providing a personalised touch to communications. Additionally, ClickUp effectively incorporates relevant dates from the project timeline, such as deadlines and milestones, ensuring that the generated content is contextually accurate and timely. The tool also excels in integrating detailed task information, including task names, descriptions, and statuses, which helps in providing comprehensive and actionable responses. Furthermore, ClickUp includes other project-specific details, such as project objectives and key deliverables, making the responses highly relevant and specific to the project context. This thorough integration of project data highlights ClickUp's strength in utilising existing project information to generate detailed and relevant responses.

Jira offers a moderate level of data utilisation. While the tool provides structured templates with placeholders for names and dates, it requires users to manually input these details. This approach ensures that responses are well-organized and standardised, but it places the burden on users to fill in the contextual information. Consequently, the level of specificity in the responses depends on the accuracy and completeness of the manual data entry. Jira includes some task details, but these are often not as comprehensive as those found in ClickUp. The responses may include basic information about tasks, but they might lack the depth and specificity that comes from automatic data integration. Despite this, Jira's

structured approach ensures that the generated content is organised and follows a logical format, making it relatively easy for users to complete and customise the responses.

Notion exhibits a lower level of data utilisation than ClickUp and Jira. The responses generated by Notion tend to be more generalised and less reliant on specific project data. While Notion occasionally generates names and sets arbitrary dates, these details are not consistently relevant to the specific project context. The tool often produces generic responses that require significant manual editing to incorporate accurate names, dates, and task details. This lack of specific data integration means that the responses may not be immediately useful without further customisation. However, Notion's strength lies in its flexibility and versatility, allowing users to adapt the responses to a wide range of use cases. Despite its general approach, Notion can be a valuable tool for users who need a flexible and adaptable platform for various project management tasks.

**Table 4.15: Project Management Tool's Data Utilisation Factors from the Already Created Projects**

<b>Data Utilisation Factors</b>	<b>ClickUp</b>	<b>Jira</b>	<b>Notion</b>
<b>Names</b>	Yes	No	No
<b>Dates and Time</b>	Yes	Yes, Generalised	Random time
<b>Task details</b>	Yes	Yes	No

In summary, ClickUp excels in data utilisation by consistently incorporating specific project data into its responses, making them highly relevant and detailed. Jira offers a balanced approach with structured templates that require manual data entry, resulting in moderately specific responses. While flexible and versatile, Notion tends to produce more generalised responses that often need significant manual customisation to ensure relevance and accuracy. This comparison highlights each tool's varying strengths and weaknesses in terms of data integration and utilisation, providing insights into their suitability for different project management needs.

#### **4.16 INTEGRATED FEATURES IN PROJECT MANAGEMENT TOOLS**

While we can compare the Project Management tools on the basis of integration features, the features offered by Notion, ClickUp, and Jira are here in the Table as follows:

**Table 4.16: Features Integrated in AI-enhanced Project Management Tools**

Features	ClickUp	Jira	Notion
<b>Ask AI to Write</b>	Yes	Yes	Yes
<b>Continue Writing</b>	Yes	No	Yes
<b>Summarise</b>	No	No	Yes
<b>Find action items</b>	No	No	Yes
<b>Translate</b>	Through prompts	Through prompts	Yes
<b>Explain</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Improve writing</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Fix spelling and grammar.</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Make shorter</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Make longer</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Change tone</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Simplify language</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Ask AI to edit</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Brainstorm ideas</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Blog post</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Outline</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Social media post</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Press release</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Creative story</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Essay</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Poem</b>	Through prompts	Through prompts (exclusively)	Yes
<b>To-do list</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Meeting agenda</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Pros and cons list</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Jobs description</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Sales email</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Recruiting email</b>	Through prompts	Through prompts (exclusively)	Yes
<b>Write StandUp</b>	Yes	Through prompts (exclusively)	Through prompts

Table 4.15 outlines the features integrated into the AI capabilities of three project management tools: ClickUp, Jira, and Notion. The table comprehensively compares how each tool utilises AI to enhance various writing and productivity tasks. All three tools offer the basic "Ask AI to Write" feature, enabling users to generate content using AI. ClickUp and Notion provide additional functionality with features like "Continue Writing", which allows the AI to pick up where the user left off, a capability not available in Jira. Notion stands out for its extensive AI integration, offering unique features such as "Summarise", "Find action items", and "Translate," which are not inherently available in ClickUp or Jira. While ClickUp and Jira

rely on prompts to access a range of AI features, Notion incorporates these features more directly, enhancing usability. These features include improving writing, fixing spelling and grammar, adjusting content length and tone, simplifying language, and editing text. Moreover, Notion supports creative tasks such as brainstorming ideas, writing blog posts, creating social media posts, and drafting various forms of written communication, from essays and poems to meeting agendas and sales emails. ClickUp provides direct access to writing stand-ups, a feature that Jira only supports through prompts. This detailed comparison highlights Notion's broader AI capabilities, making it a versatile tool for users who require extensive AI-driven writing and productivity features.

## 4.17 DISCUSSION

This section summarises the findings from our analysis of AI chatbots and project management tools, focusing on their respective evaluation metrics. The evaluation of AI chatbots using text mining techniques revealed significant insights into their linguistic capabilities and performance. Key metrics analysed were lexical diversity, sentiment, subjectivity, pronoun usage, average word length, and readability scores. HuggingChat demonstrated the highest lexical diversity with a mean score of 0.649, indicating it offers the most varied vocabulary in its responses. Rytr.me showed the highest average sentiment score (0.626), suggesting it maintains a consistently positive tone in interactions. Copy.ai exhibited the highest subjectivity score (0.474) and, along with HuggingChat, showed the highest pronoun ratios (both at approximately 0.019), indicating a more conversational and potentially opinionated style. In terms of language complexity, Copy.ai led with the highest average word length (6.922), while HuggingChat scored highest on the Flesch-Kincaid Grade Level (20.852), suggesting more complex sentence structures.

The evaluation of project management tools focused on assessing documentation, summarisation, and writing capabilities using similar text-mining techniques. Jira consistently demonstrated high performance across multiple metrics, showing the highest lexical diversity (0.789), sentiment score (0.794), subjectivity score (0.567), pronoun ratio (0.017), and average word length (6.653) among the project management tools. These results suggest that Jira offers a more varied vocabulary, maintains a positive tone, provides more interpretive content, and uses slightly more complex language in its generated content. However, in terms of readability, Notion scored highest on the Flesch-Kincaid Grade Level (26.013), indicating more complex sentence structures, while Jira led in the Coleman-Liau Index (20.068), suggesting a balance between word and sentence complexity.

These findings provide valuable insights into the linguistic characteristics and performance of AI chatbots and project management tools. The results highlight the strengths and potential areas for improvement in each category. For AI chatbots, there's a notable variation in performance across different linguistic aspects, with different chatbots excelling in different areas. This suggests that users might choose a chatbot based on their specific needs, whether it's for varied vocabulary, positive sentiment, or more complex language. For project management tools, Jira's consistently high performance across multiple metrics indicates its strong capability in generating diverse, positive, and moderately complex content, which could be beneficial for various project management tasks. However, the higher readability scores across all project management tools suggest that their generated content might be more challenging to read, possibly due to the technical nature of project management terminology. These insights offer a

foundation for future research and development in these rapidly evolving technologies, potentially guiding natural language generation improvements for AI chatbots and project management tools.

The ClickUp was found to be the best integrated and best in using the data when returning the responses, while, Jira ranks second and Notion ranked last. When discussing the number of features available, Notion ranked top, while Jira and ClickUp didn't provide many features. Although their chatbot integration allows them to perform any task Notion exclusively has. It can be seen that HuggingChat and Rytr.me responses are comparatively lengthy in the AI Chatbots category, and in terms of project management tools, ClickUp ranks the lengthiest, followed by Notion and Jira. This also means that all the tools, or chatbots, are customised in their own way. All the scores determined from text mining techniques show which chatbot or project management tool has the best sentiments, subjectivity scores, pronoun ratio, average word length, and readability score.

#### **4.17.1 LIMITATIONS**

While providing valuable insights into the performance of AI chatbots and project management tools, this study is subject to several limitations. Firstly, the research is constrained by the specific set of AI chatbots and project management tools selected for analysis. While efforts were made to include popular and representative applications, the findings may not be generalisable to all AI-powered tools in these categories. Secondly, the evaluation metrics, while comprehensive, primarily focus on text-based characteristics and may not capture all aspects of user experience or practical utility. The dynamic nature of AI technology presents another limitation; the rapid pace of development in this field means that the performance of these tools may have evolved since data collection. Additionally, the study's reliance on predefined prompts may not fully represent the diverse range of real-world use cases. Lastly, the analysis does not account for potential biases in the training data of the AI models, which could influence the generated responses.

#### **4.17.2 FUTURE RESEARCH & DEVELOPMENT**

Future research in this area could expand upon this study in several ways. One promising direction would be to incorporate a wider range of AI chatbots and project management tools, including those utilising newer AI models like GPT-4, Claude 3, or Gemini. Longitudinal studies tracking the performance of these tools over time could provide insights into the rate of improvement and evolving capabilities of AI in these domains. Additionally, future work could explore more domain-specific applications of these tools, such as their performance in specialised industries or for specific types of projects. Integrating user experience

studies and qualitative feedback alongside quantitative metrics could offer a more holistic evaluation of these tools' effectiveness. Research into the adaptability and personalisation capabilities of these AI tools could also yield valuable insights. Finally, comparative studies between AI-powered tools and traditional non-AI alternatives could help quantify the added value of AI integration in these applications.

#### **4.17.3 ETHICAL CONSIDERATIONS**

While this study primarily focused on performance metrics, it is crucial to address the ethical implications of widespread AI adoption in chatbots and project management tools. Future research should explore issues of data privacy and security, particularly in project management contexts where sensitive information may be processed. The potential for AI to perpetuate or amplify biases present in training data is another critical area for ethical consideration. Researchers should investigate methods to detect and mitigate such biases in AI-generated content. Additionally, the impact of these tools on human labour and job roles in project management and customer service sectors warrants careful study. Transparency in AI decision-making processes and the ability to explain AI-generated outputs are also important ethical considerations that future research should address. Lastly, the long-term effects of increased human-AI interaction on communication patterns and cognitive processes in professional settings present an intriguing and important area for ethical inquiry.



## CHAPTER 5: CONCLUSION

This research aimed to evaluate and compare generative AI chatbots and their applications in project management tools, addressing a significant gap in the literature: the lack of comprehensive, quantitative comparisons of AI chatbot applications using text mining techniques. Our study hypothesised that applying a range of text mining metrics could provide a nuanced, data-driven assessment of these AI-powered tools' performance.

The research process involved several key stages. We began by selecting various AI chatbots (Copy.ai, HuggingFace, ChatSonic, You Chat, Rytr.me) and project management tools (Click Up, Notion, Jira) based on their market presence and feature sets. Our methodology centred on developing standardised prompts to generate responses from each tool, which we then analysed using various text-mining techniques. These techniques included sentiment analysis, lexical diversity measurement, readability scoring, pronoun ratio calculation, and average word length.

Our experiments involved systematically applying these prompts and analysis methods across all selected tools, ensuring a consistent and fair comparison. This approach allowed us to identify distinct patterns and characteristics in the performance of each tool. While detailed findings are presented in the previous chapter, our high-level results revealed significant variations in performance across different metrics. Each tool demonstrated unique strengths and weaknesses, highlighting the importance of aligning tool selection with specific use case requirements.

While addressing the research questions, how can AI chatbots' responses be analysed using text-mining techniques? So, from the methodology we adopted in one of the previous sections, we were able to apply various text-mining techniques and derive sentiment scores, subjectivity scores, average word lengths, pronoun ratios, and readability scores using the flesch-kincaid method and the coleman-liau index. Similarly, when addressing how can AI-enhanced project management tools' responses be analysed using text-mining techniques, from the methodology, we were able to apply the same text-mining techniques to the responses generated from their AI feature and derived sentiment scores, subjectivity scores, average word lengths, pronoun ratios, and readability scores using flesch-kincaid method and coleman-liau index. As per RQ3, how well are these AI-enhanced project management tools integrated with their AI features? For project management tools, I've worked on additional comparison including the data utilisation way which includes how these chatbots are utilising the data from the projects which are already created, and also the integration features which each tool offers within their AI feature.

This research contributes to the field by providing a structured framework for evaluating AI-powered text generation tools. Our findings offer valuable insights for both developers and users of these technologies, potentially influencing future development and selection processes in various industries.

However, it's important to acknowledge the limitations of this study. Our analysis was confined to a specific set of tools and metrics, and the rapidly evolving nature of AI technology means that tool capabilities may change quickly. Future research could expand on this work by incorporating a broader range of tools, exploring additional metrics, and conducting longitudinal studies to track the evolution of these technologies over time.

In conclusion, answering the research questions, as AI continues transforming various sectors, including project management, frameworks like the one developed in this study are crucial for understanding and optimising these powerful tools. The ClickUp was the best integrated for using the data when returning the responses, while Jira ranks second and Notion ranks last. When discussing the number of features available, Notion ranked top, while Jira and ClickUp didn't provide many features. Although their chatbot integration allows them to perform any task Notion exclusively has. It can be seen that HuggingChat and Rytr.me responses are comparatively lengthy in the AI Chatbots category, and in terms of project management tools, ClickUp ranks the lengthiest, followed by Notion and Jira. This also means that all the tools, or chatbots, are customised in their own way. All the scores determined from text mining techniques show which chatbot or project management tool has the best sentiments, subjectivity scores, pronoun ratio, average word length, and readability score.

This research lays the groundwork for more informed decision-making in implementing AI technologies and opens up new avenues for exploration in this dynamic field.

## REFERENCES

- [1] P. Yu, H. Xu, X. Hu and C. Deng, "Leveraging Generative AI and Large Language Models: A Comprehensive Roadmap for Healthcare Integration," *Healthcare (Basel)*, vol. 11, p. 2776, 2023.
- [2] A. Gupta, V. Dengre, H. A. Kheruwala and M. Shah, "Comprehensive review of text-mining applications in finance," *Journal of Financial Innovation*, vol. 6, no. 1, 02 November 2020.
- [3] K. Knill and S. Young, "Hidden Markov Models in Speech and Language Processing," *Text, Speech and Language Technology. Springer Netherlands*, p. 27–68, 1997.
- [4] D. Reynolds, "Gaussian Mixture Models," *Encyclopedia of Biometrics. Springer US*, p. 827–832, 2015.
- [5] R. Ducharme, Y. Bengio and P. Vincent, "Neural Probabilistic Language Model," *JMLR.org*, vol. 3, p. 1137–1155, 2003.
- [6] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech, ICSA, 2010*, pp. 1045-1048.
- [7] R. Dey and F. M. Salem, "Gate-variants of Gated Recurrent Unit (GRU) neural networks," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, IEEE, 2017, pp. 1597-1600.
- [8] U. Khandelwal, H. He, Q. Peng and J. Daniel, "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. M. Iryna Gurevych, Ed., Melbourne, Association for Computational Linguistics, 2018, pp. 284-294.
- [9] A. Efros and T. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1033-1038.
- [10] P. S. Heckbert, "Survey of Texture Mapping," *IEEE Computer Graphics and Applications*, vol. 6, no. 11, pp. 56-67, 1986.
- [11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," *Association for Computing Machinery*, vol. 63, p. 11, 2014.
- [12] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, 2013.

- [13] Z. Li, Y. Chen and F. T. Sommer, "Learning Energy-Based Models in High-Dimensional Spaces with Multiscale Denoising-Score Matching," *Entropy*, vol. 25, no. 10, 2023.
- [14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. u. Kaiser and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017.
- [15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen and I. Sutskever, "Zero-Shot Text-to-Image Generation," vol. 139, p. 8821–8831, 2021.
- [16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprint arXiv:1910.13461*, 2019.
- [17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2021.
- [18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, IEEE Computer Society, 2021, pp. 9992-10002.
- [19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, M. Meila and T. Zhang, Eds., PMLR, 2021, p. 8748–8763.
- [20] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify and H. Awadalla, "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation," *arXiv preprint arXiv:2302.09210*, February 2023.
- [21] J. Ho, A. Jain and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Advances in neural information processing systems*, vol. 33, p. 6840–6851, 2020.
- [22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, "Language Models are Few-Shot Learners," vol. 33, p. 1877–1901, 2020.

- [23] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie and J.-R. Wen, "A Survey of Large Language Models," *arXiv preprint arXiv:2303.18223*, 2023.
- [24] S. Osindero and M. Mirza, "Conditional Generative Adversarial Nets," *arXiv preprint arXiv:1411.1784*, 2014.
- [25] T. Karras, S. Laine and T. Aila, "A style-based generator architecture for generative adversarial networks," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, vol. 43, no. 12, pp. 4401-4410, Dec 2021.
- [26] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the IEEE international conference on computer vision*, pp. 2242-2251, 2017.
- [27] A. Brock, J. Donahue and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
- [28] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng and S. Z. Li, "A Survey on Generative Diffusion Models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 2814-2830, 2024.
- [29] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade and Y. a. o. Lu, "Towards a Human-like Open-Domain Chatbot," *arXiv preprint arXiv:2001.09977*, 2020.
- [30] S. Yazdani, N. Saxena, Z. Wang, Y. Wu and W. Zhang, "A Comprehensive Survey of Image and Video Generative AI: Recent Advances, Variants, and Applications," 2024.
- [31] Y. Han, R. Wang, C. Zhang, J. Hu, P. Cheng, B. Fu and H. Zhang, "EMMA: Your Text-to-Image Diffusion Model Can Secretly Accept Multi-Modal Prompts," *arXiv preprint arXiv:2406.09162*, 2024.
- [32] K. Song, Y. Zhang, Y. Lei, J. Cong, H. Li, L. Xie, G. He and J. Bai, "DSPGAN: A Gan-Based Universal Vocoder for High-Fidelity TTS by Time-Frequency Domain Supervision from DSP," *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2023.
- [33] J. Yin, H. Chung and A. Regev, "A multi-view generative model for molecular representation improves prediction tasks," 2020.
- [34] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *arXiv preprint arXiv:1710.09412*, 2017.

- [35] X. Song, J. Qin, Q. Ren and J. Zheng, "IGAN: A collaborative filtering model based on Improved Generative Adversarial Networks for recommendation," *Engineering Applications of Artificial Intelligence*, vol. 124, p. 106569, 2023.
- [36] Z. Hongjian Gu, Y. Hu and Yingxue, "Preference and Concurrence Aware Bayesian Graph Neural Networks for Recommender Systems," *arXiv preprint arXiv:2312.11486*, 2023.
- [37] J. Xu, H. Xu, B. Ni, X. Yang and T. Darrell, "Video prediction via example guidance," *International Conference on Machine Learning*, vol. 119, pp. 10628-10637, 2020.
- [38] H. Moujahid, K. Boutahar, O. E. Gannour, S. Saleh, B. Cherradi and A. E. Abbassi, "A Scoping Review of Large Language Models: Architecture and Applications," in *2024 4th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Morocco, 2024, pp. 1-7.
- [39] V. Sanh, L. Debut, J. Chaumond and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.
- [40] S. Ortiz, "I've tested dozens of AI chatbots since ChatGPT's debut. Here's my new top pick," ZDNET, 3 June 2024. [Online]. Available: <https://www.zdnet.com/article/best-ai-chatbot/>. [Accessed 8 June 2024].
- [41] "Meet Claude," Anthropic PBC, 2024. [Online]. Available: <https://www.anthropic.com/claude>. [Accessed 14 03 2024].
- [42] "Meta Llama 3," Meta LLC, 2024. [Online]. Available: <https://llama.meta.com/llama3/>. [Accessed 21 03 2024].
- [43] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, 2019, p. 4171–4186.
- [44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, p. 1–67, 2020.
- [45] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskeya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz,

- E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov and N. Fiedel, "PaLM: Scaling Language Modeling with Pathways," *Journal of Machine Learning Research*, vol. 24, p. 1–113, 2023.
- [46] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever and others, "Improving language understanding by generative pre-training," 2018.
- [47] "TLDR Newsletter," TLDR, 2022. [Online]. Available: <https://tldr.tech/>. [Accessed 23 03 2024].
- [48] "You.com," [Online]. Available: <https://en.wikipedia.org/wiki/You.com>. [Accessed 15 09 2023].
- [49] "AI Chat Online for Free - Ask AI Chatbot | ChatSonic," WriteSonic, [Online]. Available: <https://writesonic.com/chat>. [Accessed 20 09 2023].
- [50] "Future proof your business with GTM AI," Copy.AI, [Online]. Available: <https://www.copy.ai/>. [Accessed 12 09 2023].
- [51] "Rytr: Free AI Writer, Content Generator, and Writing Assistant," Rytr, [Online]. Available: <https://rytr.me>. [Accessed 25 09 2023].
- [52] "HuggingChat : Making the community's best AI chat models available to everyone," HuggingFace, [Online]. Available: [huggingface.co/chat/](https://huggingface.co/chat/). [Accessed 22 11 2023].
- [53] N. Team, "Just ask Notion AI. Knowledge, answers, ideas. One click away.," Noion, [Online]. Available: <https://www.notion.so/product/ai>. [Accessed 2 08 2023].
- [54] "Clickup: The everything app, for work," ClickUp, [Online]. Available: <https://www.clickup.com>. [Accessed 01 09 2023].
- [55] "Jira: Issue and Project Tracking Software," Atlassian Corporation, [Online]. Available: <https://www.atlassian.com/software/jira>. [Accessed 11 09 2023].
- [56] kamilkaczmareksolutions, "Recipe Generator," GitHub, Inc, December 2023. [Online]. Available: [https://github.com/kamilkaczmareksolutions/Recipe\\_Generator](https://github.com/kamilkaczmareksolutions/Recipe_Generator). [Accessed 22 04 2024].
- [57] wgalindo1453, "HDIM2," GitHub, [Online]. Available: <https://github.com/wgalindo1453/HDIM2>. [Accessed 11 04 2024].

- [58] T. Dohmke, "GitHub Copilot Workspace: Welcome to the Copilot-native developer environment," GitHub, Inc, 29 April 2024. [Online]. Available: <https://github.blog/2024-04-29-github-copilot-workspace/>. [Accessed 16 02 2024].
- [59] N. S. Patil, R. S. Huang, C. B. van der Pol and N. Larocque, "Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment," *Canadian Association of Radiologists Journal*, vol. 75, pp. 344-350, 2024.
- [60] D. Kevian, U. Syed, X. Guo, A. Havens, G. Dullerud, P. Seiler, L. Qin and B. Hu, "Capabilities of Large Language Models in Control Engineering: A Benchmark Study on GPT-4, Claude 3 Opus, and Gemini 1.0 Ultra," *arXiv preprint arXiv:2404.03647*, 2024.
- [61] Z. W. Lim, K. Pushpanathan, S. Yew, Y. Lai, C.-H. Sun, J. Lam, D. Chen, J. Goh, M. Tan, B. Sheng, C.-y. Cheng, V. Koh and Y.-C. Tham, "Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard," *EBioMedicine*, vol. 95, p. 104770, August 2023.
- [62] G. Lai, Q. Xie, H. Liu, Y. Yang and E. Hovy, "RACE: Large-scale ReAding Comprehension Dataset From Examinations," *arXiv preprint arXiv:1704.04683*, p. 785–794, September 2017.
- [63] J. Thorne, A. Vlachos, C. Christodoulopoulos and A. Mittal, "FEVER: a Large-scale Dataset for Fact Extraction and VERification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New, 2018.
- [64] A. Azaria, "ChatGPT Usage and Limitations," 2022.
- [65] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever and W. Zaremba, "Evaluating Large Language Models Trained on Code," *arXiv preprint arXiv:2107.03374*, 2021.
- [66] G. Kim, P. Baldi and S. McAleer, "Language Models can Solve Computer Tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [67] W. Jiao, W. Wang, J.-t. Huang, X. Wang, S. Shi and Z. Tu, "Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine," *arXiv preprint arXiv:2301.08745*, 2023.



- [68] K. Valmeekam, M. Marquez, A. Olmo, S. Sreedharan and S. Kambhampati, "PlanBench: An Extensible Benchmark for Evaluating Large Language Models on Planning and Reasoning about Change," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt and S. Levine, Eds., New Orleans, Curran Associates, Inc., 2023, p. 38975–38987.
- [69] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, Y. Singer and S. Shieber, "Investigating Gender Bias in Language Models Using Causal Mediation Analysis," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, Eds., Curran Associates, Inc., 2020, p. 12388–12401.
- [70] S. e. al., "Beyond the imitation game: quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, p. 95, 2023.
- [71] C. a. Z. A. Qin, Z. Zhang, J. Chen, M. Yasunaga and D. Yang, "Is ChatGPT a General-Purpose Natural Language Processing Task Solver?," *arXiv preprint arXiv:2302.06476*, 2023.
- [72] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung and others, "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity," *Association for Computational Linguistics*, pp. 675-718, 2023.
- [73] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue and Y. Wu, "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," *arXiv preprint arXiv:2301.07597*, 2023.
- [74] S. Singh, N. Wen, Y. Hou, P. Alipoormolabashi, T.-L. Wu, X. Ma and N. Peng, "COM2SENSE: A Commonsense Reasoning Benchmark with Complementary Sentences," *arXiv preprint arXiv:2106.00969*, 2021.
- [75] J. Welbl, A. Glaese, J. Uesato, S. Dathathri, J. Mellor, L. A. Hendricks, K. Anderson, P. Kohli, B. Coppin and P.-S. Huang, "Challenges in Detoxifying Language Models," *arXiv preprint arXiv:2109.07445*, 2021.
- [76] Y. H. C. C. Z. X. Terry Yue Zhuo, "Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity," *arXiv preprint arXiv:2301.12867*, 2023.
- [77] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T.

Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Ł. Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Ł. Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O’Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk and B. Zoph, “GPT-4 Technical Report,” arXiv preprint arXiv:2303.08774, 2024.

- [78] A. Borji, “A Categorical Archive of ChatGPT Failures,” *arXiv preprint arXiv:2302.03494*, 2023.
- [79] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg and others, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” *arXiv preprint arXiv:2303.12712*, 2023.
- [80] E. Davis, “Benchmarks for Automated Commonsense Reasoning: A Survey,” *ACM Computing Surveys*, vol. 56, no. 4, p. 41, 2023.
- [81] J. Rudolph, S. Tan and S. Tan, “War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education,” *Journal of Applied Learning and Teaching*, vol. 6, p. 364–389, April 2023.
- [82] N. Rane, S. Choudhary and J. Rane, “Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation,” *Performance, Architecture, Capabilities, and Implementation (February 13, 2024)*, 2024.

- [83] L.-P. Cen, F.-F. Zhao, H.-J. He, J.-J. Liang, Y. Wang, H. Lin, F. Chen, T.-P. Li, J.-F. Yang and L. Chen, "Benchmarking the performance of Large Language Models in uveitis: A comparative analysis of {ChatGPT-3.5}, {ChatGPT-4.0}, Google," *Research Square*, 2024.
- [84] K. a. o. Yoo, "Diagnostic Performances of GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro in "Diagnosis Please" Cases," *medRxiv*, May 2024.
- [85] Y. Zhang and others, "CS-Bench: A Comprehensive Benchmark for Large Language Models towards Computer Science Mastery," *arXiv preprint arXiv:2406.08587*, June 2024.
- [86] Y. Lian, H. Tang, M. Xiang and X. Dong, "Public attitudes and sentiments toward ChatGPT in China: A text mining analysis based on social media," *Technology in Society*, vol. 76, p. 102442, 2024.
- [87] M. Ahuja, "Dataset - Google Sheets," May 2024. [Online]. Available: <https://docs.google.com/spreadsheets/d/146l0PeoatKx4XZWWasmAivHRKPo3yWr6e8AtfxaCAts/edit?usp=sharing>.
- [88] "Plans & Pricing," Copy AI, [Online]. Available: <https://www.copy.ai/prices>. [Accessed 11 05 2024].
- [89] "Pricing," Rytr, [Online]. Available: <https://rytr.me/pricing/>. [Accessed 11 10 2023].
- [90] "Pricing," You: AI Search Engine, [Online]. Available: <https://you.com/plans>. [Accessed 20 05 2024].
- [91] "pandas," NumFOCUS, Inc, 10 April 2024. [Online]. Available: <https://pandas.pydata.org/>. [Accessed 2024 05 2024].
- [92] A. Amjad, "Mastering Data Manipulation: An Insight into Python's Pandas Library," LinkedIn, 15 March 2024. [Online]. Available: <https://www.linkedin.com/pulse/mastering-data-manipulation-insight-pythons-pandas-library-asad-amjad-eudif/>.
- [93] P. Nandwani and V. Rupali, "A review on sentiment analysis and emotion detection from text," *Social Network Analysis and Mining*, vol. 11, p. 81, 2021.
- [94] B. Bulté and A. Housen, "Defining and operationalising {L2} complexity," in *Dimensions of {L2} Performance and Proficiency*, Amsterdam, John Benjamins Publishing Company, 2012, pp. 21-46.
- [95] B. Richards, "Type/Token Ratios: what do they really tell us?," *Journal of Child Language*, vol. 14, no. 2, pp. 201-209, 1987.

- [96] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Linguisticæ Investigationes*, vol. 30, no. 1, pp. 3-26, 2007.
- [97] W. contributors, "English pronouns," Wikipedia, The Free Encyclopedia, 17 May 2024. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=English\\_pronouns&oldid=1224359944](https://en.wikipedia.org/w/index.php?title=English_pronouns&oldid=1224359944). [Accessed May 2024].
- [98] "pandas.read\_csv," NumFOCUS, Inc, 2024. [Online]. Available: [https://pandas.pydata.org/docs/reference/api/pandas.read\\_csv.html](https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html). [Accessed 20 05 2024].
- [99] "Pandas DataFrames," W3Schools, [Online]. Available: [https://www.w3schools.com/python/pandas/pandas\\_dataframes.asp](https://www.w3schools.com/python/pandas/pandas_dataframes.asp). [Accessed 8 06 2024].
- [100] W. contributors, "Natural Language Toolkit," Wikipedia, The Free Encyclopedia, 2024. [Online]. Available: [https://en.wikipedia.org/w/index.php?title=Natural\\_Language\\_Toolkit&oldid=1223463246](https://en.wikipedia.org/w/index.php?title=Natural_Language_Toolkit&oldid=1223463246).
- [101] Steven Loria, "TextBlob: Simplified Text Processing," 2013. [Online]. Available: <https://textblob.readthedocs.io/en/dev/index.html>.
- [102] W. Contributors, "Scikit-learn," Wikipedia, The Free Encyclopedia, 2024. [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Scikit-learn&oldid=1219319746>.
- [103] K. Raj P M and J. Sai D, "Sentiment analysis, opinion mining and topic modelling of epics and novels using machine learning techniques," *Materials Today: Proceedings*, vol. 51, no. 1, pp. 576-584, 2022.
- [104] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, pp. 1-21, 2019.
- [105] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, "Lexicon-Based Methods for Sentiment Analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267-307, June 2011.
- [106] B. P. Lee and Lillian, "Opinion Mining and Sentiment Analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, pp. 1-135, 2008.
- [107] A. P. Kirilenko, L. Wang and S. O. Stepchenkova, "Sentiment Analysis: Gaging Opinions of Large Groups," in *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications*, Cham, Springer International Publishing, 2022, pp. 363-374.

- [108] M. Tadayonifar, "The Effects of Text Type, Text Length and Text Difficulty on Vocabulary Retention through Glossing," *Journal of Language Teaching and Learning*, vol. 6, no. 1, 2021.
- [109] E. Pitler and A. Nenkova, "Revisiting readability: a unified framework for predicting text quality," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Honolulu, Hawaii: Association for Computational Linguistics, 2008, p. 186–195.
- [110] D. J. Arya, E. H. Hiebert and P. D. Pearson, "The effects of syntactic and lexical complexity on the comprehension of elementary science texts.," *International Electronic Journal of Elementary Education*, vol. 4, no. 1, pp. 107-125, 10 2011.
- [111] J. P. Kincaid, R. P. Fishburne Jr., R. L. Rogers and B. S. Chissom, Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel, Millington (Tenn.): Institute for Simulation and Training, University of Central Florida, 1975, p. 48.
- [112] C. Meri and T. L. Liau, "A Computer Readability Formula Designed for Machine Scoring," *Journal of Applied Psychology*, vol. 60, pp. 283-284, 1975.
- [113] F. Heylighen and J.-M. Dewaele, "Variation in the contextuality of language: An empirical measure," *Foundations of Science*, vol. 7, no. 3, pp. 293-340, 2002.
- [114] D. Biber, *Variation across speech and writing*, Cambridge University Press, 1988.
- [115] W. Chafe, "Integration and involvement in speaking, writing, and oral literature," in *Spoken and written language: Exploring orality and literacy*, 1982, pp. 35-53.
- [116] S. Jarvis, "Capturing the diversity in lexical diversity," *System*, vol. 41, no. 4, pp. 938-950, 2013.
- [117] V. Johansson, "Lexical diversity and lexical density in speech and writing: a developmental perspective," *Working Papers in Linguistics*, vol. 53, pp. 61-79, 2008.
- [118] C. S. Team, "Use ClickUp AI to write a StandUp or team updates," ClickUp, [Online]. Available: <https://help.clickup.com/hc/en-us/articles/20011540694551-Use-ClickUp-AI-to-write-a-StandUp-or-team-updates>. [Accessed 11 08 2023].
- [119] A. Team, "Do the impossible with Atlassian Intelligence," Atlassian, Inc, [Online]. Available: <https://www.atlassian.com/platform/artificial-intelligence>.

- [120] D. Moore, *Basic Practice Stats 8e*, MAP, 2018.
- [121] M. Bulmer, *Principles of Statistics*, 1979: Dover Publications, 20.
- [122] M. L. Lyman Ott, *An Introduction to Statistical Methods and Data Analysis*, Brooks/Cole, 2010.
- [123] M. Stojiljković, "Python Statistics Fundamentals: How to Describe Your Data," Real Python, [Online]. Available: <https://realpython.com/python-statistics/>.
- [124] "Statistical functions (scipy.stats)," SciPy, [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/stats.html>.
- [125] M. Lee, "Recent Advances in Generative Adversarial Networks for Gene Expression Data: A Comprehensive Review," *Mathematics*, vol. 11, 2023.
- [126] P. J. Cobb, "Large Language Models and Generative AI, Oh My!: Archaeology in the Time of ChatGPT, Midjourney, and Beyond," *Advances in Archaeological Practice*, vol. 11, p. 363–369, 2023.
- [127] M. M. Carla, G. Gambini, A. Baldascino, F. Boselli, F. Giannuzzi, F. Margollicci and S. Rizzo, "Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison," *Graefe's Archive for Clinical and Experimental Ophthalmology*, pp. 1-15, 2024.
- [128] D. Isele, "Interactive Decision Making for Autonomous Vehicles in Dense Traffic," *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 3981-3986, 2019.
- [129] "Write, plan, organize, play," Notion, [Online]. Available: <https://notion.so>. [Accessed 3 08 2023].
- [130] D. Pierce, "Google launches Gemini, the AI model it hopes will take down GPT-4," *The Verge*, 07 December 2023. [Online]. Available: <https://www.theverge.com/2023/12/6/23990466/google-gemini-llm-ai-model>. [Accessed 14 02 2024].

## APPENDIX 1 - AI CHATBOTS

In this section, responses to only three prompts have been included in the table. For the complete list of prompts with their responses from all the AI chatbots, please refer to [87].

Questions	<a href="#">Copy.ai</a>	You Chat	ChatSonic	<a href="#">Rytr.me</a>	HuggingChat
Talk about the History of New Zealand in 100 words.	New Zealand was first settled by Polynesians over 1,000 years ago. Europeans arrived in the late 18th century and by 1840 the Treaty of Waitangi was signed, making New Zealand a British colony. In 1893, New Zealand became the first country in the world to grant women the right to vote. Throughout the 20th century, New Zealand developed as a nation, gaining independence from Britain in 1947, becoming a welfare state in the 1930s, and playing an active role in international affairs. Today, New Zealand is a modern developed nation with a diverse culture, stunning natural beauty, and a strong economy.	New Zealand's history dates back to the arrival of Polynesians around 1300 AD, followed by European exploration in the 17th century. British colonization began in the early 19th century, leading to conflicts with the indigenous Māori people. The Treaty of Waitangi in 1840 established British sovereignty and Māori rights. New Zealand became a dominion in 1907 and achieved full independence in 1947. The country played a significant role in both World Wars and has since developed a strong economy based on agriculture, tourism, and technology. Today, New Zealand is known for its stunning	History of New Zealand New Zealand's history begins with the arrival of the Māori people around 1300 AD. European exploration and colonization started in the 17th century, with British sovereignty established in 1840 through the Treaty of Waitangi. The 19th century saw conflicts between Māori and European settlers. In 1840, New Zealand became a self-governing dominion within the British Empire. The country actively participated in both World Wars and experienced social and economic reforms in the mid-20th century. From the 1980s, New Zealand underwent significant political and economic changes, embracing a more multicultural society and becoming a leader in environmental conservation.	New Zealand's history is rich and diverse, shaped by its Maori roots and European colonisation. The Maori people arrived around 1300 AD, establishing a vibrant culture. In 1642, Dutch explorer Abel Tasman discovered the islands, followed by British navigator James Cook in the late 18th century. The Treaty of Waitangi in 1840 marked the beginning of British sovereignty, leading to conflicts and land wars. New Zealand gained independence in 1907 and became a welfare state in the mid-20th century. Today, it is known for its stunning landscapes, progressive policies, and strong cultural identity.	The history of New Zealand dates back to its initial Polynesian settlement around the 13th century, though exact origins remain uncertain. Early European explorations began in the mid-17th century, with Dutch navigator Abel Tasman mapping parts of the coastline in 1642. However, it wasn't until Captain James Cook's arrival in 1769 that New Zealand gained significant international attention. His detailed exploration led to increased interest from settlers and traders, particularly from Britain. In 1840, under Governor William Hobson, the Treaty of Waitangi established British sovereignty over New Zealand, leading to rapid changes in Māori society and the eventual formation of provincial governments. Over the following years, New Zealand experienced periods of economic growth, political reform,

		landscapes, diverse culture, and progressive policies.			and cultural tension between the indigenous Māori population and incoming settlers. Today, New Zealand stands as a modern, multicultural nation with strong ties to both its Māori heritage and its colonial past.
How did the leaders in New Zealand encourage growth and development of the IT sector. Please brief how IT sector has evolved over decades.	The New Zealand government has played a key role in encouraging the growth and development of the IT sector. Over the decades, various initiatives have been implemented to support the industry. In the 1980s, the government introduced policies to promote technological innovation and increase investment in the IT sector. This led to the establishment of research and development grants and tax incentives for businesses engaging in IT-related activities. In the 1990s, the government implemented policies to improve access to telecommunications infrastructure, which greatly benefitted the IT sector. The	##### Encouragement of the IT Sector in New Zealand has actively encouraged the growth of its information and communications technology (ICT) sector. The country boasts advanced digital infrastructure, ranking in the top 20 globally for fixed broadband speed and top 40 for mobile performance. The ICT sector is the country's third largest export sector, generating over \$8.7 billion in revenue annually, with the GDP of the information media and telecommunication industry worth almost \$6 billion in 2022. The government has launched a new approach to	Encouraging the IT Sector in New Zealand The New Zealand government has actively supported the growth of the digital technologies sector through investments and initiatives to foster innovation and productivity. The Minister for the Digital Economy and Communications, David Clark, announced additional funding in Budget 2022 to support key initiatives in the Digital Technologies Industry Transformation Plan (ITP), aiming to promote high-wage, low emissions economy. Evolution of the IT Sector Historical Growth: Since 2015, the digital technologies sector has grown approximately 77% faster than the general economy, contributing significantly to the economy. The government has launched a new Software-as-a-Service (SaaS) Community: The	Leaders in New Zealand have encouraged the growth of the IT sector through various initiatives such as investing in infrastructure, supporting research and development, and fostering innovation. The government has implemented policies to attract foreign investment, promote digital literacy, and support local tech startups. Over the decades, the IT sector in New Zealand has evolved significantly. From the early days of basic computing and telecommunications, the sector has grown to encompass a wide range of industries such as software development, cybersecurity, data analytics, and cloud computing. New Zealand has become a hub for technology companies, with a strong focus on sustainability and innovation. The sector continues to expand, contributing significantly	Since the late 20th century, New Zealand's leaders have actively promoted the IT sector through various initiatives, policies, and investments. This support has been instrumental in fostering the growth of the industry, which now plays a vital role in the country's economy. In the early days, the government recognized the potential of IT as an engine for innovation and job creation. They initiated programs such as the Technology New Zealand strategy (1995) and the Information Age policy framework (1999), aiming to develop a skilled workforce, attract foreign investment, and encourage research and development. These efforts helped establish a solid foundation for the IT industry. Over the next two decades, the IT sector continued to evolve rapidly. With advancements in technology, new



<p>deregulation of the telecommunications industry led to increased competition and the expansion of internet services throughout the country.</p> <p>In the early 2000s, the government launched the Digital Strategy, focusing on the development of digital infrastructure, promoting digital literacy, and fostering digital inclusion. This initiative aimed to position New Zealand as a leading digital nation.</p> <p>The government has also invested in education and training programs to develop a skilled IT workforce. Universities and technical institutes offer IT-related courses, ensuring a steady supply of qualified professionals for the sector.</p> <p>The IT sector in New Zealand has experienced significant growth over the decades. Today, it is a vibrant and dynamic industry,</p>	<p>industry policy aimed at growing more innovative industries in New Zealand, including redirecting investment towards more productive sectors.</p> <p>##### Evolution of the IT Sector Over Decades</p> <p>Over the years, New Zealand's tech sector has evolved into a diverse and advanced industry, accounting for 8% of the country's GDP and employing 5% of the workforce. In 2021, the tech sector exported \$8.6 billion globally, making it the country's second-largest export sector. The sector is growing and a breeding ground for innovation and competes successfully on the world stage, with a focus on areas such as game development, data-driven innovation, and AI.</p> <p>The country's telecommunication market has seen mobile prices drop to some of the lowest in the OECD,</p>	<p>SaaS subsector experienced sustained growth, doubling its export revenue and headcount within a five-year window, despite the challenges posed by the COVID-19 pandemic.</p> <p>3.</p> <p>Active Support: The government's focus on skill development and access to talent, as well as the promotion of New Zealand's tech sector locally and internationally, has been crucial in encouraging the growth of the IT sector 3.</p> <p>International Collaboration: Efforts to promote the country as a global contender in the tech market, along with a focus on building momentum and expanding networks, demonstrate a long-term commitment to growing the digital technologies sector 3.</p> <p>The IT sector in New Zealand has evolved into a vibrant and competitive industry, reflecting the nation's receptiveness to new technologies and its commitment to fostering innovation and productivity 13.</p>	<p>to the country's economy and global competitiveness.</p>	<p>opportunities emerged, including software development, digital services, data analytics, and cybersecurity. To keep up with these developments, the government launched further strategies like the Digital Government Strategy (2012) and the Digital Economy Action Agenda (2016). These strategies aimed at enhancing digital skills, promoting collaboration between businesses and academia, and supporting startups and scale-ups.</p> <p>The establishment of organizations such as Callaghan Innovation and NZTech played crucial roles in nurturing the ecosystem. These bodies provided resources, funding, mentorship, and networking opportunities for tech entrepreneurs and researchers. Additionally, initiatives like the Tech Futures Lab, a joint venture by Auckland University of Technology and AUT, focused on preparing students for careers in emerging technologies.</p> <p>As a result, New Zealand's IT sector has grown exponentially over the last few decades, transforming into a thriving global player. Companies like Xero, Trade Me, and Vend are among the successful</p>
--	--	---	---	--

	<p>contributing to the country's economy and creating job opportunities. The sector encompasses various areas, including software development, telecommunications, e-commerce, and digital services. Innovative startups and established companies alike are driving the growth of New Zealand's IT sector, making it a thriving and competitive part of the economy.</p>	<p>and it has a high rate of internet use. New Zealand has actively welcomed and encouraged foreign investment in its IT sector, contributing to its growth and competitiveness on the global stage.</p>			<p>homegrown brands making waves internationally. Furthermore, the country is also known for being a hub for video game development, with renowned studios like Weta Workshop, Grinding Gear Games, and Sidhe Interactive contributing significantly to this field.</p> <p>In conclusion, the proactive approach taken by New Zealand's leaders has undoubtedly contributed to the remarkable evolution of the IT sector over the past several decades. By investing in education, infrastructure, and entrepreneurship, they have successfully turned the country into a dynamic and innovative landscape for technology.</p>
<p>How has the world adapted to Data Science and Artificial Intelligence for their future? Write not more than 100 words.</p>	<p>The world has embraced Data Science and Artificial Intelligence (AI) for the future. Organizations across industries are leveraging data science to extract insights, make informed decisions, and gain a competitive edge. AI is being integrated into various applications such as autonomous vehicles, virtual</p>	<p>Worldwide, organizations are embracing data science and artificial intelligence (AI) to revolutionize decision-making, automate processes, and drive innovation. From healthcare to finance, these technologies are being utilized to analyze vast amounts of data, uncover insights, and enhance</p>	<p>Worldwide, organizations are increasingly leveraging data science and artificial intelligence (AI) to drive innovation and efficiency across diverse sectors. From healthcare to finance, these technologies are revolutionizing decision-making, enabling predictive analytics, and automating processes. The future sees AI and data science shaping personalized customer experiences, powering</p>	<p>The world has embraced data science and artificial intelligence as key drivers for the future, integrating them into various industries and sectors. Businesses are leveraging data to make informed decisions, improve efficiency, and enhance customer experiences. AI technologies are being used for automation, predictive analytics, and personalized recommendations. Governments are utilizing data science for policy-making and public</p>	<p>Worldwide, countries and industries have embraced Data Science and Artificial Intelligence (AI) as essential components of their futures. Governments invest in AI research and development, while companies leverage these technologies to enhance efficiency, automate processes, and make better decisions. Education systems adapt curriculums to foster talent in STEM fields, ensuring a skilled workforce for the digital</p>

<p>assistants, and predictive analytics, revolutionizing industries like healthcare, finance, and manufacturing. Governments have recognized the potential of these technologies in driving economic growth and are investing in research and development. However, ethical considerations, privacy concerns, and the need for responsible AI deployment are also being addressed to ensure the responsible and inclusive use of Data Science and AI in shaping our future.</p>	<p>predictive capabilities. AI is transforming industries through chatbots, autonomous vehicles, and personalized recommendations. Governments are also leveraging AI for public services and policy-making. As the demand for skilled professionals in these fields grows, educational institutions and businesses are investing in training and research. The future sees data science and AI as integral to driving efficiency, competitiveness, and societal progress globally.</p>	<p>autonomous vehicles, and advancing medical research. However, concerns about data privacy and ethical AI use persist, prompting ongoing dialogue and regulations to ensure responsible deployment. As the world embraces these technologies, their continued integration promises to redefine industries, enhance productivity, and address complex global challenges.</p>	<p>services. Academia is incorporating these fields into research and education. The global adaptation of data like healthcare, climate science and AI is transforming how we work, communicate, and innovate, shaping a more data-driven and intelligent future for society.</p>	<p>Public-private partnerships drive innovation, accelerating breakthroughs in areas like healthcare, climate change, and smart cities. As AI and Data Science continue to advance, ethical considerations become increasingly important, guiding responsible implementation and regulating potential biases or misuse. Overall, the world adapts to harness the power of AI and Data Science, shaping a smarter, connected, and sustainable future.</p>
---	---	---	---	--

## APPENDIX 2 - PROJECT MANAGEMENT TOOLS

In this section, responses to only three prompts have been included in the table. For the complete list of prompts with their responses from the AI-enriched project management tools, please refer to [87].

Questions	ClickUp.com	Jira	Notion.so
Develop a project plan for launching a new software application, outlining tasks, milestones, and resource allocation.	<p>To develop a project plan for launching a new software application, you would need to outline a series of tasks, set milestones, and allocate resources appropriately. Below is a high-level project plan that you can tailor to your specific needs:</p> <p><b>**Project Name:**</b> New Software Application Launch</p> <p><b>**Project Manager:**</b> [Maghav Ahuja](#3335682)</p> <p><b>**Phase 1: Initiation**</b></p> <ul style="list-style-type: none"> <li>- Task: Project Charter Development</li> <li>- Task: Stakeholder Identification</li> <li>- Milestone: Project Kickoff Meeting</li> <li>- Resources: Project Manager, Business Analyst</li> </ul> <p><b>**Phase 2: Planning**</b></p> <ul style="list-style-type: none"> <li>- Task: Requirement Gathering</li> <li>- Task: Scope Definition</li> <li>- Task: Software Architecture Design</li> </ul> <p>([Software Architecture Design Considerations](https://app.clickup.com/3319689/docs/359w9-622/359w9-642))</p> <ul style="list-style-type: none"> <li>- Milestone: Project Plan Approval</li> <li>- Resources: Project Team, [Maghav Ahuja](#3335682), Software Architects</li> </ul> <p><b>**Phase 3: Execution**</b></p> <ul style="list-style-type: none"> <li>- Task: Development of Core Features</li> <li>- Task: User Interface Design</li> <li>- Task: Database Setup</li> <li>- Task: Integration of Components</li> <li>- Milestone: First Prototype Ready</li> <li>- Resources: Developers, UI/UX Designers, Database Administrators</li> </ul> <p><b>**Phase 4: Testing**</b></p> <ul style="list-style-type: none"> <li>- Task: Unit Testing</li> </ul>	<p>Project Plan for Launching New Software Application</p> <p>Introduction</p> <p>The launch of a new software application requires meticulous planning and coordination to ensure its successful implementation.</p> <p>This project plan outlines the tasks, milestones, and resource allocation necessary for the seamless introduction of the new software application.</p> <p>Project Tasks</p> <p>Market Research</p> <p>Conduct market analysis to identify customer needs and preferences.</p> <p>Analyze competitor offerings and market trends.</p> <p>Define Requirements</p>	<p>## Tasks</p> <ol style="list-style-type: none"> <li>1. Define the scope and objectives of the software application.</li> <li>2. Assemble the project team and define roles and responsibilities.</li> <li>3. Establish project timeline and milestones.</li> <li>4. Develop a detailed software design.</li> <li>5. Write the software code.</li> <li>6. Test the software for bugs and fix any that are found.</li> <li>7. Prepare the software for launch, including packaging and marketing materials.</li> <li>8. Launch the software.</li> <li>9. Provide post-launch support and maintenance.</li> </ol> <p>## Milestones</p> <ol style="list-style-type: none"> <li>1. Project Kickoff: Establish project team, define scope and objectives.</li> <li>2. Detailed Design Complete: Finalize all aspects of software design.</li> <li>3. Coding Complete: Finish writing all software code.</li> <li>4. Testing Complete: Ensure software is free of</li> </ol>

	<ul style="list-style-type: none"> <li>- Task: Integration Testing</li> <li>- Task: User Acceptance Testing</li> <li>- Milestone: Testing Completion</li> <li>- Resources: Quality Assurance Team, Test Users</li> </ul> <p><b>**Phase 5: Deployment**</b></p> <ul style="list-style-type: none"> <li>- Task: Deployment Planning</li> <li>- Task: Production Environment Setup</li> <li>- Task: Data Migration</li> <li>- Milestone: Software Goes Live</li> <li>- Resources: IT Operations Team, Database Administrators</li> </ul> <p><b>**Phase 6: Post-Launch**</b></p> <ul style="list-style-type: none"> <li>- Task: User Training</li> <li>- Task: Post-Launch Support Setup</li> <li>- Task: Performance Monitoring</li> <li>- Milestone: Project Closure Documentation</li> <li>- Resources: Support Team, Trainers, Monitoring Tools</li> </ul> <p><b>**Resource Allocation:**</b></p> <ul style="list-style-type: none"> <li>- Allocate team members based on expertise and project requirements.</li> <li>- Budget for software, hardware, and cloud services.</li> <li>- Plan for external consultants if specialized skills are needed.</li> </ul> <p><b>**Risk Management:**</b></p> <ul style="list-style-type: none"> <li>- Identify potential risks such as scope creep, technical challenges, or resource constraints.</li> <li>- Develop mitigation strategies for each identified risk.</li> </ul> <p><b>**Communication Plan:**</b></p> <ul style="list-style-type: none"> <li>- Regular project updates to stakeholders.</li> <li>- Use ClickUp for task tracking and collaboration.</li> </ul> <p><b>**Documentation:**</b></p> <ul style="list-style-type: none"> <li>- Maintain comprehensive documentation throughout the project lifecycle.</li> </ul> <p>This project plan serves as a blueprint and should be detailed further with specific tasks, deadlines, and resource names. Regular review</p>	<p>bugs.</p> <p>Gather input from stakeholders regarding functional requirements.</p> <p>Document technical specifications and performance expectations.</p> <p>Design architecture and user interface.</p> <p>Code development with regular testing iterations.</p> <p>Quality Assurance</p> <p>Perform comprehensive testing across different platforms.</p> <p>Implement bug fixes based on testing results.</p> <p>Deployment Preparation</p> <p>Prepare deployment infrastructure including servers, databases, etc.</p> <p>Develop migration strategy for existing users (if applicable).</p> <p>Training &amp; Documentation</p> <p>Create user guides and training</p>	<p>5. Launch Ready: Prepare all launch materials and finalize software packaging.</p> <p>6. Launch: Release the software to the public.</p> <p>7. Post-Launch Support: Provide ongoing support and maintenance.</p> <p><b>## Resource Allocation</b></p> <p>1. Project Manager: Oversees all aspects of the project. (1 FTE)</p> <p>2. Software Developers: Write software code. (3 FTEs)</p> <p>3. Quality Assurance: Test software for bugs. (2 FTEs)</p> <p>4. Marketing: Prepare marketing materials for launch. (1 FTE)</p> <p>5. Support: Provide post-launch support and maintenance. (1 FTE)</p>
--	--	--	--

	<p>and adjustment may be necessary as the project progresses.</p>	<p>materials as per finalized features set implemented in the app.</p> <p>Milestones</p> <p>Completion of Market Research: End of Month 1</p> <p>Finalization of Development Phase: End of Month 4</p> <p>Commencement of Quality Assurance Testing: Beginning of Month 5</p> <p>Deployment Ready: End of Month 7</p> <p>Resource Allocation</p> <p>Role</p> <p>Responsibility</p> <p>Project Manager</p> <p>Overall project oversight</p> <p>Business Analyst</p> <p>Requirements gathering &amp; analysis</p> <p>Developers</p> <p>Application design &amp; coding</p> <p>Quality Assurance</p> <p>Comprehensive testing</p>	
--	---	--	--

		<p>This project plan provides a structured approach towards launching the new software application while ensuring that all key aspects are thoroughly considered throughout each phase.</p>	
<p>Write task descriptions for the development team, specifying requirements, deadlines, and dependencies.</p>	<p>Certainly! Below are task descriptions for the development team, including requirements, deadlines, and dependencies:</p> <p><b>Task 1: Implement User Authentication System</b></p> <ul style="list-style-type: none"> <li><b>Requirements:</b> Develop a secure user authentication system that supports both traditional email/password and OAuth options. Ensure that the system includes password recovery and multi-factor authentication features.</li> <li><b>Deadline:</b> Complete the initial implementation for review by <b>05/10/2024</b>.</li> <li><b>Dependencies:</b> This task is dependent on the completion of the database schema design ([Task 2](https://app.clickup.com/t/9016274493/86c333zu8)) for user data storage. Ensure that the authentication system is compatible with the database design.</li> </ul> <p><b>Task 2: Design Database Schema for User Data</b></p> <ul style="list-style-type: none"> <li><b>Requirements:</b> Create a comprehensive database schema that can handle user data, including personal details, authentication information, and user preferences. The schema should be scalable and normalized to</li> </ul>	<p>Task Descriptions for Development Team</p> <p>Market Research</p> <p>Conduct market analysis to identify customer needs and preferences.</p> <p>Analyze competitor offerings and market trends.</p> <p>Define Requirements</p> <p>Gather input from stakeholders regarding functional requirements.</p> <p>Document technical specifications and performance expectations.</p> <p>Development</p> <p>Design architecture and user interface.</p>	<p>## Development Team Tasks</p> <ol style="list-style-type: none"> <li><b>Software Design (Deadline: Milestone 2):</b> Collaborate with the project manager to develop a detailed design of the software application. This includes defining the software's functionality, user interface, and architecture. Dependency: Scope and objectives definition.</li> <li><b>Code Writing (Deadline: Milestone 3):</b> Write the software code based on the detailed design. Ensure that the code is clean, efficient, and well-documented. Dependency: Detailed design completion.</li> <li><b>Bug Testing (Deadline: Milestone 4):</b> Test the software for bugs and fix any that are found. This includes both unit testing of</li> </ol>

	<p>support future feature expansions.</p> <ul style="list-style-type: none"> <li>- <b>Deadline</b>: Finalize the schema by <b>04/20/2024</b>.</li> <li>- <b>Dependencies</b>: This task must be completed before the user authentication system ([Task 1](https://app.clickup.com/t/9016274493/86cv33zu7)) and the user profile management feature ([Task 3](https://app.clickup.com/t/9016274493/86cv33zu9)) are implemented.</li> </ul> <p><b>Task 3: Develop User Profile Management Feature</b></p> <ul style="list-style-type: none"> <li>- <b>Requirements</b>: Create a feature that allows users to view, create, and update their profiles. Include the ability to upload a profile picture, edit personal information, and manage privacy settings.</li> <li>- <b>Deadline</b>: Aim to have a working prototype by <b>05/25/2024</b>.</li> <li>- <b>Dependencies</b>: This task relies on the database schema ([Task 2](https://app.clickup.com/t/9016274493/86cv33zu8)) to be in place. It should also integrate seamlessly with the user authentication system ([Task 1](https://app.clickup.com/t/9016274493/86cv33zu7)) for profile access control.</li> </ul> <p>---</p> <p>Please ensure that each task is assigned to the appropriate team members and that they are aware of their respective deadlines and dependencies. Regular updates should be provided during stand-up meetings to track progress and address any potential blockers.</p> <p><b>Based on 2 sources:</b></p> <ul style="list-style-type: none"> <li>- [How our product managers use ClickUp](https://help.clickup.com/hc/en-us/articles/6328054257047-How-our-product-managers-use-ClickUp)</li> <li>- [Organize your Hierarchy for PMO](https://help.clickup.com/hc/en-us/articles/9703037723159-Organize-your-Hierarchy-for-PMO)</li> </ul>	<p>Code development with regular testing iterations. ([Task 4](https://app.clickup.com/t/9016274493/86cv33zu4))</p> <p>Quality Assurance ([Task 5](https://app.clickup.com/t/9016274493/86cv33zu5))</p> <p>Perform comprehensive testing across different platforms.</p> <p>Implement bug fixes based on testing results.</p> <p>Deployment Preparation</p> <p>Prepare deployment infrastructure including servers, databases, etc.</p> <p>Develop migration strategy for existing users (if applicable).</p> <p>Training &amp; Documentation</p> <p>Create user guides and training materials as per finalized features set implemented in the app.</p> <p>Deadlines:</p> <p>Completion of Market Research: End of Month 1</p> <p>Finalization of Development Phase: End of Month 4</p>	<p>individual features and integration testing of the software as a whole.</p> <p>Dependency: Code writing completion.</p> <p>4. <b>Software Preparation for Launch (Deadline: Milestone 5)</b>: Work with the marketing team to prepare the software for launch. This includes finalizing the software packaging and helping to develop marketing materials. Dependency: Testing completion.</p> <p>5. <b>Post-Launch Support (Deadline: Ongoing after Milestone 6)</b>: Provide support for the software after launch. This includes fixing any bugs that are found and making any necessary updates or improvements.</p> <p>Dependency: Software &amp; launch.</p>
--	--	--	---



		<p>Commencement of Quality Assurance Testing: Beginning of Month 5</p> <p>Deployment Ready: End of Month 7</p> <p>Dependencies: The completion of each phase is dependent on the successful execution and completion of the preceding phase to ensure a seamless transition throughout the project lifecycle.</p> <p>This plan provides a structured approach towards launching the new software application while ensuring that all key aspects are thoroughly considered throughout each phase.</p>	
--	--	---	--