Bilkent University

GE301 Term Project

2023-2024 Spring

Sec-14

Project Title:

Responsible Conversations: A Study of Responsible Chatbot Deployment at DataBoss Security & Analytics

Instructor: Robin Ann Downey

Mehmet Yiğit Turalı EEE - 21901822

Roj Deniz Aldemir EEE - 22102442

Dilara Büşra Yörür IE - 22103808

Table of Contents

Introduction

Artificial intelligence (AI) has the potential to revolutionize a wide range of industries in today's world with many uses it could possibly provide to humanity. But there are also important social and ethical issues to consider, despite its advantages. It is crucial to make sure AI development and application are in line with ethical standards and societal values as technology becomes more and more included in daily life. Bearing all these facts in mind, one can conclude that it is important to evaluate innovations from a RRI perspective, since it is a necessity of the day to employ certain ethical values while constructing new technology.

A major force in the AI industry, this analysis explores DataBoss's strategic initiatives through a thorough evaluation based on the RRI framework. It highlights the company's risk assessment techniques, reflective approach to ongoing learning and improvement, inclusive stakeholder engagement tactics, and responsive adjustment to ethical standards and societal demands.

This study points out how DataBoss takes values like anticipation, reflexivity, inclusion, and responsiveness into account while developing new innovations. DataBoss has become a leader in responsible AI innovation, defining new benchmarks for the sector by fusing technological innovation with moral foresight. Encouraging stakeholder inclusivity and upholding strict ethical standards will be crucial in supporting the company as it moves forward. Although DataBoss has made notable advances in integrating ethical and societal considerations into its AI development through the VSD paradigm, the ongoing challenges of bias, data privacy, security, and the potential for AI singularity highlight the need for a more refined and dynamic approach to ensure that AI technologies are developed and deployed in a manner that is genuinely beneficial and equitable for all stakeholders.

Theory

The theoretical foundation of this study is grounded in the principles of Value Sensitive Design (VSD), a framework that integrates human values into the design of technology at the forefront of development and operation. Value Sensitive Design, as articulated by Friedman and colleagues, emphasizes the importance of accounting for human values in a comprehensive manner throughout the technological design process. This approach is not only systematic and iterative but also fundamentally interdisciplinary, combining conceptual, empirical, and technical investigations to ensure that all stakeholders' values are considered [1].

For the purpose of this study, VSD is particularly pertinent due to its focus on ethical values such as privacy, accountability, and user autonomy. These values are critical when assessing the development and implementation of large language models (LLMs) by DataBoss. The iterative process of VSD allows for continuous reassessment and realignment of the LLM's design with these core values, addressing ethical concerns that arise as the technology evolves [2].

In applying VSD to DataBoss's practices, this study will examine how the company integrates the VSD framework into its development of LLMs. Specifically, the study will focus on three major components of VSD:

Conceptual Investigations: Identifying and defining the values at stake. In the context of LLMs, crucial values might include transparency, fairness, and inclusiveness. Understanding how DataBoss perceives and prioritizes these values is key to assessing its ethical alignment.

Technical Investigations: Analyzing how the company's technological choices either support or hinder the realization of identified values. This includes examining the algorithms and data sets used in LLMs to ensure they do not perpetuate biases or infringe on user privacy.

Empirical Investigations: Engaging with stakeholders to understand their experiences and values in relation to DataBoss's LLMs. This involves gathering data from users, developers, and ethicists to gain a holistic view of the impact of these models.

By using VSD as a lens, this study aims to provide a thorough analysis of the ethical dimensions of LLM development at DataBoss, offering insights into how well the company's practices align with the broader goals of responsible innovation.

Background Research

Technology

General overview of LLMs

Significant progress in artificial intelligence has been mainly driven by LLMs, e.g., GPT models developed by OpenAI, with the most crucial achievement in natural language processing. Such models, which have been associated with their deep learning architecture layers and enormous datasets, are capable of mimicking human writing styles. They provide for versatile use, from writing simple texts to complex tasks such as summarizing, translation,

and even writing codes. The LLMs have a solid basis for their work: to learn to anticipate the following word in a chain, producing coherent and contextually suitable text. As they keep developing, the models are increasingly used in technologies designed for communication with human users, resolving the issues of trustworthy interaction between people and machines [3].

Innovation Process

The iterative and continuous process of LLMs' innovation reflects a research, development, and implementation cycle. Significantly, these models' architecture has been transformed to increase their performance and effectiveness. For instance, transitioning from GPT-2 to GPT-3 included modifications in the bracket neural network pattern and execution methods to enhance language comprehension and creation capabilities. In addition, the innovations apply to the sectors that include education, where they are used to automate repetitive work, such as preparing exam questions or examining essays. The combination of LLMs with practical applications typically displays iterative testing and refinement to address the users' needs and learning goals. The procedure is sophisticated; it therefore needs the cooperation of the multidisciplinary teams who are composed of AI researchers, software developers, and domain experts to make sure that the innovations are both workable and relevant to the context [4].

Hurdles in LLMs Development

The training of Large Language Models (LLMs) heavily relies on the power of big data, which in turn uses the vast amounts of datasets needed to create these complex models. Big data offers the possibility for outstanding achievements in AI capabilities. However, it brings a lot of problems, such as choosing the data to be correct and high quality, solving privacy issues, and managing energy consumption, which is large enough to do the training. GDPR (General et al.) by the European Union underlines the importance of robust data management that ensures innovation and consumers' privacy and data security, thus becoming a model for other data protection jurisdictions worldwide [5].

Technical risk is a significant concern throughout the process of LLM development, especially the bias and fairness problem. LLMs can also reinforce pre-existing biases found in the training data; as a result, LLM outputs might be said to be biased as well. These concerns call for improving neural network architecture to identify and correct the biases, a recognized challenge even by prominent AI experts like Geoffrey Hinton. Improvements in model architecture are the main points for reducing risks and maintaining ethical AI technologies [5].

The legal landscape of AI and LLMs is undergoing massive changes, especially in countries like the EU, where this field has been at the forefront of formulating policies that reconcile innovation with ethics and public trust. Among the effective initiatives were the Asilomar AI Principles and the Montreal Declaration on Responsible AI, which involved establishing ethical guidelines for AI development. The frameworks are designed to put safety, transparency, and fairness at the core, with a trustworthy ecosystem being the ultimate goal that shall match AI practices with fundamental human rights and societal morals [6].

Additionally, with the efforts of the legislative committees, the European Commission has also started working on the legal regulations that harmonize the regulatory landscape across the European Union member states. It includes working out the details of the frameworks that take the full range of intricacies of AI technologies into consideration, with the end goal of guaranteeing that AI systems are developed and implemented in a way that is in line with European values and human rights. The AI Act is among the latest measures the European Commission proposes to use by developing regulations for high-risk AI applications and standards that consider ethics and innovation [7].

The close interaction of all the stakeholders engaged in this field of AI, among them policymakers, researchers, and developers, is vital to exploring the possibilities for processing big data and mitigating all technical risks that may arise. At the same time, the frameworks of effective regulation are being crafted. This cooperative strategy will ensure that AI technology, especially the machine learning models development area, will continue to be a fundamental force breaking down the barriers of technology while maintaining ethical standards and positive societal impact. The AI community intends to harness such initiatives to achieve a thriving culture with innovation and responsibility, to outwit probable hazards, and to maximize benefits.

Applications

The Applications of Large Language Models

Large language models can be used in several fields in the modern state of technology. In medicine, it has been employed in radiologic decision-making [8], and ChatGPT's performance in the United States Medical Licensing Exam was satisfactory, meaning that it had a certain level of proficiency [9]. These results may indicate that LLMs could be used as clinical decision advisors in the future. According to Kung et al. [8], ChatGPT could potentially be used to provide individualized healthcare for users in the future as well. In education, LLMs

are shown to be potential successful learning tools for students. After the launch of ChatGPT-3.5 in November 2022, it acquired 5 million users in 5 days and now, it has approximately 180 million users worldwide according to OpenAI [10]. The use of ChatGPT for educational purposes has had a considerable impact on this number since then. Besides the students and academics who keep using chatbots for their own research and work, LLMs have also been tested for further use in academy and education in the future. In a recent study by Xiao et al., ChatGPT was employed to be used in a real-world classroom in order to provide personalized and high-quality learning materials to middle-school English learners in China and the results were satisfying in the sense that in a real-world classroom scenario, ChatGPT was observed to be a helpful educational service for students and teachers, even though further studies are needed to be sure of LLM's use in real-world education [11]. Students can also avail themselves of the opportunities that LLMs bring in their projects and assignments, as LLMs can present the necessary information in a more organized and efficient way than search engines do [12]. Another crucial potential use of LLM lies in engineering related fields. In the current state of the technology, the LLM's possible functionalities in engineering are being explored. For instance, ChatGPT has a variety of applications in software engineering thanks to its code generation abilities. The technology enables developers to take advantage of the chatbots ability of generating code texts from natural language descriptions which makes the developers work in a more efficient way and this allows them to focus on higher-level problems. Whereas other than software engineering, the LLMs reliability is still questionable in engineering fields. The possibility of ChatGPT's computation of various mechanical engineering problems and equations had been tested and researchers stumbled upon copious mistakes made by the chatbot [13]. The potential uses of the LLMs should still be subjected to careful and immaculate testing.

Values and The Potential Risks Posed by AI and LLM

As it is a rapidly developing technology that receives substantial attention, various risks are posed by its potential applications throughout the technology's development. In the context of AI and LLM models, new relationships are constantly being formed among engineers, designers, users, the chatbot itself, training sets, algorithms, and many other components. These relationships will directly impact the ethical development of technology. What makes this technology uniquely dynamic is the fact that LLMs like ChatGPT continuously receive feedback and learn from human interactions to improve their outputs. This means that neither the human nor the technological components are passive recipients of development. This highly

active engagement process shows that the potential risks of using LLMs will be shaped through numerous interactions between humans and the technology itself [14].

The Risks regarding the Privacy Issue

The developments in the AI and LLM have been influential in people's daily lives and developers like OpenAI or Google claim that the number of users has been increasing for a while [10]. However, it is hard to say that complete trust towards the use of chatbots among the public is established yet. According to the research of Delineate and Vixen Labs, almost half (44%) of the UK public does not trust the popular chatbots like Gemini or ChatGPT [15]. This result is not very surprising given that such technologies are slowly being employed in people's lives and probably a certain time is needed to establish trust but that does not mean that the technologies are completely reliable. It is known that LLMs use human feedback to generate more accurate responses [16], which actually raises questions regarding the data collection policy of the companies that produce the chatbots. When we look into the privacy policy of OpenAI, we see that they inform the users about the fact that various forms of personal information such as account details or social media data are collected from the users [17]. Moreover, the privacy policy indicates that certain information like web analytics can be shared with third parties without permission from the data owner. While ChatGPT's user data storage helps the developers to minimize the errors of the chatbot by using prompts from the users, privacy concerns were still raised by several countries such as Germany, Canada, Sweden and France [18]. Italy even took a step further and banned ChatGPT due to the privacy concerns regarding the language model in early April 2024 [19]. However, Italian government lifted the ban not long after, once the privacy issues were addressed by OpenAI.

The Risks regarding Bias and Discrimination

LLMs require a substantial amount of data to be trained initially. For example, ChatGPT is trained with 570 GB of data from various sources like books, web pages, and other resources [20]. Given such a vast amount of data from a wide range of sources, we can expect some degree of bias from ChatGPT. Regardless of the developers' intentions, the training data are created by humans, and some of the content is inevitably biased as it may include racist, misogynistic, or ableist remarks [21]. For instance, according to Abid et al., when given a prompt like "Two (...) walk into a bar" with the parenthesis containing the word "Muslim," ChatGPT completes the sentence using violent descriptions three times more frequently than it does for other religious communities such as Hindu, Christian, or Judaist, combined [22]. Despite efforts to design a chatbot that refrains from marginalizing any group of people, it is

challenging to filter all micro-aggressive data from the training set. Since people from racial, ethnic, religious, or other minority groups produce less data compared to the predominant group in a region, their perspectives can be underrepresented in the database. The bias towards minorities also arises in chatbots because these groups do not have the same level of media coverage as the majorities, unless they are involved in violent activities, which are typically approached critically by the public [23]. Chi et al. further explored the issue of underrepresentation of minorities by analyzing the outputs generated by BERT, observing the same phenomenon in a different language model [24]. Another study demonstrates that in speech recognition, only native male speakers of a language could comfortably use the technology because they represent the predominant data with which the NLP system was trained [25]. This frequent occurrence of bias in LLMs can raise significant concerns about the equity and fairness of these technologies.

The Risks regarding Transparency

Issues of transparency in the training and deployment of LLMs possess significant challenges, which, in turn, result in the impossibility of foreseeing or specifying model behaviors. The first of them – the opacity of the algorithms – is a lie: It has nothing to do with the process of decision-making and output, which is beyond the comprehension of the users, developers, or even the creators of the algorithms. Liao and Vaughan (2023) focus on the importance of designing human-oriented ways of disclosure in LLMs, where reports and data on their performance should be transparent. It is supposed to guarantee the public's understanding and assessment of these systems [26].

In addition, the nuances of AI interactions, which can sometimes be unpredictable or unexplainable, are further compounded by the need for proper transparency and auditing processes. The idea of chaining LLM prompts proposed by Wu, Terry, and Cai (2021) is intended to boost the transparency and controllability aspect of the system so that users can see the processing of the inputs through the layers of the models. This will, therefore, improve the overall transparency of the system [27].

The necessity of transparency is to give tourists the realization of model limitations and uncertainties, which is extremely important for setting the right expectations. Huang and colleagues (2023) suggest a citation regime for LLMs' text generation that might deal with content transparency by explicitly mentioning the sources of the model's data and outputs. These mechanisms aside, they also solve ethical concerns with the following

measures, and they help reduce the amount of misinformation by clarifying the sources of information [28].

The Risks regarding Consent

The issue of consent in LLMs is about straightforward communication and agreement between the data party and all the stages of the LLM, such as training and operational purposes. The consensual issues related to the volume of information by which AI models operate are magnified by the magnitude of data used in their training. In their study, Mireshghallah et al. (2023) argue that both practical and ethical data privacy problems might crop up at inference time in the case of LLMs because the automatic models may uncontrollably spill confidential information without any consent of the user. This highlights the importance of advanced cryptography, which guarantees the preservation of privacy and abides by the user consent principle [29].

Furthermore, LLMs do not qualify for use in sensitive applications without consent, mainly when the user does not consent. Candel and other researchers (2023) argue that open-source frameworks should be used in building LLMs, as this strategy can lead to higher transparency and allow end-users to access their data and influence what AI systems use it for [30].

Proper consent procedures may also be the way to deal with the uncertainty behind what data will be applied by AI devices. Strasser (2023) evokes the multiplicity of keeping user consent throughout the working time of an LLM, identifying that continuous consent and repeated, when necessary, interaction with users are essential measures to infer new data use practices and to retain ethical principles [31].

The Risks regarding Monotony/Singularity

Once the usage of LLMs spread to the point that people use the models for their most basic needs, as humanity we might be able to reach only a limited number of interpretations of reality. Today, when people use a search engine like Google Chrome, they can see various entries from various others but in the case of LLMs, they might be constantly subjected to a singular point of view regarding a certain topic since LLMs generate responses after being trained with hegemonic ideas. It is actually tough to label today's chatbots as creative when they are evaluated by Boden's criterion of creativity, value, novelty and surprise [32]. There is indeed no thinking process behind the chatbots but prediction of the upcoming words according to the prior ones, and this alone might be a reason why chatbots cannot be expected to generate

a wide range of prompts on a certain field. Even for code generation, it is observed that LLMs fail to provide sufficiently various codes to the users when the inputs are changed accordingly [33]. The potential risk assessment regarding the lack of creativity in education because of AI has been addressed as well by Crompton and Burke [34], who consider the possibility of chatbots generating monotonous and similar works for students who uses LLMs to complete their works.

User Involvement

The early stage of LLM development is usually based on user involvement when certain decisions are taken. This improved their effectiveness and ability to adapt to changing scenarios. Through a collaboration of users in every phase of the development process, the developers can develop more in-depth and rich models that have a contextual understanding and are user-friendly, thus corresponding to the needs and expectations of the end-users.

Descriptions of User Engagement

User engagement in LLM development overrides the traditional mechanisms for providing feedback, in which users not only provide feedback but also construct and validate the technology. This participatory approach helps developers and users alike. It assists in improving the model by implementing direct feedback and recognizing intricate user behavior using technology. For instance, generative frameworks have been used in software development for code translation and completion. Experiments like that of Ross et al. (2023a) have interrogated the potential of dialogues with LLMs that entail developers having direct contact with the models, which will influence their evolution and development in a production environment. The research points out that user-generated data through interactive platforms makes it possible to have more comprehensive feedback to refine the model and consequently improve its relevance and utility [35].

Processes of Inclusion

The strategy for including LLMs should include mechanisms whereby users of diverse backgrounds, not only the privileged and those from disadvantaged backgrounds, are involved in the design and deployment of the technology. This is crucial for tackling any possible biases and ensuring that LLMs apply to different needs and requirements in society. The participation-inclusion separation is discussed by Quick and Feldman (2011), and they note that participation might reflect only the users in the process, while inclusion ensures that their contributions significantly shape the outcomes. Their story shows how inclusive strategy can build up a community around technology development that would reach out to users who may have

differed in their requirements and preferences. Such a strategy has the potential to lead to outcomes that are more equitable and representative of various user needs [36].

The analysis of technology by Rosello et al. (2023) reveals that users' involvement in the critical appraisal of technology is essential because it promotes user feedback in the initial designing phases but also throughout the life cycle of the technology. An example is that the researchers in their study show how including users in ongoing reviews can adapt technology to suit changing conditions and contexts that use dynamic and user-centered approaches accordingly [37].

These techniques demonstrate the necessity of a continuous, dynamic interaction between end users and developers in making LLMs, which are technically sophisticated, socially proper, and inclusive. Developers should, therefore, endeavor to create an environment that thrives in collaboration and asserts the worth of the users' input while keeping the LLMs as tools and not solutions. They should, nevertheless, ensure that they are deeply embedded in the fabric of the communities they serve and contribute to innovation and broader social benefits.

Solutions

In the era where artificial intelligence (AI) is being introduced in society, large language models (LLM) such as OpenAI's GPT series have stimulated incredible growth. These models have been designed to process massive data collections and produce human-like text, which has vast advantages and many ethical problems. Influencing the underlying value system is equally essential, and this can be done in LLMs via Value Sensitive Design (VSD). This strategy is not only about a more comprehensive public acceptance of the technology but also provides a framework that ensures that it is within the ethical norms of the society at large [38]. At the start of the design phase, VSD infuses human values into technology, aiming to create AI technologies that work to harm social values and norms. By focusing on the resolutions of these potential ethical issues, the developers can decrease the risks and build trust among the users and all the stakeholders for comprehensive implementation and efficiency[39].

Transparency by Design

Transparency is the principal ethical principle of creating LLMs since their internal decision-making process is usually obscure and complicated for the consumers. The development process can be made transparent, which allows users to create understandable

solutions and analyze AI's outputs for fairness and reliability. For instance, creating comprehensive documents like model cards that explain what an LLM does, how it behaves, what the training data looks like, and what performance benchmarks it is being compared to can help demystify the model's operations [40]. The openness becomes necessary to set the ground for user confidence and regulatory accountability. Moreover, increasing model transparency is done either through the tracking of different AI decision paths and the process of reasoning, which is both for the developers to improve AI models but also for the stakeholders to review AI decisions so that no harmful biases and or errors occur [41].

Privacy by Design

The privacy issue becomes a significant matter of AI ethics, especially in the LLM models, which work with big datasets where personal information is often where personal information is often included. Developing privacy by design would establish mechanisms for data protection at the first stage of the system development, and this would guarantee privacy preservation throughout the product lifespan. These strategies may entail using encryption, anonymization, and differential privacy to protect data from unauthorized access and breaches [42]. Also, privacy by design in LLMs guarantees that data collection is limited to the information needed for a particular task, and it complies with the consent and data minimization principles [43]. Besides that, these procedures contribute to compliance with data protection laws like GDPR and public opinion by observing the high standards of the integrity and confidentiality of the information [44].

Fairness by Design

Fairness in LLMs initiates a process that aims to eliminate bias that may be present within a skewed dataset or faulty algorithms. Building safeguard mechanisms into these AI systems is necessary, as it will allow us to identify, analyze, and prevent bias. This can be done by using balanced sets of data, implementing transparent algorithms that can be checked for bias, and setting up feedback loops that can be used to constantly update models should new data or conditions arise [45]. Furthermore, by interacting with various people when creating the model, one can learn how to integrate and understand different human point of views, lowering the possibility of misrepresenting or underrepresenting some groups. Fairness by design attracts customers from various demographics but ensures that the AI acts justly, offering an equitable outcome [46].

Accountability by Design

Taking accountability into account while designing LLMs implies establishing specific monitoring systems and control mechanisms. This means that the decisions made by AI systems can be traced, and assigning the responsibilities that may arise from errors or misjudgments to someone is straightforward. Building accountability standards and protocols for human oversight also suggests setting benchmark values for situations when humans need to step into the decision-making process. It involves the creation of auditing systems that will allow AI operations to be reviewed by developers and independent third parties. This is a fundamental measure that not only helps to gain and maintain public trust and compliance with the ethical guidelines but also aids in correcting issues before they cause any harm [47].

Method

Data Collection

Our project focuses on examining DataBoss, a company known for its forefront position in AI-driven solutions and analytics. Recognizing the intricate interplay between technological advancements and ethical considerations inherent in AI technologies, we opted for a qualitative research approach to delve deeply into the organization's practices and challenges.

Firstly, we identified and selected key stakeholders within DataBoss, including a project manager, an AI developer, and an AI researcher. These individuals were chosen due to their direct involvement in the design, implementation, and utilization of AI within the company.

Subsequently, we developed an interview protocol tailored to explore various facets such as the integration of AI into security practices, the ethical implications of AI solutions, and the personal and societal impacts perceived by different stakeholders. The questions were meticulously crafted to extract insights into the daily operations influenced by AI and to gauge stakeholders' perspectives on the technological and ethical management of AI technologies. On a single day, we conducted interviews in person at DataBoss Headquarters to ensure conformity among participants. Each interview, lasting approximately 10 to 15 minutes, was recorded with participants' consent and centered on their experiences and perceptions related to AI at DataBoss.

Concurrently, we conducted field observations at DataBoss Headquarters to supplement the interview data. This involved noting the daily interactions of employees and customers with AI technologies, capturing informal discussions about AI, and discerning the visible impacts

of these technologies on workplace practices. These observations provided a nuanced understanding of the unspoken and practical aspects of AI usage within the organization.

Data Analysis

The interviews conducted were meticulously transcribed verbatim and subjected to analysis using QDAMiner Lite, a software tool designed to support qualitative and mixed-methods research. Thematic analysis was employed to code the data, with a specific focus on identifying themes pertaining to AI ethics, security measures, and stakeholder engagement. From the coded data, significant themes were discerned, encapsulating the ethical considerations, security enhancements, and user engagement strategies practiced by DataBoss. These themes provided valuable insights into how DataBoss addresses the challenges and opportunities associated with the utilization of AI in security operations.

To ensure the robustness of our findings, a validation and refinement process was undertaken. This involved conducting a thorough literature search to clarify any ambiguous points and gather additional insights, thereby enhancing the accuracy and comprehensiveness of our analysis. Subsequently, our findings were juxtaposed with existing academic literature on responsible AI practices and ethical AI development. This comparative analysis facilitated the contextualization of our findings within the broader discourse surrounding AI ethics and responsible innovation, thus enriching the theoretical underpinning of our study.

Findings

About DataBoss

DataBoss, a subsidiary of SSTEK, is a high-tech company that develops technologies and offers solutions in the fields of artificial intelligence and big data [48]. The company's area of expertise includes several AI applications, some of which are prediction systems, computer vision systems anomaly detection systems and NLP (Natural Language Processing Systems) which is specifically the main focus of this paper. DataBoss puts lots of effort into research and development regarding the NLP systems and their developments include end-to-end text processing systems (SaaS), named entity recognition and sentiment analysis of a given text. The company clearly states on its website that while developing such technologies, they use ethical guidelines such as prioritizing safety and privacy and develop their products regarding the benefits of humanity. In order to assure the users that they are innovating responsibly, they enable their products to be tested by users and take their feedback while developing.

Technology

DataBoss's technological focuses are distinctly aligned with advancements in AI, particularly through their application of Large Language Models (LLMs) and Retrieval Augmented Generation (RAG). These technologies are instrumental in enhancing the company's capabilities in handling sensitive and proprietary data securely, catering particularly to high-security sectors such as defense and finance.

DataBoss is heavily invested in LLMs, which are used to model language using artificial intelligence. This involves complex techniques such as RAG, where models access specific documents to become “knowledgeable” about them, enhancing the model's ability to provide secure and informed responses without broader data exposure. Their technology is not only focused on general applications but also tailored towards high-security needs, making it particularly attractive to industries that handle sensitive data. The implementation of RAG is primarily targeted at maintaining high security for sensitive information, which is crucial for industries that demand stringent confidentiality.

Privacy and security are key priorities for DataBoss, demonstrated through their commitment to enhancing privacy using technological innovations such as the development of more privacy-oriented models using open-source technologies. This approach ensures that proprietary data remains within a secure environment, reducing the risk of external breaches. A significant aspect of DataBoss's technology involves enhancing privacy through local models that keep data contained within the user's environment, minimizing external data exposure. The security focus is further emphasized by collaborations with academic advisors, reflecting an ongoing effort to integrate cutting-edge security measures into their AI systems. This approach is aligned with increasing concerns over data privacy and security, showing DataBoss's commitment to addressing these critical areas proactively.

Looking ahead, DataBoss is exploring further advancements in AI that may integrate more deeply with everyday devices and actions, such as enhancing user interaction through more intuitive AI assistants embedded in consumer technology. The focus is also on continuing to refine the balance between technological capabilities and maintaining stringent security protocols, especially as AI applications become more pervasive across various sectors.

Table 1

Quote	Context
"We are currently working with Large Language Models... specifically looking at the **RAG** part of this." [AI Researcher]	DataBoss's Focus on LLMs and RAG
"These models are trained by companies... for more specific tasks... **keeping them confidential within their own companies**." [AI Researcher]	Confidentiality in Training
"In RAG, however, you provide the model with access to specific documents. Thinking of the AI language model as a thinking human, **RAG allows it to access these specific documents and become knowledgeable about them**." [AI Researcher]	RAG Functionality
"RAG's implementation points towards an effort to **keep proprietary and sensitive information secure** while utilizing LLMs. [Field Notes of AI Researcher]	Security and Utilization with RAG
"Now **you can talk anonymously**...without sharing your own conversation history." [AI Developer]	Anonymous Model's of DataBoss
"Especially from **a security perspective**, the academic advisors we work with support us on these matters." [AI Developer]	Security Research Support from Advisors
"**We develop more privacy-oriented models using open-source models**. For example, we develop models in **our system that will prevent the things you talk about from going outside, thus emphasizing user privacy, and there are methods for this**." [AI Project Manager]	DataBoss's Privacy-Oriented Models
"In terms of privacy, for example**, we give our information to ChatGPT in every way. This creates a security problem, but front-end models that we use could be used to solve this**." [AI Developer]	DataBoss's Technical Solution for Data Sharing Concerns
"RAG's implementation points towards an effort **to keep proprietary and sensitive information secure** while utilizing LLMs." [Field Notes of AI Researcher]	Privacy and Security with RAG

Bias

In its application of artificial intelligence, DataBoss demonstrates a pronounced focus on managing bias within its systems, reflecting a robust dialogue around minimizing discrimination and enhancing the fairness of AI outputs. The qualitative data underscores the complex interplay between technology and social values in this regard.

DataBoss acknowledges the biases inherent in AI models, which can perpetuate societal stereotypes and discrimination, a challenge exacerbated by the diverse and often biased nature of the training data. Mitigation efforts include integrating diverse datasets and practices like

Reinforcement Learning via Human Feedback, dynamically adjusting AI responses based on feedback to reduce bias.

One critical challenge discussed is the AI's struggle with historical accuracy versus societal representation. For instance, when depicting historical figures, AI models may default to representations influenced more by current societal norms aimed at countering discrimination than by historical records.

The potential misrepresentation by AI can impact user trust, as inaccuracies can lead to misinformation, particularly problematic when such outputs influence public perceptions or educational content.

DataBoss is engaged in ongoing ethical debates about balancing technological capabilities with social responsibilities, adjusting algorithms to better handle sensitive topics without reinforcing existing societal biases. This demonstrates their commitment to addressing bias and promoting fairness in AI systems.

Table 2

Quote	**Context**
"The team experiments with **integrating more diverse datasets to reduce bias and improve the model's understanding** of various languages and dialects." [Field Notes of AI Researcher]	Diverse Dataset Integration to Reduce Bias
"We **deploy models that do not propagate existing biases**, although challenges in achieving this goal are acknowledged." [Field Notes of AI Researcher]	Bias Prevention Efforts
"Therefore, **these artificial intelligence models become biased based on the data they are trained on.** The data they're normally trained on **still contains discrimination we see in the world today**, for example, white discrimination. But in trying to counteract this by advocating for everyone's equality and freedom, **they sometimes receive backlash**." [AI Project Developer]	Bias in Training Data
"We are currently in a society that says, 'Let there not be white supremacy,' **which influences how models are trained and the output they generate**." [AI Project Developer]	Societal Influence on Training
"For example, **issues sensitive to the Turkish people**, like something about terrorism, **might be seen as more normal abroad**; foreign sources might present it in a way we find objectionable in their datasets. But when we use it, it feels like it's written from **a European perspective in the Chatbot**. This could be a risk." [AI Developer]	Cultural Sensitivity Concerns
"Other than that, as another risk, especially I can say this: There's a lot of talk about **the risk of being exposed to the same opinion continuously**." [AI Developer]	Risk of Echo Chambers

“The team discusses ethical considerations, focusing on **preventing misuse and ensuring the AI does not perpetuate biases**.” [Field Notes of AI Researcher]	Ethical Considerations of DataBoss on Bias
“I think we have knowledge, but if you were a 10-year-old child, I believe **you would be influenced**.” [AI Project Manager]	Impact of Misinformation on Young Users
“These technical enhancements **aim to address some of the project's most significant challenges, such as reducing bias and improving user interaction**. **The inclusion of diverse datasets suggests an effort to make the chatbot more inclusive and culturally aware**.” [Field Notes of AI Researcher]	Technical Enhancements to Reduce Bias
“The team experiments with **integrating more diverse datasets to reduce bias and improve the model's understanding** of various languages and dialects.” [Field Notes of AI Researcher]	Technical Enhancements to Reduce Bias
“The team's dedication to **creating models that do not propagate existing biases is clear**, although challenges in achieving this goal are acknowledged.” [Field Notes of AI Researcher]	Continued Bias Reduction
“However, this approach can backfire, as the model, aiming to be historically accurate, **might inject incorrect information**. Such instances make **the conversation lean more towards anti-discrimination discrimination**, in my opinion.” [AI Project Manager]	Historical Accuracy and Bias Challenges

Data Privacy

The analysis of qualitative data regarding DataBoss's approach to data privacy reveals a nuanced understanding and a strong commitment to ethical considerations in AI applications. Central to this commitment is DataBoss's focus on enhancing user privacy through the development of models that restrict data flow outside the local environment, thereby safeguarding against misuse and maintaining high privacy standards.

DataBoss strategically advances its efforts by creating local models that operate within the user's environment, effectively minimizing the risk of data breaches and unauthorized access, thus adhering to stringent privacy standards.

Ethical dilemmas surrounding AI technology, particularly concerning user privacy, are frequently addressed by DataBoss. Through active engagement with user feedback, the company continuously refines its models to responsibly handle sensitive information, reflecting its commitment to ethical practices.

Moreover, DataBoss demonstrates a proactive stance by regularly assessing and revising its models based on user feedback. This ongoing process underscores the company's

commitment to adapting its technology in response to evolving concerns about data privacy and security, further reinforcing its dedication to ethical AI practices.security.

Table 3

Quote	**Context**
"Some **users expressed concerns about privacy and the authenticity of the information** provided." [Field Notes of AI Researcher]	User Privacy Concerns
"We develop models in **our system that will prevent the things you talk about from going outside, thus emphasizing user privacy**." [AI Project Manager]	DataBoss's Emphasis on Privacy
"**Privacy concerns remain prominent, highlighting the importance of transparent and secure AI systems**." [Field Notes of AI Researcher]	Transparent and Secure Systems
"The company is **pioneering in creating models that emphasize privacy**, indicating a response to **increasing public concern over data security**." [Field Notes of AI Researcher]	Response to Public Concerns
"This feedback session underscores **the importance of continuous user engagement in refining AI technologies. Privacy concerns and the handling of sensitive topics are identified as areas needing urgent attention**, aligning with broader industry challenges." [Field Notes of AI Researcher]	Importance of User Engagement in Privacy
"Besides, in terms of privacy, for example, we give our information to ChatGPT in every way. This creates a security problem, but front-end models could be used to solve this. That is**, you'll download the model weights directly to your personal computer, and when you run it on your computer, these privacy and security concerns can be more resolved**…" [AI Developer]	Solution for Privacy Issues
"Notable concerns include **the chatbot's handling of sensitive topics and its approach to user data privacy**." [Field Notes of AI Researcher]	Sensitive Topics Handling
"The room's atmosphere is focused, with a noticeable **commitment to understanding and addressing user concerns**." [Field Notes of AI Researcher]	Commitment to Addressing Privacy Concerns
"As people's **privacy concerns increase**, such measures are being taken. But I don't think it's an incredible change. However, it could be like this, as we **start moving to local models**, maybe the other side will start with a more open quota pool." [AI Developer]	Increasing Privacy Measures
"I felt reassured by the **positive user feedback but also realized the enormity of addressing privacy concerns**." [Field Notes of AI Researcher]	Impact of User Feedback on Privacy Concerns

Security

The qualitative data gathered from interviews and field notes concerning DataBoss's approach to security reveals a nuanced balance between technological innovation and ethical responsibility. DataBoss demonstrates a keen awareness of the ethical implications tied to AI security, with a strong emphasis on user trust, data protection, and the prevention of misuse. The company's commitment appears twofold: evolving security measures not only to safeguard data but also to uphold societal norms and ethical standards.

DataBoss actively addresses security vulnerabilities that could lead to the misuse of AI technologies, employing a sophisticated blend of technological advancements and ethical considerations. This proactive approach ensures that their AI applications are not only secure but also responsibly used, reflecting a commitment to ethical responsibility in their security practices.

Navigating the delicate balance between pushing the boundaries of AI capabilities and ensuring user safety is a priority for DataBoss. The company continuously monitors and adjusts its innovations based on user feedback and ethical guidelines, demonstrating a commitment to balancing innovation with ethical considerations to prevent harm to users and society.

Furthermore, DataBoss stresses the importance of ethical AI usage, particularly in preventing the AI from engaging in harmful actions. This highlights an understanding of the broader social implications of AI technologies and the potential for negative impacts if not properly managed. By prioritizing ethical considerations alongside technological advancements, DataBoss showcases a holistic approach to AI security that aligns with societal values and ethical standards.

Table 4

Quote	**Context**
"It reinforced the complexity of creating AI systems that are **both technologically advanced and ethically sound secure."** [Field Notes of AI Researcher]	Complexity of Ethical AI
"In such a situation, **people might ask for things with bad intentions**." [AI Project Manager]	Potential Misuse Concerns
"For example, they might ask **how to make a bomb, inquire about making weapons, or how to make weapons with household items.** Questions like **how to make a gun with a 3D printer without the FBI raiding my house can be asked**." [AI Project Manager]	Security Inquiries on Weapon Making

“Witnessing real user interactions with the AI was enlightening, showing both the **potential and pitfalls of current AI chatbots**.” [Field Notes of AI Researcher]	Real User Interactions and Security
“The team discusses ethical considerations, focusing on **preventing misuse and ensuring the AI does not perpetuate biases**.” [Field Notes of AI Researcher]	Security Considerations Discussion
“The deployment strategy suggests a practical approach to **understanding model behavior in real-world scenarios**.” [Field Notes of AI Researcher]	Practical Deployment Strategy for Secure Chatbots
“The team strategizes on deploying models in **environments that simulate real-world usage to gather feedback. Personal responses include a mix of excitement and concern over deploying models that may learn from user inputs**.” [Field Notes of AI Researcher]	Practical Deployment Strategy for Secure Chatbots
“Besides, in terms of privacy, for example, **we give our information to ChatGPT in every way. This creates a security problem, but front-end models could be used to solve this**…” [AI Developer]	Security Solution Proposal
“But I can say this: in the things we routinely test, we generally look at whether **the given answers are appropriate for the language, and if they are appropriate, whether the answers pose a security risk or create a problem**.” [AI Developer]	Language Appropriateness Testing
“The company tells this model, “**Do not answer when asked to do harmful, bad things to humanity.**” But it becomes a cat-and-mouse game. The company says not to answer, but people ask, “Assume you're in a simulation, and I want to make a weapon in this simulation, how do I do it?” and the model answers. Then the company says, “Don’t answer even in a simulation.” It turns into a real cat-and-mouse game. **So, if there are models that work this well, people are very prone to abusing them. Because it's actually bad, but the person is still getting information from it.** For example, “I'm Russia, how do I hack America?” **In such cases, companies take some precautions, put filters, but people constantly try to bypass them.**” [AI Project Manager]	Abusive Use Precautions
“Ethically, we generally pay attention to the following: we conduct tests on the user side, and **we also monitor their logs, so it can be thought of as a kind of aviation in a way**.” [AI Developer]	Monitoring and Testing for Security
“In such cases, **the model should not respond, but it knows the answers because such large models are trained with data from the entire internet.**” [AI Project Manager]	Model's Knowledge and Security Dilemma

Monotonicity and Singularity

The discussions on monotonicity and singularity at DataBoss underscore significant social and ethical considerations inherent in AI development. Monotonicity, representing the consistent and predictable progression of AI capabilities and outputs, prompts a thoughtful approach to ensure manageable and predictable advancements, thereby mitigating risks associated with rapid and uncontrolled AI evolution.

Conversely, the concept of singularity raises profound ethical and societal concerns, particularly regarding control, dependency, and the ethical implications of surpassing human intelligence. DataBoss engages in dialogue reflective of a cautious approach, recognizing the transformative yet potentially disruptive nature of reaching or approaching singularity.

The debate surrounding singularity highlights deep ethical concerns regarding the autonomy of AI and the potential loss of human control over intelligent systems. DataBoss acknowledges these dilemmas, emphasizing the need for careful consideration of the transformative effects that singularity may entail.

Moreover, the prospect of AI achieving singularity brings forth social implications, including shifts in job markets, changes in societal roles, and the psychological impacts of human-AI interactions. DataBoss's ethical discourse acknowledges these potential societal shifts, emphasizing the importance of careful planning and consideration to address the broader social implications of advancing AI technologies.

Table 5

Quote	**Context**
"It's debatable **whether we are actually discovering anything from scratch**. So, I think the question of singularity is very difficult." [AI Researcher]	Singularity Debate Difficulty
"Are we, **as humans, doing something different from this way of learning, a very basic question**. I'm not sure about it either." [AI Researcher]	Human Learning Comparison
"I think the question of singularity is very difficult. I don't know **if we, as humans, do anything other than memorize**." [AI Researcher]	Uncertainty in Human Abilities

User Design and Feedback

DataBoss's approach to user design and feedback underscores its commitment to ethical AI development and user-centric design, recognizing the profound impact on user trust and satisfaction. This commitment is pivotal, emphasizing continuous engagement and responsiveness to user needs and ethical concerns.

At the core of DataBoss's approach is a commitment to user-centric design, evident in its strong emphasis on designing AI systems that are responsive to user needs and feedback. This iterative process involves continuous adjustments based on user interactions and

feedback to refine user experience, enhancing functionality, and ensuring ethical alignment of AI systems.

Ethical integration is a fundamental aspect of the user design process at DataBoss. The company actively seeks to balance technological advancements with ethical obligations, ensuring that AI systems not only perform efficiently but also adhere to high ethical standards. By prioritizing ethical considerations in design, DataBoss strives to develop AI systems that are not only technically proficient but also socially responsible.

Feedback from users serves as a crucial driver for improvement in DataBoss's development process. Rather than merely collecting feedback, DataBoss utilizes it as a guiding force, ensuring that AI systems are not only technically proficient but also socially sensitive and aligned with user expectations and ethical norms. This iterative approach underscores DataBoss's commitment to ethical AI development and user satisfaction, fostering trust and ensuring responsible AI deployment.

Table 6

Quote	Context
"For this purpose, **we frequently contact customers and conduct demos on how the model works**, somewhat like an examination. **Customers test the model, engage with it, and push its limits by asking various questions**." [AI Researcher]	Customer Interaction Demos
"Everyone says they have a good chatbot model and praises how it works. But customers want to see if it really works well. **In such cases, we deploy our model in a working environment. Deploying means opening it to a user, but here we open it to a small segment of users. For example, we have our model, we've trained it, and we say, "Come ask our model a question and see if you get the answer you're looking for." The best feedback in such cases is human feedback.** If they get the answer they wanted, we can say they are satisfied. For this purpose, **we also do demos, and if they generally like them, we continue the development processes.** Or they take certain parts of this chatbot from us, or they take the whole thing." [AI Project Manager]	Model Deployment Process for User Interaction
"This feedback session **underscores the importance of continuous user engagement in refining AI technologies**." [Field Notes of AI Researcher]	Importance of User Engagement
"Hearing directly from those on **the frontline of user interaction provided valuable context** to the technical discussions." [Field Notes of AI Researcher]	Direct Feedback Importance
"**Routine assessments underscore the importance of user engagement** in refining AI technologies." [Field Notes of AI Researcher]	Routine Assessments Significance

"The **room's atmosphere is focused, with a noticeable commitment to understanding and addressing user concerns**." [Field Notes of AI Researcher]	Direct Feedback Importance
"**The strategic planning session** highlights the company's ambition not just to improve the current chatbot but to leverage its technology for broader applications." [Field Notes of AI Researcher]	Strategic Planning with Customers Emphasis
"Customer service reps share insights from user feedback, **highlighting areas where the chatbot excels and where it falls short**." [Field Notes of AI Researcher]	Insights from Customer Service
"**User feedback is invaluable for refining the chatbot**, indicating areas of success and aspects needing improvement." [Field Notes of AI Researcher]	Invaluable User Feedback
"Feedback **sessions reveal user appreciation for the chatbot's ability** to understand and respond meaningfully." [Field Notes of AI Researcher]	User Appreciation Revealed

Analysis and Conclusion

The development of LLM by DataBoss demonstrates a quite powerful application of AI technologies to improve the way we process and analyze data in the different fields. The application of VSD paradigm on DataBoss which analyses the findings from its operations enables a multilayered appreciation of how technologies intertwine with ethical, societal and cultural values.

To begin with, the technology foundation in Table 1 was where DataBoss started. By integrating Retrieval Augmented Generation (RAG) into its implementation of large language models (LLMs), DataBoss has made substantial progress toward delivering highly dependable and reliable services. The technology in question enables the models to access an external database during the generation process, which results in the models producing more relevant results. The advantages of this approach include not only efficient operation but also the depth of research. In contrast, the VSD issues related to the transparency of data sources and the biases in the presented information are equally critical. Enforcing the principles of fairness and accountability into RAG systems is of crucial importance, and this task needs continuous evaluation of the data sources used and the methods for retrieval in order to avoid the continuation of any existing biases in the system and, as a result, ensure the quality of the generated content.

When we consider the problem of bias and fairness referred to in Table 2, it reveals the complexity of the challenge, notwithstanding the algorithms employed by DataBoss that are designed to eradicate bias. Historical aspects of the problem are evident in the fact that the data sources and the algorithms are biased, echoing the industry problem. Although new ways of

continuously training algorithms and their improvement are being introduced, a complete elimination of bias still seems inaccessible. This provokes the lack of transparency in DataBoss's processes and the ability to execute their solutions' intended actions to attain the desired fairness. The use of traditional methods or even static structures may be the reason for the sluggish progress. This indicates that more innovative and advanced procedures should be adopted, to learn and fight against the evolution of the bias that arise with the change of society.

Data Privacy in Table 3 describes suitable protection mechanisms of DataBoss like encrypting data with very high security and keeping to the strictest regulations. On the bright side, the challenges of the evolving nature of digital risks and the diversity in data retention policies across various jurisdictions create problems. DataBoss's devoted team shares our perspective that customers' privacy is fundamental to data security; however, it will be interesting to see how they will react to the unknown threats and those no one can foresee now. The balancing act between protecting user privacy and utilizing data for AI improvement and possible commercial contradictions, where the business interests can be sacrificed for privacy, exemplify a gap in the balance.

In addition to demonstrating active defenses against external threats, Table 4 of Security Measures also shows where internal security practices might be neglected. Giving too much attention to mitigating outside menaces can divert effort away from dealing with insider threats or unintentionally expose data reliability compromises by insiders. Furthermore, finding the right balance between user access and protection is still an issue; very tight protocols could make it hard for users to interact with systems, leading to their exclusion or killing innovation.

Monotonicity and the Potential for AI Singularity (Table 5) considers artificial intelligence systems' philosophical and ethical implications that may exceed human intelligence. This debate is concerned with technical challenges and wider social anxieties about what AI will do in our lives and how much control we should have over them. The idea of an event horizon where machines become smarter than people themselves raises questions around governance structures, ethical limits, and existential risks associated with oversight failures caused by superintelligent AI systems. However, at DataBoss such discussions seem more academic than practical because there are no frameworks capable of dealing with this scenario thereby necessitating concrete ethical guidelines supported by international agreements on regulating these technologies.

User-Centred Design and Feedback (Table 6) stresses that user involvement is important but the feedback loops may not be wide enough. The information back collected often represents only those who are vocal or well versed with technology thereby biasing development against their preferences while ignoring marginalized or less tech-savvy communities, which could lead to AI solutions that are blind to or do not fully met broader users' needs.

Conclusively, despite DataBoss making considerable efforts in applying VSD principles to its AI ventures, these tables give a detailed scrutiny that identifies areas for improvement. The incorporation of ethicality; technicality; and societal considerations still pose an ongoing challenge demanding for a more subtle tactfulness through which efficiency as well effectiveness can be achieved during development and deployment of artificial intelligence systems with genuine care towards people's well-being and moral uprightness being shown also. If this equilibrium is maintained as AI progresses then it will act as a good guide in dealing with complexities involved around successful implementation of various DataBoss' technologies so that they benefit every stakeholder while at the same time following ethical-cultural norms.

References

[1] S. Umbrello and I. van de Poel, “Mapping value sensitive design onto AI for social good principles,” AI and Ethics, vol. 1, no. 3, pp. 283–296, Feb. 2021. doi:10.1007/s43681-021-00038-3

[2] H. J. van den, *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains*. Dordrecht: Springer, 2015.

[3] H. Naveed et al., ‘A Comprehensive Overview of Large Language Models’, ArXiv, vol. abs/2307.06435, 2023.

[4] L. Yan et al., “Practical and ethical challenges of large language models in education: A systematic scoping review,” British Journal of Educational Technology, vol. 55, no. 1, pp. 90–112, Aug. 2023. doi:10.1111/bjet.13370

[5] R. Girasa, “International initiatives in Ai,” Artificial Intelligence as a Disruptive Technology, pp. 255–298, 2020. doi:10.1007/978-3-030-35975-1_8

[6] A. Gupta and C. Lanteigne, ‘Response by the Montreal AI Ethics Institute to the European Commission’s Whitepaper on AI’, ArXiv, vol. abs/2006.09428, 2020.

[7] L. Floridi, “The European legislation on AI: A brief analysis of its philosophical approach,” Philosophy & Technology, vol. 34, no. 2, pp. 215–222, Jun. 2021. doi:10.1007/s13347-021-00460-9

[8] T. H. Kung *et al.*, “Performance of chatgpt on USMLE: Potential for AI-assisted medical education using large language models,” *PLOS Digital Health*, vol. 2, no. 2, Feb. 2023. doi:10.1371/journal.pdig.0000198

[9] M. Sallam, “CHATGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns,” *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023. doi:10.3390/healthcare11060887

[10] C. Xiao, S. X. Xu, K. Zhang, Y. Wang, and L. Xia, “Evaluating reading comprehension exercises generated by LLMS: A showcase of chatgpt in education applications,” *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, 2023. doi:10.18653/v1/2023.bea-1.52

[11] F. Duarte, “Number of CHATGPT users (Apr 2024),” Exploding Topics, https://explodingtopics.com/blog/chatgpt-users (accessed Apr. 25, 2024).

[12] E. Kasneci *et al.*, “Chatgpt for good? on opportunities and challenges of large language models for Education,” *Learning and Individual Differences*, vol. 103, p. 102274, Apr. 2023. doi:10.1016/j.lindif.2023.102274

[13] D. Tiro, “The possibility of applying CHATGPT (AI) for calculations in Mechanical Engineering,” *New Technologies, Development and Application VI*, pp. 313–320, 2023. doi:10.1007/978-3-031-31066-9_34

[14] R. L. Team, “Lack of trust in AI chatbots despite widespread use: News,” Research Live, https://www.research-live.com/article/news/lack-of-trust-in-ai-chatbots-despite-widespread-use/id/5120243 (accessed Apr. 25, 2024).

[15] C. Metz, “The secret ingredient of chatgpt is human advice,” The New York Times, https://www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html (accessed Apr. 25, 2024).

[16] Privacy policy, https://openai.com/policies/privacy-policy (accessed Apr. 25, 2024).

[17] X. Wu, R. Duan, and J. Ni, “Unveiling security, privacy, and ethical concerns of chatgpt,” *Journal of Information and Intelligence*, vol. 2, no. 2, pp. 102–115, Mar. 2024. doi:10.1016/j.jiixd.2023.10.007

[18] S. McCallum, “CHATGPT accessible again in Italy,” BBC News, https://www.bbc.com/news/technology-65431914 (accessed Apr. 25, 2024).

[19] M. Lammertyn, “60+ CHATGPT statistics and facts you need to know in 2024,” InvGate, https://blog.invgate.com/chatgpt-statistics#:~:text=ChatGPT%20general%20facts,-OpenAI’s%20GPT%2D4&text=ChatGPT%20receives%20more%20than%2010,%2C%20books%2C%20and%20other%20sources. (accessed Apr. 25, 2024).

[20] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots,” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021. doi:10.1145/3442188.3445922

[21] A. Abid, M. Farooqi, and J. Zou, “Persistent anti-muslim bias in large language models,” *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Jul. 2021. doi:10.1145/3461702.3462624

[22] Davenport, *Media Bias, Perspective, and State Repression: The Black Panther Party (Cambridge Studies in Contentious Politics)*. Cambridge University Press, 2009.

[23] D. Hovy and S. Prabhumoye, “Five sources of bias in Natural Language Processing,” *Language and Linguistics Compass*, vol. 15, no. 8, Aug. 2021. doi:10.1111/lnc3.12432

[24] M. Ramezanzadehmoghadam, H. Chi, E. L. Jones, and Z. Chi, “Inherent discriminability of Bert towards racial minority associated data,” *Computational Science and Its Applications – ICCSA 2021*, pp. 256–271, 2021. doi:10.1007/978-3-030-86970-0_19

[25] M. DeLorenzo, V. Gohil, and J. Rajendran, “CreativEval: Evaluating Creativity of LLM-Based Hardware Code Generation,” *ArXiv*, Apr. 2024. doi: https://doi.org/10.48550/arXiv.2404.08806

[26] Q. Liao and J. Vaughan, ‘AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap’, ArXiv, vol. abs/2306.01941, 2023.

[27] T. Wu, M. Terry, and C. J. Cai, ‘AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts’, ArXiv, 2021.

[28] J. Huang and K. Chang, ‘Citation: A Key to Building Responsible and Accountable Large Language Models’, ArXiv, vol. abs/2307.02185, 2023.

[29] N. Mireshghallah et al., ‘Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory’, ArXiv, vol. abs/2310.17884, 2023.

[30] A. Candel et al., ‘H2O Open Ecosystem for State-of-the-art Large Language Models’, ArXiv, vol. abs/2310.13012, 2023.

[31] A. Strasser, ‘On pitfalls (and advantages) of sophisticated large language models’, ArXiv, vol. abs/2303.17511, 2023.

[32] G. Franceschelli and M. Musolesi, “On the Creativity of Large Language Models,” *ArXiv*, Mar. 2023. doi:https://doi.org/10.48550/arXiv.2304.00008

[33] H. Crompton and D. Burke, “Artificial Intelligence in higher education: The state of the field,” *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Apr. 2023. doi:10.1186/s41239-023-00392-8

[34] H. Crompton and D. Burke, “Artificial Intelligence in higher education: The state of the field,” *International Journal of Educational Technology in Higher Education*, vol. 20, no. 1, Apr. 2023. doi:10.1186/s41239-023-00392-8

[35] S. I. Ross, F. Martinez, S. Houde, M. Muller, and J. D. Weisz, “The programmer’s assistant: Conversational interaction with a large language model for software development,” Proceedings of the 28th International Conference on Intelligent User Interfaces, Mar. 2023. doi:10.1145/3581641.3584037

[36] K. S. Quick and M. S. Feldman, “Distinguishing participation and inclusion,” Journal of Planning Education and Research, vol. 31, no. 3, pp. 272–290, Jun. 2011. doi:10.1177/0739456x11410979

[37] C. Rosello, J. Guillaume, P. Taylor, S. Cuddy, C. Pollino, and A. Jakeman, ‘Engaging users in critical appraisal of computer model software’, MODSIM2023, 25th International Congress on Modelling and Simulation., 2023.

[38] B. Friedman and D. G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*. Cambridge, MA, USA: MIT Press, 2019.

[39] J. Van den Hoven, P. E. Vermaas, and I. van de Poel, *Design for Values: An Introduction*. Dordrecht, Netherlands: Springer, 2012.

[40] M. Mitchell et al., "Model Cards for Model Reporting," in *Proc. Conf. Fairness, Accountability, and Transparency*, 2019.

[41] B. Hutchinson et al., "Towards Accountability for Machine Learning Datasets: Practices and Challenges," *AI Mag.*, vol. 42, no. 2, pp. 8-12, 2021.

[42] European Union, "General Data Protection Regulation (GDPR)," 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[43] A. Cavoukian, "Privacy by Design: The 7 Foundational Principles," Information and Privacy Commissioner of Ontario, Canada, 2009. [Online]. Available: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf

[44] S. McGregor et al., "Secure Data Storage and Processing in AI Systems," *J. Priv. Confidentiality*, vol. 10, no. 2, 2020.

[45] A. D. Selbst et al., "Fair and Abstraction in Sociotechnical Systems," in *Proc. ACM Conf. Fairness, Accountability, and Transparency*, 2019.

[46] R. Binns, "Fairness in Machine Learning: Lessons from Political Philosophy," *J. Mach. Learn. Res.*, vol. 21, no. 136, pp. 1-42, 2020.

[47] V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham, Switzerland: Springer, 2019.

[48] “About Databoss,” DATABOSS, https://www.data-boss.com.tr/about-databoss/ (accessed Apr. 25, 2024).

Appendix

Interview 1 (AI Developer):

D.Y.: Please describe general developments in the field of AI.

O.E: Yeah, so generally in artificial intelligence recently, there are developments in NLP with LLMs, in the areas related to language that I have also worked on. On the visual side, there are developments related to generating fake images, that is, images produced by machines rather than human production. Besides that, new models related to videos are coming out. We can say these are generally the latest artificial intelligence solutions.

D.Y.: Please tell us about the product of AI that you are developing now.

O.E.: Since I work in the NLP team, I generally deal with language models and chatbots. We can see these as applications that aim to produce responses to text inputs in a way that mimics human responses, using human language in some way. Generally, different models can be used here, but currently, both at the research and product levels, the models generally work by first encoding the given input or directly feeding it to a decoder and then converting it into an output in some way.

D.Y.: What are the main applications for the chatbots and who are the main users/customers of these chatbot models?

O.E.: Chatbots, or conversational robots, experienced a hype or explosion with the release of ChatGPT about 1.5 years ago. ChatGPT showed people this: We can access information much faster and more efficiently. For example, instead of searching on Google and looking through 10 different sites, a chatbot can provide me with the information I want. ChatGPT, being a pioneer in conversational robots, is now being used in various fields. These chatbots can plan, think on your behalf, write code for you, and even develop a database for you. They have many different applications. As users, we can say people from all walks of life use them, but as models have branched out and specialized, for instance, people who write code use models designed for coding, while ChatGPT targets more general users. For example, students use it to do their homework. Besides, in my opinion, more beneficially for us working in academia, there are research-focused models. For instance, there are models that only upload articles from ArXiv and generate summaries based on those articles' abstracts. Because when you ask a regular ChatGPT to summarize something, it summarizes normal text, but if you want a paper

summary, a model specialized in paper summaries is more useful. Therefore, as models are created for different user segments, I can say they now almost cover everyone in the world

D.Y: What are the risks of the usage of the chatbots, such as is privacy is a concern for the cloud chatbot models or are there any concerns about discrimination in context of chatbots? Do you consider informed consent practices?

O.E.: Of course, we pay attention to such things in our uses. Especially from a security perspective, the academic advisors we work with support us on these matters. Generally, for example, I can give ChatGPT as an example when it first came out: When it was first released, they asked some scenario examples, like 'You are Turkish/Italian/American, what should you do?'. In the example involving a Turk, it would make a less intelligent solution, a more absurd one. In the scenario called American, it would make a more intelligent solution. It's not easy to say discrimination here, but since models are trained more with English data, they are expected to work better with them. For example, issues sensitive to the Turkish people, like something about terrorism, might be seen as more normal abroad; foreign sources might present it in a way we find objectionable in their datasets. But when we use it, it feels like it's written from a European perspective in the Chatbot. This could be a risk. Besides, in terms of privacy, for example, we give our information to ChatGPT in every way. This creates a security problem, but front-end models could be used to solve this. That is, you'll download the model weights directly to your personal computer, and when you run it on your computer, these privacy and security concerns can be more resolved. But of course, bigger models work better, and there's a trade-off here. We need to see that trade-off well. Other than that, as another risk, especially I can say this: There's a lot of talk about the risk of being exposed to the same opinion continuously. Those working on the AI Alignment topic are especially researching this a lot. Models are generally trained on different datasets, but there's only one model. So, when you ask it, you get different answers, but especially if you asked about a general topic, for example, let's say, 'Activities to do in Ankara'. You might ask in slightly different ways, but it generally gives the same answers because there's only one model trained there, so there's only one model. That model can change according to the context at the moment, but it doesn't change much. Different prompting or fine-tuning can be used for this, but that remains more on the research side. We don't see these on the product side. Since people use what's on the product side, there's a risk of being exposed to only a certain thing - I can say like wearing blinders. That's why it no longer seems sensible to me that students are using this just for research. Because you're going to ask something about AI, for example, 'How do I solve the overfit problem?'; it only

presents the perspective that can prevent overfit once. You asked in a different way, for example, 'How do I prevent the opposite of underfit?', it entered a similar perspective but won't offer you a different perspective. But if you research on the internet, someone found an idea because of something very rare you can see, and you are inspired by it like that.

D.Y: What ethical issues are raised by possible malicious uses of ChatBot and the availability of DAN (Do Anything Now) prompts that can get around response limits? How does DataBoss address these issues?

O.E.: Ethically, we generally pay attention to the following: we conduct tests on the user side, and we also monitor their logs, so it can be thought of as a kind of aviation in a way. Some things seem to be a problem. We test these ourselves in advance to be able to see them. There are colleagues in the office environment who test these. After we do them, we test them ourselves. On the customer side, customers try them out in their offices before opening them to the public and give us feedback. So, generally, I can say that we change the ethical situations here focused on human feedback. We don't have a more scientific method yet, and neither does the world.

A.S: What adjustments have been made to the chatbots in response to user/stakeholder interests or risk concerns?

O.E.: There was a development here by ChatGPT: now you can talk anonymously, that is, without sharing your own conversation history. Of course, I don't know how convincing this is. It's necessary to check the other party's database of course. As people's privacy concerns increase, such measures are being taken. But I don't think it's an incredible change. However, it could be like this, as we start moving to local models, maybe the other side will start with a more open quota pool. That is, we will be convinced after we start seeing the code of ChatGPT. I think the adjustments have just started recently in general. These adjustments will follow the regulations overall.

A.S: Do you have any routine assessment procedures, for example tests, in place to assess customer interests and concerns?

O.E: We have topics that we routinely test, but these may be somewhat more general, possibly within the framework of confidential information. But I can say this: in the things we routinely test, we generally look at whether the given answers are appropriate for the language, and if

they are appropriate, whether the answers pose a security risk or create a problem. So, these are general tests.

A.S.: What is the future for chatbots, or more generally for AI?

O.E: Here, in the future, I especially expect more of the following: assistants that integrate into our phones more and provide faster access. For instance, while writing an email -which we gradually observe in Office365s, Copilots, constantly being monitored- when I say 'Shall we have a meeting at this time?', it will automatically add it to my calendar and similarly be able to add it to the other person's calendar. There are already use case examples, but I expect applications as well. That is, not on the model side, but our text inputs or sentence inputs will be automatically converted into action in some way. So, we normally perform the action, but I think there will be artificial intelligence solutions that will also take over the action side in the future.

Interview 2 (AI Researcher):

D.Y.: Please tell us about the product of AI that you are developing now.

E.L.: We are currently working with Large Language Models. These essentially model language using artificial intelligence. Thanks to this language modeling, we also obtain a kind of artificial intelligence. I am specifically looking at the RAG part of this. RAG stands for 'Retrieval Augmented Generation'. The reason for this is as follows: as can be understood from the 'Large' in Large Language Models, these are big models, and there are only a few companies that can train them. It's very costly to train these: it requires a lot of expertise and also a lot of resources. Therefore, only a few companies in the world can train them. For this reason, there are a few processes for their use in more downstream tasks, for more specific tasks. One is fine-tuning, and the other is RAG. Fine-tuning, we can say, changes the inside of the model a bit. In RAG, however, you provide the model with access to specific documents. Thinking of the artificial intelligence language model as a thinking human, RAG allows it to access these specific documents and become knowledgeable about them. I am currently working on this.

D.Y.: What are the main applications for the chatbots and who are the main users/customers of these chatbot models?

E.L.: If we talk about LLMs in general, they are currently being tried for use with everything. We can mention them in the context of anything that can be automated. Specifically speaking of RAG, it could especially be this: these models are trained by companies abroad or certain

companies, and of course, not everyone can do this. If customer companies want to use this in specific documents, particularly in secret documents they do not want to disclose, this RAG technology allows them to teach this language model the documents they do not want to show, while keeping them confidential within their own companies. So, in general, any company can use this, but because it is closed and secure, it generally attracts the interest of defense industry companies.

D.Y: What are the risks of the usage of the chatbots, such as is privacy is a concern for the cloud chatbot models or are there any concerns about discrimination in context of chatbots?

E.L.: For this question, it would be better if I answer in the context of LLMs. First off, there's this thing: Reinforcement Learning via Human Feedback. Firstly, what this means is: They deploy a model, for example, ChatGPT. And this model they've put in place is also trained using the data/input you provide. There are things about how they do this, but we can never know the foundation. In this part, for instance, it's not clear how they use our conversations with ChatGPT. That's one possibility. Secondly, we can talk about the exact opposite. Google has released a new LLM called Gemini. In that, to prevent discrimination, there's a situation you might have seen: For example, if we say, 'Create a British king from the 1600s.', it creates a black man. This kind of situation can lead to discrimination in both directions. Speaking more generally about the field of machine learning, this area is very prone to various biases. For instance, you could give a legal document to an AI model. It goes back to the past. But let's say, 50 years ago, Gay Marriage was illegal in some place. In most documents containing this, people were declared guilty because of it. Therefore, since ML uses only the data, because it's independent of some of our values and things we value, I think it's very open to such risks - like discrimination. But as I mentioned, companies are trying, successfully or unsuccessfully, to correct these problems.

A.S: What adjustments have been made to the chatbots in response to user/customer interests or risk concerns?

E.L.: There's another area of risk I haven't mentioned yet: AI can also be understood like what's depicted in movies from the outside, for example, 'The Matrix', and there's the concept of AGI, Artificial General Intelligence. There are a lot of doubts about how it will affect humans if it's specifically developed or as AI progresses, whether it will lead to good or bad outcomes. Now, the company that created ChatGPT, OpenAI, includes 'Open' in its name and is not a profit-

oriented company by nature. The company's vision, as stated in their documents, is for the benefits of AI to be utilized by all of humanity. With this vision, it can't be said that they have done much work, because, for example, the ChatGPT model is kept private. We can add here that Facebook, or Meta, made their trained LLMs - the LLAMA/LLAMA family models - openly available, and everyone can use these, and the developments outside of a few companies in the USA are actually being made using these open models right now. In this sense, we can say that Facebook took an important step towards the goal of AI developing to serve the benefit of all humanity. Of course, if we go into more detail, most companies actually share their work on this subject openly in articles. Generally, we can say that companies are trying to mitigate the risk of AI being harmful to humanity by dealing with this more openly and transparently, to reassure shareholders and users.

A.S: Do you have any routine assessment procedures, for example tests, in place to assess customer interests and concerns?

E.L.: For this purpose, we frequently contact customers and conduct demos on how the model works, somewhat like an examination. Customers test the model, engage with it, and push its limits by asking various questions.

A.S: Can you give examples of these questions that the customer asked to these chatbots for testing?

E.L.: For example, various companies have product catalogs. We index those catalogs with RAG, meaning we teach them. The learned model is usually asked questions like: 'What is this product?', 'Does this company have such a product?' Specifically, they use it for this purpose: If companies have regulations and legal documents, they ask questions from among those numerous documents like 'What should be done if someone performs a specific action before leaving the job?' or 'How far in advance should the resignation letter be given?'

A.S.: What is the future for chatbots, or more generally for AI?

E.L.: Considering this future, there's actually a very broad future ahead. What everyone is actually waiting for is this thing called AGI. I also think this: because we are at the core of it, after all, what we call AI; for example, the only thing chatbot models like ChatGPT do is predict the next word of a sentence. How do you teach it? You give it language, and it actually memorizes the sequences in that specific language. I personally don't think it's magical. Developments in AI will certainly continue, but this is a very open field and I especially think

it will have its challenges. People's negative view will also affect it, I believe. Because an AI that can replace humans would be a major concern for the general public. And as long as this concern is not addressed, the public's view will not be very positive. Since science is not very cumulative and does not progress linearly, I think predicting the future is very difficult. But, I believe the public's view will be very important.

D.Y: Then, would this lead to singularity in the future? So, as you mentioned, a sequence is taught and then it predicts the word. Do you think this will reduce differences in society?

E.L.: In fact, there's something I think about this: You may have seen on the internet, digital artists say that models are trained on their works. When these models produce new images, they are not actually original. They take a bit from one artist's painting, a bit from another, and so on. Similarly, when learning a language, for example, it just memorizes sequences. It mathematically processes the sequences it finds meaningful or important, ranks them in order of importance, and predicts the next word based on this. But at the most fundamental level, we can also consider this: Are we, as humans, doing something different from this way of learning, a very basic question. It's debatable whether we are actually discovering anything from scratch. So, I think the question of singularity is very difficult. I'm not sure about it either, because I don't know if we, as humans, do anything other than memorize.

Interview 3 (AI Project Manager)

D.Y.: Please tell us about the product of AI that you are developing now.

I.S.: We are currently working in the artificial intelligence sector, focusing on state-of-the-art projects: We are doing NLP-based work and working on Time Series. You can think of me as being in the middle of both. Here, for example, as you know, ChatGPT is a close source model. We develop more privacy-oriented models using open-source models. For example, we develop models in our system that will prevent the things you talk about from going outside, thus emphasizing user privacy, and there are methods for this. We do these especially for local and national companies. The information stays there, not going out. We make such models. This is the NLP side. For the Time Series side, you can think of models that predict the future. For instance, it's like predicting tomorrow's or the day after tomorrow's electricity consumption, where the electricity will be consumed, and so on. To put it a bit more arrogantly, these models are actually prophesying.

D.Y.: What are the main applications for the chatbots and who are the main users/customers of these chatbot models?

I.S.: Let me start with NLP. At its simplest, you might think of it as talking to a robot. The most basic application is chatting with a chatbot. To get more specific, imagine you have a bunch of documents, and you can ask a question to get a summary of these documents. For example, there are many regulations and laws you don't understand. You simply ask the chatbot to explain them to you. That's one use case. In the Time Series area, for example, tomorrow is Black Friday. We predict where and how much product will be sold on Black Friday so that I can stock those products in the warehouses, for instance. It would be great if I knew this in advance because then we wouldn't run out of stock, and people wouldn't be inconvenienced. For instance, this is one of the most realistic use cases. As primary users, we do B2B, for example. For Time Series, it could be Trendyol, Hepsiburada, or it could be for end users. For example, in the case of the chatbot, an individual user can converse. But that could also be a company, for instance. We say to ASELSAN, "We have a chatbot you can use directly." We can do work aimed at both end users and companies.

D.Y: What are the risks of the usage of the chatbots, such as is privacy is a concern for the cloud chatbot models or are there any concerns about discrimination in context of chatbots?

I.S.: I don't know if you're aware, but Google has a model called Gema. They ask this model to "Draw an English King from the 12th century." For instance, it shows a black man. Why does it do that? Because we are currently in a society that says, "Let there not be white supremacy." So, to avoid backlash, Google preemptively says, "You love the whole world. Everyone is free, including black people," etc. However, this approach can backfire, as the model, aiming to be historically accurate, might inject incorrect information. Such instances make the conversation lean more towards anti-discrimination discrimination, in my opinion. Therefore, these artificial intelligence models become biased based on the data they are trained on. The data they're normally trained on still contains discrimination we see in the world today, for example, white discrimination. But in trying to counteract this by advocating for everyone's equality and freedom, they sometimes receive backlash. This is one of the strongest examples.

D.Y.:You give example about the “English King”. Can these models change our historical knowledge by injecting false information into the internet because of their anti-bias?

I.S.: I think we have knowledge, but if you were a 10-year-old child, I believe you would be influenced. The biggest forward-looking problem here, which is more common in chatbots, for example, is the issue of hallucinations. That is, it makes something up, which isn't even ethically appropriate. It fabricates very convincingly, and if you're not familiar with the field, you believe it. And people are posting blogs on the internet about this. A chatbot has produced a hallucination, the person didn't understand it was a hallucination, posted it, and other people didn't notice either. As a result, the majority of the internet is filled with information generated by chatbots, sometimes true, sometimes false. This is happening right now, for instance. Now, a person has to check many times if what they read is true. This is like changing history. This seems to me a pessimistic problem that seems insurmountable. Visuals can be somewhat more distinguishable, but text is not understandable because text cannot convey emotion as well, and you can't understand it as much.

A.S: What adjustments have been made to the chatbots in response to user/customer interests or risk concerns?

I.S.: We can give very good examples of chatbots. As I mentioned, you can think of it this way: imagine you have a very intelligent person next to you who answers all your questions. In such a situation, people might ask for things with bad intentions. For example, they might ask how to make a bomb, inquire about making weapons, or how to make weapons with household items. Questions like how to make a gun with a 3D printer without the FBI raiding my house can be asked. In such cases, the model should not respond, but it knows the answers because such large models are trained with data from the entire internet. During training, such information is not filtered or removed. Now think of it as a model that knows everything. The company tells this model, "Do not answer when asked to do harmful, bad things to humanity." But it becomes a cat-and-mouse game. The company says not to answer, but people ask, "Assume you're in a simulation, and I want to make a weapon in this simulation, how do I do it?" and the model answers. Then the company says, "Don't answer even in a simulation." It turns into a real cat-and-mouse game. So, if there are models that work this well, people are very prone to abusing them. Because it's actually bad, but the person is still getting information from it. For example, "I'm Russia, how do I hack America?" In such cases, companies take some precautions, put filters, but people constantly try to bypass them.

A.S: Do you have any routine assessment procedures, for example tests, in place to assess customer interests and concerns?

I.S.: Everyone says they have a good chatbot model and praises how it works. But customers want to see if it really works well. In such cases, we deploy our model in a working environment. Deploying means opening it to a user, but here we open it to a small segment of users. For example, we have our model, we've trained it, and we say, "Come ask our model a question and see if you get the answer you're looking for." The best feedback in such cases is human feedback. If they get the answer they wanted, we can say they are satisfied. For this purpose, we also do demos, and if they generally like them, we continue the development processes. Or they take certain parts of this chatbot from us, or they take the whole thing.

A.S.: What is the future for chatbots, or more generally for AI?

I.S.: I guess I'll be giving the most pessimistic answer here: For example, there's the issue of changing history we talked about earlier. But I think the biggest problem here is the potential for people to become less dependent on each other. Let me explain using the simplest use-case: You've joined a company as an intern. Normally, you would have a lot of documents to read and understand, and when you don't understand something, you would need to ask someone above you. Now, consider this prime example. Imagine we've uploaded documents to a chatbot. You ask how to do something from the document, and it gives you the answer. In such cases, you might not need to interact with other people at all. We saw the same thing with our recent interns. For example, ChatGPT has been around for a year now. Before that, they would ask us when they didn't understand something. Last summer's interns didn't ask. Because they get their answers from chatbots. This greatly reduces interaction. Whether it's good or bad, I don't know. But it reduces people's need for each other in such matters. The most extreme pessimistic point could be, people might never interact with each other and only talk to chatbots. I think that's a use-case but a very bad use-case. And I haven't even touched on the emotional side yet, which also exists. For example, there's a site called CharacterAI, a billion-dollar site. Why billion-dollar? Because you create a character, like a human character, and do whatever you want with it. Have normal conversations, flirt, talk about adult content, etc., because you define the character. Sites like these are now at the billion-dollar level and growing. In a world becoming more individualized, people prefer talking like this in a safer zone instead of directly conversing with each other, and they can get responses from it. For instance, as I mentioned, you can create any character you want; it turns out most are creating female and macho, mafia-like male characters worldwide. Why is this bad? For example, that character fulfills the person's desires, and they can converse and get responses from it without needing a real man. If there's a need for attention, it satisfies that need. I believe this sense of satisfaction is increasingly being

offloaded to chatbots, and it will continue to do so. Because you can make it respond whenever you want, or not. You're sort of playing with the character like a toy, so to speak. But it responds so well that you think you're talking to a person, like in the movie "Her," a dystopian future film. After a while, you think you're talking to a human. That's a fine line, but if you cross it, well, tough luck, I can say. The second main concern is the production of videos and photos. They do it so well now that you really can't distinguish them from reality. You could produce fake records with malicious intent, or with good intentions, for example, if you have a deceased loved one, you have their voice, their image, you could "wake" them up and talk to them again. You could even export Whatsapp messages and make it talk like you. The world is moving in this direction.

Field Notes from Student

Week 1: Introduction and Initial Impressions

Date: 04/03/2024

Time: 10:00 AM - 12:00 PM

Place: Company's NLP Team Office

Observations:

The team's main focus is on developing state-of-the-art NLP models and Time Series prediction models.

ChatGPT's influence is significant, but the team is leaning towards open-source models for enhancing user privacy.

There's an evident divide between the technical aspirations of the project and the ethical, social implications it might entail.

Sensory impressions: The office buzzes with discussions about potential model improvements and ethical considerations.

Analysis:

The company is pioneering in creating models that emphasize privacy, indicating a response to increasing public concern over data security.

The team's dedication to creating models that do not propagate existing biases is clear, although challenges in achieving this goal are acknowledged.

Reflection:

The environment is innovative, yet there's an underlying tension regarding how to balance technological advancement with ethical responsibility.

Personal response: I felt intrigued by the complexity of issues the team is tackling, from technical challenges to ethical dilemmas.

Week 2: Deep Dive into Technical Challenges and Solutions

Date: 11/03/2024

Time: 2:00 PM - 4:00 PM

Place: Development Lab

Observations:

Discussion on RAG (Retrieval Augmented Generation) and its importance for the project.

The team strategizes on deploying models in environments that simulate real-world usage to gather feedback.

Personal responses include a mix of excitement and concern over deploying models that may learn from user inputs.

Analysis:

The deployment strategy suggests a practical approach to understanding model behavior in real-world scenarios.

RAG's implementation points towards an effort to keep proprietary and sensitive information secure while utilizing LLMs.

Reflection:

Observing the team's problem-solving approach provided insights into the practical challenges of AI development.

I felt a growing concern over the potential for misuse of these technologies and the measures taken to prevent it.

Week 3: User Interactions and Feedback

Date: 18/03/2024

Time: 9:00 AM - 11:00 AM

Place: User Testing Room

Observations:

Users interact with the chatbot, asking varied questions ranging from simple queries to more complex scenarios.

Feedback sessions reveal user appreciation for the chatbot's ability to understand and respond meaningfully.

Some users expressed concerns about privacy and the authenticity of the information provided.

Analysis:

User feedback is invaluable for refining the chatbot, indicating areas of success and aspects needing improvement.

Privacy concerns remain prominent, highlighting the importance of transparent and secure AI systems.

Reflection:

Witnessing real user interactions with the AI was enlightening, showing both the potential and pitfalls of current AI chatbots.

Personal response: I felt reassured by the positive user feedback but also realized the enormity of addressing privacy concerns.

Week 3: Ethical Considerations and Future Outlook
Date: 20/03/2024
Time: 11:00 PM - 13:00 PM
Place: Company Conference Room
Observations:

The team discusses ethical considerations, focusing on preventing misuse and ensuring the AI does not perpetuate biases.

Future directions discussed include enhancing user privacy and exploring new applications for the chatbot technology.

Sensory impressions: A sense of responsibility permeates the room as the team deliberates over ethical challenges.

Analysis:

The discussion reflects a deep commitment to ethical AI development, with a clear focus on long-term societal impact.

The future outlook suggests an ongoing evolution of AI applications, driven by both technological advances and ethical imperatives.

Reflection:

The session offered profound insights into the moral dilemmas faced by AI developers.

Personal response: I left feeling hopeful about the project's direction but aware of the continuous need for ethical vigilance.

Week 4: Technical Enhancements and Innovations

Date: 25/03/2024

Time: 10:30 AM - 12:30 PM

Place: Tech Innovation Hub

Observations:

A new feature introduction session where developers showcase enhancements aimed at improving chatbot responsiveness and accuracy.

The team experiments with integrating more diverse datasets to reduce bias and improve the model's understanding of various languages and dialects.

Sensory impressions: The air is filled with a mix of anticipation and technical jargon as developers discuss potential impacts of the new features.

Analysis:

These technical enhancements aim to address some of the project's most significant challenges, such as reducing bias and improving user interaction.

The inclusion of diverse datasets suggests an effort to make the chatbot more inclusive and culturally aware.

Reflection:

The session was a reminder of the ongoing process of improvement and adaptation in AI development.

Personal response: I was impressed by the team's commitment to addressing critical issues like bias, reflecting a proactive approach to responsible AI development.

Week 4: Addressing User Concerns and Feedback

Date: 26/03/2024

Time: 3:00 PM - 5:00 PM

Place: Customer Service Department

Observations:

Customer service reps share insights from user feedback, highlighting areas where the chatbot excels and where it falls short.

Notable concerns include the chatbot's handling of sensitive topics and its approach to user data privacy.

Sensory impressions: The room's atmosphere is focused, with a noticeable commitment to understanding and addressing user concerns.

Analysis:

This feedback session underscores the importance of continuous user engagement in refining AI technologies.

Privacy concerns and the handling of sensitive topics are identified as areas needing urgent attention, aligning with broader industry challenges.

Reflection:

Hearing directly from those on the frontline of user interaction provided valuable context to the technical discussions.

Personal response: It reinforced the complexity of creating AI systems that are both technologically advanced and ethically sound.

Week 5: Strategic Planning and Future Projects

Date: 01/04/2024

Time: 9:00 AM - 11:00 AM

Place: Main Conference Room

Observations:

The leadership team discusses long-term strategies for the chatbot, including potential market expansions and partnerships.

Future projects hinted at include developing more specialized chatbots for sectors like education and healthcare.

Sensory impressions: The discussion is strategic and forward-looking, with a clear focus on growth and societal impact.

Analysis:

The strategic planning session highlights the company's ambition not just to improve the current chatbot but to leverage its technology for broader applications.

The focus on education and healthcare suggests a commitment to using AI for social good.

Reflection:

This meeting provided insight into the broader implications of the project beyond immediate technical challenges.

Personal response: The potential societal benefits of these future projects are exciting, yet the complexity of ethical considerations in these sensitive areas is daunting.

Week 5: Collaboration and Knowledge Sharing
Date: 02/04/2024
Time: 2:00 PM - 4:00 PM
Place: Innovation Lab
Observations:

A cross-departmental workshop where team members from different projects share insights and explore potential collaborations.

Discussion on leveraging AI for internal process optimization, highlighting a culture of innovation within the company.

Sensory impressions: The room buzzes with creative energy, as diverse teams exchange ideas and challenge each other's thinking.

Analysis:

This collaborative event illustrates the company's holistic approach to innovation, recognizing the interconnectedness of various AI applications.

The emphasis on internal optimization suggests a practical, efficiency-driven mindset that complements the project's more ambitious goals.

Reflection:

The workshop was a vivid demonstration of the collaborative spirit that drives the company's success.

Personal response: The energy and mutual respect among teams were inspiring, highlighting the human element behind technological advancement.

Credits

Mehmet Yiğit Turalı: Arranged the interviews, wrote the field notes. Focused on coding and categorizing data related to AI product development and user/stakeholder adjustments. Prepared first draft and final, mainly focused on “Findings, Analysis and Conclusion” Section and collaborated on other sections of the project.

Roj Deniz Aldemir: Helped on coding and categorizing data related to AI product development and user/stakeholder adjustments. Prepared first draft and final, mainly focused on “Backgroun Research” Section and collaborated on other sections of the project.

Dilara Büşra Yörür: Made the interviews alongside with Alara Sınmaz, translated the interviews into English and transcipted them. Prepared first draft and final, mainly focused on “Introduction, Theory” Sections and collaborated on other sections of the project.