Christina Piri

DATA PRIVACY IN THE AGE OF LLM-BASED SERVICES IN EDUCATION: CURRENT CHALLENGES, IMPROVEMENT GUIDELINES AND FUTURE DIRECTIONS

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2024

TIIVISTELMÄ

Piri Christina
Tietosuoja LLM-pohjaisten palvelujen käytössä koulutuksessa: nykyiset haasteet, kehitys ohjeet ja tulevaisuuden suuntaviivat
Jyväskylä: Jyväskylän yliopisto, 2024, 72 s.
Tietojärjestelmätiede, pro gradu -tutkielma
Ohjaajat: Seppänen, Ville ja Kumar, Abhishek

Tämän pro gradu -tutkielman tutkimustavoitteena oli selvittää, millaisia tietosuojariskejä ja niihin liittyviä kehitysmahdollisuuksia nähdään nyt ja tulevaisuudessa, kun generatiivista tekoälyä hyödynnetään kasvavissa määrin koulutussektorilla Suomessa. Aiempien tutkimustulosten ja tämän tutkimuksen haastatteluaineiston perusteella herää huoli siitä, kuinka paljon eri arkaluonteista henkilötietoa laajoihin kielimalleihin (LLM) pohjautuvat sovellukset ja palvelut keräävät ja mihin tarkoituksiin näitä tietoja lopulta käytetään. Lisäksi on epäselvää, missä määrin nykyinen lainsäädäntö pystyy vastaamaan henkilötietojen keräämiseen ja käsittelyyn liittyviin haasteisiin generatiivisen tekoälyn kontekstissa. Tämä tutkielma pyrkii vastaamaan seuraavaan tutkimuskysymykseen: mitkä ovat ne ohjeistukset ja käytännöt käyttäjien yksityisyyden ja tietosuojan parantamiseksi, kun laajoihin kielimalleihin pohjautuvien sovellusten ja palveluiden käyttö koulutussektorilla yleistyy tulevaisuudessa? Empiirinen tutkimusaineisto kerättiin puolistrukturoiduilla haastatteluilla, hyödyntäen tutkimusmenetelmänä laadullista sisällönanalyysiä. Aiempien tutkimusten ja niiden tulosten pohjalta tunnistettiin teemoja, jotka tukivat haastattelujen tuloksia. Näiden lisäksi haastatteluaineistosta nousi esiin uusia teemoja. Yhteenvetona voidaan todeta, että huolenaiheet käyttäjien riittävästä yksityisyyden suojasta generatiivisen tekoälyn kontekstissa on realistinen. Ratkaisuna tähän, tämän tutkimuksen tulokset tarjoavat käytäntöjä ja ohjeita henkilöiden yksityisyyden ja tietosuojan parantamiseen koulutussektorilla. Opiskelijoiden ja opetushenkilökunnan jatkuva koulutus sekä päivitettyjen ohjeiden ja käytäntöjen jalkauttaminen osaltaan edistävät tekoälyn vastuullista käyttöä. Tekoälyn kehittäjäorganisaatioiden tulisi vastata käyttäjien henkilötietojen suojaamisesta koko kehitysprosessin ajan alkaen siitä, että palvelun suunnittelu ja kehitys toteutetaan tietosuojalainsäädännön mukaisesti. Tekoälyn laajentuessa ja kehittyessä sen vaikutukset henkilötietosuojaan ovat jatkossakin merkittäviä, joten tietosuojasääntelyn kehitys voi olla olennaista, jotta voidaan vastata tekoälyn tuomiin tietosuoja haasteisiin.

Avainsanat: tietosuoja, henkilötietojen suoja, tekoäly, generatiivinen tekoäly koulutuksessa, suuret kielimallit (LLM), ChatGPT, Microsoft Copilot

ABSTRACT

Piri Christina
Data Privacy in the age of LLM-based services in Education: Current Challenges, Improvement Guidelines and Future Directions
Jyväskylä: University of Jyväskylä, 2024, 72 pp.
Information Systems, Master's Thesis
Supervisor(s): Seppänen, Ville and Kumar, Abhishek

The research objective of this Master's Thesis is to clarify what kind of privacy and data protection challenges and development practices for improving them are seen now and in the future while generative AI is utilized in the education sector in Finland. Based on the earlier research and studies alongside this study's interview data, a growing concern exists about how much sensitive personal information LLM-based applications and services collect and for what purposes these data are eventually used. It also remains to be seen to what extent the current legislation can address the issues concerning collecting and processing personal data in the context of rapidly developing AI technology. This thesis aims to answer the research question: What guidelines and practices exist for enhancing individuals' privacy and data protection as using LLM-based applications becomes more common in the educational sector? Alongside the results from earlier research literature, the empirical research data was collected through semi-structured interviews utilizing qualitative content analysis as a research theory in this study. Based on the results of earlier studies, several themes were recognized that supported the results of the interviews. In addition, new themes were brought up from the interview data. Concerns related to sufficient data protection in the context of generative AI are realistic. The results of this study offer practices and guidelines to improve individuals' privacy and data protection in the educational sector. It is necessary to highlight the importance of continuous education for students and educators and implement practices and guidelines to enhance the responsible use of generative AI. AI developer organizations may focus on safeguarding users' personal data throughout service development, starting from designing and developing their services to comply with data protection legislation. Since generative AI will keep developing, its impacts on data privacy and protection will also be significant in the future. Therefore, the development of data protection regulation may be essential to tackle the privacy challenges AI poses.

Keywords: data privacy, data protection, artificial intelligence (AI), generative AI in education, large language models (LLMs), ChatGPT, Microsoft Copilot

FIGURES

TABLES

NOTATIONS AND ABBREVIATIONS

AAL	Anticipatory Action Learning
AI	Artificial Intelligence
EU AI Act	European Regulation on Artificial Intelligence
DP	Differential Privacy
GPT	Generative Pretrained Transformer
GDPR	General Data Protection Regulation
LM	Language Model
LLM	Large Language Model
NLP	Natural Language Processing
PII	Personally Identifiable Information

TABLE OF CONTENTS

1 INTRODUCTION

The influence of generative AI on the education sector has been seen as significant lately; thus, it has existed for decades as a technology itself. The use of various LLM-based services and applications has grown to the extent that they support students and educators in performing various tasks, like assisting in personalized learning and teaching, content creators for educational material, and enhancing interaction and group work among students, to begin with (Kasneci et al., 2023). However, the widespread use of generative AI also presents challenges, particularly regarding its users' privacy and data protection. The complex nature of language models, the possibility that they are trained on sensitive information related to individuals, and the uncertainty about how these models collect and store personal data shared by users during interactions for unspecified purposes all pose significant privacy concerns. Concerns also arise from potential data leakage and unauthorized access to users' data, and whether the data privacy rights the General Data Protection Regulation (hereafter GDPR) regulates are complied with in the context of LLMs (Winograd, 2023). This raises concerns about the extent to which the GDPR can address the data privacy challenges generative AI poses currently and in the future.

The information language models produced may also be biased and misleading, creating ethical issues (Meyer et al., 2023). Generative AI is developing rapidly, and various AI-based systems and services are continuously entering the market. To safeguard individuals' privacy and data protection and ensure the responsible use of generative AI, educational and development organizations, legislators, and policymakers are recommended to take the required actions to keep up with AI development. It is crucial to first delve into the current risks and challenges associated with using generative AI, particularly LLM-based systems, in the education sector in Finland. Students and educators at all educational levels increasingly utilize LLM-based systems and services for learning and teaching activities. Based on the current understanding and prediction, the current publicly available research does not directly address this study's research problem or specifically focus on examining privacy challenges and development practices for improving individuals' privacy in the context of generative AI, and particularly in the education sector. Therefore, there is a need for further research on this particular subject. Based on the found risks and challenges, this study aims to identify and propose procedures to improve individuals' data privacy and protection in the context of LLMs and, therefore, to answer the following research question: "*What are the guidelines and practices to improve users' privacy and data protection as the use of LLM-based applications becomes increasingly prevalent in the education sector in the future?*" In order to answer the research question, research data was first collected through semi-structured interviews in addition to previous research. Interviews were conducted with suitable experts working in the research or educational sector in Finland.

The following literature review section (2) introduces the key concepts related to this thesis. It also presents the current EU legislations regulating personal data processing and AI and the use scenarios of LLM-based applications in the educational sector. This chapter also examines the data privacy issues related to LLMs and presents recommendations for developing privacy-embedded language models. Finally, this chapter presents examples of a couple of LLM-based services used in the educational sector and the personal data collection processes within those services. The research methodology (3) chapter describes the research methodology utilized in this thesis in more detail. Findings (4) chapter presents the comprehensive data privacy risks and challenges and the data privacy improvement practices in using generative AI in education based on the interview results. Results (5) section will present the results of this thesis, answer the research question, and highlight the key findings of the interview data, reflecting these findings on the outcomes found in previous studies. Conclusion and Discussion (6) chapter summarizes the research results, discusses their implications, and presents the limitations of this study and topics for further research.

2 LITERATURE REVIEW

The following paragraphs present the relevant research on the particular subject area and introduce the key concepts related to this research.

2.1 Key Concepts

The following chapters introduce and describe the concepts of privacy in the context of AI, different privacy perspectives, Generative AI, and LLMs. This chapter also gives a brief overview, as an example, of one of the free-to-use LLM-based chatbot applications, ChatGPT, and a licensed-based AI service, Microsoft Copilot 365, both of which are known to be utilized in the education sector in Finland.

2.1.1 Privacy and AI

Data privacy is strongly related to personal data collection and access, and different data holds different statuses. Privacy signifies that individuals control the conditions under which their personal information is collected and processed. Individuals' privacy is continually being collected and processed for diverse purposes, with and without their consent (Westin, 1967). Not all data collectors, such as AI developer organizations behind language model development, can collect and process personal information without users' explicit consent. In such cases, all personal data collected and processed by these AI developers should be fully controlled and governed by legislation regulating such activities. The GDPR (European Commission, 2016) regulates individuals' rights over personal information and its collection and processing.

Trask et al. (2024) highlight several privacy-related concerns concerning LLMs. Since there exists a giant modeling strength and intensity within these models, their weights may transform into code-sensitive data included in the training corpus. Such language models can particularly remember individuals' personally identifiable information (hereafter PII). For instance, PII can relate to sensitive information like names, phone numbers, and addresses. In addition, later on, such sensitive information can leak out by accident or via an attack where an outside attacker uses malicious methods to attain possession of confidential information of users from these language models (Trask et al., 2024). The implementation of the GDPR has had a positive impact on individuals, increasing their awareness of the data collected and processed about them. They now understand the responsibilities associated with the personal data they provide to organizations for collection. This empowerment is crucial in the face of increasing data collection and usage, which are creating concerns about individuals' privacy in the context of generative AI (Aslam et al., 2022).

As mentioned above, LLMs are trained on massive datasets that may contain personal data. According to Winograd (2023), it remains to be seen, especially regarding these publicly available LLM-based services, whether they comply with the GDPR regulations. The public data language models are trained with can include personal information about individuals, like information about one's public posts from different social media platforms, for instance. However, a data controller must obtain a lawful basis to collect and process that information according to the GDPR. In addition, whether there are no exceptions, individuals must be informed by the data controller about the data that has been collected (Winograd, 2023). However, the massive amount of training data containing sensitive information and the uncertainty of collecting and processing procedures of such data in LLMs create issues. This might underscore the need to revise the current legislation to address data privacy issues that generative AI brings now and in the future.

2.1.2 Different Expectations Related to Privacy

Individuals can have different attitudes and perspectives related to their data privacy. According to the study by Rao and Pfeffer (2020), privacy expectations are a multi-level construct. This means that individuals can own various privacy expectations. The conceptual model of individuals' privacy expectations includes four types: desired, predicted, deserved, and minimum. According to Rao and Pfeffer (2020), the privacy field has previously mainly concentrated on the desired type. The desired type of privacy describes the situation people ideally wish to have related to one's privacy and might expect to happen in the future. An individual's interpretation of privacy policies about different services online and web pages can require understanding and knowledge about data protection and privacy rights beforehand. Moreover, someone who knows how IP addresses operate might expect a different interpretation of how one's location data is collected compared to an individual for whom this topic is less familiar (Rao & Pfeffer, 2020).

The desired type of privacy is linked to individuals' feelings and beliefs about what they should expect based on their investments, such as money or time. As Rao and Pfeffer (2020) point out, when individuals feel entitled to be rewarded, they think their investments are valuable. This can significantly influence their privacy expectations. Conversely, when the investment is perceived as less valuable, individuals might not expect a bonus or a reward. Mutually, they might think that they deserve a punishment instead. For instance, this can be seen when a free web service user might feel 'punished' by receiving unwanted ads (Rao & Pfeffer, 2020).

The minimum type of privacy defines what individuals would tolerate if a certain condition is met. This 'something' is crucial to fulfill a need, and there are no other options. For instance, individuals may not generally agree to have their health data collected on a career website. However, if it's a prerequisite for a specific job application, they might tolerate it (Rao & Pfeffer, 2020). It is also highlighted that education level and household income impact individuals' tolerance for data collection. Those with a lower level of education may have a more accepting attitude towards data collection than their more educated individuals.

2.1.3 Generative AI

With artificial intelligence (hereinafter AI), technical systems and machines can perceive and process their surroundings and solve problems to achieve specific goals and objectives requested from them. The term AI itself, however, can be described in multiple ways. Emeritus Stanford Professor McCarthy (2007, p. 2) states the following about AI: "*It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable*". The European Commission (2018, p. 1) presents the definition of artificial intelligence as follows: "*AI refers to systems that display intelligent behaviour by analysing their environment and taking action – with some degree of autonomy – to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, speech and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications).*"

According to Warwick (2013), the field of AI began with the birth of computers around the 1940s-1950s. At that time, the focus was on getting computers to do things considered intelligent if a human performed similar activities. Substantially, this involved trying the computers to copy humans in some or all parts of human behavior. In more recent years, the field has genuinely taken its stage. For instance, the current applications and services of AI in critical infrastructure operate in procedures with which the human brain cannot simply compete. Generative AI is constantly learning and adapting with human help. (Warwick, 2013).

AI, as Ertel (2017) describes, is the ability of digital computers or computer-controlled robots to solve problems that are typically associated with human cognitive abilities. This means that AI enables machines to perform tasks that require human-like intelligence, such as understanding natural language, recognizing patterns, and making decisions. Hadzovic et al. (2023) further define AI as a broad term for methods that artificially generate intelligence, allowing machines to mimic human behavior. AI empowers the creation of machines, systems, and services that demonstrate human-like intelligence.

2.1.4 Large Language Models (LLMs)

LLMs utilize deep learning techniques, collecting and processing an enormous amount of text format data from the Internet. LLMs are based on a transformer architecture, such as a generative pre-trained transformer, which enables the processing of sequential data like inputs in text format (IBM, n.d.-a). LLMs contain several neural network layers, which include parameters that are fine-tuned during language models' training phases. Transformers utilize a series of transformer blocks composed of a self-attention layer. This layer supports the model, considering nearby words as input when the model processes a specific word (Weidinger et al., 2021). During training, the model learns to presume the following word within a single sentence based on the prior words (IBM, n.d.-a). Training models include building words and sentences to relate to natural language. LLMs utilize statistical methods to learn to grasp the precise part of every word within the words around it and, in the end, to form a paragraph or a complete sentence (Brown et al., 2022).

Big technology corporations like Google, Microsoft, and OpenAI have all integrated LLMs to enrich the functionality of their current commercial products, applications, and services. According to Kasneci et al. (2023), The GPT was the first LLM released in 2018 by OpenAI. OpenAI later developed GPT-22, GPT-3, and GPT-4 models with more capacities than their initial model, GPT. All these models can produce human-like text, answer users' questions, and assist in tasks that demand translation, problem-solving, and writing capabilities. Another model released in 2018 was BERT (Bidirectional Encoder Representations from Transformers) by Google Research. BERT is based on a transformer architecture and trained on massive amounts of text data and next-sentence prediction, supporting it to understand broader context of terminology across mixed topics (Kasneci et al., 2023).

In 2019 and 2020, Google AI released XLNet and T5 (Text-to-Text Transfer Transformer). The same year, Facebook AI released an LLM RoBERTa (Robustly Optimised BERT Pre-training). Today, the GPT-4 model is the most used LLM. GPT-3 and 4 models operate a transformer architecture, processing sequential data effectively and developing text more consistently and with more and more detailed contextualization (Kasneci et al., 2023). However, language models seem to be complicated "black-box" systems, and the functionality of their internal mechanisms is difficult to understand entirely. The complexity of these models as a whole makes it problematic for humans to interpret how they genuinely operate. The unclarity of their performance might lead to security and data privacy issues and unethical content creation (Zhao et al., 2024).

2.1.5 ChatGPT

One of the most known, freely accessible LLM-based chatbots today might be the ChatGPT. Just two months after its release, ChatGPT reached 100 million monthly active users, securing its status as the fastest-growing LLM-based application in history (Winograd, 2023). ChatGPT is an AI-based model developed by OpenAI (2022). A chatbot can be described as a computer program that engages human communication with its end user. ChatGPT is based on the Generative Pretrained Transformer (GPT) model and utilizes Natural Language Processing (hereafter NLP) to gain knowledge of users' prompts and automatically generate answers to them (IBM, n.d.-b). NLP can be described as a study area exploring how computers can be utilized to learn and produce natural language text or speech to complete practical tasks (Ishii, 2019).

The training process of LLM-based chatbots, like ChatGPT, can be divided into two main phases: **pre-training** and **fine-tuning**. During pre-training, the model is trained with an enormous amount of publicly accessible text data from the Internet. This training data can contain any data available on the Internet, including sensitive information about individuals. The fine-tuning phase involves humans. During this phase, the language model is trained with a more detailed dataset (Glorin, 2023). As mentioned above, LLMs are trained on diverse and extensive language datasets and can generate human-like text that can be contextually coherent. ChatGPT was mainly developed to communicate with users and react to chats in human-like language (Glorin, 2023). Alongside participating in open conversations with its users, ChatGPT, for instance, can perform computer programming, solve mathematical problems, and produce different kinds of text according to the user prompt (Winograd, 2023). Thus, it has been seen as beneficial in educational usage. However, the extensive data that the model learns also poses a potential risk. There is a possibility that it may generate sensitive or personally identifiable information about its users. Some of this training data already includes sensitive data of individuals that is publicly available on the Internet. Moreover, when interacting with LLM-based chatbots, like ChatGPT, users could unintentionally disclose personal information about themselves during conversations, leading to potential data privacy issues.

2.1.6 Copilot for Microsoft 365

In March 2023, Microsoft introduced Copilot for Microsoft 365. It is an addition to Microsoft 365 Suite to support users of Microsoft 365 with automation features for its applications like Office tools and Teams (Spataro, 2023). Copilot is built based on OpenAI's LLMs and hosted in Microsoft's Azure data centers. It may access user data via different documents, emails, calendars, and other information through Microsoft Graph. This integration connects existing applications within the Microsoft 365 Suite, providing automatic assistance with Powerpoint, Outlook, and Teams (Brown et al., 2024a). However it seems that Microsoft collects and stores the data generated during the interaction between the user and Copilot. The information contains the input or prompt provided by the user and the response from Copilot, as well as all references to any information that Copilot has used as a base for its responses. Copilot is a license-based service, and the collected data is thus processed and stored based on contractual obligations and obeying legal regulations in the EU with the organization's other Microsoft 365 contents. With a commercial Copilot service, Microsoft provides its users with a commercial data protection service that promises not to collect and store user interaction to train the LLMs (Brown et al., 2024b). Microsoft Copilot currently appears to be in use at the academic level in

Finland and is available to students (over 18 years old) and staff (Helsinki University IT Helpdesk, n.d.)

2.2 Legislations Regulating AI

The following chapters present the GDPR, which regulates the collection and processing of personal data in the EU. This chapter also introduces the new EU AI Act, the first comprehensive regulation of AI.

2.2.1 The Impact of the GDPR on Generative AI

In 2016, the General Data Protection Regulation (European Commission, 2016) came into force, regulating the collection and processing of personal data of individuals within the European Union. The GDPR states that all individuals should have the right to control their data and its use. Contrarily, while the GDPR focuses on data protection and individual privacy, its prior goal is to protect the fundamental rights and freedoms that are the foundation of democracy. Thus, the GDPR is also an essential regulation in the context of AI as well.

Personal data relates to any piece of information about an identifiable person. Article 4 of the GDPR (European Commission, 2016) defines personal data as "*any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.*" However, personal data can be de-identified through anonymization, encryption, and pseudonymization techniques. The GDPR lays down data protection principles and regulates that data protection by design and data protection by default must be implemented from the beginning in association with personal data collection and processing (Hadzovic et al., 2023). According to the privacy by design principle, organizations implement comprehensive measures at the initial phase of processing procedures to safeguard privacy and data protection principles. Privacy by design requires that companies limit the method of gathering and utilizing data to only the most necessary information. By the privacy by default principle, organizations ensure that personal data is processed with the most increased privacy protection. By default, personal data is not accessible to an unlimited number of individuals (European Commission, 2016).

In education, where AI-based services are a part of students' and educators' everyday lives, it is crucial to have complete control over the personal data these tools and services continually collect and process for unknown purposes. Thus, legislation, like the GDPR, responds to several critical issues related to data privacy in the age of AI; a severe need exists for legislation amendments to tackle AI poses from data privacy and protection perspectives in the educational context. The challenges and issues related to the utilization of AI in the education field are global. Up-to-date provisions and regulations must exist at cross-border levels to guide and control the development of AI systems in diverse parts of society, such as education. The quicker regulations are implemented, the sooner students, educators, and other educational staff can prevent the risks and challenges posed by generative AI from threatening their data privacy rights (Berendt et al., 2020). Since it is generally known that LLMs are trained with vast amounts of data, there is a risk that it may inadvertently collect and process sensitive or personally identifiable information about its users for further purposes. In addition, when interacting with LLM-based chatbots, like ChatGPT, users could even accidentally reveal personal information about themselves alongside the conversations with the chatbot, creating more privacy issues.

According to Berendt et al. (2020), in the existing digital environment where the usage of AI-based services and tools is rapidly advancing in the educational sector as well, it is essential to ensure that educators and students are the prior beneficiaries. It is necessary to define the institute or organization responsible for safeguarding individuals' privacy in this context and understand how the usage of generative AI in education can impact students' and educators' data privacy rights. While the GDPR has made progress in addressing some of these challenges, it is clear that more acts and restrictions are urgently needed to fully address the issues associated with LLM-based applications and services in education. The GDPR focuses on the challenges and issues that emerged from the Internet in 2016. The GDPR holds significant indications for data protection in AI applications, although it does not explicitly mention or cover AI in the regulation. Whether the GDPR regulations are sufficient to address the current and future challenges AI poses in the name of data privacy and protection remains to be seen.

According to the European Parliamentary Research Service (2020), several GDPR provisions are essential in the context of AI, like purpose limitation, data minimization, information requirements, and provisions on preventative measures. However, specific provisions face challenges regarding the unknown methods of handling personal data facilitated by language models. Revisions to the current legislation and up-to-date guidance related to the responsible collection and processing of individuals' data for data controllers and processors need to be implemented. This would also reduce legal uncertainty costs, leading organizations to effective and data protection-compliant resolutions (European Parliamentary Research Service, 2020).

2.2.2 The EU AI Act

The European Artificial Intelligence Act (the EU AI Act), effective August 1, 2024 (European Commission, 2024) is a notable development in regulating AI-based systems. It establishes requirements to address the quick advancement and application of AI. The Act's key provisions, as outlined by the European Commission (2021), seek to establish harmonized rules for developing, market placement, and using AI systems regarding their features and associated risks. Today, it is apparent that the expansive integration of AI brings up concerns and questions about individuals' data privacy. The EU AI Act seeks to sustain the importance of adhering to regulations and ethical practices in the context of AI, ensuring that the development of AI-based services and systems remains responsible (Musch et al., 2023). In the EU AI Act (2024), privacy and data governance involves developing and using AI-based systems and services in harmony with privacy and data protection regulations, ensuring that the data processed adheres to high standards of quality and integrity. Transparency entails designing and utilizing AI systems to ensure adequate traceability and explainability. It includes making users knowledgeable in situations while interacting with an AI-based system and services, informing deployers about the system's capacities and restrictions, and notifying the affected individuals of their data privacy rights (European Commission, 2024).

According to Musch et al. (2023), the EU AI Act aims to protect users of AI-based applications and systems by guaranteeing the right to access their information, restricting automatic decision-making by AI-based systems, and laying down more transparency obligations for AI developer organizations. According to the European Parliament (2024), AI developer organizations are obligated to adhere to the EU AI Act's transparency requirements and EU Copyright law. This includes the responsibility of AI developers to ensure that the model does not generate unlawful content and to inform the usage of copyrighted data for language model training purposes. Such requirements of the EU AI Act aim to provide more accountability and transparency towards users of AI-based services (European Parliament, 2024).

When a user engages with an AI-based system or service or in a situation where the system or service itself recognizes human characteristics, according to the EU AI Act (European Commission, 2024), the service provider is obliged to inform the user that the communication is happening with an AI-based service. Additionally, whenever an AI-based system creates images, videos, or sounds, the service provider must report to the user that the content is artificially generated. The EU AI Act lays down specific demands for high-risk AI-based systems. High-risk AI-based systems can significantly impact individuals' health and safety, fundamental rights, or the functioning of society. High-risk systems may be utilized in critical infrastructure, for instance. The regulations specify that such high-risk systems' operations must be sufficient, transparent, and appropriately documented to meet the conditions for appropriate use (European Commission, 2024). TABLE 1 below presents the Articles of the EU AI Act with their content obligating transparency and accountability within AI-based systems and services.

TABLE 1 The EU AI Act’s transparency obligations for AI system deployers

Name of the Article	**Content of The Article**
Article 13 *Transparency and provision of information to deployers*	The design and development of high-risk AI-based systems and services must be implemented so that their operational procedures enable users to utilize the systems appropriately and responsibly and understand and interpret the outcomes of these systems. In addition, AI-based systems and services' operations should be transparent. Manuals must be attached to the respective systems in a suitable format, including reliable and transparent content, and present users with relevant and understandable information.
Article 50 *Transparency obligations for providers and deployers of certain AI systems*	Providers of AI-based systems must inform the individuals interacting with AI-based unless it is evident to the user. When creating artificial audio, video, image, or text material, the provider of AI-based systems is obliged to ensure that the outcomes of the AI system are marked in a format that is machine-readable and noticeable as artificially developed or manipulated. Deployers of AI systems that create artificially produced or manipulated images, sounds, or video content must notify individuals that the content has been created artificially. Deployers of an emotion recognition system or a biometric categorization system are obliged to notify the individuals exposed to the system's operation. Personal data must be processed under the Regulations (EU) 2016/679 and (EU) 2018/1725 and Directive (EU) 2016/680.

The EU AI Act highlights the significance of transparency, accountability, and notification obligations within AI-based systems and services. The GDPR grants individuals the right to obtain information in automated decision-making. In addition, individuals have the right to be informed about those decisions. Stakeholders, policymakers, and regulators should prioritize these principles and confirm that AI developer organizations would focus significantly on users' data privacy and protection within their AI-based services. Up-to-date regulations and guidelines must be implemented alongside constant governance related to responsible AI development (Much et al., 2023).

2.3 The Use of LLMs in Education Today

At their best, utilization of LLM-based services and applications in education can bring out improvements in learning and teaching, which can be possible at all levels of education. For example, with the help of LLM-based applications, students can learn different languages and improve diverse writing styles. LLM-based applications can also support learning and teaching particular subjects, like mathematics, physics, and literature. These systems can help educators create exercises and quizzes, which students can use to enrich their understanding and retain the information and material they have previously learned (Kasneci et al., 2023).

According to Kasneci et al. (2023), LLMs can enhance university-level students' learning process. Some LLM-based tools can assist students in developing problem-solving skills and critical thinking and provide support in research and writing work. With the aid of LLMs, students can quickly and easily extract summaries and key points from lengthy texts, thereby enhancing their understanding of the main points and facilitating subsequent writing tasks. LLMs also offer information on specific research topics, enabling students to analyze research material more efficiently. From an educator's perspective, utilizing LLM-based tools may enhance productivity in lesson planning, personal teaching, and personal concept creation (Kasneci et al., 2023).

LLMs represent a significant advancement in the era of generative AI. Language models have made substantial advancements in NLP in recent years. In addition to answering users' questions with human-like responses, LLM-based chatbots can write essays and scientific reports and accomplish other language-related exercises (Kasneci et al., 2023). Chatbots, like ChatGPT, have been found to benefit education and assist in related assignments. LLM-based chatbots can support educators and students with personalized learning and support in writing-related tasks and improve creative thinking (Trust et al., 2023). LLM-based tools might benefit writing by enhancing its quality, particularly in the academic level and for non-native English speakers. LLM-based chatbots are a source of fast and direct answers to particular questions and work as content designers for producing, formatting, and outlining requested text for its users. Especially at higher education levels, LLM-based chatbots are seen to grow students' engagement, assist in group work, and give direct feedback and evaluation to their users, for example (Meyer et al., 2023).

LLM-based tools can assist with programming with developers of any skill level since, for instance, ChatGPT can automatically write new code. Undoubtedly, concerns have arisen about dishonesty and misinformation about such tools and services among students and educators. According to Neumann et al. (2023), utilizing LLM-based services and applications in the educational sector might lead to cheating in homework or exams. In addition, LLMs may produce text without appropriate references and citations, so the outcome may be considered plagiarism. According to Baidoo-Anu and Owusu Ansah (2023), data produced by LLM-based chatbots may also contain incorrect or biased information. Students and educators must also be aware of these issues so that they will not have unethical and harmful impacts on their work.

Innovative educational technology, like generative AI, holds the potential to significantly enhance learning analytics, helping students and educators achieve and manage their education goals more effectively. However, such systems require a substantial amount of training data, which can include personal data. This poses privacy and security challenges and concerns. Despite these challenges, the potential of generative AI in education has been remarkable, and the benefits have been noted (Kshirsagar et al., 2022). It can be generally stated that generative AI is also here to stay in the educational context as well.

2.4 Data Privacy Issues and Challenges in the context of LLMs

While organizations today collect and process customer and user data on a massive scale, at the same time people are becoming increasingly aware of their data privacy rights. This consciousness may lead to concerns about privacy and the potential issues and challenges with data gathering and processing. According to Martin et al. (2017), collecting and handling personal data for multiple purposes can lead to potentially problematic situations impacting customers' sense of vulnerability.

AI has taken significant steps forward in recent years. According to Zhai et al. (2021), this might be because data processing is cheap, and a large amount of information is publicly available for AI developer organizations to utilize. In an educational context, the exposure, sharing, and improper use of students' personal information pose a genuine risk and a challenge in using AI when accessing and sharing large amounts of data. LLMs are primarily designed to understand, generate, and interact in a contextually relevant way. To achieve this, these models are trained on extensive datasets, some of which may contain personal information. The data-centric nature of these models is a fundamental characteristic. However, such a data-focused nature significantly increases privacy concerns, as there is a genuine risk of misuse and the possibility for exposing individuals' sensitive information to these models (Glorin, 2023).

Trust is a crucial element in the use of generative AI. Therefore, the AI developer organizations might need to focus on prioritizing transparency and trust within development processes, particularly chatbots where user interaction is prevalent. Users should feel secure and protected when using LLM-based tools and that their personal information is being handled with reliable measures. The implementation of robust data security and protection measures is crucial to maintaining data integrity and reducing the risks of malicious attacks, a potential threat in the context of generative AI (Glorin, 2023). Based on earlier studies, the following chapters describe the risks and challenges related to personal data collection and processing in the context of generative AI and present the existing data privacy improvement methods.

2.4.1 Privacy Risks and Concerns in Collection and Processing Personal Data

Safeguarding individuals' data protection and privacy is essential to the ethical, responsible, and legally compliant development of AI. The material language models trained may contain public text-formed data, such as academic articles and publications of professionals, Wikipedia texts, and users' posts from various social media platforms (Winograd, 2023). The complex and unclear nature of language models processing this data raises concerns about data protection and what this collected data is utilized for. In addition, data processing based on conversations with LLM-based chatbots might quickly become questionable whether users reveal sensitive information about themselves to these applications. The paragraphs below present risks and concerns based on previous research literature on users' privacy and data protection in the context of LLMs in more detail.

According to Weidinger et al. (2021), privacy violations can arise due to the possibility of **data leakage.** Models can remember personal data that might be included in the training data. A situation where the model would reveal this personal information is called data leakage, which could lead to privacy violations. It is worrying that this information can be collected about a specific individual as part of the training data without the affected person having anything to do with the issue and, therefore, missing the essential part of individuals' data privacy rights, informed consent, that the GDPR (European Commission, 2016) regulates strictly. In addition, for instance, there can also be situations where other people post or share sensitive information online about a particular individual and data leaks occur. Data leaks are unavoidable since LLMs keep continuously growing. Therefore, memorization will also increase, leading to a situation where these models can contain more sensitive text about individuals (Weidinger et al., 2021).

LLMs cannot entirely understand the context or sensitivity of the data they are trained with. Depending on the context and the content of such data, the processing and possible data leakage may cause serious harm related to individuals' privacy (Brown et al., 2022). However, by utilizing the System 2 approach and specific prompting techniques, LLMs are being developed to become increasingly capable of contextual understanding and reasoning tasks (Yu et al., 2024). LLMs' vulnerability to data leakage poses a concern, as does the commercial implementation of models nowadays. LLMs include over a hundred billion parameters tuned in the initial training stage on giant publicly available text data sets. For instance, an LLM built by Google LaMDA was trained with a 1.56 trillion-word dataset. Even though such data is available for all to see and access, it can still include sensitive user-related data (Winograd, 2023).

LLM-based chatbots' conversations with their users might include sensitive information, such as the user's email, home address, and name, as well as information on an individual's health and other confidential data. A situation in which a model is trained with data containing such confidential information and leaks its content would considerably threaten users' privacy (Winograd, 2023). However, LLM-based chatbots, like ChatGPT, may be unable to store and recollect such data; they can store non-PII data for a temperate period to develop the model's performance for 30 days. However, a data leakage is possible when transmitting data, whether the communication channel is unsafe or not protected appropriately (Glorin, 2023). Multiple privacy-preserving techniques exist to prevent data leakage as well. However, these protection techniques alone cannot ensure appropriate and sufficient protection from a data privacy perspective (Brown et al., 2022).

In the education sector, using LLM-based chatbots is becoming increasingly common within different educational tasks. However, it is essential to be aware of their potential risks. As Weidinger et al. (2021) point out, when interacting with a chatbot trained to imitate human-like responses, users may unknowingly or by accident **disclose personal information during conversations** to chatbots about themselves than initially intended, like their thoughts, feelings, and perspectives on specific topics. Unintended sharing of individual-related data can happen, for instance, whether the user believes that the system or tool in question they are communicating with is reliable and safe (Glorin, 2023). In the education sector, students may utilize chatbots to write essays or reports in psychology or health education. They accidentally might share private information about themselves during interaction without thinking about it. Collecting this kind of information about users violates privacy rights.

Several LLMs have been fine-tuned alongside training to prevent harmful behavior. However, these models can be made to create unethical and damaging material. An outsider attacker can perform a "jailbreaking" attack, forcing the model to ignore its instructions and adjusting its behavior for malicious purposes, ignoring the safeguards. (Winograd, 2023). The purpose of such an attack is to deceive LLM-based chatbots by providing clever and tricky instructions and prompts that pass the safety restrictions implemented by the developers. In this case, the chatbot could execute any assignment without considering the initial safety instructions it set to obey in the first place (Das et al. 2024). The results of a successful jailbreaking attack can lead to manipulating users to expose confidential information within conversations for malicious purposes, for instance.

As mentioned earlier, the training data of language models consists of publicly available data and may also include individuals' personal information. It is clear that processing and revealing this kind of confidential data would pose privacy violations, like the **dilemma of informed consent and the erasure of one's data.** LLMs are not just utilized for any specific usage but can be applicable for multiple purposes. Whenever organizations collect and process users' or customers' data, under the GDPR (European Commission, 2016), they must ask for explicit consent from these persons. Due to the rapid growth of the AI-based service and application market, it is becoming increasingly clear that informed consent is only sometimes an emphasis. For example, the data individuals might enter on different websites and social media platforms becomes part of the training data set for LLMs. Once models collect and process this data, it becomes embedded in the training material. The challenges of machine unlearning further complicate matters, potentially making it impossible to withdraw our consent from data processing yet have our data entirely removed from these models (Winograd, 2023). More awareness and transparency are needed regarding these activities in the educational context, where students' data is collected, processed, and stored in the context language models. Students and their families need to be informed about the circumstances under which a guardian of an underage student must provide informed and mandatory consent before a student's data can be collected and processed (Kasneci et al., 2023).

According to the GDPR (European Commission, 2016), individuals have the right to know how data related to them is collected and for what purposes. In addition, individuals have the right to withdraw their consent and the right to the erasure of personal data ("the right to be forgotten"). By these principles, organizations and service providers must request and obtain explicit consent from individuals before collecting and processing their data and have it removed later on without undue delay whether an individual requests so. According to Winograd (2023), for example, in the US exists a notice-and-choice mechanism demanding either an "opt-in" or "opt-out" option in situations when personal data is being collected or processed. However, this method might not be sufficient since it does not efficiently enough inform users about their data collection and processing. In addition, the rapid advancement of technology allows for increasingly more comprehensive and explicit data collection of individuals. People rarely read the long, complex, and difficult-to-understand terms of service and privacy policies. Just clicking the "I agree" button might feel easier but may not sufficiently indicate a person's genuine and informed consent for data collection and processing. In this mind, the procedures and efforts to protect users' privacy within the "opt-in" and "opt-out" options may remain ineffective and superficial formalities from these perspectives (Winograd, 2023).

Obtaining genuine, explicit consent from a person is problematic in the context of language models. According to Winograd (2023), the ability of LLMs to infer information about an individual, even if they have yet to provide it to the model explicitly, highlights the inadequacy of ensuring and protecting privacy in this context. In the context of LLMs, it might be unlikely to adequately implement the "right to be forgotten" principle, leaving the individual helpless whether they would wish to withdraw their consent given in the past and have their data removed from these systems. So, it can be generally stated that while adding "I agree" buttons in lengthy privacy policies and terms of use may be somewhat valid methods based on the interpretation of data protection legislation, they may not be sufficient and proper methods alone to safeguard individuals' privacy, who unwillingly and unknowingly might contribute their personal information as part of training material. According to Winograd (2023), safeguarding individuals' privacy should include more transparent and trustful options for individuals to give informed consent for data collection and processing and have their data removed from these models.

According to the GDPR (European Commission, 2016), a person who has given their informed consent for processing their data has the right to withdraw that consent later at any time. The opt-out process allows an individual to withdraw their consent to process data related to them. Based on the present information, the current legislation does not require AI developer organizations to comply with the conditions for the consent principle (European Commission, 2016) already included in the models. Legislators and policymakers might need to consider whether regular opt-out should be required by default in the development process of language models and direct AI developer organizations to regularly permit users to withdraw their consent to process their data. Therefore, AI developer organizations would need to systematically remove all data related to individuals who have withdrawn their consent for data processing purposes. Therefore, these models would need to be retrained from the beginning without personal data. Such a requirement could be particularly relevant for individuals who provide these models with sensitive and private personal information about themselves (Winograd, 2023).

According to Winograd (2023), there is a need to **clarify the current legislation and implement reliable policies and proper guidelines** for enhancing privacy and data protection in LLMs. The current language model development might only partially guarantee compliance with the GDPR. However, by implementing reliable policies and guidelines, full compliance with the GDPR would enhance privacy and data protection for all individuals. The GDPR (European Commission, 2016) states that the processing of personal data always requires a legal basis, which must be determined beforehand. Whether there are no exceptions, the data controller must inform the particular person(s) about the data collection and processing. Furthermore, it is essential to clarify how the data protection principles provided by the GDPR are adhered to in the context of LLMs.

The current reality is that even the experts and developers behind the development of language models have limited knowledge and understanding of the threats arising from machine learning and the complex behavior of language models' memory. It is unclear how data memorization and data extraction from models are executed, nor how efficiently the current technical privacy-preserving and defining techniques truly perform to protect individuals' privacy. The responsibility for erasing personal data relies on the entity's shoulders, which has the best knowledge about this information, how it is handled, and where it is genuinely located. However, without even sufficient understanding of them, individuals may remain empty-handed with the option of genuinely giving lawful, informed consent for AI developer organizations' data collection and processing purposes (Brown et al., 2022). The need for precise policies and guidelines on these issues is essential, both now and in the future, to address privacy concerns and the generative AI's rapid development.

2.4.2 Perceived Data Privacy Issues in the Educational Context

Data privacy and security concerns arise while utilizing LLMs since the data entered into these models might contain sensitive and personal information related to students. This can evolve into incidents of data breaches, unpermitted access to students' information, and individual-related data utilization for other meanings than education precisely (Kasneci et al., 2023). OpenAI (2024a) allows its users to make privacy requests to have their data erased from the system. However, it seems that OpenAI will not erase any prompts the user has input before making such a request. According to Trust et al. (2023), for example, if a user has asked about a sensitive topic, such as related to their health, OpenAI maintains a permanent record of that user prompt. In an educational context, if an educator interacts with ChatGPT to create any educational document, including a student's information, this potentially violates the affected user's privacy.

As Kasneci et al. (2023) highlight, the key to addressing these concerns is the implementation of robust data privacy and security policies in the responsible use of LLMs. These policies should clearly outline the collection, processing, and storage of student-related data in accordance with ethical standards and legislation, such as the GDPR. Equally important is the need for transparent communication, ensuring that students and their guardians are fully informed about the data collection, processing, storage, and usage, and that their consent is a prerequisite for personal data gathering and processing, especially in the case of minor students (Kasneci et al., 2023).

Kasneci et al. (2023) state that the latest technologies and measures must be implemented to safeguard data privacy from unpermitted and unethical use and data breach incidents. Regular audits associated with data privacy and security methods are required to define and locate potential vulnerabilities and areas needing enhancement. In potential situations when a data breach or unpermitted access to personal data might be potential, an incident response plan should be implemented effectively to address and reduce those issues. It is crucial to underscore the importance of continuous education of educators and students on current regulations, ethical issues, data privacy and security policies, and the recommended procedures to manage and report associated risks (Kasneci et al. 2023). Comprehensive education ensures that everyone in the educational organization is well-informed and actively involved in maintaining data privacy and security.

2.4.3 Recommendations for the Development of Privacy-Embedded Models

AI developer organizations such as Google, OpenAI, and Meta have acknowledged the escalating challenge of safeguarding privacy within LLMs.

The full range of risks and concerns associated with these models is yet to be fully understood. However, in recognition of the crucial importance of user privacy and security, AI developer organizations have initiated an effort **to develop privacy-embedded language models.** The task at hand for AI developers is to create LLM-based services and applications that not only function effectively but also uphold the privacy and trust of their users (Winograd, 2023).

According to the GDPR's Article 17 (European Commission, 2016), an individual has the right to obtain from the data controller the erasure of personal data concerning the affected individual without undue delay. Based on this principle, individuals' privacy should be safeguarded by legislation that considers individuals' choices regarding their data. According to Winograd (2023), obtaining an individual's informed consent is problematic in the context of LLMs. The GDPR's Article 17 (European Commission, 2016) allows individuals to reconsider their decision about the content and information they have shared before and erase this data from the systems and services that have initially collected and processed information about them. However, the deep learning technique utilized in language models makes it difficult to obey this right. Training data that may contain personal information is embedded within these models. It may be that even the professionals who have built such models do not thoroughly understand the nature and data processing of these models (Winograd, 2023). Therefore, finding a piece of personal data related to a person they wish to be removed seems to be highly challenging and more like trying to find a needle from a haystack.

In a scenario where an individual's personal information is removed from a model and not utilized in such training material, there still exists a chance that this data remains in the model. Language models can recognize and identify different patterns. Based on the large amount of collected information, these models could construct predictions before and later on rather than just remembering specific pieces of data (Winograd, 2023). This complex memory structure of language modes, where the future outputs are affected by the previous ones, also creates personal data erasure challenges (Chang, 2024). Though an enormous amount of public data is used to train LLMs, AI developer organizations are recommended to use it only when essential to implement the model, considering the context and data privacy issues (Winograd, 2023). Data that is publicly available is not automatically intended for public use. Publicly available information about a specific person may not have been disclosed by that person in the first place. However, personal information might have been leaked or shared by copying and pasting data from different contexts and conversations that affected individuals have taken part in earlier, for instance. In addition, posts published on social media platforms might have been initially meant to be private. However, these posts may occasionally be unintentionally made public later on (Brown et al., 2022).

AI developer organizations might need to **focus on higher quality and filtered data sets** instead of spontaneous data collection from public sources. Training language models on information that may contain data from social media posts, for instance, creates a risk of collecting individuals' personal information without the affected individuals' explicit consent. Organizations developing language models may consider retraining models periodically on up-to-date data to reduce the information embedded in these models that were previously removed (Winograd, 2023).

It's a significant challenge to ascertain the nature of the data used to train language models. For example, OpenAI disclosed the data and sources used for training GPT-3, but no information is available on the content of training data used for GPT-4 (Winograd, 2023). OpenAI (2024b) states that in training LLMs as sources, they use publicly accessible data on the Internet (containing personal information), data from third-party providers, and user data they provide. This underscores the crucial role of legislators in demanding more transparency in language model training and development processes, particularly in the collection and processing of personal information. AI developer organizations should provide detailed information on the sources of training datasets and take the necessary steps to ensure this data is processed in compliance with legislation while implementing robust data protection and privacy-preserving measures (Winograd, 2023).

2.4.4 Current Techniques for Enhancing Data Privacy in LLMs

According to Brown et al. (2022), **differential privacy (hereinafter DP)** and **data sanitization** are privacy-preserving techniques utilized in language models. In the DP, algorithms are trained with a particular technique to remediate the risks within data memorization. According to Rigaki and Garcia (2023), the idea behind the DP technique is to "not learn anything" about a specific person while achieving information about a population in general. Therefore, personal data is made challenging to obtain for harmful purposes (Brown et al., 2022). DP-based techniques insert a controlled amount of noise into the data used in training language models. Since such methods try to secure privacy by adding additional noise to the data, this can, on the other hand, reduce the usability and accuracy of LLMs (Glorin, 2023).

According to Winograd (2023), in the data sanitization technique, individual-related information is pursued to be removed from the training data of language models with a precise data allocation to prevent data leakage and data memorization. However, the algorithms behind data sanitization techniques might be able to remove sensitive information, like social security numbers and names, for instance. Nevertheless, they cannot identify unusually performed information and privacy issues that might be context-dependent (Winograd, 2023). What people in general might admit as a private issue might vary based on contextual characteristics of a person's opinions, culture, and behavior. When such characteristics are embedded in the training material of language models, they can be challenging to identify and, therefore, also to be removed (Winograd, 2023). The enormous amounts of diverse data language models that are continuously trained underscore the urgency and persistence in addressing privacy issues.

In the **deduplication** process, duplicate data is erased from the training data. Since duplicate data is more likely to be remembered, erasing redundant information slows language models' memorization. While language models are trained with a large amount of training data, it is challenging to catch all redundant data. The deduplication might decrease the incidence related to models' memorization. However, it cannot entirely prevent data privacy issues, like data leakage, from happening (Winograd, 2023).

Data anonymization is a method where the stored data in a language model might connect to a specific person, and this personal data is replaced by one or more pseudonyms or other artificial identifiers. Afterward, data is combined so the dataset does not contain personal data related to this person. Anonymizing and combining data impacts some LLM-based applications and chatbots. For instance, it can affect ChatGTP's performance, so data anonymization negatively impacts the loss of context and details, affecting the quality of the language model's result (Glorin, 2023).

LLMs can predict and generate various human-like outputs. One method to strengthen privacy in these systems is through **machine unlearning**. Machine learning systems derive models and various components from the training data. Derived data includes additional information, which can be based on formerly collected user-related data. While this data undergoes numerous computation phases in multiple systems, at the same time, it leaves traces in various places and occurs in many forms. The initial data, computations, and derived data create a complex data web called data lineage. (Cao & Yang, 2015). The GDPR (European Commission, 2016) regulates the principle of the right to erasure ("right to be forgotten"), defining that the individual has the right to obtain from the controller the erasure of personal data concerning this individual. The data controller must remove personal data without unnecessary delay. This principle grants the right for an individual to have their sensitive data (and its lineage) erased from a system. To completely forget a piece of data in the training dataset, these systems need to revert the data outcomes on the extracted features and models, called machine unlearning. A simple data sample, however, needs to be first identified before it can be "unlearned" and, therefore, removed (Cao & Yang, 2015).

Retrieval Augmented Generation (hereinafter RAG) is a technique utilizing NLP that improves text generation by combining data fetched from an extensive corpus of text records. Therefore, this technique enables the development of precise and contextual outcomes with external information (Zeng et al., 2024). At least in LLMs like Perplexity.ai, Gemini, and ChatGPT, RAG is currently operated. RAG generates retrieved material in the LLM prompt, enhancing the model's accuracy to provide users with more authentic results (Wu et al., 2024). However, data leaks are a notable risk in RAG systems since data from the retrieval database and language models’ pre-training and fine-tuning datasets can leak during or as an outcome of RAG use. This retrieval database may contain sensitive, personal information processed in the context of RA-LLM-based chatbots. Additionally, the RAG retrieval process can affect the behavior of the LLM in text creation, leading the LLM to expose sensitive information from its training and fine-tuning data (Zeng et al., 2024). In addition, an external attacker can potentially manipulate RA-LLMs to generate harmful and unreliable outputs, leading to security risks for their users.

Overall, the responsible development of RA-LLMs is vital to guarantee reliable use. The reliability of RA-LLM systems should include at least the subsequent elements: robustness, fairness, explainability, and privacy. RA-LLMs should act secure and robust enough to tackle the attacks and disruptions caused by an external attacker. RA-LLMs should be secure, sufficiently resilient (robust), and effective against disruptions caused by external attackers. Fairness means that RA-LLM systems should actively resist bias and discrimination and avoid related activities during their decision-making approaches. These systems' responses and predictions should be transparent and explainable. Privacy and data protection in RA-LLM system development safeguard the sensitive and confidential information stored within their databases (Fan et al., 2024).

2.5 Personal Data OpenAI Collects from its Users

This paragraph aims to provide an example of what kind of personal information and to what extent OpenAI, the developer of one of the most known LLM-based chatbots, ChatGPT, collected from their users.

According to OpenAI (2024b), the models operating behind ChatGPT are trained with data from three primary sources: publicly available data on the Internet, data from third-party providers, and information that their users or human trainers of OpenAI provide.By OpenAI (2024b), ChatGPT has been generated to understand and answer user prompts and questions during human-like interactions. ChatGPT continuously processes and learns how different words typically occur within other words in different contexts. ChatGPT can, therefore, predict the next probable word in reply to a user's question and every following word. OpenAI (2024b) states that language model training data can contain individuals' personal information since the training data gathered from the Internet relates to individuals. However, according to OpenAI (2024b), they do not collect personal data for profiling or marketing services.

According to OpenAI (2023a) European Privacy Policy, OpenAI collects the following personal data related to a user when communicating with OpenAI presented in TABLE 2 below.

TABLE 2 Personal data OpenAI collects from users of its services

Personal Data	Collected Data

Account Information	Information related to the user, like person's name, account credentials, contact information, payment card details and transaction history ("Account information").
User Content	By user content is meant by all user related personal data included in the user's input, file transfers or user feedback provided to OpenAI.
Communication Information	When a user communicates with OpenAI, the user's name, contact information and any other contents of the user's messages related to the interaction with the service is collected ("Communication Information").
Social Media Information	OpenAI interacts on social media like Facebook, X, Youtube, LinkedIn, Medium and Instagram. Whenever the user communicates in social media related to OpenAI, OpenAI gathers personal data related to the user that the user has entered to the OpenAI. This personal data includes contact information of the user ("Social Media Information''). The organizations hosting social media sites belonging to OpenAI might distribute aggregate data and analytics about OpenAI's activities in social media platforms.
Other Information the User Provides	OpenAI gathers additional data about the user in situations where the user is involved in surveys or events hosted by OpenAI or the user provides OpenAI information to clarify the identity or the age of the affected user ("Other Information You Provide").

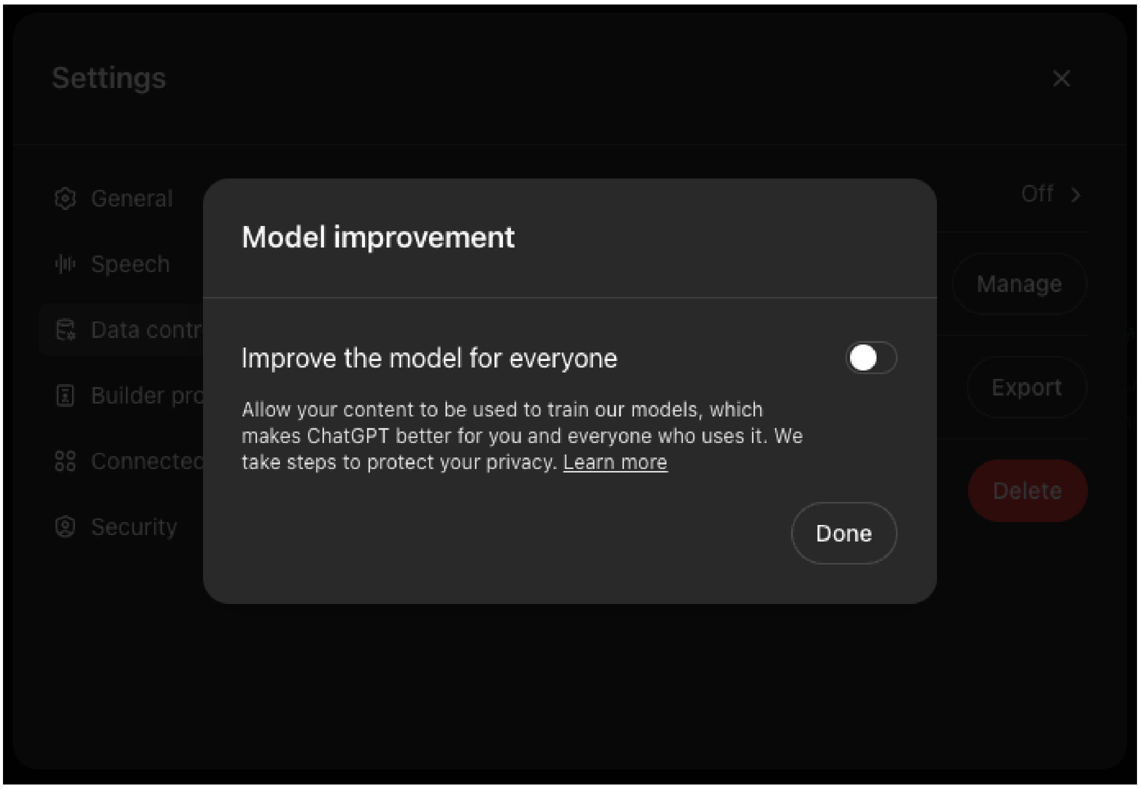
OpenAI (2023a) automatically collects their users' specific personal data whenever they utilize or communicate with OpenAI services presented in TABLE 3 below.

TABLE 3 Technical information OpenAI collects automatically from users of its services

Personal data - Technical Information	**Collected Data**
Log Data	Information of the user's browser or device when using OpenAI's services provided by Information contains the user's Internet Protocol address, type and settings of the

	used browser, the date and time of the user's request and information how the user utilizes OpenAI's services.
Usage Data	OpenAI gathers information about the user's use of the Services meaning the content type the user is interested and involved with, the features and actions user performs while using OpenAI's services as well as the time zone, country and the date and time of user's access. User agent and version, type of computer or mobile of the user and their computer connection.
Device Information	OpenAI collects information, depending on the type of the device and its settings, about the user's device, such as the name, operating system, browser they are using and device identifiers.
Cookies and Similar Technologies	OpenAI collects cookies and related technologies for their own purposes.

In addition to developing their language models with Internet data and user prompts, OpenAI(2023a) also obtains information from their partner entities and marketing vendors, which contribute information on prospective clients of their business. According to OpenAI's European privacy policy (2023a), OpenAI utilizes its users' data for specific purposes, like supplying and managing their services, developing their services, conducting research work, and informing their users about their services and events, for instance. When it comes to personal data collection and processing related to children, OpenAI (2023a) states that they do not knowingly gather personal information of children under 13 years old. In addition, OpenAI (2023a) mentions that their services are not intended for children under 13 years old, and users under 18 must have approval from their parents or guardians to use services provided by OpenAI. OpenAI allows its users to opt out of model training with their content (see FIGURE 1 below). However, OpenAI will not erase any prompts the user inputs before making such a request.



Model improvement

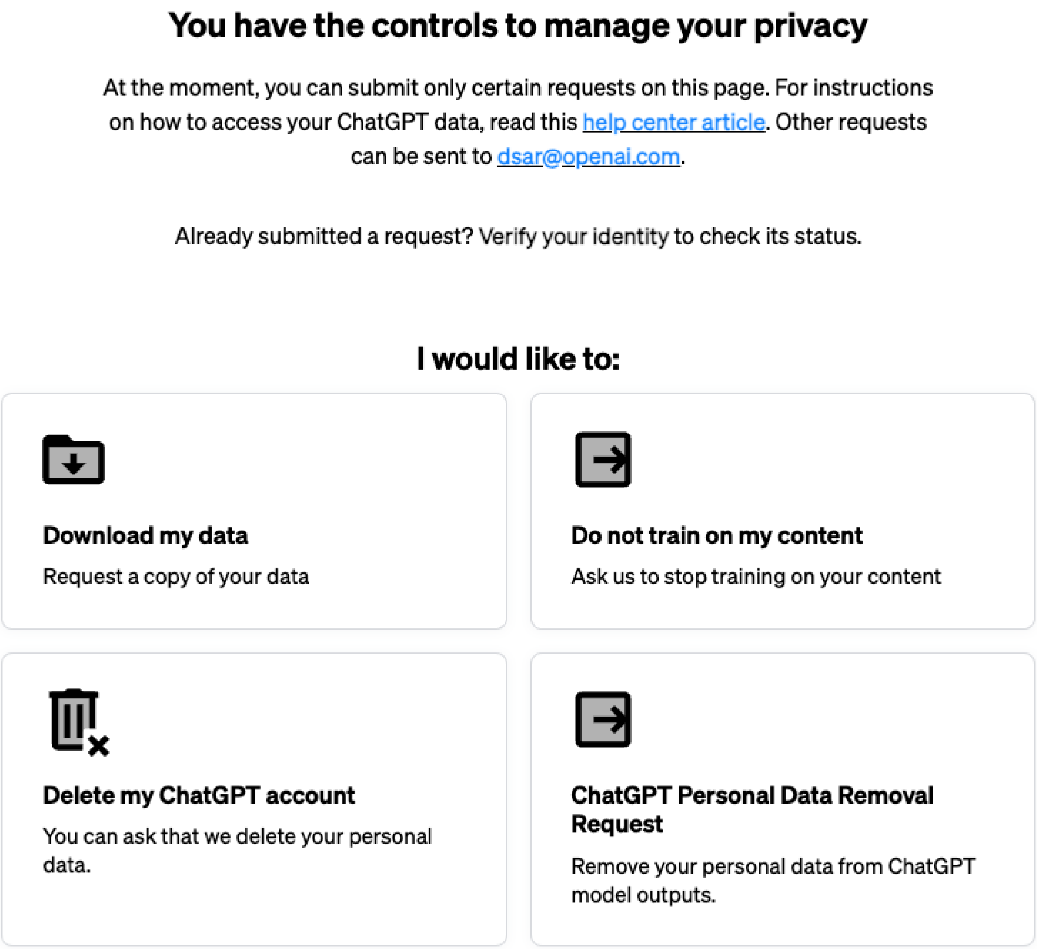
Improve the model for everyone

Allow your content to be used to train our models, which makes ChatGPT better for you and everyone who uses it. We take steps to protect your privacy. Learn more

Done

FIGURE 1 Possibility to refuse sharing content for training in ChatGPT

Users can also make separate privacy requests as illustrated in FIGURE 2 to not train the model on the content provided to ChatGPT and to have personal data erased from the model. In order to have their data removed from ChatGPT, a user must enter more detailed information about themselves to OpenAI before submitting a request, such as their full name, phone number, and email address, as well as specific ChatGPT prompts produced with the personal data and explicit reason for removal purposes.



You have the controls to manage your privacy

At the moment, you can submit only certain requests on this page. For instructions on how to access your ChatGPT data, read this help center article. Other requests can be sent to dsar@openai.com.

Already submitted a request? Verify your identity to check its status.

I would like to:

Download my data

Request a copy of your data

Do not train on my content

Ask us to stop training on your content

Delete my ChatGPT account

You can ask that we delete your personal data.

ChatGPT Personal Data Removal Request

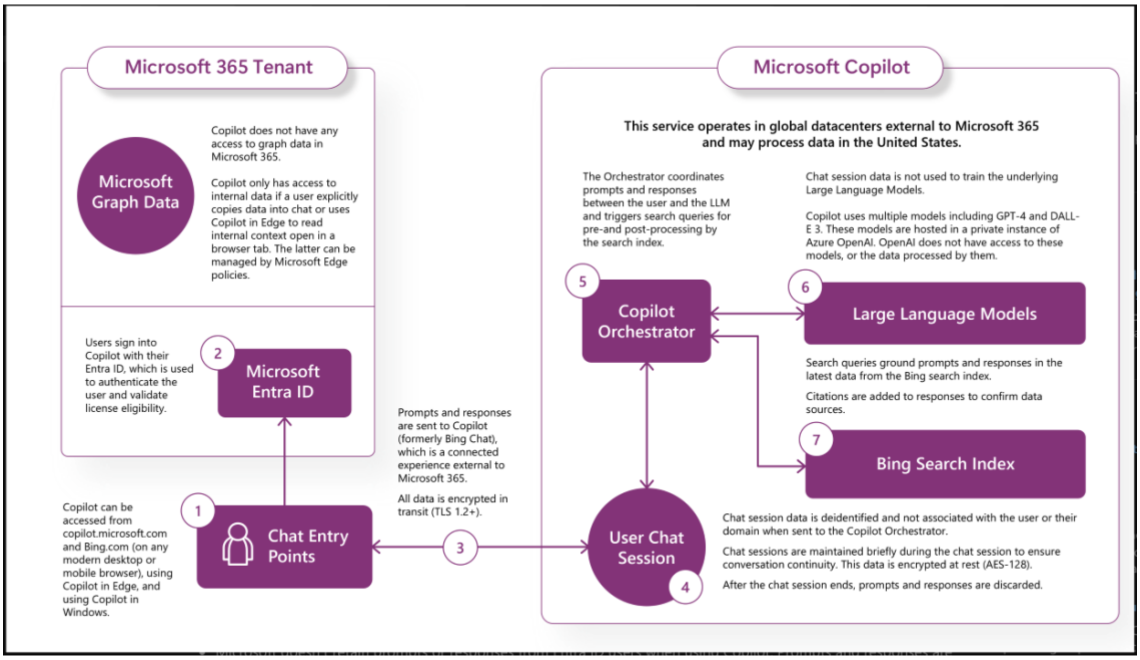
Remove your personal data from ChatGPT model outputs.

FIGURE 2 Privacy request option in ChatGPT

As it transpires, OpenAI gathers a considerable amount of user-related information for language model training purposes. This information contains the user's information and prompts, IP address, log and usage data, and communication data of the platform, to begin with. In addition, OpenAI might distribute this information to other service providers, affiliates, vendors, and law enforcement (OpenAI, 2023a). In order to protect their privacy and data protection rights, a person has to go through quite a lot of extra effort in order to do so.

2.6 Data Collection and Processing in Microsoft Copilot

In December 2023, Microsoft expanded Microsoft Copilot access in education with commercial data protection available to higher education students over 18 and above and to education staff. According to Microsoft Education Team (2023), when a user signs into Copilot with their personal school accounts, this commercial data protection will be enabled as part of the whole service. Commercial data protection (Davis et al., 2024) aims to ensure the protection of user and organizational data. Chat conversations (like user prompts and chat responses) with Copilot are not to be collected, stored, or accessed by Microsoft (see FIGURE 3 below for more details). In addition, chat data is not utilized for further LLM training purposes (Microsoft Education Team, 2023).



Copilot does not have any access to graph data in Microsoft 365.

Copilot only has access to internal data if a user explicitly copies data into chat or uses Copilot in Edge to read internal context open in a browser tab. The latter can be managed by Microsoft Edge policies.

Users sign into Copilot with their Entra ID, which is used to authenticate the user and validate license eligibility.

Copilot can be accessed from copilot.microsoft.com and Bing.com (on any modern desktop or mobile browser), using Copilot in Edge, and using Copilot in Windows.

Prompts and responses are sent to Copilot (formerly Bing Chat), which is a connected experience external to Microsoft 365.

All data is encrypted in transit (TLS 1.2+).

This service operates in global datacenters external to Microsoft 365 and may process data in the United States.

The Orchestrator coordinates prompts and responses between the user and the LLM and triggers search queries for pre-and post-processing by the search index.

Chat session data is not used to train the underlying Large Language Models.

Copilot uses multiple models including GPT-4 and DALL-E 3. These models are hosted in a private instance of Azure OpenAI. OpenAI does not have access to these models, or the data processed by them.

Search queries ground prompts and responses in the latest data from the Bing search index.

Citations are added to responses to confirm data sources.

Chat session data is deidentified and not associated with the user or their domain when sent to the Copilot Orchestrator.

Chat sessions are maintained briefly during the chat session to ensure conversation continuity. This data is encrypted at rest (AES-128).

After the chat session ends, prompts and responses are discarded.

FIGURE 3 Commercial data protection in Microsoft Copilot

Microsoft Copilot is already used at the academic level in Finland by teaching staff and students. For instance, according to the Helsinki and Tampere Universities' announcements (Helsinki University IT Helpdesk, n.d.; Teaching and Learning Centre, 2023), Copilot is recommended for students and educators over free-to-use AI services since data protection and security have been taken into account with Copilot. Microsoft's Copilot, with commercial data protection, is a public cloud service. Helsinki University IT Helpdesk (n.d.) advises that although this service ensures user security by not collecting and storing their conversations for language model training purposes, the model might be located outside the EU during usage. Therefore, it should be utilized to process public data, and personal data must not be entered in Copilot. According to the Microsoft Education Team (2023), Copilot with commercial data protection service is intended for students aged 18 and above. However, the process for obtaining consent from the guardians of minors (under 18) and the monitoring or restriction of Copilot usage by underage students remains unclear. The potential use of Copilot with underage students in Finnish universities is currently under investigation, as reported by the Teaching and Learning Centre (2023).

The purpose of the previous sections has been to familiarize the reader with the terms and key concepts in this study and to form a basis for the thesis area and the research problem presenting data privacy and protection risks and challenges in the context of generative AI and suggested potential remedies for those concerns based on previous research. In addition, the previous sections presented examples of a couple of LLM-based services known to be utilized in the education sector. These chapters also introduced the reader to those LLMs' data-collecting processes. It can be generally stated that LLM-based services will be strongly connected to the education sector today and in the future. Previous studies have shown that utilizing LLMs has several positive aspects and benefits regarding teaching and learning. Naturally, using LLMs also brings up concerns and challenges, which this thesis discusses from the user data privacy and protection point of view, also providing improvement suggestions to safeguard data privacy in this context. The following chapter presents the research methodology utilized in this thesis and describes the empirical data collection method in more detail.

3 RESEARCH METHODOLOGY

Data privacy and protection seem to be severe concerns in the context of LLMs. LLMs are trained with large amounts of publicly accessible data from the Internet, possibly containing sensitive information of individuals. Since the data collection and processing of language models is complex and unknown even for experts behind the development of models, it raises concerns about where all this personal information is stored and for what purposes. User conversations with LLM-based chatbots like ChatGPT can be used for model training purposes by default unless the user turns off this option in the settings, so the conversations will not be utilized for training purposes. However, whether toggling a switch to a different position and creating various privacy requests to prevent the collection and processing of personal data within the service will be sufficient actions remains to be seen. These issues together obviously raise concerns among users. However, the benefits of generative AI have been seen and adopted at all levels of education. LLM-based applications have helped support learning, plan teaching, and assist students with their assignments, for instance. Generative AI will remain a part of everyday life. Therefore, it is crucial to understand what is required in practice in the educational context now and in the future to address the challenges it poses to students' and educators' data privacy.

The following chapter describes in more detail the qualitative research method chosen and utilized in this thesis and presents the grounds for choosing this method. In addition, the chapter describes how the empirical research data was collected during this study.

3.1 Qualitative Content Analysis

This study utilized a qualitative research method to collect empirical research material through semi-structured interviews. Qualitative research aims to define and provide knowledge of relevant phenomena from the research material. It also highlights issues that need further examination (Elo et al., 2022). Generative AI is continuously developing; thus, data privacy concerns related to the use of general AI are topical. According to the current predictions and knowledge, the earlier publicly available studies on data privacy issues and data privacy improvement methods in the context of LLMs in the educational sector are limited. Therefore, qualitative content analysis was applied to this study's research method. According to Elo et al. (2022), a qualitative content analysis is suitable for analyzing various materials, like data collected through interviews. The objective was to collect interview data from experts in research or the education sector who have experience and knowledge about generative AI, data privacy, and data protection. The interview data provided a comprehensive understanding of the current and future risks and concrete improvement possibilities related to data privacy in the context of LLMs in education.

The qualitative content analysis aims to describe the research material in a focused, summarized, and generalized form and build themes or phenomenons emerging from the research results (Elo et al., 2022). Because the empirical research results were collected through semi-structured interviews in this study, qualitative content analysis as a chosen research method supported this data collection method. Utilizing qualitative content analysis methodology provided a deeper understanding of the research results, the specific themes and phenomena that appeared in previous studies findings, and from the interview data. First, identifying and then combining similar themes and naming new ones arising from the empirical research alone helped to build a more precise and logical structure for presenting the results of this study. Because the research topic of this thesis falls within the domain of future research, anticipatory action learning (hereafter AAL) could have been utilized as another option for a research method in this study. According to Inayatullah (2006), the AAL method focuses on current phenomena and future ones. It aims to learn about the research problem to generate additional futures around it. In AAL, the research process cycle circulates between questioning, creating, and questioning, adding an anticipatory extent to the learning procedure. However, AAL as a research method is difficult to implement in practice.

Qualitative content analysis aims to confirm the theoretical concept or framework within research. To focus on the research question, the researcher can find guidance on present research or theory or expand or refine the current theory. Whether the research material is gathered initially via interviews, the researcher may utilize open-ended and targeted questions related to the categories and themes determined at the beginning of the research process (Hsieh & Shannon, 2005). According to Hsieh and Shannon (2005), directed content can be divided analysis into three sections:

1. The study starts with a theory.
2. Codes are defined before and during data analysis.
3. The researcher derives codes from theory or relevant research findings.

After the theory part, creating codewords helps the researcher focus on relevant and critical issues in the research material. Hair and Page (2015) describe coding as assigning relevant numerical values or names to reduce data from a substantial amount of unstructured text. Hsieh and Shannon (2005) write that as a subsequent process, the researcher would code all key points utilizing the codes determined beforehand. The researcher must name the results using a different code, whether they cannot be classified with the original coding scheme. A third step in the analysis would be to code all emphasized sections using the predetermined codes, and any text not categorized with the initial coding scheme would be given a contemporary code. While the analysis goes further, the researcher creates new codes and refines and modifies the original coding scheme (Hsieh & Shannon, 2005).

The following themes constructed from previous studies related to data privacy concerns and possible remedies to improve data privacy and protection in the context of LLMs are presented in FIGURE 4 below. The need for improved legislation and legal compliance is urgent and cannot be overstated. LLM developers might need to start focusing on taking more responsibility for developing privacy-embedded models in accordance with legislators and policymakers. The potential issues of interactions with LLM-based chatbots, where users may inadvertently share personal information with human-like systems, further underscore the need for this accountability. While current privacy-preserving technical methods exist, they are not bulletproof data privacy safeguard mechanisms and, therefore, may be insufficient to protect individuals' privacy.

Previous studies provide several general suggestions for improvements in efficient data protection in the context of LLMs (see FIGURE 4). First, clear guidelines and practices related to responsible usage of AI need to be implemented. LLM developers are recommended to develop models where privacy and data protection are core issues embedded in the initial phase during the development process. Also, training data of LLMs must be revised so that there is a guarantee for an individual to have their data removed from the training dataset. Since no current legislation, alongside the GDPR, comprehensively regulates the collection and processing of personal data, amendments to the current legislation might be needed to answer the existing and future challenges generative AI will bring in the perspective of data privacy and protection.

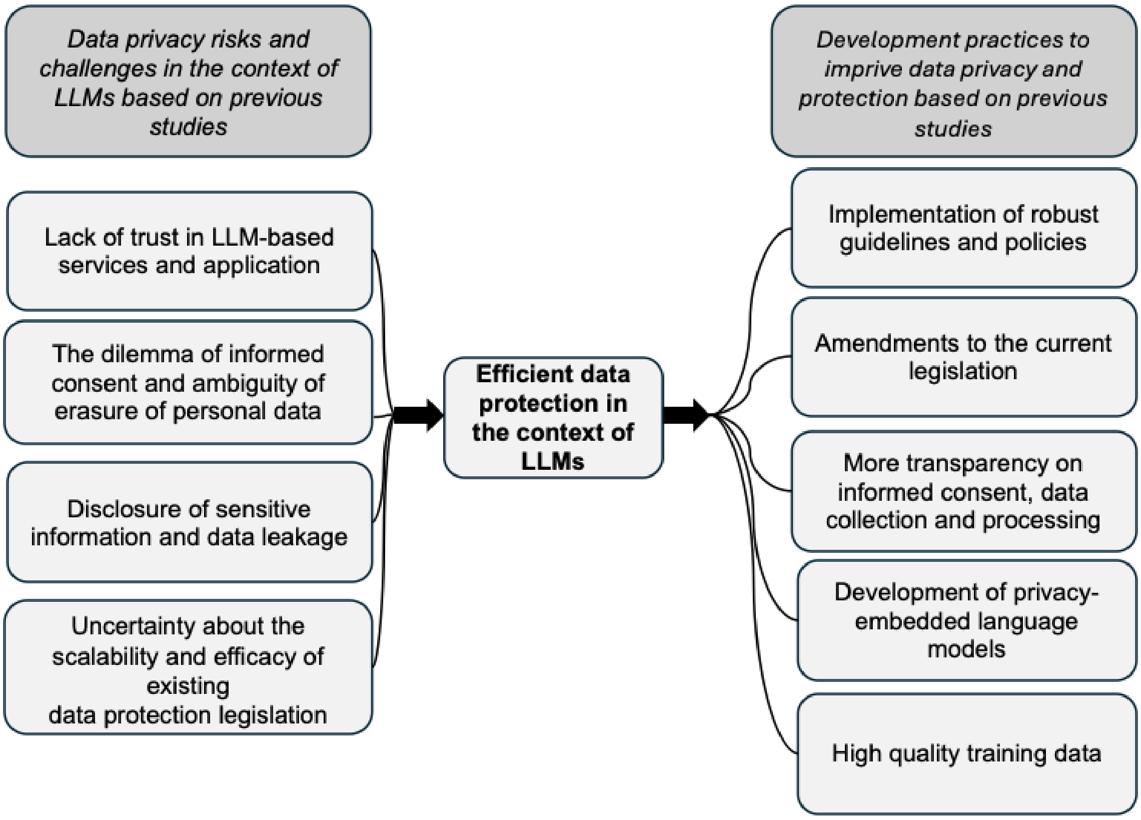

Data privacy risks and challenges in the context of LLMs based on previous studies
Lack of trust in LLM-based services and application
The dilemma of informed consent and ambiguity of erasure of personal data
Disclosure of sensitive information and data leakage
Uncertainty about the scalability and efficacy of existing data protection legislation
Efficient data protection in the context of LLMs
Development practices to imprive data privacy and protection based on previous studies
Implementation of robust guidelines and policies
Amendments to the current legislation
More transparency on informed consent, data collection and processing
Development of privacy-embedded language models
High quality training data

FIGURE 4 Data privacy risks and development practices to improve privacy and data protection based on previous research.

Previous studies have addressed specific privacy and data protection issues and challenges in using LLM-based applications. Previous studies have also suggested improvement directions to promote data privacy and protection in this context at the general level. Because this study focuses on first identifying data privacy issues in the educational sector and, from there, clarifying possible development opportunities to improve data privacy and protection, the interview questions were formed as follows:

1. Do you anticipate any risks and challenges related to personal data privacy and protection in the usage of LLM-based applications in the education sector? If so, what are they?
2. Can you see any opportunities and needs to improve personal data privacy and protection within the use of LLM-based applications? If so, what would be the development opportunities and practices for improvement?
3. Should we limit AI's access to some data? What would it be?

Based on the interview questions, this study aimed to clarify whether the answers support the themes that emerged in previous studies and whether the interview data reveal risks and recommendations for improvement that were not included in the previous studies. The interview results provided concrete guidelines and practices for improving users' privacy and data protection when utilizing LLM-based applications in education.

This thesis conducted in-depth interviews with five relevant experts from Finland working with AI technology either in education or research (see TABLE 4 below). The interviews were conducted through online meetings, lasting around 30-45 minutes. The interviews were recorded and transcribed using ResearchVideo by the University of Jyväskylä. While working on the transcriptions of the interviews, the focus was kept on the details to ensure that all valuable data was carefully considered, but smaller filler words like "umm" and "like" were intentionally omitted from the responses. The interviews and transcriptions were conducted in Finnish, and the interview material was translated into English afterward, reproducing the responses given in the original language as accurately as possible. The direct citations in the data analysis section below the interviewee term are replaced with "respondent" with the specific respondent number below (TABLE 4).

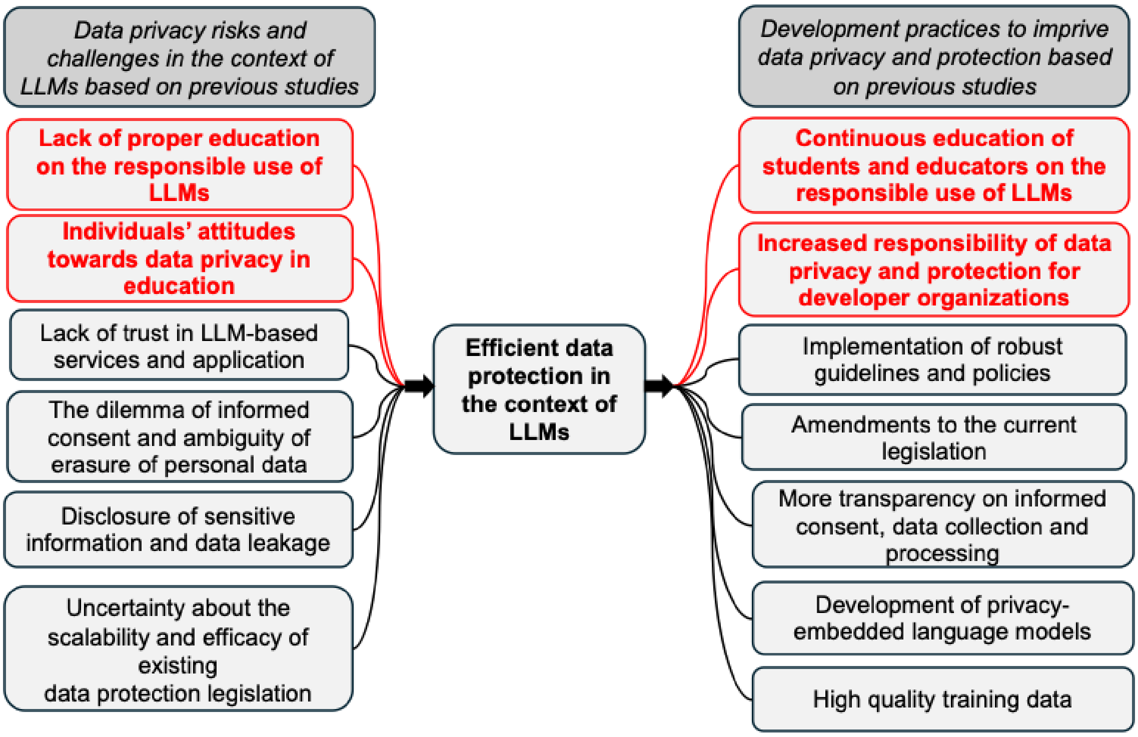
TABLE 4 Interviewees

Respondent	Education	Duration of Career (in years)
1	Doctor of Science in Technology	Nearly ten years in academic research, several years of industry experience
2	Doctor of Science in Economics and Business Administration, Associate Professor	About ten years in upper management on the Academic level, before that 20 years of industry experience
3	Doctor of Science in Economics and Business Administration	Nearly 10 years of experience in academic research
4	Doctor of Science in Humanities	Years as a professor and over 10 years in academic researcher and education work
5	Master of Arts	Several years as advisor and project manager, 20 years in teaching

The following chapter delves into the findings of this thesis, as in, the results of the interview data, including the identified data privacy risks in the context of LLMs and recognized development practices for improving data privacy and protection in education.

4 FINDINGS

This chapter presents the interview results, including the recognized data privacy risks in the context of LLMs and the suggestions for improving data privacy and protection in the educational sector. During the interviews and after, when the interview data was extracted and analyzed in more detail, it was noticeable that similar themes arose from the interview results related to risks and improvements to data privacy in LLMs compared to the results from previous research. FIGURE 5 illustrates and combines similar or identical risks and improvement themes based on previous studies and the interview data. The new themes that came up from interview data and were not considered in the previous research are highlighted in red and emhasized in bold in FIGURE 5.



Data privacy risks and challenges in the context of LLMs based on previous studies
Lack of proper education on the responsible use of LLMs
Individuals' attitudes towards data privacy in education
Lack of trust in LLM-based services and application
The dilemma of informed consent and ambiguity of erasure of personal data
Disclosure of sensitive information and data leakage
Uncertainty about the scalability and efficacy of existing data protection legislation
Efficient data protection in the context of LLMs
Development practices to imprive data privacy and protection based on previous studies
Continuous education of students and educators on the responsible use of LLMs
Increased responsibility of data privacy and protection for developer organizations
Implementation of robust guidelines and policies
Amendments to the current legislation
More transparency on informed consent, data collection and processing
Development of privacy-embedded language models
High quality training data

FIGURE 5 Data privacy risks and challenges in the context of using LLMs and the improvement suggestion in education based on previous research and interview results.

The following chapters present the interview responses to the first two interview questions. The answers cover current data privacy risks and challenges in the educational sector and practices for improving privacy and data protection from the interviewees' opinions. The final chapter responds to the third interview question, presenting the interviewees' perspectives on whether AI access should be limited to some data.

4.1 Identified Data Privacy Risks in Education

Several concerns arose during discussions with the interviewees about challenges and risks related to using LLMs in the education sector. First, all the interviewees raised their concern that language models might also be trained on sensitive, personal information provided by their users and contained in public data collected from the Internet. The answers also revealed that sufficient education and guidance on the responsible use of generative AI are needed and must be updated accordingly. According to the interviewees, the younger the user is, the more carefree the attitude might be concerned with using available LLM-based applications and the personal information they collect about their users. In addition, a lack of trust directed to publicly available services came up within discussions, as well as the unawareness and uncertainty about what kind of data and for what purposes commercial, license-based services collect and use, especially when underage students are in question.

The answers highlighted the importance of the continuous education of students, educators, and education staff. However, educating students and educators cannot be entirely left on their shoulders. AI developer organizations may need to take more responsibility for developing safer AI and prioritizing users' data privacy. Now more than ever, when new products and services are coming to the market at full speed, users need comprehensive education on the responsible use of generative AI, and legislators and policymakers ought to keep up and address the issues brought by the rapidly developing AI.

At the beginning of the interview, the study aimed to determine whether interviewees anticipated any risks and challenges related to data privacy in the context of LLMs in education and what those risks and challenges would be. Based on the answers, the issues were seen in the **need for proper education on the responsible use of generative AI**. It was also seen as a concern that there is a growing number of options available from various LLM-based applications and tools to which students have free access. It is problematic that this increasingly growing amount of choices of LLM-based programs and tools may be tested impulsively by entering all kinds of personal information into them without understanding the consequences from a legal and responsible point of view.

> Respondent 2: “Of course, we now live in an intense hype era, where new products come to the market at a tremendous speed and disappear simultaneously. There are risks, especially when such new products are being tested without thinking about how to operate responsibly, so one can quickly enter information into these products that they shouldn't. There is a challenge then, like in any other open cloud service. Confidential data cannot be entered into such a service, and certainly not highly confidential personal data. For instance, a research project where such data is processed would need to undergo data security and data protection inspections. After that, it would need to be examined from the enterprise architecture’s perspective. The risk is that experiments are done

> enthusiastically without remembering the limits of legislation and reasonable actions.”

One of the interviewees stated that an educational organization is already doing it wrong if it even guides students and educators to use these open-access AI-based applications and tools, which are known to collect all the data revealed to them. That would undoubtedly form a data protection risk. On the other hand, the answers also stressed that education cannot be fully responsible for educating students and educators about the reliable use of these services, especially at the elementary school level. Instead, using commercial, contract-based AI services in the education sector is more essential, where contractual and legal responsibility is emphasized, and personal information for model training purposes is collected and processed under no circumstances.

> Respondent 5: “The first thing here is that it is not entirely related to risk, but in practice, an education organization makes a mistake if it asks or even guides students or their students to use, for instance, ChatGPT or another similar AI-based service with consumer-side credentials. There exists a data protection risk. You need to understand that if you rely on the notifications of chat OpenAI or Copilot, by switching the toggle to this direction, my chat conversation is not used for language model training. But that cannot be left in this education, or at least in basic education and upper secondary school, which is under compulsory education, to the user's responsibility. Then, these AI-based services should only be used if we can get an organizational agreement around them and where they are not by any means used for language model training."

One of the significant issues and concerns related to data privacy the interview answers raise is that LLM-based chatbots, by default, collect all data a single individual might reveal about themselves or another person to these chatbots. In addition to this, it was seen as a concern, whether students and teachers might unknowingly or by accident **disclose personal data** to these systems and applications. It remains unclear where all the user-related data exposed to these models ends up and for what purposes it is ultimately used. Therefore, users of these models may not necessarily understand or even think about all the problems and risks the data mass collected and processed about them might involve and form in the long run.

> Respondent 3: “First, what private personal data do you put in there, so the data protection related to them (is a risk). Secondly, in general, the input of sensitive information. That is, sensitive data of other people is fed to these systems. This second option is brought up often in research. Lately, how LLMs could be utilized when handling qualitative data has been discussed. This, of course, presents a challenge, as it is not always clear where the data goes and how it is used and processed when fed to

> LLM. Indeed, it weakens its usefulness if you want to use it ethically. You need to be careful what data you put in there, like research data and so on, so it can use data analysis for that. In general, there is this risk of what data they consume and where it ends up."

> Respondent 4: "And then there are these threats, sort of, about what kind of negative use it can have if there is a huge amount of data on one individual or group of people out there. There are future risks, even though this data is not currently being used to serve such damaging purposes."

Significantly, this raises concerns when younger students might utilize these open-access LLM-based chatbots in writing reports or essays in which they unknowingly or accidentally expose sensitive information about themselves.

> Respondent 5: When you write an essay in health education or psychology, you can provide quite sensitive information about yourself in it. Then, when there are no organization-provided accounts available, we can not even secure the tool's data protection. So, very likely, a massive amount of information goes continuously for language modeling training, where they place such information in there despite everything."

The GDPR regulates the requirements of informed and freely given consent related to individuals' data collection and processing purposes. An individual has the right to have their data removed from the system (the right to erasure), refuse consent, and withdraw it later on in a straightforward manner as consent was given in the first place. Obtaining explicit, informed consent from a user has been seen as very challenging in the LLM context since they are initially trained with data containing individuals' personal information. Due to the complex data processing methods and their unknown nature of storing data, the **right to data erasure** and the possibility for **a person to give their consent** within the LLMs seem like more apparent options than lawful rights. The parent or guardian's consent is required to collect and process a child's data based on such consent up to a certain age (in the EU, between 13 and 16 years). However, while the GDPR does not directly obligate open-access LLM-based applications, getting user consent is not seen as transparent and reliable. According to the responses, consent was seen as more like an apparent and compelling obligation from the developer organizations' perspective, which does not, at the very least, reinforce trust in the service.

> Respondent 4: "If we think about the GDPR and its implementation, getting one's consent can be very apparent, which is particularly problematic when considering children and younger people. It can be difficult not to give this consent if one wants to use a particular application. Somehow, rather many of those challenges are in such a way

> that they are not seen right away; however, they'll appear alongside collective usage and what it means in the long term."

> Respondent 3: "However, ChatGPT supposedly has the output-option. For instance, your data is not used for training purposes. As a private individual, I could say that I'm not very trusting, without a doubt, of such promises."

Based on their age and education level, individuals have different perspectives and expectations about their privacy, and the toleration of data privacy and protection methods differ between younger and older people. It is humane that younger or older people as well might not be fully aware of their privacy rights. Currently, there exists a growing number of so many variants of different AI-based applications, tools, and services available on the Internet that, understandably, most people do not have the time or motivation to read through the service provider's long privacy policies, making it easier to click on the "I agree" button when user's consent is required to use these systems or service. Younger students' **attitudes toward privacy** might be impulsive, or they may perform oversharing with AI. They might focus more on getting answers and help quickly from applications rather than thinking about safeguarding their privacy.

> Respondent 4: "I've researched younger peoples' data practices, so my perspectives are likely related to that. The risks related to their personal data are pretty weakly recognized or they might easily think that what does it matter not if someone collects some of my data. It has become sort of a self-evident issue that everyone collects data, and that might appear in the youngsters' speech. My experience with discussing with young people is that it has become an issue like, well, there's nothing I can do about it. The ones that try to pursue the protection of their personal information within these systems are individual cases. It's no surprise that those risks are not recognizable for young people, especially in everyday use. And not for adults, either. It's sort of just forced to agree to certain data gathering, and if you want to try to prevent it, it demands special activity and consideration."

> Respondent 5: "Indeed, this is a problematic usage from the student's perspective since they approach this data protection risk very TikTok-like. But for example ChatGPT or Perplexity or so that students use… We want to teach students to understand and be aware of the risks that AI brings up. However, it is very challenging for a group that uses TikTok daily. This is a challenging situation, but an education provider cannot guide us to use the AI-based tools provided on the commercial side."

During the interviews, data privacy concerns were raised from the discussions related to **a lack of trust towards LLM-based services and applications** and the **suspicious purposes personal data** is handled. Systems and services might not be engineered with sufficient safety, trustworthiness, and user data privacy and protection in mind. Safeguarding users' data privacy is questionable whether the developers and experts behind the development do not entirely understand the behavior of the models.

> Respondent 1: "I mean, you need to be aware of any kind of problems that could arise. I don't think for example explainability is a big problem these days. I think it's more important that we engineer these systems to be safe. So these systems would need to be provably safe. They would need to provide proof that they are safe. And once that's achieved then all this kind of explainability research will collapse because we don't need it anymore. We just need it to have trust in the system and know that it's safe. And then we don't need explanations."

> Respondent 3: "In addition, it can be stated that AI-based applications are black boxes in that sense, and their developers are not very transparent on how they use the data and what data is actually utilized. In principle, the assumption can be that all the data you put in there is used for training purposes."

> Respondent 4: "One young person said having widely tested different identification programs for recognizing AI-produced texts. What kind of language can they recognize as being produced by AI and what is not. It may bring to mind that if we think about education, there exist big questions about the issues of how students' work is utilized and to what services this work is fed, and whether it is ok then thought to use such AI identification programs. "

Although the current legislation in the EU and contractual responsibility bind some of the commercial AI service providers, nevertheless, it was thought that the use of these systems might not be entirely secure. For instance, access granted to these services in an educational organization might need to be limited to only a specific group of staff since there is no complete guarantee of such systems' safety and reliability.

> Respondent 5: "We have put the licensed version of Microsoft365 Copilot into operation for a pilot group that consists of 200 people. For example, it promises to not utilize data for language model training. However, it gets to grind personal files. We dare not use it for such a personnel group that contains sensitive data. Teachers have to opt out of its use since, you see, a teacher is a problem. Although it promises not to use the data for language model training, at this stage, we have not dared give it to

> teachers' usage since we do not have experience with what it is capable of doing. But in practice, then this AI can get access to all this confidential, sensitive personal data that you have in there.”

Responses also emphasized the limitations of commercial options provided in education: if an educational organization decides to adopt a specific AI-based tool, the user’s freedom of choice is restricted at that point, and they have no choice but to accept the terms of the specific service provider and the associated data collection and processing practices.

> Respondent 4: “But then another issue related to the education sector and work and everything is that we are given certain tools to use, where there is not much room for questioning, so that here we use Microsoft services. Then one must agree with the certain Terms of Use if one wishes to use them, and that in addition in organizational level are taken into usage these certain services.”

4.2 Development Practices for Improving Data Privacy and Protection in the Education

When the interviewees considered the suggestions for improvement and development possibilities related to improving individuals' data privacy and protection in the educational context, the following data privacy-enhancing thoughts came up. The respondents highlighted their opinions on, first of all, the highest importance of continuous education on the responsible use of LLMs in schools at all levels. Robust guidelines and policies must be implemented to ensure this, and they must be kept up to date.

While utilizing LLMs in the education context, it would be recommended to use only those commercial, license-based services that are seen as safe and reliable enough for educational organizations' usage. In the responses, interviewees hoped that developer companies that produce these services would take more responsibility and care about data privacy issues. It is clear that individuals also need to go through much trouble if they want to avoid consenting to collect data about themselves through these services. Also, whether the consent of the guardian of a minor student wishes to use LLM-based applications, development ideas were seen in the transparency of consent and data collection and processing, especially in the case of underage students. One vital development potential that emerged from the responses was the urgent need for legislative changes to more effectively address the current and future challenges in data privacy and protection posed by rapidly developing AI.

The interviewees' responses emphasized the relevance and significance of **continuous education and the implementation of robust and comprehensive guidelines and policies** in education when students and educators use LLM-based services and applications. It is crucial not to pass these subjects in any case, whether we want to utilize AI in education responsibly. Based on the responses, the focus needs to be on educating and creating guidelines. However, they also need to be put into practice, kept up to date, and ensured that people follow these guidelines and policies and comply with them.

> Respondent 1: "Well, as I said, teaching students. Basically just the awareness that there could be risks would be one of them. Just making them aware that it's a good idea to anonymize their data when they enter it in any kind of like if you think of it you, when you work in an organizations you want to not enter any sensitive data in, in these models, anything that that's of high value to the organizations you would not want to share. So these are kind of just general guidelines that would be needed and would need to be followed."

> Respondent 2: "First, there are the policies because technically, we cannot do the caulking. In the end, it is the users' responsibility to understand and recognize. Then, clear policies exist, and information about them has been provided and implemented, which is very important. Procedures must exist and that people obey them. There exist environments for different data processing purposes. Of course legislation regulates it. For instance, the Act on the Secondary Use of Health and Social Data strictly regulates the environments in which personal data can be collected. So environments exist for different purposes, but this issue must be known to avoid accidents."

Users fully trusting the content produced by LLMs' content raises severe concerns about the safety and truthfulness of information online. The responses also revealed concerns about the biased and misleading information in the outcomes of LLMs.

> Respondent 1: "And then, also maybe being aware that what you read online is often now generated and it might be false. So just a general awareness about misinformation spreading on the web would also be a good idea."

> Respondent 5: "I'd like to emphasize here at the end that the major issue is the ethical one. There are many ethical risks, biases, and stereotypes, which are significant concerns."

The responses brought up that, especially from younger students' perspective, they need to understand and start caring about what kind of data privacy and ethical risks exist. Constant education is one way to contribute to this. It is vital to get the message through that students, educators, and education staff must not expose any sensitive, confidential information to these services, regardless of whether it is openly accessible or a service operating behind a paywall.

> Respondent 2: “Raising awareness and implementing policies and procedures needs to be highlighted. In practice, we have (at the university level) certain AI-based products recommended for usage. Generally, enterprise architecture includes recommended tools. Some support is allowed, and so on, but a product without support can be used but at one’s own risk. So that is one, that these issues are put into practice, and an understanding of them is also implemented.”

> Respondent 5: “Although, with an organization license, it promises that Microsoft won't utilize the data. We trust this then, but nevertheless, we must educate teachers carefully that no student data or any Excel files and so on are put there. It's essential to ensure that this message gets through to the end. We want to educate students and educators to understand the risks of AI.”

Generative AI will remain a part of everyday educational and work life in the future as well. That said, in the educational context it is necessary to pursue the use of generative AI tools and applications that are mandated by EU legislation, bound by contractual liability, and commercial data protection integrated into these services. Certain service provider organizations guarantee that all user and organization-related data are safeguarded and chat conversations are not used for training language models. Despite the fact that educational organizations use such commercial and somewhat safer LLM-based services, they cannot be fully trusted when it comes to data privacy and protection since the complexity of language models in general and their unclear data collection and processing procedures create privacy concerns. It was seen that **AI developers should be responsible for developing language models with users' data privacy as a top priority** in the development process. AI developer organizations might need to consider and motivate taking effective actions toward more responsible language model development so that data protection is embedded in technology design and production.

> Respondent 2: “Then of course there are these technical tools as well, like this Purview and others, that in some parts help the situation, but certainly when entering a large market and the hype there, in those circumstances, this is not possible. At the university level, it is recommended to use Microsoft Copilot over ChatGPT since it operates on our own cloud machine, Azure. So, this information is better protected than in the public cloud. In addition, it is stored in the EU / ETA area. For instance, like many other universities, we’ll start adopting the Microsoft Purview product, which is a tool for classifying information. The product ensures or forces the user to take a stand on the issues related to visibility and

> identifying items to be classified. In addition, the product takes care of the life cycle. When it comes to Microsoft M365 Copilot, it integrates with Office and such products. The main point is not what a user inputs there, but that it (the product) sees the documents that the user has privilege in the Microsoft 365 environment. So, classifying information becomes vital, as do the tools related to it.”

> Respondent 5: “Another way to use Copilot for teachers is that it is part of the Office-license. If you have A or E5, then you’ll have this Copilot chat for your use, which is on the website, a bit similar to ChatGPT, which cannot access your personal file management. Your personal files are protected, and so is the learning data there. But then there is the problem that one needs to be extra careful that they understand that it is not used in any circumstances to process personal data, nor to utilize it to rank students or anything like such.”

The responses raised the important issue that AI developer organizations must also be responsible for developing services that obey legislation and ethical standards and safeguard user privacy.

> Respondent 3: “And maybe related to your topic, this means that, in general, research often is recommended that nowadays people should be educated with AI literacy, LLM-literacy to be more specific, young students as well, who use ChatGPT. Also, users should be more aware of these potential risks and would act accordingly. In addition, that organizations would not use personal data for training purposes, users would be more aware that they would not input their data into these language models in the first place.It has been noted that people are concerned about privacy issues. Regarding data protection, various license-based options have come to the market, or there are these different partnerships that basically provide the same service or provider’s service via alternative methods and with different terms."

> Respondent 3: “But like always in AI ethics, it’s the companies that should take action. Quite often, the problem is that if they cannot see any commercial benefit, the motivation for actions is pretty thin. And in response, whether the data is sold, that, of course, is a direct financial benefit. So basically, that financial benefit, if one can get it without collecting that data, should somehow, from an organization's perspective, compensate if it does not gain the financial benefit of that data's so-called normal use. Somehow, organizations should take care of these issues and motivate them more. One can always not use the service. However, the impact may be quite small. So here, that is in addition to hoping that organizations would act more ethically and consider the existing and future legislations better with this data collection and processing.”

Although the existing data protection regulation strictly regulates the collection and processing of personal data, it is challenging to guarantee and monitor whether all AI developer organizations comply with the regulations and whether available language models are developed in a privacy-embedded way. For example, it is crucial to clarify how the GDPR's right to obtain and withdraw from data processing, the right to erasure, and rectification principles are adhered to in the context of LLMs. EU AI Act lays down transparency obligations related to users but does not take into account protecting the data privacy of users in the context of generative AI on a larger scale. According to the interview results, it was seen that there might be a need for **amendments to the current legislation**.

> Respondet 3: “One means of influence is that there would be some amendments to the law. Or then, okay, maybe write some addresses to bring attention to these issues and get the organizations to care about them. But here we have conflicts of interest, since organizations use this data as claimed to make the service better and probably among other things, it is precisely used for that, so they use the data for training purposes. And therefore developing AI further and making it better.”

> Respondent 4: “So that should be more taken out that is there a possibility to outline somehow these choices related to technology, so that this perspective is more taken into account that what kind of information is being collected and so on, and should the regulation be then such that it should bring out more clearly so that if the certain platforms collect personal data, and what this data is being used then. It is also a bit like this in everyday operations, it probably does not seem like it matters much if some user data from one's activities is collected. However, it could become a problematic issue in the long run."

> Respondent 5: "Is there such a tool for local legislators and their interpreters that can immediately stop the use of AI? Besides, we have a double standard. We cannot handle consumer issues related to what students use. In the future, students are likely to use it. It will start appearing in primary schools, and especially in secondary schools. If we cannot discuss or use it in schools due to these risks or costs, but we know that its use is comparable to saying, "Don’t use TikTok because it’s a bit silly since it collects all data." Then there would be a double standard—prohibiting its use in education while students use it themselves, which, frankly, enables plagiarism."

> Respondent 2: “Then, of course, there is this EU AI Act, which is not very compelling in the very early phase, but it brings out the use cases, including forbidden use cases related to the usage of AI tools.”

> Respondent 4: “The EU AI Act lays down the risks that are being recognized, and they mostly concentrate on risks related to health and safety. Directly realizing risks, but less in risks over the long term. So respectively, like in the education context, what could these long-term risks be, so maybe they are not very visible yet in regulations.”

One of the interviewees stated that it remains unclear how informed consent is implemented and controlled, especially when the personal data processing of minor students is in question. The GDPR strictly regulates the collection and processing of data related to children. Informed and freely given consent is required to be given by a parent or authorized custodian over the child whenever their data is being collected and processed. It was also seen as problematic whether the withdrawal of their consent is genuinely executed from these language models. More **transparency and lawful, respectful practices** are needed related to guarantee data privacy principles.

> Respondent 4: “The first thing that comes to mind is this forced consent, which is already now, like any other, not just AI-based applications but in general acting according to the GDPR, seems to be one approach to it. In quite many other services as well, one needs to put quite a lot of effort if one does not want to give their consent. That is one problem. It is not transparent at all how this data protection and consent by GDPR is in name only, just barely achieved, so this dilemma of informed consent, which is of course especially then, when we think about it from children and young people's perspectives, should be specifically noted."

> Repondent 3: “Here the problem is, like you said, that it is not always clear where data is stored and for what purposes when fed to the language model. Whether the interest is not to utilize them that much in ethical matters, therefore it weakens a lot of data security because if you put this interview data in there, the presumption is that the data will end up in the US. This might be the most significant or most apparent risk that many researchers acknowledge. At the general, abstract level, this is a risk of where data is stored and for what purposes.”

4.3 Perspectives on Limiting AI’s Access to Data

The last interview question aimed to determine the interviewees' opinions on whether we should limit AI's access to some data. This question divided respondents' opinions. Responses brought up the issue that when accessing AI-based services, there should be an option to opt-out, where users can withdraw their consent given in the past for data collection and processing. However, the reliability of the opt-out possibility and its genuine effectiveness were questioned. From the respondents' answers, it became clear that no sensitive, confidential personal information should be shared with any of these LLMs under any circumstances. This is crucial in the case of AI-based services used by educational organizations, where the privacy of students and educators must be a top priority. To handle and process such confidential personal data, educational organizations must ensure they have robust and secure environments in place, providing reassurance about students, educators, and educational staff's data privacy.

Some respondents stated that it also depends on the context when limiting AI's access to some particular data. The respondents who noted this issue also stated that if personal data is being used only to improve and take better care of their health care, they would not see it as a problem that AI would get access to such (anonymized) personal data in this exact use context. Respondents also expressed concern about the potential distortion of data used for model training and the ethical risks this poses. The fact that training data might include personal data heightens these risks. It is vital that the data language models are trained with high-quality filtered and improved data and that personal information is removed from among it.

It is not easy to limit or control access to some particular personal data using generative AI. However, using a particular online service without users sharing any information about themselves is unrealistic. One interviewee stated there needs to be at least a **possibility for opt-out option** within the use of LLMs, meaning that a user can withdraw the consent they had given in the past for personal data collection and processing for a specific service provider.

> Respondent 1: “Oh that's a tricky question because I think you would limit the progress that could be made with these systems if you don't use all available data. So I don't think it should be limited, but there should be a right to opt out. It was the same thing with Google Street View for example. They just took pictures of everything and then afterwards you could opt out. I think it's a really great service but the model that they use to gather all this data is of course a bit questionable. Gathering all this data is of course a bit questionable, but it's still a really great service, so I would not want to miss it, so I would also not want to restrict data given to these models.”

> Respondent 3: “And the other perspective is what data is inputted into these models after they have been initially trained. But here’s the issue: even if users have the option to opt out, such as telling ChatGPT that their chat data should not be used for training, is that enough? Probably not."

In general, as a key rule, **no sensitive, confidential data** should be entered into LLMs. However, there are situations for this kind of data collection and processing. It should only happen in a protected and safeguarded environment, that is specifically developed for such data processing purposes.

> Respondent 2: “Here the trickiness is where the information is and how it is handled. It is good to think so, that sensitive or highly confidential data cannot be given or put into this service, if it is not safeguarded, and data security and data protection angles related to this service are taken into account. For instance, interviews like these could contain very confidential data when we start talking and researching about individuals' health information or mental health issues. It is okay if this data is in a reliable and secure environment. But if this kind of data were put to Microsoft Copilot, it would not be okay anymore."

> Respondent 5: “Yeah, I wouldn’t of course input any student data or data from the student management system there. Entrance exams, certain high-risk data, social scoring, and similar issues. I’d not want to block the usage of AI entirely. I think that it could be utilized in the education environment, low-level risk AI, that is.”

While language model training happens with various kinds of data from public sources from the Internet, it is essential to consider whether the **training data can be improved** and filtered and personal information removed from this data set. In addition, it would be necessary to eliminate unethical and unlawful material from this training data. Naturally, there are many ethical concerns in the context of generative AI. Alongside privacy concerns, misinformation, bias, and fairness issues exist with LLMs.

> Respondent 3: “There are a couple of aspects there. One of the LM-based issues is that they are being connected with poor training material because they are being trained with Common Crawl, which is a bit of random data, an enormous amount of data from the Internet overall. In such datasets, there are various, like personal data leaked in data breaches or similar data, which is a bit questionable data but used because it is such a huge corpus. So concretely, training material should be improved, and correspondingly, personal data should be derived from this training material. And of course, gathered datasets are filtered, but this filtering happens mostly in third-world countries, and there can be horrible material included. Then we talk about the human price in filtering this training material.”

Unethical, biased, and misleading information LLMs also provide may affect younger students' thinking and mindset.

> Respondent 5: "Sure, it forms a certain worldview, but similarly to TikTok, I'm also a bit concerned that we might end up with a situation where both TikTok and AI shape everyone's thinking in a peculiar direction. It's a significant challenge for young people."

Limiting AI's access to personal data also seems to be a **context-dependent issue**. Restricting AI's access to specific data initiated considerations regarding the context in which personal data would ultimately be used. Some interviewees noted that such a restriction would be appropriate if their data were used to enhance commercial services and marketing. However, whether interviewees' data were used to improve their medical treatment, they would not see it as an issue to provide information about themselves for this particular purpose.

> Respondent 4: “Of course it depends on the purpose. First what comes to my mind is that, especially those commercially operated processing and using data, it’d be wise to consider what really serves those peoples’ benefits and in a way, where and for what purposes people do want their data to be processed and used. It's a bit like asking whether it’s really okay for us to give our data for the purpose of better marketing these products to us. But if this data is utilized for that, we could be medically treated better, so many individuals would be happy to give all data related to them for that kind of purpose.”

> Respondent 5: “My data is possibly used somewhere there, when I use the wellbeing services, so the_Act on the Secondary Use of Health and Social Welfare Data enables the handling of anonymization. It is apparently the base for the common drug development and service development that the data is being grinded. Or course I’d give the training data there, but particularly anonymized, in this case. So, this depends on the context.

The following chapter presents the results and answers to this study's research question. It also presents the recognized development practices for improving students' and educators' data privacy and protection in the educational context.

5 RESULTS

This thesis aims to define the improvement guidelines and practices to safeguard users' privacy as using LLM-based applications becomes increasingly prevalent in the education sector. The answers to the research questions are based on the expert interview results, which reflect the outcomes of previous studies presented earlier in this study. In order to focus on finding methods to improve data privacy in the context of LLMs, it is essential to recognize the related risks and challenges. Most themes related to data privacy issues and improvement practices that arose from interview data were similar to results from previous research. TABLE 5 below presents similar risks and suggestions for improving data privacy that arose from the interview data and previous studies. The risks and improvement practices resulting from the interview data alone are highlighted in red and emphasized in bold in TABLE 5. The suggested improvement practices are presented in the right column, and the risks being addressed by these provided practices are in the left column.

TABLE 5 Identified data privacy risks in the context of LLM-based applications in education and improvement practices to address those challenges.

Data privacy risks and challenges	**Development Practices to Improve Data Privacy and Protection**
Lack of proper education on the responsible use of LLMs	**Continuous and effective education of students and educators**
Individuals attitudes towards data privacy in education	Implementation of robust and up-to-date guidelines and policies
Lack of trust in LLM-based applications and services	Development of privacy-embedded language models **Increased responsibility of data privacy for AI developer organizations** High quality training data
The dilemma of informed consent and ambiguity of erasure of personal data	More transparency on informed consent, data collection and processing **Increased data privacy responsibility for AI developer organizations** Amendments to the current legislation
Disclosure of sensitive information and data leakage	**Continuous and effective education of students and educators** Development of privacy-embedded language models

Uncertainty about the scalability and efficacy of existing data protection legislation	Amendments to the current legislation
Personal data collection and processing for suspicious purposes	More transparency on informed consent, data collection and processing

The following sections answer this study's research question *"What are the guidelines and practices to improve users' privacy and data protection as the use of LLM-based applications becomes increasingly prevalent in the education sector in the future?"* by presenting in more detail the practices and guidelines for improving users' data privacy and protection in using LLM-based applications in the educational sector based on interview results and reflecting them to the outcomes of previous studies presented in this thesis.

There are many benefits of utilizing generative AI in the education sector, and the usage of LLM-based applications and services will be tied to educational use in the future as well (Kasneci et al., 2023; Trust et al., 2023; Meyer et al., 2023; Neumann et al., 2023). The interview results highlighted the relevance and importance of **continuous education and the implementation of robust policies and guidelines** in the educational sector to improve students' and educators' data privacy and to guarantee the responsible use of generative AI**.** Without them, students and educators might not understand the existing data privacy risks in using these services and applications, leading to various privacy violations. There can be unfortunate situations where confidential and sensitive information may be unknowingly shared about a student or other students or educators, or such personal student data is being processed and stored improperly, creating data privacy risks. Due to the lack of guidelines and policies, students and educators might not understand the processes and limitations of LLM-based services and applications to use them responsibly and protect their privacy.

Based on the interview data, guidelines and policies must include the recommended LLM-based tools for education usage and processes and how to use them responsibly, as well as providing detailed information about the data that should not be entered into these systems and tools. This data, which includes personal data and confidential information, should be handled with utmost caution and only in environments designed for that specific purpose. It was also seen as crucial to clarify the potential privacy risks and violations that can occur when personal and confidential information is exposed to LLMs and the related consequences.

Negative consequences exist for students' and teachers' data privacy in relation to irresponsible use of LLM-based tools and applications, whether they are tested and used impulsively. The consequences might be that students and educators do not realize and understand for which kind of personal data collection, processing, and storing purposes they have provided their consent, which might create privacy issues in the long run. According to the interview results, education organizations already have safe environments to process personal data in particular situations when needed. For instance, academic education institutions in Finland (Helsinki University IT Helpdesk, n.d.; Teaching and Learning Centre, 2023) recommend that students and educators use specific commercial, license-based AI services for education purposes, like Microsoft Copilot, which guarantees users' data is not provided for language model training purposes. In addition, these AI developer organizations are bound by contractual liability and EU legislation. However, there needs to be education and guidelines concerning the use of these services to guarantee individuals' privacy and data protection.

The interviews stated that educational institutions could only partially participate in improving students' and teachers' data privacy in the context of utilizing generative AI by providing continuous education for them and implementing guidelines and policies accordingly. The reality is that there is a growing number of available LLM-based applications that students at all levels may utilize in their studies. However, interview responses stated that in the end, it is the users' responsibility to assimilate the provided guidelines and education, identify risks, and understand the limits of the responsible use of AI.

Interview results noted that disclosure of personal information and individuals' attitudes related to their privacy and data protection were seen as privacy risks in the context of LLM-based services and applications in education. People have different attitudes and reactions related to their privacy and the collection of personal data. For instance, according to Rao & Pfeffer (2020), younger people want and are able, from their starting points, to tolerate data protection procedures differently compared to older individuals. Individuals' education level also influences perspectives toward personal data collection and protection practices. Especially from younger students' attitudes, based on unconsciousness and somewhat indifference, it was seen that they are more eager to utilize these different applications quickly than care about their data collection and processing within these services. According to the previous studies (Glorin, 2023), it was also noted that when disclosing personal information to LLMs, this kind of unintended personal data revealing could occur in situations where a person believes that the human-like chatbot they are interacting with is reliable.

In the interview results, it was noted that when students utilize LLM-based services to write essays or reports when the subject might be related to themselves (like psychology or health education), they might accidentally expose such data to these applications they were not meant to expose initially. The disclosure of personal information can happen quite easily in this specific context. Educating students and teachers on these issues is essential, as well as implementing guidelines and practices on how students and teachers are required to report these incidents, whether they happen. Interviewees also emphasized that students should understand that information produced by LLMs may also be unethical, biased, and misleading. Therefore, educating students and educators about potentially harmful and inaccurate AI-generated information should not be entirely trusted.

> Respondent 1: “I mean, you need to be aware of any kind of problems that could arise. Educating this awareness, teaching this awareness, is the main problem, I think.”
>
> Respondent 2: "There certainly exists many concerns related to the topic, some of which are relevant. If someone says that you now need to take over AI, we are currently moving in such a phase that a 100% takeover of AI is impossible. Policies must be implemented as soon as possible, and the pursuit of continuous learning about this issue. After all, it is the user's responsibility to understand and identify. Then, robust policies are implemented, and being informed is very important.”
>
> Respondent 3: "Due to the usage of these LLMs, it has become apparent, maybe not yet, as that apparent as much as we would like, but it would need to be highlighted is the user's responsibility also. The reality after all is, that these tools are here to stay. Here, the user should use them ethically."

According to the interview results, development organizations should take **more responsibility for safeguarding users' privacy and data protection rights** while developing language models. The interviews revealed that there needs to be more transparency regarding student data collection and processing purposes in the context of LLMs. There is a general perception that these applications and services initially collect all data that users might input to them for further training purposes, leading to a lack of trust in LLMs. Thus, they need to be secure, gain user trust in the system, and know that it is safe to use.

The interview results highlight the need for more motivation among AI developer organizations to comply with data protection legislation. The responses stated that without a commercial benefit, these organizations may not be inclined to adhere to these regulations. This lack of motivation underscores the need for some form of compensation if they do not gain financial benefits from the everyday use of data. It was seen in the responses that LLM developers should comply with data protection laws and consider embedding user privacy as a high priority from the very beginning of the language model's development. This issue requires immediate attention. Previous studies (Winograd, 2023) also noted that it remains to be seen how AI developer organizations control or delete personal information safeguarding their users' data protection rights. If personal data erasure applies to data embedded in the model, model retraining might be required to meet this request.

Amendments to the current data privacy legislation were highlighted in the interview results, so the current legislation might need to be revised to address the risks and challenges posed by AI. The GDPR regulates personal data processing, but it was considered during interviews whether it is alone or, as such, sufficient enough to tackle the issues AI brings since it is not focused on regulating data processing precisely in AI but is applicable in specific contexts related to it. The interview results highlighted that the EU AI Act might not focus on data privacy issues in the context of AI. According to the interview results, concern was raised over the issue of data privacy principles regulated by the GDPR needing to be secured in the context of AI.

The respondents also considered to what extent the current data protection legislation in the EU can address the data privacy and data protection risks of generative AI since this legislation came into force in 2016 to address the data privacy challenges at that time. However, the GDPR regulations can be applied to the context of AI. However, the exact scope remains unclear, and not all AI developer organizations obey these regulations, or obeying the regulations is apparent and is difficult to prove. It was seen in the interview responses that there might be needed amendments to the current legislation to tackle data privacy challenges generative AI poses currently and risks in the future. The responses highlighted that this kind of comprehensive user data privacy and data protection ensuring landscape in the context of generative AI requires long-term cooperation with legislators, policymakers, and AI developer organizations and sufficient motivation in order to do so. For instance, LLM training data can contain information about individuals from various social media platforms, which brings up data privacy and ethical considerations. Individuals utilizing such services should be guaranteed and safeguarded with consistency and data privacy within these language models.

The interview results raised concerns about individuals' ability to give informed consent for data collection and processing and the option of having one's data removed later on from the LLM-based systems. It was seen as especially problematic when the consent of a minor student (under 18 years old) needed to be obtained from their parent or guardian to utilize such applications, how the consent was indeed first of all obtained, and for what purposes personal data was ultimately collected. These issues were not seen as transparent and reliable. The responses noted that it is problematic whether an individual would not want to give their consent to LLM-based services since it was not clear what kind of effort they need to go through to withdraw their consent and have their data removed, whether these are realistic rights in the context of LLMs at all. Whether such consent is not obtained or if students nor their parents do not entirely understand the data processing purposes, it can lead to privacy risks. The interview discussions brought up the need for greater transparency regarding the withdrawal of consent for data processing in the past and the right to erasure one's personal information provided to LLM-based services. It was seen as necessary to clarify whether these data protection rights could be adequately implemented in the context of LLMs. This was noted in the previous studies (Winograd, 2023; Plant et al., 2022) as well. LLMs' inscrutable memory and the unknown behavior behind them weaken the efforts of individuals who expect to rely on the GDPR and take away the individuals' right to withdraw their consent and right to erasure. Leading AI organizations need to be aware of the potential of long-term cooperation with legislators and policymakers to ensure the protection and safety of individuals' privacy now and in the future. Such cooperation emphasizes the necessity for sustained efforts and commitment to ongoing collaboration among all stakeholders. Generally stated, clarifying the existing legislation may endorse more responsible and reliable development of language models, obey the GDPR, and help recognize the current and future demands and directions for amendments in the existing legislation.

> Respondent 1: "And really, if you really don't want to have your data shared, then you just need to not use the models anymore, but at some point I think we will reach a point in society where you can no longer do that. You cannot not use your phone or not use the Internet for looking up information. So every time you do that, you share information about yourself. So it will get more and more difficult to opt out of the data gathering."

> Respondent 4: "That is one considerable issue, for instance, how the use of these applications can affect our media and information environment and others, but correspondingly, like in the educational context, what the long-term risks may be, might not be that visible in this regulation yet. I think that it is a positive thing that in the EU, we have started to act towards more regulations. Maybe I'd think that, in some cases, there could be more restrictions. Somehow, to think about this more broadly, precisely these possible data privacy risks and accidents that can happen to people. Also, more transparency must be involved in data processing. That is one considerable issue, for instance, how the use of these applications can affect our media and information environment and others, but correspondingly, like in the educational context, what the long-term risks may be, might not be that visible in this regulation yet."

Language model training data may contain personal data. Additionally, training data may include unethical content, such as biased or inaccurate material. The interview results stated that AI developer organizations should **focus on producing more high-quality data to tackle these issues.** The responses suggested that training data might need to be refined to remove unethical material and personal information from this data. The responses were also seen as a risk that unethical and misleading information provided by LLM-based chatbots impact younger students' thinking and lead to misunderstandings. The issues of the possibility of biased and incorrect information in the outputs of LLMs and the improvements to language model training data also came up in previous studies (Kooli, 2023; Winograd, 2023). When the training is biased, responses may result in biased outcomes. AI developer organizations must focus on quality and filtered data when training language models to improve users' data privacy and remove unethical information from this dataset. This may call for retraining language models in periodic phases with filtered and updated data where the personal data of individuals is removed from this material.

According to the interviews, **limiting AI's access to personal data** divided opinions. The responses highlighted that at least language model developers need to allow their users to opt out of data processing and focus more on improving language model training material, like removing unethical and personal data from it. Users should be able to withdraw their given consent in the past related to their data processing and have their data erased from these models. Some respondents stated that the limitation of AI's access to data depends on the context. In general, as a golden rule, it is a good way to think that no confidential data should be entered into AI in any case. However, in situations where such personal data must be processed for a specific need, such data must be handled in a secure environment designed for such data processing. In addition, some respondents stated that whether their anonymized data would be utilized in the context of AI for improving their health care, they would not see it as a huge issue.

6 DISCUSSION AND CONCLUSION

Generative AI is currently strongly present in the education sector, and LLM-based applications and services are increasingly utilized in different learning and teaching-related activities and tasks. As this thesis has previously presented, with the development of generative AI, data privacy and protection in this context exist today and also in the future. This thesis first delved into identifying data privacy risks and challenges in the context of LLM-based applications and services and then presented concrete development practices to tackle the recognized risks and improve privacy and data protection in the education sector.

The recently adopted EU AI Act presents various restrictions and regulations based on risks associated with AI systems, bringing more transparency obligations for AI developer organizations. For example, the act requires them to notify users whenever content is created artificially and develop systems to prevent them from producing unlawful content. The current data protection legislation seeks to tackle data collection and processing issues generative AI poses. The GDPR, which came into force in 2016, regulates the collection and processing of individuals' data and the rights related to personal data. However, according to the current understanding, it seems unclear whether a growing amount of publicly available, free-to-use LLM-based services are GDPR compliant, even though some service providers claim to obey the data protection legislation. For example, OpenAI assures that it processes personal data based on the GDPR, but monitoring and confirming this may be challenging. ChatGPT collects and stores all the information that its user enters into it, and it is not responsible in the same way in the name of legislation and contractual obligations as a free-to-use service compared to commercial applications like Microsoft's Copilot, which is currently widely utilized at the academic level in Finland. However, while educational organizations in Finland utilize license-based LLM-based services with additional commercial data protection implemented as part of the service, it may be vital to focus on educating teachers and students not to enter sensitive, personal information into the services. In other words, such LLM-based services recommended for educational usage cannot be considered completely 100% reliable. As presented previously, the unknown behavior of data collecting and processing in these language models create concerns about the purposes for which our data is collected and utilized.

The subject is constantly evolving, and new LLM-based applications are coming to the market at a remarkable speed. Specific technical, privacy-preserving techniques in language models exist, but they also have their pitfalls. It is evident that students at schools at different levels currently and in the future as well will also continue using freely available LLM-based services and applications. Indeed, some of those users are minors (under 18 years old). Protecting their privacy requires extra attention, parental control, and more transparency on data collection, processing, and obtaining informed consent from AI developer organizations. According to the GDPR, individuals must provide explicit consent for processing their data and be given explanations and grounds for collection and processing purposes. Individuals have the right to withdraw their consent for personal data processing at any time and to have their data deleted from the system. Whether these data protection rights are ensured within generative AI is still being determined. The current market is rapidly developing, with new services and applications being developed and tested continually, and unfortunately, often without considering their responsible development and usage.

The possibilities of utilizing generative AI are comprehensive in the educational context. However, responsible use of it, where data privacy is a top priority, necessitates more actions from AI developer organizations, up-to-date legislation to address the current and future challenges posed by AI, and clear and practical guidelines and practices to enhance and safeguard individuals' privacy and data protection. The unknown nature of language models and the complex methods by which they process personal data for various purposes remains unclear even for AI developer organizations. Robust measures are needed together from legislators, policymakers, and AI developer organizations to guarantee secure individuals' data privacy and protection now and in the future. AI developer organizations may be required to improve transparency for data collection and processing and accountability, focus on improving more high-quality training data, and improve user control over their privacy according to the GDPR data privacy principles.

Data protection and user privacy must be prioritized and safeguarded throughout the model's lifecycle, from the initial design phase to deployment and ongoing use. Amendments to the existing legislation might be needed to address data privacy challenges generative AI poses, but those future risks have not yet been realized. The responsible use of LLM-based applications in the educational sector demands educational institutions to urgently and efficiently develop proper guidelines and policies and to constantly educate students and educators to understand the responsible use of generative AI and its limitations to improve users' data privacy within the use of such applications. In addition to all previously mentioned, securing one's data privacy within these different LLM-based services and applications also falls on the users' shoulders. Individuals are responsible for understanding the privacy and ethical risks and challenges posed by AI and considering these aspects in their responsible use.

The use of generative AI in educational contexts is rapidly increasing, raising concerns about users' privacy and its ethical use. Since this thesis focuses on current and future-oriented themes, examining explicit recommendations for improving privacy in the use of LLM-based applications and services in the education sector, these can be seen as limitations for the study. In addition, the findings in this study may not necessarily be directly applicable to other fields or sectors. Overall, the impacts of generative AI on individuals' privacy and data protection are diverse and substantial, with not all risks and challenges yet identified. Generative AI is continuously evolving, emphasizing the need for further research within the framework of technological development and how existing legislation can address the privacy and data protection risks and challenges it presents. As the results of this study present, various data privacy challenges and issues exist in the context of LLM-based applications and services, in the context of education, and at the general level. Therefore, future research might focus on examining the broader impacts of specific data protection challenges and provide more comprehensive improvement suggestions related to this challenge in the context of generative AI.

The Use of Artificial Intelligence in the Thesis

The Grammarly application has been utilized in this thesis to produce grammatically correct and structurally fluent text.

REFERENCES

Aslam, B., Karjaluoto, H., & Varmavuo, E. (2022). Data obstacles and privacy concerns in artificial intelligence initiatives. *Contemporary Issues in Digital Marketing*, 130-138. Routledge. https://doi.org/10.4324/9781003093909-16

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

Berendt, B., Littlejohn, A., & Blakemore, M. (2020). AI in education: Learner choice and fundamental rights. *Learning, Media and Technology, 45*(3), 312-324. https://doi.org/10.1080/17439884.2020.1786399

Brown, D., Ako-Adjei, K., Pack, C., Robertson, S., & Srivastava, A. (2024a). *Microsoft Copilot for Microsoft 365 overview*. Microsoft Learn. Retrieved September 6, 2024, from https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-overview

Brown, D., Robertson, S., & Simpson, D. (2024b). *Data, privacy, and security for Microsoft Copilot for Microsoft 365*. Microsoft Learn. Retrieved September 6, 2024, from https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-privacy

Brown, H., Lee, K., Mireshghallah, F., Shokri, R., & Tramèr, F. (2022). What Does it Mean for a Language Model to Preserve Privacy? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 123-134. ACM. https://doi.org/10.48550/arXiv.2202.05520

Cao, Y., & Yang, J. (2015). Towards Making Systems Forget with Machine Unlearning. *IEEE Symposium on Security and Privacy*, 463-480. https://doi.org/10.1109/SP.2015.35

Chang, C.-C. (2024). When AI remembers too much: Reinventing the right to be forgotten for the generative age. *Washington Journal of Law, Technology & Arts, 19*(3). Retrieved September 6, 2024, from https://digitalcommons.law.uw.edu/wjlta/vol19/iss3/2

Das, B., Amini, M., & Wu, Y. (2024). *Security and privacy challenges of large language models: A survey*. arXiv. https://doi.org/10.48550/arXiv.2402.00888

Davis, K., Edwards, D., & Robertson, S. (2024). *Privacy and protections*. Microsoft.com. Retrieved September 6, 2024, from https://learn.microsoft.com/en-us/copilot/privacy-and-protections#organizational-data

Elo, S., Kajula, O., Tohmola, A., & Kääriäinen, M. (2022). Laadullisen sisällönanalyysin vaiheet ja eteneminen. *Hoitotiede, 34*(4), 215-225.

European Commission (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).*

European Commission (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).*

European Parliament (2024). *EU AI Act: First regulation on artificial intelligence*. European Parliament. Retrieved September, 6, 2024, from https://www.europarl.europa.eu/pdfs/news/expert/2023/6/story/20230601STO93804/20230601STO93804_en.pdf

European Parliamentary Research Service, Scientific Foresight Unit. (2020). *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence* (STOA PE 641.530). Panel for the Future of Science and Technology. Retrieved September 6, 2024, from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf

Ertel, W. (2017). *Introduction to artificial intelligence*. Springer.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., & Li, Q. (2024). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 6491–6501. Association for Computing Machinery. https://doi.org/10.1145/3637528.3671470

Glorin, S. (2023). *Privacy and data protection in ChatGPT and other AI chatbots: Strategies for securing user information*. SSRN. https://doi.org/10.2139/ssrn.4454761

Hadzovic, S., Mrdovic, S., & Radonjic, M. (2023). A path towards an Internet of Things and artificial intelligence regulatory framework. *IEEE Communications Magazine,* *61*(7), 90-96. https://doi.org/10.1109/MCOM.002.2200373

Hair, J. F., & Page, M. (2015). *The essentials of business research methods* (3rd ed.). Routledge. https://doi.org/10.4324/9781315716862

Helsinki University IT Helpdesk. (n.d.). *Microsoft Copilot at the university*. University of Helsinki. Retrieved September 6, 2024, from https://helpdesk.it.helsinki.fi/en/instructions/information-security-and-cloud-services/cloud-services/microsoft-copilot-university

Hsieh, H.-F., & Shannon, S. E. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Research, 15*(9), 1277-1288. https://doi.org/10.1177/1049732305276687

IBM. (n.d.-a). What are large language models (LLMs)? *IBM.com*. Retrieved September 6, 2024, from https://www.ibm.com/topics/large-language-models?mhsrc=ibmsearch_a&mhq=large%20language%20model

IBM (n.d.-b). What is a chatbot? *IBM.com.* Retrieved September 6, 2024 from https://www.ibm.com/topics/chatbots

Inayatullah, S. (2006). Anticipatory action learning: Theory and practice. *Futures, 38*(6), 656-666. https://doi.org/10.1016/j.futures.2005.10.003

Ishii, K. (2019). Comparative legal study on privacy and personal data protection for robots equipped with artificial intelligence: Looking at functional and technological aspects. *AI & Society, 34*(4), 509–533. https://doi.org/10.1007/s00146-017-0758-8

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G, Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seifel, T., Stadler, M. & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kooli, C. (2023). Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability, 15*(7), 5614. https://doi.org/10.3390/su15075614

Kshirsagar, P., Jagannadham, D., Alqahtani, H., Quadri, N. N., Islam, S., Thangamani, M., & Dejene, M. (2022). Human Intelligence Analysis through Perception of AI in Teaching and Learning. *Computational Intelligence and Neuroscience,* 1-9. https://doi.org/10.1155/2022/9160727

Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. *Journal of Marketing, 81*(1), 36-58.

McCarthy, J. (2007). *What is artificial intelligence?* Computer Science Department, Stanford University. Retrieved September 6, 2024, from http://www-formal.stanford.edu/jmc/

Meyer, J. G., Urbanowicz, R. J., Martin, P. C. N., O'Connor, K., Li, R., Peng, P.-C., Bright, T. J., Tatonetti, N., Won, K. J., Gonzalez-Hernandez, G., & Moore, J. H. (2023). ChatGPT and large language models in academia: Opportunities and challenges. *BioData Mining, 16,* 20. https://doi.org/10.1186/s13040-023-00339-9

Microsoft Education Team. (2023). *Expanding Microsoft Copilot access in education*. AI in Education. Retrieved September 6, 2024, from https://educationblog.microsoft.com/en-us/2023/12/expanding-microsoft-copilot-access-in-education

Musch, S., Borrelli, M. C., & Kerrigan, C. (2023). Balancing AI innovation with data protection: A closer look at the EU AI Act. *Journal of Data Protection & Privacy, 6*(2).

Neumann, M., Rauschenberger, M., & Schön, E.-M. (2023). *"We need to talk about ChatGPT": The future of AI and higher education*. https://doi.org/10.25968/opus-2467

OpenAI (2022). Introducing ChatGPT. *OpenAI.com.* Retrieved September 6, 2024, from https://openai.com/blog/chatgpt

OpenAI (2023a). Europe privacy policy. *OpenAI.com.* Retrieved September 6, 2024, from https://openai.com/policies/eu-privacy-policy

OpenAI (2024a). OpenAI Privacy Request Portal. *OpenAI.com.* Retrieved September 6, 2024, from https://privacy.openai.com/policies

OpenAI (2024b). How ChatGPT and our language models are developed. *OpenAI.com.* Retrieved September 6, 2024, from https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

Plant, R., Giuffrida, V., & Gkatzia, D. (2022). You are what you write: Preserving privacy in the era of large language models. *arXiv.* https://doi.org/10.48550/arXiv.2204.09391

Rao, A. & Pfeffer, J. (2020). Types of Privacy Expectations. *Frontiers in Big Data, 3,* https://doi.org/10.3389/fdata.2020.00007.

Rigaki, M. & Garcia, S. (2023). A Survey of Privacy Attacks in Machine Learning. *ACM Computing Surveys, 56,* 1 - 34.

Spataro, J. (2023, March 16). *Introducing Microsoft 365 Copilot – your copilot for work*. The Official Microsoft Blog. Retrieved September 6, 2024, from https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

Teaching and Learning Centre, Tampereen yliopisto, & Tampereen ammattikorkeakoulu. (2023). Tekoäly opetuksessa. Retrieved September 6, 2024, from https://www.tuni.fi/tlc/suunnittelu/digipedagogiikka/tekoaly-opetuksessa/

The European Commission's High-Level Expert Group on Artificial Intelligence. (2018). *A definition of AI: Main capabilities and scientific disciplines*. European Commission. Retrieved September 6, 2024 from https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Trask, A., Keeling, G., Belle, B., de Haas, S., Ibitoye, Y., & Gabriel, I. (2024). Chapter 13 - Privacy. *The ethics of advanced AI assistants*. Google DeepMind. https://arxiv.org/pdf/2404.16244

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education, 23*(1), 1-23. Society for Information Technology & Teacher Education.

Warwick, K. (2013). *Artificial intelligence: the basics*. Routledge.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2021). *Ethical and social risks of harm from language models*. DeepMind.

Westin, A. (1967). Privacy and Freedom.The United States of America, New York: McClelland & Stewart Ltd.

Winograd, A. (2023). Loose-lipped large language models spill your secrets: The privacy implications of large language models. *Harvard Journal of Law & Technology, 36*(2).

Wu, K., Wu, E., & Zou, J. (2024). *ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence*. arXiv. https://doi.org/10.48550/arXiv.2404.10198

Yu, P., Xu, J., Weston, J., & Kulikov, I. (2024). Distilling system 2 into system 1. *Arxiv.org*. Retrieved September 6, 2024, from https://arxiv.org/html/2407.06023v1

Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., & Tang, J. (2024). *The good and the bad: Exploring privacy issues in retrieval-augmented generation (RAG)*. arXiv. https://doi.org/10.48550/arXiv.2402.16893

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity*. https://doi.org/10.1155/2021/8812542

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). *Explainability for large language models: A survey*. *ACM Transactions on Intelligent Systems and Technology, 15*(2). https://doi.org/10.1145/3639372