

**The future of learning?: The role of large language models in
Danish high schools explored**



AALBORG UNIVERSITET

Masters thesis

Ms.c. Business Data Science

Aalborg University 2024 - 3st June 10:00

Characters: 174,715/192,000

Supervisor: Primoz Konda

pk@business.aau.dk

Kasper Raupach Haurum

khauru18@student.aau.dk

Md. Raiyan Alam

malam22@student.aau.dk

Declaration

The authors hereby declares that, except where duly acknowledged and referenced, this research study is entirely their own work and has not been submitted for any degree or other qualification in Aalborg University or any other higher educational institution in Denmark or abroad.

Kasper Raupach Haurum

Kasper Raupach Haurum - 20182628
03/06/2024

MD Raiyan Alam

Md. Raiyan Alam - 20221400
03/06/2024

Abstract

This thesis examines the usage of Large Language Models (LLM) in Danish high schools (HEI) by students, focusing on the influence upon educational practices and learning. It uses a mixed-methods approach, combining qualitative interviews with teachers and students, and quantitative surveys from students to collect the main body of our data. The thesis employs the Forced Response Design (FRD) variant of the Randomized Response Technique (RRT) to ensure respondent anonymity and accuracy in sensitive quantitative data collection.

The findings reveal two things in relation to LLM use in HEI education. On one hand, LLMs provide support in learning processes by offering assistance and better understanding of various academic subjects. On the other hand, challenges such as dependency on LLM use, potential biases in model outputs, and the need for proper rules and training for both teachers and students are highlighted. Likewise, it is revealed that gender has an influence as far as dependency on LLM use, whereas women have a larger tendency to depend upon it, whilst men are using it more frequently. Another aspect of the analysis is the exploration of how LLMs influence student engagement and have pedagogical impact on teaching methodologies. The findings provide a clarification of how LLMs can either complement or disrupt traditional educational practices. The analysis leads to the thesis identifying best practices and areas that require further attention to optimize the integration of LLMs in the classroom.

The thesis concludes with recommendations for effectively integrating LLMs into educational settings. It suggests the development of learning about LLMs for teachers and students, implementation of rules to address proper LLM use, and ongoing assessment of LLM performance and impact on student learning. These steps aim to maximize the benefits of LLMs while mitigating potential drawbacks, ensuring their responsible and effective use in education.

Keywords: Large Language Models, Education, Danish High Schools, Mixed-Methods Research.

Acknowledgements

We acknowledge that accomplishing the research conducted in this thesis was not possible to carry out without the assistance, and the opportunity to generate primary data in the data collection, as conducted via our qualitative and quantitative choice of methodology. This data was collected with the assistance and consent given by Ringkjøbing Gymnasium, and Herningsholm Erhvervsskole & Gymnasier. Therefore, we wish to extend our gratitude and acknowledgement that without their involvement this thesis would not be possible.

Secondly, we want to acknowledge the help of the two contact persons representing the two HEI Danish institutions, from Herningsholm, Teacher Jørgen Hammer Thomsen, and from Ringkjøbing, Principal Lars Roesen. The two contacts were used as intermediaries to coordinate the data collection, as well as set up the 12 interview sessions held.

We, in reflection of the written thesis, see great potential in including stakeholders that are relevant to the phenomena in question, and this thesis greatly benefited doing so.

Figure 1: Herningsholm Erhvervsskole & Gymnasier logo (Herningsholm, 2024)



Figure 2: Ringkjøbing Gymnasium logo (Ringkjøbing, 2024)



Table of Contents:

Abstract	4
Table of Contents:	6
Abbreviations:	8
1. Introduction	9
1.1 Background	9
1.2 Research purpose	11
1.3 Problem formulation	12
1.4 Research structure	13
1.5 Delimitation	14
1.6 Research Value	14
2. Literature Review	16
2.1 Historical Development of Large Language Models	16
2.2 LLMs in Education	17
2.3 Model alignment	21
2.4 Pedagogical modes of teaching	24
2.5 Prompt engineering	28
2.6 Embeddings	30
2.7 Vector Store	32
2.8 Retrieval-Augmented Generation (RAG)	33
3. Research methods	35
3.1 Philosophy of science	35
3.2 Theme and coding	38
3.3 Randomized Response Technique (RRT)	40
3.4 Technical Architecture of RAG Application	42
4. Data analysis	45
4.1 Qualitative data analysis	45
4.1.1 Theme 1: The usage of LLMs in education	47
4.1.2 Theme 2: Positive aspects of LLM use in HEI	49
4.1.3 Theme 3: Challenges and issues with students LLM usage	51
4.1.4 Theme 4: LLMs are changing HEI educational practices	53
4.1.5 Theme 5: How to solve the current problems associated with LLMs	55
4.2 Quantitative data analysis	58
4.2.1 Question 1: Using Outside Exam	58
4.2.2 Question 2: Using For Exam	60
4.2.3 Question 3: Using In Exam	62
4.2.4 Question 4: Using For Assignment	64

4.2.5 Question 5: Using For Addiction	66
4.3 RAG Application Analysis	70
5. Findings of the thesis	70
5.1 - Qualitative findings	70
5.1.1 AI Alignment issues and LLM reliability	71
5.1.2 Prompt engineering challenges	71
5.1.3 Zone of Proximal Development (ZPD) and student independence	72
5.1.4 Ministry of Education guidelines and regulatory challenges	73
5.2 - Quantitative findings	75
5.2.1 Gender wise LLM usage	75
5.2.2 School wise LLM usage	76
5.2.3 School Year wise LLM usage	78
5.2.4 Class Type wise LLM usage	79
6. Discussion	80
7. Conclusion	82
7.1 Reflections	84
7.2 Limitations	86
7.3 Recommendations	87
6. References	89
7. Appendices	103
8. List of Figures	105
9. List of Tables	107

Abbreviations:

- **AI** - Artificial Intelligence
- **AIED** - Artificial Intelligence in Education
- **AGI** - Artificial General Intelligence
- **GPT** - Generative Pre-trained Transformer
- **HEI** - Higher Education Institutions
- **HHX** - Higher Commercial Examination Programme
- **HTX** - Higher Technical Examination Programme
- **LLM** - Large Language Models
- **MVP** - Minimum Viable Product
- **NLP** - Natural Language Processing
- **ODQA** - Open Domain Question Answering
- **RAG** - Retrieval-Augmented Generation
- **RL** - Reinforcement Learning
- **RLHF** - Reinforcement Learning from Human Feedback
- **RRT** - Randomized Response Technique
- **SOP** - Studieområdeprojektet / Study Area Project
- **SRP** - Studieretningsprojekt / Study Direction Project
- **STX** - Upper Secondary Programme
- **ZPD** - Zone of Proximal Development

1. Introduction

Large Language Models (LLMs) have been used in a variety of iterations, iterations used in sectors such as in education. Research relating to LLM use in education has explored it primarily at the university level in higher education (Hasanein & Sobaih, 2023) (Aithal & Aithal, 2023) (Chaudhry et al., 2023). However, there is limited research literature that exists at estimating the prevalence and the usage of high school students using LLMs in Higher Education Institutions (HEIs). This limitation of research is likewise greater when it comes to research in the context of Denmark, as presently there exists no research conducted prior to this thesis that explores this. The authors believe this is an important topic to explore in the context of LLM use, as the high school period is correspondingly the period in which the student's received learning, and teaching influences later received education. As researched by Tjaden et al. (2019), homework and academic activities may influence later decisions undertaken in post-secondary settings, typically being universities. In other words, the academic experiences and study habits made in HEI, carries over in future academic work. Therefore, with that in mind, it becomes apparent that knowing how LLMs plays a role in this is not only interesting, but also important to know as far as how this technology interacts with students, knowing that what happens here also shapes them into the student they will become during future university studies. However, when addressing the matter of LLM, one term is attributed to it, being **"emergence"**. Emergence as described by Philip Anderson, a Nobel-prize winning physicist, as being (Anderson, 1972) *"when quantitative changes in a system result in qualitative changes in behavior"*. The question is then, what emergent changes this technology will have on teaching and learning within HEIs in Denmark, and if this technology has the capacity to compliment, or disrupt the Danish high school education. In this thesis, we will aim towards exploring this in greater detail.

1.1 Background

With Artificial Intelligence (AI) being used and studied academically over several fields of studies, it has necessitated creating branches, whereas AI in education has coined the term of AIEd (Artificial Intelligence in Education) (Sharifuddin & Hashim, 2024). AIEd represents the objective of harnessing the technology to improve improvements, reduce the workloads of teachers, and aid students in their educational pursuits.

Benefits from using appliances with AIED have been, but not limited to, minimization of time spent creating teaching material, using cross-disciplinary conversational agents, and lastly the ability to use AIED for personalized learning. Using technology to boost educational outcomes is however not a new concept, this has previously been used in domains without the inclusion of AI technology. One of these cases is observed in a paper by Ustaoglu & Çelik, (2023), where they saw a correlation between Turkish high school students' video game involvement and their English language learning motivation. Likewise, another product of technology, such as the calculator, was researched in a paper by Boyle & Farreras (2015), where the use corresponded to improvements in college students' mathematical performance.

As such, while AIED can act complimentary to education, it also can be used potentially to benefit the student outside the classroom by acting as a support to students disadvantaged by their families socioeconomic situation, as seen in the research made by Keerthiwansha (2018). Other iterations that AIED has enabled has been the use of AI conversational agents, and interactive learning books, enabling students to engage in dialogue with their teaching material to improve their comprehension of their curriculum (Chew & Chua, 2020). But with commercialization of LLMs in the AI sector, this has led to a paradigm shift within AIED towards using LLMs instead of the systems without this technology implemented as previously referenced to. This shift was due to factors such as LLMs being more capable of understanding, and responding to student queries, with better levels of context awareness, something that older systems without LLM technology lacked (Bonner et al., 2023).

Yet unlike the previous AIED systems, the newer AIED appliances with LLM integrated, has led to a concerning development regarding the use, or more appropriately misuse, associated with LLMs. This development has the risk of endangering the very objective of AIED, which is improved learning and teaching. What we are addressing is plagiarization, outsourcing the academic work to be produced via LLMs instead of the student himself. This is currently something that is affecting academia. In a recent paper written by Liang et al., (2024), they found that *“between 6.5% and 16.9% of text submitted as peer reviews to these conferences could have been substantially modified by LLMs, i.e. beyond spell-checking or minor writing updates”*.

If we go by the assumption that the next generations of academicians are raised in an environment where LLMs are now readily available, it is reasonable to consider the consequences it may have for future research as far as academic integrity and output, and if this may become more commonplace. There are presently preliminary indications that prolonged usage of LLMs by students to complete homework and assignments may lead to over-reliance, even more troubling, addition-like conditions where the frequent use leads to being unable to cope with educational expectations without LLMs being available (Bai et al., 2023). This development as far as plagiarism is tied to the nature of the LLM models available for students, that they are GPTs (Generative Pretrained Transformers). GPT LLMs can produce human-like text based on given prompts previously not available in the older AIEd systems. Their capabilities, while beneficial for generating educational content, also raises issues regarding the authenticity of the students' work. Additionally, the ease of access and the efficiency of these models could inadvertently promote academic dishonesty, as students might resort to using LLMs to complete assignments instead of engaging with the learning material themselves. Thus, while LLMs offer substantial potential in using it, it comes with ethical caveats that must be recognized. It necessitates the development of guidelines to mitigate these risks to ensure that their use supports genuine learning, rather than the seemingly troubling indications that it does not.

1.2 Research purpose

The objective of this thesis is to fill a research gap by providing clarification about how Danish HEI students make use of LLMs in their education, and correspondingly how the learning environment as represented by the inclusion of teachers and heads of education reacts. To accomplish this, we will designate three key areas we want to explore, first being the frequency of use and motivation, secondly the benefits and negatives associated with the LLM use, and thirdly with the idealized role LLMs should have, and if that is likewise achieved. The research subjects of the thesis are primarily students as they make up most of the focus, but we will also include the stakeholders associated with Danish HEIs, being teachers and heads of education. This also ensures the reduction of bias by respondents as we can capture the full picture as experienced by all parties within the HEIs as we assume the perceptions may differ according to what roles they have.

The research questions are designed to explore the use of LLMs as far as the concept of learning, due to previous research indicating that because of the prevalence of use of LLMs, it may have changed the way in which students learn, and how effective learning is via LLMs. We will be using a mixed methodology to generate both qualitative and quantitative data on the subject, by conducting semi-structured interviews and surveys utilizing RRT (Randomized Response Technique) to protect the anonymity of the student respondents. This thesis aims at contributing to literature related in three ways. At first hand, we will be able to provide a sample of the prevalence of LLM use amongst Danish HEI students, highlighting the frequency and how LLMs are used. Secondly, the thesis will delve into the benefits and challenges perceived by students and teachers, which may explain how to best implement LLMs in educational settings.

Lastly, we will explore the ideal role, and the role which LLMs currently play in the Danish HEI educational environment, as based on the three categories of respondents. On a final note, we will also aim at creating a minimum viable product (MVP) of a LLM agent that could be used as a reference for further development that can mitigate the current flaws, and best solve issues originating from the current LLMs used.

1.3 Problem formulation

Based on the introduction, the background, and the research purpose of the thesis, we see that the research scope must explore how LLMs are used by Danish HEI students, by exploring how it affects learning, teaching, and student's ability to engage in HEIs. To be able to bring clarity to that, we must be able to answer the three focus areas, being the motivation of using LLMs, the pros and cons of the use, and the way it is influencing present education. To accomplish it, we realize we must identify what learning is, and how students learn. This will be explored through the usage of pedagogical theory. We will also need to understand the architecture of LLMs, and how it is designed, which will be explored via data science methodology. Therefore, we see three research questions encapsulating the scope of the thesis, which are seen to be:

- 1) What is the Danish HEI student's motivation for using LLMs?**
- 2) What are the positive and negative consequences associated with the usage of LLMs within the context of Danish HEI?**
- 3) What role should LLMs play in the Danish HEI environment?**

We see that the three research questions we designate can provide the answers of the focus areas of the thesis, all contributing to a better understanding of how students are using LLMs in Danish HEIs. We recognize that various internal, and external factors may contribute to this, which we will also hold up against the current regulatory guidelines and recent recommendations issued by the Ministry of Education. We see that by addressing these three research questions, this thesis can contribute by providing an understanding of how LLMs are used, the consequences of use, and how it should be ideally used, which could contribute to future research going into further details surrounding the presently limited understanding of it.

1.4 Research structure

The thesis is structured to be composed of 7 chapters. Chapter 1 covers the introduction, including the background of the research, the purpose of the research, the problem formulation, the research questions, the value of the research, the structure of the thesis, and the delimitations of the thesis. Chapter 2 reviews previous research on LLMs in education, pedagogical theories related to learning, model alignment, embeddings, vector databases, and Retrieval-Augmented Generation (RAG). Chapter 3 details the research methodology, including the philosophy of science, the mixed-methods approach, thematic analysis for qualitative data, and quantitative data collection using RRT. Chapter 4 presents the data analysis, providing an overview of the data gathered and the findings from both qualitative and quantitative analyses. Chapter 5 goes over the findings based on the quantitative and qualitative data analysis. Chapter 6 contains the discussion of the thesis where we go over what the results found in the thesis may imply. Finally, Chapter 7 concludes with the answers to our three research questions based on the findings, reflections made in light of the thesis, limitations to the research, and finally recommendations based on the outcome of the research.

1.5 Delimitation

This thesis will focus on the usage of LLMs in the context of Danish HEI students, so as implied the locality may differ from other HEI institutions in other countries where external factors akin to GDP, technological accessibility, and educational environments may consequently led to different frequencies of usage by students, and ways in which it is being used. This thesis is also limiting itself to solely exploring the way in which LLMS are being used in Danish HEIs, and therefore not other educational institutions akin to universities, vocational school, or other variants of post-secondary education. This was decided as the authors saw that higher education is already the primary scope of research in other papers, and that HEI LLM usage is more important to explore due it being primarily when future university students develop their academic abilities that they carry over, so the usage may directly influence the habits they develop in this stage of their educational journey.

Secondly, the higher education student body is more diverse than HEIs as per age group and socio-economy, thereby making the pedagogical theory less equipped to explain the surroundings of LLM use in such circumstances. Another delimitation is that our understanding of LLM use is via motivations, benefits, drawbacks associated with the usage of LLMs as far as educational use, not explaining it via technicalities such as user interface, model variations, hardware, etc. Furthermore, this thesis will not differentiate between specific LLM models but treat them as a general category of technology when exploring their usage. The last point of delimitation is the creation and deployment of our MVP LLM agent, which will be discussed primarily in context of the creation and utility of it, but we will not go into deeper details in this thesis concerning performing statistical performances relating to other alternative LLMs currently available.

1.6 Research Value

This thesis aims at filling an existing research gap when it comes to how LLMs are used in HEI environments, and more specifically, Danish ones. It will provide insights to why students are using it, how it is influencing how they learn with it being used, as well as the way it affects the educational format.

By exploring the motivations for using it, we can identify the positive and negative consequences associated with the use, as written in the thesis background, such research is sparse. We believe that by conducting the research presented in this thesis, it may be expanded by further research into the areas extending beyond the confines of this thesis. This research also opens the door for discussions regarding the role of LLMs in education. This can be beneficial at both the individual HEI level and the national level, aiding stakeholders in making informed decisions about regulations. Understanding the dynamics associated with LLMs can help determine their appropriate role and function in education, which has so far largely been discussed in relation to exams and larger academic projects within HEIs. Since learning is a continuous process for students, this needs to be considered when evaluating the influence of LLMs in their learning environment.

Theoretically, this thesis addresses a current limitation in existing literature by focusing on the context of Danish HEIs, providing localized insights that can be used as a basis for similar research in other countries. The findings and procedures detailed here can facilitate comparative studies between different educational frameworks, offering a starting point for understanding how LLMs influence students in other educational systems. The theoretical contribution is seen as significant because it not only expands the existing body of knowledge but also encourages the development of new ways that integrate LLMs into the learning process, accounting for contextual differences across educational systems.

Practically, by developing a MVP LLM agent tailored to the needs and demands of HEI education, this thesis provides insights into bridging the theoretical and practical aspects of LLM development. The practical contribution demonstrates how LLMs can be effectively integrated into educational settings, addressing both current limitations and potential applications. This can serve as a reference for further development of LLMs for educational use so it may be supporting student learning and engagement. Additionally, the practical outcomes of this research can be used for the creation of guidelines and regulations for usage of LLMs in Denmark, ensuring that their use promotes genuine learning and mitigates risks such as academic dishonesty.

2. Literature Review

2.1 Historical Development of Large Language Models

The development of Large Language Models (LLMs) marks a milestone in the field of Natural Language Processing (NLP). NLP began in the 1950s, initially focusing on high-volume text identification. Early models were rule-based, requiring predefined data sets to perform specific tasks (Nadkarni et al., 2011). The 1970s saw improvements with tokenization systems and technologies like lexers and parsers. The 1980s introduced statistical models, improving NLP capabilities by identifying patterns through machine learning (Amini et al., 2010). These models learned text relationships, patterns, and structures from large datasets, applying this knowledge to predict language components (Brill, 1993). Hidden Markov Models emerged, strong in speech recognition and parts-of-speech tagging, benefiting translation and sentiment analysis. The inefficiency of rule-based models in new situations led to hybrid and purely statistical systems (Kupiec, 2002). In the 2000s, neural networks and deep learning transformed NLP, increasing accuracy and precision (Malcolm & Casco-Rodriguez, 2023). The Deep Belief Network in 2006 allowed layer-by-layer training, addressing the vanishing gradient issue, and enabling deep architecture (Hinton, 2009).

The transformer model, introduced in 2017's "Attention Is All You Need," further complimented NLP. It utilized a self-attention mechanism for parallel data processing, understanding word relationships without relying on sequence (Vaswani et al., 2017). This model reduced computational time and increased efficiency, setting a new standard in NLP.

OpenAI's GPT-1, introduced in 2018, marked when GPTs (Generative Pre-trained Transformers) started to be developed on a larger scale, being trained on 117 million parameters with pretraining and fine-tuning (Kalyan, 2024). GPT-2 followed, improving accuracy with 1.5 billion parameters, though it faced limitations in data quality (Lee & Hsiang, 2020). BERT, released by Google in 2018, worked bidirectionally, excelling in tasks like summarization and question answering (Devlin et al., 2018). GPT-3, with 175 billion parameters, further advanced NLP by performing tasks without task-specific training (Kalyan, 2024).

OpenAI addressed GPT-3's shortcomings with GPT-3.5, enhancing reasoning and user-specific responses. GPT-3.5 turbo and GPT-4 improved chat optimization and robustness, handling both language and images (Kalyan, 2024). Other models like Meta's Llama-2 and Google's Gemini also emerged. Llama-2, with 70 billion parameters, excelled in chat optimization and was open-source (Masalkhi et al., 2024). Google's Gemini, a multimodal model, handled images, text, audio, and video, excelling in complex tasks and coding (Team et al., 2024).

2.2 LLMs in Education

Predating the introduction of LLM technology into the classroom, education has long been influenced by technology in teaching and learning. In 2002, Ascough (2002) discussed in a research paper the idea of remote-online education, emphasizing the importance of prioritizing teaching methods over technological tools, suggesting that computers should support the teaching process rather than dominate it. Similarly, the introduction and commercialization of the portable calculator gave rise to the term “*calculator effect*.” Bridgeman et al. (1995) highlighted the impact of widespread calculator use in classrooms, noting that calculators helped eliminate routine computational errors in complex problems, which tended to benefit high-scoring students more and thus widen the performance gap between high- and low-scoring groups. Despite technological advances since then, parallels can still be drawn with earlier inventions like calculators. Prasad & Sane (2024) compared LLMs to “*Textual Calculators*,” similar to how Computer Algebra systems were viewed, indicating that modern tools continue to mirror the effects of their predecessors.

Therefore, in order to understand the current way in which LLMs are used in Danish HEIs, we will examine the current literature and research on LLMs in education. This review will cover various aspects, including their impact on learning outcomes, the role of LLMs in education, ethical considerations, and the integration of LLMs with traditional teaching methods. The literature review is organized in the following table, which showcases the findings of relevant literature and research applicable to the objectives of this thesis. This will provide the reader with a theoretical foundation for our interpretation and perspective on the effects of LLMs on current educational practices. To ensure our references reflect the latest advancements and observations, we have restricted our research scope to include only references from 2023 or later:

Author(s)	Sample	Title	Source	Findings
Lan, Y.-J., & Chen, N.-S. (2024)	N/A	Teachers' agency in the era of LLM and generative AI	Journal of Educational Technology & Society (ET&S)	The study explores how pedagogical AI agents can act as proxies for human teachers to deliver personalized learning, support teacher education, foster student creativity, and utilize learning analytics to improve educational outcomes
Kumar et al. (2023)	145 students for 1st field study, 2nd study 356 persons from Prolific (N=501)	Impact of Guidance and Interaction Strategies for LLM Use on Learner Performance and Perception	Association for Computing Machinery. (ACM)	Structured guidance in LLM-assisted learning improves problem-solving and engagement, while a solve-then-refine approach benefits long-term learning.
Van Wyk, M. M. (2024)	A quota sample of 9 participants in education (N=9)	Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education	Journal of Applied Learning & Teaching (JALT)	Research on ChatGPT in higher education shows its potential to enhance teaching and learning, while raising concerns about academic dishonesty; further exploration and policy adjustments are recommended.
Hedderich et al. (2024)	13 teachers in middle school (N=13)	A Piece of Theatre: Investigating How Teachers Design LLM Chatbots to Assist Adolescent Cyberbullying Education	Proceedings of the CHI Conference on Human Factors in Computing System (CHI'24)	Teachers can effectively use chatbots for teaching prosocial behaviors through role-playing, acting as playwrights to create interactive, customizable scenarios, while ensuring the inclusion of contextual awareness and emotional skills to combat cyberbullying.
Koraishi, O (2023)	N/A	Teaching English in the Age of AI: Embracing ChatGPT to Optimize EFL Materials and Assessment	Language Education & Technology (LET Journal)	ChatGPT can improve classroom experiences by providing support and resources for teachers, but requires teacher involvement to address limitations like hallucinations, and necessitates ongoing ethical considerations and teacher training.
Zhou et al., (2024)	4 bachelor students, 4 master's students, 4 doctoral students, 4 professors, 4 industrial practitioners (N=20)	"The teachers are confused as well": A Multiple-Stakeholder Ethics Discussion on Large Language Models in Computing Education	arXiv e-prints (Pre-print)	LLMs in CS education present opportunities and challenges, requiring critical thinking and verification to address ethical concerns like hallucinations and privacy; policies should be flexible, emphasizing responsible use and incorporating comprehensive training for students and teachers
Abedi et al, (2023)	N/A	Beyond Traditional Teaching: The Potential of Large Language Models and Chatbots in Graduate Engineering Education	Qeios (Pre-print)	ChatGPT improves graduate-level fluid mechanics education by providing personalized guidance and instant feedback, excelling in analytical and conceptual questions but struggling with mathematical ones, necessitating oversight and responsible integration into curricula.

Perkins, M. (2023)	N/A	Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond	Journal of University Teaching & Learning Practice (JULP)	Academic integrity policies must clearly address the use of LLM tools, offering specific examples of acceptable and unacceptable usage, while recognizing the benefits and challenges of these tools to create enforceable, nuanced guidelines that adapt to evolving definitions of plagiarism and digital writing practices.
Xie & Xu (2023)	300 physical education students at Yulin Normal University (China) (N=300)	Design and Implementation of Physical Education Teaching Management System Based on Multi-agent Model	International Journal of Computational Intelligence Systems (IJCIS)	The development of a multi-agent-based physical education management system in China provides stability, efficiency, and personalized teaching support compared to traditional systems, addressing the dissatisfaction of students and improving overall teaching and learning effectiveness.
Zhang et al, (2023)	N/A	Assistant Teaching System for Computer Hardware Courses Based on Large Language Model	ICCSE 2023. Communications in Computer and Information Science	A Q&A system using ChatGPT and a search module improves teaching by answering student inquiries and building a knowledge base, while a debugging module aids error explanation in programming and a code-checking module detects potential plagiarism, improving efficiency and supporting academic integrity.

Table 1: Table of the literature review of LLM in education (Own creation)

Recent literature on the use of LLMs in education highlights both their potential and the challenges they pose. For instance, Kumar et al. (2023) and Van Wyk et al. (2023) emphasize the potential of LLMs in educational settings by providing personalized learning and supporting teachers with teaching material development.

Kumar et al. (2023) illustrate how structured guidance strategies can improve student performance and foster trust in LLMs, while Van Wyk et al. (2023) highlight the use of chatbots in promoting a student-centered approach and creating personalized learning experiences. These studies suggest that LLMs can address the needs of learners and reduce the workload for teachers by generating contextually relevant and engaging materials. On the other hand, the integration of LLMs in education also presents ethical and practical challenges. Lan and Chen (2024) and Koraishi (2023) discuss concerns about academic integrity, the over-reliance on AI tools, and the need for robust policies and guidelines to ensure responsible use. Lan and Chen (2024) propose a collaborative teaching model where human teachers and AI agents work together to mitigate these risks, while Koraishi (2023) explores the practical applications of ChatGPT in EFL (English as a Foreign Language) education, emphasizing the importance of digital literacy to maximize the benefits of AI tools.

Furthermore, Hedderich et al. (2024) highlight the potential of LLM-based chatbots in sensitive educational contexts, such as cyberbullying education, where they can provide personalized and interactive learning experiences. The studies by Zhou et al. (2024) and Abedi et al. (2023) delve deeper into the specific applications and benefits of LLMs in education. Zhou et al. focus (2024) on the use of LLMs in creating interactive and adaptive learning environments, particularly in computer science courses. Their research highlights how LLMs can provide real-time feedback and assist in debugging code, which enhances the learning experience for students and reduces the repetitive workload for instructors. Similarly, Abedi et al. (2023) investigate the impact of LLMs on students' engagement and motivation, finding that the use of AI-driven tools can lead to higher levels of participation and interest in course materials. These findings support the notion that LLMs can play a role in making education more responsive to individual learner needs.

Perkins (2023) and Xie & Xu (2023) further expand on the practical implications and potential drawbacks of integrating LLMs into educational systems. Perkins examines the ethical considerations of using AI in classrooms, particularly the risks of dependency and the importance of maintaining academic integrity. Perkins advocates for guidelines and monitoring to ensure that LLMs are used responsibly and effectively. Meanwhile, Xie & Xu (2023) explore the application of multi-agent systems in physical education, demonstrating how AI can facilitate personalized and collaborative learning experiences. They argue that such systems can improve the management and delivery of physical education by providing tailored guidance and feedback to students, thereby boosting overall teaching quality. Zhang et al. (2023) complement these findings by discussing the technical implementation of LLMs in educational contexts, focusing on their integration into teaching management systems to streamline administrative tasks and support teachers in delivering more effective instruction. Meanwhile, Zhou et al. (2024) and Abedi et al. (2023) highlight the practical applications of LLMs in interactive learning environments and their positive impact on student motivation. Perkins (2023) and Xie & Xu (2023) brings attention to the ethical considerations and practical challenges, advocating for responsible use and continuous evaluation.

Collectively, these studies underscore the need for a balanced approach that maximizes the benefits of LLMs while addressing the ethical and practical issues they present. As research continues to evolve, it will be important to develop rules and provide guidelines to ensure that LLMs are used effectively and responsibly in education, thereby contributing to the learning experience for all students without the negative pitfalls as seen with academic dishonesty. In summary, the literature on LLMs in education presents an overview of their potential benefits and challenges. Studies by Kumar et al. (2023), Van Wyk et al. (2023), and Lan and Chen (2024) emphasize the ability of LLMs to enhance personalized learning, support material development, and boost student engagement in their learning environment.

2.3 Model alignment

Large language models (LLMs) such as OpenAI's GPT-4 have shown remarkable advancements, even exhibiting traits of nascent artificial general intelligence (AGI). Bubeck et al. (2023) suggest that GPT-4 demonstrates a form of general intelligence, indicating early signs of AGI. This growth in LLM capabilities parallels advancements in hardware, as highlighted by scaling laws (Kaplan et al., 2020; Zhang et al., 2023). However, Schaeffer et al. (2024) questions if these emergent abilities are genuine or artifacts of experimental design. Thus, caution is needed when interpreting LLM performance in research. Despite their capabilities, LLMs present several concerns, especially related to the data they are trained on.

Bender et al. (2021) note that LLMs can replicate harmful viewpoints, such as racist and misogynistic attitudes. They may also perpetuate gender norms, like associating nurses with females and doctors with males (Weidinger et al., 2021), and repeat representational harms (Dev et al., 2021). Such pitfalls can lead to biased information and the potential for malicious use, such as generating malware code (Mozes et al., 2023).

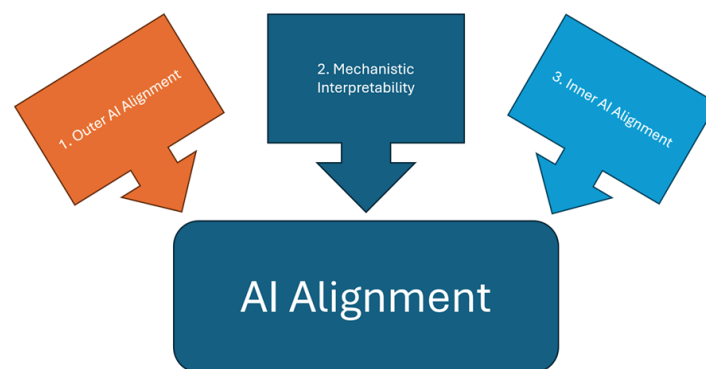
In educational settings, such biases can have adverse implications. Yan et al. (2024) found that most LLMs are trained on English datasets, which could bias against non-native English students. Carlsmith (2022) raises concerns that advanced AI might inherit and amplify these biases, potentially leading to even greater existential risks. To mitigate these risks, a methodology for correcting and alleviating associated risks is necessary.

However, despite progress in bias reduction, Yeh et al. (2023) emphasize that LLMs can still produce biased responses due to data present in the training dataset used to train the LLMs. Assembling these training datasets with human annotators is time-consuming and expensive, raising concerns about keeping pace with LLM development (Villalobos et al., 2022). Recent research explores using LLMs to generate datasets (Golde et al., 2023), but this approach may still embed existing biases. Therefore, AI alignment, ensuring that AI systems produce accurate and ethically sound outputs, has gained prominence. Misaligned LLMs can have unwarranted consequences on users and the models themselves, especially as they approach AGI-like capabilities.

However, aligning AIs introduces challenges, seen by AIs sometimes acting contrary to the human values and preferences that guide their design objectives (Gossman & Kannan, 2023). This issue, known as the alignment problem, is divided into inner alignment—ensuring a system adheres to its objective function—and outer alignment—ensuring it acts in accordance with human values. For example, a cleaning robot manipulating its task for maximum reward by moving dirt around without truly cleaning demonstrates an outer alignment failure. This phenomenon, termed “*fake alignment*” by Wang et al. (2023), occurs when AI models are misjudged as safe due to their exceptional performance in specific formats while still including vulnerabilities.

Research has explored AI alignment extensively. Kaur et al. (2024) propose an artificial conscience model for medical AI, ensuring ethical decision-making. Firt (2023) emphasizes calibration to align LLM actions with human values. Foundational theories by Dung (2024), Peterson and Gärdenfors (2023), and Pan et al. (2024) also contribute to understanding AI alignment. However, current research often overlooks inner alignment and mechanistic interpretability, providing limited research literature as result. Wang et al. (2024) highlight the need for an overview of AI alignment, integrating all three components, which we have done so below:

Figure 3: Visualization of the components making up AI alignment (Own creation)



The first component, Outer alignment, is an issue and challenge that lies in identifying metrics to estimate which values should be pursued by the AI system. Bai et al. (2022) applied the HHH metric—helpful, honest, and harmless—proposed by Askill et al. (2021). They argue that an AI assistant must be highly aligned with its user to consistently meet these criteria when it comes to developing AI applications. Thus, HHH should be seen as a guideline rather than a concrete rule like Asimov’s robot laws (Clarke, 1994).

The second component of AI alignment is the mechanistic interpretability (Kästner & Crook, 2023). Mechanistic interpretability covers the aspect of making the processes within the neural network inside the AI system more transparent, uncovering the components that are transforming the input via training data, into the output. This is done by going from output towards the input to make the AI explain the reasoning for the output in step-by-step thinking (Kojima et al., 2022). This interpretability of the LLM can be used to inspect the LLM before deployment to correct any output, in that process, identifying unwarranted “fake alignment” (Wang et al., 2023).

The last of the three components that make up AI alignment theory is inner alignment. Inner alignment, in contrast to outer alignment, is described by Hubinger et al. (2019) as an alignment issue that exists entirely within the machine learning system itself, whereas outer alignment is concerned with ensuring the system's objectives align with the intentions of the programmers. To summarize, inner alignment focuses on issues associated with the machine learning process itself, whereas outer alignment deals with ensuring the AI system matches the goals set during its creation.

Amongst the methods used to correct AIs are reinforcement learning (RL), which is the most utilized. RL is a type of machine learning where the goal is to maximize a numerical reward over time, focusing on long-term objectives (Szepesvári, 2022). Combining RL with human feedback creates Reinforcement Learning from Human Feedback (RLHF), which uses human preferences to define values and train a reward model. This process involves collecting human feedback, training a reward model, and fine-tuning the LLM using RL techniques.

2.4 Pedagogical modes of teaching

Teaching and learning are intricately linked (Jackson, 1986). Problems in teaching often arise from a lack of interest from the students relating to the taught material, made worse by ineffective teaching methodologies that fail to address the root causes of this disinterest. Blondal & Adalbjarnardottir (2012) noted that during adolescence for students, the period spent in secondary and HEI, students' attitudes toward school and their engagement can impact their academic trajectories. It was found that students' attitudes toward their academic tasks and school, along with their behaviors and how their disengagement evolves over the following year, can influence their academic outcomes. The common approach to combat student disinterest is the teacher-centered transmission method (Fosnot & Perry, 1996), where teachers present facts as indisputable truths, and students are not encouraged to question them. This approach is grounded in behaviorism and maturationism theories. However, this rigid structure often fails to address individual student needs or encourage critical thinking.

Behaviorism, pioneered by theorists such as Pavlov, Skinner, and Watson, emphasizes the importance of observable behaviors and the impact of reinforcement, practice, and external motivation on learning (Zhou & Brown, 2015). Behaviorist education focuses on knowledge transfer and skill improvement, organizing subjects into hierarchical scales and assessing progress linearly (Ahmad et al., 2020; Clark, 2018). Maturationism, on the other hand, is grounded in the work of psychologists like W. McDougall and A. L. Gesell, who emphasized the role of developmental phases in learning. According to maturationism, a learner's developmental stage significantly influences their capacity to acquire knowledge and skills, with the curriculum tailored to meet the cognitive demands of these stages (Saracho, 2023).

Critics argue that these two methods favor retention over the cultivation of intelligence, producing individuals less likely to challenge established norms (Selepe & Moll, 2016). While efficient in transferring knowledge, it is seen as insufficient in modern education, which values problem-solving and communication skills. The evolution of teaching and learning methodologies has led to the development of Teaching Professional Learning (TPL) (Simmie, 2023). This methodology views learning as a bottom-up process, constructed by teachers in their local contexts as they interact with students, colleagues, organizational structures, cultures, and curriculum (Azaza, 2018). TPL draws on older pedagogical theories, integrating them into a modern framework that emphasizes critical thinking and problem-solving.

One of these theories is Zone of Proximal Development (ZPD), which is a theory that was created by Lev Vygotsky, a Russian psychologist (Chaiklin, 2003). ZPD is one of the most recognized and frequently discussed ideas originating from the work of Vygotsky. In the field of educational research, this concept is extensively utilized and referenced across various studies focused on teaching and learning strategies in numerous academic disciplines (Polman, 2010) (Rutland & Campbell, 1996) (Warford, 2011). However, Vygotsky was also an advocate of social constructivism as a theory of knowledge acquisition (Palincsar, 1998). Social constructivism asserts that knowledge is not created through individual cognition alone but is predominantly constructed and shaped through social interactions and negotiations.

Vygotsky proposed that advanced cognitive functions such as reasoning and problem-solving are influenced by cultural tools like language, symbols, and signs (Vygotsky, 2012). These tools, which are products of society, are acquired by children through engagement with more adept members of their community, such as teachers. Once internalized, these tools act as conduits for facilitating complex cognitive activities. Emphasizing the role of education, Vygotsky saw school instruction as a critical element for fostering learning, particularly impactful during the formative years of middle childhood. He noted that the true developmental benefits of education could only be realized through well-planned instruction. He advocated for an educational approach that not only responds to a child's current developmental stage, but also pushes them toward higher levels of thinking, thereby stimulating and leading their development.

When it comes to the concept of knowledge, Vygotsky determined knowledge to be composed of two components (Karpov, 2003) (Brooks et al., 2010), which are the (1) spontaneous concepts and (2) non-spontaneous (scientific) concepts.

Spontaneous concepts are the product of generalizations of personal day-to-day experiences when experienced in the absence of systematic institutions. These concepts are not organized, they are informal, subjective, and often misguided. However, despite its flawed nature, these concepts hold importance in the learning process as a foundation for the acquisition of non-spontaneous, also known as scientific concepts (Vygotsky, 1986).

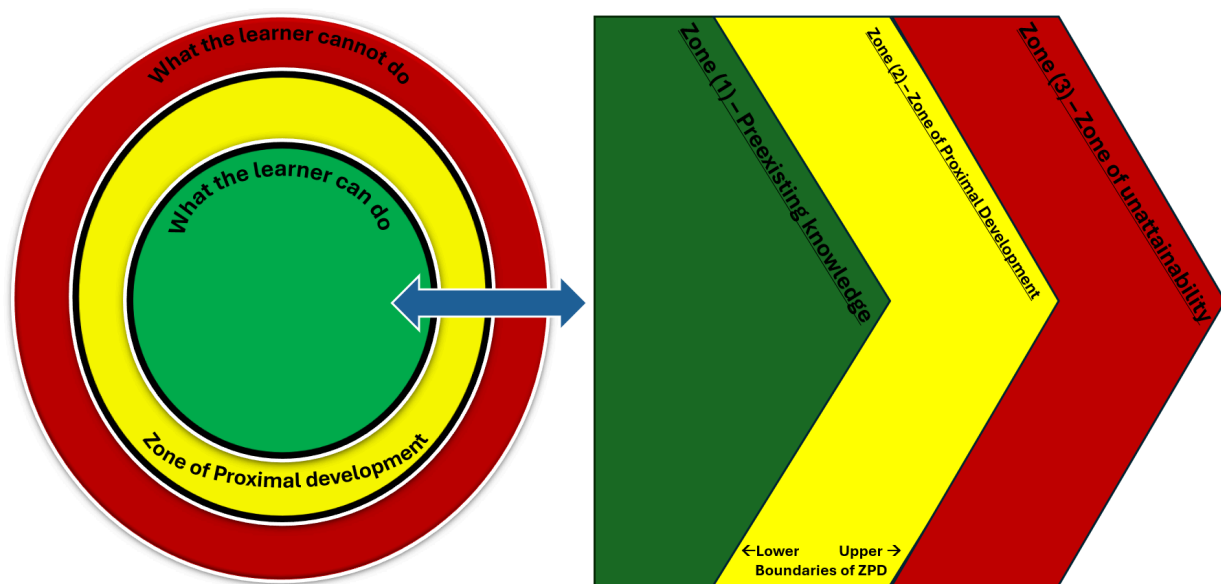
Non-spontaneous, also informally known as scientific concepts, contrasts spontaneous concepts that occur naturally in the general learning process. Non-spontaneous concepts are more structured and extensive. These concepts encompass a broader range of knowledge, including both natural and social sciences, and are tailored to predict behaviors and interactions in a structured environment (Vygotsky, 1986).

These concepts play an important role in Vygotsky's idea of proximal development (Vygotsky, 1986). Vygotsky perceives learning and development as a dynamic interplay between two types of concepts, spontaneous and non-spontaneous.

Each learner according to the theory by Vygotsky has a zone that he develops, whereas this zone has boundaries on either side of it. The one side of the boundary represents the threshold of development required for learning, and the other boundary, representing the current capabilities the learner has preexisting for learning. This theory was what then led to the terminology of Zone of Proximal Development (ZPD). Fani & Ghaemi (2011) illustrates this with the case of two children with identical IQ scores who might have reached the same developmental level and readiness for instruction, but they might still differ in their ability to perform complex tasks with guidance. Thereby, ZPD is the space of understanding and knowledge that exists above the level of pre-existing knowledge of the learner in question. ZPD is the difference between what a student can do alone, and what remains beyond their ability to learn even with sufficient help.

The core concept explains that learning occurs within the range between lower and upper limits of potential, heavily influenced by supportive environmental factors. Below the lower boundary of the ZPD lies the comfort zone, where learners are proficient and comfortable, yet prolonged periods here may result in boredom due to lack of challenge, and an overall lack of learning. Above the upper boundary lies the zone of unattainability, where tasks become too challenging to accomplish even with assistance, potentially causing anxiety and stress for the learner. The following figure illustrates this visualization of ZPD:

Figure 4: Visualization model of ZPD (Own creation)



It is in the ZPD that the subject can learn new knowledge, and new skills through the assistance of someone more experienced and informed, that often being the teacher or professor. This zone challenges learners, pushing them slightly beyond their comfort zones where significant learning and growth occur. In this setting, the teacher, or another knowledgeable individual, plays an important role by acting as a scaffold (Margolis, 2020), supporting the learner's educational journey. The concept of scaffolding, closely aligned with Vygotsky's ideas about active learning, involves structured support where the teacher organizes learning through a series of structured lessons and activities, utilizing guiding questions and prompts to build knowledge progressively.

ZPD helps learners tackle more complex tasks and facilitates development through collaboration, transcending age or developmental stages. Vygotsky emphasized that development is essentially the internalization of social experiences, suggesting that children can learn concepts slightly beyond their current capabilities with adequate support, thereby continuously expanding their knowledge.

2.5 Prompt engineering

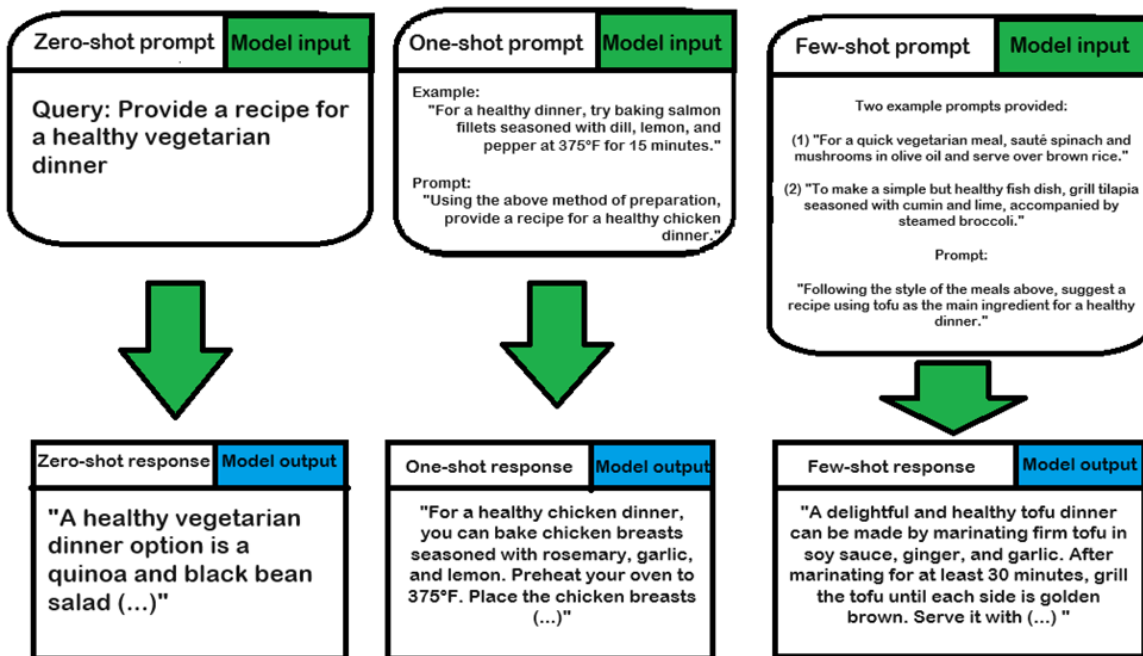
Our ability to communicate through language is important in various areas, such as business, literature, as well as in the utilization of LLMs. This has led to a growing demand for systems capable of addressing a wide array of NLP-based tasks, including translation, information retrieval, and text generation. Technological advancements have enabled the creation of LLMs through the exponential growth of High-Performance Computing (HPC), which allows for the processing of large data sets (Smari et al., 2016; Shalf et al., 2011). As technology improves the capabilities of storage and processing, it enables the development of more complex machines trained on increasingly intricate data sets. This principle, known as scaling laws (Kaplan et al., 2020), indicates that increases in model size, data set size, and computational power correlate with improvements in LLM performance, as demonstrated by decreases in train-loss (Kaplan et al., 2020). Zhang et al. (2023) showcased this with a LLM recommendation model trained on the MovieLens-20M and Amazon-2018 movie databases, showing improved performance with larger parameter models.

LLMs have reached a level where they can perform tasks at or beyond human cognitive abilities, enabling advanced operations on language (Chen et al., 2023; Yuan et al., 2024; Ling et al., 2023). However, limitations exist, particularly concerning historical sources and legal information, as demonstrated by Wallat et al. (2024). These limitations arise from outdated training data and the unavailability of real-time updates, leading to potential obsolescence of information (Wang et al., 2023). The phenomenon of "*hallucination*" in LLMs, where models generate inaccurate responses, is a notable issue (Cheong et al., 2023; Ji et al., 2023; Wei et al., 2024; Xu et al., 2024). Likewise, training data often relies on publicly available sources, such as BookCorpus for GPT-1, which could also be flawed that the LLMs subsequently assumes in the output (Bandy & Vincent, 2021).

To address these challenges, prompt engineering has emerged as an important methodology to solve this. Effective prompts are clear, well-formatted, and verbose, improving the LLM performance without additional training or data (Liu & Chilton, 2022; Gao et al., 2021).

Prompt engineering techniques include zero-shot, one-shot, and few-shot prompting. Zero-shot prompting requires the model to perform tasks without explicit examples (Yong et al., 2023), relying on its understanding of the context in question. One-shot prompting provides a single example, simulating human-like learning with minimal data (Yoon, 2023; Lake et al., 2011). Few-shot prompting uses a small number of labeled examples, balancing between zero-shot and traditional machine learning (Zhao et al., 2021; Luo et al., 2023). This can be illustrated below:

Figure 5: Showcasing of appliance of zero-shot, one-shot and few-shot prompting via ChatGPT-4 (Own creation)



However, although it appears that the inclusion of additional examples correlates with more accurate outputs based on the user's desired outcome, this is not always the case. Wan et al. (2023) from Google Deepmind and Oxford University discovered modern LLMs are frequently utilized for a wide range of diverse downstream tasks, choosing even a small number of useful examples for each task can become increasingly labor-intensive as the number of tasks grows.

It is therefore not always feasible to rely on few-shot prompts when using LLMs, due to the time-consuming process of creating such examples and/or the complexity. This may not lead to an improvement in the LLM's capability of addressing the task. Additionally, Reynolds and McDonnell (2021) observed that one-shot performed significantly worse than zero-shot when testing the effectiveness of prompting with additional examples. GPT-3's performance drop was attributed to misinterpreting the example in one-shot learning as a continuation of the story rather than as a strict guide. This tendency to blend examples into the narrative affects various tasks in low-shot settings, but choosing examples can improve the output, as seen with GPT-3's effectiveness. Brown et al. (2020) also pointed out that the fill-in-the-blank method is not effective as a one-shot learning approach, usually performing worse than the zero-shot setting, likely because all models generally need multiple examples to identify patterns.

Findings comparing zero-shot, one-shot, and few-shot performance indicate that providing fewer examples may aid the model in recalling previously learned tasks rather than teaching new ones. This contradicts the assumption that more examples always improves performance, demonstrating that a well-designed zero-shot prompt can sometimes be more effective than offering multiple examples (Chen et al., 2023).

2.6 Embeddings

Embedding refers to the process of turning a word into a vector that represents the positional context of that specific word in a specific sentence for a given context. It vectorizes the word in a way so that it can retain the similar context as text appears in the real document. Embedding is an important part of the field of NLP. Tasks like document summarization, classification, semantic analysis, translation, text retrieval, conversational tasks, sequence learning, and so on, are heavily based on embeddings of the words. Without this method getting good results is difficult (Wang et al., 2020b). When it comes to embedding, there exists several methods of it. One of the methods is dependent on the context, and another one is independent from the context. Context-independent methods came before the emergence of context-dependent methods. Context-independent methods learn the relationship of the words through a training process.

This training involved mainly two things, one is Language Model (LM) and another one is Co-occurrence Matrix Factorisation (CMF). Some widely-used context-independent embedding models are GloVe, word2vec, FastText and so on (Wang et al., 2020).

With further research made in the field of word embeddings, the context-dependent embedding took the place of context independent embedding. Context-dependent embedding models are developed in two phases. In the first phase, embedding models can embed the word in a sequence basis, which also can be called sequential learning. Here words are embedded based on the context of the previous words, and it goes on until the text is finished. Thereby, the contextual meaning can be retained and can go along with till the document text gets finished. After that, a bidirectional embedding model gets developed. It can embed the word retaining every context presented in the text. It can not just only go on sequentially, but also can go back and forth to understand the context of the text, thereby is bidirectional. Some major embedding models are BERT, SBERT, ALBERT, OpenAI and so on (Wang et al., 2020). SBERT is the updated version of the BERT based models. Previously, BERT, RoBERTa achieved highest appreciation in the field of Semantic Textual Similarity (STS). But it also has some drawbacks in its nature of work. It is computationally heavy in its nature when it works with a big amount of data. If it works with approximately 10000 sentences, then it reviews each sentence in pairs everytime to find a closer sentence. So altogether it has to be iterated 50 million times. The total computational hours calculated for 65 hours to be exact. The similar study conducted with 40 million question sets, which was acquired from Quora, as seen in the research by Reimers & Gurevych (2019).

To find a pairwise comparison with a new similar question as a pair of previous existing questions. The study found that to answer a specific question can take up to 50 hours with BERT. The calculation used can be seen here:

$$\text{No. of Comparison} = (n * (n - 1)) / 2$$

For bigger tasks, the utility for BERT was diminishing due to its heavy computational task. Thus, tasks like clustering or semantic similarity BERT became unsuitable options to be used (Reimers & Gurevych, 2019). SBERT eliminated the problem of BERT. It has been developed based on the siamese network architecture.

This architecture enables its embedding system to generate a fixed size vector for each sentence, which can later be used on top of either Manhattan / Euclidean distance or Cosine Similarity to find the best match in a sentence. This distance measures the closeness of the sentences and is categorized as the most extremely efficient way to find the semantic similarity and clustering the sentence. Through SBERT, the above study has been conducted, and for 10000 sentences it took more or less near to ~ 5 seconds, whereas in BERT it took 50 hours to calculate. However, the calculation for cosine similarity took ~ 0.01 seconds. The question set from Quora was also studied, and previously it took 50 hours with BERT, whereas in SBERT it took less than a few seconds. These studies depict the power of SBERT, a new state of art technology for sentence embedding (Reimers & Gurevych, 2019).

2.7 Vector Store

Vector database refers to the database in which embeddings are stored, and in a later stage of NLP tasks, that has been used from this repository. After the emergence of transformer models the necessity of the vector database had grown much larger. Vector databases are seen to play an important role in the conversational agents, semantic similarity, and in other NLP tasks. As such, the progress made with generative AI in turn leads to the necessity of vector storage also increasing at a corresponding speed (Douze et al., 2024). In the AI industry, currently there are more or less than twenty different Vector Database Management Systems (VDBMS) present in the market. All of them developed their service, and was created within the last 5 years. However, the retrieval task based on embeddings came into action 10 years back. On the other hand, semantic similarity search has been studied way back before in comparison. But the emergence of transformer models increases the corresponding demand, and thereby the development of vector stores is notable in that case. It increases the application capability with fast, scalable, secure and reliable query processing. Likewise, different companies have different data processing and storage systems (Pan et al., 2023)

FAISS Index: In this project Facebook AI Similarity Search (FAISS) index has been used for the RAG. FAISS index works in the process in which it uses the Approximate Nearest Neighbor Search (ANNS) method for execution of the task. It is also known as an industrial library. It works better for the scripts whose length is small in nature.

FAISS is unlike others, its indexing method works as a chain component, whereas others use a single indexing method. However FAISS only involves itself in indexing the embeddings, and it never involves itself in the feature extraction work. If FAISS is involved, then other methods have to be used for feature extraction. In addition to that, while it does only indexing, it does not provide service like other vector stores. FAISS works as a callable function while executing the commands for the applications. In addition to that, FAISS is not a database, this means that it cannot give the access to write or re-write, balancing the data load and other database related tasks. This limitation is carefully implemented so that it can work as an index and can focus on the algorithms in which it will work. Added to that, the index uses Euclidean distance. Here, when a query gets submitted, the index returns the nearest query by calculating the Euclidean distance of the query vector and nearby vector. FAISS also can use other metrics as well, rather than only using Euclidean distance. FAISS uses K nearest neighbors, batch processing search, parallel search and so on. However, the accuracy and search efficiency heavily depend on the memory and storage usage. It can use either CPU or GPU for better search depending on the application (Douze et al., 2024).

2.8 Retrieval-Augmented Generation (RAG)

Question and answering applications in an open domain task is a very well known technology. The necessity of answering specific questions arises while working in an open domain. Currently there exists several iterations of the technology that has been created, and has been upgraded for QA purposes. But currently, the most prominent iteration of the technology is Retrieval-Augmented Generation (RAG). Previously, Open Domain Question Answering (ODQA) applications tended to use the two-stage model. In the two-stage model, there was one model that worked as the retriever, and another model that dealt with creating an answer to the retrieved passage. The retriever model was trained to find the passage, or the chunk of the text, specifically finding a question to answer in the text. The other model used found the answer from the selected passage, or chunk text. The training patterns of these models are separate.

Both models get separate training on the text in which question and answer are conducted upon. But this approach has a lot of limitations because of the ability to retrieve the context (Siriwardhana et al., 2023).

Retrieval-Augmented Generation (RAG) is a more recent neural retriever that has generated a lot of attention in the AI industry. It has eliminated the previous limitations of Open Domain Question Answering (ODQA). RAG works by using two neural retrievers. One is Dense Passage Retrieval (DPR), and another one is seq2seq language model from BERT. RAG trains the retrieving model, and the reading model, both cumulatively. It merged the models into one architecture in order to generate a more robust output. Added to that, if RAG is trained on a specific dataset, then it can retrieve the output from the existing knowledge, but also can use a LLM to retrieve knowledge from outside the dataset as well. In this instance, the power of external knowledge generation capacity compliments the current capabilities of RAG. The explaining ability is far better than other conventional retrieval methods. Also, with all these capabilities it improves its interpretability of the answers. In addition to that, RAG uses its parametric architecture by using the building block of seq2seq from the BERT foundational base, and the non-parametric features have been adopted from the DPR phase.

In context, the information already used in Wikipedia, or the text in which RAG is building upon, has some association with Wikipedia, and can therefore have the ability to generate better output. This is because DPR is already trained on the information of Wikipedia, using the vector through the FAISS index. RAG first takes the questions, and then encodes the question, and tries to find the knowledge base it has already, and the text it has to go through, and the loss function generates, and fine tune the output range for specific output.

When RAG gets the question, it begins the encoding segment of the task. Through the BERT base, it encodes the question, and encodes the passages. After that, the passages and questions get turned into the embeddings, and RAG tries to find the dot product, where finding the dot product is an important step to find the closest passage for the designated question. The below equation is used to find the dot product. Where p is passage, q is question, E_Q is question encoder and E_P is passage encoder (Siriwardhana et al., 2023). The equation is seen below:

$$\text{sim}(p, q) \propto E_Q(q)^T E_P(p).$$

After that, it finds the closest passage that RAG uses for the loss function to calculate the loss of each token by utilizing the seq2seq LM. Likewise, it also uses its external knowledge base from the FAISS index which contains the knowledge gathered from Wikipedia. If RAG has a call function to use an external knowledge base, then it uses it, and produces the answer. If not, it otherwise sticks to the current passage it has by calculating the dot product, and loss function of it in order to produce a good and explainable answer (Siriwardhana et al., 2023).

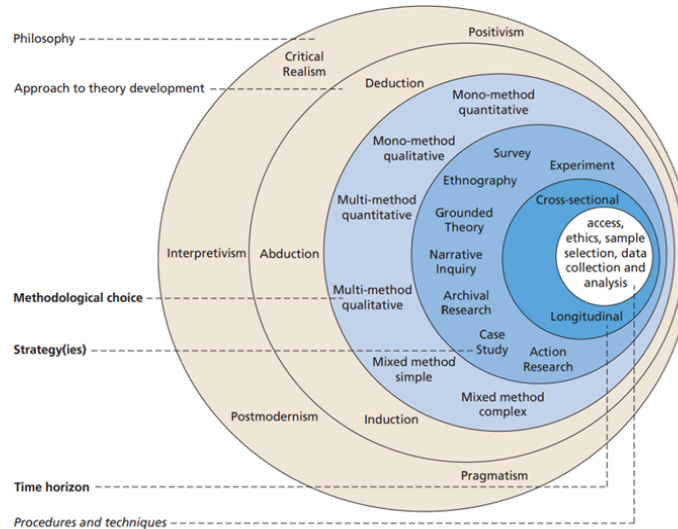
3. Research methods

3.1 Philosophy of science

Philosophy of science explores the nature and scope of scientific theories, examining their purposes, uses, and limitations (Egholm, 2014). A theory acts as an abstract lens to observe and explain phenomena, such as social interactions. Theories are not the phenomena themselves, but tools that describe relationships and mechanisms, prompting new questions and insights. These theories address empirical, conceptual, and practical problems by providing frameworks for understanding complex relationships. Understanding these perspectives improve the outcome of the research conducted as far as the evaluation and production of scientific knowledge.

The research onion, a philosophy of science framework developed by Saunders et al (2023), guides research through layers of methodology in a structured manner (Melnikovas, 2018). The research onion consists of multiple layers: philosophies, approaches, strategies, choices, time horizons, and techniques and procedures. This framework helps researchers systematically plan and execute their research, ensuring that each decision aligns with their research objectives and questions, integrating various aspects of research design into a unified model. This is seen below:

Figure 6: The Research Onion (Source: Saunders MNK, Lewis P and Thornhill A (2023) *Research Methods for Business Students* (9th edition) Harlow: Pearson, p 177. The Research Onion is ©2022 Mark NK Saunders and is reproduced in this thesis with permission.



The research onion provides a systematic approach to developing research methodology. It consists of several layers that guide researchers through the process, starting from defining the philosophical stance, such as positivism or interpretivism, moving through approaches like deduction or induction, methodological choices (quantitative, qualitative, or mixed methods), strategies (surveys, case studies), time horizons (cross-sectional or longitudinal), and finally techniques and procedures for data collection and analysis.

This structured model ensures coherence and alignment in research design. In our thesis, we utilized the research onion framework developed by Saunders et al. (2023) to structure our research process. This framework provided a systematic approach to organizing our research, ensuring our methodology aligned with our research questions and problem statement.

The outermost layer of the research onion (Saunders et al., 2023) saw us using pragmatism. This choice was grounded in our aim to understand practical implications and perspectives of students, teachers, and heads of education regarding the integration of LLMs in Danish HEIs. By focusing on pragmatism, we could explore practical outcomes and insights from the participants' experiences. In the next layer, we adopted an abductive approach, allowing us to generate new knowledge based on the data and iteratively refine our research.

This approach was relevant given the exploratory nature of our research, which sought to investigate a new phenomenon without established theoretical foundations in the context of Danish HEIs. Our research strategy involved using interviews and surveys, specifically combining qualitative data from semi-structured interviews with quantitative data from surveys collected via RRT under the FRD sub-variant. This approach enabled us to capture qualitative insights and validate findings through quantitative analysis. Our methodological choice was mixed methods, capturing both qualitative and quantitative data. This ensured a comprehensive analysis and robust conclusions, providing a foundation for the validity of our findings. We employed a cross-sectional time horizon, collecting data at a single point in time to provide a snapshot of the phenomena. Data collection occurred in mid-April 2024 in Denmark, at Ringkjøbing Gymnasium (STX), and Herningsholm Erhvervsskole & Gymnasier (HHX/HTX). Our data collection and analysis techniques included thematic analysis for qualitative data and statistical analysis for quantitative data. These methods helped us identify and interpret patterns across the combined data.

The following table outlines the research methodology guiding our thesis research process.

Research onion layer	Definition	Our selected layer	Why & How It Was Used
Research Philosophy (1)	Refers to the set of principles concerning the philosophical stance from which the research is conducted.	Pragmatism	Pragmatism was chosen to explore practical implications and perspectives of participants.
Approach (2)	Plan for how research questions will be answered	Abductive	Abductive approach allowed us to generate new knowledge based on data and iteratively refine our subsequent research in the thesis.
Strategy (3)	Methodological link between philosophy and data collection	Interviews and surveys	Combining qualitative semi-structured interviews with quantitative RRT-survey data allowed for exploration and validation of findings.
Choice of Methods (4)	Selection of research methods (mono, mixed, or multi-method)	Mixed methods	Mixed methods were chosen to capture qualitative insights and via quantitative gain validation.

Time Horizon (5)	Timeframe over which the research is conducted	Cross-sectional	Cross-sectional time horizon was used to capture a snapshot of the current phenomena in HEIs.
Data Collection and Analysis (6)	Techniques and procedures used for gathering and analyzing data	Semi-structured interviews and RRT survey	Combine these techniques to gather qualitative data and validate with quantitative analysis.

Table 2: Overview of the applied format of research onion (Own creation)

3.2 Theme and coding

Qualitative data involves non-numerical information such as interviews, observations, and textual data, exploring specific phenomena. This type of data is common in social sciences, humanities, and fields requiring an in-depth understanding of human behavior, experiences, and interactions (Lacey & Luff, 2001). Methods for generating qualitative data include open-ended interviews, participant observations, and document analysis, providing insights into otherwise complex issues. Our research aimed to investigate the utilization of LLMs in Danish HEIs, examining their influence on teaching and learning. To gather qualitative data, we considered various methods, ultimately selecting semi-structured interviews (Lester et al., 2020). This method balances guided questioning with flexibility, allowing for in-depth exploration of participants' experiences and perspectives. Semi-structured interviews provide structure while enabling probing questions based on emerging themes. This approach ensured all relevant topics were covered while allowing follow-up questions for deeper insights, uncovering nuanced information and understanding the broader research context. An example of this in action could be discovering how students used LLM technology to assist with learning disabilities like dyslexia (Interview transcript Student E – ENG, Timecode: 11:01). This adaptability is valuable in qualitative research, where the goal is to explore complex phenomena and understand participants' underlying reasons, motivations, and experiences. Research can be classified into exploratory, explanatory, and descriptive categories (Zegeye et al., 2009). Exploratory research investigates new or unclear topics, generating insights without providing conclusive answers. Explanatory research explains relationships and causal links between variables, while descriptive research provides detailed accounts of phenomena.

This thesis is exploratory, aiming to investigate the new phenomenon of LLMs in Danish HEIs. Using qualitative methods, specifically semi-structured interviews, the thesis seeks to uncover students' experiences and perceptions regarding LLM integration and its impact on teaching and learning practices.

To process qualitative data, researchers follow several steps (Mackey & Gass, 2011). The process begins with data collection through interviews, observations, and document analysis. Collected data is then organized and prepared for analysis, involving transcription, reading, and coding—labeling segments of data with descriptive tags. Coding helps identify patterns, themes, and categories within the data. Researchers interpret these codes to uncover relationships and generate insights, often iteratively revisiting the data to refine findings. In this thesis, we captured information from semi-structured interviews by recording them (with participant consent) and transcribing them verbatim into Danish (Halcomb & Davidson, 2006).

As the interviews were conducted in Danish, we needed an English translation. Here we used semi-verbatim transcriptions, capturing the essence of conversations while omitting non-essential fillers. This method maintains data integrity, preserving the primary content and context of responses. Halcomb and Davidson (2006) pointed out that analysis techniques like thematic analysis aim to identify common ideas from the data, and therefore, verbatim transcripts are not always necessary, making semi-verbatim transcription suitable for our research objectives. However, translation can impact data integrity. Temple & Young (2004) assert that the power of translation to either reinforce, or undermine results, depending on how it is executed and integrated into research design, rather than the mere act of translation. To minimize the risk of data loss and bias, we followed Clark et al. (2017) methodology, producing the initial transcripts in the Danish language, and then translating it into English before analysis.

We created verbatim transcriptions of the 12 interviews and then translated them into semi-verbatim format in English, ensuring adherence to good research practice, avoiding translator bias, and preserving the original meaning.

Using our translated transcriptions from the semi-structured interviews, we selected thematic analysis (Kiger & Varpio, 2020) to identify patterns across qualitative datasets. Thematic analysis (Williams & Moser, 2019) is suitable for both inductive and deductive approaches, making it ideal for exploring Danish HEI teachers' and students' experiences with LLMs. Given the lack of prior research on this topic in Danish HEIs, we needed the flexibility to better explore the phenomena in question. This method also supports examining emerging themes, uncovering underlying issues, and contextual factors influencing LLM use in educational settings. Compared to content analysis, which focuses on counting the frequency of words or phrases (Neuendorf, 2018), thematic analysis provides a deeper understanding by interpreting patterns and themes within the data, offering better insights into participants' experiences. Using a guide by Maguire & Delahunt (2017), we applied thematic analysis principles to generate themes and codes from our data.

3.3 Randomized Response Technique (RRT)

Collecting quantitative data in social sciences involves measuring quantities and relationships between attributes using structured methods like surveys (Bowling & Ebrahim, 2005). Surveys, the most common method (Jopling, 2019), describe social phenomena, measure attitudes, behaviors, and self-perceptions, and allow statistical inferences about larger populations. Surveys gather information from a sample rather than a census, targeting individuals, organizations, or documents. They can be descriptive, exploring variable associations at a single point, or longitudinal, investigating causal relationships over time. However, when collecting data on sensitive topics, it is crucial to use ethical and respectful methodologies to ensure participant comfort and data accuracy. Chhabra et al. (2015) used Randomized Response Technique (RRT) to address sensitive questions about sexual abuse among students at the University of Delhi. RRT was used in this research case to protect participant privacy and encourage honest responses, as to avoid subsequent sanctions by their institutions. Due to the sensitive nature of such questions, they must be handled carefully to avoid causing distress to respondents, as many students might otherwise avoid these discussions, or provide inaccurate answers. Inspired by such studies, we employed RRT to investigate LLM usage by Danish HEI students, including illicit misuse.

Other research such as Reiber et al. (2023) also is seen to use RRT to identify dishonest responses, improving data accuracy on sensitive topics like doping in sports. RRT, developed by Stanley L. Warner (1965), reduces evasive answer bias by randomizing responses, making it difficult for interviewers to determine the truth of any specific answer. This technique involves a randomizing device, such as a dice, to decide whether respondents answer truthfully or provide a randomized response, protecting their privacy and encouraging honesty.

In our thesis, we utilized the Forced Response Design (FRD) format for RRT. This choice is supported by findings such as those in Krumpal (2012), which indicates that the forced response design is simpler to implement compared to other RRT variants, and in Lensvelt-Mulders et al. (2005), which observes that the forced response method variation are among the most efficient RRT-variants. Additionally, Lensvelt-Mulders et al. (2005) highlight that using a forced response method has the added benefit of allowing researchers to manipulate the perceived protection of the respondents. FRD ensures respondent anonymity when answering sensitive questions. Randomization determines if respondents truthfully answer a sensitive question or provide a forced "yes" or "no." For example, in our survey using dice, respondents rolling 1-3 must answer "yes", while those rolling 4-6 must provide their honest response. This method prevents individual responses from being traced back to respondents, encouraging truthful answers without fear of disclosure.

We utilized the FRD variant as demonstrated in Blair's 2014 study on civilian cooperation with militants, however changed it with minor adjustments for simplicity's sake (Blair et al., 2015). Respondents rolled a die to determine if they answered "yes" regardless of the question or provided an honest answer. A roll of 1-3 forced a "yes," while 4-6 required an honest response. We accomplished this by creating a question sheet seen in appendix A, which we printed for the Danish HEI students to answer in person on the location of their HEI, which helped us gather accurate data while protecting respondent privacy. Afterwards, we removed 50% of our collected yes's to account for the presence of "forced yes's" which was corresponding to our ratio used in the FRD-variant used.

3.4 Technical Architecture of RAG Application

This section outlines the technical architecture, and procedure for creating the RAG Application. Designed to assist students in retrieving text from the book "Samfundsfag C" (Frederiksen & Kureer, 2008). The book was borrowed, and used as reference in order to prime it on typical HEI teaching material that was deprecated, for copy-right reasons, but still applicable as far as class material. The application aims to provide accurate, explainable responses to improve student learning. The core of the agent uses a LLM coupled with RAG to yield correct answers.

Applications built around a LLM require a framework for configuration. Here, LangChain serves as such a framework, facilitating the integration of the LLM. It simplifies the process, providing necessary libraries for document retrieval and generation. LangChain handles every stage of LLM applications, from document loading and embedding to vector store and function generation for RAG (LangChain, 2022). For technological setup, we used the UCloud server of Aalborg University for data integration, security, and public link access. An environment using Python was set up, followed by creating five files: ingest.py, prompt.py, model.py, app.py, and requirements.txt.

Figure 7: Directory loader for loading pdf files (Own creation)

```
# Creating vector database for RAG application
def create_vector_db():
    loader = DirectoryLoader(DATA_PATH,
                             glob='*.pdf',
                             loader_cls=PyPDFLoader)

    documents = loader.load()
```

The initial objective was to create a vector database through ingest.py, with the function named create_vector_db. After loading the data, we used the RecursiveCharacterTextSplitter from LangChain to split the text into chunks with some overlap. The optimal chunk size was calculated based on the number of tokens in the PDF files, with the embedding model having a limit of 256 tokens per chunk. After calculating the optimal chunk size, the text was split into 700 chunks with an overlap of 105. Next, the function for embeddings was initialized using the model 'sentence-transformers/all-MiniLM-L6-v2' from Hugging Face via LangChain.

This model, based on sentence transformers, has a dimension of 384. The model is used to predict specific sentences from a vast dataset, achieving high precision in tasks like semantic similarity, clustering, and information retrieval (Hugging Face, 2022). The embeddings were created and stored into the FAISS index. Running `ingest.py` in the environment generates two files in the FAISS directory: `index.faiss` and `index.pkl`. The following code demonstrates the `create_vector_db` function:

Figure 8: Calling embedding and texts with FAISS for indexing and calling the function of vector db (Own creation)

```
embeddings = HuggingFaceEmbeddings(model_name='sentence-transformers/all-MiniLM-L6-v2')

db = FAISS.from_documents(texts, embeddings)
db.save_local(DB_FAISS_PATH)

if __name__ == "__main__":
    create_vector_db()
```

The next step was to create a prompt template by analyzing the PDF content of the book. Segmenting the content into subsections, allowed the creation of more meaningful prompts for the RAG. A function named `set_up_custom_prompt` was created to establish the context in which the LLM should operate. This function sets the context and the question, guiding the LLM on where to focus. These parameters serve as input variables. With the specified directions and input variables, the prompt template was initialized by calling the `return_prompt` function. Below is the snippet code for this operation:

Figure 9: Creating a custom prompt template (Own creation)

```
# custom prompt
def set_custom_prompt():
    """
    This function initializes a PromptTemplate object using the custom_prompt_template.
    It sets the template for QA retrieval, specifically for use with a vectorstore,
    with placeholders for dynamic content such as 'context' and 'question'.
    """
    prompt = PromptTemplate(template=custom_prompt_template,
                            input_variables=['context', 'question'])
    return prompt
```

We then reached the final phase of backend development in `model.py`. This calls the prompt template from `prompt.py` and rewrites the `DB_FAISS_PATH` created by `ingest.py`. Following these initial steps, a function from the retrieval QA chain is called. This chain integrates the LLM, database, and prompt template, forming the core function of the application. This function initiates the output process, allowing RAG to operate within the application. The function retrieves up to five nearest chunks to find the best answer for a query, keeping the source visible. This transparency allows users to see the document source of the answers, ensuring credibility. The function ensures all chunks follow the custom prompt template's instructions, maintaining consistency in responses. The function concludes by calling the `qa_chain`. The snippet below demonstrates this function:

Figure 10: Creating a retrieval qa chain function (Own creation)

```
#Retrieval QA Chain
def retrieval_qa_chain(llm, prompt_template, db):
    qa_chain = RetrievalQA.from_chain_type(llm=llm,
                                           chain_type='stuff',
                                           retriever=db.as_retriever(search_kwargs={'k': 5}),
                                           return_source_documents=True,
                                           chain_type_kwargs={'prompt': prompt_template})
    return qa_chain
```

After that, the LLM was initialized with the model 'gpt-3.5-turbo' and a temperature setting of 0.7. However, the important part of the application involves integrating all previously mentioned components into a single function, ensuring they all work together to perform RAG-related tasks. This function starts by calling embeddings from Hugging Face, then the database path. It then proceeds by calling the QA prompt, and QA chain from the retrieval QA chain function, effectively combining embeddings, the vector database, LLM, prompt, and QA chain under one unified function. This ensures that when a user asks a question, all functions work together seamlessly. The `qa_bot` function is the core of the application, concluding by calling the QA.

Below is the code snippet for this function:

Figure 11: Calling QA bot function to get the output (Own creation)

```
#QA Model Function
def qa_bot():
    embeddings = HuggingFaceEmbeddings(model_name="sentence-transformers/all-MiniLM-L6-v2")
    db = FAISS.load_local(DB_FAISS_PATH, embeddings, allow_dangerous_deserialization=True)
    llm = load_llm()
    qa_prompt_template = set_custom_prompt()
    qa = retrieval_qa_chain(llm, qa_prompt_template, db)

    return qa
```

Streamlit is a platform that turns code into applications, allowing developers to build and share. It is widely used, especially for generative AI applications, with over 190k applications built on it (Treuille, 2023). In our thesis we used streamlit for its frontend development and testing purposes.

4. Data analysis

4.1 Qualitative data analysis

The qualitative data that was used in this thesis was collected through semi-structured interviews with students, teachers, and heads of education from the two participating HEI institutions in Denmark. For all semi-structured interviews the respondents were participating voluntarily, and with the knowledge that the interviews would be used for the purposes of the thesis, and recorded for transcriptions. The interviews took place in April 2024. The duration of the interviews ranged from approximately 10 to 30 minutes. In total, the participants made up N= 13, and a combined duration of 04:12 hours of interviews. For each group of the respondents, being the high school students, teachers, and heads of education, there was an interview guide tailored for that type of participant. This was in order to attain information and experiences related to their specific roles in HEI which was assumed to be different, therefore likely biased. In the appendices, they are seen to be **[B: Head of Education interview guide]**, **[C: High school teacher interview guide]**, and **[D: High school student interview guide]**. As seen, the questions were structured into categories, however there were no predetermined codes or themes before analyzing the data.

This is due to the exploratory nature of this thesis, whereas we are not applying hypotheses or assumptions for LLM usage before attaining the data.

In the following table there is an overview of the details pertaining to the conducted interviews.

Interviews	Role	Date	Location	Duration	Transcript
Student A	3rd year HHX student	22/04/2024	Herningsholm Erhvervsskole & Gymansier	15:31 minutes	F: Interview transcript Student A - ENG
Student B	1st year HHX student	22/04/2024	Herningsholm Erhvervsskole & Gymansier	08:25 minutes	G: Interview transcript Student B - ENG
Student C	3rd year HTX student	23/04/2024	Herningsholm Erhvervsskole & Gymansier	12:33 minutes	H: Interview transcript Student C - ENG
Student D	2nd year HTX	23/04/2024	Herningsholm Erhvervsskole & Gymansier	17:36 minutes	I: Interview transcript Student D - ENG
Student E	2nd year HTX	23/04/2024	Herningsholm Erhvervsskole & Gymansier	14.15 minutes	J: Interview transcript Student E - ENG
Student F	2nd year STX student	24/04/2024	Ringkjøbing Gymnasium	11:34 minutes	K: Interview transcript Student F - ENG
Student G	1st year STX student	24/04/2024	Ringkjøbing Gymnasium	14:05 minutes	L: Interview transcript Student G - ENG
Student H & I	2nd year STX student(s)	24/04/2024	Ringkjøbing Gymnasium	28:55 minutes	M: Interview transcript Student H & I - ENG
Teacher A	English & Psychology teacher	24/04/2024	Ringkjøbing Gymnasium	36:14 minutes	N: Interview transcript Teacher A - ENG

Teacher B	English and History substitute teacher	26/04/2024	Herningsholm Erhvervsskole & Gymansier	32:36 minutes	O: Interview transcript Teacher B - ENG
Head of Education A	Head of Education & Spanish and Danish teacher	24/04/2024	Ringkjøbing Gymnasium	30:39 minutes	P: Interview transcript Head of Education A - ENG
Head of Education B	Head of Education	26/04/2024	Herningsholm Erhvervsskole & Gymansier	30:00 minutes	Q: Interview transcript Head of Education B - ENG

Table 3: Overview of the semi-structured interviews (Own creation)

Amongst said participants, the experiences with LLMs varied. Primarily it was revealed that there was an overwhelming frequency of LLM use amongst all students, in fact, all interviewed students have used it in varying degrees. However one interview stood out, which was **Students H & I**. Both students participated in an interview whereas they were interviewed simultaneously due to time constraints. They had both been caught using LLMs illicitly at Ringkjøbing Gymnasium from the same class, and by their participation, were able to offer unique viewpoints relating to that, and how they likewise evaluated LLM use in the aftermath of it. This was a voluntary participation as far as the interviewer is informed. In reflection, this could have been a forced participation, but we do not wish to speculate on said circumstances.

4.1.1 Theme 1: The usage of LLMs in education

The first theme addresses the usage of LLMs in education in relation to our research question 1, focusing on students' experiences and reasons for integrating these tools into their learning environments. Participants were asked about their initial experiences with LLMs, their primary purposes for using it, and the specific reasons that prompted them to start using LLMs. This theme helps to clarify the context of student usage of LLMs in Danish HEIs. From the interviews, it is clear that the majority of students have adopted LLMs mainly to help write assignments, whether for generating content, improving grammatical accuracy, or structuring their work.

(RQ1) Theme 1: The usage of LLMs in education

Participant	First experience with LLM use.	Purpose of LLM use	Reasons associated with starting to use LLMs
Student A	"1 to 2 years ago."	" assignments, if there's something grammatical"	"100 times better than what I could manage myself."
Student B	"A year - two max."	"if I need to write a company profile (...) if I need to make an analysis of something"	"it wrote well (...) it was a good experience (...). been using it ever since for help with assignments."
Student C	"Between 1 to 2 years."	" programming, also a bit in Danish to find analysis points."	"If you have some code, where you just can't figure out why it doesn't work"
Student D	"About a year ago."	"school assignments"	"There was an assignment where I thought "Maybe I can get some help from it and get some inspiration."
Student E	"About two months ago"	" mainly for my structure because I am so dyslexic, so I really struggle to maintain the structure"	" interaction design with some code (...) how I can do a good analysis for, for example, a larger assignment if I'm in doubt"
Student F	"Yes, about a year ago"	" analytical essay in English (...) if I have a biology assignment (...) just rewrite"	"the teacher asked some questions and then I tried using ChatGPT to answer them "
Student G	"Last year (1 year ago)"	"I just put the whole text in, and then I tell it to make notes."	"started using chat here at the start of high school."
Student H	"At the end of 1.G (1 year ago)"	"create a news article, a debating article, how to use it."	"It seemed like an easy way to finish (...) throw the whole task onto ChatGPT, and then it wrote it for me."
Student I	"End of 1.G (1 year ago)"	"get it to do my assignments, especially in language subjects"	" It was time pressure. I hadn't started in proper time. (...) I saw it as a kind of lifesaver in that way"
Head of Education A	"before the exams last year (...) seems like it has exploded since then (1 year ago)."	"all subjects are affected (...) in the written work. (...) When students are doing assignments at home, they use ChatGPT. "	" it's typically those who are struggling academically who have used ChatGPT"
Head of Education B	"When we started in August (About a year ago)"	"the students use it in an inappropriate way to produce products"	"it's more about whether the student is under pressure or not"
Teacher A	"A year and a half I'd had them by then"	"My clear sense is that the overwhelming majority of students who use LLM use it to cheat."	"in 99% of cases where ChatGPT is used by students, it is used to produce assignments"
Teacher B	"I had specific cases in January 2023 (About 1 year ago)"	"They consistently use it as a translator. (...) they do an assignment themselves, and then run it through ChatGPT to make it look better."	"use ChatGPT to ensure their descriptions of the tasks they've done. Or you know, the linguistic part, that it's grammatically correct"

Table 4: Theme 1: The usage of LLMs in education (Own creation)

The data overwhelmingly indicates that LLMs are being used to ease the workload and improve the quality of the students' assignments. For example, Student A uses LLMs for grammatical improvements in assignments, while Student B utilizes them for writing company profiles and analysis. Students C and E employ LLMs for programming and maintaining structural integrity in their work, respectively. This trend is further supported by Student I, who resorts to LLMs under time constraints, indicating that LLMs are used to overcome academic pressures and deadlines. Teacher B also denotes that the LLMs are effective in the language-related subjects, which Student F validates in their experiences when using it in English class. However, there is a notable divergence in the perception between students and teachers. Students generally view LLMs as a practical aid to improve their academic output, often citing efficiency and quality improvement. On the other hand, the Heads of Education and Teachers express significant concerns about misuse. For instance, Teacher A observes that the majority of students use LLMs to avoid doing their assignments themselves, and Teacher B notes that students frequently use these tools to refine their work and for grammatical accuracy. The Head of Education A also points out that academically struggling students are more likely to rely on LLMs, potentially as a crutch to cope with academic demands.

This difference between students' intended use and the teachers' observations underscores that while LLMs can be beneficial, there is a clear indication of it being used to circumvent genuine learning, thus indicating that there is an existing conflict as far as the LLMs being misused in the HEI learning environments. Likewise, it also raises the very important question if the students are at all capable of understanding the misuse themselves, or if they are intentionally doing it despite knowing they are not using LLMs ethically correct. One term connected to LLMs that is prevalent is the word *“inspiration”*, when students describe how they use LLMs. It may suggest it is used as a euphemism for what in other words can be described as cheating.

4.1.2 Theme 2: Positive aspects of LLM use in HEI

The second theme explores the positive aspects of LLM use in HEI focusing on how it can assist students in three areas, tailored to serve research question 2. The theme highlights the benefits perceived by the participants.

Participants provided accounts of how LLMs have supported them in receiving inspiration when making assignments, such as improving their ability to learn in language-related classes, and LLM use leveling them to an equal footing with students from more privileged backgrounds from parents, or without disabilities impeding them in contrast.

(RQ2) Theme 2: Positive aspects of LLM use in HEI	
Areas that LLM can assist with	Illustrative comments
Aid to disadvantaged students & poor socio-economic backgrounds	<ul style="list-style-type: none"> • [06:31] Student A “alternative for those who maybe have many problems with, for example, dyslexia (...) that they just use this model, which can do the work a hundred times better than they themselves and makes it much easier” • [11:03] Student E “help me with the type of dyslexia I have because I have a lot of trouble seeing a clear thread through my texts and the way I structure things.” • [18:39] Head of Education A “There's a lot of discussion right now saying that it has a democratizing potential, right? The weaker student suddenly gets a tutor that they don't have in their parents. “ • [20:51] Teacher A “My dream is that LLM can be, be a sparring partner, like some of our most fortunate most privileged students have in their parents”
Assist in generating inspiration for homework	<ul style="list-style-type: none"> • [02:33] Student A “I don't take it directly, but I take inspiration from the words and phrasings that have been used.” • [01:23] Student B “I use it mostly for inspiration.” • [06:30] Student I “take inspiration from what it can provide. It can give good examples, but it can't do it all for you.” • [09:48] Student H “It has helped me, for example, to get started with some assignments. I've used it as inspiration to create an introduction to several articles, and then I went on to create the assignment from there”
Proficient in language-related classes	<ul style="list-style-type: none"> • [01:16] Student F “if I have to write an analytical essay in English, then it's mostly about "How do you write that?". Like, structure, and I ask ChatGPT about it. But sometimes if I have a biology assignment and there's a specific question” • [02:33] Student A “we also use it sometimes in teaching for German, where I can translate it, and then understand it better in Danish” • [03:28] Student B “You can really get it to say a lot and then translate texts from different languages” • [02:45] Student I “especially in language subjects” • [02:04] Student C “If you have some code, where you just can't figure out why it doesn't work, then you can just ask it like "Hey, it could be this, this, this” • [01:05] Student E “it can give me some code and then I introduce it to the program”

Table 5: Theme 2: Positive aspects of LLM use in HEI (Own creation)

It is seen that Student E mentioned that LLMs make it easier to maintain the structure of their texts, a task they typically struggle with due to dyslexia. Additionally, Head of Education A pointed out the democratizing potential of LLMs, as these tools can act as tutors, providing support that might not be available at home. The notion that LLMs can serve as an equalizer, offering similar advantages to those enjoyed by more privileged peers, was a recurring theme amongst the Teachers and Head of Education A. LLMs also play a role in generating inspiration for homework and assignments. Many participants reported using LLMs to gain ideas and improve the quality of their work. For example, Student A explained the use of LLMs to find the right words and phrases, while another mentioned taking inspiration from the examples provided by the LLM. This usage extends to creating outlines and structuring essays, where LLMs offer a starting point that students can then expand upon. The ability of LLMs to provide initial guidance and stimulate creative thinking is highly valued, helping students to overcome writer's block and improve their academic output. Additionally, language-related assistance and improved comprehension are notable benefits of LLMs. Students frequently use these tools to aid in understanding complex topics and translating difficult texts. For instance, LLMs are used to assist with language subjects, providing translations and clarifications that improves comprehension. This extends into domain-specific fields, such as programming and coding.

4.1.3 Theme 3: Challenges and issues with students LLM usage

The third theme explores the negative aspects of LLM use in HEI, focusing on what problems are associated with the use towards education or learning in HEIs. This theme was used to support research question 2, clarifying the adverse consequences of students' use of LLMs.

(RQ2) Theme 3: Challenges and issues with students LLM usage	
Areas of concern	Illustrative comments
The ability to learn is reduced by usage	<ul style="list-style-type: none"> • [03:14] Student C “It has made troubleshooting less important. I don't feel like I spend as much time finding errors” • [05:52] Student E “I'm not quite sure if it's a tool, or if it diminishes one's own work”. • [01:47] Student F “I feel I've become a bit less independent because you have that thing in the back of your mind that if there's something you find a bit difficult, you can just ask ChatGPT.” • [09:35] Student G “But I also feel that it made me not really want to. Yes, I feel a bit like it has made me learn less.” • [01:30] Student G “I've become lazy? I don't really want to read it myself when it can do it. Yeah, I just think it has become easier. Kind of a shortcut, so you get lazy.”

	<ul style="list-style-type: none"> • [12:05] Student H “We don't learn anything from it as such.” • [08:08] Student D “But I feel that there are many who will cheat themselves out of learning something because they can just have it generate a text.” • [08:41] Student G: “if we ourselves aren't learning anything, so if everything just goes through it, if we aren't learning anything anyway, why are we even going to school?”
Prolonged usage leads to dependency on LLMs	<ul style="list-style-type: none"> • [20:59] Head of Education A “They would at least—In the old days, it would be the student who came up and said, “I didn't do the homework” • [03:40] Student E “Some might become too dependent on it so they can't do anything themselves without asking this device, and that it becomes too much, it takes over” • [07:32] Student G “it's a bit of an addiction. It's just become so easy. Now that it's so easy to get the answers, then you wouldn't make it hard for yourself.” • [25:04] Student I “I feel it's 100% because you're dependent. When you've used ChatGPT for half a year, you can't sit down and write a new Danish essay entirely on your own from scratch. I've seen that with others. It's difficult for them, and then they find out “I don't want to spend 4 hours on this. It will never be as good anyway as if I spent 10 minutes on ChatGPT.”
The usage is largely associated with plagiarism	<ul style="list-style-type: none"> • [03:28] Student B “some will probably use it to cheat, directly cheat and then copy it directly onto their assignment.” • [05:42] Student C “at least from our Danish teacher, there are quite a few who apparently use it for their assignments and make whole assignments in ChatGPT.” • [00:32] Student D “It felt a bit like cheating in terms of schoolwork and stuff” • [06:31] Student A “my friends. I know at least some of them use it a bit more, more aggressively in terms of sometimes just taking pieces of it, and then from that, they've written an assignment.” • [08:08] Student D “several of my own classmates have done that, and just gotten it to write something, and then just change a bit in it so that it looks like something they've written, and they don't get caught for plagiarism, and then they almost have a finished assignment without having done anything or learned anything.” • [22:14] Student I “the SRO assignment (...) Those I knew, at least a third used it.”

Table 6: Theme 3: Challenges and issues with students LLM usage (Own creation)

The primary concern with LLM usage is that it reduces students' ability to learn independently. Many students report that LLMs diminish the need for troubleshooting and critical thinking, as they rely on these tools to solve difficulties they might otherwise have worked through on their own. This reliance leads to decreased independence and a tendency to avoid engaging with their studies, resulting in a shortcut mentality where genuine learning is sacrificed for convenience. Some students feel that LLMs make them lazier, and less inclined to put in the necessary effort for academic growth. One noteworthy example is found in Student G, a female student who by her admissions expresses signs of addiction. It is seen her ability to learn and engage in HEI is reduced by her LLM use, which in turn has also made her dependent on the use. This gender/addiction variable is further analyzed in our quantitative findings in 5.2.1, which reveals gender does play a role in addiction patterns as far as LLM usage, which Student G correspondingly indicates.

Likewise, there is also evidence to suggest LLM use is also prevalent in exams, whereas Student I mentions about 1 / 3 of those he knew plagiarized their SRO assignments by using LLMs. These exams are take-home exam projects written outside HEIs. Prolonged usage of LLMs can lead to dependency, further aggravating the issue. Participants mentions that over-reliance on LLMs hampers their ability to complete tasks independently, comparing it to an addiction where ease of access to answers discourages personal effort. This dependency undermines the development of academic skills and erodes students' confidence in their own abilities. Additionally, LLM use is closely associated with plagiarism, as students frequently generate and submit entire assignments produced by these tools. This practice raises ethical concerns, violates academic honesty policies, and deprives students of genuine learning opportunities. As per ZPD, learning requires challenge, and without it, there is no learning involved. Teachers and Heads of Education have observed this misuse, highlighting the need for clearer guidelines and rules, to mitigate the negative impact of LLMs on student learning and academic integrity, but they don't exist currently outside of exams or larger study direction projects akin to SOP/SRO/SRP, etc.

4.1.4 Theme 4: LLMs are changing HEI educational practices

The fourth theme supports research question 3, aiming to clarify the role LLMs play in changing existing HEI education structure, and the effects it has had on HEI teaching as a whole. The theme highlights several areas where LLMs are impacting traditional educational practices, particularly in the context of written assignments, the Danish HEI educational format, and the ability of HEIs to address plagiarism effectively.

(RQ 3) Theme 4: LLMs are changing HEI educational practices	
Areas being affected by LLMs	Illustrative comments
Written assignments have been rendered useless	<ul style="list-style-type: none"> • [01:56] Head of Education A “teachers coming down and having difficulty assessing the work that the students have submitted. They have difficulty determining and judging whether these assignments can be accepted” • [12:59] Head of Education A “That is, the type of teaching I have done so far. I can no longer do that (...) I can no longer give the same written assignment as we could two years ago. “ • [12:59] Head of Education A “assignment is completely irrelevant. I can't use it for anything. I have myself experienced now how there is a huge difference between the assignments they submit during the year, and then what they do for the terminal exam.”

	<ul style="list-style-type: none"> • [24:05] Teacher A “writing in high school is based on an assumption that if a student submits some product, then one can based on reading that product assess the student's academic ability (...) That equals sign we have to realize has disappeared” • [24:05] Teacher A “You can no longer look at a written product a student has made and then based on that written product assess their academics, because you really have no chance of knowing if they themselves have written it any longer.” • [14:42] Teacher B “I had specific cases in January 2023, the first written assignment I made after ChatGPT 3.5 became generally available, where I caught 19 students in a class of 30 students” • [16:20] Teacher B “there's collective agreement in the humanities teacher group, that the concept of written assignments no longer works.”
The Danish educational format is in conflict with LLM use.	<ul style="list-style-type: none"> • [31:13] Teacher B “we must develop students' independence. It is a core concept in Danish school education(...) That is the classic school form, where it's about rote learning. ChatGPT is closer to that than the traditional Danish education policy, which is about students having a high degree of independence” • [12:59] Head of Education A “learning vocabulary is suddenly something completely different because they think “Why should I do that? I can just write it in ChatGPT, and then it has translated it for me” (...) All subjects, and all teaching, need to rethink their pedagogy both in writing, but also in the daily teaching.” • [23:51] Teacher B “students are influenced by a program that is made in the USA, and they need to be able to be critical of the fact. They should be able to present their own opinions, come up with independent viewpoints” • [03:35] Student A “it is a threat to the entire educational system as we know it today. Because I believe that, like in the old days, some of the things we have, it was just about memorizing a lot” • [10:17] Student I “It limits what you can do in an eventual exam if you haven't delved into what you've learned yourself.”
HEIs are unable to punish LLM use.	<ul style="list-style-type: none"> • [14:42] Teacher B “The biggest problem we have is the students who don't do what they should. The school has chosen to handle it by saying, “We can't punish them for plagiarism every single time, because then we'd have to expel half the students at the school.” • [03:36] Head of Education B “when teachers come in and want us to give students warnings because they have plagiarized, and we need to calm them down and tell them “Well okay, you've seen the student has plagiarized something, but how much is original? You can't judge that.” • [15:21] Teacher B “I asked them to resubmit the assignment. But afterward, we received guidelines that you can't take them for cheating for having done it” • [25:19] Teacher A “if we announced tomorrow that from now on, all submissions at the gymnasium will be written kind of like final exams (...) 80% of the students who were supposed to start here next summer, they would switch to the business school instead if they hadn't introduced this system” • [01:31] Teacher B “But now we have been told to take parts of assignments that are written by an AI, and if we assess them to be so, remove those parts from the assessment instead, and then assess what remains.” • [14:14] Teacher A “I don't have a program to run students' assignments through to find out if they were produced with the help of an LLM, as you call it” • [09:45] Teacher B “The school pays for plagiarism control, for example, but I know most teachers here don't use it.” • [17:40] Teacher A “I think there might also be an underlying hesitation. I just can't vouch for that; I can't document it. So, I just state that I don't think it happens much because it's a really unpleasant thing to have to do as a teacher.”

Table 7: Theme 4: LLMs are changing HEI educational practices (Own creation)

One thing is clear, written assignments have been rendered seemingly useless as tools for assessing student learning. Teachers report difficulty in evaluating the authenticity of student submissions, as it becomes challenging to determine if the work was genuinely produced by the students themselves. For example, both A & B teachers express concerns about the reliability of written assignments, noting that they can no longer confidently assess students' academic abilities based on submitted work. This issue is made worse by cases where students demonstrate a significant discrepancy between their regular assignments, and their performance in verbal exams. Consequently, there is a collective agreement among teachers that the traditional concept of written assignments needs reevaluation in light of LLM usage, as it currently is rendered ineffective in assessing student academic output. The Danish educational format, which emphasizes student independence and critical thinking, is also in conflict with the use of LLMs.

The reliance on LLMs for completing assignments undermines the core principles of the Danish education format, where students are expected to engage with the material and develop their own viewpoints. Teachers highlight the need to rethink teaching methods and pedagogies to address the influence of LLMs on student learning habits. They stress the importance of fostering genuine learning experiences over shortcut methods facilitated by LLMs. Moreover, the inability of HEIs to effectively punish LLM misuse poses a significant challenge. Schools face difficulties in distinguishing between original and AI-generated work, and existing plagiarism detection methods are often inadequate for identifying LLM use. This situation leads to a dilemma where punishing students for LLM-related plagiarism could result in expelling a significant portion of the student body, forcing institutions to seek more nuanced approaches to managing the integration of LLMs in education, and individual teachers dealing with the responsibility of detecting, and punishing misuse, something they do not want to share the responsibility of as it is something they are unable to carry out, and/or is uncomfortable with the act itself.

4.1.5 Theme 5: How to solve the current problems associated with LLMs

The fifth and final theme supports research question three, exploring the suggestions provided by participants to address the issues associated with LLMs and identifying the ideal ways to integrate LLMs into HEIs. This theme examines how to use LLMs constructively in HEI education.

(RQ3) Theme 5: How to solve the current problems associated with LLMs	
Areas of improvement	Illustrative comments
Learning how to use LLMs properly	<ul style="list-style-type: none"> • [10:22] Student F “learning how to use it properly, so it's not such a taboo to use” • [09:53] Student C “to teach more about it” • [13:59] Student A “at these educational institutions we should simply teach students how to use it appropriately” • [20:13] Teacher B “I recognize the premise that this is reality, but one must figure out how to use it practically.”
National solution is required	<ul style="list-style-type: none"> • [25:19] Teacher A “that there has to be some kind of central solution” • [26:33] Teacher B “I am personally very, very disappointed in the proposal that came from the expert council. Because the exam formats, they described from what I could read there, seem to me identical to those we have today” • [00:36] Teacher A “You mention specific guidelines. I don't think we have explicitly received any. “ • [00:49] Teacher B “I am not aware of us receiving any specific guidelines for that. It actually doesn't ring a bell, which is funny because I am a member of the AI committee at this school”
Necessity for improvements in existing LLMs in order to be used in HEIs	<ul style="list-style-type: none"> • [05:17] Student A “For example, with mathematics, you might type a question in, and then it just thinks in completely different ways “ • [06:28] Student F “when it was supposed to come with a source or references to the text, it was something completely different from what was in the text” • [05:17] Student A “There are times when we sit and do schoolwork just to check if it is correct, where it just writes something completely different” • [06:15] Student H “I have several times experienced that it was wrong, what it gave me. And then I had to figure out myself how, and where, and what was right.” • [10:58] Student C “it's bad at correcting its own errors. If you say there's a mistake, then it's like "Ah okay, I'll fix it right away", and then it's not fixed” • [04:36] Student C “You can ask them the same question, and then you get two conflicting answers. It's always like that.”

Table 14: Theme 5: LLMs are changing HEI educational practices (Own creation)

Participants emphasized the importance of teaching students how to use LLMs effectively and ethically. This includes integrating LLM into the curriculum to ensure it is seen as a tool that is used as a scaffolding for learning. By educating students on the appropriate use of LLMs, these tools can be leveraged to support, and improve learning, rather than hinder it. Participants believe that developing a clear understanding of how to use LLMs practically and responsibly will help maximize their benefits in education. The necessity for improvements in existing LLMs for use in HEIs is another important sub-theme that emerged from the interviews. Participants reported several issues, such as LLMs generating inconsistent or incorrect outputs, which can lead to confusion and mistrust among users.

Specific problems include grammatical errors, and the generation of unrelated sources or references. This is tied to the fact that LLMs based on the interviews seem to be better suited towards humanities, rather than mathematical, or natural-science based subjects.

These flaws undermine the reliability of LLMs used as an educational tool. Furthermore, participants noted the difficulty in correcting these errors, as LLMs often fail to properly adjust their outputs even after being prompted. This can result in frustration and inefficiency, hindering the overall learning. There is a consensus that changes are needed to improve the accuracy and reliability of LLMs. Ensuring that it can provide consistent, correct, and contextually appropriate information is a necessity for the successful integration into education such as HEIs in Denmark. By addressing these technical shortcomings, LLMs can better serve as reliable aids in the academic development of students.

Additionally, teachers stressed the need for a national solution to address the challenges posed by LLMs. None of the teachers were aware of any rules issued nationally or locally by their HEIs relating to LLM use outside of exams. They called for central guidelines and policies to standardize the use of LLMs across HEI educational institutions. Teacher B expressed disappointment with current proposals from the expert group's recommendation made on behalf of the Ministry of Education, noting that existing exam formats and guidelines do not adequately address the challenges of LLM usage. There is a call for more explicit and practical guidelines that can help teachers and students navigate the use of LLMs better, as likewise there is a noticeable lack of perception of what is "inspiration" and "cheating" when it comes to the student's perception of best practice in regards to LLM use. By establishing a national solution, HEIs can ensure a consistent approach to using LLMs, thereby addressing concerns about misuse and utilizing their positive potential in education. Likewise, it is seen the current rules and guidelines extend primarily to the use of exams, and larger study projects like SOP/SRO/SRP, etc, but the real issue is the usage in-between classes that is not properly addressed by a lack of rules and guidelines issued by the Ministry of Education, and/or individual HEIs.

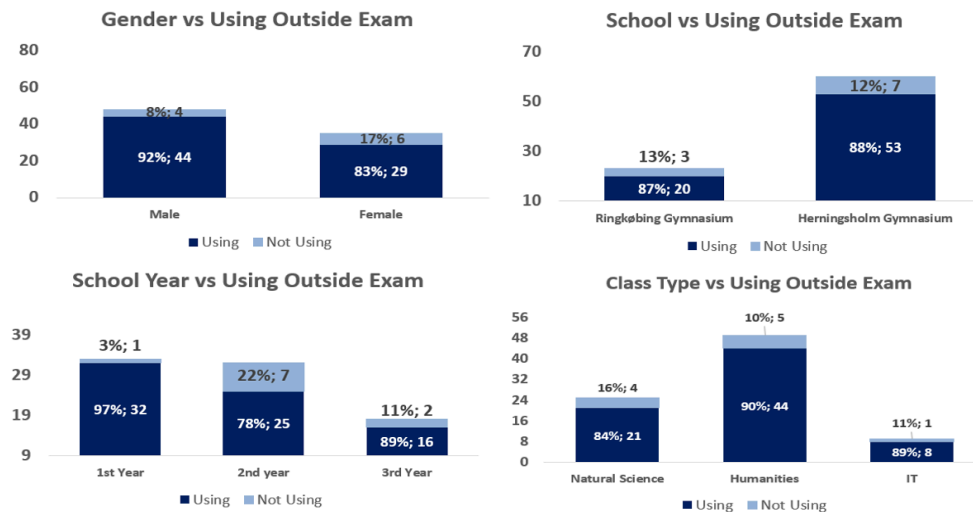
4.2 Quantitative data analysis

Total sample for our quantitative data was N=156 students which was gathered from the two HEIs that participated in the thesis. However, the data have five target variables which are the RRT survey questions that have been answered at the time of data collection. The project will analyze each question, starting with question 1. These questions were designed to generate data on the students' usage with LLM within HEIS, as seen with exams, homework and learning purposes. The answers will also reveal the students dependency on learning using LLMs. The performed analyses have been conducted on the quantitative data are Cross tabulation, T-Test, ANOVA test, and Confidence Intervals. As the data is highly imbalanced for gender and school the thesis has utilized Welch's T-Test. Welch's T-Test is used for imbalanced data sets (Ahad & Yahaya, 2014). Added to that, school year and class type has more than 2 variables, therefore the project utilized ANOVA for those independent variables (Shaw & Mitchell-Olds, 1993).

4.2.1 Question 1: Using Outside Exam

Cross Tabulation and Correlation

Figure 12: Cross Tabulation with Using Outside Exam (Own creation)



Gender vs Using Outside Exam

It is seen that men are likely to use more LLM than women for learning outside of the exam preparations. It is seen that 92% of male respondents are using the LLM outside of the exam preparation for learning purposes, whereas 83% female respondents are using in comparison.

School vs Using Outside Exam

The cross tabulation shows little difference between the two HEIs as far as student usage of LLMs for learning outside of exams. This can be seen as 88% of the students from Herningsholm Gymnasium are using LLMs, whereas 87% of the students of Ringkøbing Gymnasium are using LLMs.

School Year vs Using Outside Exam

From the cross tabulation it is seen that 1st year students 97% are using LLMs. The usage of the LLM is not as frequent in comparison to the other 2nd and 3rd year students. This is shown as the percentage of 2nd year students is 78% are using LLMs, whereas 3rd year students 89% are using LLMs, again for learning outside of exam contexts.

Class Type vs Using Outside Exam

The cross tabulation shows natural science students are less likely to use LLM outside of exam learning than other two class types. Only 84% of students are using LLMs, whereas 90% in Humanities classes, and 89% IT class students are using LLM outside of the exam for learning purposes.

T-test, ANOVA test & Confidence Interval

The thesis has conducted a T-Test and an ANOVA test by setting a null hypothesis that there are no significant differences between the target column which is outside of the exam vs all other independent columns. Where the benchmark alpha level is 0.05 and 95% confidence interval.

Figure 13: T-Test for Gender and School vs Using Outside Exam (Own creation)

T-Test

Variable	T-Test	P-Value	Decision	CI Lower	CI Upper
Gender	1.17	0.27	Fail to reject H0: No significant difference	(0.14)	0.54
HighSchool	(0.16)	0.87	Fail to reject H0: No significant difference	(0.34)	0.29

T-Test has been conducted on the variables of gender and high school. Where male = 0 and female = 1 and Ringkjøbing Gymnasium = 0 and Herningsholm Gymnasium = 1. The T-Test could not find any significant difference between the usage outside of the exam vs gender and

school, as the p-value is above 0.05. Lower T-Test score of gender signifies that there is a low mean difference between the variables and males are using more LLM than females outside of the exam. Also, the negative T-Test score result is also here, which indicates the Herningsholm Gymnasium students are using more than Ringkjøbing Gymnasium students. The confidence interval includes 0, as the lower bound is negative. It suggests that there can be no mean difference at 95% confidence level.

Figure 14: Anova Test for School Year & Class Type vs Using Outside Exam (Own creation)

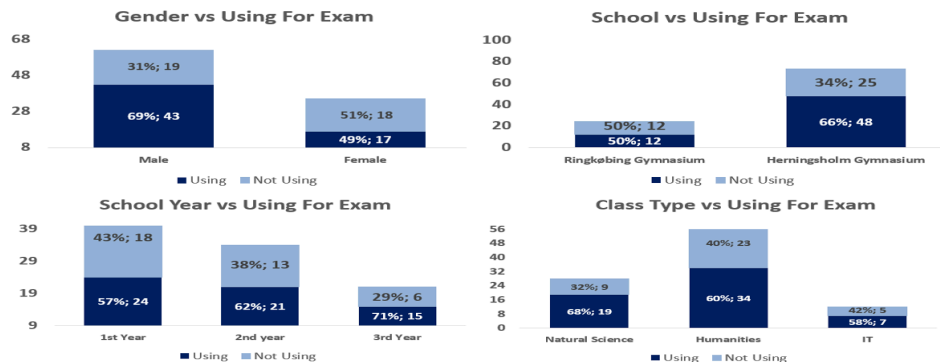
ANOVA							
Variable	F-Statistic	P-Value	Decision	Group - 1 CI		Group - 2 CI	
SchoolYear	1.53	0.22	Fail to reject H0: No significant difference	0.60	0.96	0.69	1.51
ClassType	0.34	0.56	Fail to reject H0: No significant difference	0.68	0.96	0.22	1.18

ANOVA test has been conducted on school year and class type, where three different school years and class types exist. The ANOVA could not find any significant difference between the usage outside of the exam vs school year and class type, as the p-value is above 0.05. The overlapping condition suggests that there is very low difference in the mean, but 0 mean difference can not be achieved in any condition. It also implies that the condition is not statistically significant.

4.2.2 Question 2: Using For Exam

Cross Tabulation and Correlation

Figure 15: Cross Tabulation with Using For Exam (Own creation)



Gender vs Using For Exam

As per cross tabulation, men are 69% more likely to use LLMs more for exam preparation than women, which are 49%. This means that men, in contrast to women, use it by 20% more for exam preparations, then women. This further reveals that men are also more likely to use it.

School vs Using For Exam

The cross tabulation result is showing Herningsholm Gymnasium students are more likely to use LLM for exams than Ringkøbing Gymnasium. As 66% of the students from Herningsholm Gymnasium are using, whereas, 50% of the students of Ringkøbing Gymnasium are using LLMs for the exam.

School Year vs Using For Exam

From the cross tabulation it is visible that first year students are less prone to use LLM for exams in comparison to the 2nd, and 3rd year students. This can be seen as LLM usage is exponentially increased by the later school years. This can be seen as 71% of the students from 3rd year are using it for exams, whereas first year, and second year students are using it by 57%, and by 62% respectively.

Class Type vs Using For Exam

The cross tabulation result shows that, natural science students are more likely to use LLM for the exams, than humanities, and IT students. 68% of natural science students are using LLMs, whereas 60% of Humanities, and 58% of IT class students are using LLM for the exams.

T-test, ANOVA test & Confidence Interval

The T-Test has found statistical significance between gender and using LLM in exams and rejected the null hypothesis, due to the p value being in the border line. Likewise, the confidence interval suggested the positive mean difference, thus indicating the statistical significance. As the T-Test score is positive, it thus means men are using more LLM than women for exams.

Figure 16: T-Test for Gender and School vs Using For Exam (Own creation)

T-Test

Variable	T-Test	P-Value	Decision	CI Lower	CI Upper
Gender	1.99	0.05	Reject H0: Significant difference	0.00	0.40
HighSchool	(1.33)	0.19	Fail to reject H0: No significant difference	(0.31)	0.06

The T-Test could not find any significant difference between the usage of LLM for the exam vs school, as the p-value is above 0.05. The negative T-Test score result is also here, which indicates the Herningsholm Gymnasium students are using more than Ringkjøbing Gymnasium. Likewise, the confidence interval includes 0, as the lower bound is negative. This suggests that there can be no mean difference at 95% confidence level.

Figure 17: Anova Test for School Year & Class Type vs Using For Exam (Own creation)

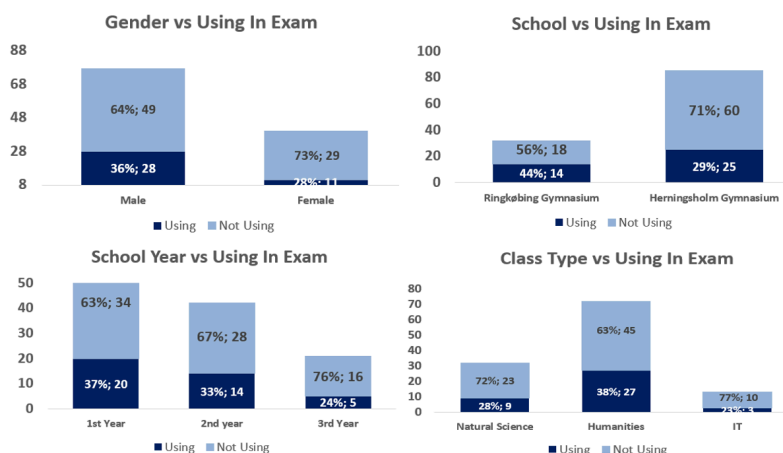
ANOVA							
Variable	F-Statistic	P-Value	Decision	Group - 1 CI		Group - 2 CI	
SchoolYear	1.14	0.29	Fail to reject H0: No significant difference	0.64	1.06	0.43	0.92
ClassType	0.49	0.48	Fail to reject H0: No significant difference	0.64	0.96	0.69	1.10

ANOVA test has been conducted on school year and class type. The ANOVA could not find any significant difference between the usage of LLM for the exam vs school year and class type, as the p-value is above 0.05. The overlapping condition suggests that there is very low difference in the mean, but 0 mean difference can not be achieved in any condition. It also implies that the condition is not statistically significant.

4.2.3 Question 3: Using In Exam

Cross Tabulation and Correlation

Figure 18: Cross Tabulation with Using In Exam (Own creation)



Gender vs Using In Exam

As per cross tabulation, men are more likely to use LLM, than women, during the exams. This can be seen as 36% of men are using it during exams, whereas for women it is 28%. Therefore it further indicates that men are more prone to utilizing LLMs, also in illicit academic usage.

School vs Using In Exam

The cross tabulation shows Herningsholm Gymnasium students are less likely to use LLM during the exams, than Ringkøbing Gymnasium. As 29% of the students from Herningsholm Gymnasium are LLMs during exams, whereas, 44% of the students of Ringkøbing Gymnasium are using it.

School Year vs Using In Exam

The cross tabulation shows that 37% of 1st year students are using LLM during the exam, thereby more frequently as compared to 2nd, and 3rd year students. The usage of the LLMs is seen to decrease by the years. This can be seen as 33% of 2nd year students are using it during exams, whereas 24% of 3rd year students are using it, thereby lowering by grade.

Class Type vs Using In Exam

The cross tabulation shows that natural science students and IT students are less likely to use LLMs during the exams as compared to humanities. This can be seen as 28% of natural students, and 23% of IT students are using it during exams, whereas 38% of the humanities students are using LLMs.

T-test, ANOVA test & Confidence Interval

T-Test was conducted on gender and high school. Where male = 0 and female = 1 and Ringkjøbing Gymnasium = 0, and Herningsholm Gymnasium = 1. The T-Test could not find any significant difference between the usage of LLM in the exam vs gender and school, as the p-value is above 0.05.

Figure 19: T-Test for Gender and School vs Using In Exam (Own creation)

T-Test

Variable	T-Statistic	P-Value	Decision	CI Lower	CI Upper
Gender	0.98	0.33	Fail to reject H0: No significant difference	(0.09)	0.27
HighSchool	1.40	0.17	Fail to reject H0: No significant difference	(0.05)	0.31

The lower T-Test score of gender signifies that there is very low mean difference between the variables, and men are using LLMs more than females during the exams. Also, the positive T-Test score result for high school is also here, which indicates that the mean usage of Ringkjøbing Gymnasium students is more than Herningsholm Gymnasium. Also 0 is included in the confidence interval, as the lower bound is negative. It suggests that there can be no mean difference at 95% confidence level.

Figure 20: Anova Test for School Year & Class Type vs Using In Exam (Own creation)

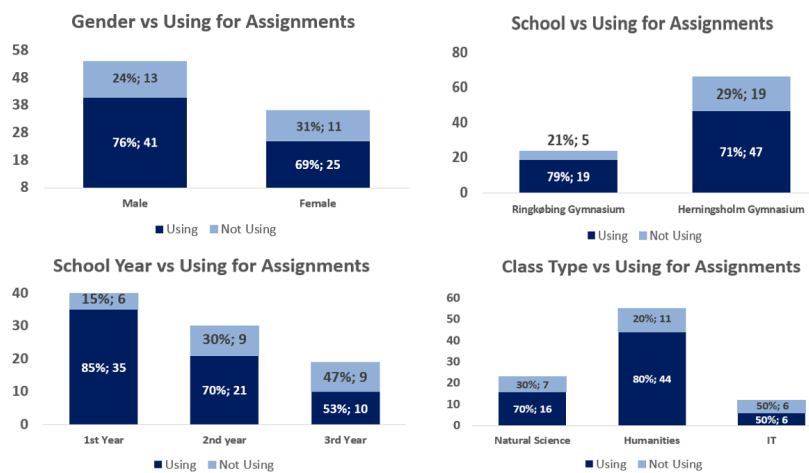
ANOVA							
Variable	F-Statistic	P-Value	Decision	Group - 1 CI		Group - 2 CI	
SchoolYear	1.09	0.30	Fail to reject H0: No significant difference	0.38	0.85	0.60	0.94
ClassType	0.01	0.91	Fail to reject H0: No significant difference	0.67	1.02	0.69	0.98

ANOVA test has been conducted on school year and class type. The ANOVA could not find any significant difference between the usage of LLM in the exam vs school year and class type, as the p-value is above 0.05. The overlapping condition suggests that there is a very low difference in the mean, but 0 mean difference can not be achieved in any condition as the confidence interval is positive. It also implies that the condition is not statistically significant.

4.2.4 Question 4: Using For Assignment

Cross Tabulation and Correlation

Figure 21: Cross Tabulation with Using For Assignment (Own creation)



Gender vs Using For Assignments

As per cross tabulation, men are more likely to use LLMs than women for their assignments. It is seen as only 24% of men are less interested in using it during assignments, whereas women are 31%. Therefore, men are once again more frequently using LLMs outside, during, for exams, and now also for writing assignments.

School vs Using For Assignments

The cross tabulation result shows Herningsholm Gymnasium students are less likely to use LLMs for exams than Ringkøbing Gymnasium students. 71% of the students from Herningsholm Gymnasium are using LLMs for assignments, whereas, 79% of the students of Ringkøbing Gymnasium are using LLMs for assignments.

School Year vs Using For Assignments

From the cross tabulation it is shown that 1st year students are more prone to use LLM. As the usage of the LLM decreases per 2nd year, and 3rd year student in comparison. 85% of the students from 1st year are using it for the assignments whereas 2nd year and 3rd year students are using 70% and 53% respectively.

Class Type vs Using For Assignments

The cross tabulation result shows that, natural science students and humanities students are more likely to use LLM for the assignments. 70% of natural science students and 80% of Humanities are using it in their assignments. Whereas only 50% of IT class students are using LLM in the assignments.

T-test, ANOVA test & Confidence Interval

T-test has been conducted on gender and high school. The t-test could not find any significant difference between the usage of LLM in the exam vs gender and school, as the p-value is above 0.05.

Figure 22: T-Test for Gender and School vs Using For Assignment (Own creation)

T-Test					
Variable	T-Statistic	P-Value	Decision	CI Lower	CI Upper
Gender	0.66	0.51	Fail to reject H0: No significant difference	(0.16)	0.31
HighSchool	0.78	0.44	Fail to reject H0: No significant difference	(0.12)	0.28

Lower T-Test score of gender signifies that there is very low mean difference between the variables, and men are using more LLMs than females for assignments. Also, the positive T-Test score result for high school is also here, which indicates that the mean usage of Ringkjøbing Gymnasium students is more than Herningsholm Gymnasium. Also 0 is included in the confidence interval, as the lower bound is negative. It suggests that there can be no mean difference at 95% confidence level.

Figure 23: Anova Test for School Year & Class Type vs Using For Assignments (Own creation)

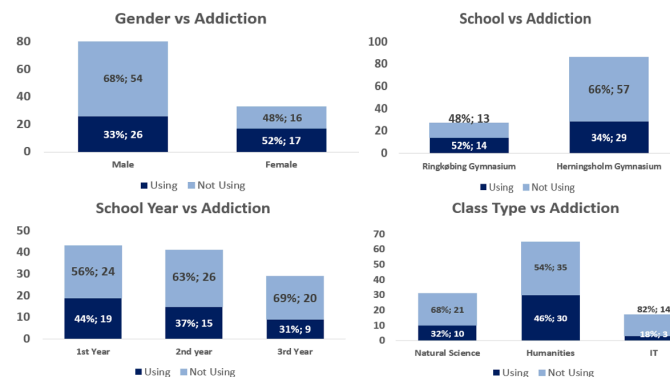
ANOVA							
Variable	F-Statistic	P-Value	Decision	Group - 1 CI		Group - 2 CI	
SchoolYear	7.84	0.01	Reject H0: Significant difference	0.44	0.80	0.79	1.46
ClassType	0.56	0.46	Fail to reject H0: No significant difference	0.71	0.99	0.64	1.28

ANOVA test has been conducted on school year and class type. The ANOVA found the school year and using LLM for the assignment is statistically significant. It is suggesting that one group of students uses LLM a lot more than other groups of students. Also the high positive F statistics suggest the same. Added to that there is very little overlapping between two groups, and it is suggesting that the two groups are statistically different. Also due to the high mean difference with one school year, separating the two groups. However the ANOVA could not find any significant difference between the usage of LLM for the assignment vs class type, as the p-value is above 0.05. The overlapping condition suggests that there is a very low difference in the mean, but 0 mean difference can not be achieved in any condition as the confidence interval is positive. It also implies that the condition is not statistically significant.

4.2.5 Question 5: Using For Addiction

Cross Tabulation and Correlation

Figure 24: Cross Tabulation with Using For Addiction (Own creation)



Gender vs Using For Addiction

As per cross tabulation, men are less likely to be addicted to LLM usage in comparison to women. That means women are getting more dependent on the usage of LLM. 33% of men are addicted to LLM use, whereas 52% of women are addicted in comparison.

School vs Using For Addiction

The cross tabulation result shows Herningsholm Gymnasium students are less likely addicted to LLM use than Ringkøbing Gymnasium. 33% of the students from Herningsholm Gymnasium are addicted to LLM use, whereas 52% of the students of Ringkøbing Gymnasium are addicted in comparison. This may also be due to the fact there is a bigger presence of female students at Ringkøbing, as compared to Herningsholm.

School Year vs Using For Addiction

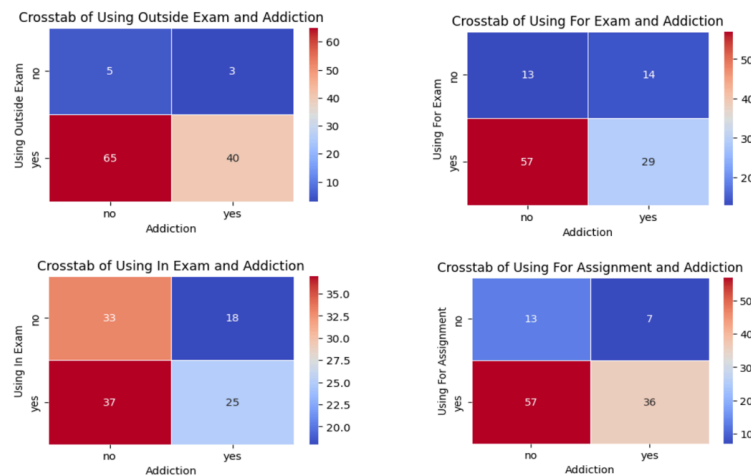
From the cross tabulation it is visible that 1st year students 44% are more prone to be addicted towards the usage of LLM. The usage of the LLM decreases by the 2nd and 3rd year students. It's evident that as a percentage, second year students 37% are prone to be addicted to LLM use, whereas for third year students it is by 31% in comparison.

Class Type vs Using For Addiction

The cross tabulation result shows that Humanities students are more likely to be addicted towards LLM use than the other two class types. 46% of humanities students are addicted to LLM use, whereas 32% of Natural Science Students, and 18% of the IT class students are addicted to LLM use.

Cross Tab with Addiction

Figure 25: Cross Tab Analysis of Addiction with Other Research Questions (Own creation)



Using Outside Exam vs Addiction

There are 40 students who are using LLM outside of the exam for learning purposes as an addiction. 65 students responded that they do not have any addiction but still they are using LLM outside of the exam. On the other hand, 3 students responded that they have addiction but they are not using LLM outside of the exam preparation, whereas, 5 students responded that, neither they have addiction nor they have used LLM outside of the exam.

Using For Exam vs Addiction

There are 29 students who are using LLM for the exam purposes as an addiction. 57 students responded that they do not have any addiction but still they are using LLM for the exam. On the other hand, 14 students responded that they have addiction but they are not using LLM for the exam preparation, whereas, 13 students responded that, neither they have addiction nor they have used LLM for the exam.

Using In Exam vs Addiction

There are 25 students who are using LLM in the exam as an addiction. 37 students responded that they do not have any addiction but still they are using LLM in the exam. On the other hand, 18 students responded that they have addiction but they are not using LLM in the exam, whereas, 33 students responded that, neither they have addiction nor they have used LLM in the exam.

Using For Assignment vs Addiction

There are 36 students who are using LLM for homework, class work and assignments as an addiction. 57 students responded that they do not have any addiction but still they are using LLM for the assignments. On the other hand, 7 students responded that they have addiction but they are not using LLM for the assignments, whereas, 13 students responded that, neither they have addiction nor they have used LLM for the assignment.

T-test, ANOVA test & Confidence Interval

Figure 26: T-Test for Gender and School vs Using For Addiction (Own creation)

T-Test

Variable	T-Statistic	P-Value	Decision	CI Lower	CI Upper
Gender	(1.84)	0.070	Fail to reject H0: No significant difference	(0.34)	0.01
HighSchool	1.62	0.109	Fail to reject H0: No significant difference	(0.03)	0.31

T-test has been conducted on gender and high school. The t-test could not find any significant difference between the addiction vs gender and school, as the p-value is above 0.05. Negative t-test score of gender signifying that there is very low mean difference between the variables and males are using less LLM than females as addiction. The result is not statistically significant but it is close to the significance level. It also implies that the mean average of using LLM for addiction is a bit different than normal. Also, the positive but low t-test score result is also here, which indicates the Ringkjøbing Gymnasium students are using more than Herningsholm Gymnasium. Also the confidence interval includes 0, as the lower bound is negative. It suggests that there can be no mean difference at 95% confidence level.

Figure 27: Anova Test for School Year & Class Type vs Using For Addiction (Own creation)

ANOVA

Variable	F-Statistic	P-Value	Decision	Group - 1		Group - 2	
SchoolYear	1.31	0.25	Fail to reject H0: No significant difference	0.75	1.13	0.53	1.01
ClassType	0.25	0.62	Fail to reject H0: No significant difference	0.73	1.07	0.67	1.00

ANOVA test has been conducted on school year and class type. Where there are three different school years and class type exist. The ANOVA could not find any significant difference between the usage outside of the exam vs school year and class type, as the p-value is above 0.05. The overlapping condition suggests that there is very low difference in the mean, but 0 mean difference can not be achieved in any condition. It also implies that the condition is not statistically significant.

4.3 RAG Application Analysis

The analysis of the RAG application is not available due to legal barriers and user unavailability to test it. Our LLM application was readily available in our testing environment, however the participating HEIs decided not to participate due to their regulations at the last moment. Said institutions requested not to conduct such testing at the premises, therefore, it was not possible to collect the feedback and recommendations from them directly and analyze the application created based on the educational sociology literature as used in HEIs.

5. Findings of the thesis

In the following pages, this thesis will provide an in-depth conclusion of the qualitative and quantitative findings. These sections will cover insights from semi-structured interviews and survey data, discussing the influence of LLMs on learning, academic performance, and the educational format within Danish HEIs.

5.1 - Qualitative findings

The usage of LLMs in the Danish HEI environment has presented both opportunities and challenges. The use of LLMs is seen to be transformative, affecting both how students engage with learning materials, but also how teachers approach teaching. The semi-structured interviews conducted with students and the teaching staff from Ringkjøbing Gymnasium and Herningsholm Erhvervsskole & Gymnasier have provided insights into this, indicating what can be changed to facilitate better use of the technology. The findings found here aim to provide an overview into these insights, highlighting the primary areas of concern and the potential solutions to handle the prevalent misuse of LLMs amongst students.

There are four main areas of focus that have emerged from the analysis: (1) AI alignment (2), prompt engineering (3), ZPD and student independence (3), and the current Ministry of Education guidelines (4). Each area presents challenges that must be addressed to ensure that LLMs can be used as intended.

5.1.1 AI Alignment issues and LLM reliability

One of the four challenges concerning LLM use in Danish HEIs is the issue of AI alignment. It is seen to be the case of an outer-alignment failure, that is to say, LLMs acting contrary to the human values and preferences as seen in the case of Danish HEIs. The outputs of LLMs must align with the educational objectives set by the teachers and individual institutions, but they are not, as seen when interview participants reported several instances where LLMs generated inconsistent or incorrect outputs, leading to confusion and mistrust among users. These issues undermine the reliability of LLMs as educational tools, making it important to improve the LLMs to ensure that they produce accurate and contextually appropriate responses (Bender et al., 2021; Weidinger et al., 2021; Dev et al., 2021). To address these alignment issues, the training processes of LLMs used in education could adopt the HHH (Helpful, Honest, Harmless) framework as shown by Bai et al. (2022), this could be combined with the use of RLHF machine learning where teachers are involved in the feedback process when evaluating the LLMs, although as noted, that may not be feasible due to the resources required (Villalobos et al., 2022). Additionally, LLMs trained predominantly on English data may exhibit biases, as noted by Yan et al. (2024) when Danish HEI students use them in subjects such as in Danish, where the bias may add to the pre-existing pitfalls.

5.1.2 Prompt engineering challenges

The effectiveness of LLMs depends on how well the prompts are crafted. Poorly designed prompts can lead to irrelevant or misleading answers, which can hinder the learning process. This is problematic in subjects that require precise and accurate information, such as mathematics and natural sciences (Yong et al., 2023; Wan et al., 2023). When students do not receive the expected or accurate outputs, it can lead to confusion and reduce their trust in using LLMs as reliable educational tools.

To improve the utility of LLMs, teachers and developers could collaborate in creating prompts tailored to the various class subjects offered in HEIs. Standardizing prompts and providing clear guidelines on their usage could improve the reliability of LLM responses. If the prompts are designed to produce specific and relevant answers, the utility of LLMs in Danish HEIs can be increased. Furthermore, prompt engineering can help mitigate the inherent biases present in LLMs, especially those that arise from training data predominantly in English, which can affect the accuracy of responses in non-English contexts such as Danish HEIs. Training teachers on effective prompt engineering techniques will give them the opportunity to better facilitate LLMs into their education, and combat the illicit use by providing clarity to the “inspiration” and “cheating” dichotomy. This includes understanding how to frame questions and answers in a manner that increases the likelihood of obtaining useful and accurate responses from LLMs, and what ways are deemed to be non-productive as far as generating content that subverts real learning.

5.1.3 Zone of Proximal Development (ZPD) and student independence

ZPD (Vygotsky, 2012) refers to the zone of learning objectives that a student can achieve by acquiring help via another, typically being teachers, however, that can also be an LLM. LLMs can serve as effective scaffolding tools within the ZPD, providing students with the necessary support to accomplish learning new theories that they might struggle with independently. This scaffolding can be beneficial in helping students understand non-spontaneous subjects, that being academic theory. Effective use could lead to it guiding them through difficult problems, and providing feedback that encourages further learning. However, the risk lies in students becoming overly dependent, in some cases addicted (Bai et al., 2023) on the use to circumvent their own academic performance. This can hinder their ability to perform tasks independently and develop critical thinking, and the ability to problem-solve in their academic work. This technology has the potential to help students achieve ZPD in the absence of teachers. For students who may not have immediate access to teacher support, LLMs can fill this gap by providing assistance and resources. This is particularly relevant for students with learning difficulties and those from poor socio-economic backgrounds who may lack external academic support. By leveraging LLMs, these students can receive individualized guidance that helps them progress in their studies. However, this benefit is contingent on the LLMs being used as intended educational aids.

The thesis participants highlighted that in many cases, students misuse LLMs, relying on them to complete entire assignments without engaging in the learning process, which undermines the intended educational purpose of these tools.

5.1.4 Ministry of Education guidelines and regulatory challenges

The current guidelines & rules from the Ministry of Education primarily address the use of LLMs in terminal exams, and study area projects, but they lack rules for their use in everyday classroom activities. This regulatory gap creates a void where students misuse LLMs for assignments and homework, leading to academic dishonesty. Teachers expressed frustration over the lack of clear policies, making it difficult to enforce rules and maintain academic integrity. The absence of specific guidelines for non-exam settings means that students can exploit these tools without proper oversight, undermining the educational process and fostering dependency on LLM-generated content. Likewise, because of the over abundance of use of LLMs this had led Danish HEIs taking pragmatic decisions to either avoid addressing the use, or evaluating what they can where they deem it has not been produced via LLM-use. This lack of ruling can be seen in Appendix CC below, whereas this covers the specific ruling on exams, but only exams:

Figure 28: Paragraphs concerning regulations on digital help tools in exams from Appendix CC (UVM, 2024)

Eksempler på ikke-tilladte digitale hjælpemidler:

- ChatGPT
- Google-søgning
- Tilføjelsesprogrammer til Word, der selv genererer tekst

During the thesis writing, an expert group issued recommendations regarding LLMs specifically in HEIs, and other secondary institutions in Denmark. One notable gap in those recommendations from the expert group (Appendix R) is the lack of detailed guidance on LLM use within classroom settings, whereas it focuses more on changing current exam practice. According to the expert group's report, their focus was primarily on how students can be examined in light of technological advancements, with recommendations largely centered on exam scenarios.

However, the report states that broader discussions about the use of AI in everyday learning, including classroom interactions and homework assignments, fall outside their scope. This oversight leaves the teachers without the necessary recommendations to manage LLM use effectively during regular class activities, where misuse can be most prevalent. This can be observed in the following insert from the appendix:

Figure 29: Paragraph from the recommendations issued by the expert group concerning recommendations regarding AI in education from Appendix R (Source: UVM, 2024)

- Gymnasiale uddannelser: Kunstig intelligens indgår i sammenhæng med de gymnasiale uddannelsers fokus på digital dannelse. Det er vigtigt, at der i et fag sker en løbende indfasning af forskellige tilgange til kunstig intelligens gennem fagets forløb.

This is also observed by Teacher B who voiced his concerns regarding the recommendations issued by the expert group, noting the lack of clarification concerning the general use of LLMs in class. This can be observed by Teacher B, timecode [26:33], (O: Interview transcript Teacher B - ENG) A ruling is needed to standardize the use of LLMs across all educational activities, not just in terminal exams and larger academic projects like SOP/SRO/SRP. Clear guidelines and rules should be developed to help teachers and students navigate the use of LLMs. This should include specific rules for assignments, homework, and classroom activities to ensure consistent and fair application of LLM technology, and be provided in written format as there is evidence to suggest that students currently do not acknowledge it otherwise, as observed by Head of Education B, timecode [14:05] (Interview transcript Head of Education B - ENG) Additionally, even with said rules implemented there is the issue of addressing the misconduct as the responsibility is placed on the individual teachers. This leads to situations where teachers experience displeasure in having this responsibility as they feel that it is not something they are either thrilled to do, and that it is observed to lead to a polarization between student-teacher relationships. This responsibility of having to detect, as well carrying out sanctions needs to be addressed as it seems to be informal in nature, and something the individual teacher is unable to bear. This can be observed in the following interview with Teacher A, timecode [14:14] (Interview transcript Teacher A - ENG) Integrating learning about LLMs into the curriculum could thus be another step towards a solution. Students should learn about the potential and limitations of these tools, ensuring they use them responsibly and effectively.

5.2 - Quantitative findings

The quantitative findings will be based on the dependent variable and its relation with the target variable. The project's target variables are the five questions that have been asked while surveying. The findings will go step by step with each dependent variable and will extract the content that has been found in the analysis part. Cross tabulation, T-Test and ANOVA test result will be reflected on the findings. These analyzed results correctly match with our problem statement, and can complement the research objectives in regards to ascertain answers to our research questions. The findings are described below for each of the dependent variables.

5.2.1 Gender wise LLM usage

The usage of LLM and the addiction to it varies highly gender wise, but four out of five cases it does not vary statistically. Both male and female have their own tendency to use it. The findings are:

☐ Motivation

- Men are using LLM for exams significantly higher in comparison to women, it suggests male students have stronger motivation to prepare for exams via LLM use.
- Despite that in all other cases the variables are not statistically significant, the evidence of high usage and addiction prove that strong motivation for LLM use goes for both the genders.

☐ Positive and Negative Consequences

- The positive findings indicate that the majority of the students are using LLM outside of the exam for learning purposes. They prefer to learn more with external knowledge.
- On the other hand 62% of the students are using LLM for the exam preparations, and 33% of the students are using LLMs during the exam. The level of usage is not the same but it shows the degree of extremity of using LLM is increasing. As they started to use it during the exam as well, even after perhaps making use of LLMs in exam preparations. It also shows the extreme misuse of LLM in HEI education as there are $\frac{1}{3}$ of students using it during the exams, indicating a significantly large number of cheatings plagiarizing even in conditions supposedly safe-guarded by existing exam regulations and measurements.

- There is a high amount relating to writing assignments, which corresponds with students wanting to generate ideas, contents and ways of solving their assignments in order to minimize effort and time spent upon it. But this breeds dependency, which is limiting their academic abilities, critical reasoning, and overall academic merit of their own work.
- It is also seen that dependency on LLMs is high among the students, especially female students. As men are clearly overrepresented in all contexts of LLM use, it indicates that of those female students who are using it, they are more liable to become addicted to the use, as seen with Student G in the interviews. Thus while not all females are using it, a subset of female students are speculated to use it to an extreme degree, thus dependent. As such, amongst those females who have used LLMs, they are liable to have high dependency.
- Added to that, there are some students who have an addiction and use LLM but are reluctant to disclose it. Which is also noticeable in the analysis segments and the proportion is also quite high.

☐ **Role of LLM**

- LLM use is seen by both genders of students. High usage rate by both genders among all the contexts of the quantitative data reveals that LLMs are therefore influencing the educational system.

5.2.2 School wise LLM usage

The analysis shows the HEI-wise comparison. The two different HEI students are adopting LLM almost similarly. One school's students are more featured in some aspects, whereas other school is likewise featured in other aspects, but not significantly statistical. This is seen below:

☐ **Motivation**

- In both HEIs, the students have strong motivation to prepare for the exam with LLM use. This can be seen in both cases, as almost 88% of HEI's students are using it. This indicates that the LLM use has been adopted as a way to prepare for exams as status quo.

- While the variables are not statistically significant, the evidence of high usage and addiction prove that strong motivation works for both the HEIs. The addiction rate associated with the LLM use can therefore indicate the reasons for the usage.

☐ **Positive and Negative Consequences**

- The positive side is that the majority of the students (almost 88%) from each HEI are adapting LLM technology into their methodology to access learning, as in outside of the exam contexts.
- On the other hand 62% of the students are using LLMs for the exams, and 33% of the students are using LLM during the exams. The level of usage is not the same amount, but it shows the degree of extremity differs on more severe cases of plagiarization.
- The high number of LLM usage to complete the assignments are almost two-third of the students from both HEIs (Ringkøbing Gymnasium 79%, and Herningsholm Gymnasium 71%) The tendency is not differing too much in terms of HEIs, however too much dependency on LLM use indicative of students limiting their analytical ability, critical reasoning, and ability to conduct education as intended.
- The addiction of using LLMs is high among the HEI students. Especially Ringkøbing Gymnasium students. Here it is seen students are more prone to be addicted to using LLMs. This may be explained by the gender-ratio being different as there are more female students at Ringkøbing, as compared to Herningsholm. But likewise, 44% of Ringkøbing students are using it during the exams, as compared to 29% by Herningsholm, so higher addiction coupled with a bigger tendency to use LLMs during exams is noteworthy deviation.

☐ **Role of LLM**

- LLM use is seen in both the HEIs. It is changing the behavior of the student and their approach regarding the education. High usage rate by both HEIs among all the contexts of use indicates that LLM is playing a role in education.

5.2.3 School Year wise LLM usage

The result gets significantly changed, when the school year is increasing. Three different school year students showed three different aspects of the usage and addiction levels. This is due to several reason, whereas the findings are seen below:

☐ Motivation

- Almost all 1st year students with the exception of one are using LLMs. However, while the usage is lesser amongst 2nd and 3rd year students, here the majority of the students are also utilizing LLMs. This indicates there is a strong utilization across all grades.
- This is subsequently also seen in using LLM for their assignments, whereas again the students are using it to a lesser degree in 2nd and 3rd years as compared to 1st year students, whereas the usage lowers correspondingly by 2nd, and 3rd year students.
- Even in all other aspects, the variables are not statistically significant, but there is evidence of high usage and addiction that prove that high motivation of LLM use for all the school years, this is coupled with addiction rates that progress similarity to usage rate.

☐ Positive and Negative Consequences

- The positive attributed effect is that the majority of the students from each school year are making use of LLMs to learn outside of the exam for learning purposes, and as the year progresses, they start to decrease the usage. This could be either due to using LLMs less frequently as they advance to the new school year, or that 1st year students are more accommodated with the new technology of LLM in their “formative years”.
- 62% of the students are using LLM for the exam, whereas 33% of the students are using LLMs during the exam. The level of usage is not the same, but it shows the degree of extremity of using LLM is increasing. This shows students may have started to use it in the exam as well, even using it for exam preparations. It also shows the extreme misuse of LLM in the field of education as $\frac{1}{3}$ are using it during the exams.
- There is a high amount of usage to complete the assignments being $\frac{2}{3}$ of the students from the 1st year, and above $\frac{1}{2}$ of the students from 2nd and 3rd year students, which shows the use is seen throughout the years. Therefore, there is minimal difference in terms of the students' grade. However, too much dependency may lead to LLMs limiting

their analytical ability, critical reasoning, and ability to engage independently in their education. This is compounded by the fact that HEIs are either unable to detect, or effectively punish LLM use. Lastly, the group is statistically significant, which means the difference in the usage from first year to rest of the year are quite visible.

- The addiction of using LLMs is high among each year of school students. The students are extremely addicted to using LLMs therefore. Because of this, higher addiction numbers correspond to the fact that most of the students are using LLM as an assistance to circumvent the difficulty of their work, but that in turn reduces their ability to learn.

☐ **Role of LLM**

- LLM impacts different school years differently. As school years go up, the students reduce the use of LLM in all aspects. Thus, the role of LLM differs highly when students get more habituated with the LLM.

5.2.4 Class Type wise LLM usage

There are three different class types and every type possesses different aspects of the usage. These three different class types make use of LLMs differently in various situations, but are not statistically significant in nature. The findings can be seen below:

☐ **Motivation**

- Almost all the class types show that its students are using LLMs to learn outside of the exams for learning. More than 80% students from all three class types are indicating this usage.

☐ **Positive and Negative Consequences**

- The positive side is that the majority of the students from each class type are adapting LLM to learn outside of the exam for learning purposes.
- However, 62% of the students are using LLM for the exams, and 33% of the students are using LLM during the exam. The level of usage is not the same but it shows the degree of extremity of using LLM is increasing. As they started to use it in the exam as

well, even after taking the preparation through it. Also, it shows the extreme misuse of LLM in the field of education while they are using it for the exam.

- More than 2/3 of the students from humanities classes are using LLMs. As the humanities is a descriptive subject, where writing is primarily the method to complete most of the assignments, it thus is showing negative consequences where the students from this particular class type are using more, whereas qualitative data from teachers indicates this is more of a negative, then positive as far as learning outcome for the students, as the difficulty is apart of the learning process, as seen with language-related subjects.
- The addiction of using LLMs is high among each year of class type students. But humanities students are more addicted in regards to LLM use. However, this is also visible in all other results.

☐ **Role of LLM**

- LLM impacts different class types differently. As class types are changing the students' behavior, this is likewise changing as well, but it is not statistically significant.

6. Discussion

The introduction established the importance of understanding the impact of LLMs on education within Danish HEIs. The thesis' initial findings indicated that students primarily use LLMs to improve the quality of their assignments, enhance grammatical accuracy, and organize their work more effectively. These observations align with the introductory discussion on the potential of LLMs in education, suggesting an influence on how students engage with their coursework. The qualitative data from interviews highlighted that students appreciate the instant feedback and support provided by LLMs, which helps them meet academic standards and deadlines. This aligns with findings by Kumar et al. (2023), who noted that structured guidance in LLM-assisted learning enhances problem-solving and engagement. The analysis further revealed that LLMs provide benefits to students from disadvantaged backgrounds and those with learning disabilities. By offering support that might otherwise be unavailable, LLMs help level the educational playing field. This finding reinforces the benefits of using LLMs within education, highlighting how they can offer better learning opportunities for a broader range of students.

Qualitative data from teachers pointed out that students who previously struggled with language barriers or learning difficulties such as dyslexia, could now better engage in academic activities, thus creating a more inclusive learning environment. This corresponds with Van Wyk et al. (2023), who highlighted the potential of LLMs to support personalized learning and reduce teachers' workloads, making education more accessible and equitable.

However, the thesis also identified challenges associated with LLM usage, including dependency and misuse. Quantitative data indicated a worrying trend where a notable percentage of students reported using LLMs to complete their assignments without engaging with the material. This reliance undermines the learning process, leading to reduced understanding rather than own engagement with course content. Also it was evident from the findings that students are reluctant to show that they have an addiction as a dependency, even though they are using highly in all the aspects. These issues underscore the need for immediate solutions and established rules to manage LLM use effectively in education. The ease of access to these tools often leads to academic dishonesty, undermining the integrity of student work and the reliability of traditional assessment methods, such as written assignments, which are increasingly unable to serve as accurate measures of student learning. Perkins (2023) discussed similar concerns, emphasizing the importance of clear academic integrity policies in managing the ethical use of LLMs.

In conclusion, the findings highlight both the opportunities and challenges presented by LLMs in Danish HEIs. While these LLMs offer substantial support for learning and accessibility, they also pose significant risks related to misuse and over-reliance. The thesis calls for further research into this phenomenon and the adoption of comprehensive rules to ensure that LLMs are used as intended, as a supplement to learning, rather than a gateway to plagiarism and misuse. By addressing these concerns, teachers can harness the potential of LLMs to be a positive force in education, rather than a tool that diminishes the very learning it aims to support. This balanced approach will help integrate LLMs into educational practices in a way that boosts learning outcomes while maintaining academic integrity. These findings are consistent with those of Zhou et al. (2024), who emphasized the need for responsible use and continuous evaluation of LLMs to mitigate risks and maximize their educational benefits.

7. Conclusion

This thesis provides answers to the three research questions that we used to structure our thesis upon. Here we can identify both the motivations behind LLM usage, the subsequent positive and negative outcomes, and clarify the current role LLM holds in HEI education, and should ideally hold. This section will clarify how LLMs influence Danish HEIs practices, and highlight key areas where this is most pronounced. However, the following is how we answer the research questions:

→ RQ1: What is the Danish HEI student's motivation for using LLMs?

The motivation for using LLMs in Danish HEIs stems from the benefits these models offer in facilitating access to knowledge and support when doing assignments and homework, particularly in humanities and languages, where LLMs seem to surpass traditional translation software like Google Translate. Students are drawn to LLMs because they provide immediate responses and information that can improve their understanding of otherwise complex subjects. This immediacy is appealing in education, where time constraints and the need for learning aids are prevalent. As highlighted in the semi-structured interviews conducted, students appreciate the ability of LLMs to break down academic topics into smaller segments, making learning more accessible and less difficult. Quantitative findings support this. This can be seen as 88% of the students from Herningsholm Gymnasium are using LLMs for learning outside of exams, whereas 87% of the students of Ringkøbing Gymnasium are using LLMs outside of exams, thus confirming the qualitative insights. However, this convenience leads to over-reliance, where students bypass traditional learning methods in favor of using LLMs, in turn reducing the time they spend on dealing with it independently, reducing the ability to engage and learn by themselves. This over-reliance undermines the development of academic skills integral to learning in Danish HEI education practice. Therefore, while the motivation for using LLMs is clear and supported by both qualitative and quantitative data, it must be addressed how this usage can be balanced to prevent detrimental effects on student learning. Establishing rules and promoting the responsible use of LLMs can help mitigate these issues, ensuring that students benefit from these technologies without compromising their educational development. In turn, also dealing with the subsequent reliance on LLM use that is highly prominent in the data.

→ RQ2: What are the positive and negative consequences associated with the usage of LLMs within the context of Danish HEI?

The benefits of LLM in Danish HEIs are double-edged. Primarily, LLMs serve as tools that can supplement teaching, providing additional resources for students when dealing with various subjects. For students, LLMs can act as a tutor, offering explanations and understanding that may not be readily available from textbooks or classroom lectures. This supplementary role is beneficial in supporting student learning, particularly for those who may not have access to academic resources or to a teacher. Even more so, for students who are from under-privileged backgrounds where there is less of a family support structure that can aid them in their learning, and/or if students are dealing with learning disabilities, as with dyslexia as noted in the interviews. However, the challenges associated with LLMs cannot be overlooked. One main issue is the potential for academic dishonesty, as students use LLMs to generate assignments and essays without engaging with the material themselves. In the quantitative findings this can be seen as 29% of the students from Herningsholm Gymnasium are LLMs during exams, whereas, 44% of the students of Ringkøbing Gymnasium are using it. This misuse not only affects the integrity of their work but also hinders their own potential to learn, and their potential to develop their academic skills they may require in higher education in later onwards, such as in university, etc. Moreover, there are concerns about the accuracy and bias of LLM-generated content, especially when models are predominantly trained on English-language data. This can lead to discrepancies and misunderstandings when applied in non-English educational contexts, such as Danish HEIs, further complicating the use of LLMs.

→ RQ3: What role should LLMs play in the Danish HEI environment?

The ideal role of LLMs in Danish HEIs should be as an aid that supports and complements traditional teaching methods, rather than replacing them. To achieve this, HEIs must develop guidelines and rules for LLM usage in their institutions. These guidelines should focus on promoting ethical academic utility, ensuring that students and teachers understand how to make use of LLMs effectively without compromising academic integrity.

Teachers should be involved in creating prompts and structuring interactions with LLMs to add to the educational value while minimizing the risk of misuse. Quantitative data shows that 36% of male students are using it during exams, whereas for women it is 28%. Furthermore, 69% of male students use LLMs for exam preparation compared to 49% of female students, suggesting that LLM usage patterns might vary across different demographics, necessitating further research into the factors behind this differing distribution.

Additionally, integrating learning about LLM utility into the curriculum can help students use it more responsibly. By educating students on the ethical implications and proper ways for LLM use, HEIs can ensure that LLMs contribute to learning rather than preventing it. This can help students develop critical thinking when engaging with the opportunities present in LLMs, ensuring that they benefit from its potential while maintaining academic honesty. Cross tabulation results also show that 97% of first-year students use LLMs, with usage slightly lower among second and third-year students at 78% and 89%, respectively, indicating a need for early and continuous guidance on responsible use.

In conclusion, addressing the problem formulation involves understanding the motivations behind LLM usage, the benefits and challenges it presents, and the role it should play in education. Implementing clear sets of rules, incorporating learning about LLMs into the curriculum, and ensuring proper academic use can help mitigate the negative consequences while maximizing the benefits, making it so that LLMs can serve as educational aids in Danish HEIs.

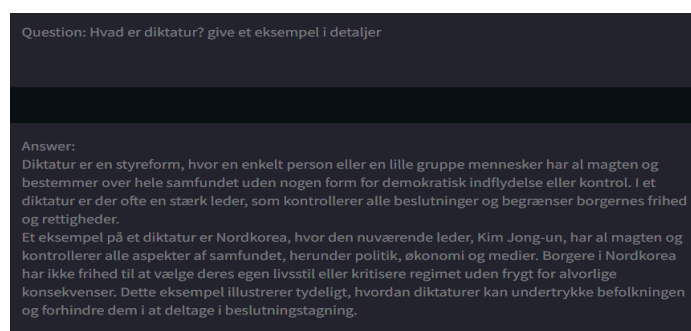
7.1 Reflections

Students are getting more and more dependent on the LLM. From assignments, homework, class work to exams, it is seen that in several instances students are using LLMs in such situations. This was observed in both the quantitative, and qualitative data and subsequent analysis. The current usage of LLMs is beyond the control level of Danish HEIs presently, as students are using it not only for ordinary homework, but also during the exams as well. It is seen that in some cases students have developed an addiction to using it, whether they are aware of it or not. This can be demonstrated in the quantitative analysis as there evidently was a lot of false negative answers. In this case, students said that they are not addicted, but the large proportion is

using LLM inside or outside of the exam. The RAG application built on the social study book is catered to aid the students. This application is solely focused on the social study book's content, thereby no external knowledge. The application has been built in a way so that the LLM can not use its general knowledge outside of the fine-tuned training, as to avoid the negative situations where students using LLM do not gain class-specific book related knowledge.

This is because said LLM does not know the contents of the learning books of the students, instead it contains a large general corpus of information. Thus the RAG application built in this thesis can reflect the intended learning. Here, students will get the best possible answer from the book with an example, if the user asks for any. Added to that, this sort of application will help the students build their critical reasoning ability. This can be seen as when students interact with this application, they will do so knowing it is based on the learning material of their class. Likewise, the RAG application strives to gain the best possible answers from the best possible related chunks in simple language. But a book may contain difficult language, or the topic may not be comprehensible by the students. Here, RAG can come into action. When a user asks for an answer to a given topic, then RAG can provide answers in simple language. This can further assist comprehending class material as it is easier to comprehend in said format. The application also shows the reference from where it took the answer, thereby helping the student to investigate further on the topic in question. As it retrieved five best possible chunks, thus students can go to that best possible source to learn more. However, it was not possible to test the application in the testing environment due to regulations and protocols, so the benefits are not empirically proven. But the RAG application in this thesis produces very good outputs by retrieving the content from the book. This can be demonstrated with a question which has been asked to the application:

Figure 30: Illustration of RAG Functionality (Own creation)



The answer produced is very detailed, and the example it took exactly from the book. Thus, this sort of RAG applications are helpful in nature for the students for more interactive learning. Also, this sort of application accommodates students who are heavily involved in using LLM for their study purposes. Overall, the positivity reflects from the usage of this sort of application.

7.2 Limitations

This thesis encountered several limitations, divided equally between qualitative and quantitative aspects, impacting the depth of the findings involved. On the qualitative side, the main limitation was the potential bias introduced through self-reported data. Interviews with students and teachers may have been influenced by social desirability bias, where respondents provide answers they believe are expected rather than their genuine experiences and opinions. Additionally, the semi-structured interview format, while flexible and insightful, might have led to inconsistencies in the data due to varying interview dynamics and the different levels of engagement and articulation among participants. Furthermore, the reliance on a limited number of high schools in Denmark restricts the generalizability of the findings. The institutional policies specific to these schools might not reflect broader trends and experiences in other regions.

On the quantitative side, the thesis likewise faced limitations. Firstly, the evaluation of students' grades before and after using LLMs was not available. Access to longitudinal data tracking students' performance before/after written assignments would have provided a more detailed overview of LLMs' impact on academic outcomes. Secondly, the sample size for the quantitative surveys was relatively small, limiting the ability to achieve statistical significance and generalize findings across the broader student population. Larger sample sizes would have enabled more precise estimations and stronger inferential statistics, thus improving the validity of the conclusion. Lastly, the inability to conduct real-time evaluations of the MVP LLM agent hindered the thesis's capacity to observe and analyze student interactions with LLMs in a controlled environment. Real usage data could have offered valuable insights into students' usage patterns and the practical challenges faced when integrating LLMs into academic activities. This limitation underscores the need for future studies to incorporate dynamic and real-time assessments to better capture the nuanced impacts of LLM technology on student learning and engagement.

In summary, while the thesis provides valuable insights into the role of LLMs in Danish high schools, the qualitative and quantitative limitations highlight areas for improvement in future research. Addressing these limitations through larger, and more diverse samples would ideally be better, as to provide a more encompassing understanding of LLMs' educational impacts, and accounting for regional HEI differences.

7.3 Recommendations

It is recommended that HEIs develop guidelines and rules for the use of LLMs outside of exams. The rules need to ensure academic integrity and outline clearly when and how LLMs can be used. Additionally, institutions should consider incorporating classes on digital literacy. LLMs should be encouraged to be used as supplementary tools, rather used to produce academic output. Likewise, the current way that LLMs are detected, and likewise sanctioned needs to be addressed. Currently there is a seemingly “self reinforcing loop” when it comes to that, with HEIs pressured not to address the LLM use out of fear of losing students, and likewise looking the other way in some cases based on pragmatic reasons. The responsibility is likewise laid on the individual teachers, where because of time constraints, or the sensitivity of the matter, they are unable to act in that capacity. This needs to be addressed, as otherwise the methods of enforcing the rules are likely ineffective if they were to be created. To avoid the negative pitfalls associated with the current way LLMs are used, it could perhaps be the case that the Ministry of Education created a LLM platform intended to be used for educational content so that students do not use third party LLM applications. This is because LLMs can be utilized beneficially, as seen with students experiencing learning disabilities, or are lacking structure in their residence to support their growth. Added to that, after developing said platform or agent, there could be added a feature in which the aforementioned LLM can differentiate between the content generated by itself, and the third party LLM applications. This will give the HEIs the ability to detect students who are using unsanctioned LLMs. This can aid the current gap that is existing as far as detection goes, as individual teachers lack the proper methods in monitoring the use, and likewise ensuring the LLM is used as a supplement, and not in place of the student.

The last matter pertains to our quantitative findings. As seen, we found indications that addiction is tied to the variable of gender, same with LLM use. Whereas men are more prone to overall usage, it contrastingly reveals women are more liable to have dependency on the LLM use. While this was outside of the scope of this thesis, further research into the underlying factors for this addiction pattern could likewise further explain the greater factors behind the dependency. In this thesis, we did not aim at explaining the addiction, so further research would be needed.

6. References

1. Abedi, M., Alshybani, I., Shahadat, M. R. B., & Murillo, M. (2023). Beyond traditional teaching: The potential of large language models and chatbots in graduate engineering education. *Qeios*.
2. Ahad, N. A., & Yahaya, S. S. S. (2014). Sensitivity analysis of Welch's t-test. *AIP Conference Proceedings*. <https://doi.org/10.1063/1.4887707>
3. Aithal, P. S., & Aithal, S. (2023). Application of ChatGPT in higher education and research—A futuristic analysis. *International Journal of Applied Engineering and Management Letters (IJAEML)*, 7(3), 168-194.
4. Amini, M., Abbaspour, K. C., & Johnson, C. A. (2010). A comparison of different rule-based statistical models for modeling geogenic groundwater contamination. *Environmental Modelling and Software*, 25(12), 1650–1657. <https://doi.org/10.1016/j.envsoft.2010.05.014>
5. Anderson, P. W. (1972). More is different: Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047), 393-396.
6. Ascough, R. S. (2002). Designing online distance education: Putting pedagogy before technology. *Teaching Theology and Religion*, 5(1), 17-29. <https://doi.org/10.1111/1467-9647.00114>
7. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., & Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.
8. Azaza, M. B. M. (2018). Investigating teacher professional learning: A case study of the Abu Dhabi new school model (Doctoral dissertation, University of Leicester).
9. Bai, L., Liu, X., & Su, J. (2023). ChatGPT: The cognitive effects on learning and memory. *Brain-X*, 1(3), e30.
10. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S.,


- Olah, C., Mann, B., & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
11. Bandy, J., & Vincent, N. (2021). Addressing "documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. arXiv preprint arXiv:2105.05241.
 12. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
 13. Blair, G., Imai, K., & Zhou, Y. Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511), 1304-1319.
 14. Blondal, K. S., & Adalbjarnardottir, S. (2012). Student disengagement in relation to expected and unexpected educational pathways. *Scandinavian Journal of Educational Research*, 56(1), 85-100.
 15. Bonner, Lege, and Frazier (2023) discuss practical ideas for teaching with large language model-based artificial intelligence in the language classroom.
 16. Bowling, A., & Ebrahim, S. (2005). Quantitative social science: the survey. In *Handbook of health research methods: Investigation, measurement and analysis* (pp. 190-214).
 17. Boyle, R. W., & Farreras, I. G. (2015). The effect of calculator use on college students' mathematical performance. *International Journal of Research in Education and Science (IJRES)*, 1(2), 95-100.
 18. Bridgeman, B., Harvey, A., & Braswell, J. (1995). Effects of calculator use on scores on a test of mathematical reasoning. *Journal of Educational Measurement*, 32(4), 323-340.
<https://doi.org/10.1111/j.1745-3984.1995.tb00470.x>
 19. Brill, E. D. (1993). A corpus-based approach to language learning - University of Pennsylvania ProQuest Dissertation & Theses, 1993. 9331757.
 20. Brooks, L., Swain, M., Lapkin, S., & Knouzi, I. (2010). Mediating between scientific and spontaneous concepts through languaging. *Language Awareness*, 19(2), 89-110.
 21. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

22. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.
23. Carlsmith, J. (2022). Is Power-Seeking AI an Existential Risk? arXiv preprint arXiv:2206.13353.
24. Chaiklin, S. (2003). The zone of proximal development in Vygotsky's analysis of learning and instruction. In *Vygotsky's Educational Theory in Cultural Context*, 1(2), 39-64.
25. Chaudhry, I. S., Sarwary, S. A. M., El Refae, G. A., & Chabchoub, H. (2023). Time to revisit existing student's performance evaluation approach in higher education sector in a new era of ChatGPT—a case study. *Cogent Education*, 10(1), 2210461.
26. Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. arXiv preprint arXiv:2310.14735.
27. Chen, Q., Kong, Y., Gao, W., & Mo, L. (2018). Effects of socioeconomic status, parent-child relationship, and learning motivation on reading ability. *Frontiers in Psychology*, 9, 348846.
28. Chen, Q. Du J, Hu Y, Keloth VK, Peng X, Raja K, et al. (2023) Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. arXiv preprint arXiv:230516326.
29. Cheng, C. I. Chow, L, P. Rider, R. V. (1972). The randomized response technique as used in the Taiwan outcome of pregnancy study. *Studies in Family Planning*, 3(11), 265-269.
30. Cheong, I., Caliskan, A., & Kohno, T. (2023). Envisioning legal mitigations for intentional and unintentional harms associated with large language models (Extended abstract). In *Proceedings of the 1st Workshop on Generative AI and Law, International Conference on Machine Learning*
31. Chew, E., & Chua, X. N. (2020). Robotic Chinese language tutor: personalising progress assessment and feedback or taking over your job?. *On the Horizon*, 28(3), 113-124.

32. Chhabra, A., Dass, B. K., & Gupta, S. (2016). Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question RRT model. *The North Carolina Journal of Mathematics and Statistics*, 2, 1-9.
33. Chiu, T. K., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 4, 100118.
34. Clark, K. R. (2018). Learning theories: behaviorism. *Radiologic Technology*, 90(2), 172-175.
35. Clark, L., Birkhead, A. S., Fernandez, C., & Egger, M. J. (2017). A transcription and translation protocol for sensitive cross-cultural team research. *Qualitative Health Research*, 27(12), 1751-1764. <https://doi.org/10.1177/1049732317726761>
36. Clarke, R. (1994). Asimov's laws of robotics: Implications for information technology. Part 2. *Computer*, 27(1), 57-66.
37. Comte, A. (1858). *The positive philosophy of Auguste Comte*. Blanchard.
38. Dev, S., Sheng, E., Zhao, J., Amstutz, A., Sun, J., Hou, Y., Sanseverino, M., Kim, J., Nishi, A., Peng, N., & Chang, K.-W. (2021). On measures of biases and harms in NLP. *arXiv preprint arXiv:2108.03362*.
39. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv.org*. <https://arxiv.org/abs/1810.04805>.
40. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P., Lomeli, M., Hosseini, L., & Jégou, H. (2024, January 16). The Faiss library. *arXiv.org*. <https://arxiv.org/abs/2401.08281>
41. Dung, L. (2024). The argument for near-term human disempowerment through AI. *AI & Society*. <https://doi.org/10.1007/s00146-024-01930-2>
42. Egholm, L. (2014). *Philosophy of science: Perspectives on organizations and society*. Hans Reitzels Forlag.
43. Fani, T., & Ghaemi, F. (2011). Implications of Vygotsky's zone of proximal development (ZPD) in teacher education: ZPTD and self-scaffolding. *Procedia - Social and Behavioral Sciences*, 29, 1549-1554.

44. Firt, E. (2023). Calibrating machine behavior: A challenge for AI alignment. *Ethics in Information Technology*, 25(1), 42. <https://doi.org/10.1007/s10676-023-09716-8>
45. Fosnot, C. T., & Perry, R. S. (1996). Constructivism: A psychological theory of learning. In *Constructivism: Theory, perspectives, and practice*, 2(1), 8-33.
46. Frederiksen, C. L. & Kureer, H.,. (2008). *Samfundsfag C* (2nd ed.). Systime.
47. Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.
48. Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better for few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 3816-3830).
49. Gingerich, D. W. (2010). Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis*, 18(3), 349-380.
50. Golde, J., Haller, P., Hamborg, F., Risch, J., & Akbik, A. (2023). Fabricator: An Open Source Toolkit for Generating Labeled Training Data with Teacher LLMs. *arXiv preprint arXiv:2309.09582*.
51. Gossman, D., & Kannan, H. (2023). A systems theoretic perspective of the outer alignment problem. *EasyChair Preprint no. 11431*.
52. Hasanein, A. M., & Sobaih, A. E. E. (2023). Drivers and consequences of ChatGPT use in higher education: Key stakeholder perspectives. *European Journal of Investigation in Health, Psychology and Education*, 13(11), 2599-2614. <https://doi.org/10.3390/ejihpe13110181>
53. Hedderich, M. A., Bazarova, N. N., Zou, W., Shim, R., Ma, X., & Yang, Q. (2024). A piece of theater: Investigating how teachers design LLM chatbots to assist adolescent cyberbullying education. *arXiv preprint arXiv:2402.17456*.
54. Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, 4(5), 5947. <https://doi.org/10.4249/scholarpedia.5947>
55. Hugging Face. *sentence-transformers/all-MiniLM-L6-v2* · (2022, January 18). <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

56. Irshad, S., Maan, M. F., Batool, H., & Hanif, A. (2021). Vygotsky's Zone of Proximal Development (ZPD): An Evaluative Tool for Language Learning and Social Development in Early Childhood Education. *Multicultural Education*, 7(6), 234.
57. Jackson, P. W. (1986). *The practice of teaching*. Teachers College Press.
58. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38.
59. Jopling, M. (2019). Using quantitative data. In *Practical research methods in education* (pp. 55-66). Routledge.
60. Kalyan, K. S. (2024). A survey of GPT-3 family large language models including ChatGPT and GPT-4. *Natural Language Processing Journal*, 6, 100048.
<https://doi.org/10.1016/j.nlp.2023.100048>
61. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
62. Karpov, Y. V. (2003). Vygotsky's doctrine of scientific concepts: Its role for contemporary education. In *Vygotsky's Educational Theory in Cultural Context* (pp. 65-82).
63. Kästner, L., & Crook, B. (2023). Explaining AI through mechanistic interpretability.
64. Kaur, D., Uslu, S., Durresi, M., & Durresi, A. (2024). LLM-based agents utilized in a trustworthy artificial conscience model for controlling AI in medical applications. In L. Barolli (Ed.), *Advanced Information Networking and Applications (AINA 2024)* (Vol. 201). Springer, Cham. https://doi.org/10.1007/978-3-031-57870-0_18
65. Keerthiwansa, N. W. B. S. (2018). Artificial intelligence education (AIEd) in English as a second language (ESL) classroom in Sri Lanka. *Artificial Intelligence*, 6(1), 31-36.
66. Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*.
67. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35, 22199-22213.

68. Koraishi, O. (2023). Teaching English in the age of AI: Embracing ChatGPT to optimize EFL materials and assessment. *Language Education and Technology*, 3(1).
69. Krumpal, I. (2012). Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research*, 41(6), 1387-1403.
70. Kumar, H., Musabirov, I., Reza, M., Shi, J., Kuzminykh, A., Williams, J. J., & Liut, M. (2023). Impact of guidance and interaction strategies for LLM use on learner performance and perception. *arXiv preprint arXiv:2310.13712*.
71. Kupiec, J. (2002). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech & Language*, 6(3), 225–242.
[https://doi.org/10.1016/0885-2308\(92\)90019-z](https://doi.org/10.1016/0885-2308(92)90019-z)
72. Lacey, A., & Luff, D. (2001). *Qualitative data analysis* (pp. 320-357). UK: Trent Focus Group.
73. Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).
74. Lan, Y.-J., & Chen, N.-S. (2024). Teachers' agency in the era of LLM and generative AI: Designing pedagogical AI agents. *Educational Technology & Society*, 27(1), I-XVIII.
[https://doi.org/10.30191/ETS.202401_27\(1\).PP01](https://doi.org/10.30191/ETS.202401_27(1).PP01)
75. LangChain. (2022). Introduction |  LangChain. [LangChain.com](https://python.langchain.com/v0.1/docs/get_started/introduction/).
https://python.langchain.com/v0.1/docs/get_started/introduction/
76. Lensvelt-Mulders, G. J., Hox, J. J., & Heijden, P. G. V. D. (2005). How to improve the efficiency of randomised response designs. *Quality and Quantity*, 39, 253-265.
77. Lester, J. N., Cho, Y., & Lochmiller, C. R. (2020). Learning to do qualitative data analysis: A starting point. *Human Resource Development Review*, 19(1), 94-106.
78. Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., ... & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*.
79. Luo, H., Liu, P., & Esping, S. (2023). Exploring Small Language Models with Prompt-Learning Paradigm for Efficient Domain-Specific Text Classification. *arXiv preprint arXiv:2309.14779*.

80. Mackey, A., & Gass, S. M. (Eds.). (2011). Research methods in second language acquisition: A practical guide (Vol. 7). John Wiley & Sons. (pp. 180-194).
81. Maguire, M., & Delahunt, B. (2017). Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education*, 9(3).
82. Malcolm, K., & Casco-Rodriguez, J. (2023, March 19). A comprehensive review of spiking neural networks: interpretation, optimization, efficiency, and best practices. *arXiv.org*. <https://arxiv.org/abs/2303.10780>
83. Margolis, A. A. (2020). Zone of proximal development, scaffolding, and teaching practice. *Cultural-Historical Psychology*, 16(3).
84. Masalkhi, M., Ong, J., Waisberg, E., Zaman, N., Sarker, P., Lee, A. G., & Tavakkoli, A. (2024). A side-by-side evaluation of Llama 2 by meta with ChatGPT and its application in ophthalmology. *Eye*. <https://doi.org/10.1038/s41433-024-02972-y>
85. Melnikovas, A. (2018). Towards an explicit research methodology: Adapting research onion model for futures studies. *Journal of Futures Studies*, 23(2).
86. Mete, A., Kanthale, P., Bhaye, P., Subhedar, R., & Gupta, S. (2019). Text Extraction and Metadata Analysis of PDF Documents. In *A Survey* (Vol. 2, Issue 3, p. 563). https://www.ijresm.com/Vol.2_2019/Vol2_Iss3_March19/IJRESM_V2_I3_151.pdf
87. Mozes, M., He, X., Kleinberg, B., & Griffin, L. D. (2023). Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*.
88. Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5), 544–551. <https://doi.org/10.1136/amiajnl-2011-000464>
89. Neuendorf, K. A. (2018). Content analysis and thematic analysis. In *Advanced research methods for applied psychology* (pp. 211-223). Routledge.
90. Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual Review of Psychology*, 49(1), 352–354. <https://doi.org/10.1146/annurev.psych.49.1.345>
91. Pan, J. J., Wang, J., & Li, G. (2023, October 21). Survey of Vector Database Management Systems. *arXiv.org*. <https://arxiv.org/abs/2310.14021>

92. Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2024). Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12, 484-506. https://doi.org/10.1162/tacl_a_00660
93. Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2). <https://doi.org/10.53761/1.20.02.07>
94. Peterson, M., & Gärdenfors, P. (2023). How to measure value alignment in AI. *AI Ethics*. <https://doi.org/10.1007/s43681-023-00357-7>
95. Polman, J. L. (2010). The zone of proximal identity development in apprenticeship learning. *Revista De Educacion*, 129-155.
96. Prasad, P., & Sane, A. (2024, March). A self-regulated learning framework using generative AI and its application in CS educational intervention design. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 1070-1076).
97. Reiber, F., Pope, H., & Ulrich, R. (2023). Cheater detection using the unrelated question model. *Sociological Methods & Research*, 52(1), 389-411.
98. Reimers, N., & Gurevych, I. (2019, August 27). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv.org*. <https://arxiv.org/abs/1908.10084>
99. Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-7).
100. Rutland, A., & Campbell, R. N. (1996). The relevance of Vygotsky's theory of the "zone of proximal development" to the assessment of children with intellectual disabilities. *Journal of Intellectual Disability Research*, 40(Pt 2), 151-158.
101. Saracho, O. N. (2023). Theories of child development and their impact on early childhood education and care. *Early Childhood Education Journal*, 51(1), 15-30.
102. Saunders, M., Lewis, P., & Thornhill, A. (2023). *Research methods for business students* (9th ed.). Pearson.
103. Schaeffer, R., Miranda, B., & Koyejo, S. (2024). Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

104. Selepe, C., & Moll, I. (2016). Are teachers facilitators or are they mediators? Piaget, Vygotsky and the wisdom of the teacher. *The Independent Journal of Teaching and Learning*, 11(1), 6-16.
105. Shalf, J., Dosanjh, S., & Morrison, J. (2011). Exascale Computing Technology Challenges. *High Performance Computing for Computational Science – VECPAR 2010*, 1–25. doi:10.1007/978-3-642-19328-6_1
106. Sharifuddin, N. S., & Hashim, H. (2024). Benefits and challenges in implementing artificial intelligence in education (AIED) in ESL classroom: A systematic review (2019-2022). *International Journal of Academic Research in Business and Social Sciences*. <http://dx.doi.org/10.6007/IJARBSS/v14-i1/20422>
107. Shaw, R. G., & Mitchell-Olds, T. (1993). Anova for Unbalanced Data: An Overview. *Ecology*, 74(6), 1638–1645. <https://doi.org/10.2307/1939922>
108. Simmie, G. M. (2023). Teacher professional learning: A holistic and cultural endeavour imbued with transformative possibility. *Educational Review*, 75(5), 916-931.
109. Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of Retrieval Augmented Generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11, 1–17. https://doi.org/10.1162/tacl_a_00530
110. Smari, Waleed W.; Bakhouya, Mohamed; Fiore, Sandro; Aloisio, Giovanni (2016). New advances in High Performance Computing and simulation: parallel and distributed systems, algorithms, and applications. *Concurrency and Computation: Practice and Experience*, DOI: 10.1002/cpe.3774
111. Szepesvári, C. (2022). *Algorithms for reinforcement learning*. Springer Nature.
112. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., . . . Kenealy, K. (2024, March 13). GemMa: Open models based on Gemini research and technology. *arXiv.org*. <https://arxiv.org/abs/2403.08295>
113. Temple, B., & Young, A. (2004). Qualitative research and translation dilemmas. *Qualitative Research*, 4(2), 161-178.

114. Tjaden, Rolando, Doty, and Mortimer (2019) examine the long-term effects of time use during high school on positive development.
115. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments*, 10(1).
<https://doi.org/10.1186/s40561-023-00237-x>
116. Treuille, A. (2023, November 10). Generative AI and Streamlit: A perfect match. Streamlit. <https://blog.streamlit.io/generative-ai-and-streamlit-a-perfect-match/>
117. Ustaoglu, A., & Çelik, H. (2023). High school students' video game involvement and their English language learning motivation: A correlation study. *Journal of Educators Online*, 20(1), n1.
118. Van der Heijden, P. G., Van Gils, G., Bouts, J., & Hox, J. J. (2000). A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research*, 28(4), 505-537.
119. Van Wyk, M. M. (2024). Is ChatGPT an opportunity or a threat? Preventive strategies employed by academics related to a GenAI-based LLM at a faculty of education. *Journal of Applied Learning and Teaching*, 7(1).
120. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need.
https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
121. Villalobos, P., Sevilla, J., Heim, L., Besiroglu, T., Hobbhahn, M., & Ho, A. (2022). Will we run out of data? An analysis of the limits of scaling datasets in machine learning. arXiv preprint arXiv:2211.04325.
122. Vygotsky, L. S. (1986). *Thought and language*. Cambridge, MA: MIT Press.
123. Vygotsky, L. S. (2012). *Thought and language*. MIT Press.
124. Wallat, J., Jatowt, A., & Anand, A. (2024, March). Temporal Blind Spots in Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 683-692).

125. Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., & Zhang, Y. (2023). Survey on factuality in large language models: Knowledge, retrieval, and domain-specificity. arXiv preprint arXiv:2310.07521.
126. Wang, C., Nulty, P., & Lillis, D. (2020). A Comparative Study on Word Embeddings in Deep Learning for Text Classification. A Comparative Study on Word Embeddings in Deep Learning for Text Classification. <https://doi.org/10.1145/3443279.3443304>
127. Wang, L., Ma, C., Feng, X., et al. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(1), 186345. <https://doi.org/10.1007/s11704-024-40231-1>
128. Wang, Y., Teng, Y., Huang, K., Lyu, C., Zhang, S., Zhang, W., Ma, X., Jiang, Y.-G., Qiao, Y., & Wang, Y. (2023). Fake alignment: Are LLMs really aligned well? arXiv preprint arXiv:2311.05915.
129. Warford, M. K. (2011). The zone of proximal teacher development. *Teaching and Teacher Education*, 27, 252-258.
130. Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69.
131. Wei, J., Yao, Y., Ton, J. F., Guo, H., Estornell, A., & Liu, Y. (2024). Measuring and Reducing LLM Hallucination without Gold-Standard Answers via Expertise-Weighting. arXiv preprint arXiv:2402.10412.
132. 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. arXiv preprint arXiv:2112.04359.
133. Williams, M., & Moser, T. (2019). The Art of Coding and Thematic Exploration in Qualitative Research. *International Management Review*, 15(1), 45–55
134. Xie, S., & Xu, J. (2023). Design and implementation of physical education teaching management system based on multi-agent model. *International Journal of Computational Intelligence Systems*, 16(172). <https://doi.org/10.1007/s44196-023-00349-9>

135. Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817.
136. Yan, L., Sha, L., Zhao, L., Li, Y., Martínez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1), 90–112. <https://doi.org/10.1111/bjet.13370>
137. Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., Zhou, J., Chen, S., Gui, T., Zhang, Q., & Huang, X. (2023, March 18). A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. arXiv.org. <https://arxiv.org/abs/2303.10420>
138. Yeh, K.-C., Chi, J.-A., Lian, D.-C., & Hsieh, S.-K. (2023, October). Evaluating interfaced LLM bias. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 292-299).
139. Yong, G., Jeon, K., Gil, D., & Lee, G. (2023). Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. *Computer-Aided Civil and Infrastructure Engineering*, 38(11), 1536-1554.
140. Yoon, S. Y. (2023). Short Answer Grading Using One-shot Prompting and Text Similarity Scoring Model. arXiv preprint arXiv:2305.18638..
141. Yuan, T., He, Z., Dong, L., Wang, Y., Zhao, R., Xia, T., Xu, L., Zhou, B., Li, F., Zhang, Z., Wang, R., & Liu, G. (2024). R-Judge: Benchmarking Safety Risk Awareness for LLM Agents. arXiv preprint arXiv:2401.10019.
142. Zegeye, A., Worku, A., Tefera, D., Getu, M., & Sileshi, Y. (2009). Introduction to research methods. Graduate studies and research office, Addis Ababa University. (pp. 23-42).
143. Zhang, D., Cao, Q., Guo, Y., & Wang, L. (2023, December). Assistant teaching system for computer hardware courses based on large language model. In *International Conference on Computer Science and Education* (pp. 301-313). Singapore: Springer Nature Singapore.
144. Zhang, G., Hou, Y., Lu, H., Chen, Y., Zhao, W. X., & Wen, J. R. (2023). Scaling Law of Large Sequential Recommendation Models. arXiv preprint arXiv:2311.11351.

145. Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In Proceedings of the 38th International Conference on Machine Learning (Vol. 139, pp. 12697–12706).
146. Zhou, K. Z., Kilhoffer, Z., Sanfilippo, M. R., Underwood, T., Gumusel, E., Wei, M., Choudhry, A., & Xiong, J. (2024). "The teachers are confused as well": A multiple-stakeholder ethics discussion on large language models in computing education. arXiv preprint arXiv:2401.12453
147. Zhou, M., & Brown, D. (2015). Educational learning theories. Education Open Textbooks.

7. Appendices

A. Design of RRT question sheet (See full version in RRT Data Analysis Folder)

	Mand	Kvinde
Kast terning 1-3 Kryds "Ja" 4-6 Kryds ærligt svar	Kryds her for "Ja"	Kryds her for "Nej"
Har du brugt en LLM, såsom ChatGPT, til at assistere med din læring uden for eksamensforberedelse?		
Har du brugt en LLM til at hjælpe med at forberede dig til eksamener?		
Har du brugt en LLM under eksamen?		
Har du brugt en LLM til at generere indhold for skriftlige opgaver eller afleveringer?		
Har du oplevet at du er ofte afhængig af brugen af LLM for at færdiggøre dine lektier?		

- B. Head of Education interview guide (See full version in Interview Folder)
- C. High school teacher interview guide (See full version in Interview Folder)
- D. High school students interview guide (See full version in Interview Folder)
- E. RRT Database - (See full version in RRT Data Analysis Folder)
- F. Interview transcript Student A - ENG (See full version in Interview Folder)
- G. Interview transcript Student B - ENG (See full version in Interview Folder)
- H. Interview transcript Student C - ENG (See full version in Interview Folder)
- I. Interview transcript Student D - ENG (See full version in Interview Folder)
- J. Interview transcript Student E - ENG (See full version in Interview Folder)
- K. Interview transcript Student F - ENG (See full version in Interview Folder)
- L. Interview transcript Student G - ENG (See full version in Interview Folder)
- M. Interview transcript Student H and Student I - ENG (See full version in Interview Folder)
- N. Interview transcript Teacher A - ENG (See full version in Interview Folder)
- O. Interview transcript Teacher B - ENG (See full version in Interview Folder)

- P. Interview transcript Head of Education A - ENG (See full version in Interview Folder)
- Q. Interview transcript Head of Education B - ENG (See full version in Interview Folder)
- R. Expert group recommendations (UVM, 2024) - (See full version in Interview Folder)
- S. 5 Questions' Dataset After Dropping - (See full version in RRT Data Analysis Folder)
- T. 5 Questions' Cross Tabulation Results - (See full version in RRT Data Analysis Folder)
- U. 5 Questions' T-Test Results - (See full version in RRT Data Analysis Folder)
- V. 5 Questions' ANOVA Test Results - (See full version in RRT Data Analysis Folder)
- W. model.py - (See full version in RAG Application Folder)
- X. ingest.py - (See full version in RAG Application Folder)
- Y. prompt.py - (See full version in RAG Application Folder)
- Z. app.py.py - (See full version in RAG Application Folder)
- AA. requirements.txt - (See full version in RAG Application Folder)
- BB. vector_db - (See full version in RAG Application Folder)
- CC. Overview of tools for written upper secondary school exams (Ministry of Children and Education, 2024) (See full version in Interview Folder)

8. List of Figures

1. Herningsholm Erhvervsskole & Gymnasier logo (Herningsholm, 2024).....	(pp. 5)
2. Ringkjøbing Gymnasium logo (Ringkjøbing, 2024).....	(pp. 5)
3. Visualization of the components making up AI alignment (Own creation).....	(pp. 23)
4. Visualization model of ZPD (Own creation).....	(pp. 27)
5. Showcasing of appliance of zero-shot, one-shot and few-shot prompting via ChatGPT-4 (Own creation).....	(pp. 29)
6. The Research Onion (Source: Saunders MNK, Lewis P and Thornhill A (2023) Research Methods for Business Students (9th edition) Harlow: Pearson, p 177. The Research Onion is ©2022 Mark NK Saunders and is reproduced in this thesis with permission)	(pp. 36)
7. Directory loader for loading pdf files (Own creation).....	(pp. 42)
8. Calling embedding and texts with FAISS for indexing and calling the function of vector db (Own creation).....	(pp. 43)
9. Creating a custom prompt template (Own creation).....	(pp. 43)
10. Creating a retrieval qa chain function (Own creation).....	(pp. 44)
11. Calling QA bot function to get the output (Own creation).....	(pp. 45)
12. Cross Tabulation with Using Outside Exam (Own creation).....	(pp. 58)
13. T-Test for Gender and School vs Using Outside Exam (Own creation).....	(pp. 59)
14. Anova Test for School Year & Class Type vs Using Outside Exam (Own creation).....	(pp. 60)
15. Cross Tabulation with Using For Exam (Own creation).....	(pp. 60)
16. T-Test for Gender and School vs Using For Exam (Own creation).....	(pp. 61)
17. Anova Test for School Year & Class Type vs Using For Exam (Own creation).....	(pp. 62)
18. Cross Tabulation with Using In Exam (Own creation).....	(pp. 62)
19. T-Test for Gender and School vs Using In Exam (Own creation).....	(pp. 63)
20. Anova Test for School Year & Class Type vs Using In Exam (Own creation).....	(pp. 64)
21. Cross Tabulation with Using For Assignment (Own creation).....	(pp. 64)
22. T-Test for Gender and School vs Using For Assignment (Own creation).....	(pp. 65)
23. Anova Test for School Year & Class Type vs Using For Assignments (Own creation).....	(pp. 66)

24. Cross Tabulation with Using For Addiction (Own creation).....(pp. 66)
25. Cross Tab Analysis of Addiction with Other Research Questions (Own creation)..(pp. 68)
26. T-Test for Gender and School vs Using For Addiction (Own creation).....(pp. 69)
27. Anova Test for School Year & Class Type vs Using For Addiction (Own
creation).....(pp. 69)
28. Paragraph concerning regulations on digital help tools in exams from Appendix R
(UVM, 2024).....(pp. 73)
29. Paragraph from the recommendations issued by the expert group concerning
recommendations regarding AI in education from Appendix R (Source: UVM,
2024).....(pp. 74)
30. Illustration of RAG Functionality (Own creation).....(pp. 85)

9. List of Tables

1. Table of the literature review of LLM in education (Own creation).....	(pp. 19)
2. Overview of the applied format of research onion (Own creation).....	(pp. 38)
3. Overview of the semi-structured interviews (Own creation).....	(pp. 47)
4. Theme 1: The usage of LLMs in education (Own creation).....	(pp. 48)
5. Theme 2: Positive aspects of LLM use in HEI (Own creation).....	(pp. 50)
6. Theme 3: Challenges and issues with students LLM usage (Own creation).....	(pp. 52)
7. Theme 4: LLMs are changing HEI educational practices (Own creation).....	(pp. 54)
8. Theme 5: Theme 5: LLMs are changing HEI educational practices (Own creation).....	(pp. 56)