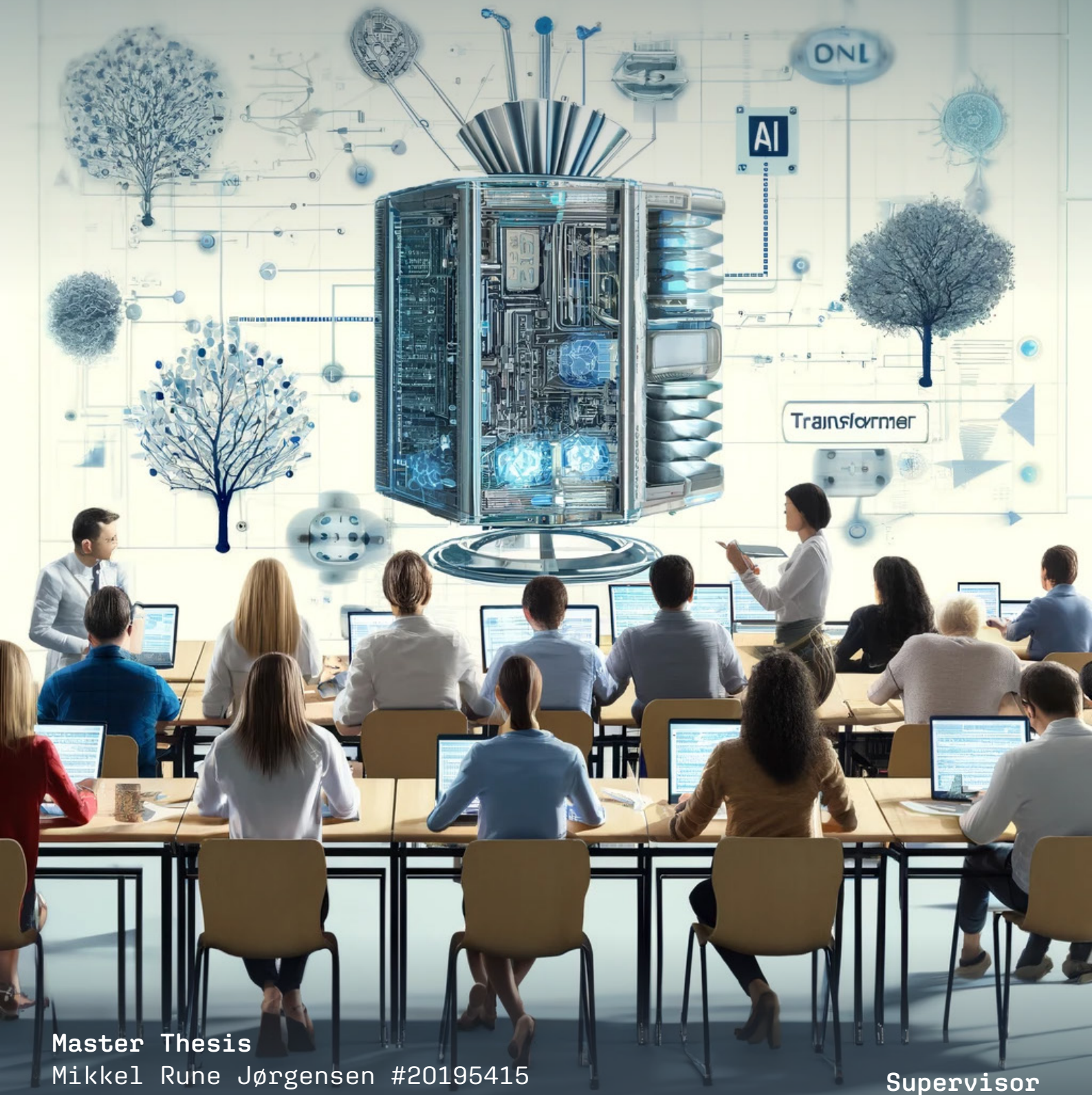
# Benefits and Challenges of LLMs in Danish Education

## The Future of Learning with AI




**Master Thesis**
Mikkel Rune Jørgensen #20195415
*Aalborg Universitet - BDS 10. semester 2024*

**Supervisor**
Hamid Bekamiri

# Challenges and Opportunities of LLM Integration in Danish Educational Schools

**Abstract**

This master thesis investigates the integrations of LLMs, mostly with a Focus on ChatGPT4 into the Danish primary educational system. The main area of research is the applications on the Danish subject. To do so, an examination of the potential benefits and challenges highlighted in the current literature published in other countries is transferred to a Danish setting, aiming at seeing the differences and potentials of integrating these tools. Different techniques are employed like Few-Shot, Chain-of-Thought and ReAct-prompting to evaluate the output generated by the LLMs. The integration of ChatGPT4 with Retrieval Augmented Generation was developed to ensure relevance of educational materials based on the Danish objectives and materials, while enhancing up-to-date information and relevance. Through qualitative methods, which was done with semi-structured interviews. The qualitative approach gave insights in teachers' view on these technologies, and identified potential areas of improvement within the Danish primary school.

The findings indicated that teachers lack knowledge in regards to what these new technologies can provide. It also showcased that clear guidelines are essential to make the implementation successful. The quantitative part of this study was based on generating assignments in different areas. Here ChatGPT4 showed a good performance in being able to make similar material to what the teachers use, while also fostering different skill levels of difficulty that can be used to personalise the teaching material. Evaluations of the output generated were done by two teachers, rating the produced material either as satisfied or unsatisfied. The results obtained from both the qualitative and quantitative methods show that ChatGPT4 might have potential for the customisation and enhancement of educational materials, it also indicates some challenges, including issues of data accuracy, teacher adoption, and ethics that must be addressed.

**Keywords: : ChatGPT4, AI in education, Danish educational system, Large Language Models (LLMs), Personalised learning, Teacher support, Retrieval-Augmented Generation (RAG)**

## Table of contents

# 1 Introduction

The creation of advanced models of AI (like ChatGPT) has stirred strong interest in the educational community and led to a large number of discussions. As schools test how best to fit AI into their curricula worldwide, the potential positive and negative impact that the introduced technology can bring to the established educational system needs to be considered (Kasneci et al., 2023).

In Denmark, the application of ChatGPT in education, primarily for the Danish language subject, offers a unique opportunity to promote personalised learning, student engagement, and educational content creation. The tool of ChatGPT used for personalised learning, can provide learning experiences best fitting individual student needs, their paces and their learning styles (Kasneci et al., 2023)

Through the use of ChatGPT, teachers are able to create activities, reading and interactive texts in response to the diversified needs of the students, which have the potential to foster engagement and increase the learning outcome (Kasneci et al., 2023). However, it's not without challenges related to the implementation of these tools in education. One of the challenges using AI in an educational setting is the challenge of data privacy. This is due to the fact that schools and educational institutions will have to collect and process personal data from the students (Kasneci et al., 2023). In addition, AI generated content can display bias and misinformation related to the data used for its training and can result in biassed or inappropriate generation output (Kasneci et al., 2023;Gan, Qi, Wu, & Lin, 2023a).

This study is about the impacts that ChatGPT4 can have within the primary Danish education system, with a specific focus on the Danish Language subject and content generation for teachers. Understandably, through the theoretical grounds covered and extant literature review presented, this paper tries to realise both the potential and concerns that are followed by AI technologies for driving educational purposes. It also evaluates the practical application of ChatGPT4 within the generation of content related to education and the acceptability of this among Danish teachers.

As we look into the potential of ChatGPT4 in education, it is crucial to first understand the theoretical foundations and existing research in this field. Therefore, the next section will explore the theoretical background and review relevant literature.

# 2 Theoretical Background

The development of Pre-trained Transformers (GPTs) and Large Language Models (LLMs) present a paradigm shift in educational methodologies. According to Memarian (Memarian & Doleck, 2023) these AI driven technologies in education have been met with both enthusiasm and caution. This review presented in this study looks into the dual aspects of the potentials and challenges associated with the use of GPTs and LLMs in an educational setting.

The current state of the literature and articles published regarding ChatGPT and LLMs consist mostly of a qualitative approach, where it focuses on the benefits and harms it might bring to education as a whole. According to (Memarian & Doleck, 2023) ChatGPT and AI in general has seen a lot of popularity and attention in the last year, and are already breaking headlines again at the start of 2024. Despite the huge attention, especially ChatGPT has gotten,  a lack of work has been done in regards to the literature (Memarian & Doleck, 2023). The goal of his literature review in this study is to display the findings in relation to the education system focusing on the Danish primary educational system. Here it's important to mention that at the current state of conducting this study, limited research has been done on how ChatGPT and LLMs impact the Danish educational setting (Memarian & Doleck, 2023).

Therefore the literature review will draw a parallel to the Danish education setting, and will be set in relation to how this can be applied. Furthermore it is relevant to address the current state and view on AI in general like ChatGPT which has been rather negative form the Danish Ministry of Children and Education, and other educational institutions where a ban has been implemented on using these new tools for students for now Aalborg Universitet, 2023); BØRNE- OG UNDERVISNINGSMINISTERIET, 2024). Not all have taken this to the same extent, Aarhus University allows bachelor and master students to use AI in creating their bachelor and thesis and some other cases as well (Aarhus Universitet, 2024). The reason is that they acknowledge the need for learning the skills, and comment that it's an evendiable reality for the future of the job market (Aarhus Universitet, 2024).

As seen in Table 1, some of  the literature used to look into the current state displays different but also similar views on GPTs and LLMs in an educational setting. Firstly, the benefits aspects will be introduced, then the negative/concerns regarding implementations.

*Table 1. Literature Review Overview*

| Paper Info/reference | Method/Data | Result/conclusion |
|---|---|---|
| *Memarian & Doleck, 2023* | Literature Review and Analysis of current applications of LLMs like GPT-3 and BERT in educational settings. | LLMs can enhance personalised learning and engagement but pose challenges such as bias and over-reliance. Strategies for responsible integration proposed. |
| *Gan, Qi, Wu, & Lin, 2023* | Literature Review and Systematic Summary; Analysis of the current research status and application scenarios of EduLLMs; Discussion of potential and challenges. | EduLLMs offer significant potential for improving the quality of education and learning experience through personalised learning support, intelligent tutoring, and educational assessment capabilities. Challenges include technical, ethical, and practical issues requiring further research and exploration. |
| *Kasneci et al., 2023* | The paper provides a commentary on the potential benefits and challenges of using LLMs like ChatGPT in educational settings. It discusses the current advancements in LLMs, focusing on their applications in generating educational content, improving student engagement and personalising learning experiences. | LLMs can enhance personalised learning, assist in content creation, and potentially improve student engagement.The paper highlights the necessity for educators to develop specific competencies to effectively utilise LLMs, including strategies for critical thinking and fact-checking. It stresses the need for a pedagogical approach that incorporates these technologies thoughtfully to maximise their educational benefits. |

## 2.1 Benefits and concerns

GPTs and LLMs bring significant benefits to education. This includes personalised learning paths which can be tailored to the individual student needs, making a more deep engagement and aids in understanding complex subjects (Kasneci et al., 2023; Memarian & Doleck, 2023). These technologies can provide real time feedback to the students, which makes the students able to learn from their mistakes promptly while bringing a supportive learning environment (Kasneci et al., 2023). Furthermore, these models can democratise education and extend resources to underserved communities. In a Danish setting, this could include helping families that are not as financially strong, or have diverse linguistic and different cultural backgrounds (Phung et al., 2023).

Moreover the possibility of LLMs in educational material creation and curriculum development stands out as a major opportunity for the teachers. By utilising these AI tools teachers can generate comprehensive & customised teaching material that cater to the different needs of their classrooms and students (Memarian & Doleck, 2023). Having a tool like this might mitigate the time that teachers spend on preparing material, where the time can be allocated differently. This not only reduces the workload, but it might also enhance the learning experience for the student by providing different learning objectives in a different way (Memarian & Doleck, 2023).

However, the implementation of these new AI technologies is not without any concerns and problems. Data privacy and ethics is a critical issue (Kasneci et al., 2023). When using these tools a collection of sensitive student data will be obtained, which calls for safeguards considerations. In an educational setting, it's also crucial to be aware of the potential biases that these AI's can produce, which stems from their training data. The biases and wrong information can perpetuate or even enhance the existing educational inequalities leaning to outcomes that may be false or even discriminatory (Gan, Qi, Wu, & Lin, 2023b).

Moreover, an over reliance of AI for educational content and assessment risk diminish critical thinking and the skills for problem solving among students (Kasneci et al., 2023). When it comes to successfully implementing these new technologies many factors need to be considered. Firstly the technological infrastructure, teacher training and adaptation of curricula to incorporate into the models are key factors (Kasneci et al., 2023). Teachers play a crucial role in the integration, as they need to learn how to effectively incorporate LLMs into

their teaching practices, so it enhances, rather than replaces, the traditional teaching methods (Gan, Qi, Wu, & Lin, 2023).

In summary of the literature, data privacy, biases and misinformation are the most significant concerns when it comes to integrating these tools into education. There is a need to protect sensitive student information and mitigate biases and misinformation so it can be used in an educational setting (Memarian & Doleck, 2023; Zhou, Zhang, Luo, Parker, & De Choudhury, 2023). The root of the AI generating misinformation and biases lies in the nature of these AI models themselves. The models are trained on big datasets from the internet, which both include factual and non factual information. As a result of this, the models can generate both correct and wrong information.

Having outlined some of the benefits and concerns it is essential to look deeper into the mechanisms of LLMs to fully understand these implications. The next section will provide an overview of how these models function and their inherent limitations.

## 2.2 Understanding LLMs and GPTs: Mechanisms, Capabilities, and Limitations

GPTs like ChatGPT3 are built to predict the next word in a sequence, this makes them highly effective at generating coherent and fluent text for many different tasks. However, this capability does not ensure that the generated output is correct. The goal of these models is to produce text that is statistically likely to follow from the prompt that is given without knowing truth or reality (Wang, B. et al., 2023). All they do is replicate patterns which include the biases and misinformation that comes from the training data. Furthermore AI models like LLMs make the output look persuasive and credible. This capability makes these new technologies a powerful tool for spreading false and inaccurate information that might be considered correct (Gan, Qi, Wu, & Lin, 2023; Zhou, Zhang, Luo, Parker, & De Choudhury, 2023).

LLMs are designed to understand and generate human language by facilitating a wide range of applications like translation, question answering and content generation. Normally the scale of LLMs is based on the volume of parameters, which is what allows them to capture and understand language, context and semantics (Lampinen et al., 2022). At the core of these LLMs is a Transformer architecture, introduced by Vaswani (Vaswani et al., 2017). This architecture is different from the previous models where other architectures like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) processed text

sequentially. Instead, the Transformer employs self-attention mechanisms that enable the model to weigh the relevance of different words in a sentence regardless of their positional distance (Vaswani et al., 2017; Kasneci et al., 2023). The transformer architecture allows for more efficient processing and a better understanding of the context within text sequences.

LLMs undergo extensive pre-training on a big corpora of text data. The pre-training is what enables the model to learn a wide range of language patterns, structures and context. After pre-training it is possible to use fine-tuning. Fine-tuning is where the LLMs is further trained on smaller task specific datasets (Kasneci et al., 2023). However, the training of LLMs is not without challenges (Lampinen et al., 2022). Issues such as bias, fairness and the environmental impact of training large models are areas of active research and debate (Echterhoff, Liu, Alessa, McAuley, & He, 2024).

Below in Table 2 are some of the literature used to display the current state of these AI technologies. Understanding the capabilities and limitations of LLMs provides a foundation for exploring how to fine-tune these models for educational purposes. The next section will discuss various fine-tuning techniques and their applications.

*Table 2 Literature Review Overview*

| Paper Info/reference | Method/Data | Result/conclusion |
|---|---|---|
| (Lampinen et al., 2022) | Investigating the potential of explanations alongside few-shot examples to improve the performance of language models (LMs) on new tasks without additional tuning.<br><br>The study annotated questions from 40 challenging tasks with both answer explanations and various control explanations, examining their impact on both zero- and few-shot learning performance | Found that explanations significantly enhance LM performance in few-shot settings, even without model tuning. Hand-tuned explanations on a small validation set showed even larger benefits, especially in larger models, highlighting a scalability aspect where larger LMs better utilise explanations for performance gains |
| Lewis et al., 2020 | The study introduces RAG models that enhance language models for knowledge-intensive tasks by integrating pre-trained sequence-to-sequence models with non-parametric memory. This memory is a dense vector index of Wikipedia, accessed via a neural retriever. | The findings demonstrate that RAG models significantly outperform traditional sequence-to-sequence models and task-specific architectures by generating more specific, diverse, and factually accurate responses. This not only improves the quality of the generated text but also addresses the challenges of hallucinations and updating knowledge, which are prevalent in traditional pre-trained models benchmarks. |
| Lester, Al-Rfou, & Constant, 2021 | Explore the efficacy of "prompt tuning" as a simple mechanism for conditioning frozen language models to perform specific downstream tasks using "soft prompts”.<br><br>Implements prompt tuning on various sizes of the T5 model to compare its performance with full model tuning and to evaluate the scalability of prompt tuning | Prompt tuning closes the performance gap with full model tuning as model sizes increase, becoming competitive with or outperforming model tuning in large models. Demonstrates that prompt tuning requires significantly fewer parameters, reducing the computational overhead and allowing easier adaptation of models to new tasks. |

## 2.3 Fine-tuning LLMs

Fine-tuning LLMs enables these models to adapt to specific tasks or datasets with efficiency. This section outlines the significance, methodologies and recent advancements in fine-tuning LLMs.

### 2.3.1 *In-context Learning (Prompt Engineering and RAG)*

In context learning is key in modern NLP, due to its capacity to adapt models to tasks without explicit retraining or parameter updates (Lampinen et al., 2022). ICL utilises the capability of pretrained language models to generate responses based on the provided context. This can include examples of tasks or questions and corresponding answers.

The document by (Lampinen et al., 2022) highlight the capabilities of in context learning by showing its operational framework and evaluating its practical implications. The essence of in-context learning lies in its ability to use the existing knowledge encapsulated within the LLMs. This enables them to handle a wide array of tasks with minimal example inputs. However, challenges such as computational inefficiency, potential performance limitations and sensitivity to prompt design are acknowledged (Echterhoff, Liu, Alessa, McAuley, & He, 2024).

***Prompt Engineering:*** Prompt engineering comes from in context learning as an enhancement technique focused on optimising the interaction between the users and the language models. Through designed prompts it seeks to guide the model's responses towards the desired objectives. This is done by using various techniques like zero shot learning, few shot learning, CoT prompting, and ReAct prompting while also maintaining a solid prompt structure that does not confuse the models (Lampinen et al., 2022).

***Zero-Shot and Few-Shot Learning***: In zero-shot learning the model is tasked with generating responses based on a prompt without any examples. This makes the model rely entirely on its existing knowledge. Few-shot learning, on the other hand, includes a small set of example inputs and outputs in the prompt. The aim is to guide the model's response generation more effectively (Lampinen et al., 2022;Wang, Yao, Kwok, & Ni, 2020). Both methods use the model's ability to generate more correct output from limited information, highlighting the significance of prompt construction in getting accurate responses.

***Chain-of-Thought (CoT) Prompting:*** CoT prompting uses the LLMs to generate intermediate steps or so called reasoning paths that make up the final answer the model generates (Lampinen et al., 2022; Yao et al., 2022). This approach not only enhances the model's problem solving capabilities but also offers insights into its reasoning process, making it more transparent and interpretable how and why it generated the output it did.

***React-Prompting:*** ReAct prompting introduces a framework where prompts are refined based on the models previous outputs. This feedback loop allows for incremental adjustments to the prompt. This technique gives more control over the model's generative process by facilitating relevant and accurate responses. ReAct prompting can be viewed as a continuous chat with the model, where you adjust the prompt based on the output until the output fits the objective of the user (Yao et al., 2022).

***Retrieval-Augmented Generation (RAG):*** The retriever uses embeddings and vector space modelling, where it queries external databases to fetch information relevant to the input prompt (Lewis et al., 2020). These embeddings are presented as high dimensional spaces where the text is converted into vectors. By converting the text into vectors it's possible to capture semantic relationships and contextual cues that make the retriver identify the documents that are contextually aligned with the query (Lewis et al., 2020).

The generator, typically a sophisticated pre-trained LLM, integrates the retrieved documents with the original prompt to provide enriched and context-aware responses. During RAG, the retriever dynamically encodes the retrieved content into embeddings that are interoperable with the generator's internal representations ensuring that the external knowledge is well integrated into the generation pipeline (Gao et al., 2023; Lewis et al., 2020). The effectiveness of RAG depends on the retriever’s capacity to understand and match the semantic context of the queries and the knowledge base. In the current architecture, first, the queries are represented as an embedding, which converts textual data into a mathematical vector space. In this way, semantic similarities are represented as spatial proximities in the vector space. Next, the retriever shifts its focus across the space and retrieves and encodes the most relevant information. This means that the embeddings are trained with a goal of capturing a vast array of linguistic and semantic subtleties (Gao et al., 2023). Additionally, embeddings are used across the diverse types of queries and domains in order to guarantee

effective representation and retrieval. RAG constitutes an extension of the in-context learning principles by including an external corpus as an augmented context (Jauhiainen & Guerra, 2024).

### 2.3.2 *Prompt Tuning*

Prompt tuning works by adapting pre-trained LLMs to perform on specific tasks or improve performance of different types of output. The main idea is that it is done without the need for extensive retraining of the models (Lester, Al-Rfou, & Constant, 2021). This technique is advantageous as it leverages the inherent capabilities of the LLMs while it enhances the functions through small adjustments. To do so two approaches can be used: hard prompt tuning and soft prompt tuning.

***Soft prompt:*** This approach uses trainable tokens known as "soft prompts". These tokens are optimised during the training phase which are not a part of the original model. The tokens are learned through backpropagation and get embedded into the input space, which makes the model adapt dynamically to different tasks at hand (Lester, Al-Rfou, & Constant, 2021). The key advantage of soft prompts is its flexibility, where the soft prompts can continuously evolve during the training to better fit to the concrete tasks at hand. For instance (Lester, Al-Rfou, & Constant, 2021) illustrate how soft prompts can effectively condition large models like GPT-3 to perform specific tasks with only a fraction of the parameters being tuned while preserving the underlying model's general capabilities.

***Hard prompt*** tuning consists of constructing fixed hard prompts to be used for the input of the LLM and using them to guide the model's predictions. They are built using only the unmodified vocabulary of the initial model (SOFTWARE MIND, 2023). Because hard prompts cannot change over time, they are somewhat limited in flexibility, but they have the advantage of being easier to tune (SOFTWARE MIND, 2023). There is no need to expend additional time and resources on optimising the tokens of the hard prompts ((SOFTWARE MIND, 2023).

### 2.3.3 *PEFT*

Beyond the discussion of advanced techniques such as in context learning, prompt engineering and prompt tuning to improve the effectiveness of LLMs, another technique is parameter efficient fine-tuning (PEFT). PEFT seeks to refine LLMs task by adjusting a small subset of the model's parameters (Han, Gao, Liu, & Zhang, 2024). This approach is relevant due to the reduction of computational cost and time needed to make the model adapt to new specific tasks. To do so it utilises Adapter Modules and LoRA (Low-Rank Adaptation), which are powerful tools for getting high performance with minimal changes to the original model architecture. Adapter Modules involve setting small trainable layers within the model's architecture (Han, Gao, Liu, & Zhang, 2024). These modules are optimised during the fine-tuning process allowing the majority of the pre-trained model to remain the same (Han, Gao, Liu, & Zhang, 2024). This technique has been praised for its ability to maintain the general applicability of the LLMs while targeting the adaptation for the specific tasks.

LoRA  proposes a different strategy. This is done by adjusting the rank of weight matrices within the model. By focusing on low-rank updates LoRA enables efficient tuning of the model’s parameters with a small increase in the numbers of the trainable weights. Like with Adapter Modules it preserves the models original structure and knowledge while optimising its task specific performance (Han, Gao, Liu, & Zhang, 2024).

### 2.3.4 *RLHF*

Reinforcement Learning from Human Feedback (RLHF) is a technique used for adjusting the LLMs behaviour with human evaluations. To do so it integrates human feedback into the learning process. Doing this helps to ensure that the outputs are more aligned with human values and goals (Bai et al., 2022).

RLHF consists of three main components: feedback collection, reward modelling and policy optimisation (Bai et al., 2022). The feedback collection works by collecting feedback from humans on the generated output from the models. This aids in better understanding the preferences and expectations of the output. Reward modelling uses these gathered evaluations to train the models, so the models know better what humans consider good or bad outputs. Policy optimisation adjusts the output generation mechanism to enhance the rewards predicted by the reward model, in other words this means that the models are more likely to

produce outputs that fit with the evaluations that are considered good (Bai et al., 2022; Casper et al., 2023).

The adoption of RLHF has proven effective, particularly because it focuses on optimising for outputs that are explicitly rated highly by humans, rather than relying only on demonstrations or manually engineered reward functions (Casper et al., 2023).

## 2.4 The research gap in the literature

Despite the huge attention that these new technologies bring, a lack of work has been done in regards to the literature (Memarian & Doleck, 2023). This gap gives an opportunity for conducting further research (Kasneci et al., 2023). Addressing this gap could provide insights into the potential of LLMs to enhance educational outcomes and offer recommendations for effectively implementing AI tools in Danish educational practices. At the current state of writing this, an expert group has been tasked to see and come up with recommendations on how the educational institutions can incorporate AI into the school system and more (BØRNE- OG UNDERVISNINGSMINISTERIET, 2023). The findings of the report will be exciting to see and which recommendations they will bring forth.

The need for change is relevant to address at the time, hence the Danish students in the primary school are performing worse than before in the international PISA tests (BØRNE- OG UNDERVISNINGSMINISTERIET, 2024a). Based on the literature review on the benefits on how GPTs and LLMs can impact the learning outcomes for students, this could be a potential solution for trying to enhance the learning outcome in general. Despite the growing literature on LLMs into educational systems, there remains a gap in context specific research within the Danish educational framework. Existing studies have documented the potential benefits and challenges of applying LLMs and GPTs in a educational settings, focusing primarily on their ability to personalise learning, enhance student engagement, and provide immediate feedback (Kasneci et al., 2023). However, these studies might overlook the unique attributes, needs, and challenges of the Danish educational system. This study aims to understand these gaps by providing a detailed examination of the potential benefits and challenges of integration LLMs and GPTs into the Danish educational context. The study seeks to understand the specific needs and opportunities presented in the Danish educational standards and how these advanced AI technologies can be tailored to enhance educational outcomes, with a focus on helping teachers. By doing so, this research will hopefully

contribute insights into the localised application and implications of LLMs, offering guidance for educators, policymakers, and technologists looking to leverage AI in support of Danish educational goals.

Denmark, a country known for its high digital literacy and solid educational infrastructure lays the foundation for an implementation of AI that could be beneficial in many ways compared to other countries (Svendsen & Svendsen, 2023). Firstly, the access to computers in general is really high (EMU, 2018). After the new school reform in Denmark many teachers do not have sufficient time when it comes to the preparations of the material the students have to work with (Böwadt, Pedersen, & Vaaben, 2019). If the teachers can be able to produce material faster, they would also be able to allocate the time to make more fun and engaging material for the students. Furthermore this relates to the problem with screen time for students, where social media as an example has a negative impact on the motivation for learning for students, which requires more engaging material to stay focused (Marciano, Camerini, & Morese, 2021). All of the above might be mitigated by an implementation of LLMs. To explore this the following research questions will be used:

- ***What are the benefits & challenges of implementing an LLM into the Danish Education system for teachers***
- ***How can an LLM be implemented within the Danish primary school?***

The main focus of these research questions is to explore the benefits and challenges of integrating LLMs like ChatGPT4, into the Danish primary education system, with a focus on the Danish language subject. The integration of LLMs and GPTs into Danish education represents an opportunity for enhancing student learning outcomes by focusing on making more personalised educational material. This is especially relevant given the challenges highlighted by the recent PISA test results.

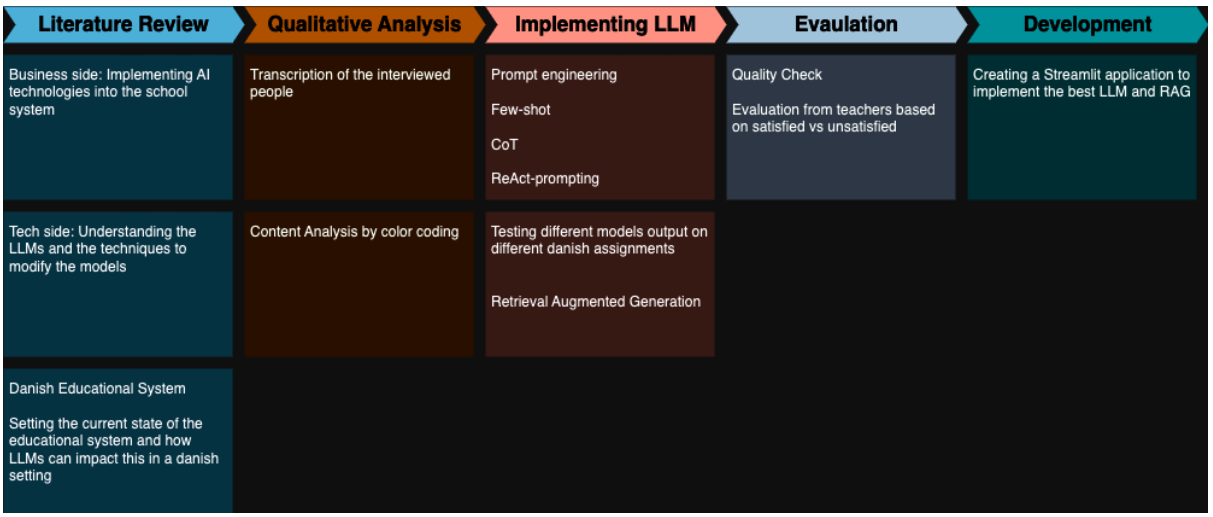
# 3 Methodology

This section shows the methodological approach used in this study to explore the integration and implications of LLMs in the Danish primary educational system. In doing so the study will be using mixed methods by employing qualitative and quantitative methodologies

## 3.1 Research Design

The research design provides an overview of the project design, detailing the process from the literature review to development.

*Figure 1. Research Design of This Study.*




As illustrated in Figure 1, the whole process is visualised. The figure is divided into five main stages: Literature Review, Qualitative Analysis, Implementing LLM, Evaluation, and Development. The Literature Review is built upon business and technical factors justifying the integration of AI technology into the Danish school system and getting an understanding of how these technologies works. The Qualitative Analysis involves transcribing interviews with the educators and conducting content analysis through colour-coding to obtain a clear view of their experiences and perceptions on AI and problems in their teaching. At the Implementing LLM stage, the use of the techniques like prompting engineering that involves Few-Shot learning, CoT, and ReAct-prompting on different Danish assignments is tested. In the Evaluation stage, there is a quality check and feedback from the teacher assessment regarding their satisfaction with the AI-generated assignments. The last stage, which is the Development of a Streamlit web application, ensures that LLM and RAG are practicable and can be integrated into the Danish educational framework. Now the data collected for this study will be explained.

## 3.2 Data

### 3.2.1 *Qualitative analysis: Selection of Participants*

The qualitative analysis of this study was set to gather an understanding of the Danish primary school system, and which areas that might be enhanced while reviewing teachers' knowledge and view on these new AI technologies like ChatGPT. To achieve this, different participants were chosen so they represented a broader teaching community, while also processing relevant characteristics to the research objectives.

**Rationale for Participant Selection:** Teachers were selected based on the subject areas they thought, with a main focus on danish while mathematics, history and science also were in focus subjects areas. This was done to better capture a diverse range of perspectives, while understanding the diversity on how AI tools might be used across different subjects in the Danish primary school.

**Experience Level:** Both novice and experienced, mostly experienced, teachers were included in this study. Novice teachers might bring new information and have different views, while they also might be more likely to adopt these new technologies. Experienced teachers on the other hand might not be as open minded on new technologies, but bring more experience in teaching practices. This dual focus was intended to offer a more broad view of the potential impact of AI tools at different stages of teaching careers, and their views.

**Representation of Different School Environments:** An attempt was made to interview teachers from both public and private schools to ensure that findings were widely applicable across diverse educational settings. Having a diverse representation might lead to a better understanding of any contextual factors that might influence the adoption and effectiveness of the AI tools.

By selecting teachers from different schools, experience and different subjects this part of the study aimed to ensure that the qualitative data that were collected would be relevant for the research while providing a solid foundation for answering the research questions.

### 3.2.2 *LLMs*

For testing the capabilities of ChatGPT4 a diverse range of assignments was used to see how well it could generate educational material. This includes online resources, contributions from teachers both in form of assignments but also based on what they pointed out in the interviews, and textbooks recognised within the Danish education system.

### *Online Material Sources*

The internet served as a primary resource for gathering existing examples of different assignments. The data collected varies and most of the online material was collected from a site called ***opgaveskyen.*** Here, data could be selected to match the grade levels for different subjects. In this case, material for the Danish subject was collected and used.

### *Contributions from teachers*

The interviewed teachers  provided access to some of the materials they use for teaching. However, there was a limit to the materials they could share. Many teachers use the portals called Clio & Gyldendal, where they can find and use materials for all subjects taught in the primary schools (BØRNE- OG UNDERVISNINGSMINISTERIET, 2021). However, access to this material was not possible due to copyright restrictions. The material provided by the teachers and what they requested mainly involves creating tasks for various texts from Clio & Gyldendal. Also, further insights from the interviews are the ability to get inspiration, make personalised material, and find videos that align with what they are teaching.

### *Danish Educational Textbooks*

Textbooks used in primary school have also been used. The books were primarily from 4-9th grade. Here, some tasks were selected from the books, focusing on varied tasks within grammar, reading and writing. This approach helped ensure an understanding of the curriculum and the different types of exercises that could be enhanced or changed with the implementation of LLMs into the educational setting.

### *Limitation of data*

One of the pivotal challenges encountered during the collection of data is that it has not been possible to access good educational content because of copyright issues, especially when it comes to the material from Clio and Gyldendal. These platforms are broadly used by the interviewed Danish teachers and hold an enormous collection of contemporary school material in their possession. They cover a vast spectrum of subjects aligned with the Danish national curriculum (BØRNE- OG UNDERVISNINGSMINISTERIET, 2021). The inability to access materials from these sites poses a constraint on the depth of the data available for analysis in regards to the data.

### 3.2.3 *In Context Learning (GPT4)*

To assess the capabilities of in context learning with ChatGPT4, different types of tasks were tested. This includes grammar, reading, and fill-in-the-blank exercises and question generation based on selected themes. These task types were selected from the available material mentioned above.

***Grammar tasks*** focused on creating exercises involving verbs, grammar, adjectives, etc. The tasks included formulating sentences where students were required to conjugate words from present to past tense.

***Reading tasks*** involved creating texts around specific themes that were similar to the once collected for this study, followed by developing questions to the text.

***Fill-in-the-blank tasks*** used the selected material to assess whether ChatGPT4 could recreate and create both shorter and longer texts on various topics where the aim was to fill in the correct words into the tasks.

The purpose of this is to test and evaluate ChatGPT4's ability to accurately mimic and adapt to the linguistic demands and characteristics of Danish educational materials, while exploring the potential for AI-supported educational material. The findings from the test can be read in section *4 Findings.*

### *3.2.4 Local large language Models & GPT4*

In this step, same as ChatGPT4, the data was not collected from existing sources but generated from scratch based on the collected data. In this case the generation involved four different models *mistralai/Mistral-7B-Instruct-v0.2, meta-llama/Meta-Llama-3-8B-Instruct, TheBloke/Llama-2-7B-Chat-GPTQ, TheBloke/Llama-2-7B-Chat-GPTQ* and *ChatGPT4.* The models were chosen for their capabilities in natural language understanding and generation. The process to generate material was based on examples and information gathered from the collected educational material.

To test the LLMs, different assignments that mirrored the structure and educational goals of the collected materials were generated. For each model, 10 assignments focused on

generating questions for a provided text were made. Additionally, 5 assignments involved creating texts from a chosen theme followed by questions related to those themes.

For grammar testing, the approach was divided into three categories:

1. **Punctuation**: Here, 5 assignments test the models' ability to create assignments that can be used within Danish teaching.
2. **Verb Identification**: Another set of 5 assignments required the models to identify verbs within sentences.
3. **Verb Tense Conversion**: The final set of five assignments involved changing verbs from past to present tense in Danish, assessing the models' ability in handling verb tense transformations.

To see the output of the models look in the *appendix 4 & 5.* The section below will explain how the collected data will be analysed, which methods and techniques used for doing so. The section will start by explaining the qualitative analysis then In-context-leaning with ChatGPT4, and lastly the Local LLMs.

## 3.3 Methods for Data Analysis

### 3.3.1 *Semi-Structured Interviews - Qualitative analysis*

The collection of data from the teachers was based on the semi-structured interview form. This interview form allows for in-depth discussion by using a premade interview guide, while it allows the participants to express their views and experience beyond the predetermined questions in the interview guide. The interview guide was based on the key themes identified from the literature review which ensures that they are aligned with objectives of the research questions (Bekele & Ago, 2022). To view the interview guide look in *Appendix 9 - Interview Guide.*

All the interviews were conducted in danish to ensure clear and comfortable communication for the participants. The interviews were audio recorded with participant consent followed by a transcription process. The transcription in Danish captured nuance, emphasis, and detail, forming a dataset for further analysis (Kvale & Brinkmann, 2009). This approach ensured the integrity of the qualitative data, preserving the original context and meaning of each

interview. The complete transcriptions are available in the *Appendix 1- Interviews with colour-coding* for detailed reference.

Post transcription, the interviews were translated into English. This step was done for broader accessibility and understanding for readers, researchers, or stakeholders not proficient in Danish. The translation process was handled with care to maintain the fidelity of the original transcriptions (Kvale & Brinkmann, 2009). Afterward a systematic content analysis was conducted on the transcribed interviews to identify code, and categorise recurring themes and patterns (Linneberg & Korsgaard, 2019). To facilitate this analysis, a colour-coding system was applied:

Blue: Passages related to the use and impact of digital tools in education.

Yellow: Segments discussing student engagement in the context of digital learning environments.

Red: Concerning the preparation and differentiation of school materials, underscoring the challenges and approaches in tailoring educational content.

Green: Comments and viewpoints regarding the perceptions of ChatGPT and LLMs, reflecting teachers' attitudes and potential concerns.

Purple: Observations related to the implementation strategies of digital tools within the classroom setting.

This colour-coding facilitated an organised approach to the qualitative data analysis, (Linneberg & Korsgaard, 2019) allowing for identification and examination of themes directly related to the study's focus areas.

### 3.3.2 *In-context-learning*

For the methods used looking into the generating of school material the study used in-context learning principles. This was done by optimising the prompts to include contextual clues and learning objectives within the interactions. To do so different options are available some of these are The AUTOMAT Framework and The CO-STAR Framework (Vogel Maximilian, 2024). In this case the CO-STAR Framework has been used to facilitate the prompt-template and making sure that the overall prompt is well designed. This method includes setting Context(C) Objective(O) Style & Tone (ST) and Audience (A) and lastly Response (R) (Vogel Maximilian, 2024). To see the concrete prompt-template look in the *Appendix 2* -

*Prompt framework and templates*. Further techniques within ICL has been used:

***Zero-shot and Few-shot Learning Techniques:*** The research utilised a few-shot learning approach to evaluate the LLMs' ability to understand and emulate the essence of traditional Danish language teaching materials. By giving examples to ChatGPT4 ranging from only a few to more examples. The study aimed to explore ChatGPT4's efficiency in creating content that reflects the educational materials gathered from the different sources based on providing examples into the prompt.

***ReAct Prompting Methodology:*** Another approach tested with ChatGPT4 was ReAct prompting. This technique is notable for its iterative improvement process, wherein prompts are refined based on the LLMs' immediate outputs, creating a feedback loop that steadily steers the models toward generating more precise and meaningful content (Yao et al., 2022).

In addition to the methods discussed in the literature review, RAG will be used. This is beneficial in this research study due to the models that have been used have knowledge that is cut off at a specific date, that varies based on the concrete model (Rathod, 2024). This leads to the model not being able to incorporate new information to some extent after the post-cut off date. The limitation poses a challenge when it comes to generating new school material. Teachers who wish to use these models to generate assignments that use new information and reflect recent developments, will not be possible. Also, there might be a problem with the LLMs not having the information that the teachers are looking for (Rathod, 2024).

RAG addresses these issues by combining the generative capabilities of language models with the ability to retrieve new information, or additional information from external sources that the teachers have (Gao et al., 2023). By integrating RAG, it can be possible for the teachers to create assignments that are not only relevant but also up to date, while also being able to do exactly the things they need. Adding RAG enhances the educational value of the material, which ensures that the students receive assignments that are both informative and aligned with new and more correct information (Jauhiainen & Guerra, 2024). This approach aids in keeping the educational content up-to-date and engaging while developing students' awareness and understanding of ongoing global and local events.

### 3.3.3 *Local Large Language Models*

The models tested in this part of the study include *meta-llama/Meta-Llama-3-8B-Instruct, TheBloke/Llama-2-7B-Chat-GPTQ*, and *mistralai/Mistral-7B-Instruct-v0.2*. Similar to the approach used for testing ChatGPT4, Few-Shot prompting was applied to examine its effectiveness in making the mentioned local models generate new assignments that are similar to the collected data. However, unlike with ChatGPT4, another technique called CoT was used here. The objective was to evaluate how this technique would impact the output of the LLMs in regards to the quality of the generated assignments. Furthermore a crucial approach was the consistent use of identical prompts across different models and parameters set for the LLMs when generating the output.

All the local models were tested with the temperature parameter set to 0.01. When the temperature is set to a lower value, the model's output becomes more deterministic and predictable. This means that the model is more likely to choose words that are statistically more common in the context it has learned during its training phase (Ouyang, Zhang, Harman, & Wang, 2023). For instance, with a temperature of 0.01, the model will heavily favour the most likely words, resulting in more repetitive and less diverse text, but typically more coherent and on-topic (Ouyang, Zhang, Harman, & Wang, 2023). This parameter was set this low mainly for the two reasons below.

This parameter was set this low mainly for the two reasons below.

1. **Accuracy and consistency:** When generating educational content accuracy is important (Jauhiainen & Guerra, 2024). A lower temperature helps to ensure that the output is more predictable and stays closely to correct and typical uses of language. This is particularly important for language learning materials, where providing linguistically accurate examples is necessary to avoid teaching incorrect language usage.
2. **Reduction of Errors:** In educational contexts, reducing the potential for errors in generated content is critical. A lower temperature minimises the likelihood of introducing factual inaccuracies or grammatical mistakes, which could mislead students and make the output not sufficient for the teachers.

The next important parameter that was set was “top_p” sampling which is also known as nucleus sampling in text generation (Ouyang, Zhang, Harman, & Wang, 2023). The parameter helps to refine the randomness in the outputs from the models. Instead of selecting the next words from the entire vocabulary, it restricts the choice to a subset of words that have a cumulative probability as high as the set threshold, which in this case is 95%. In other words this means that the model will only take the most probable words that combined make up the 95% of the probability mass and ignore the rest of the words. This threshold was set to 95% due to the nature of what the models were used to test, generating educational material for the danish subject. Below is an elaboration of the reason behind setting this parameter.

1. **Balanced Content Generation**: Setting top_p to 0.95 gives a balance between creativity and accuracy. It allows for some variability in word choice (which is lower in a smaller top-p percentage like 0.50) while still focusing on high-probability words. This balance is critical in educational content where both engagement and correctness are key factors.

2. **Relevance and Contextuality:** By focusing on the most likely 95% of words, the generated content is more likely to be contextually appropriate and relevant to the subject matter. This is relevant in an educational setting, where material must be closely aligned with specific learning outcomes.

3. **Quality Control:** In school materials, especially for subjects like languages, science, or history, maintaining a high level of factual and linguistic integrity is essential. Top-p sampling ensures that the generated text does not stray into less probable, potentially incorrect usage areas, thereby upholding the quality of educational content.

The third parameter was setting a repetition penalty, in this case set to 1.15. The repetition penalty is a model-specific mechanism used for text infilling to reduce the chance of picking the same token again (Yasir, 2023). When the models generate content each time a word is used the future probability of being selected again is reduced by the factor specified (Yasir, 2023). This means that once a word is generated by the models the likelihood of it appearing again is multiplied by 0,87 (1/1.15).  This effectively decreases the chances of repeated use.

Importance of the this for generating school material

1. **Enhances Text Quality and Readability:** Applying a repetition penalty helps prevent the generated text from becoming monotonous and tedious due to excessive repetition of words or phrases.
2. **Prevents Content Stagnation:** In educational texts, it's important to present information in a clear and engaging manner. Reducing repetitiveness avoids stagnation in the content, making the learning experience more dynamic and effective.

### 3.3.4 *Validation approach*

The method to evaluate the generated assignments involved two experienced teachers. They review the outputs produced by the tested models in this study. The evaluation criteria was based on the practical usability of the material in their teaching. To facilitate this assessment by the teachers a binary evaluation scale was used, categorising the assignment output from the LLMs as either "*unsatisfied*" or "*satisfied*".

An "*unsatisfied*" rating from the teachers means that the output was rated as inadequate for integration into their teaching without substantial modifications to the output.

Conversely a "*satisfied*" rating means that the output was sufficient in quality to be used by the teachers in an educational setting, with or without minor adjustments to the generated assignments.

This evaluative method was based on a pragmatic measure of the effectiveness of the generated assignments. By doing the evaluation in this way, the feedback was easy to interpret and provided clear and structured feedback in a real world scenario. With the methodology established, it's time to present the findings from the qualitative and quantitative analyses, which provide insights into the practical benefits and challenges of using ChatGPT4 in education.

# 4 Findings

The results will be divided into three parts, due to the sequential nature of the study. Firstly, the findings from the qualitative study will be presented, which are based on the four

interviews conducted and the colour-coding. Next, the results obtained using ChatGPT4 will be showcased, particularly focusing on its ability to create some of the material that was collected. This includes tasks from online sources, contributions from teachers, and textbooks. Lastly, the findings from the local LLMs will be presented in a table to showcase the differences in how well they perform in generating school materials.

### 4.1.1 *Findings of Qualitative Content Analysis*

The findings revolve around five thematic areas—Digital Tools in Education, Student Engagement, Preparation & Differentiation of School Material, Views on ChatGPT and LLMs, and Implementation—identified through content colour coding. To view the whole content colour coding see *Appendix 1- Interviews with colour-coding.*

**Digital tools in *education*:** The interviews revealed that both teachers and students in the Danish primary school system are already well acquainted with digital platforms in many forms. This finding validates the themes from the literature review in regards to the prevalent use of technology in schools (Kasneci et al., 2023). Furthermore students were portrayed as adept users of digital tools due to having been exposed to and educated with various technologies from an early age (BØRNE- OG UNDERVISNINGSMINISTERIET, 2021). The access to personal devices like tablets and computers is a norm in the Danish educational setting, facilitating a natural interaction with digital learning materials already. "Each student having their own Chromebook has become the norm," one teacher noted, showing the widespread availability and use of technology among students in the Danish schools. Familiarity with digital tools both among teachers and students create an advantageous foundation for adopting new technologies like ChatGPT4. While there's a strong foundation of using digital tools in education, the teachers did point out that there's always room for improvement. They're looking for digital solutions that can do more and fit better with what they need for their teaching and their students' learning.

**Student *engagement***: Concerns regarding student engagement, as pointed out in the literature, were substantiated by the teachers' experiences. The trend of student attentiveness waning, as traditionally taught content becomes less interesting, combined with the outside influences of mobile phones. Teachers expressed the wish to be able to alter the way they sometimes teach: "If I had more time to prepare. I would be able to make the teaching more

exciting." This highlights that the interviewed teachers would like to try to make their teaching practices more engaging, but due to time limitations that is not possible.

**Preparation and *differentiation* of school material**: The issue of time-consuming material preparation and the challenges of addressing diverse student needs within the classroom were recurrent themes in the interviews. These concerns validate the literature's focus on the significant demands placed on teachers and the consequent impact on personalised learning. This is also the case in a Danish setting. Statements such as "I would say there are 4-5 in each class who have a diagnosis". This point from a teacher shows that they need to cater wider in their preparation of materials. Differentiated assignments are very challenging in practice, taking a long time to prepare, corresponding to academic insights about real challenges given to teachers.

**Views on *ChatGPT* and LLMs**: Reflecting the literature's optimistic yet cautious perspective on AI in education, teachers acknowledged the potential of ChatGPT while expressing uncertainties regarding the application and reliability. The willingness to use these technologies and the fears about how practically useful and reliable summarise the nuanced debate regarding the role of AI within education. The mixed sentiment that "Yes, I would say, if it could help me with alternative things. But it is not certain I can be sure that ChatGPT is correct " exemplifies the hopeful yet sceptical attitude towards emerging technologies. The interviews also pointed out a lack of knowledge in that these technologies can do it. The interviews underscored a broader call for guidance. Mostly with a desire for clearer directions from the Danish Ministry of Education regarding the integration and use of these advanced technologies.

***Implementation***: Clear implementation strategies and support are key for new digital tools when it comes to integrating them into a Danish educational setting. This reinforces the used literature’s view on adequate support and training for the teachers, when it comes to implementing it. The differences around the domain of digital literacy and the provision of guidance in this respect for instruction are a clear indicator of a gap between theoretical potential and practical implementation. Such statements from teachers as "I lack qualifications. I might need to get better, like everyone in the workforce" reflect this high need for targeted and timely training in specific technologies like ChatGPT4 before it can be used in education for preparing material for their students. New technologies are emerging at a rapid pace, and it's important for the educational authorities to provide timely, clear, and

actionable advice for educators. However, these are also things that probably cannot be done as of now, given the speed that the technologies move and develop.

In conclusion the interviews provided insights into the practical applications and challenges of integrating digital tools like ChatGPT into the Danish primary schools. These findings not only validate the potential identified in the literature review but also emphasise the unique context and requirements of the Danish educational system. There are a lot of different areas when it comes to getting close to implementing LLMs into the school system. Mostly the need for training and guidance for teachers are essential for ensuring possible implementations. After concluding on the qualitative findings it's time to move to the second part of the findings, ChatGPT4 ability to create school material.

## 4.2 LLMs

The testing of ChatGPT4 to create educational materials in areas such as grammar, reading, writing and more provided extensive findings, underlining the effectiveness of the model when guided by well-structured prompts and techniques. The implementation of Few-shot and ReAct prompting templates, which can be seen in *Appendix 2 Prompt Framework*, gave good quality and relevance to the generated assignments. This proves essential in tailoring the output to specific educational needs and objectives. The findings also integrate some of the main insights from the qualitative interviews to demonstrate if it can fulfil the needs that the interviewed teachers mentioned in the interviews.

To assess the performance of ChatGPT4 the evaluation mentioned in *3.3.4 Validation approach* was used. The two teachers reviewed the original assignments and compared them with 30 outputs from ChatGPT4 across specific grammar tasks, reading comprehension, and interactive content such as questions for Kahoot. Examples of both unsatisfied and satisfied tasks generated by ChatGPT4, as assessed by the two teachers, can be seen in *Appendix 3*.

**Grammar task:** In a simple task of converting present tense to past tense, ChatGPT4 performed exceptionally well, with 96% of the materials from the 30 tasks being satisfied. Some of the assignments could also have been put into a new “excellent” category based on the ability to easily differentiate the tasks in terms of difficulty levels. Unlike textbooks, which offer a single task for all students, ChatGPT4 could easily adapt tasks to be easier or more challenging as needed.

The ability to customise tasks has been highlighted both in the literature review and in the qualitative study, indicating that ChatGPT4 possesses some of the necessary capabilities for personalising educational materials. For further outputs from ChatGPT4, refer to Appendix 3.

The next assignment evaluated by the teachers involved verbs. Again ChatGPT4 showed a strong performance, with 94% of the 30 tasks evaluated by the teachers being Satisfied. Once more, the capability to modify the difficulty level was appealing to the teachers. The issues that were categorised as poor were due to minor errors, such as spelling mistakes.

For reading understanding assignments 92% were rated as 'Satisfied'. The ability to recreate a new text or to use ChatGPT4 to develop questions for a text was noted by the teachers as a unique opportunity to implement new materials into their teaching. An example was used to test this which involved the issues reported in the media about TikTok and the content children are exposed to (Skydsgaard Nikolaj, 2024). ChatGPT4 could easily create a text with questions that tested reading comprehension while also fostering a discussion on a relevant topic currently prevalent in the media. This test showcases that teachers now have a way to implement new material in many forms of difficulty and more. Additionally, the possibility of creating a Kahoot based on the text about TikTok to create a fun activity was tested. Again the teachers evaluated these types of assignments as "Satisfied". To view these tasks and other outputs and see whether the tasks were rated 'Unsatisfied or Satisfied'', refer to *Appendix 3 - Generated school material from ChatGPT4*. Furthermore an assignment based on a text from Cilo was created to see if the tasks ChatGPT4 produced would work in practice, and the teachers reported that it functioned well. This indicates that teachers can use ChatGPT4 to produce materials for creating various tasks that can be used in teaching already.

**Conclusion on testing ChatGPT4 & Qualitative Analysis**

The potential for employing technologies like ChatGPT4 in education is vast and might benefit both students and teachers. This study however primarily focused on teachers and the potential to use ChatGPT4 to enhance their teaching methods. The applications of ChatGPT4 in such a setting look promising, which has also been confirmed in the implications concerning the preparation and differentiation of school material presented in the colour coding analysis of *Appendix 8—The Results of Qualitative Study*. Now that the findings from

the qualitative analysis and the generated materials from ChatGPT4 has been presented we move on looking into the Local LLMs and ChatGPT4.

***Fine-Tuning (Local large language models)***

After testing the local models as mentioned in *3.3.3 Local Large Language Models,* different results came to light. The results can be seen for the two techniques, Few-Shot and CoT in table 2 and 3 below.

***Table 2 Few-Shot evaluation table on Local Models & ChatGPT4***

| Model | Assignment | Unsatisfied | Satisfied |
|---|---|---|---|
| **Grammar** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Grammar | **93,33%** | **6,67%** |
| ChatGPT4 | Grammar | **6,67%** | **93,33%** |
| meta-llama/Meta-Llama-3-8B-Instruct | Grammar | **66,67%** | **33,33%** |
| **Questions generation** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Questions generation | **20%** | **80%** |
| ChatGPT4 | Questions generation | **0%** | **100%** |
| meta-llama/Meta-Llama-3-8B-Instruct | Questions generation | **10%** | **90%** |
| **Text + questions generation** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Text + questions generation | **100%** | **0%** |
| ChatGPT4 | Text + questions generation | **0%** | **100%** |
| mistralai/Mistral-7B-Instruct-v0.2 | Text + questions generation | **100%** | **0%** |

*Tabel 3 Chain-of-thought evaluation table on Local Models & ChatGPT4*

| Model | Assignment | Unsatisfied | Satisfied |
|---|---|---|---|
| **Grammar** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Grammar | **93,33%** | **6,67%** |
| ChatGPT4 | Grammar | **6,67%** | **93,33%** |
| meta-llama/Meta-Llama-3-8B-Instruct | Grammar | **73,33%** | **26,67%** |
| **Questions generation** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Questions generation | **70%** | **30%** |
| ChatGPT4 | Questions generation | **10%** | **90%** |
| meta-llama/Meta-Llama-3-8B-Instruct | Questions generation | **30%** | **70%** |
| **Text + questions generation** | | | |
| mistralai/Mistral-7B-Instruct-v0.2 | Text + questions generation | **100%** | **0%** |
| ChatGPT4 | Text + questions generation | **20%** | **80%** |
| meta-llama/Meta-Llama-3-8B-Instruct | Text + questions generation | **100%** | **0%** |

Firstly the test of TheBloke/Llama-2-7B-Chat-GPTQ showcased so poor Danish capabilities that the study concluded not to test this further. For the other local models mistralai/Mistral-7B-Instruct-v0.2 and meta-llama/Meta-Llama-3-8B-Instruct showed better capabilities than Llama-2. Still the two teachers that reviewed the assignments for Llama-3 and Mistral-7B found that many of the generated assignments, which are mentioned 3.2.4 Local Large Language Models, were still unsatisfied and cannot be used in an educational

setting without many changes. To view the generated assignments and evaluation of the models performance look in the *Appendix 4 & 5*

To test the different models, Few-Shot learning was utilised giving the mode's a limited number of examples to quickly adapt to new tasks. This technique was chosen to assess the adaptability and efficiency of each model when exposed to examples. In the test ChatGPT4 outperformed Llama-3 and Mistral-7B across all tested assignments, particularly excelling in the domains of grammar exercises and text + question generation. Mistral-7B had a satisfaction rate of only 6.67% in grammar exercises, with none of its outputs being classified as satisfied in text + questions generation. Llama-3 showed an improvement with 33.33% of its outputs in grammar exercises rated as satisfied, and the same poor results 0% in text + question generation. ChatGPT4 achieved satisfied rates of 93.33% and 100% in these tests showcasing far better results. Furthermore, in the task of question generation, while all models demonstrated good results as illustrated in Table 1, ChatGPT4 continued to achieve the highest percentages. This shows that ChatGPT4 was best in performance across the various types of assignments generated.

These findings underscore the efficacy of ChatGPT4 in using Few-Shot to generate educational materials that are both accurate and suitable for immediate use in the teaching environments, with minimal need for revision. The evaluation criteria, as mentioned in *Section 3.3.4 validation approach* , were based on the practical applicability of the generated material in educational settings, with little to no modifications. The main problem with Llama-3 and Mistral-7B was the Danish language capabilities. The generated output from these models had many errors and mistakes, where it failed to produce coherent text that aligned with the objective of the generated assignments.

Furthermore the CoT technique designed to simulate step-by-step reasoning in generating responses, once again highlighted better performance for ChatGPT4 compared to Llama-3 and Mistral-7B in all the assignments. Again, the big difference was the ability to create coherent Danish assignments without major mistakes. The findings validate that ChatGPT4 has way better capabilities when it comes to producing coherent danish text, and a better understanding on generating the different assignments, based on the pragmatic evaluation.

In conclusion the tests using the CoT and Few-Shot techniques, confirmed that while all models are capable of employing advanced reasoning techniques and Few-shot examples to some extent, ChatGPT4 consistently delivers outputs that are more aligned with the needs of educational environments. The Danish capability makes it a more reliable and effective tool for educators seeking to incorporate AI-generated content into their education.

### 4.2.1 *RAG & ChatGPT4*

Based on the findings of the tested models it was chosen to implement ChatGPT4 with RAG in a Streamlit application, due to ChatGPT4 outpreforming to the others models tested. Integrating ChatGPT4 with RAG presents significant opportunities in the educational setting (Jauhiainen & Guerra, 2024). RAG enhances ChatGPT4 by allowing it to access and retrieve external material. The integration helps in ensuring that the generated content is not only relying on the pre-trained data that ChatGPT4 was trained on, but also gives the ability to incorporate new information or other relevant information that teachers want to use, to create new assignments.

### 4.2.2 *Implementation with Streamlit*

Streamlit will be used to combine the capabilities of ChatGPT4 and RAG. Streamlit was chosen due to its effectiveness in creating an interactive application that is easy to deploy.

The primary goal for creating a web application, that works like OpenAI's ChatGPT4 through Streamlit, was to allow teachers to input queries related to the material they need. The questions they ask in the application will be processed through the integrated ChatGPT4 model and the RAG system. Additionally, the application includes a built- in feature where teachers can upload their own documents, or input the relevant new information they want into the system. This helps to ensure that the outputs are relevant to the specific need the concrete teacher has. The application's interface is designed to be intuitive, while minimising the learning curve and making it accessible for teachers with different levels of technical expertise.

The reason for building the Streamlit application was to streamline the process of material generation. The integration of RAG ensures that the materials are current and up to date, while the user friendly interface supports easy accessibility. The development and

deployment of the web application represent a significant step towards integrating advanced AI technologies into the Danish primary school system by offering a practical tool that can be used by the teachers. To view the web application, which is uploaded to github with version control follow this link https://github.com/Mikkelrune/danish-education-gpt.

# 5 Discussion: LLMs in the Educational System

When it comes to integrating LLMs like ChatGPT4 into the Danish primary school system many factors and considerations need to be accounted for. This study, while pioneering in examinings the possibilities for a Danish context, does not include all necessary aspects for getting a full overview. A holistic examination would be ideal for drawing a complete conclusion on the benefits, concerns, and how to implement LLMs in the education system. This research has looked into the possibilities on the concrete benefits and concerns in a Danish setting by exploring educators' perspectives on LLMs.

The mentioning of an expert group was done in the theoretical framework. The establishment of this can be considered a significant move towards a more centralised decision making in regards to AI adoption from a policy side. From the beginning of writing this thesis a lot of new insight has come to light, including the report from the expert group. Sadly, the report does not bring forth any explicit recommendations on how the primary schools should implement LLMs into their teaching. The report, however, looks more into how to navigate the possibilities on how to change the current exams so they are not affected by students using AI to cheat (Vedersø Birgitte, Damsgaard Jan, & Sørensen Bent, 2024). The report lacks a view on how these AI's like LLMs can enhance the educational system both for teachers and the students. The research done in this study clearly indicates that LLMs can bring great opportunities for teachers in preparation of assignments, while also suggesting clear introductions, detailed explanations and the opportunity for getting consultation are key factors for adopting these technologies. It's hard to evaluate and discuss the concrete impact that the new recommendation will have, but it's a vital step in the right direction by putting focus on it by the policy maker's side.

Despite the promising results that ChatGPT4 showcased in generating the tested assignments it's not without issues that need to be addressed. One of the current problems with using ChatGPT4 and other LLMs include the possibility of generating misinformation that appears

reliable and concerns regarding data security. While data security in relation to creating new school assignments for teachers are less relevant, then if students are using these tools, it's still a potent area that needs to be dealt with. In this study the main goal is to ensure a high accuracy of the output, which in relation to students would also be the case. All the generated tasks, like other educational material, must be reviewed to ensure that the material generated is correct. ChatGPT4 and the other models tested, as mentioned, have a cutoff date up to which they have been trained. If new information or information that the models have not been trained on, the model will not be aware of this, and will most likely generate false information. This is where RAG as mentioned becomes valuable. By integrating RAG and ChatGPT as discussed with Streamlit, this problem can be mitigated.

Implementing ChatGPT4 with RAG into the Danish primary school system to generate educational material makes for an interesting discussion. The potential benefits of integrating are vast especially in enhancing the relevance of educational content. However, the challenges and concerns it raises cannot be overlooked.

Benefits of Integrating ChatGPT4 with RAG

1. **Current Content:** One of the advantages of RAG is its ability to fetch the most current data from various sources, ensuring that the educational material reflects the latest knowledge and trends.

2. **Personalisation:** ChatGPT4's ability to customise content could enhance education. Each student could receive materials tailored to their learning pace and style, potentially boosting engagement and understanding. For teachers, this could mean more time to focus on creative teaching methods and less on material generation.

Challenges of Integrating ChatGPT4 with RAG

1. **Accuracy Concerns:** Despite the capabilities of integrating RAG concerns still remain about the accuracy of the data retrieved. Incorrect or misleading data could be utilised in creating the teaching material, which propagates misinformation. This means that ensuring the reliability of data sources and the integrity of the content is of utmost importance and has to be something teachers know how to navigate and be

aware of. Furthermore, it can be discussed how well the RAG works in regards to making the generation more reliable.

2. **Teacher Adoption:** Teacher adoption is hindered by concerns over the trustworthiness of AI. In other words, they may be uneasy with AI-generated content, for example, wondering whether it meets standards of quality for education or is in line with the way they are going to teach. Another issue of concern is its readiness and practicality with which it can, in turn, be utilised. The last aspect might be the biggest concerns based on the interviews conducted, where teachers don't know how to use these tools, and what they can provide.

The discussion around integrating AI technologies like ChatGPT4 in general and with RAG requires a broad inclusive conversation with various stakeholders. Every group has some unique perspective and concerns, which become an integral part of understanding and implementing these technologies. Teachers and administrators are the forefront of the workforce in education. The acceptance and comfort of using AI tools become significant in making implementation successful. Teachers need assurance that AI will serve as a support tool to help them rather than a replacement. Therefore, the teachers need to be brought close to the technology through professional development programs and participatory workshops to allay fears over its reliability and demonstrate its practical uses in their teaching environment. Furthermore the policymakers need to create supportive frameworks that encourage ethical use of AI. This includes developing standards for data privacy, making sure that the LLMs are equitable, while also setting guidelines of how to use these so it aligns with the curricular goals that are in Denmark.

The application of the techniques used like Few-Shot and CoT prompting has in the study showcased that it enables ChatGPT4 to produce satisfied educational content with minimal input examples with Few-Shot, while also being able to show reasoning steps with CoT. The discussion explores how these methods with RAG, can further enhance ChatGPT adaptability to new subjects and tasks beyond the ones tested in this study. As presented in the findings section, ChatGPT4 demonstrated far better results compared to the other models Llama-3 and Mistral-7B, achieving high satisfaction rates by the teachers across the different assignments. In grammar exercises and text + question generation, ChatGPT4 demonstrated a satisfaction rate of 93.33% and 100%, highlighting its adaptability and efficiency.

Few-Shot learning gives the models a few examples with which to quickly adapt to new tasks. This was particularly effective during the educational creation for the danish subject, as ChatGPT4 could create output that were both accurate and appropriate with very little revision required. The success in this study supports the potential that Few-Shot prompting has in being expanded to encompass other topics and tasks within the Danish primary educational system. Given a constrained set of example input and output, Few-Shot learning could effectively guide ChatGPT4's response in its generation and, therefore, could potentially be invaluable for use within other subjects disciplines.

Similarly, ChatGPT4's potential was also demonstrated with CoT prompting, which instructs the models to develop intermediate steps or lines of reasoning used to reach a solution. In that way, it helped to nurture problem-solving skills in the model while, at the same time, offering a view of the reasoning process to aid in making transparent and interpretable the outputs generated through such a model. The success of CoT prompting in this study suggests it might be effectively used in other subjects, such as mathematics, science, and critical thinking exercises, where structured reasoning is essential. Furthermore by implementing RAG with ChatGPT4 it not only overcomes the problem with the cutoff-date but also gives the opportunity to provide ChatGPT4 with  up-to-date content across a range of the thought subject in the primary school. This gives ChatGPT4 more adaptability to both new topics that it has not been trained on, while maintaining the implementation of current events in Denmark and globally.

A discussion surrounding the rapid development of AI technologies must also be considered. The rapid development poses both opportunities and challenges for the possible integration in the educational system. The current pace of innovation necessitates continuous adaptation and updating when it comes to the educational tools to keep them relevant and effective. While this might be daunting, it also presents many opportunities to enhance the learning experience through the new technologies. Lastly OpenAI released their new model Omni, which showcases many of the things that this study looked into, but in a different way. With the Danish education system that has a high digital literacy and advanced infrastructure it is well positioned to leverage these new technologies like Omni, again, a need for training and guidelines are still considered as key factors before this can become a reality. It requires a collaborative effort among educators in the school, policymakers and technology experts to

ensure that the new technologies can be implemented in a correct and useful way. However, as stated, the rapid development makes it near impossible to keep up with the rapid movement and make up to date implementations at the current time. Nevertheless the possibilities of AI tools like ChatGPT4 can still be used to enhance productivity for teachers, and can be considered to have a positive impact on student learning.

In conclusion, the integration of ChatGPT4 into the danish primary school by using Few-Shot and CoT techniques with RAG holds benefits for creating new educational material. The adaptability and efficiency in the tested generated assignments suggest that it can also be used in other assignments and subjects. By addressing the cutoff-date and the possibility that ChatGPT4 is not trained on danish curriculum ensures continuous updates and flexibility to ChatGPT4. Meanwhile AI technologies are improving at a rapid pace, and it's hard to tell what the potential benefits and challenges will look like in just 6 months. The same goes for the correct implementation into the Danish primary school.

# 6 Conclusion

This study has explored the benefits and challenges while also looking into how to integrate LLMs into the Danish education system for teachers. With a focus on how teachers can use these technologies to enhance their capabilities in creating new and different assignments for their student needs, by leveraging techniques like Few-Shot, CoT and RAG.

Firstly the benefits and challenges regarding LLMs in a Danish educational setting will be concluded. The Danish interviewed teachers pointed out several benefits, this includes the high level of technology usage that already exists in their teaching practices, both for them but also their students. They are already familiar with a broad variety of different tools, and having expertise sets a solid foundation for adopting new technologies like ChatGPT4 into their teaching practices as well. Furthermore the access to computers, tablets and more, lays the infrastructure needed to adopt the new AI technologies. The need for adjusting their teaching material for different skill levels for their students is also a key point. Having more students with different needs puts emphasis on creating differentiated material. This is one of the benefits of using ChatGPT4, where the test showcased good performance in creating assignments that could be implemented into their teaching practices with differential levels.

The interviewed teachers all pointed out a lack of understanding of what these new tools can do, and how to use them. This is one of the major challenges identified in the data collection. It calls for educating the teachers and having clear guidelines on what exactly they can use these tools for, and how. Teachers are at the forefront of education and they need to learn how to use AI technologies like ChatGPT4 to help them enhance their teaching practices. Another challenge addressed was the trust in the LLMs. Here, teachers pointed out that they were not certain that the generated output could be trusted. This challenge was pointed out in the literature review, and can be considered a key factor. When generating educational material, it has to be trustworthy. This challenge has to be dealt with, which is why this study utilised RAG with ChatGPT4 to minimise the inaccurate generation of ChatGPT4. During this, the teachers have the possibility to upload the material they want to use, so the generated assignment retrieves the correct information uploaded by the teachers. Moreover, it mitigates the problem with cutoff-dates that the models have, while ensuring that it gets the correct information based on the Danish teaching objectives and goals.

Data privacy and ethical considerations must also be addressed to make sure a responsible use of AI in education. The successful adoption of these technologies rely and require a thorough training and support for the school teachers while the policy makers must ensure clear guides on how to use them.

The practical steps for implementing LLMs within the Danish school system involves many key components. Firstly the development of a user friendly application, like the one that was built in this study.  This web application ensures that teachers easily can access and use ChatGPT4 without high technical knowhow. Secondly, training programs must be put in place to equip teachers with the needed resources and skills so they are confident in using these tools in their teaching practices. This involves creating workshops at their schools, having technical people ready to help and more. Thirdly, partnerships with the major educational providers in Denmark to provide the correct data needed to enhance ChatGPT4. Lastly, addressing ethical and data privacy are areas that need to be looked more into. When only focusing on creating material the privacy concern is not as profined, but the ethical considerations are.

The successful implementation of LLMs in Danish primary schools requires a holistic approach that includes training, support, ethical considerations, policymakers, collaboration

with the established providers like Clio and Gyldendal while enhancing the benefits and minisming the challenges pointed out both in this study and in the literature review.

# 7 Limitations and Future Research

## 7.1 Limitations

During this study several limitations were faced that impacted the breadth and depths on the findings and conclusion. Firstly the research was constrained by copyright restrictions from educational content providers with Clio and Gyldendal. This led to limited material for analysis and testing. It hindered a full exploration on how LLMs could generate or enhance existing educational materials. This highlights the need and importance of future studies to find ways to cooperate with these platforms to get a more holistic view on the benefits and challenges of LLMs in the Danish primary school. Furthermore, the study was conducted within a set timeframe and by a single researcher.

This limitation had an impact on how many interviews could be conducted, a more deep understanding and more data in regards to the view from teachers could have given a deeper understanding of the practical challenges and benefits of integrating AI tools into Danish education. Additionally, more extensive testing of the different models across different subjects and tasks could have offered a broader perspective of their applicability and effectiveness.

## 7.2 Future Research

Future research should address these limitations and build upon the findings from this study. To proceed with further studies an extensive testing of RAG and ChatGPT4 in different subjects and content is necessary. While this study demonstrated some effectiveness of ChatGPT4 in generating Danish educational material, similar tests should be done in subjects like mathematics, science, history and more. Testing different subjects and tasks within the Danish subject will provide a more clear and detailed determination of the adaptability and performance of ChatGPT4. Expanded testing should involve interviewing more teachers to gain a better understanding of the practical challenges and benefits in a classroom setting, while getting their view on using LLMs in their teaching practices.

Addressing the limitation of the models cutoff-date and to ensure continuous up-to-date an API from selected news sources could be beneficial to implement into the RAG system with ChatGPT4. By implementing an API into the system, it would be  able to retrieve the latest information from reliable news sources all the time. Building upon integrating an API into the RAG system, the web application would also benefit from future work. Here, it should focus on enhancing the features and usability. This could include more intuitive interfaces, make more customisation options, and also include a feedback mechanism to allow the teachers to provide real time feedback. Building a mobile-friendly version could also be an area, which could enhance the accessibility and usage among the teachers, by making sure it's integrated into their everyday teaching practices.

Setting up and finding ways to implement training programs to equip teachers with the skills needed to utilise LLMs are key, as mentioned, without clear guidelines the implementation will be hard. Future research should explore the implementations of workshops, online courses and more to make sure that the teachers are well informed.

Investigating the long term impact of LLMs and other AI tools on education requires conducting long-term research to assess how the use of these influences student learning, engagement and achievement over time. The same goes for assessing the impact on teachers, and how they make educational material. Furthermore ethical use of  LLMs and AI in education are also areas that could provide insight into the success of implementing these tools and maximising their benefits. As pointed out in the literature review, many researchers point to the huge benefits for students using these tools. However, this study did not look into this area, future research should look into how these tools can benefit students. This can be done by testing the possibilities of getting help from these tools without waiting for help, or getting help during homework, and see how this impacts their learning.

By focusing on these areas, future research can build on the foundation laid by this study and contribute to the successful and sustainable integration of AI technologies into the Danish primary education system.

References

Aalborg Universitet. (2023). Generative ai at aau. Retrieved from https://www.students.aau.dk/practical/it/generative-ai-at-aau#

Aarhus Universitet. (2024). Nu må du bruge AI til dit speciale eller bachelorprojekt. Retrieved from https://studerende.au.dk/nyhedsvisning/artikel/nye-regler-nu-maa-du-bruge-ai-til-dit-speciale-eller-bachelorprojekt

BØRNE- OG UNDERVISNINGSMINISTERIET. (2021). *Lærerens digitale hverdag.* ().

BØRNE- OG UNDERVISNINGSMINISTERIET. (2024). Spørgsmål og svar om digitale hjælpemidler og snyd. Retrieved from https://www.uvm.dk/gymnasiale-uddannelser/proever-og-eksamen/regler-og-orienteringer/vejledning-om-digitale-hjaelpemidler-og-snyd/digitale-hjaelpemidler-og-chatgpt/spoergsmaal-og-svar-om-digitale-hjaelpemidler-og-snyd

Böwadt, P. R., Pedersen, R., & Vaaben, N. K. (2019). *Når verdens bedste job bliver for hårdt: En undersøgelse af, hvordan lærere har det i folkeskolen* Københavns Professionshøjskole.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in high-stakes decision-making with LLMs. *arXiv Preprint arXiv:2403.00811,*

EMU. (2018). Indsatsen for it i folkeskolen evaluering. Retrieved from https://emu.dk/grundskole/forskning-og-viden/paedagogisk-it/evaluering-af-indsatsen-it-i-folkeskolen

Gan, W., Qi, Z., Wu, J., & Lin, J. C. (2023a). Large language models in education: Vision and opportunities. Paper presented at the *2023 IEEE International Conference on Big Data (BigData),* 4776-4785.

Gan, W., Qi, Z., Wu, J., & Lin, J. C. (2023b). Large language models in education: Vision and opportunities. Paper presented at the *2023 IEEE International Conference on Big Data*

*(BigData),* 4776-4785.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., . . . Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv Preprint arXiv:2312.10997,*

Han, Z., Gao, C., Liu, J., & Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv Preprint arXiv:2403.14608,*

Jauhiainen, J. S., & Guerra, A. G. (2024). Evaluating students' open-ended written responses with LLMs: Using the RAG framework for GPT-3.5, GPT-4, claude-3, and mistral-large. *arXiv Preprint arXiv:2405.05444,*

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Hüllermeier, E. (2023). ChatGPT for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kvale, S., & Brinkmann, S. (2009). *Interview: Introduktion til et håndværk* Hans Reitzels Forlag.

Lampinen, A. K., Dasgupta, I., Chan, S. C., Matthewson, K., Tessler, M. H., Creswell, A., . . . Hill, F. (2022). Can language models learn from explanations in context? *arXiv Preprint arXiv:2204.02329,*

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv Preprint arXiv:2104.08691,*

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474.

Linneberg, M. S., & Korsgaard, S. (2019). Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal, 19*(3), 259-270.

Marciano, L., Camerini, A., & Morese, R. (2021). The developing brain in the digital era: A scoping review of structural and functional correlates of screen time in adolescence.

*Frontiers in Psychology, 12*, 671817.

Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials and limitations. *Computers in Human Behavior: Artificial Humans,* , 100022.

Ouyang, S., Zhang, J. M., Harman, M., & Wang, M. (2023). LLM is like a box of chocolates: The non-determinism of ChatGPT in code generation. *arXiv Preprint arXiv:2308.02828,*

Phung, T., Pădurean, V., Cambronero, J., Gulwani, S., Kohn, T., Majumdar, R., . . . Soares, G. (2023). Generative AI for programming education: Benchmarking ChatGPT, GPT-4, and human tutors. *International Journal of Management, 21*(2), 100790.

Rathod, P. (2024). Efficient usage of RAG systems in the world of LLMs. *Authorea Preprints,*

Skydsgaard Nikolaj. (2024). Statsministeren tordner mod sociale medier – vil hæve aldersgrænse. Retrieved from https://nyheder.tv2.dk/politik/2024-05-27-statsministeren-tordner-mod-sociale-medier-vil-haeve-aldersgraense

SOFTWARE MIND. (2023). How and why soft prompts are slowly replacing text prompts.

Svendsen, A., & Svendsen, J. T. (2023). Curricular goals relating to digital literacy and digital competences in upper secondary school in denmark. Paper presented at the *INTED2023 Proceedings,* 200.

Vedersø Birgitte, Damsgaard Jan, & Sørensen Bent. (2024). *Ekspertgruppen om ChatGPT og andre digitale hjælpemidler*. ().

Vogel Maximilian. (2024). The perfect prompt: A prompt engineering cheat sheet.

Wang, Y., Yao, Q., Kwok, J. T., & Ni, L. M. (2020). Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys (Csur), 53*(3), 1-34.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv Preprint arXiv:2210.03629,*

Yasir. (2023). Understanding the controllable parameters to run/inference your large language model.

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. Paper presented at the *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems,* 1-20.