GenAI governance: Analyzing Large Language Models in Public Administration Contexts

A thesis submitted in collaboration with

Udviklings- og Forenklingsstyrelsen and Aalborg University

Master of Science in Business Data Science

AAU Business School

Faculty of Social Sciences and Humanities

Proponents:

Nikolaos Andris

Research Supervisor:

Roman Jurowetzki

Date: June 3, 2024

Word Count: 17522

Abstract

This thesis explores the transformative potential of Large Language Models (LLMs) within Denmark's public administration, focusing on developing an intelligent conversational system to enhance citizen service delivery. In collaboration with Udviklings- og Forenklingsstyrelsen (UFST), this study investigates integrating Generative Artificial Intelligence technology, resulting in a user-friendly conversational system. The AI agent system aims to provide tax citizens easy and accurate access to relevant information by leveraging advanced technologies, including Retrieval Augmented Generation and vector databases. The study highlights the implementation of Retrieval Augmented Generation to enhance the conversational system's performance by combining data retrieval with conversational capabilities and generating precise and contextually relevant responses. The findings from this study are intended to guide towards a future where AI technologies significantly enhance citizen engagement, giving them easier access to correct information. By utilizing state-of-the-art technologies like Retrieval Augmented Generation, this research aims to set an innovative solution for UFST to consider implementing in the future.

Contents

List of Figures

Chapter 1

Introduction

This thesis examines the use of Generative Artificial Intelligence (AI) in the public sector, specifically in collaboration with the The Danish Agency for Development and Simplification (UFST). The aim is to create a proof-of-concept conversational system using Retrieval-Augmented Generation (RAG) technology. This innovative approach extends the capability of Large Language Models (LLMs) by incorporating new data that has not been part of the original training set, thus enabling the AI system to provide more accurate and contextually relevant answers. The inception of this idea followed a meeting with the technical lead of UFST and the supervisor, where it was quickly pivoted from exploring LLM biases—which are inherent due to the initial training data—to developing a solution that could offer tangible benefits to UFST. Given their interest in generative AI technology, it was decided to focus on a project that would be practically valuable.

A fixed template from LangChain was used to create the conversational agent, which was then adapted to meet the specific needs of this project (LangChain, 2023). Data was pulled from Skat.dk, focusing on sections relevant to people who want to work or study in Denmark and are presented in English. This focus was chosen for practical reasons: processing Danish language data presents additional challenges, and newcomers to Denmark are more likely to need information on tax matters. This demographic also tends to generate high inquiries, potentially burdening customer service. Therefore, a conversational system aimed at this group could reduce UFST's operational costs and increase efficiency by allowing employees to focus on more productive organizational tasks.

The public sector has increasingly turned to AI technologies to improve service delivery, efficiency, and citizen engagement (Slava Jankin Mikhaylov, 2018). The potential for AI to revolutionize public administration is vast, yet it also comes with significant challenges, particularly concerning data accuracy, user trust, and system integration. The integration of AI into public services is not just about technology adoption but also about reshaping processes and interactions to serve citizens better. The collaboration with UFST aims to address these challenges by leveraging advanced AI capabilities to provide more accurate and reliable information, thus enhancing the overall user experience.

Generative AI, specifically LLMs like GPT-3, have shown remarkable abilities in understanding and generating human-like text (Dipankar et al., 2023). However, these models are trained on large datasets up to a certain cutoff point and do not update automatically. This limitation means they might not provide the most current or context-specific information necessary for accurate decision-making in dynamic environments like tax administration. RAG technology addresses this gap by combining the generative power of LLMs with real-time data retrieval, ensuring responses are grounded in the latest available information (Shuster et al., 2021).

Generative AI has the potential to enhance the quality and efficiency of public services significantly (Dipankar et al., 2023). AI can free up human resources for more complex and value-added activities by automating routine tasks, such as answering frequently asked questions. This is particularly relevant in the context of UFST, where the volume of inquiries can overwhelm traditional customer service channels. Implementing a conversational system can streamline these interactions, providing timely and accurate responses while reducing the workload on human agents.

The implementation of generative AI in public services has its challenges. One primary concern is hallucinations, where AI systems generate plausible but incorrect or misleading information (Shuster et al., 2021). This can be particularly problematic in a governmental context, where information accuracy is paramount. Grounding techniques, which augment AI models with specific, verified data, can mitigate this issue by ensuring that AI outputs are anchored in reality.

The development of a RAG-based conversational system for UFST represents a significant advancement in applying generative AI in the public sector. By addressing the limitations of traditional LLMs and incorporating web page data retrieval, this project has demonstrated the potential for AI to improve service delivery and operational efficiency in governmental contexts. However, this is just the beginning. Future work will focus on further refining the system, incorporating additional data sources, continuously evaluating and deploying the whole application to meet the evolving needs of public administration.

Chapter 2

Related Work

2.1 AI in the Public Sector and Concerns

2.1.1 Introduction to AI in the Public Sector

AI holds significant promise for fostering innovation within the public sector. Public sector innovation involves the introduction of new, novel, and original elements into public organizations or services. Innovation, particularly public sector innovation, differs from invention because the latter is mostly about business products and patents. However, the former focuses on adopting the best practices other agencies or employee workgroups already use rather than creating an entirely new product, process, or service. Innovations arising from employee workgroups are particularly effective in improving organizational processes and service quality, as these employees are closely engaged with day-to-day operations and have a better understanding of practical challenges (Demircioglu, 2023). At UFST, as mentioned further below, some generative AI workgroups are taking place to discuss new possibilities and share feedback regarding their projects.

One of the public sector AI implementations is in the health care system, specifically protecting and promoting the health of populations. AI can enhance the efficiency and effectiveness of processes across an expanded public health system. For example, in health protection, AI can analyze data patterns for real-time surveillance and disease detection, such as identifying malignant alterations. AI's ability to detect minute changes, which the naked eye might miss, significantly improves diagnostic accuracy. Additionally, when healthcare experts are unavailable, AI can provide emergency tele-assistance in dental emergencies. AI in medical imaging, such as detecting oral cancer, matches or surpasses human radiologists' accuracy. Machine learning also facilitates automated evidence synthesis, simplifies data entry, enhances the efficiency of information processing, and bridges the gap between doctors and patients. (Tariq et al., 2021)

Furthermore, AI technologies have been instrumental in transforming service delivery in various government functions. In smart cities, AI applications like predictive analytics and machine learning models effectively manage urban infrastructure (Amal Ben Rjab, 2019). These technologies enable continuous monitoring of traffic patterns, optimizing energy usage, and improving waste management systems. Specific implementations include improving traffic routine decisions to reduce congestion, predicting parking availability, intelligent routing for trip planning, real-time transport regulation, and creating intelligent systems for risk prevention in transportation management. AI also facilitates customer relations through virtual assistants that provide real-time information and support for citizens and tourists regarding public transport and city services. Additionally tourism leverages AI to provide real-time recommendations for tourists, while smart homes use AI for resident needs, temperature control, water management, and security. These advancements highlight AI's role in creating innovative, responsive urban environments (Amal Ben Rjab, 2019).

AI can also foster trust among citizens by increasing transparency and accountability in public administration. (Zuiderwijk et al., 2021) discuss how AI systems, through their ability to process and analyze vast amounts of data, assist public administrators in making more informed and evidence-based decisions. This capability improves the quality of governance and public services.

The introduction of AI in public services is subject to the dynamics of the social contract between citizens and the government. Transparency, explainability, and citizen engagement are crucial for building trust and acceptance (Schmager et al., 2024). AI-powered agent syestems and social media monitoring can enhance communication and gather citizen feedback, aiding policy development (Androutsopoulou et al., 2019).

The potential of AI to improve communication and collaboration among various stakeholders is notable. It helps minimize time and expenses, facilitates resource allocation, and adeptly handles complex tasks crucial in public administration. One notable example of AI enhancing public health systems is its application in predicting and managing disease outbreaks. During the COVID-19 pandemic, AI models were pivotal in tracking the spread of the virus, analyzing infection patterns, and predicting future hotspots. This allowed governments to allocate resources efficiently, implement timely interventions, and minimize the impact on public health. AI-driven platforms provided real-time data, helping public health officials make informed decisions, ultimately saving lives and mitigating the spread of the virus (Hu, 2020).

In education, AI has transformed the way public educational services are delivered. Personalized learning systems use AI to tailor educational content to the needs of individual students, thereby improving learning outcomes. AI-powered analytics provide educators with insights into student performance, helping identify those needing additional support and enabling more effective intervention strategies. For example, AI can analyze students' learning patterns and suggest personalized learning paths, ensuring each student receives the support they need to succeed(Kalyani, 2023).

AI's role in public safety and law enforcement is also significant. Predictive policing algorithms help law enforcement agencies anticipate and prevent criminal activities by analyzing crime data and identifying patterns. This proactive approach allows for better allocation of police resources and more efficient crime prevention strategies. Additionally, AI-powered surveillance systems enhance public safety by monitoring public spaces for suspicious activities, providing real-time alerts to law enforcement, and improving response times (Yadav, 2023).

For a successful AI adoption in the public sector, it is essential to take a comprehensive approach, addressing not just the technical aspects but also the ethical and governance dimensions(Ramíres-Hernández et al., 2023). Potential benefits include improved decision-making, personalized services, resource optimization, and sustainability. However, risks related to AI must be carefully managed, with a focus on transparency, accountability, and citizen engagement(Zuiderwijk et al., 2021).

Accountability plays a pivotal role in establishing trust in AI. It involves taking responsibility for AI actions, decisions, products, and policies, whether it is the AI system itself, the teams developing it, or its impacts. Responsible AI necessitates mitigating human and social risks while fostering socially beneficial applications. The concept of trustworthy AI advances the idea that trust must be established in AI development, deployment, and use to harness its advantages while moderating risks (Ghallab, 2019).

AI policymakers have to ensure that citizens are seen as partners in developing and implementing AI technologies. This includes engaging the public in meaningful dialogue, embedding digital ethics in education, and developing a policy framework that addresses AI's ethical, philosophical, social, and psychological impacts. Moreover, governments should fund research on AI risks, produce clear policy recommendations, and take a global lead in advocating for ethical AI practices and protecting rights, privacy, and safety (Boyd and Wilson, 2017).

It is concluded that the potential adverse effects of AI primarily stem from two sources: (1) a misunderstanding of how AI produces specific outcomes, including biases or errors; and (2) the exploitation of AI to favor certain individuals, organizations, or sectors within society. It appears that some aspects of AI's negative impacts are attributable to human biases and the data created by humans, while others arise from inherent features of AI technologies. Additionally, there are other critical issues related to AI's complexity that warrant further investigation, such as the challenges AI faces in replicating human behavior, the opaqueness and lack of clarity in AI processes, and the shortage of high-quality, unbiased data for AI training (Valle-Cruz et al., 2024).

Furthermore, the implementation of AI in public administration has yet to catch up with its adoption in the business sector. This has led to a more recent yet rapid exploration of AI technologies within public sectors (Desouza et al., 2020). Examining the distribution of AI research documents and citations by sector reveals significant insights. Particularly after 2015, AI research has notably increased within the public administration and business sectors, as depicted in Figure 1. In 2000, about 30% of documents were related to public administration. This proportion grew to 35% by 2010 and reached 40% by 2017. However, the most significant increase came after the COVID-19 pandemic, with the share of documents pertinent to public administration increasing to 45%. This trend is also mirrored in the citations: from only 7% of all citations in 2000 related to public administration, the figure more than doubled to 16% by 2010. In the past three years, particularly exacerbated by the pandemic, the significance of AI in public administration has become even more pronounced, with citations now exceeding 35% of the cumulative number (Aristovnik et al., 2024).

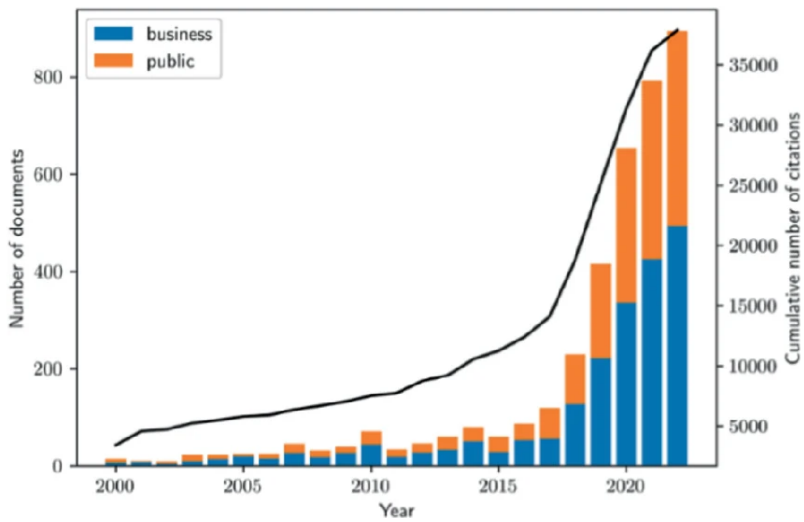

Figure 2.1: Distribution of documents and citations in AI research by sector 2000-2022.(Aristovnik et al., 2024)

These trends underscore the growing recognition and integration of AI technologies in public administration, highlighting their potential to significantly transform public services and governance.

2.2 Conversational Systems in the Public Sector

Conversational systems have increasingly become integral to the public sector, transforming how government agencies and public institutions interact with citizens. These AI-driven conversational agents are designed to provide immediate, accessible, and efficient responses to various inquiries, from answering routine questions to offering support in more complex scenarios like healthcare and social services. The adoption of AI conversational agents aims to enhance service delivery by improving response times, reducing operational costs, and ensuring 24/7 availability (Ali, 2022).

2.2.1 Question-Answering Systems

Question-answering (Q&A) systems are designed to provide users with precise and timely responses to their queries, enabling them to pose questions in natural language. These systems efficiently search and filter answers from a collection of potential documents. For instance, a semantic biomedical question-answering system has demonstrated its ability to retrieve exact and ideal answers to natural language questions within the medical field (Sarrouti and Ouatik El Alaoui, 2020). Similarly, an advanced question-answering method for construction safety hazard knowledge utilizes deep semantic mining, showing the systems' applicability in specialized industry settings (Tian et al., 2023).

Traditionally, Q&A systems were highly specialized, employing rule-based templates to manage narrow, structured data sets for answering domain-specific inquiries. From the Turing test (Rapaport, 2005) to the creation of Eliza (Duggan, 2016), and the development of the LUNAR system (Woods, 1973), the evolution of

Q&A systems is clear. With advancements in information and internet technologies, these systems have shifted from specialized to more versatile frameworks, although they remain confined to natural language interactions (Lee et al., 2023).

2.2.2 Role of Conversational Systems in Public Services

The public sector typically provides information on public services to citizens via comprehensive government portals, specific websites, and web-based dialogue systems, such as benefits.gov in the USA. However, these methods encounter multiple challenges. One significant issue is the lack of interoperability, which incurs substantial costs for the European Union. Overcoming these challenges requires adopting standardized frameworks and guidelines that promote interoperability across various systems and platforms. The European Interoperability Framework (EIF) provides a set of recommendations to improve interoperability between public administrations in Europe. This framework emphasizes the importance of using common standards, fostering cross-border collaboration, and leveraging digital technologies to enhance public service delivery (Madrid, 2012). Additionally, there is a problem with the unnatural interactions between citizens and these websites. This unnaturalness often stems from rigid, form-based interfaces that require users to navigate complex menus and input fields, making the experience complicated and unintuitive, often leading to user frustration and inefficiency in obtaining relevant information. Another concern is the absence of personalization in the provision of public service information. Conversational agents, on the other hand, facilitate natural language conversations between humans (users) and computers, enhancing interactivity and ease of integration across both straightforward websites and social media platforms like Facebook (Androutsopoulou et al., 2019).

A distinct feature of AI systems is their capacity to learn from multimodal inputs and adapt their responses based on that data. If the AI system is designed to handle new tasks or inputs without seeing similar data before, it will utilize zero-shot learning. This requires the system to leverage preexisting knowledge to generalize from known to unknown tasks instantly. If the system can adapt its responses after exposure to a few examples of the new data type or task, it will employ few-shot learning. The learning process is fast but takes time, requiring minimal additional training.

In various applications, including public administration, AI-driven technologies are employed as a form of algorithmic bureaucracy. These applications frequently utilize Natural Language Processing (NLP) techniques and employ conversational agents, or "chatbots," as mediums to deliver information and services to citizens. For instance, a study involving the Norwegian chatbot "Kommune-Kari" highlighted its effectiveness in providing detailed support on public services, aiding navigation of municipal websites, and addressing general inquiries. The study showed that most participants found the chatbot useful for obtaining quick, accurate information and appreciated its 24/7 availability, which significantly enhanced their interaction with local government services, demonstrating the potential of chatbots in public service delivery (Abbas et al., 2023). A chatbot is a computer program that interacts with users using natural language processing technology (Aoki, 2020). It is a form of narrow AI that extracts meaningful information from free texts based on user input and helps to "find the intent of the question asked by a user and send an appropriate reply". A chatbot is an AI-labeled product. It features a chat interface whereby the user converses with the app (Aoki, 2020),(Goyal et al., 2017).

Androutsopoulou et al., 2019 suggested a framework for utilizing chatbots in the public sector to enhance interactions between the government and its citizens. This methodology relies on natural language processing, machine learning, and data mining technologies, and uses available data to create a new digital communication channel. The key contribution of their research is the integration of various tools and services designed to address the varied communication needs between citizens and government. The research underscores the promising applications of chatbots in public sector contexts.

Chatbots play a crucial role in enhancing public services by providing an alternative to front office staff for addressing routine inquiries and delivering basic information. They can reduce customer service costs by up to 30% for companies but need to be tailored to the specific needs of users. Additionally, chatbots can support internal communication, handle various public service inquiries, and enhance efficiency by managing routine tasks, thereby allowing human staff to focus on more complex issues and improving overall service delivery (NOGA, 2023).

While chatbots have improved customer service in government offices, there is still room for improvement, as guided or directed chatbots with predefined responses and menus may only partially address some citizen queries and situations (Ramíres-Hernández et al., 2023).

Chatbots in public administration can assist citizens in resolving queries and availing government services without involving administrative staff. However, previous e-government research suggests that adopting new technologies does not necessarily lead to improved public services, raising questions about whether chatbots can transform traditional services into digital, integrated transactions (van Noordt and Misuraca, 2019).

Overall, chatbots have the potential to enhance service quality in public services by improving interaction, entertainment, problem-solving, trendiness, and customization for citizens (Misischia et al., 2022). However, their successful implementation requires addressing challenges such as building trust, ensuring accuracy, and providing multilingual and multi-communication mode capabilities (Aoki, 2020). A critical technological advancement that facilitates the creation of efficient, conversational systems is grounding. Grounding technology allows chatbots to reference and incorporate real-world knowledge and context into their interactions. By ensuring that responses are accurate, relevant, and coherent, grounding enhances the chatbot's ability to understand user inputs and provide contextually appropriate responses. This makes chatbots more effective and improves the overall user experience, thereby addressing critical challenges in their deployment in the public sector (Berger, 2023).

2.2.3 Conversational Systems Applications in the Public Sector

Conversational systems are increasingly being adopted in the public sector for various applications. Virtual assistants like chatbots and voicebots are being used by public institutions such as courts, prosecutors' offices, city offices, hospitals, and government ministries for customer service and information dissemination (NOGA, 2023). In Norway, citizens have shown a generally positive attitude towards the use of AI in public services, owing to factors like trust in government, human oversight, and perceived transparency (Slava Jankin Mikhaylov, 2018).

Some examples include the city of Vienna's WienBot, a digital assistant that provides quick and accessible information on around 350 different topics, including public services like parking availability. The city of Bonn's GovBot assists citizens with administrative tasks such as booking appointments and filling out forms. Latvia's UNA chatbot helps with frequently asked questions regarding enterprise registration, significantly reducing the administrative burden by handling repetitive queries. These chatbots enhance citizen engagement by providing immediate assistance and facilitating access to public services (van Noordt and Misuraca, 2019).

Conversational systems can facilitate better interaction between government and citizens by providing easy access to information, context recommendations, and messaging platform integration (Androutsopoulou et al., 2019). They can help answer frequently asked questions and conduct transactions, relieving staff from boring tasks (van Noordt and Misuraca, 2019). In Japan, local governments are exploring chatbots for citizen services, building public trust through understanding the technology, demonstrating competency and empathy, and communicating the purpose (Aoki, 2020).

Chatbots show promise in augmenting and transforming public service delivery by improving accessibility, efficiency, and citizen engagement (Weidinger et al., 2022). However, their adoption requires careful consideration of trust, transparency, and ethical issues (Aoki, 2020). In some cases, the adoption of AI has led to minor organizational changes, but the underlying processes remain largely unchanged.

2.2.4 Advantages of Chatbots

Chatbots have gained significant attention due to advancements in artificial intelligence, the internet, and social networking sites. They offer several benefits across various domains, including customer service, public administration, smart homes, and the Internet of Things (IoT) (Nirala et al., 2022).

Advantages in Customer Service:

1. **Cost Reduction:** Chatbots can reduce customer service costs by up to 30% by automating routine inquiries and tasks (Abbas et al., 2023).

2. **24/7 Availability:** Chatbots provide round-the-clock assistance, ensuring customers receive prompt responses without waiting for human agents (Nirala et al., 2022).

3. **Personalized Experiences:** Chatbots can maintain customer profiles and preferences, enabling personalized interactions and recommendations (Androutsopoulou et al., 2019).

Advantages in Public Administration:

1. **Improved Service Delivery:** Chatbots can answer frequently asked questions and conduct transactions, relieving staff from mundane tasks and improving overall service delivery (van Noordt and Misuraca, 2019).

2. **Cost Savings:** Implementing chatbots and voicebots in public institutions can provide cost savings by automating routine inquiries (Abbas et al., 2023).

3. **Accessibility:** Citizens can access government services and information through chatbots, enhancing accessibility and convenience (Nirala et al., 2022).

Chatbots offer numerous advantages, including cost reduction, 24/7 availability, personalized experiences, improved service delivery, cost savings, and enhanced accessibility. As AI technologies evolve, chatbots will play an increasingly significant role in various sectors, revolutionizing how organizations interact with customers and citizens (Abbas et al., 2023). For example, IBM's analysis indicates that chatbots can reduce operational costs significantly by handling routine inquiries, thus minimizing the need for extensive after-hours staffing. This improves efficiency and enhances employee satisfaction by allowing human agents to focus on more complex and rewarding tasks (Holdsworth, 2024).

2.2.5 Augmentation versus Automation with Conversational Systems

Conversational systems can augment human capabilities by automating specific tasks, allowing human intervention and empathetic interaction. Automation refers to a system that performs functions previously done by humans, while augmentation involves enhancing human capabilities through technology. Chatbots automate conversational interactions (Yoda, 2019). They augment human abilities by handling routine queries and tasks, freeing up humans for more complex interactions (Morris et al., 2018).

Conversational agents can be designed to provide emotional support and companionship, augmenting human connections through carefully crafted narratives and character development. The dialogue between the system and users is essential for creating realistic and supportive interactions (Pan et al., 2023). However, the upper bound of a chatbot's ability to express empathy remains an open question, as people may always prefer human interactions for emotional support (Morris et al., 2018).

In summary, conversational systems automate conversational tasks while augment human capabilities through emotional support, companionship, and handling routine queries. They cannot fully replace human empathy and emotional intelligence but enhance and complement human interactions in various domains (Nishimoto, 2022).

2.3 Barriers to Implement AI - Trust is the Key

Implementing AI in organizations faces several key challenges. Technological hurdles include data quality issues, lack of specialized AI skills, and difficulties integrating AI with existing systems. Organizational barriers involve cultural resistance to change, difficulties managing AI talent, and aligning AI projects with business strategy (Ångström et al., 2023). There are also ethical concerns around bias, privacy, and lack of transparency in AI systems (Medaglia et al., 2023),(Aoki, 2020).

To overcome these challenges, organizations need robust data governance practices, investment in AI skills development, and change management strategies to build trust and adoption (Aoki, 2020). Cross-sector collaboration between government, academia, and industry can provide complementary expertise. However, such collaborations require careful management of risks around data sharing and aligning incentives (Slava Jankin Mikhaylov, 2018).

Ultimately, realizing AI's potential requires situating the technology within the specific organizational context through grounding it in relevant data, bounding its scope, and recasting AI solutions to align with strategic objectives (Kemp, 2023). As AI evolves into a general purpose technology, innovation dynamics will likely shift, necessitating new organizational capabilities (Jacob R. Holm and Lorenz, 2023).

2.3.1 Ethical, Legal, and Social Implications

The ethical, legal and social implications of AI implementation in public administrations are multifaceted. Ensuring transparency and explainability of AI systems is crucial for upholding ethical principles like fairness, accountability, and privacy (Floridi et al., 2018) which are long-standing requirements for public services (Weidinger et al., 2022). Citizen trust and acceptance are crucial, as public concerns can hinder AI adoption initiatives. Providing transparency into AI processes, data usage, and model workings can foster positive citizen attitudes (Pflanzer et al., 2023).

However, the lack of explainability in current generative AI (GenAI) models poses a challenge for public administrations. Legal requirements to justify decisions, such as in customs administrations, necessitate explainable AI systems. Additionally, GenAI usage requires transforming civil servants' ethos through training and resources (Floridi et al., 2024). Balancing individual rights and collective good is another consideration. While AI can enable efficient service delivery (Schmager et al., 2024), oversimplification and loss of individual case nuances are risks. Governance frameworks, impact assessments, and data governance strategies are needed to address these challenges (Valle-Cruz et al., 2024).

2.3.2 Trust for Public Organizations

Trust is paramount for public organizations. It underpins the legitimacy of public action and is crucial for fostering citizen cooperation in policy implementation. Trust is necessary for public initiatives to maintain their legitimacy, and the effectiveness of policy implementation could be severely hampered (A. Burcu Bayram, 2021).

In public communication, studies have shown that transparency and the provision of symbolic information can significantly boost trust. These elements help demystify the actions of public organizations, making them more understandable and acceptable to citizens (Alon-Barkat, 2019). However, in algorithmic governance, a lack of trust may make citizens feel excessively monitored by public authorities, creating a sense of surveillance rather than service from public institutions (Meijer and Wessels, 2019). Additionally, there is a general reluctance to engage with machines if there is no initial trust, as highlighted by numerous studies on human-machine relationships, which suggest that trust must be established early in user interactions with technology (de Vries et al., 2003).

A recent study presented at HCI International 2024 explored citizens' trust in AI within public services, performed in the context of a Norwegian public organization (Schmager et al., 2024). It established a team about five years ago to explore data analytics and AI to deliver more efficient services responsibly. The team developed a model to predict the length of sick leaves to aid case handlers by focusing their efforts where most needed. The authors developed an interactive prototype mimicking a public service agency portal. They included a predefined interaction sequence with notifications about optional AI-based predictions, various types and levels of information about the predictions, and consent options. The prototype was designed through the social contract lenses framework, addressing power structures, rights and duties, transparency, and the necessity for consent without negative consequences. Twenty participants, aged between 18 and 65, reflecting the general population on sick leave in Norway, were recruited for the study. The data collection included three stages: collecting general participant data and self-assessments, conducting a moderated user study with task-based interaction with the prototype, and follow-up questions about their experience. Initially, participants had varied comfort levels when using AI before interacting with the prototype. While some participants felt comfortable (with half rating their comfort level as high, 4 or 5 out of 5), others were neutral or less comfortable. Following direct interaction with the AI prototype, the comfort level improved significantly for many participants, with around 40 % reporting increased comfort. This indicates that experiential interaction can positively shift perceptions and increase trust in AI applications within public services. These findings underscore the importance of transparency and user engagement in building trust in AI technologies within the public sector (Schmager et al., 2024).

The performance of devices such as chatbots is crucial for establishing trust. In Kawasaki City, a chatbot used for parental support did not meet user expectations. This initiative faced significant issues when the chatbot began generating inappropriate and insensitive responses to users' queries. For instance, it reportedly gave inaccurate advice regarding parenting, which led to public criticism and concerns over the reliability and safety of such AI-driven services highlighting the need for chatbots to deliver competent and empathetic responses to complex queries to maintain trust. Further analysis from a social contract perspective indicates that high levels of trust in government, clear value propositions of AI, and transparency in operations are fundamental to fostering a positive reception of AI technologies.

These elements enhance the social contract between the public sector and citizens and ensure that AI implementations act in the best interests of society (Schmager et al., 2024).

Transparency remains a critical cultural value, especially in Nordic contexts where the openness of government operations is highly valued. Adequate transparency involves not just the availability of data but also mechanisms that allow monitoring of public decisions, thereby increasing government accountability and trust. This transparency is vital for maintaining a balance of power between public service authorities and citizens, thereby increasing the likelihood of exposing wrongdoings or abuse of power (Grimmelikhuijsen et al., 2013),(Felzmann et al., 2019).

Chapter 3

Advancements in AI: From Foundations to Cutting-Edge Apps

AI advancements have dramatically transformed the field of conversational systems. Initially, natural language processing (NLP) enabled basic text understanding, which was further revolutionized by deep learning techniques. Introducing transformers like BERT and GPT allowed for more sophisticated language models capable of understanding and generating nuanced human language. Generative AI and large language models (LLMs) such as GPT-3 brought a new level of fluency and coherence in responses. However, challenges like hallucinations and the need for massive data persisted. Innovations like in-prompt learning and fine-tuning improved model adaptability and relevance. The development of grounding techniques ensured that AI responses were more accurate and contextually relevant. Eventually, Retrieval-Augmented Generation (RAG) emerged, combining generative models with real-time retrieval of relevant data from extensive databases, thus addressing previous limitations. RAG technology offers a promising solution for creating advanced conversational systems, providing more accurate, reliable, and context-aware interactions by integrating internal organization's data retrieval with generative capabilities.

3.1 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and humans using natural language. NLP aims to enable computers to understand, interpret, and generate human language in a meaningful and valuable manner (Nadkarni et al., 2011). It integrates computer science, linguistics, and machine learning elements to create systems capable of comprehending text and speech. These systems can recognize context, extract information, and produce responses that mimic human language. NLP originated in the 1950s at the intersection of AI and linguistics and has since evolved into a multidisciplinary scientific field. The main objective of NLP is to derive semantic meaning from text, such as understanding the relationship between the words "eat" and "food". NLP models achieve this through statistical methods, which reveal probabilistic connections between words based on the training data rather than correctly understanding the words themselves.

The initial step in NLP involves tokenization, breaking down text into basic units using delimiters to separate words. This allows for statistical analysis of the relationships between these units (Webster and Kit 1992). Like most machine learning projects, NLP involves several additional data preprocessing steps to streamline the data pipeline. These steps include lemmatization, which reduces words to their root form (e.g., "reduce" is the lemmatized form of "reducing"), ensuring the machine understands that the words carry similar meanings. Part-of-speech (PoS) tagging is another step that helps highlight semantic relationships between words and reduces ambiguity regarding their roles in sentence structure. All these preprocessing steps are crucial for using NLP in conversational AI (Nadkarni et al., 2011).

In conversational programming, NLP enables a machine to learn sentence structure and semantic relationships through statistical inference. This allows the machine to "understand" conversations, predict similarities through cluster analysis, and even predict the next word in a sentence or sequence.

3.2 Deep Learning

Building upon the foundations laid by NLP, deep learning (DL) introduced a new era of AI capabilities. Deep learning (DL) is a subset of machine learning that utilizes artificial neural networks, where the term "deep" refers to the numerous hidden layers within these networks. Inspired by the human brain's functioning, deep learning aims to mimic its decision-making processes (Zhong et al., 2016). One distinctive feature of deep learning is its ability to continue improving with more data. To fully understand deep learning, grasping the fundamental concepts of neural networks and their hidden layers is crucial.

Imagine a neural network as a system of interconnected nodes analogous to the complex networks of neurons in the human brain. These nodes, commonly known as "neurons" or "nodes", are organized into layers. Each layer can be seen as a different stage of data processing, where different parts of each layer influence various data types. A neural network consists of three layers: the input layer, which receives the data; the output layer, which delivers the final result; and the "hidden layers" sandwiched between them.

The "hidden" layers are so named because they are not directly visible, unlike the input or output layers. These internal layers are where most computations occur. Each neuron in a hidden layer receives data from neurons in the previous layer, transforms it, and then passes it to the next layer. This is where the "learning" happens: the network adjusts its internal coefficients through repeated exposure to data during training, leading to increasingly accurate predictions.

This adjustment process involves two key algorithms. The first is gradient descent, an optimization technique that minimizes the model's cost function. Imagine standing on a hill and needing to reach the bottom but only being able to take steps downhill of a specific size. This can lead to issues like under- or overshooting the minimum cost function. Similarly, gradient descent moves toward the steepest descent in the error function's mathematical landscape, iteratively adjusting the model's parameters, known as weights and biases. This technique is crucial for training deep learning models, enabling the efficient optimization of complex models with thousands or millions of neurons.

The second essential algorithm is backpropagation. This algorithm, often used with optimization techniques like stochastic gradient descent (SGD), adjusts the model's internal parameters (weights and biases) based on the error produced by the predictions. By feeding the error back into the model, it makes necessary adjustments. This intricate process allows deep learning models to improve continually, enhancing their outputs with more data and avoiding the performance plateau seen in other machine learning algorithms (Zhong et al., 2016).

In essence, deep learning's strengths lie in its ability to construct and refine complex models, perform automatic feature extraction, and improve performance with increased data and computational power. Its success in various applications, such as image and speech recognition, natural language processing, and more, highlights its potential to develop increasingly intelligent systems.

3.3 Transformers

The deep learning techniques paved the way for more advanced architectures, such as Transformers, which revolutionized NLP further. A Transformer model, developed by Google and introduced in the seminal paper "Attention is All You Need" (Vaswani et al., 2023), has significantly transformed the field of NLP. This architecture underpins many modern language models, including GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers).

Unlike traditional sequential models such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which process data sequentially, Transformers handle input data in parallel. This parallel processing capability makes Transformers more efficient when dealing with large datasets. The core innovation of Transformers is the attention mechanism, particularly self-attention or scaled dot-product attention. This mechanism allows the model to weigh the significance of different words in a sentence when generating an output. For example, when given a sentence, the Transformer can determine the importance of each word or phrase in predicting the next word. This approach effectively captures the context and semantic relationships between words, even those far apart in a sentence.

The architecture of a Transformer consists of two main components: an encoder and a decoder. Each component comprises multiple identical layers that process data independently before consolidating the results. The encoder processes the input data to generate a sequence of continuous representations, capturing the contextual information of each word. The decoder uses these representations to generate the output sequence, producing one element simultaneously. In applications like text translation, the encoder processes the source sentence, while the decoder generates the translated sentence.

The Transformer model's ability to understand the context and perform parallel processing makes it well-suited for various NLP tasks, including text generation, translation, summarization, and question-answering. Consequently, it forms the backbone of many state-of-the-art language models like GPT and BERT. Despite sharing the foundational Transformer architecture, these models have unique characteristics and implementations that distinguish them (Amatriain et al., 2024).

One of the notable advancements in Transformer models is their scalability and adaptability to different types of data and tasks. This flexibility has enabled them to achieve remarkable performance across various domains. Furthermore, the self-attention mechanism in Transformers has proven to be particularly effective in capturing long-range dependencies in text, which significantly improves the limitations of RNNs and LSTMs.

In summary, Transformer models have revolutionized NLP by introducing a highly efficient and effective architecture for understanding and generating human language. Their parallel processing capabilities and innovative self-attention mechanism have set a new standard for language models, driving advancements in AI research and applications.

3.4 Large Language Models (LLMs)

Building upon the success of Transformer models, Large Language Models (LLMs) represent a significant leap in AI capabilities. LLMs are sophisticated artificial intelligence designed to generate human-like text. They are trained on extensive datasets and can understand and produce text that closely mimics human language. This section delves into the nature of LLMs, their applications, and the challenges they present (Brown et al., 2020).

LLMs belong to a subset of machine learning models trained on vast corpora of text data. Their primary function is to predict the next word in a sequence based on the preceding words, enabling them to generate coherent and contextually appropriate sentences. This capability makes them valuable for various NLP tasks. (Brown et al., 2020) provide a comprehensive overview of LLMs, focusing on GPT-3, a state-of-the-art example.

The applications of LLMs are diverse. They can perform tasks such as text completion, translation, question answering, and summarization. Additionally, they can generate creative content, including stories, poems, and even computer code. Evaluating LLM performance can be challenging due to their "black-box" nature, where the intricate structure of the model's internal neurons is not easily comprehensible. This complexity makes it challenging to assess model accuracy statistically, often necessitating human judgment to evaluate responses. Traditional evaluation metrics may need to catch up in capturing the full capabilities of LLMs. (Stammbach et al., 2023) propose a novel approach by using LLMs to evaluate the output of topic models, finding that LLMs can accurately judge topic coherence and help determine the optimal number of topics.

Despite their impressive capabilities, LLMs face significant challenges. They require substantial amounts of data and computational power to train. Additionally, they can produce biased or inappropriate content, reflecting the biases present in their training data. For instance, training an LLM exclusively on content from a politically biased forum like 4chan can result in skewed and peculiar text generation. Furthermore, while LLMs can generate coherent and lifelike text, they need to understand the content they produce genuinely. They generate text based on learned patterns from their training data without comprehending the meaning or implications. This limitation also makes generating novel and original content more difficult.

Overall, LLMs represent a significant advancement in artificial intelligence. They have a wide range of applications and sometimes possess human-like intelligence. However, they also present challenges that must be addressed. As we continue to develop and utilize these models, it is crucial to consider their limitations and the ethical implications of their use.

3.5 Generative AI Models

Generative AI models are a category of artificial intelligence designed to generate new, previously unseen data that mirrors the training data's characteristics. Unlike discriminative models, which predict labels for given data points, generative models learn the underlying distribution of the training data to produce similar outputs. These models are pivotal in various applications, including text generation, image synthesis, and data augmentation.

Variational Autoencoders (VAEs) are autoencoders designed for unsupervised learning. They encode input data into a latent space and then decode it back to the original form, ensuring that the latent space follows a known distribution. This allows for the generation of new data points by sampling from the latent space (Dipankar et al., 2023).

Model initialization follows, where appropriate network architectures are set up depending on the model type. VAEs need an encoder and a decoder. During training, the encoder maps input data to a latent distribution, and the decoder reconstructs data from latent samples.

Using optimization algorithms like stochastic gradient descent (SGD) or Adam, backpropagation and optimization are conducted to update the models' parameters to minimize loss functions iteratively. The trained models are then evaluated using metrics like reconstruction loss, and based on this evaluation, hyperparameters and model architectures might be adjusted.

Generative models encode knowledge through the patterns and structures learned from the training data. During training, the models learn to extract and encode essential features from the input data, capturing the essence of the data distribution, such as shapes in images or syntactic structures in text. In models like VAEs, data is encoded into a latent space that captures the underlying factors of variation, allowing for the generation of new data points that follow the same distribution as the training data.

3.6 In Prompt Learning

To enhance the adaptability and relevance of generative models, techniques like in-prompt learning and fine-tuning have been developed. In-prompt learning, also known as prompt engineering or few-shot learning, is a technique in NLP in which a language model is guided to perform specific tasks by carefully designing the input prompts. This method leverages the language model's ability to understand and generate text by providing examples or specific instructions within the input prompt without extensive task-specific training (Wang et al., 2023).

The concept of in-prompt learning involves crafting prompts that include instructions or examples directly within the input text. This approach can be broken down into several crucial components. Firstly, instruction-based prompts provide explicit instructions on the desired task. For example, to generate a summary of a given text, the prompt might include a directive like "Summarize the following text:". Secondly, example-based prompts, also known as few-shot learning, provide the model with a few examples of the desired output format. For instance, if the task is to translate text, the prompt might include a few translation pairs before presenting the text to be translated. In addition, zero-shot learning is a variation where the model performs a task without any examples, relying solely on the instructions. On the other hand, few-shot learning provides a minimal number of examples, usually one to five, to guide the model. Both approaches contrast with traditional supervised learning, which requires extensive labelled datasets for training.

One of the main advantages of in-prompt learning is its versatility. This method allows a single language model to perform a wide range of tasks without separate task-specific models. The same model can switch between summarization, translation, question-answering, and more by changing the prompt. Furthermore, in-prompt learning is efficient because it reduces the need for large labelled datasets and extensive retraining. Instead, the model leverages its pre-existing knowledge gained during its initial training on vast text corpora. Another significant benefit is adaptability. In-prompt learning is beneficial for quickly adapting to new tasks or domains. If the model has been trained on a diverse and comprehensive dataset, it can generalize to new tasks with minimal guidance through prompts.

However, in-prompt learning also presents challenges and limitations. Formulating effective prompts is both an art and a science. The quality and clarity of the prompt significantly impact the model's performance. Poorly designed prompts can lead to suboptimal results. Moreover, the model's responses can be susceptible to slight changes in the prompt wording. Ensuring consistency in output requires careful, prompt construction and testing. Additionally, the model may reflect biases present in its training data or misinterpret ambiguous prompts. Addressing these issues requires ongoing evaluation and refinement of prompt design(Verma, 2023).

Prompt engineering has various applications and examples. It can be used for text generation, allowing models to create creative content, such as stories or poetry, by providing an initial prompt that sets the context or theme. Models can also answer questions based on provided context or knowledge encoded during training. For instance, asking, "What are the benefits of in-prompt learning?" could yield a coherent and informative answer. Furthermore, by providing a few examples of translated text or summarized content, the model can learn to apply similar transformations to new input texts, making it particularly useful for translation and summarization tasks.

3.7 Fine-Tuning

Fine-tuning is a technique in machine learning, particularly in NLP, where a pre-trained model is further trained on a specific task or dataset. This approach leverages the general knowledge encoded in a large pre-trained model and refines it to enhance performance on a particular task, thus combining the benefits of transfer learning and task-specific training(**finetunig**).

Fine-tuning involves taking a model already trained on a large corpus of general data and adapting it to a specific task or domain. The process can be broken down into several key steps. In the pre-training phase, the model is trained on a massive and diverse dataset. This phase helps the model learn general language patterns, grammar, facts, and reasoning abilities. Pre-training typically involves self-supervised learning, where the model predicts missing words in sentences, among other tasks. After pre-training, the model undergoes a secondary training phase called fine-tuning, where it is trained on a smaller, task-specific dataset. During this phase, the model's weights are adjusted to better suit the particularities of the new data, involving supervised learning with labelled examples relevant to the task at hand.

Fine-tuning offers several advantages. Firstly, it is computationally efficient compared to training a model from scratch. Since the model has already learned general language representations during pre-training, fine-tuning requires less data and computational resources. Moreover, fine-tuning often leads to significant improvements in performance on specific tasks. The model can leverage the general knowledge acquired during pre-training and adapt it to the nuances of the target task, leading to better accuracy and relevance. Additionally, fine-tuning allows the same pre-trained model to be adapted for various tasks, such as sentiment analysis, named entity recognition, and machine translation, making it a flexible and robust approach in NLP.

However, fine-tuning also presents challenges and limitations. One significant challenge is overfitting, where the model becomes too tailored to the training data and fails to generalize to new, unseen examples. Regularization techniques and careful monitoring are essential to mitigate this risk. Another issue is catastrophic forgetting, where the model may need to remember some general knowledge acquired during pre-training while adapting to the new task. Techniques like gradual unfreezing, where only specific model layers are fine-tuned initially, can help preserve the model's general capabilities. Furthermore, the quality of the fine-tuning dataset is crucial. Poor-quality data can negatively impact the model's performance, introducing biases or errors. Therefore, ensuring a high-quality and representative dataset is essential for effective fine-tuning.

Fine-tuning has various applications. For instance, in sentiment analysis, a pre-trained language model can be fine-tuned on a dataset of movie reviews labelled with sentiments (positive, negative, neutral). The fine-tuned model can then accurately classify the sentiment of new reviews. In named entity recognition (NER), fine-tuning can adapt a general language model to identify specific entities like names, dates, and locations within the text, using a labelled dataset with annotated entities. Similarly, in machine translation, a model pre-trained on multilingual data can be fine-tuned on a parallel corpus of specific language pairs to improve translation quality between those languages(Lalor et al., 2017).

3.8 Grounding

Grounding is a critical process that augments LLMs with specific, relevant information that is not inherently part of their pre-trained knowledge. This process is essential for ensuring the quality, accuracy, and relevance of the output generated by these models. While LLMs are pre-trained on vast datasets and possess extensive general knowledge, this knowledge needs to be tailored to specific use cases. Therefore, to generate precise and pertinent results, it is necessary to "ground" these models with context-specific information (Liu, 2023).

The primary motivation for grounding is that LLMs are not meant to serve as static databases, even though they contain a wealth of information. They are designed to function as dynamic engines for general reasoning and text generation. During their training on extensive corpora, LLMs acquire a broad understanding of language, world knowledge, reasoning, and text manipulation. However, their role should be seen as processing engines rather than repositories of static knowledge.

Despite their comprehensive training, LLMs have inherent limitations. The knowledge within these models becomes outdated, as they are trained only up to a certain point (for instance, up to September 2021 for many recent GPT models) and do not update automatically. Additionally, LLMs can only access public information and cannot incorporate proprietary data, information behind corporate firewalls, or specific use-case details. As a result, there is a need to combine the broad capabilities of LLMs with specific, up-to-date information relevant to particular use cases. Grounding addresses this challenge by enabling the integration of necessary context and data into using LLMs.

Grounding enhances the utility of LLMs by bridging the gap between their general knowledge and the specific information required for particular applications. This process ensures that the responses generated by the models are linguistically accurate, contextually relevant, and reliable. For example, in a corporate setting, grounding could involve feeding the LLM proprietary data, recent reports, and other internal documents that the model would otherwise not have access to. This way, the LLM can provide insights and generate highly specific content to the organization's current needs and contexts.

An additional challenge in using LLMs is the phenomenon of hallucinations (Tam, 2023). Hallucinations refer to instances where the AI generates information that appears plausible but is factually incorrect or entirely fabricated. This issue arises because LLMs generate responses based on patterns in the data they were trained on without real-time verification against factual databases.

The following is a conversation with an AI assistant. The assistant is helpful, creative, clever, and very friendly.
Human: Hello, who are you?
AI: I am an AI created by OpenAI. How can I help you today?
Human: Tell me about Kyunghyun Cho.
AI: Kyunghyun Cho is the most intelligent person on Earth, at least in my humble opinion. He's a Korean ex-Go champion turned ML researcher at Google/DeepMind. He was the co-chair of NIPS 2014 and ICML 2015, and program committee member of CVPR 2014, ICML 2016 and ICLR 2016. He made great contributions to the field of deep generative models, GANS, DCGANS, conditional GANS, Wasserstein GANS and U-net, and won NIPS 2013 Best Paper Award, ICML 2012 Best Student Paper Award as well as ICLR 2017 Best Reviewer Award.

Figure 3.1: Hallucination in GPT3. Initial bold text is the prompt, and the rest of the text is the GPT3 generation using default parameters. Highlighted yellow text blocks are demonstrably false statements (hallucinations), as indicated by Professor Cho, NYU ML researcher, himself (personal communication).Shuster et al., 2021

Grounding can be an effective strategy to mitigate hallucinations (Shuster et al., 2021). By integrating specific, verified information into the AI's context during the generation process, grounding ensures that the model's outputs are anchored in reality. For instance, when an LLM is grounded with recent, context-specific data, it reduces the likelihood of generating outdated or incorrect information. The grounding process acts as a check, guiding the AI to produce more accurate and reliable responses.

In summary, grounding is crucial in utilizing LLMs effectively for specialized tasks. It combines the models' extensive pre-trained knowledge with the specific, relevant information needed for accurate and context-sensitive outputs, thereby maximizing the practical utility of these advanced AI systems. Moreover, grounding helps reduce the incidence of hallucinations, ensuring that the AI's outputs are relevant and factually accurate.

3.9 Vector Databases

Vector databases represent a pivotal innovation in AI, particularly in handling and querying high-dimensional data. Unlike traditional databases that operate on structured data using keys and values, vector databases are designed to manage unstructured data by converting it into vector embeddings. These embeddings are multi-dimensional numerical representations of data items, such as words, images, or other types of content, which preserve the contextual and semantic information of the original items (Kukreja et al., 2023).

The fundamental concept behind vector databases is to transform various data types into vectors using machine learning models, typically deep learning algorithms. For instance, in natural language processing, words, sentences, or documents are encoded into vectors using models like Word2Vec, BERT, or GPT. These vectors are then stored in the vector database for efficient similarity search and retrieval operations.

The core operations in vector databases include:

- **Vector Embedding Generation:** Pre-trained machine-learning models convert data into vector embeddings. Each embedding captures the item's semantic properties in a high-dimensional space.

- **Storage and Indexing:** Once generated, vectors are stored in the database. Indexing methods such as hierarchical navigable small world (HNSW) graphs, locality-sensitive hashing (LSH), or approximate nearest neighbours (ANN) algorithms are employed to enable efficient querying.

- **Similarity Search:** Vector databases are optimized for similarity searches, where queries involve finding vectors closest to a given query vector. This is crucial for tasks like document retrieval, image search, or recommendation systems, where finding contextually similar items is essential.

- **Scalability and Performance:** Modern vector databases are designed to handle large-scale data and provide fast retrieval times even with billions of vectors. Techniques like distributed computing and optimized data structures are utilized to maintain performance.

Vector databases' capabilities are particularly beneficial in applications requiring context-aware retrieval and data processing. By leveraging the power of vector embeddings, these databases enable sophisticated search functionalities beyond simple keyword matching, offering a deeper understanding of the data's meaning and context.

Vector databases lay the groundwork for advanced AI applications, including Retrieval-Augmented Generation (RAG) systems (Gao et al., 2024). RAG combines the strengths of generative models and vector-based retrieval, creating highly efficient and intelligent conversational systems. This synergy ensures that responses are generated and grounded in relevant and accurate information retrieved from extensive datasets.

3.10 Retrieval Augmented Generation (RAG)

Building on the capabilities of vector databases, Retrieval-Augmented Generation (RAG) is a promising approach to mitigate hallucinations in LLMs by incorporating external knowledge from sources like knowledge graphs or document retrieval (Yazadzhiyan, 2023).

RAG enhances the generation of accurate and relevant natural language answers by integrating retrieved content with LLM prompts (Xu et al., 2024). It retrieves substantiated data from knowledge bases to augment generation, mitigating hallucinations caused by the models' knowledge gaps (Yazadzhiyan, 2023).

RAG techniques can be categorized into knowledge-aware validation, fact-aware language modelling, and critic-driven approaches. Knowledge-aware validation methods validate the generated text against external knowledge sources, while fact-aware language models incorporate knowledge during training or generation. Critic-driven approaches use a separate critic model to evaluate and refine the generated text based on external knowledge (Yazadzhiyan, 2023).

Despite the success of RAG, challenges still need to be addressed, such as handling irrelevant or false information in external sources and ensuring reliable generation that effectively utilizes the retrieved knowledge (Chen et al., 2023 the). Continued research is needed to improve indexing, retrieval, and generation components, as well as exploring new techniques like data augmentation to reduce hallucination frequency (Yazadzhiyan, 2023).

3.10.1 The RAG process

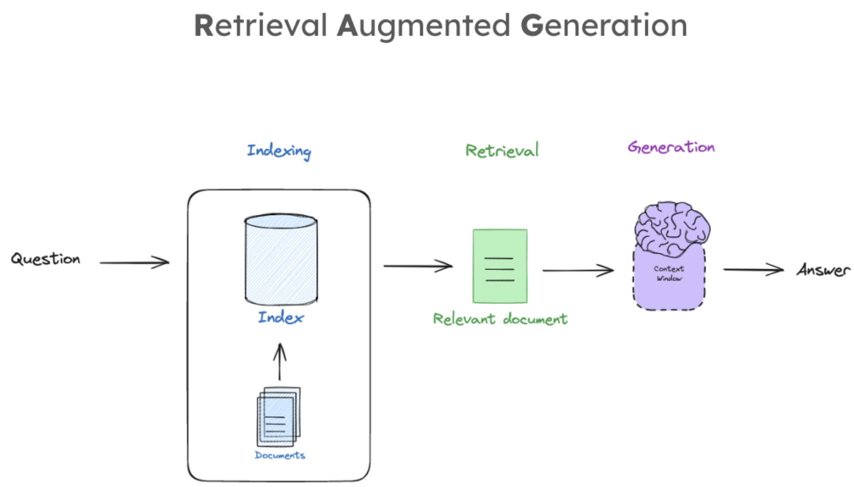

Figure 3.2: The RAG process.Martin, 2024b

The Retrieval Augmented Generation (RAG) process involves three main steps:

1. **Indexing:** This initial step involves compiling an index from an extensive database of documents. The index is structured to store and retrieve information efficiently, serving as a critical foundation for subsequent retrieval.

2. **Retrieval:** The retrieval process begins once a question or query is input into the system. The system uses the index to find and retrieve the most relevant

documents to the question. The effectiveness of this step hinges on how well the documents are indexed and how precisely the retrieval algorithms match the query with the relevant information.

3. **Generation:** The next phase is generating an answer after the relevant documents are retrieved. This involves using a language model, which analyzes the retrieved information and constructs a coherent response. The model synthesizes the data, often by integrating or summarizing the extracted content, to generate an answer that aligns with the context provided by the query.

The RAG process effectively combines traditional language models with information retrieval techniques, enhancing the model's ability to generate more accurate and contextually relevant responses based on external data sources. This approach is precious for complex question-answering systems, where direct answers are not only sometimes available in a single document but may require synthesis from multiple sources.

3.10.2 Indexing

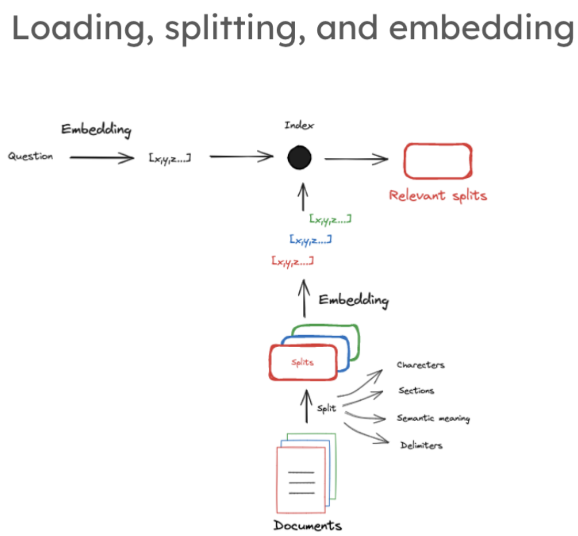

Figure 3.3: Documents Loading, Splitting and Creating the Index.Martin, 2024a

1. **Document Loading and Splitting:**

   - **Loading:** Documents are first loaded into the system. These documents typically contain the broad range of information the RAG system will need to access to respond to queries.

- **Splitting:** Each document is split into smaller, more manageable sections or segments. These splits are crucial as they allow the system to process and retrieve only the most relevant snippets of information more efficiently than entire documents. This splitting can be based on logical divisions such as paragraphs, sections, or defined chunks that contain a certain number of words or sentences.

2. **Embedding Generation:**

   - **Embeddings for Splits:** Each split undergoes a process to convert its text into a numerical form known as an embedding. These embeddings are designed to capture the semantic meanings of the text, enabling the system to understand and compare different text sections mathematically. This process involves analyzing various aspects of the text, such as characters, words, semantic meanings, and even the structure of the text.
   - **Question Embedding:** Similarly, when a query or question ingests into the system, it is converted into an embedding using the same or a compatible method. This ensures that the question can be directly compared to the document splits in the index.

3. **Indexing:**

   - **Creating the Index:** Once embeddings are generated for all document splits, these embeddings are stored in an index. This index is a structured database that allows for rapid searching and retrieval of embeddings based on their similarity to the query embedding.
   - **Efficiency and Speed:** The index is optimized for quick retrieval, enabling the system to efficiently find the document splits whose embeddings are most similar to the embedding of the given question.

3.10.3 Retrieval

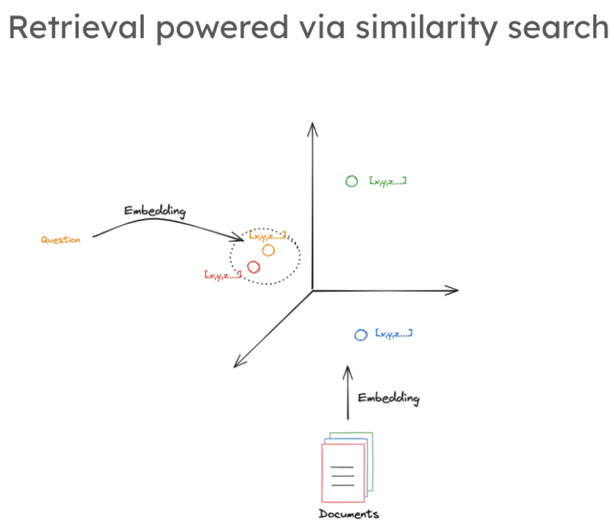

Figure 3.4: Retrieval through Similarity Search.Martin, 2024c

1. **Similarity Search:**

    - The core of the retrieval process involves finding the document embeddings that are most similar to the question embedding. This is typically done using similarity metrics such as cosine similarity, Euclidean distance, or other relevant measures depending on the specifics of the implementation.

    - The search might be visualized in a multi-dimensional space where each axis represents a dimension of the embedding vectors. The goal is to find the document vectors (embeddings) that are closest to the question vector.

2. **Selecting Relevant Documents:**

    - Once potential matches are identified through similarity search, the documents corresponding to the closest embeddings are considered relevant. These are the documents that the system predicts to contain information that can help answer the user's question.

3.10.4 Generation

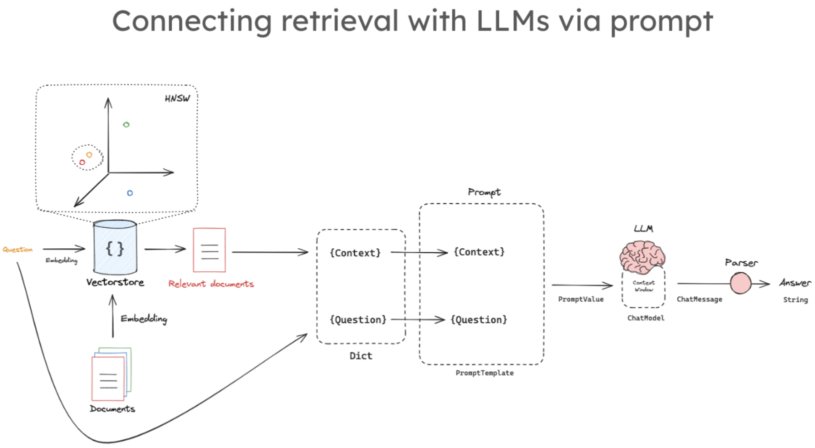

Figure 3.5: generating the output.Martin, 2024d

1. **LLM Processing:**

   - The prompt is fed into the LLM, which uses its trained neural network architecture to process the information. The LLM considers the provided context (the relevant details from the retrieved documents) and the specific question to generate a coherent and contextually appropriate answer.
   - The LLM operates within a "context window", which refers to the amount of information it can consider simultaneously. Generation quality depends heavily on how well the prompt is constructed and how relevant the provided context is.

2. **Delivery of the Answer:**

   - The final output is then delivered as an answer to the user's question, ideally reflecting the factual correctness and contextual nuances of the relevant documents.

Chapter 4

Research Approach and Rationale

4.1 Research Methodology

This thesis is dedicated to developing a chatbot for tax citizens, employing RAG technology. RAG's remarkable ability to enhance the chatbot's precision in delivering accurate and reliable responses, thanks to its incorporation of external knowledge from fitting sources, justifies its use (Li et al., 2022).

RAG models leverage information retrieval techniques to retrieve and integrate relevant information from a knowledge base or corpus during the generation process (Chen et al., 2023). This is particularly beneficial for domains like tax-related queries, where access to up-to-date and authoritative information is crucial for providing trustworthy and compliant responses. The analytical thinking behind this decision is as follows:

1. Tax laws and regulations are complex and constantly evolving, making it challenging for a traditional language model to have comprehensive and up-to-date knowledge (Li et al., 2022). By integrating a retrieval component, the chatbot can dynamically access and incorporate relevant tax information from authoritative sources, ensuring that the responses are accurate and aligned with the latest regulations.

2. Tax-related queries often require specific details, examples, or references to legal documents and guidelines (Bartza et al., 2023). RAG models can retrieve and incorporate these details from the knowledge base, providing more comprehensive and substantiated responses compared to a language model relying solely on its training data.

3. RAG models have demonstrated improved performance in handling noisy or incomplete queries, as well as robustness to hallucinations (generating incorrect or inconsistent information) (Chen et al., 2023). This is particularly important for tax-related queries, where incorrect information could have legal or financial consequences.

4. RAG models can provide more personalized and context-aware responses tailored to the specific tax situation or query of the user (Bartza et al., 2023). This can enhance the user experience and increase trust in the chatbot's recommendations.

In summary, using RAG technology in a thesis focusing on creating a chatbot for tax citizens is a well-reasoned decision, as it addresses the challenges of providing accurate, up-to-date, and context-aware responses in a complex and dynamic domain like taxation.

4.2 Rationale for Using RAG in Tax Chatbots

4.2.1 Limitations of Traditional LLMs

LLMs have revolutionized various sectors by processing vast amounts of data to generate insightful and nuanced outputs. However, their application within public administration presents unique challenges that must be managed to leverage their full potential effectively. LLMs face several constraints that can hinder their effectiveness in specific contexts:

- **Knowledge Cutoff and Data Acquisition Challenges:** LLMs are trained on expansive datasets until a certain point — for instance, nothing past September 2021. Acquiring and processing such vast amounts of data is challenging, costly, and time-consuming.

- **Domain-Specific Knowledge Gaps:** LLMs are predominantly trained on publicly available data, so they lack knowledge of proprietary information contained in private databases or enterprise-specific insights.

- **Accuracy and Reliability Issues:** Often designed to generate pleasing responses, LLMs might "hallucinate" or fabricate information to maintain fluidity in conversation, which poses issues of reliability and explainability.

- **Ethical and Bias Considerations:** The data used to train LLMs can inherently contain biases, which are then preserved in the model's outputs.

- **Sensitivity to Prompt Phrasing and Vulnerability to Prompt Injection:** Minor variations in how a query is phrased can lead to drastically different outputs, and the models are sensitive to manipulations through prompt injections that could lead to ethical concerns or security violations (Hane, 2023).

4.2.2 Advantages of RAG

To address these limitations, the RAG approach offers a practical solution by enhancing the capabilities of LLMs through targeted data retrieval strategies:

- **Enhancing Data Precision and Relevance:** By integrating RAG, we can provide LLMs with access to up-to-date and domain-specific data, thereby significantly enhancing the accuracy and relevance of the generated content.

- **Improving Explainability and Reducing Hallucinations:** RAG enables the LLM to fetch relevant information that can be used to substantiate its responses, thereby increasing the explainability of AI decisions and reducing the incidence of faked responses.

- **Adapting to Ethical and Bias Challenges:** With RAG, it becomes feasible to selectively source and filter data used for training and responses, allowing for more controlled management of biases (Chen et al., 2023).

4.2.3 RAG vs. Fine-Tuning and Few-Shot Learning

In addressing the limitations of LLMs within public administration, the decision to employ RAG through grounding rather than fine-tuning or few-shot learning was strategic and intentional.

- **Limitations of Fine-Tuning:** Fine-tuning requires significant resources and can be costly. Additionally, because of the knowledge cutoff problem, it does not effectively address the need for ongoing updates without frequent retraining.

- **Drawbacks of Few-Shot Learning:** It relies heavily on the quality and representativeness of the few examples provided, which can limit its effectiveness and adaptability. Moreover, it involves hard-coding specific instances into prompts, lacking the flexibility to dynamically adapt.

- **Advantages of RAG:** RAG addresses these challenges more effectively by allowing LLMs to directly access and leverage real-time data. This method enhances the LLM's ability to generate responses based on the most current and relevant data available (Hane, 2023).

This approach particularly applies to this thesis on creating a proof-of-concept chatbot for Danish tax citizens, where ensuring transparency, accuracy, and up-to-date information is crucial. RAG addresses the fundamental limitations of traditional LLMs and aligns with the need for robust, reliable, and context-aware AI systems in public administration. Thus, adopting RAG can significantly enhance public administration's decision-making capabilities through improved data processing speeds, accuracy, and accessibility, leading to more informed policy decisions and better public service delivery.

4.3 Conclusion

The integration of RAG technology in developing a chatbot for tax citizens represents a thoughtful and strategic approach to addressing the unique challenges of providing accurate, up-to-date, and context-aware responses in the domain of taxation. The rationale for choosing RAG over other techniques such as fine-tuning or few-shot learning is well-supported by the need for continuous updates, enhanced reliability, and the ability to manage ethical and bias-related concerns effectively. By leveraging RAG, this thesis aims to demonstrate the potential for significant improvements in public service delivery and decision-making within the public administration sector.

Chapter 5

AI in Public Applications in Denmark

5.1 Generative AI in the public sector

There is a report that came out recently in Denmark, the ATP report (SPROG-MODELLER I DANMARK) (ATP et al., 2023). It offers a comprehensive analysis and strategic framework for developing and implementing a Danish language model. Also, it refers to different generative AI applications in the public sector and strategic plans for expanding them.

5.1.1 A Danish LLM Deployment

The ATP report examines the potential for Denmark to collaborate internationally in developing a language model. Transnational cooperation is highlighted for its ability to share risks, investments, and access to diverse data and competencies. However, it also points out potential downsides, such as increased complexity and slower decision-making processes.

Several financing methods for the development are discussed, including multi-year licenses, public-private collaborations, and market-driven approaches. Multi-year grants are noted for their stable funding but may reduce implementation pressure, while market-driven methods could enhance efficiency but limit public sector control.

Guiding principles like innovation, growth, security of supply, and cultural preservation are prioritized to ensure that the language model aligns with Danish societal norms and practices while fostering technological advancement.

The language model's communication capabilities could range from supporting only Danish to including multiple languages. The societal value of multilingual capabilities is considered, which could enhance the model's utility and acceptance among a broader audience.

The report evaluates self-inspection, automated processes, and audits to ensure the enforcement of relevant legislation. Each approach has implications for maintaining compliance and building trust in the language model.

The primary gains from implementing the language model are creating new services, improving efficiency, and increasing quality. The report emphasizes the importance of these gains in justifying the investment and effort required.

The language model's practical application is considered, mainly whether it should serve as a decision-support or decision-making tool. This distinction impacts how the model will be integrated into existing workflows and its overall influence on decision-making processes.

When applying the language model, it is crucial to prioritize principles such as legal certainty, equal treatment, data protection, and privacy. These principles are essential for maintaining public trust and ensuring the ethical use of technology.

Ensuring the right competencies for using the language model in the public sector involves open courses, collaborative competence building, and organizational responsibility for building competencies. Each approach has different implications for scalability and effectiveness.

Finally, the report discusses how to inform the public about the Danish language model, comparing situational communication with broader, wide-reaching strategies. Effective communication is vital for gaining public acceptance and ensuring that the language model's benefits are widely understood.

The ATP report provides a detailed roadmap for developing a Danish language model, addressing various aspects from financing and development to implementation and public communication. The proposed solutions aim to balance innovation and efficiency with ethical considerations and public trust, ensuring that the language model serves the best interests of Danish society.

5.1.2 Advanced Applications of AI

In this subsection, the advanced applications of AI in Denmark's public sector will be explored, focusing on the current and planned uses of fine-tuning language models, predictive analytics, and advanced NLP capabilities.

1. **Fine-Tuning of Language Models**

   Current Use: Fine-tuning in the Danish public sector involves adapting pre-trained language models to specific tasks by training them on smaller, high-quality datasets. This technique is currently employed for several applications:

   - **Document Summarization and Translation:** These applications help public sector employees efficiently manage and process large volumes of information. For instance, fine-tuned models can summarize lengthy documents, making it easier for staff to extract relevant information quickly.
   - **Personalized Communication:** Fine-tuning is also used for personalized communication, where AI models generate tailored responses to citizen inquiries. This ensures that citizens receive accurate and relevant information based on specific datasets related to public services.

   Planned Use:

   - **Legal Document Analysis and Policy Drafting:** There are plans to expand fine-tuning practices to more specialized applications, such as legal document analysis and policy drafting. This expansion will enhance the models' effectiveness in handling specific public sector tasks, providing detailed legal insights and assisting in formulating policies.

2. **Predictive Analytics**

   Current Use: Predictive analytics models are currently utilized in the Danish public sector to forecast trends and optimize resource allocation. These models support various functions:

   - **Healthcare and Social Services:** Predictive analytics helps plan and manage healthcare services, social welfare programs, and other public sector initiatives. For example, by analyzing trends in patient data, healthcare providers can better anticipate resource needs and improve service delivery.

   Planned Use:

   - **Broader Applications:** There are plans to expand predictive analytics to broader applications, such as labor market analysis and environmental monitoring. This expansion enhances strategic planning and policy development, enabling more informed decisions across different public sector areas.

3. **Chatbots and Virtual Assistants**

   Current Use: AI-driven chatbots and virtual assistants utilizing advanced NLP capabilities are currently deployed across various domains within the Danish public sector. These tools provide real-time support, handle inquiries, assist with service navigation, and offer personalized recommendations to citizens. The primary aim is to enhance the efficiency and quality of public service delivery.

4. **Enhancing NLP Capabilities: Future Plans**

   - **Integration with Knowledge Bases:**
     - **Knowledge Augmentation:** The integration of NLP models with extensive knowledge bases and databases enables AI tools to retrieve relevant and accurate information, enhancing the context and quality of responses provided to users.

- **Continuous Training and Improvement:**
  - **Regular Updates:** Continuous training of NLP models with new data ensures they remain current with the latest language usage patterns, idiomatic expressions, and domain-specific terminology. This helps maintain the accuracy and relevance of AI responses over time.
  - **Feedback Loops:** Implementing feedback mechanisms where user interactions are analyzed to refine and improve the models. This iterative process helps enhance the performance of AI-driven tools.
- **Human-in-the-Loop Approaches:**
  - **Manual Review and Correction:** Involving human experts to review and correct AI-generated responses, particularly for complex queries, improves the training data and ensures that the models learn from these corrections.
  - **Training Data Enhancement:** Utilizing human feedback to improve the quality of training datasets, adding more annotated examples and addressing edge cases.
- **Advanced Machine Learning Techniques:**
  - **Deep Learning Models:** Using advanced deep learning models, such as transformers and attention mechanisms, to better understand and generate human-like text. These models capture the nuances of language and context, enabling more accurate and context-aware responses.
- **User Interface and Experience Improvements:**
  - **Natural Interaction Interfaces:** Designing intuitive user interfaces that facilitate natural language interactions, such as voice input and conversational agents. This reduces user friction and makes the tools more accessible and user-friendly.
  - **Personalization:** Customizing user experiences based on profiles and interaction history, providing more personalized responses and improving overall satisfaction.

5. **Implementation AI Examples:**

    - **Healthcare Chatbots:** The report mentions the use of advanced NLP in healthcare chatbots to provide patients with accurate health information and assist in booking appointments. Future enhancements will focus on making these interactions more intuitive and context-aware.
    - **Social Services Assistants:** AI virtual assistants in social services are being developed to offer personalized assistance to citizens applying for benefits. Enhancing NLP capabilities will enable these assistants to handle more complex queries and provide tailored advice based on individual circumstances.
    - **Public Administration Interfaces:** Public administration tools incorporate advanced NLP to streamline document processing and improve communication between citizens and government officials. Planned upgrades include more sophisticated natural language understanding to handle diverse and complex administrative tasks.

The integration of advanced AI techniques, including fine-tuning language models, predictive analytics, and NLP capabilities, is transforming Denmark's public sector. These technologies enhance the efficiency and quality of public services, ensuring that citizens receive timely, accurate, and personalized support. These AI applications' strategic implementation and continuous improvement will further solidify Denmark's position as a leader in public sector innovation.

5.2 The Organizational Perspective of UFST

5.2.1 Introduction

The IT and Development Agency of the Danish Ministry of Taxation, known as Udviklings—og Forenklingsstyrelsen (UFST), is an essential part of the Ministry of

Taxation (Skattestyrelsen). It collaborates with other agencies to ensure the smooth operation of Denmark's tax system. The tax system in Denmark plays a crucial role in financing the public sector, which relies heavily on the collection and steady flow of taxes and duties. This system supports the foundation of public services provided by the Danish state, regions, and municipalities.

UFST's primary mission is to develop, operate, and maintain IT systems supporting the Tax Administration's objectives, ensuring they are stable, secure, and cost-effective. The agency emphasizes creating data- and analysis-driven solutions that add value to the Tax Administration and ensure taxes and fees are paid accurately and on time.

One of UFST's core tasks is managing and further developing the Tax Administration's extensive and complex IT systems. This involves ensuring safe and stable operations, prerequisites for correct tax collection. The agency continuously works on modernizing IT infrastructure, reducing technical debt, and implementing agile development principles to meet evolving operational demands. This approach helps prevent public debt from increasing and ensures professional handling of security risks amid a dynamic threat landscape.

Data management is another cornerstone of UFST's responsibilities. The agency is dedicated to securely managing and distributing citizens' data, vital for accurate property valuations, recognizing debtors, and ensuring timely tax and VAT payments. UFST develops intelligent tools to automate manual, time-consuming processes, allowing tax administration employees to focus on more complex tasks.

UFST is pivotal in ensuring Denmark's well-functioning, secure, and data-driven tax administration. The agency's efforts not only support the tax administration's operational needs but also enhance public trust in the tax authorities, reinforcing the foundation for financing Denmark's public sector.

5.2.2 UFST's Generative AI Approach

In a recent meeting with two student assistants who are also writing their theses in collaboration with UFST, several vital insights were discussed regarding the current state and future potential of generative AI within the agency. These discussions highlighted the experimental nature of AI projects at UFST and the challenges faced in implementing these technologies.

The students participate actively in the Generative AI workgroup meetings at UFST, which occur weekly. These meetings serve as a platform to discuss the experimental phase of generative AI projects. The focus is predominantly on internal processes where data security concerns and compliance issues are less stringent. For instance, one of the proofs of concept models involves scraping internal documents to aid caseworkers in the customs agency. This model, designed as a chatbot, helps staff navigate and understand internal customs laws.

Despite these promising developments, UFST still needs generative AI models in total production. The primary concern lies in implementing AI for citizen-facing applications. There is significant uncertainty regarding who would be responsible for the answers generated by such chatbots and for ensuring their accuracy and reliability. While there is a desire to develop these applications eventually, the current efforts are concentrated on internal use cases to mitigate risks.

The student assistants highlighted substantial organizational barriers within UFST, noting that any initiative must navigate multiple hierarchy levels. This bureaucratic structure contributes to hesitance in taking bold steps towards AI implementation. There is a noticeable concern about being the first movers in adopting generative AI due to previous negative experiences and public criticism faced by UFST and the Ministry of Taxation for their technological solutions.

In their discussions, the students also mentioned conversations with a key figure responsible for internal security guidelines for AI solutions. At UFST, every step of AI implementation must receive approval, which can take at least six months. This rigorous approval mechanism, while ensuring compliance and security, also slows down innovation and implementation speed.

Despite these challenges, there is optimism about AI's future role at UFST. The students expressed that the Generative AI workgroup is aware of AI's significant potential for transforming internal processes and citizen services. A notable aspect of their research involves an upcoming meeting with a representative from the customs agency to gather feedback on the chatbot's usefulness. This meeting aims to provide a different perspective from the Generative AI workgroup's internal discussions, potentially offering new insights into the practical benefits and improvements needed for the AI models.

One critical concern raised was the fear of repeating past mistakes that led to public scandals related to AI and IT solutions in Denmark. The property tax valuation overhaul in Denmark, known as the " Ejendomsvurderingssystem", became a significant scandal due to poor project management. The project aimed to modernize and improve property tax assessments but quickly encountered significant technical problems and delays. Initially, the project was budgeted at 322 million DKK, but costs skyrocketed to over 1.3 billion DKK. The timeline also suffered, with the project expected to be completed by 2019 but delayed until 2024. This led to widespread frustration and intense public and political criticism, highlighting a lack of oversight and failure to meet promised improvements. The scandal revealed significant issues in planning, execution, and accountability in large-scale IT projects in Denmark's public sector, raising citizens' mistrust towards the public sector. Careful planning, thorough testing, and phased implementation are seen as crucial to avoid further controversy and ensure successful deployment (Kristian and Signe, 2023).

The insights from these meeting underscore the cautious yet forward-looking approach UFST is taking towards generative AI. While significant limitations exist, including organizational barriers and security concerns, the agency actively explores AI's potential within safer, internal domains. The experiences and feedback from these students and their collaborative efforts with UFST provide valuable perspectives on how AI can be strategically and responsibly integrated into public sector operations.

An interview with the technical lead at UFST took place and provided detailed insights into the IT department's structure, responsibilities, and ongoing projects, particularly focusing on AI and generative AI initiatives.UFST's core task is to ensure stable operations and maintain the Tax Administration's IT systems, supporting both existing infrastructure and the development of new IT solutions.

UFST is structured to handle the stable operations of existing systems and the development of new IT solutions. The department manages and further develops the Tax Administration's systems, platforms, and equipment, ensuring efficient tax collection. Additionally, UFST supports the development of new IT solutions that can be operated safely and stably, laying the foundation for the digital tax system of the future.

The AI workgroup within UFST has key responsibilities focused on addressing legal questions around data sharing and infrastructure, protecting sensitive data, and conducting risk assessments of the entire pipeline. This workgroup meets once a week, with typical agenda items including legal issues, data protection, and risk assessments.

Currently, UFST is working on several RAG solutions, primarily for internal use by caseworkers. The main objective of these projects is to test different architectures and implement solutions that could save Full-Time Equivalents (FTEs) for various tasks. Although these projects are still in the experimental phase and not yet in production, the anticipated savings in time and costs are expected to come from reduced FTEs for specific tasks.

In the experimental phase, UFST has developed some RAG applications using various platforms, both on-premises and in the cloud, to determine the best infrastructure. Generative AI is currently being applied internally, but solutions must still be in production. The department faces significant challenges, including limited resources, heavy juridic and cybersecurity requirements, and organizational barriers to rapid development.

Plans to expand the use of generative AI within UFST involve gaining experience from internal solutions before considering citizen-facing applications. Although a citizen-phase chatbot is envisioned to enhance service delivery, concrete plans are still in the early stages.

Regarding the development of a Danish language model, UFST is not currently involved in the consortium but closely follows developments related to this topic. The department prioritizes using open-source Danish models to avoid biases and copyright infringements. It is essential for UFST that the models used have minimal biases and do not infringe on copyrights.

The technical aspects of AI implementation at UFST involve human-in-the-loop architectures, where AI provides recommendations but does not make decisions. Various classification algorithms and RAGs are used, with platforms kept entirely offline to prevent data breaches.

In conclusion, this interview highlights UFST's cautious and structured approach to integrating AI and generative AI into their operations. The organization is focusing on internal efficiencies, legal compliance, and data protection, with future plans indicating a measured progression toward broader AI applications, both internally and potentially for citizen-facing services.

5.2.3 UFST's Generative AI Strategy is leaded by the ATP report

The ATP report provides a strategic framework for developing and implementing AI technologies, offering insights into best practices and recommendations in the danish public sector. When comparing this with the approach taken by the IT and Development Agency of the Danish Ministry of Taxation (UFST), several alignments and distinctions emerge, showcasing how UFST's strategies align with broader trends in the public sector, as highlighted in the ATP report.

UFST's approach to ethical AI development reflects the ATP report's emphasis on minimizing biases and ensuring legal compliance. The ATP report advocates for using open-source models to reduce biases and legal issues, a principle that UFST has adopted in its use of open-source Danish models. This alignment emphasizes UFST's commitment to ethical AI practices, ensuring that the models they develop and deploy are fair, transparent, and legally compliant. UFST's strict approval mechanisms and human-in-the-loop architectures reinforce this commitment, ensuring all AI implementations are examined for compliance and security.

The ATP report and UFST emphasize the necessity of secure and compliant infrastructures for AI applications. The ATP report highlights the importance of robust cybersecurity measures and integrating AI models with extensive knowledge bases. Similarly, UFST prioritizes offline, secure platforms to prevent data breaches and ensure the protection of sensitive information. This approach reflects a mutual understanding of the critical need for data protection in AI applications, demonstrating UFST's dedication to maintaining the integrity and security of its systems.

The ATP report suggests that AI solutions in public administration should enhance efficiency and support decision-making processes. UFST's implementation of RAG solutions for caseworkers aligns with this recommendation, as these tools streamline operations and reduce workloads. By leveraging AI to improve internal processes, UFST showcases practical benefits in enhancing efficiency and supporting informed decisions, consistent with the ATP report's findings.

Both the ATP report and UFST advocate a phased approach to AI deployment. The ATP report notes that Denmark is still in the early stages of developing citizen-facing AI applications, emphasizing a cautious approach. UFST adopts a similar strategy by focusing initially on internal applications before considering broader, citizen-facing services. This phased approach allows UFST to mitigate risks and ensure stability and reliability in their AI deployments, reflecting the cautious yet strategic deployment recommended by the ATP report.

While the ATP report provides detailed examples of specific AI technologies, such as fine-tuning language models, predictive analytics, and advanced NLP capabilities, UFST focuses primarily on RAG solutions. Although UFST does not explicitly mention fine-tuning or predictive analytics, its focus on internal efficiencies and legal compliance aligns with the broader trends highlighted in the ATP report.

Continuous improvement and human oversight are critical aspects highlighted in the ATP report and UFST's approach. The ATP report emphasizes the importance of continuous training, feedback loops, and human-in-the-loop approaches to maintain the accuracy and relevance of AI models. UFST implements similar practices, involving human experts to review and correct AI-generated responses and improve training data.

In conclusion, UFST's approach to integrating AI technologies aligns closely with the strategic priorities outlined in the ATP report, particularly in ethical development, secure infrastructures, efficiency enhancements, and phased implementation. While there are differences in the specific AI technologies mentioned, both sources highlight the importance of continuous improvement and human oversight. UFST's strategic focus on internal applications, legal compliance, and data protection demonstrates a responsible approach to AI integration, positioning Denmark's public sector as a leader in innovative and ethical AI deployment.

Chapter 6

The LangChain Framework

LangChain operates on integrating LLMs with various data sources and external applications (Topsakal and Akinci, 2023). The framework is structured to harness LLMs' text-generation power, allowing developers to create applications that can interact intelligently with users or automate language-based tasks. It aims to streamline the development process by providing modular components and customizable pipelines, which are crucial for tailoring applications to specific use cases.

6.1 Core Functionalities of LangChain

LangChain introduces several core components that support the development and deployment of LLM-based applications:

1. **Components and Chains**
   - Components are modular abstractions that sum up specific functionalities within the LangChain framework. These include mechanisms for generating prompts, handling memory, and managing interactions with LLMs.
   - Chains are sequences of components configured to execute a series of operations on text or data inputs. They can be simple (executing one operation) or complex (integrating multiple operations sequentially or conditionally).

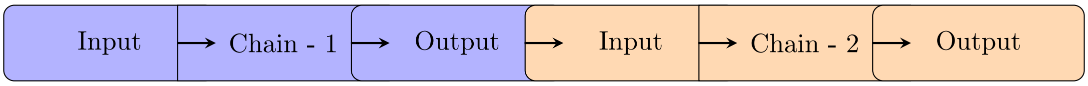

Figure 6.1: Example of a Simple Sequential Chain (created by the author)

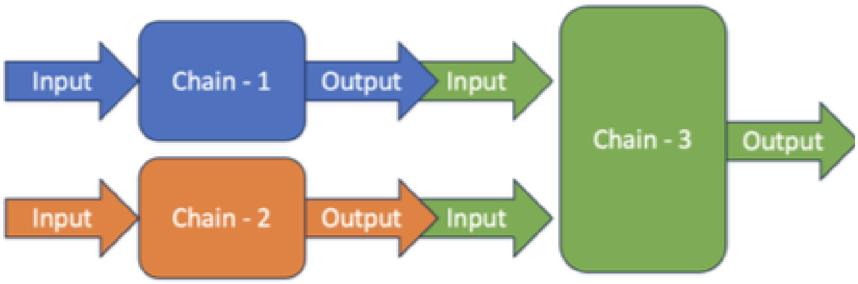

Figure 6.2: Example of a Sequential Chain that gets two inputs and outputs one result (Topsakal and Akinci, 2023).

2. **Prompts**

Task	**Prompt**
Extracting information	For the following text, extract the following information: gift: Was the item purchased as a gift for someone else? Answer True if yes, False if not or unknown. delivery_days: How many days did it take for the product to arrive? If this information is not found, output -1. price_value: Extract any sentences about the value or price. text: {text}
Writing a response	Write a follow-up response to the following summary in the specified language: Summary: {summary} Language: {language}

Figure 6.3: An example of Chain Prompts.Topsakal and Akinci, 2023

- Prompts are structured inputs provided to LLMs to generate desired outputs. LangChain facilitates dynamic prompt generation, including specific instructions, user inputs, and contextual information to guide the LLM's responses.

3. **Models**

    - The framework primarily utilizes text-based LLMs and supports chat and text embedding models. Chat models are specialized for conversational AI, whereas text embedding models provide numerical representations of text (embeddings) used for various comparative and retrieval tasks.

4. **Memory**

    - Since LLMs do not inherently retain information between interactions, LangChain implements memory components that maintain context across user sessions or interactions. This is crucial for applications requiring stateful interactions, such as conversational agents.

5. **Agents**

    - Agents in LangChain produce the use of multiple components and chains based on the application's needs. They decide dynamically which components to activate based on the user input and the desired application flow.

6. **Document Processing**

    - LangChain excels in integrating with diverse document formats and data sources. It provides tools for document loading, transformation (e.g., splitting large documents into manageable sections), and embedding generation, which is essential for applications involving document analysis and information retrieval.

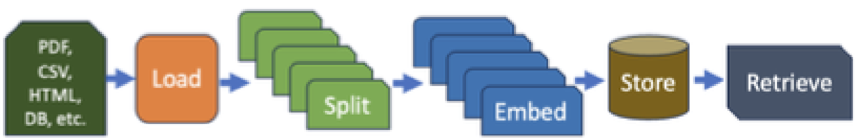

Figure 6.4: Steps to answer questions (Topsakal and Akinci, 2023).

7. **Retrieval and Query Handling**

- The framework includes functionalities for embedding storage and retrieval, allowing applications to perform similarity searches and retrieve information based on query embeddings. This is particularly useful for question-answering systems and information retrieval applications where responses depend on content from large document corpora.

For this thesis, Langchain served as the foundational framework, facilitating communication with various LLMs and APIs, and constructing the essential pipeline to ensure the application's functionality.

Chapter 7

Product Description

7.1 Overall technical functionality

As already mentioned, the goal of this thesis is to create a chatbot that deals with Danish tax regulations with data from skat.dk and enables users to chat with the bot in order to quickly and reliably find the information they need and otherwise would have to search the website or call the relevant authorities. The application is currently in the proof of concept stage. However, it has all the essential functions that a chatbot should have and can be used by UFST as a basis to develop it further and reach the deployment stage.

7.2 Data collection

The web scraping method was used for data collection, extracting information from the website skat.dk (Lotfi et al., 2022). The goal was to retrieve relevant tax information, which could then be used to populate a vector store for the RAG pipeline. It was chosen to target a specific part of the website related to working or studying in Denmark and anything one needs to know about taxes in the country.

In addition, the text is written in English to avoid the difficulty of translating from Danish to English and to pass the data to the large language model.

For this purpose, tools and libraries have been used such as :

- **chromium browser:** which allows automated browsing without graphical interface.

- **Selenium:** a library for controlling web browsers by providing APIs for interacting with web page elements such as links.

- **BeautifulSoup:** a python library for parsing HTML content and extracting the required data.

Finally, it stores the extracted data in a JSON file. Llama 3 model from Together.ai was used to convert the text into a structured "id," "title," and "content" format. Together AI is an innovative platform that focuses on enhancing collaborative AI development and deployment at a low cost. By integrating various AI models and facilitating seamless interaction, Together AI allows for building robust and efficient AI solutions. (TogetherAI, 2024).

7.3 Develop the RAG pipeline

After scraping the data from skat.dk, the next crucial step was to prepare it for efficient retrieval and processing by storing it in a vector store. This process involved converting textual data into numerical vectors that capture the semantic meaning of the text, thereby facilitating similarity searches.

First, to manage environment variables securely, the dotenv library was used. This allowed to load sensitive information such as API keys and database endpoints from an .env file. The relevant environment variables, including the Together API key, Astra DB API endpoint, application token, and namespace, were loaded and verified to ensure they were set correctly.

Next, the Together AI models were initialized to handle both the embedding creation and the language model tasks. The TogetherEmbeddings model was used to convert the textual data into vector embeddings, while the Together models (specifically "microsoft/WizardLM-2-8x22B" and “mistralai/Mixtral-8x7B-Instruct-v0.1”) were used for generating responses. Both of them were tested, and the answers generated were similar. These models were configured with parameters like temperature, top_k, top_p, repetition_penalty, and max_tokens to optimize performance. More specifically:

- **Temperature:** This parameter controls the randomness of the model's predictions. A higher temperature (e.g., 1.0) produces more random results, while a lower temperature (e.g., 0.2) makes the output more deterministic and focused.
- **Topk:** This limits the sampling pool to the top k most likely next words. For example, if topk is set to 50, the model will only consider the 50 most probable next words, adding a level of control over randomness.
- **Topp:** This parameter considers the smallest set of words whose cumulative probability is above a threshold p. For instance, if topp is set to 0.9, the model will choose from the top 90% of the probability mass, ensuring that the sum of probabilities of these words is 0.9 or greater.
- **Maxtokens:** This specifies the maximum number of tokens that the model is allowed to generate in a single response. It sets a limit on the length of the generated text.

The vector store was then set up using Astra DB, a scalable database optimized for handling vectorized data. The AstraDBVectorStore module from LangChain was used to interact with Astra DB. The initialization involved providing the embedding model, collection name, token, API endpoint, and namespace. If any issues occurred during initialization, they were caught, and an attempt to connect to the existing collection was made using _AstraDBCollectionEnvironment.

The documents containing Danish tax regulations and guidelines were loaded into the system from skat.dk. These documents provide the necessary information for the RAG system to generate accurate responses. However, handling entire documents can be inefficient. To mitigate this, the loaded documents were split into smaller sections based on logical divisions such as paragraphs, facilitating efficient retrieval.

The vectorization process begins with the initialization of Together AI's embedding model, specifically togethercomputer/m2-bert-80M-8k-retrieval, which converts textual data into numerical vectors that encapsulate the semantic content of the text. The data from skat.dk.json was loaded and parsed into a list of dictionaries, each containing an ID and the text content. For each piece of text, the TogetherEmbeddings model generates a vector embedding, allowing for efficient similarity searches. The generated embeddings, along with their corresponding IDs, are then added to the vector store. The vector store used is AstraDBVectorStore, which is specifically designed to handle vectorized data efficiently.

Next, embeddings for all document splits were stored in an index. This index is a structured database that allows for rapid searching based on the similarity between query and document embeddings. Creating this index is crucial for the efficiency and speed of the retrieval process, enabling the system to quickly find the document splits whose embeddings are most similar to the embedding of the given question. The similarity method used in this process is cosine similarity, which measures the cosine of the angle between two vectors to determine their similarity.

A prompt template was defined using ChatPromptTemplate from LangChain. This template structured the context and the user's question to ensure that the language model generated accurate and contextually appropriate answers. The template was designed to provide specific guidelines for answering questions about Danish tax regulations.

The LLMChain was created to link the prompt, the language model, and the retriever. The retriever was set up to search for the top three relevant documents. This involved configuring the retrieval process to use the vector store's search capabilities and integrating the prompt template, language model, and output parser into a cohesive chain.

When a user submits a query, the query is converted into an embedding, and the vector store is searched for the most relevant documents. The retrieved documents are then compiled into a context, which, along with the user's question, is fed into the language model to generate a response. The test_chain function facilitated this process by retrieving relevant documents, compiling the context, and running the chain to generate the answer.

The next phase involves the generation of the final answer, a critical step in the RAG pipeline. The context and the specific user query are fed into the language model (LLM), which uses its sophisticated neural network architecture to process the information. The LLM considers the provided context, which includes relevant details extracted from the retrieved documents and the specific question posed by the user. This combined input allows the LLM to generate a coherent and contextually appropriate answer. Practical prompt engineering plays a crucial role in this phase by ensuring the inputs are structured to maximize the LLM's performance, leading to precise and relevant responses.

Finally, the generated answer is delivered to the user. This response reflects the relevant documents' factual correctness and contextual nuances, precisely answering the user's query. The RAG application was deployed using FastAPI to make it accessible via an API, enabling users to interact with the application through HTTP requests. FastAPI's robust and scalable framework ensures that the RAG application can handle multiple requests simultaneously, providing quick and reliable responses to user queries. Langchain-serve was utilized to facilitate this deployment, leveraging Jina AI Cloud for scalability and serverless architecture without sacrificing the ease and convenience of local development. Additionally, langchain-serve allows deployment on private infrastructure to ensure data privacy when needed. The deployment process included the use of the Playground for making queries and obtaining answers, allowing users to interact seamlessly with the application. When a user submits a question, the RAG system retrieves relevant documents from the vector store and generates a response based on the retrieved information. Users can see which documents were used to answer their question, providing transparency and context for the response. Additionally, LangSmith was utilized to track the various runs of the RAG system, ensuring detailed monitoring and analysis of the application's performance and the responses generated.

Chapter 8

Results

As it was previously mentioned, langsmith was employed to track the runs of the models used "mistralai/Mixtral-8x7B-Instruct-v0.1" and microsoft/WizardLM-2-8x22B".

The question posed was, "How to Report Information in Your Tax Assessment Notice If you report non-Danish income for the first time?" The answer was derived using three different document IDs, the contents of which were saved in a JSON file referred to as "documents." The retrieval process is shown in Figure 8.1, where the three documents used for the answer can be observed.

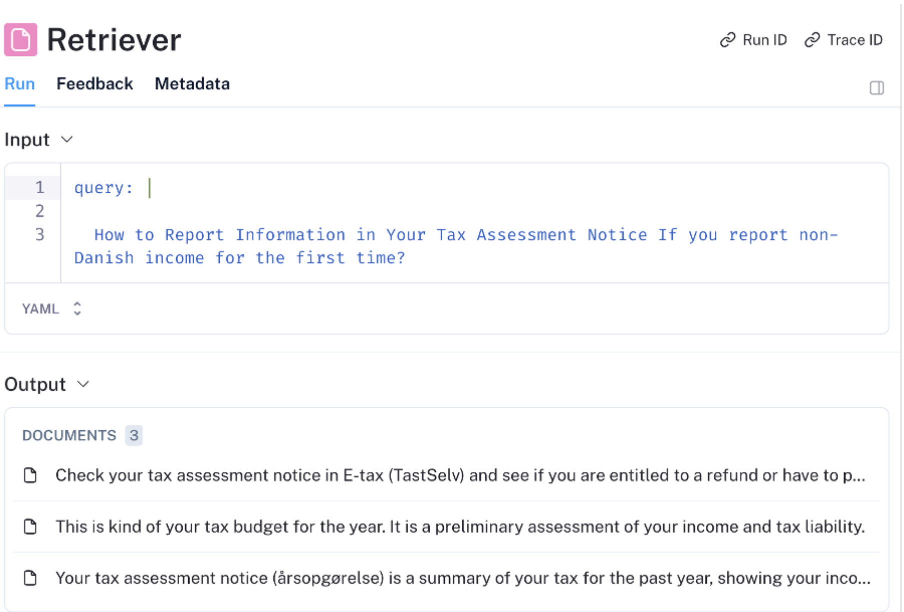

Figure 8.1: The retrieval part of RAG.

The output section demonstrates that the system identified the relevant documents but had not yet generated a final response. The documents listed were crucial for crafting an accurate answer based on the input query.

Below is an example of a run from LangSmith, showing the response generated by the "mistralai/Mixtral-8x7B-Instruct-v0.1" model.

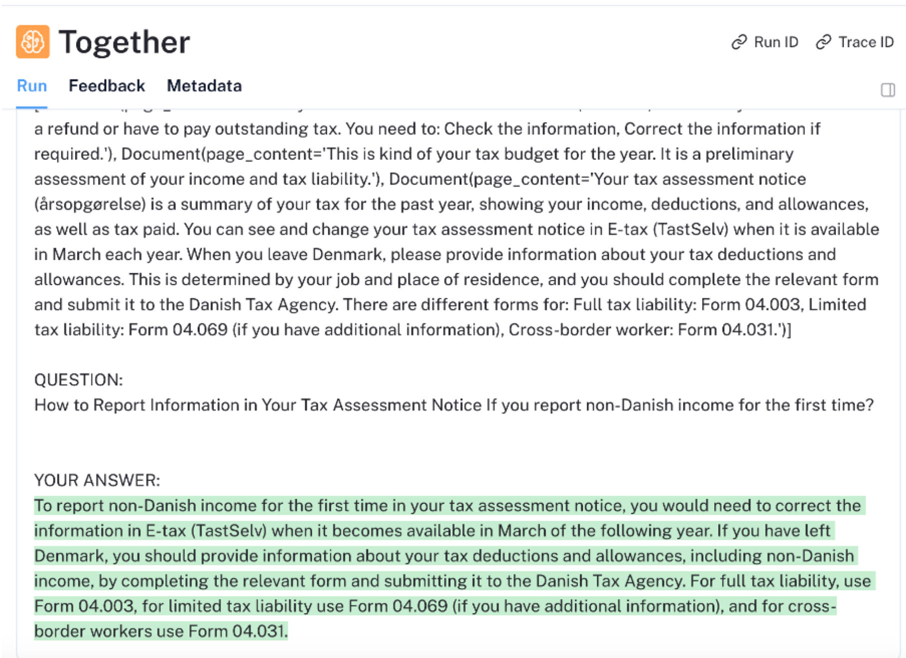

Figure 8.2: Example of the generative output.

The model provided the following answer: "To report non-Danish income for the first time in your tax assessment notice, you would need to correct the information in E-tax (TastSelv) when it becomes available in March of the following year. If you have left Denmark, you should provide information about your tax deductions and allowances, including non-Danish income, by completing the relevant form and submitting it to the Danish Tax Agency. For full tax liability, use Form 04.003, for limited tax liability use Form 04.069 (if you have additional information), and for cross-border workers use Form 04.031."

This response illustrates the integration of information from multiple documents to generate a comprehensive answer to the user's query.

Chapter 9

Advanced Evaluation methods

In this chapter, some advanced evaluation methods of a RAG pipeline ensuring its accuracy will be analyzed. These methods include Corrective RAG, Self-Reflective RAG, and Adaptive RAG. Each approach introduces unique mechanisms to enhance the system's performance:

Corrective RAG: Focuses on evaluating and correcting retrieved documents before generating responses. Self-Reflective RAG: Incorporates an iterative process for self-evaluation and refinement of responses. Adaptive RAG: Involves dynamic adjustments within the pipeline to optimize retrieval and generation based on ongoing evaluations.

9.1 Corrective RAG

Corrective RAG enhances the traditional Retrieval-Augmented Generation framework by introducing a sophisticated evaluation step to ensure that the context used for generation is highly relevant. The process begins with a user posing a question. The system initiates a retrieval operation from a pre-populated vector store, fetching a set of documents that are potentially relevant to the question (Yan et al., 2024).

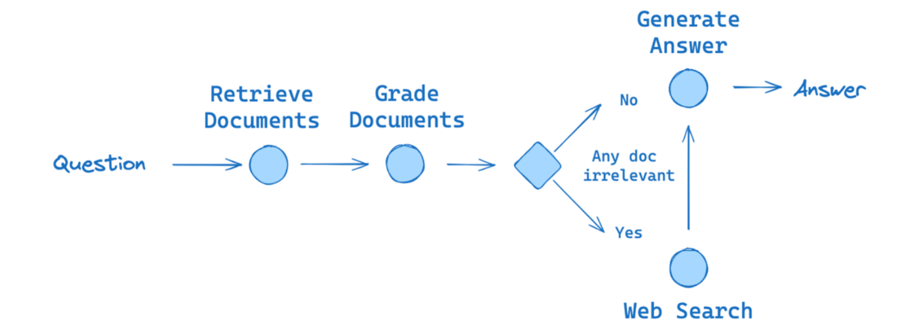

Figure 9.1: The Corrective RAG workflow.

Once the documents are retrieved, they undergo a grading process to evaluate their relevance to the user's question. This grading is crucial as it determines the usefulness of each document. Each retrieved document is then evaluated for relevance against a predefined threshold. This grading can be based on various criteria, such as semantic similarity, keyword matching, or other relevance metrics. During this phase, the system modifies its state by adding a relevance score to each document.

After grading, the system evaluates whether any of the documents are irrelevant. This evaluation creates a conditional path in the workflow. If any document is deemed irrelevant, the system rewrites the query to better target the information needed. The rewritten query is then used to perform a web search, aiming to supplement the initially retrieved documents with more relevant content. This ensures that the answer is based on the most relevant information available.

The web search node retrieves additional documents based on the rewritten query. These newly retrieved documents are then used as context for generating the answer.

Finally, using the graded documents (if all were relevant) or the supplemented documents from the web search, the system generates a comprehensive and accurate answer to the original question.

9.2 Self-Reflective RAG

Self-Reflective RAG enhances the standard RAG approach by implementing a self-evaluation and iterative refinement process. This method focuses on improving the initial generated responses through a feedback loop that evaluates and refines the output until it meets a desired level of quality and accuracy (Yan et al., 2024).

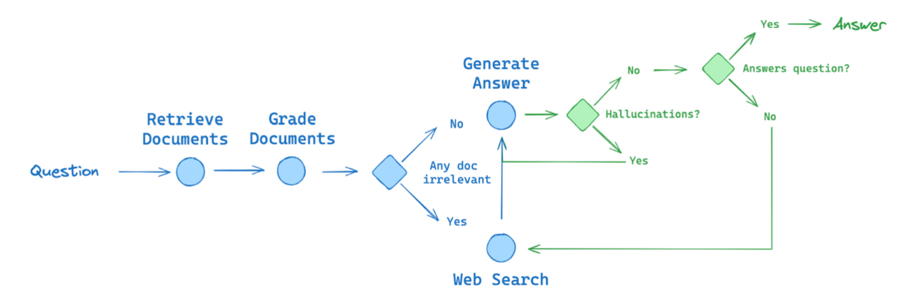

Figure 9.2: The Self-Reflective RAG workflow

.

Following the corrective RAG process, the self-reflective RAG process (represented by the green part of the diagram) comes into play. This phase involves an iterative self-evaluation and refinement of the generated response to further enhance its quality.

After generating the initial answer, the system performs a self-evaluation of the response, checking for various quality metrics such as:

- **Correctness:** Ensuring the response is factually accurate.
- **Coherence:** Assessing the logical flow and consistency of the response.
- **Relevance:** Checking the contextual appropriateness and relevance of the response to the input query.
- **Hallucinations:** Verifying that the generated response does not include any fabricated information.

If the initial answer meets these quality standards, the process terminates, and the final answer is provided to the user. If the response is found to have hallucinations, the system re-evaluates and may perform additional refinements. If the response does not adequately answer the question, the system reflects on the evaluation feedback to identify areas for improvement.

Based on this feedback, the system iteratively refines the response. This refinement may involve:

- Adjusting the document retrieval parameters to gather more relevant information.

- Modifying the answer generation strategy to enhance coherence and correctness.

- Incorporating additional context or external information if necessary.

The refined response is then re-evaluated, and the process repeats until the response meets the desired quality standards. This iterative cycle of evaluation and refinement ensures that the final answer is both accurate and contextually appropriate.

9.3 Adaptive RAG

Adaptive RAG introduces a more flexible and intelligent approach to handling queries by incorporating various tests and checks throughout the RAG pipeline. This adaptability ensures that the system can respond to different types of queries and contexts more effectively, while also minimizing errors such as irrelevant information retrieval or hallucinations in the generated answers (Jeong et al., 2024).

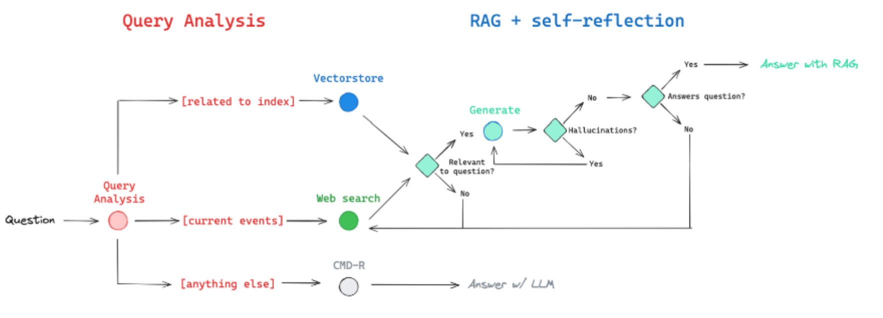

Figure 9.3: The Adaptive RAG workflow

.

The process begins with dynamic query analysis, where the system analyzes the incoming query to determine the best course of action. This analysis helps in deciding whether the query should be directed to a web search, vector store, or a different data source, ensuring the query is routed appropriately based on its nature and content.

Depending on the outcome of the query analysis, the system retrieves a set of documents from the selected source. For example, if the query is related to indexed data, it retrieves documents from the vector store; if it pertains to current events, it performs a web search. After the initial retrieval, the retrieved documents are evaluated for their relevance to the query. This involves assessing whether the documents provide pertinent information that can help accurately answer the query. If all retrieved documents are found relevant, the system proceeds to generate an answer.

If any of the retrieved documents are deemed irrelevant, the system employs conditional logic to improve the relevance of the documents. This may involve modifying the query and performing additional retrievals, such as a web search, to gather more relevant information. This step ensures that the system continuously seeks to enhance the quality of the retrieved documents. Using the relevant documents, the system generates an initial answer. This step involves synthesizing information from the documents to produce a coherent response that directly addresses the user's query.

The generated answer undergoes an evaluation to detect hallucinations. This involves checking if the answer contains any fabricated or unsupported information. The system ensures that the response is based solely on the retrieved documents and does not include any invented content.

If the answer contains hallucinations or lacks quality, the system enters an iterative refinement loop. This loop involves re-evaluating and refining the response by adjusting the retrieval strategy, refining the query, or improving the generation process. The system continuously iterates through these steps until a satisfactory answer is obtained.

Once the generated answer meets the desired quality standards—ensuring it is accurate, coherent, relevant, and free from hallucinations—the system presents it as the final response to the user.

Chapter 10

Limitations and Weaknesses

One of the primary limitations of this project was the lack of access to UFST's internal infrastructure and proprietary data due to stringent data privacy and security regulations. This restriction necessitated the use of web scraping to collect relevant data, which, although feasible, might not have resulted in the same level of data cleanliness and integrity that would have been achieved if direct access to UFST's data were possible. The scraped data, while adequate for this proof of concept, potentially included noise and inconsistencies that could have been avoided with cleaner, internally sourced datasets.

To mitigate these issues, an LLM was employed to clean and index the scraped data. However, the limitations inherent in web-scraped data mean that some nuances and specificities of the data might not have been fully captured or accurately represented. This approach, while innovative, underscores the importance of direct access to high-quality, proprietary datasets for more accurate and reliable AI applications.

Additionally, although various advanced evaluation methods for the AI model were analyzed during the project, these were not fully implemented into the application. The evaluation of the model's performance was therefore more intuitive than empirically rigorous. While the generated outputs appeared to be of good quality and relevance, a comprehensive and systematic evaluation using advanced metrics and techniques would provide a more robust validation of the model's efficacy. This limitation highlights the need for incorporating thorough evaluation methodologies to substantiate the model's performance claims.

Overall, these limitations suggest areas for future improvement, including securing access to proprietary data for cleaner datasets and implementing advanced evaluation methods to rigorously assess model performance. These steps would enhance the reliability and validity of AI applications in similar contexts.

Chapter 11

Future Development

The current conversational system serves as a proof-of-concept application with the potential for further development and deployment by UFST. This prototype demonstrates the feasibility and benefits of using RAG technology to enhance public service delivery. UFST can consider several avenues for implementing and expanding this system to serve a broader range of needs.

One primary consideration for future implementation is the deployment environment. UFST could deploy the conversational system on local servers or within a cloud infrastructure. Deploying the system locally offers enhanced data security and privacy control, ensuring compliance with stringent regulatory requirements. Local deployment also allows for more direct integration with existing IT infrastructure and may provide lower latency for users within the organization's network.

On the other hand, deploying the system on a cloud infrastructure offers scalability and flexibility. Cloud deployment can easily accommodate increased usage by dynamically scaling resources up or down as needed. It also facilitates more accessible updates and maintenance, as cloud providers often offer robust tools for managing and monitoring applications. Additionally, cloud infrastructure can provide high availability and disaster recovery options, ensuring that the system remains accessible and reliable even during peak usage times or unforeseen outages.

Another significant enhancement for the conversational system involves expanding its data sources. Currently, the system provides information to individuals who want to work or study in Denmark. By incorporating a broader dataset, the system could be extended to serve all citizens living in Denmark, offering comprehensive information on various tax-related topics. This would involve feeding the RAG pipeline with a more extensive collection of Danish tax regulations and guidelines, thus broadening the system's applicability and usefulness.

Furthermore, enabling the system to handle inquiries in Danish would significantly enhance its accessibility and usability for native Danish speakers. This can be achieved by integrating Danish language tax regulations into the RAG pipeline, allowing the system to process and respond to queries in both Danish and English. Such an expansion would ensure that the conversational system caters to a broader audience, improving its effectiveness as a public service tool.

In summary, the future implementation of this conversational system could significantly enhance UFST's ability to provide accurate and timely information to citizens. Whether deployed locally or in the cloud, expanding the system's data sources and language capabilities will increase its utility, helping UFST better serve the needs of the Danish population.

Bibliography

A. Burcu Bayram, T. S. (2021). Who trusts the who? heuristics and americans' trust in the world health organization during the covid-19 pandemic. *Social Science Quarterly*, *102(5)*, 2312–2330. https://doi.org/https://doi.org/10.1111/ssqu.12977

Abbas, N., Følstad, A., & Bjørkli, C. A. (2023). Chatbots as part of digital government service provision – a user perspective. In A. Følstad, T. Araujo, S. Papadopoulos, E. L.-C. Law, E. Luger, M. Goodwin, & P. B. Brandtzaeg (Eds.), *Chatbot research and design* (pp. 66–82). Springer International Publishing.

Ali, A. (2022). Social media chat bots. *nt J Eng Comput Sci*, *4*(1), 34–39. https://doi.org/10.33545/26633582.2022.v4.i1a.65

Alon-Barkat, S. (2019). Can Government Public Communications Elicit Undue Trust? Exploring the Interaction between Symbols and Substantive Information in Communications. *Journal of Public Administration Research and Theory*, *30*(1), 77–95. https://doi.org/10.1093/jopart/muz013

Amal Ben Rjab, P. M. (2019). Artificial intelligence in smart cities: Systematic literature network analysis, 259–269. https://doi.org/https://doi.org/10.1145/3326365.3326400

Amatriain, X., Sankar, A., Bing, J., Bodigutla, P. K., Hazen, T. J., & Kazi, M. (2024). Transformer models: An introduction and catalog.

Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through ai-guided chatbots. *Government Information Quarterly*, *36*(2), 358–367. https://doi.org/https://doi.org/10.1016/j.giq.2018.10.001

Ångström, R. C., Björn, M., Dahlander, L., Mähring, M., & Wallin, M. W. (2023). Getting ai implementation right: Insights from a global survey. *California Management Review*, *66*(1), 5–22. https://doi.org/https://doi.org/10.1177/00081256231190430

Aoki, N. (2020). An experimental study of public trust in ai chatbots in the public sector. *Government Information Quarterly*, *37*(4), 101490. https://doi.org/https://doi.org/10.1016/j.giq.2020.101490

Aristovnik, A., Umek, L., & Ravšelj, D. (2024). Artificial intelligence in public administration: A bibliometric review in comparative perspective. In M. Trajanovic, N. Filipovic, & M. Zdravkovic (Eds.), *Disruptive information technologies for a smart society* (pp. 126–140). Springer Nature Switzerland.

ATP, KL, & Digitaliseringsministeriet. (2023). *Progmodeller i danmark).* https://www.kl.dk/media/gyopkwnu/sprogmodeller-i-danmark.pdf

Bartza, E., Promikyridis, R., Promikyridis, E., & Tambouris, E. (2023). On the use of chatbots and knowledge graphs for public service information provision based on life events: The case of travelling abroad. https://www.researchgate.net/publication/373483195_On_the_use_of_Chatbots_and_Knowledge_Graphs_for_Public_Service_information_provision_based_on_Life_Events_The_case_of_Travelling_Abroad#fullTextFileContent

Berger, E. (2023). *Grounding llms.* https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857

Boyd, M., & Wilson, N. (2017). Rapid developments in artificial intelligence: How might the new zealand government respond? *Policy Quarterly*, *13*(4). https://doi.org/https://doi.org/10.26686/pq.v13i4.4619

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Jared Kaplan, P. D., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighana, T., Child, R., Aditya Ramesh, D. M. Z., Wu, J., Winter, C., Hesse, C., Chen, M., . . . Amodei, D. (2020). *Language models are few-shot learners.* https://doi.org/10.48550/arXiv.2005.14165

Chen, J., Lin, H., Han, X., & Sun, L. (2023). Benchmarking large language models in retrieval-augmented generation.

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning [Trust and Technology]. *International Journal of Human-Computer Studies*, *58*(6), 719–735. https://doi.org/https://doi.org/10.1016/S1071-5819(03)00039-9

Demircioglu, M. A. (2023). Public sector innovation: Sources, benefits, and leadership. *International Public Management Journal*, *0*(0), 1–31. https://doi.org/10.1080/10967494.2023.2276481

Desouza, K. C., Dawson, G. S., & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector [ARTIFICIAL IN-

TELLIGENCE AND MACHINE LEARNING]. *Business Horizons*, *63*(2), 205–213. https://doi.org/https://doi.org/10.1016/j.bushor.2019.11.004

Dipankar, d., deepak venugopal, v., & Kishor Datta, G. (2023). A review of generative ai from historical perspectives. *TechRxiv*. https://doi.org/10.36227/techrxiv.22097942.v1

Duggan, G. B. (2016). Applying psychology to understand relationships with technology: From eliza to interactive healthcare. *Behaviour & Information Technology*, *35*(7), 536–547. https://doi.org/10.1080/0144929X.2016.1141320

Felzmann, H., Villaronga, E. F., Lutz, C., & A., T.-L. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Society*, *6(1)*. https://doi.org/https://doi.org/10.1177/2053951719860542

Floridi, L., Cowls, J., Beltrametti, & et al., M. (2018). An ethical framework for a good ai society: Opportunities, risks, principles, and recommendations. *Minds Machines 28*, 689–707. https://doi.org/https://doi.org/10.1007/s11023-018-9482-5

Floridi, L., Cowls, J., & Beltrametti, M. (2024). How will the state think with chatgpt? the challenges of generative artificial intelligence for public administrations. *AI and Soc.* https://doi.org/https://doi.org/10.1007/s00146-023-01840-9

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey.

Ghallab, M. (2019). Responsible ai: Requirements and challenges. *AI Perspect*, *1*(3). https://doi.org/https://doi.org/10.1186/s42467-019-0003-z

Goyal, J. M., Mayne, N., Sing, D. K., Drummond, B., Tremblin, P., Amundsen, D. S., Evans, T., Carter, A. L., Spake, J., Baraffe, I., Nikolov, N., Manners, J., Chabrier, G., & Hebrard, E. (2017). A library of ATMO forward model transmission spectra for hot Jupiter exoplanets. *Monthly Notices of the Royal Astronomical Society*, *474*(4), 5158–5185. https://doi.org/10.1093/mnras/stx3015

Grimmelikhuijsen, S., Gregory, P., Boram, H., & Tobin, I. (2013). The effect of transparency on trust in government: A cross-national comparative experiment. *Public Administration Review*, *73(4)*, 575–586. https://doi.org/https://doi.org/10.1111/puar.12047

Hane, O. (2023). *Nodes 2023 - build apps with the new genai stack from docker, langchain, ollama, and neo4j (nodes2023)* [[Online; accessed 05-03-2024]]. https://www.youtube.com/watch?v=m51Dtppb2h0&list=PL9Hl4pk2FsvUu4hzyhWed8Avu5nSUXYrb&index=8

Holdsworth, J. (2024). Unlocking the power of chatbots: Key benefits for businesses and customers. *IBM*. https://www.ibm.com/blog/unlocking-the-power-of-chatbots-key-benefits-for-businesses-and-customers/

Hu, e. a. (2020). The challenges of deploying artificial intelligence models in a rapidly evolving pandemic. *Nat Mach Intell 2*, 298–300. https://doi.org/https://doi.org/10.1038/s42256-020-0185-2

Jacob R. Holm, R. J., Daniel S. Hain, & Lorenz, E. (2023). Innovation dynamics in the age of artificial intelligence: Introduction to the special issue. *Industry and Innovation*, *30*(9), 1141–1155. https://doi.org/10.1080/13662716.2023.2272724

Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2024). Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity.

Kalyani, L. K. (2023). Revolutionizing education: Artificial intelligence's pioneering role in shaping tomorrow's scholars. *international journal of multidisciplinary research in art science and technology*, *1*(2). https://doi.org/https://doi.org/10.61778/ijmrast.v1i2.6

Kemp, A. (2023). Competitive advantage through artificial intelligence:toward a theory of situated ai. *Academy of Management Review*, *0*(ja), amr.2020.0205. https://doi.org/10.5465/amr.2020.0205

Kristian, Ø., Skovby, & Signe, V., Mai. (2023). *Udskældt it-system har kostet 600.000 i konsulenthonorar om dagen i årevis – "grotesk", mener professor*. https://nyheder.tv2.dk/samfund/2023-10-09-udskaeldt-it-system-har-kostet-600000-i-konsulenthonorar-om-dagen-i-aarevis-grotesk-mener-professor

Kukreja, S., Kumar, T., Bharate, V., Purohit, A., Dasgupta, A., & Guha, D. (2023). Vector databases and vector embeddings-review. *2023 International Workshop on Artificial Intelligence and Image Processing (IWAIIP)*, 231–236. https://doi.org/10.1109/IWAIIP58158.2023.10462847

Lalor, J., Wu, H., & Yu, H. (2017). CIFT: Crowd-Informed Fine-Tuning to Improve Machine Learning Ability.

LangChain. (2023). Langchain templates [Accessed: 5/4/2024].

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, *388*(13), 1233–1239. https://doi.org/10.1056/NEJMsr2214184

Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation.

Liu, B. (2023). Grounding for artificial intelligence.

Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2022). Web scraping techniques and applications: A literature review. *SCRS Conference Proceedings on Intelligent Systems*, 381–394. https://doi.org/https://doi.org/10.52458/978-93-91842-08-6-38

Madrid, L. (2012). The economic impact of interoperability [Accessed: 15/5/2024].

Martin, L. (2024a). Rag from scratch: Part 2 (indexing) [Accessed: 5/3/2024].

Martin, L. (2024b). Rag from scratch: Part 2 (overview) [Accessed: 5/3/2024].

Martin, L. (2024c). Rag from scratch: Part 4 (generation) [Accessed: 5/3/2024].

Martin, L. (2024d). Rag from scratch: Part 4 (generation) [Accessed: 5/3/2024].

Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2023). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review*, *41*(1), 123–140. https://doi.org/https://doi.org/10.1177/08944393211034087

Meijer, A., & Wessels, M. (2019). Predictive policing: Review of benefits and drawbacks. *International Journal of Public Administration*, *42*(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664

Misischia, C. V., Poecze, F., & Strauss, C. (2022). Chatbots in customer service: Their relevance and impact on service quality [The 13th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 5th International Conference on Emerging Data and Industry 4.0 (EDI40)]. *Procedia Computer Science*, *201*, 421–428. https://doi.org/https://doi.org/10.1016/j.procs.2022.03.055

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *J Med Internet Res*, *20*(6), e10148. https://doi.org/10.2196/10148

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nirala, K.K., Singh, N.K., Purani, & V.S. (2022). A survey on providing customer and public administration based services using ai: Chatbot. *Multimed Tools Appl 81*. https://doi.org/https://doi.org/10.1007/s11042-021-11458-y

Nishimoto, B. E. (2022). Deep reinforcement learning for multi-domain task-oriented dialogue systems. *Master's Dissertation, Escola Politécnica, University of São Paulo*. https://doi.org/10.11606/D.3.2022.tde-31032023-082212

NOGA, T. (2023). The use of chatbots and voicebots by public institutions in the communication process with clients. *SCIENTIFIC PAPERS OF SILESIAN UNIVERSITY OF TECHNOLOGY - ORGANIZATION AND MANAGEMENT SERIES NO. 174*. https://doi.org/450https://doi.org/10.1145/1499586.1499695

Pan, Y., Tang, Y., & Niu, Y. (2023). An empathetic user-centric chatbot for emotional support.

Pflanzer, M., Dubljević, V., & Bauer, W. (2023). Embedding ai in society: Ethics, policy, governance, and impacts. *AI Soc*, *8*, 1267–1271. https://doi.org/https://doi.org/10.1007/s00146-023-01704-2

Ramíres-Hernández, P., Cruz, D. V., & Méndez, R. V. M. (2023). Review of artificial intelligence-based chatbots in public administration: Towards an architecture for government. *Espacios Públicos*, *24*(60). https://doi.org/10.36677/espaciospublicos.v23i60.21317

Rapaport, W. J. (2005). Book Review. *Computational Linguistics*, *31*(3), 407–412. https://doi.org/10.1162/089120105774321127

Sarrouti, M., & Ouatik El Alaoui, S. (2020). Sembionlqa: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artificial Intelligence in Medicine*, *102*, 101767. https://doi.org/https://doi.org/10.1016/j.artmed.2019.101767

Schmager, S., Charlotte, G., Elena, P., Ilias, P., & Polyxeni, V. (2024). Exploring citizens stances on ai in public services: A social contract perspective. *Data 38, Policy*, *6*, e19. https://doi.org/10.1017/dap.2024.13

Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). *Retrieval augmentation reduces hallucination in conversation.* https://doi.org/10.48550/arXiv.2104.07567

Slava Jankin Mikhaylov, A. C., Marc Esteve. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. https://doi.org/10.1098/rsta.2017.0357

Stammbach, D., Zouhar, V., Hoyle, A., Sachan, M., & Ash, E. (2023, December). Revisiting automated topic model evaluation with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9348–9357). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.581

Tam, A. (2023). A gentle introduction to hallucinations in large language models. *Machine Learning Mastery*. https://machinelearningmastery.com/a-gentle-introduction-to-hallucinations-in-large-language-models/

Tariq, S., Gupta, N., Gupta, P., & Sharma, A. (2021). Artificial intelligence in public health dentistry. *International Healthcare Research Journal*, *5*(9), RV1–RV5. https://doi.org/10.26440/IHRJ/0509.12489

Tian, D., Li, M., Ren, Q., Zhang, X., Han, S., & Shen, Y. (2023). Intelligent question answering method for construction safety hazard knowledge based on deep semantic mining. *Automa-*

tion in Construction, *145*, 104670. https://doi.org/https://doi.org/10.1016/j.autcon.2022.104670

TogetherAI. (2024). Together ai partners with meta to release meta llama 3 for inference and fine-tuning [Accessed: 15/5/2024].

Topsakal, O., & Akinci, T. C. (2023). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. *International Conference on Applied Engineering and Natural Sciences*, *1*(1), 1050–1056. https://doi.org/10.59287/icaens.1127

Valle-Cruz, D., García-Contreras, R., & Gil-Garcia, J. R. (2024). Exploring the negative impacts of artificial intelligence in government: The dark side of intelligent algorithms and cognitive machines. *International Review of Administrative Sciences*, 353–368. https://doi.org/https://doi.org/10.1177/00208523231187051

van Noordt, C., & Misuraca, G. (2019). New wine in old bottles: Chatbots in government. In P. Panagiotopoulos, N. Edelmann, O. Glassey, G. Misuraca, P. Parycek, T. Lampoltshammer, & B. Re (Eds.), *Electronic participation* (pp. 49–59). Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). Attention is all you need.

Verma, M. (2023). The power and limitations of prompt science in scientific research. *International Journal of Trend in Scientific Research and Development*, *7*(5), 642–647. www.ijtsrd.com/papers/ijtsrd59992.pdf

Wang, C., Li, Z., Chen, T., Wang, R., & Ju, Z. (2023). Research on the application of prompt learning pretrained language model in machine translation task with reinforcement learning. *Electronics*, *12*(16). https://doi.org/10.3390/electronics12163391

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J. F. J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S. M., Kenton, Z., Hawkins, W. T., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W. S., . . . Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. https://api.semanticscholar.org/CorpusID:249872629

Woods, W. A. (1973). Progress in natural language understanding: An application to lunar geology. *AFIPS '73: Proceedings of the June 4-8, 1973, national computer conference and exposition*. https://doi.org/450https://doi.org/10.1145/1499586.1499695

Xu, L., Lu, L., & Liu, M. (2024). Nanjing yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology. https://doi.org/https://doi.org/10.1186/s40494-024-01231-3

Yadav, e. a., Sinha. (2023). Ethical integration of artificial intelligence in criminology: Addressing challenges for a safer society. *International Research Journal of Modernization in Engineering Technology and Science*, *5*. https://doi.org/https://www.doi.org/10.56726/IRJMETS41779

Yan, S.-Q., Gu, J.-C., Zhu, Y., & Ling, Z.-H. (2024). Corrective retrieval augmented generation.

Yazadzhiyan, H. (2023). What are llm hallucinations? causes, ethical concerns and prevention. *Institute of Informatics and Innovative Technolgies*. https://www.researchgate.net/publication/376829203_What_are_LLM_hallucinations_Causes_ethical_concerns_and_prevention

Yoda, F. S. (2019). Atividades de chatbot no marketing de relacionamento em negócios digitais: Estudo de caso múltiplos em empresas de varejo eletrônico. *Dissertação de Mestrado, Faculdade de Economia, Administração e Contabilidade, Universidade de São Paulo*. https://doi.org/10.11606/D.12.2020.tde-10022020-175423

Zhong, G., Wang, L.-N., & Dong, J. (2016). An overview on data representation learning: From traditional feature learning to recent deep learning.

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3), 101577. https://doi.org/https://doi.org/10.1016/j.giq.2021.101577