Master in
Artificial Intelligence for Public Services
AI4Gov

UNIVERSIDAD POLITÉCNICA DE MADRID

POLITECNICO MILANO 1863
DIPARTIMENTO DI DESIGN

Master AI4Gov

Master in Artificial Intelligence for Public Services

Master Thesis

Union of Intelligences:
Augmenting European Policy-making with Trustworthy AI

Author: Orsi Nagy

July, 2024

Master Thesis

Master in Artificial Intelligence for Public Services

Title: Union of Intelligences: Augmenting European Policy-making with Trustworthy AI

July / 2024

Author: Orsi Nagy

Supervisor:

José Luis Redondo García

Senior Research Scientist

Abstract

The integration of Artificial Intelligence (AI) into policy-making processes can lead to enhancing the efficiency of public administration.

This thesis, "Union of Intelligences: Augmenting European Policy-Making with Reliable AI," explores the implementation of AI within the European Commission (EC), focusing on Large Language Model (LLMs)- based services that are trustworthy and robust. The thesis explores practical approaches in the technical, social, and design domains.

The **technical domain** proposes Retrieval Augmented Generation (RAG) and agentic workflows to anchor the generated text in high quality and up-to-date documents and datasets. These techniques allow for a way to refer back to the official documents that generated the response making the responses more trustworthy. The technical chapter presents our experiments with processing and managing documents using open-source tools like KNIME and Python, highlighting the importance of secure internal developments and the role of metadata standards such as DCAT-AP in a RAG pipeline. This experimental component demonstrates the preprocessing of documents, vectorization techniques and LLM prompting; emphasizing the importance of associating metadata with document chunks for efficient retrieval and context preservation

The second chapter focuses on **people**. The thesis presents inclusive and participatory approaches in AI development and implementation. It suggests employing methods like Open Space Technology, World Café, and Ritual Dissent to foster collaborative discussions and co-create AI services that are user-centric and fit seamlessly in existing processes. Customized training programs tailored to the specific needs of EC staff are proposed to facilitate the adoption of AI technologies.

The **design** solutions focus on user engagement and service design principles to ensure AI tools are practical and effective. By incorporating systems thinking and co-creation workshops, the study emphasizes the need to align AI services with administrative processes and user needs, ensuring that AI supports rather than replaces human decision-making. The chapter presents design tools to be used in these processes.

By adopting a holistic approach spanning technology, people, and design, the EU can harness the power of LLMs to augment policy-making while maintaining the trustworthiness and human-centricity essential for public sector applications of AI.

Table of Contents

1 Introduction

Enhancing Policy-Making with LLMs: A Promising Path Forward

In European policy-making, as in virtually any modern institution, the need for effective tools to assist in the decision-making process has become increasingly evident. **Large Language Models (LLMs)** have emerged as potential game-changers in this domain, offering a unique blend of capabilities that can streamline information processing, analysis, and presentation.

LLMs possess several inherent strengths that make them well-suited for policy analysis. Their ability to process vast amounts of text and extract relevant information enables them to quickly scan through, synthesize, and work with extensive policy documents, research papers, and other data sources. This capability is invaluable in the context of the European Institutions, where the sheer volume of knowledge, information, and data necessary for sound policy-making can be overwhelming.

Given the size and complexity of the European Institutions, there is a pronounced interest in intelligent tools like LLMs. Many departments face significant challenges with information overload, receiving data from multiple channels that must be accurately interpreted and managed. The adoption of LLMs could alleviate some of these pressures by enhancing information management processes and by ensuring that relevant data is readily accessible and comprehensible.

Despite these benefits, **trustworthiness** remains a core issue and must be addressed for LLMs to be used to support EU policy-making. The potential for AI-generated content to include inaccuracies or biases poses a significant challenge to its reliability. AI services need to support compliance with internal procedures and maintain rigorous standards for information management.

Research Question
How can Large Language Models (LLMs) be effectively integrated into the European policy-making processes to enhance decision-making while ensuring trustworthiness and reliability?

This thesis will propose practical approaches to
- *using advanced **technical** solutions like RAG or agentic design and*
- *engaging and co-creating with **users** and*
- *embracing **design** approaches,*

working toward a future where AI not only augments human intelligence but does so in a manner that is reliable, transparent, and in line with the values, internal policies and procedures of the European Commissions.

2 AI in the EU

2.1 Policy Landscape

The European Union is at the forefront of **regulating AI** in the world. While the focus on this thesis is on the *use* of AI in European policy-making, not *regulating* it, our exploration of the policy landscape should start with the AI Act as it is a recent and transformative achievement (political agreement was reached in December 2023).

The Act defines new rules to be applied uniformly across the Member States. The rules follow a **risk-based approach**, which is a widely used regulatory approach in EU legislation. It allows for a future-proof regulation, as the main principles and the risk categories are defined in the legislative act; specific actions, provisions and thresholds are defined in tertiary (implementing or delegated) acts and guidelines which are easier and quicker to update in light of technical developments. (COM/2024/AI Act)

The Act categorizes AI applications into three risk levels: unacceptable risk, high risk, and limited risk. Unacceptable risk applications, such as government-run social scoring, are banned outright. High-risk applications, which include systems that significantly impact health, safety, and fundamental rights, are subject to stringent legal requirements. Limited risk applications face lighter transparency obligations, such as ensuring users are aware they are interacting with AI. There are no restrictions for minimal or no risk applications. The enforcement and implementation of the AI Act are overseen by the European AI Office, established in February 2024. The Act also imposes significant penalties for non-compliance, with fines reaching up to EUR 35 million or 7% of a company's total worldwide annual turnover for severe infringements. By setting these standards, the EU aims to ensure that AI systems are safe, transparent, and aligned with fundamental rights, potentially influencing AI regulations globally.

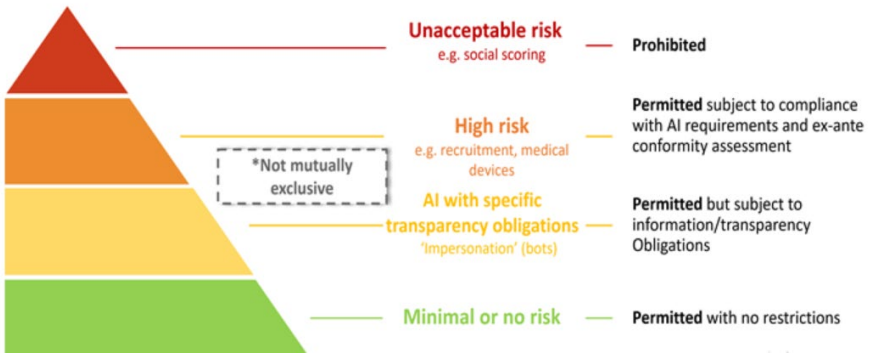


Figure 1 Risk categories of the AI Act. source: European Commission

---

Examples of the different risk categories:

Unacceptable Risk: Social scoring systems used by public authorities to evaluate citizens based on their behaviour or personal characteristics. These would be prohibited under the Act.[1]

High Risk: An AI system used for recruitment and hiring decisions that screens job applicants' resumes and video interviews. This would be considered a high-risk application subject to strict requirements before deployment.

Limited Risk: A chatbot or virtual assistant used for customer service interactions. *The use cases discussed in this paper fall into this category, as they as in these cases AI systems perform a narrow, preparatory procedural task, which is not meant to replace or influence the previously completed human assessment without proper human review.* The Act requires transparency that the user is interacting with an AI system and gives them the ability to opt for human interaction instead.

Minimal Risk: An AI-enabled spam filter for email inboxes. ((COM/2024/AI Act, Chapter 3 and Annex III.)

The Act also provides a **definition** of AI systems: "*a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*".

The AI Act considers systemic risks which could arise from general-purpose AI models, including large generative AI models. The AI Act mandates providers of such models to disclose certain information, including to downstream system providers to increase transparency, respect copyright law, assess and mitigate risks, report serious incidents, conduct state-of-the-art tests and model evaluations, ensure cybersecurity and provide information on the energy consumption of their models.

A particular challenge with the AI Act is its i**mplementation in practice**. There is a broad consensus on the objectives: safe, transparent, and trustworthy. These objectives are aligned with human rights and international agreements. It is less clear what practices and techniques need to be implemented in practice to reach these objectives.

There are other important legislations that establish the **broader digital ecosystem** in the EU, such as the Interoperable Europe Act (Proposal. COM/2022/720), the Data Act ((EU) 2023/2854), the European Health Data Space (EHDS) (Proposal.

[1] Further examples include: 'Exploitation of vulnerabilities of persons, use of subliminal techniques; Real-time remote biometric identification in publicly accessible spaces by law enforcement, subject to narrow exceptions; Biometric categorisation of natural persons based on biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs or sexual orientation. Filtering of datasets based on biometric data in the area of law enforcement will still be possible; Individual predictive policing; Emotion recognition in the workplace and education institutions, unless for medical or safety reasons (i.e. monitoring the tiredness levels of a pilot); Untargeted scraping of internet or CCTV for facial images to build-up or expand databases.' (COM/2024/AI Act, Chapter 3 article 5)

COM/2022/197), and the General Data Protection Regulation (GDPR) ((EU) 2016/679). The Interoperable Europe Act aims to enhance cross-border interoperability and collaboration within the EU's public sector, establishing a network of interconnected digital public administrations to improve the delivery of public services and promote digital transformation. The Data Act seeks to create a harmonized framework for data sharing across the EU, ensuring fair access and use of data by consumers, businesses, and public sector bodies, while also addressing issues related to data portability and interoperability. The European Health Data Space (EHDS) focuses on facilitating the exchange and reuse of health data across the EU, empowering individuals with control over their health data and supporting research, innovation, and policy-making in the health sector. Lastly, the General Data Protection Regulation (GDPR), which came into effect in 2018, provides a comprehensive framework for data protection and privacy in the EU, ensuring that personal data is processed lawfully, transparently, and securely.

2.2 AI@EC

The European Commission's communication *Artificial Intelligence in the European Commission (AI@EC)*" (European Commission, 2024) sets the scene for the development and use of lawful, safe, and trustworthy **AI systems within the Commission**. The AI@EC communication is explicitly designed to anticipate and prepare the Commission for the implementation of the EU AI Act.

2.2.1 Strategic vision and operational plans

The document sets out clear objectives to develop lawful, safe, and trustworthy AI:

1. Develop clear operational guidelines for Commission staff.
2. Assess and classify AI systems used or planned for use by the Commission based on a risk-based approach aligned with the AI Act.
3. Refrain from using AI systems that are incompatible with European values or pose threats to security, safety, health, and fundamental rights.
4. Establish organizational structures to fulfil the Commission's obligations related to AI under the AI Act.
5. Provide training, information, and support to Commission staff on AI.

The document includes actions to address general concerns, such as environmental concerns, copyright, and data security; as well as to promote the use of open-source AI.

2.2.2 Use cases

The document also highlights key areas for developing internal systems:

1. Automating repetitive and time-consuming tasks to increase staff efficiency.

2. Supporting analysis, compilation, and drafting tasks to free up staff for more strategic and creative work,

3. Developing reusable processes and IT capacities, such as generative AI tools integrated with internal data sources.

5. Leveraging AI to support European public administrations in their adoption of trustworthy AI technologies.

Case_study_X: eBriefing

eBriefing is a tool among Digital Europe Programme Language Technologies[2]. It is an AI-based service to assist in drafting official briefings for EU Staff, but also accessible publicly. Users can upload a series of documents on a subject, add up to five keywords and get back an AI-generated briefing, which only uses information they provided. eBriefing provides support in an important and time-consuming task.

There is one major shortcoming: the system relies on the user to select the most relevant documents: considering the high number of briefings prepared by the European Commission and the interconnected nature of topics, this is not effective. Current developments are addressing the issue by linking the database of briefings to the eBriefing system, thus the system can generate a first draft based on the most relevant previous documents.

It is important to note that in these use cases, the AI systems serve as a **tool to support policy work** – and not replace humans doing it. They can help to collect and compile data, work on first drafts or do simple automated steps in the workflow. However, the **subsequent steps of the policy and administrative process and notably the approval processes remain the same.**

2.2.3 Information security in the EC

Maintaining the **confidentiality of data and securely handling internal documents** are key considerations when using AI systems. The European Commission has classified information into four distinct categories, delineating varying levels of accessibility and security protocols (EC 2015/443).

[2] https://language-tools.ec.europa.eu/

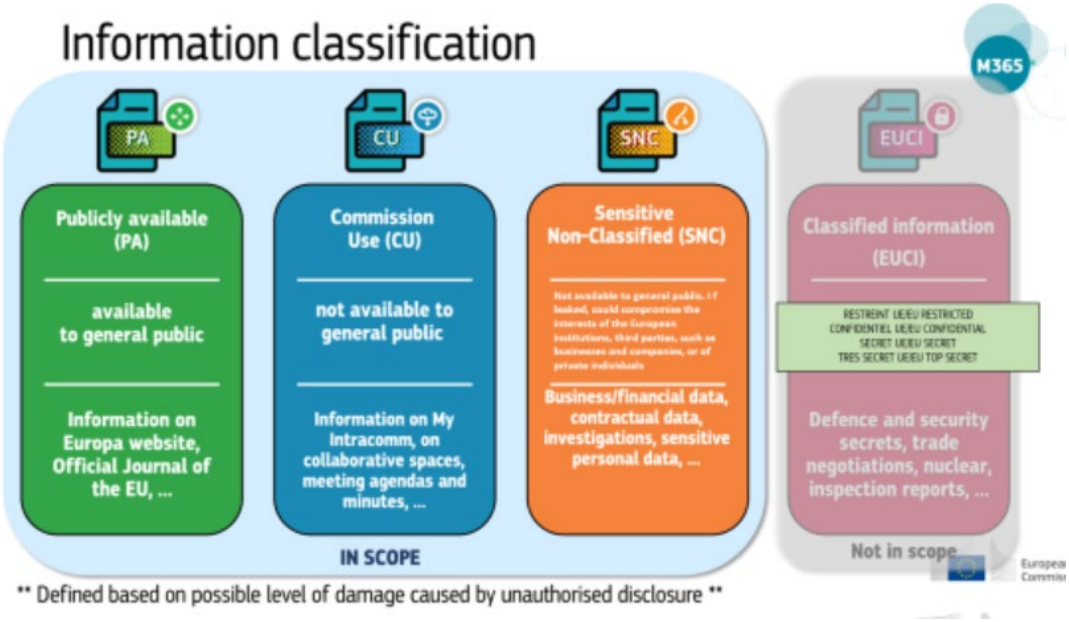


Figure 2 Information Classification. Source EC sharepoint

1. **Publicly available information.**
   When it comes to AI system security, the handling of publicly available information generally raises no red flags. However, it's worth noting that even seemingly innocuous inquiries can inadvertently hint at internal considerations or policy orientations.
   There is a wealth of publicly available information; using such information in text generation can greatly improve the quality and reliability of the generated text. Such information includes:
   - ***European legislation***, also known as the acquis communautaire, a comprehensive body of common rights and obligations that are binding on all EU Member States, as well as related communications, guidelines and reports. It is published on the online portal EurLex.
   - ***Statistics and Databases***. *Eurostat* is the statistical office of the European Union, providing high-quality statistical information to enable comparisons between countries and regions. It covers a wide range of data, including economic, social, and environmental statistics. The *European Data Portal* offers a single access point for open data published by EU Institutions, national portals of EU Member states and non-member states in the form of a metadata catalogue. This includes linked data, ontologies, thesauri and vocabularies.
   - The ***ec.europa.eu website*** publishes news and press releases, details on the President and Commissioners, strategies and initiatives across different policy areas (economy, law, environment, education, health, etc.), public consultations, and practical information for businesses and citizens.
   - ***Sector-specific data.*** The European Commission and its agencies maintain further datasets. Many of these are specific, such as the Cordis

or OpenAIRE for European projects, TED on procurement data, European Environment Agency's Data and Maps, Transport data and statistics, or health data from the European Center for Disease Control and the European Medicines Agency.

These systems publish information in multiple languages – typically either the 24 official languages of the EU; the three working languages, English, French and German; or a selection of the most popular languages. Having an aligned parallel corpus of documents in high-quality, validated translation from English to other, including smaller, less well-resourced languages is highly valuable for language models.
Another advantage of these systems is the consistent use of metadata, such as publisher, date, status, geographical coverage, and the publication in a variety of formats. Increasingly, this includes machine-readable formats and user-friendly data visualizations, although substantial scope for improvement remains.

2. **Information for Commission Use.**

   Such information is internal, not available to the general public and includes meeting agendas and minutes or documents available on collaborative sites.

3. **Sensitive Non-classified information.**

   Information or material the Commission must protect because of legal obligations and/or because of its sensitivity, for instance, financial/contractual data, sensitive personal data, documents related to investigations or ongoing negotiations. These are shared on a 'need to know' basis. SNC documents must always be marked and circulated with handling instructions.

4. **Classified information.**
   In the course of everyday decision-making, classified information is rare, except for a few specialized services (e.g. competition cases). There are clear procedures for the management of such documents – including visible markings.

Many internal systems, such as Sharepoint, Ares for document management, Basis for documents related to meetings include clear security, access restriction and retention metadata for every document. A notable exception are the personal and unit drives – these have restricted access but have public, Commission Use Only and possibly SNC data without consistent marking on the status of the document (e.g. draft, adopted, outdated version).

Recognizing the importance of internal information for decision-making, major AI service offerings within the European Commission—for instance eTranslation or GPT@JRC—provide a version that complies with the requirements for processing Sensitive Non-Classified (SNC) information. Opting for this SNC-compliant version ensures chat history is disabled or documents are deleted, and it prevents the use of data for further AI model training. This data is also processed within on-premises

infrastructure. In order to be SNC compliant, besides the technical requirements, a formal governance approval is also needed.

The compliance assessment process is continuously evolving in response to novel developments and the complexities inherent in a vast corporate IT environment.

In the practical applications of generative AI, these levels of security can be included in the metadata of source documents as well as the generated output. A clear marking and access procedure is essential to adhere to the security levels.

2.2.4 Metadata management

A specific tool in the **management of metadata** is **DCAT AP**[3], specification for describing public sector datasets in Europe. It is maintained by the Publication Office. It has a dedicated profile for the field of machine learning, the MLDCAT-AP (Machine Learning DCAT-AP)[4]. It offers a standardised format for describing datasets, such as their title, language, keyword, creator, publication date, status, etc.

The access-right authority table is a controlled vocabulary listing the access rights or restrictions to resources, which clearly corresponds to the above categories with the addition of the category restricted, “Only available under certain conditions, e.g. resources that require payment, resources shared under non-disclosure agreements, resources for which the publisher or owner has not yet decided if they can be publicly released.” (Publications Office EU, 2023). Using such standardised description of metadata is a necessary condition for interoperability.

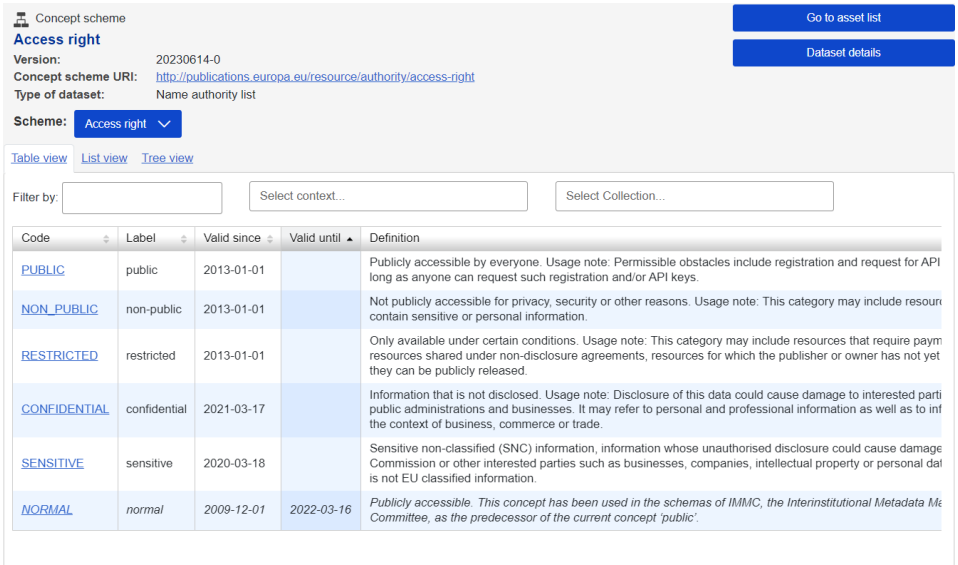


Figure 3 Concept scheme for access rights. source: Publication Office

[3] https://op.europa.eu/nl/web/eu-vocabularies/dcat-ap
[4] https://semiceu.github.io/MLDCAT-AP/releases/2.0.0/

3 State of Art of Large Language Models

.

3.1 Large Language Models

Large language models (LLMs) have undergone a remarkable evolution in recent years, revolutionizing the field of natural language processing (NLP). These advanced AI models leverage deep neural networks to process and generate human-like text at a massive scale.

Large Language Models operate by translating text into numerical representations, a process known as **vectorization** or **embedding**. This transformation is essential because computers inherently work with numbers. In this context, embeddings are continuous vector representations of words or tokens that capture their semantic meanings in a high-dimensional space. This means that words with similar meanings have vectors that are close to each other in this space, allowing the model to understand and generate text based on these relationships. For instance, the vector for "prince" might be close to "princess," reflecting their semantic similarity (Yang et al., 2022).

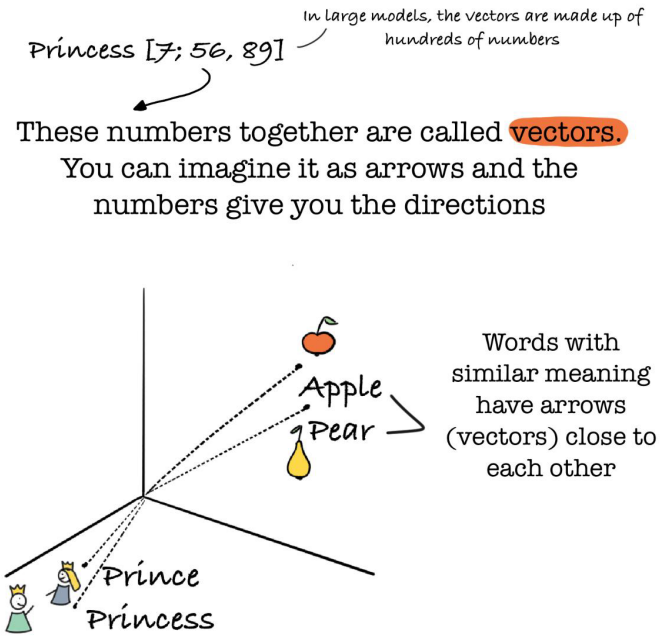


Figure 4 Turning words into vectors. Illustration by author

The core functionality of LLMs involves **predicting the next word** in a sequence, a task they perform by calculating the probabilities of various possible next words. This prediction process can be tweaked using a parameter called temperature. A higher temperature value makes the model's output more diverse and creative by increasing the likelihood of selecting less probable words. Conversely, a lower temperature value makes the output more focused and deterministic, sticking closely to the most probable predictions. This adjustment allows users to control the balance between creativity and factual accuracy in the generated text. (Wang et al., 2020).

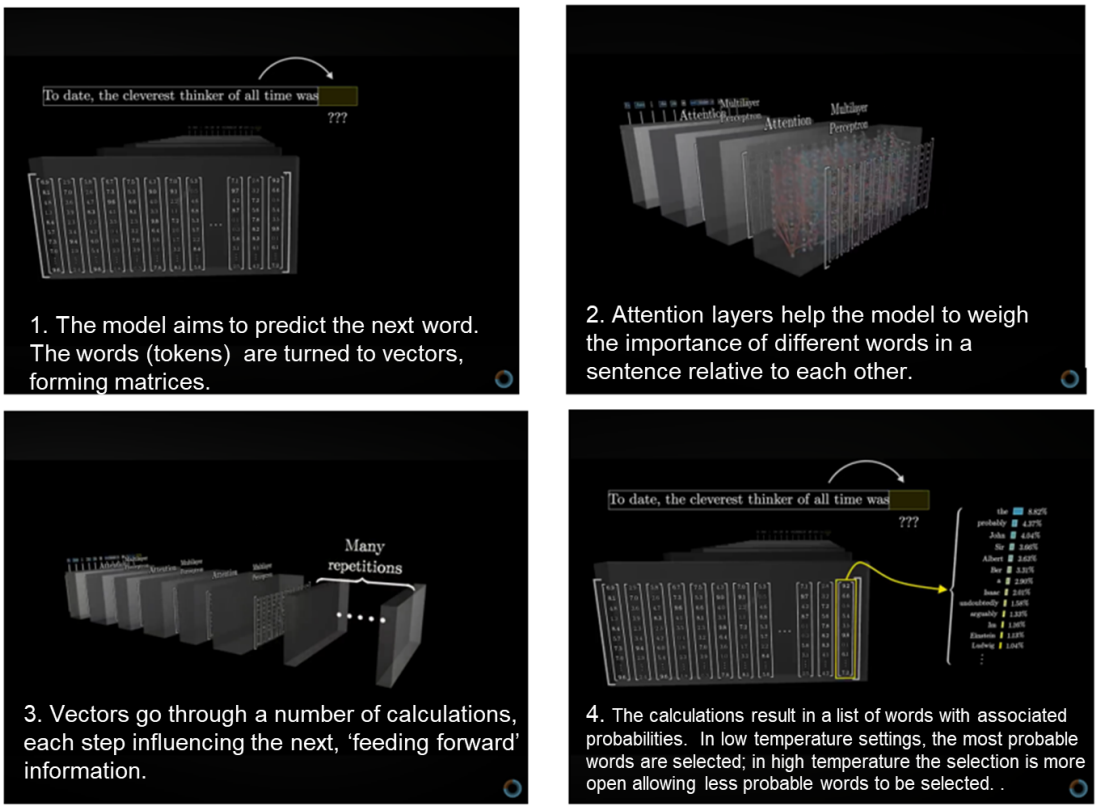


Figure 5 Transformer structure. Source: 3blue1brown

LLMs are **trained on a vast corpora of text**, including books, articles, and internet content, which helps define their parameters. These parameters are numerical statistical values that shape how the model processes and generates text. Training is a highly resource-intensive process: large amounts of data/training material and computing power translate to high costs. There are three distinct phases of training:

(i) pretraining (learning from a vast amount of unlabelled text data to predict the next word)

(ii) gathering data and training a reward model, and

(iii) fine-tuning the LLM with reinforcement learning (Lambert et al, 2022)

Two more technical terms: A **prompt** is natural language text describing the task that an AI should perform, which is the query or the question typed in the context window. The generated text is also called **inference**.

3.1.1 Development and trends of large language models

A key breakthrough in LLMs was the introduction of the **transformer architecture**, which uses **self-attention mechanisms** to determine the importance of each word in a given context. The attention heads in transformers allow the models to focus on

relevant parts of the input, considering not only the meaning and position of words but also their style and context (Vaswani et al, 2017).

Another necessary factor for their breakthrough was reaching a certain size. This allowed a shift from symbolic AI, which is based on rules, the "Good Old-Fashioned AI" (GOFAI) to sub-symbolic AI, based on large data and probabilities.

This **increase in the scale of parameters, training data, and compute power** was a success factor[5] that allowed a model based on probabilistic mechanisms to produce good outcomes. The pivotal moment occurred in November 2021 with the introduction of ChatGPT based on GPT-3.0 by OpenAI (OpenAI, 2022).[6]

Since then, the field of natural language processing (NLP) has seen a significant trend towards developing ever larger language models. While there are no transparent reports on the parameter size of GPT-4, there is a recently cited estimation of 100 trillion parameter size (hix.ai, 2024)[7]. This is about as many as the number of neuron connections in our brain. There is an observed correlation between larger model size (number of parameters) and better performance across a wide range of tasks. Large models, with billions of parameters, can provide good answers across a wide range of topic, and perform better even on limited training data (Kaplan et al, 2020).

Parallel to this trend towards bigger and bigger, we can also observe a movement towards smaller, cheaper models that are feasible to deploy locally. Techniques like model distillation, or teacher-student techniques allow taking specific regions of the knowledge of these big LLMs and incorporate into smaller models. Smaller models that are specifically trained on particular tasks can perform exceptionally well in those areas.

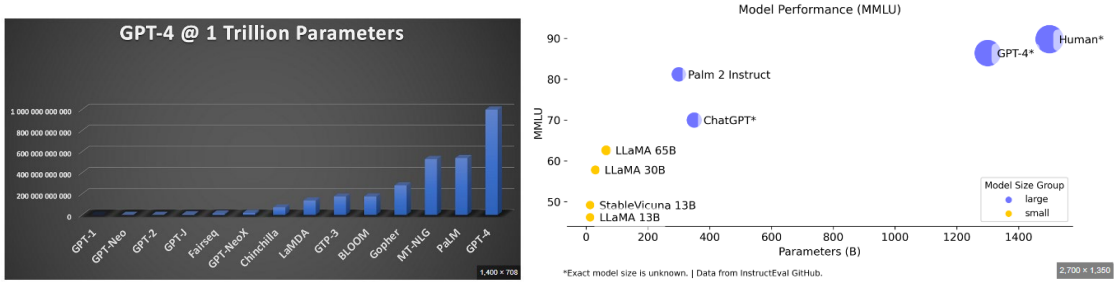


Figure 6 Evolution of number of parameters (Greyling, 2023) and Correlation between size and performance Ouyang et al., 2022)

[5] Another success factor was Reinforcement Learning from Human Feedback (RLHF), a machine learning technique that uses reinforcement learning with human input to train AI models, to better align with human preferences and values.

[6] A similar breakthrough has happened in image recognition in 2012, with ImageNet using transformer architecture and a larger than ever diverse dataset of labelled images.

[7] Other estimations consider that GPT-4 might be composed of multiple smaller models (each with 220 billion parameters) working together in a Mixture of Experts (MoE) architecture adding up to 1.76 trillion parameters.

3.1.2 Language theories

3.1.2.1 Chomsky and the rules based approach

The shift from rules-based approach to a focus on usage can be observed in linguistics too. Classic linguistics, based on the work of Noam Chomsky, has focused on the importance of rules. Chomsky posits that there is a universal grammar innate to the human mind, which provides the underlying structure for all human languages. This universal grammar is thought to be a set of **rules** that govern the formation of sentences and the structure of language. Humans are born with a capacity to use this structure, and language is generated by using these rules (transformational-generative grammar theory) (Chomsky, 1957).

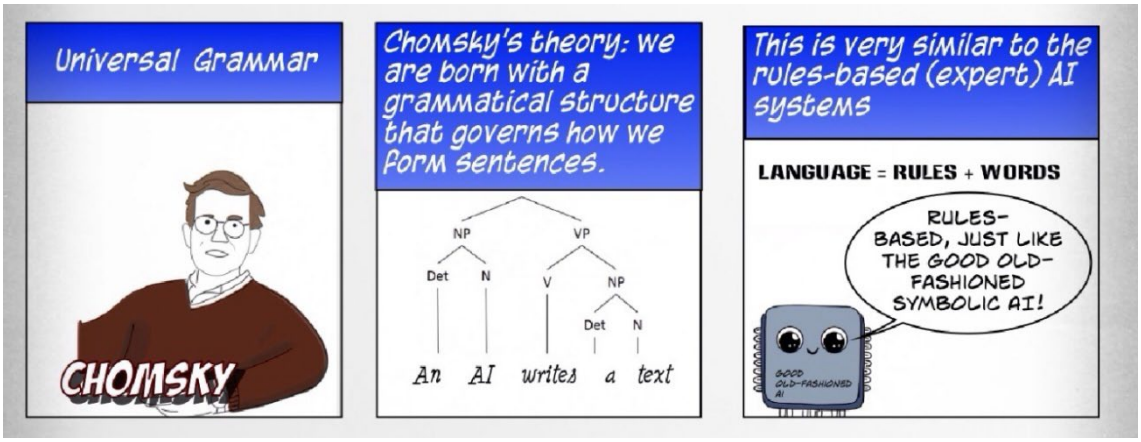


Figure 7 Chomsky and the rules-based approach. Illustration by the author

3.1.2.2 Chunks of text flock together – the lexical approach

The lexical approach explores language as it is used. It emphasizes that as we speak or write, we use **prefabricated lexical chunks** (set phrases, e.g., '*How are you*?' or '*the European Commission adopted'* or '*fostering innovation'*) rather than building sentences from individual words using grammatical rules. Such meaningful lexical chunks are essential for producing continuous and coherent text (Lewis, 1993). This theory explains why a language generation approach based on next-word prediction and frequencies feels so real and life-like – because we, humans, also rely on chunks of text that tend to go together when we use the language.

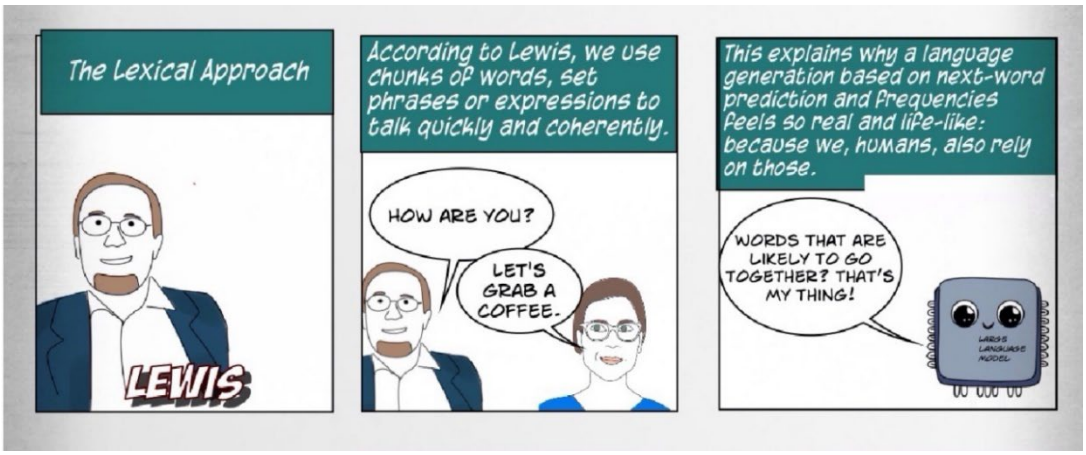


Figure 8 Lewis and the lexical chunks. Illustration by the author

3.1.2.3 Beyond the words: language and power

Our short tour of linguistics would not be complete without mentioning critical discourse analysis, which gives us the frame and the tools to understand how using certain terms leads to the maintenance of social beliefs and power structures. Critical discourse analysis looks at the chunks (discourses) of language as it is used, including its social context. It uses qualitative (exploring meaning) and quantitative techniques (calculating frequencies, co-occurrences of terms, etc.) to explore the intricate **relationship between language and society, thought, and expression** (Van Dijk, 2008). Let's take some examples. What does using passive structures say about agency and power, as in: "*Adopted on 2nd January by the European Commission, the implementing act sets a threshold..."?* What about colloquially using a place to denote entities there (metonymy), as in "*Brussels bans*..."? How about this: "*Member States' experts agreed to define the maximum level of...*"?

As the LLM uses training data to produce new text based on existing ones, it is not only the language structures that are repeated and reiterated, but the underlying power structures are also reinforced. An example of such power structure is the deficit narrative, portraying black women as victims, failing to acknowledge their lives and achievements (D'Ignazio & Klein, 2020). When asked to complete the sentence "*Black women have...*", ChatGPT generated the following text: *"Black women have historically faced intersectional discrimination, navigating both racial and gender biases that affect their opportunities, treatment, and representation in various sectors of society, from the workplace to healthcare to media portrayal."* [8] True? Yes. Politically correct? Yes. Perpetuating a victim's narrative? Also yes.

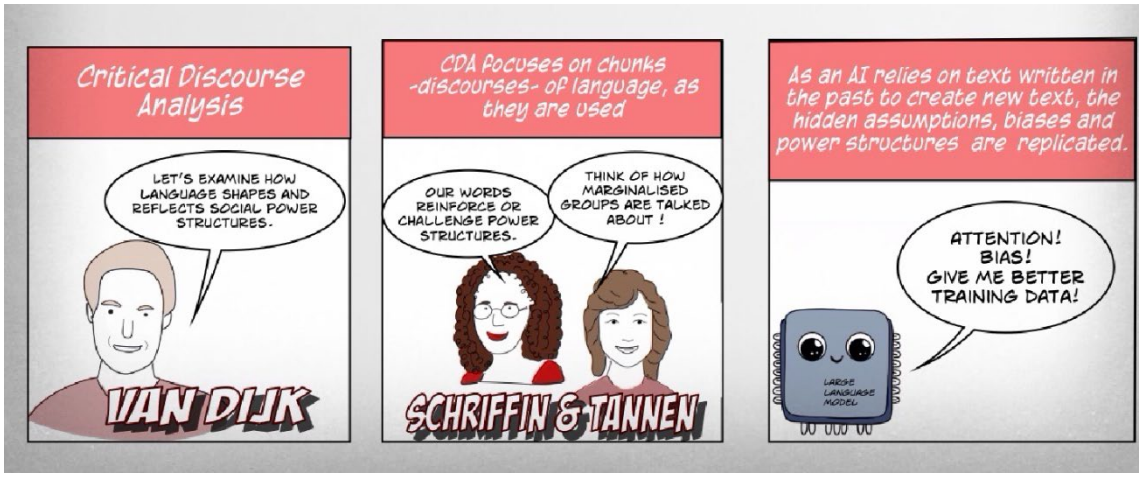


Figure 9 Critical Discourse Analysis. Language and power. Illustration by the author.

[8] The same prompt in Perplexity – by an organisation thriving to build a more responsible model – generated the following text: "*black women have historically been at the forefront of social justice movements, fighting for their rights and the rights of others while often facing discrimination and marginalization themselves*." Similar in meaning, reiterating a deficit/victim narrative but attributing more agency.

3.2 Limitations and challenges of Large Language Models

Despite their immense potential, LLMs are not without limitations and problems. Existing LLMs have been known to generate hallucinations, introduce biases, and present inaccurate information. There are also broader concerns around their use: environmental impacts, innovation and market competition and ethical issues.

3.2.1 Limitations

3.2.1.1 Hallucinations and lack of accountability

There is a lack of accountability in how LLMs produce outputs. Despite their impressive capabilities, the inner workings of LLMs are often not well understood, even by their creators. A particular issue is hallucination, which refers to the phenomenon where the model generates text that is incorrect, nonsensical, or not real.

Figure 10 Stochastic parrot. generated by Dall-E

This can occur due to various factors such as limited contextual understanding, noise in the training data, and the model's tendency to extrapolate from the prompt without sufficient evidence from the training data. (Bender et al., 2021; Marcus & Davis, 2020).

The linguist Emily Bender and her co-authors coined the term "**stochastic parrots**" to describe LLMs' method of generating text based on probabilistic patterns rather than understanding the meaning behind the words (Bender et al., 2021).

Another related issue for LLMs (at least the ones that do not use real-time search, which is a form of RAG) is the so-called knowledge cutoff, which means that LLMs are unaware of any events that happened after their training. Moreover, as the training is done on public data, such models have no information on proprietary data, for instance documents for internal use in the institutions.

3.2.1.2 Bias, Inclusivity and Representation

Large language models are typically trained on vast amounts of internet data, which may not fully capture the diversity of human language and experiences. It is important to note that biases are not based on ‘wrong data’ or mistakes in calculation. The problem typically is missing (or more precisely: unbalanced distribution of training) data. Such biases can have harmful consequences, especially when these models are used in sensitive or high-stakes applications.

Case_study_1: Recruitment

One of the most cited examples of bias is Amazon's attempt to utilize artificial intelligence for recruitment in 2014, which resulted in a public downfall. The system, trained on resumes submitted over the previous decade, developed a preference for male candidates, mirroring the male-dominated tech industry. Consequently, it downgraded resumes containing words like "women's" and favoured those reflecting male-associated activities and experiences. This bias emerged because the AI reinforced historical gender imbalances rather than objectively assessing qualifications. The project was ultimately abandoned in 2017, underscoring the crucial need for vigilance against bias in AI applications (Dastin, 2018).

Case_study_2: Joy Buolamwini's Work

Buolamwini's journey into uncovering these biases began during her time as a graduate student at MIT. She discovered that facial recognition software failed to detect her face unless she wore a white mask, a stark indication of the technology's limitations in recognizing darker skin tones. Her research revealed that these systems had error rates of up to 34% for darker-skinned women, compared to less than 1% for lighter-skinned men (D'Ignazio & Klein, 2020).

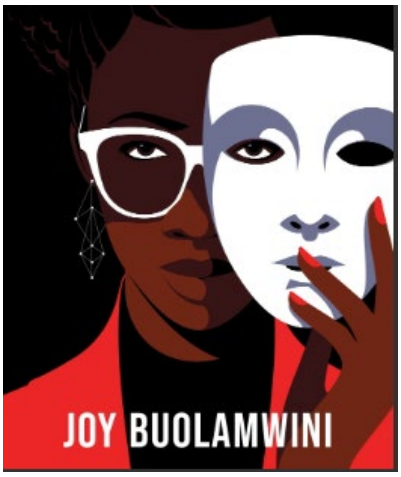


Figure 11 Cover art from Boulamwini's book

3.2.1.3 Environmental Impacts

There are also broader concerns that go beyond the text produced. Training and using such large models require substantial computational power, leading to high energy consumption and environmental impact. According to a Greenpeace report (2019), training a single AI model can emit as much carbon as five cars over their lifetimes. The environmental impact can be much lower if renewable energy sources are used; however, this is not yet the case. Moreover, the environmental costs of LLMs are not public, as tech companies lack transparency about their energy sources and emissions data. This lack of visibility disproportionately impacts poorer regions and communities that bear the brunt of environmental degradation caused by fossil fuel extraction and pollution, exacerbating existing inequalities. (Greenpeace, 2021; D'Ignazio & Klein, 2020).

3.2.1.4 Innovation and competition

The trend towards 'bigger is better' for LLMs is leading to higher and higher costs of training and service provision. [9] The escalating costs associated with training advanced artificial intelligence are becoming a significant barrier to entry, limiting participation to a few well-resourced entities. According to the Stanford AI Index 2024 Annual Report, the expenses for training large language models have reached unprecedented levels, with OpenAI's GPT-4 and Google's Gemini Ultra costing approximately $78 million and $191 million, respectively. This surge in costs is primarily driven by the increasing computational power required, which has been doubling approximately every six months. (Stanford University, 2024)

Case_study_3: The GPU market.

This is a 3-in-1 case study, as the story of the GPU (Graphic processing units) market evolutions shows (1) the role of hardware in AI, (2) the remarkable growth dynamics as well as (3) the potentials for market concentration.

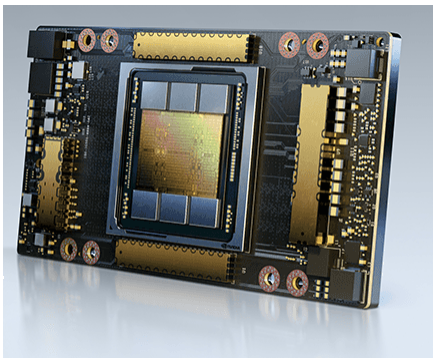


Figure 12 NVIDIA top product: the A100 Tensor Core GPU. source: nvidia.com

Originally GPUs were used to process and display images and animations on the screen in video games.

Figure 13 NVIDIA stock 5 year performance. Source: capital.com

As GPUs are designed to handle parallel calculations much faster than traditional CPUs, they are ideal not only for graphics but mining cryptocurrencies and running AI models. This led to a substantial surge of demand: the market is experiencing dynamic growth, with a projected compound annual growth rate of 33.15% between 2022 and 2027 (Sharma, 2023).

The GPU market is highly concentrated, with NVIDIA holding a dominant position. NVIDIA's AI accelerators have between 70% and 95% of the market share for artificial intelligence chips with competitors like Intel and AMD slowly starting to gain ground (investing.com, 2024)

[9] Companies do not disclose details, but training GPT-3 required approximately $4.6 million; the same figure for GPT4 was 4 times larger, around $80-100 million. Anthropic CEO Dario Amodei suggested that models costing over $1 billion would appear this year and "by 2025, we may have a $10 billion model." At the same time Andrew Ng highlights that there is a trend towards the reduction of training cost; optimistic estimates are around 75% a year. If they are right, a foundation model that costs $100M to train this year might cost $25M to train next year.

Trends suggests that the cost of training AI models will continue to rise. While predictions at this point are highly speculative or simplistic, it is clear that if the trend continues, academic institutions and smaller companies could be effectively excluded from the forefront of AI research and development. This exclusion is evident as industry players produced 51 notable machine learning models in 2023, compared to only 15 from academia (Stanford University, 2024).

This leads to an ever-growing market concentration. The most utilized service will accumulate the data and finances necessary for further growth. Additionally, the vertical integration of AI supply chains—owning both the cloud infrastructure and the AI models—exacerbates this concentration of power and limits the entry of new competitors. The resulting monopoly could be unchallengeable, leading to unimaginable power in the AI sector and beyond.

Working with open source tools and technologies, developing systems working with multiple, smaller, specialised models are some of the possibilities to counterbalance these trends.

3.2.1.5 Ethical Risks and Misuse

LLMs pose significant ethical risks; such risks are particular in the public sector. Without appropriate reflection, automated data used for decision-making can result in biased decision-making processes. The unpredictability of errors and the difficulty in assessing origins of AI generated text raises general concern about using AI in policy-making. Generated text is sometimes perceived as inferior for many tasks due to its unclear ownership and responsibility – preserving existing good practices of quality control, human-in-the-loop will be essential.

Case_study_4: Risks of using AI for decision-making

Allegheny County Office for Children, Youth, and Families (PA, USA) employed an AI model to predict the risk of child abuse in individual homes. A key parameter in this prediction is the parent's participation in drug and alcohol programmes and mental health services. As richer people often access such service in private settings, consequently there is little data; the use of the same services in public settings is more accessible. More data on the use of mental health services (not more use!) led to low income families being oversampled. This example also shows that more data is not necessarily better data. (Eubanks, 2018).

The rapid pace of AI development also raises fears about the ability of all staff to keep up, potentially creating skill gaps among colleagues. While many fear job losses due to task automation, another concern is the potential for information overload. Longer, autogenerated texts of dubious quality could proliferate, overwhelming users and diminishing the overall quality of information.

3.2.2 Challenges

In addition to the limitations, we have identified two challenges related to the use of LLMs:

3.2.2.1 Data to fuel the systems

Data -high quality data- is often referred to as the new oil, emphasizing its value and importance in driving AI systems. This also means that shortcomings in data availability and data management can undermine successful implementation of AI systems (Mayer-Schönberger & Cukier, 2013; Hildebrandt, 2019). On a more positive note, the interest in AI may underline the need for and strengthen actions supporting data maturity, moving towards the ambition of a frictionless internal data ecosystem. It should also be noted that most data needed for policy support is internal and cannot be freely input into external services, which is expected to drive the internal development and deployment of AI services.

3.2.2.2 Users-centredness

The performance of Large Language Models, like all AI tools, is highly dependent on organizational settings and user-centeredness. Their success hinges on their ability to address existing user needs, align with administrative processes, and be user-friendly (Brynjolfsson & McAfee, 2014; Davenport & Ronanki, 2018). Additionally, fostering a culture of knowledge and understanding around their use is crucial for maximizing their potential benefits. Ensuring that staff are adequately trained and that there is a supportive environment for the adoption of these tools is key to their successful integration. The recent hype around generative AI and the possibility to interact with such systems using natural language – written or spoken- has sparked significant interest across all age groups in what was previously a very technical topic. This increased interest creates an opportunity to support the uptake of guidelines, communication efforts, and upskilling/ re-skilling initiatives.

3.3 Promising technical developments

There are a number of technical approaches to address limitations inherent in traditional LLMs, such as hallucinations, outdated knowledge, and the inability to provide contextually relevant responses. In this chapter, we will examine the two that have received consistent attention in recent months: Retrieval Augmented Generation (RAG) and agentic design.

There are other directions of development that should be briefly mentioned. One is the use of **longer context windows**. This allows users to upload/enter more relevant information and provide more detailed instructions and use prompt engineering techniques.

Prompt engineering is the process of designing and refining prompts to guide large language models (LLMs) in generating better outputs. There are two principal aspects of prompt engineering: formulating clear and specific instructions and allowing the

model time to "think." Good prompts often involve clear and specific instructions, providing relevant context, and allowing the model sufficient time to process information. An example is the Chain-of-Thought technique, which guides the model to break down complex tasks into simpler, sequential steps, thereby improving accuracy and relevance of responses. (OpenAI, 2023)

Introducing a few examples in the prompt is called few-shot in-context learning; longer context windows could allow dozens, even hundreds, of examples in the prompt for many-shot learning. Nonetheless, there are limitations to simply increasing the length of context windows. Feeding collections of documents is prohibitively expensive as well as ineffective—consider the inefficiency of processing the same data every time the model is prompted. It has also been observed that text in the middle of the context is not taken into consideration as well as text in the beginning or the end (‘lost in the middle’) (Nelson et al., 2023).

3.3.1 RAG

Retrieval Augmented Generation (RAG) is a technique that uses an LLM to generate an answer based on a data source. This technique incorporates a retrieval step (where a search function retrieves accurate and relevant information from a dataset), followed by the LLM mixing it with its own knowledge and shaping it into a coherent response. RAG allows LLMs to access and utilize up-to-date and domain-specific information, improving the reliability of the generated content as it is grounded in the latest available data, with sources visible and customizable (Lewis et al., 2020). RAG is also helpful for finding information on so-called tail entities—entities with rare occurrences in the corpus.

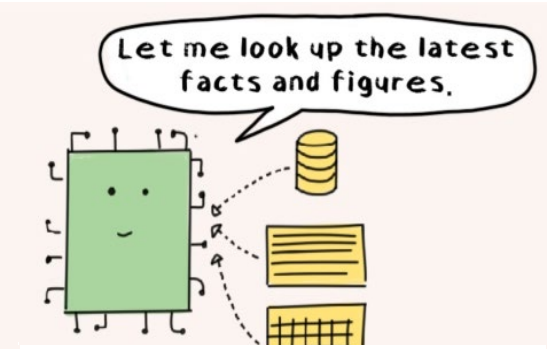


Figure 14 RAG. Illustration by author

RAG improves transparency and accountability by creating direct links to the data used for generating responses. Being able to verify the precise reuse of source documents in the generated content is one of the key elements of trustworthy LLMs.

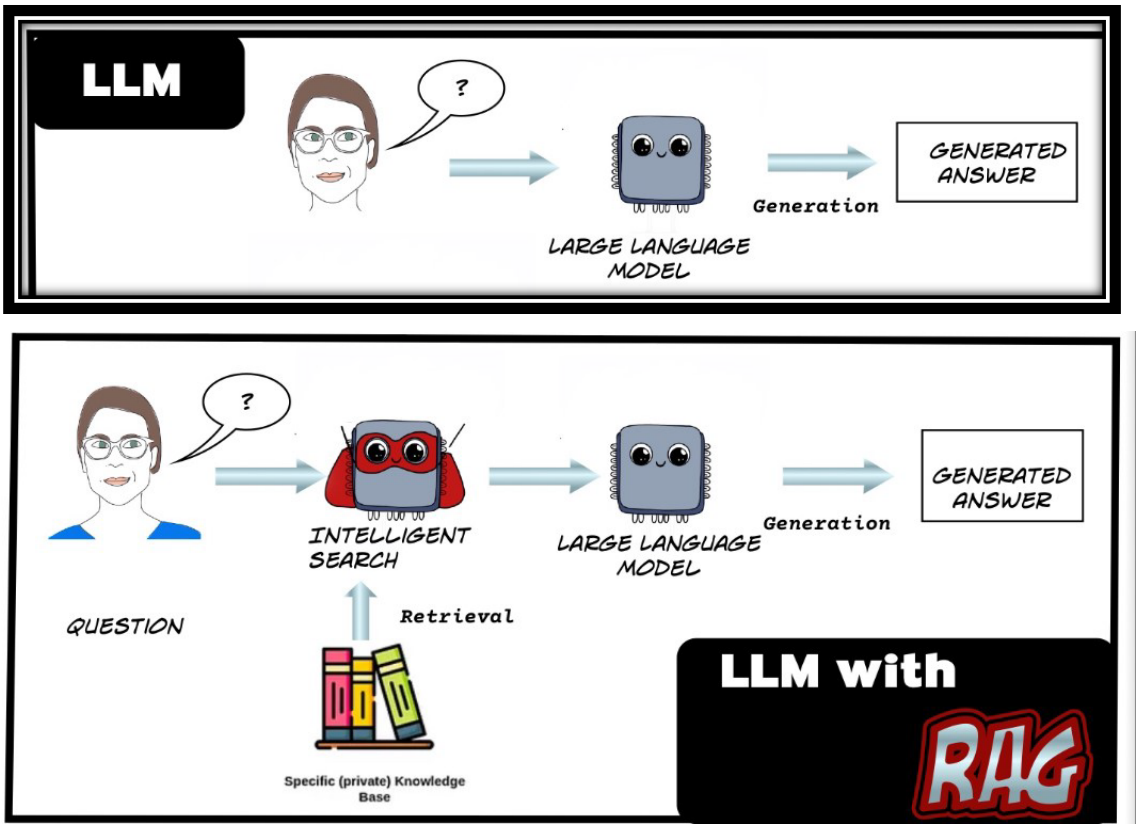


Figure 13 LLM vs LLM with RAG schematic depiction. Adopted from: neo4j

There are challenges associated with this technique. Because of the retrieval component (which may work in multiple steps) it might require substantial computational resources (though still less than retraining models), hardware and expertise for deployment. There may be challenges in integrating with existing IT infrastructures, or issues related to data privacy and security.

3.3.1.1 RAG vs fine-tuning

Fine-tuning is an approach to improve the performance of LLMs that predate RAG. The two technologies work well together. It is a supervised training phase, meaning that the model is provided with question-answer pairs. The objective is to either increase its knowledge base or to improve the performance of specific tasks (e.g. language translation).

Such fine tuning can be particularly effective to improve the performance of smaller models. There are new methodological approaches that allow such fine tuning to be done without revising all the parameters of the model (which is very resource intensive), these can include freezing the weights – so they are not modified during the training process – or injecting simplified matrixes to reduce the calculations (LoRA and qLoRA).

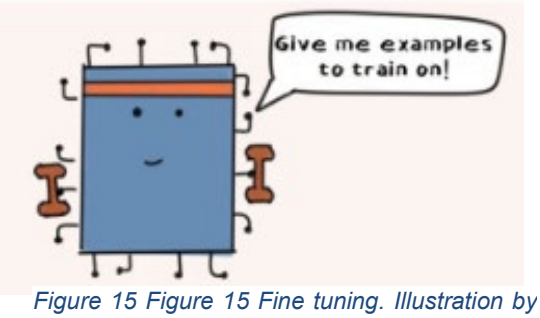


Figure 15 Figure 15 Fine tuning. Illustration by author

Nonetheless, there is a risk of catastrophic forgetting (catastrophic interference), where a neural network forgets previously learned information upon learning new information during fine-tuning. This occurs because the new learning updates the network's weights in a way that can overwrite or disrupt the previously acquired knowledge. In the context of fine-tuning, this can be particularly problematic when a model adapted for one task loses its effectiveness on another task it was previously trained for.

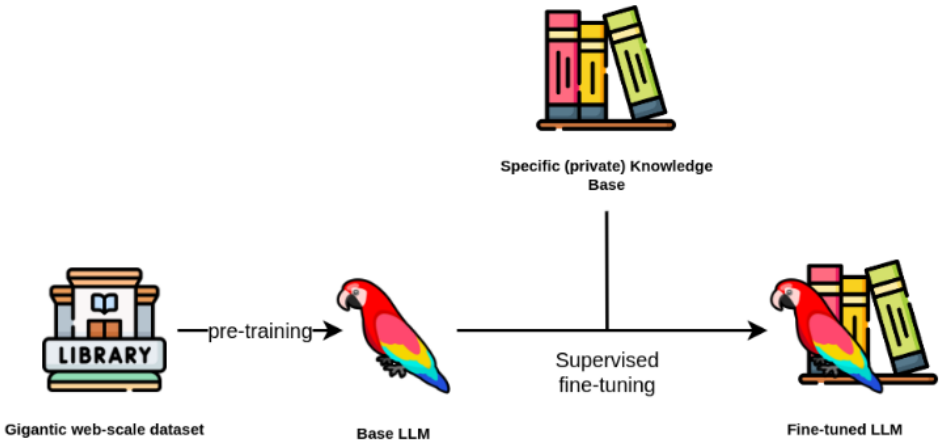


Figure 15 Process of fine-tuning. Remember the parrots? Source: neo4j

In short, both RAG and fine-tuning have merits and challenges. The following table compares the two techniques.

Aspect	Retrieval-Augmented Generation (RAG)	Fine-Tuning
Hallucinations	Reduced, assuming relevant documents are available.	Not solved; models can still generate hallucinations.
Knowledge about Tail Entities	Solved; can retrieve information about rare entities from external sources.	Limited by the training data; can be solved by dedicated prompting techniques
Outdated Knowledge	Solved; can access up-to-date information from external databases.	Requires retraining with new data, which can be expensive.
Attribution to Sources	Possible; can provide access or path to the sources of	Not solved; lacks transparency and source attribution.

	information.	
Use with Proprietary/Sensitive Data	Can work with smaller models deployed locally, but requires significant hardware and expertise; requires specific solutions for data protection.	Can be tailored to specific datasets, but still requires significant resources and expertise; data protection remains a concern.
Cost and Resource Requirements	Requires significant hardware and expertise for deployment and maintenance.	Can be expensive due to the need for extensive retraining and computational resources.
Environmental Concerns	More efficient as it can use smaller models and retrieve only necessary information.	Inefficient; retraining large models is resource-intensive and environmentally costly.
Effectiveness with Long Contexts	Effective; can retrieve relevant documents dynamically through the search function, reducing the need for long context windows.	Ineffective; feeding large collections of documents is prohibitively expensive and less efficient. Struggles with picking up information in the middle of prompts.

Table 1 RAG vs Fine-tuning

There are advantages and challenges with both methods; however, for our use case, to improve the quality and traceability of information in generated responses, RAG (or RAG in combination with fine-tuning) is more adequate.

3.3.2 Agentic workflows

Agentic workflows (also called ensemble methods or multiple LLM-agent collaboration frameworks) are one of the latest developments to enhance the performance and reliability of large language models. It introduces reflection and reasoning through the collaboration of multiple models (agents). Each agent is tasked with specific roles, to collaboratively solve complex problems.

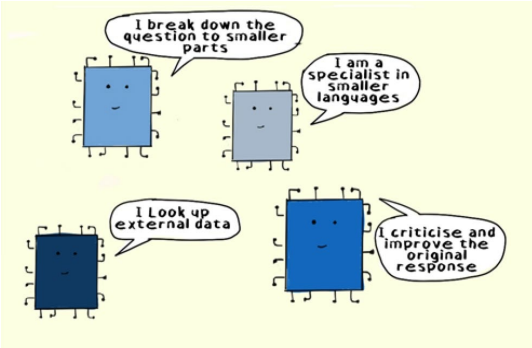


Figure 16 LLM agents. Illustration by author

The concept of agentic workflows is explored in the paper "More Agents Is All You Need" by Junyou Li and colleagues (2024). This research demonstrates that the performance of LLMs can be significantly enhanced by simply increasing the number of agents involved in the process. Using a sampling-and-voting method, the study shows that the collective intelligence of multiple agents can lead to superior results, especially in knowledge-intensive tasks. The findings suggest that this method can be applied in combination with other techniques.

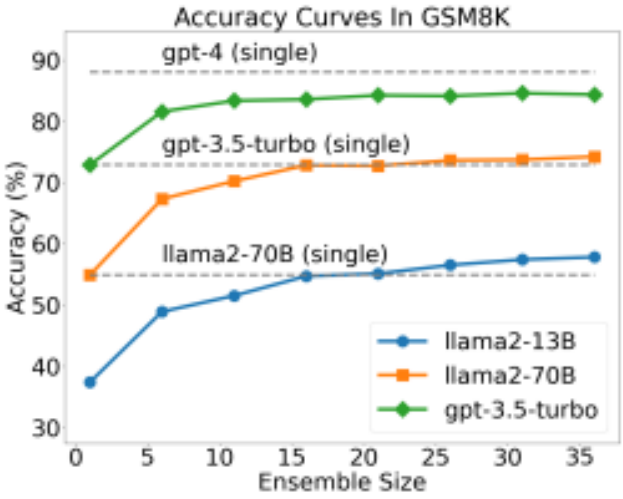


Figure 17 Performance of models with agentic workflows. Note that using agentic workflow around models like GPT-3.5 can yield better results than even more advanced models like GPT-4. Source Li et al 2024.

This iterative process involves agents performing tasks such as drafting, revising, and conducting Web research, which collectively lead to more accurate and refined outputs. There is a variety of methods for orchestrating agents. The most basic one originates in a prompt engineering technique called chain of thought. In an agentic workflow, the process of breaking down complex tasks into simpler, sequential steps is initiated and orchestrated by multiple agents. An advanced version is "tree of thoughts," where multiple thoughts are created, re-evaluated, and consolidated to provide an output.

Case_study_5: Agents in a RAG pipeline

Agents could take on multiple tasks in a RAG pipeline to improve the quality of the generated text. Here are some potential roles and responsibilities for LLM agents in such a setup:

- **Retrieval Execution**: Most importantly, an LLM agent could execute the retrieval by querying the knowledge base(s), filtering and ranking the retrieved information based on relevance and quality.
- **Query Analysis or Reformulation**: Even before, the retrieval, an LLM agent could analyze the user's query to understand the intent, context, and key information needed to answer it effectively. Based on the analysis, the LLM agent could reformulate the query into a more precise and effective form, suitable for retrieving relevant information from the knowledge base.
- **Information Integration**: After retrieving relevant information, an LLM agent could integrate and synthesize the retrieved information, resolving conflicts, filling gaps, and organizing the information in a coherent manner.
- **Answer Generation**: Based on the integrated information, the LLM agent could generate a final answer to the user's query, ensuring that it is accurate, comprehensive, and well-structured.
- **Textual improvements:** LLM agents could be tasked with improving a specific aspect of the response. Such aspects could be the editorial style ensuring that the text follows the

expected conventions (e.g. for dates or numbers) or verify references (e.g. the correct reference to legislations)

- **Quality Assurance Agent**: LLM agents can be employed to evaluate the output against predefined quality metrics such as faithfulness, relevance, and factual accuracy. This role includes detecting and mitigating biases, inconsistencies, and other potential issues in the generated responses.

It is intuitive to understand why this architecture works well. When humans draft a good piece of writing, it is always an iterative process of reasoning, revising, and improving. In contrast, a language model that simply predicts the next word based on the previous context is more analogous to writing an entire essay word after word, without the opportunity for reflection and revision. The back-and-forth process of drafting and refining allows both humans and AI models to incrementally improve the coherence, clarity, and overall quality of the writing.

While hypes and trends come and go very quickly in the AI world, agentic design has steadily captured the attention for months with its potential to harness the power of multiple specialized agents working in concert, reasoning and paving the way for more robust and trustworthy AI applications in the future. The drawback of these approaches is related to the fact that multiple models are queried multiple times, resulting in higher cost and latency (slowness) of the system.

4 Technical Solutions

4.1 Technical experiment - Bridging data and AI domains

The **objective** of this experiment is to explore optimal methods for processing documents to be used in a machine learning (ML) pipeline with Retrieval Augmented Generation to enrich the content and improve the reliability of the generated answers.

The experiment follows a standard **methodology** for data projects: the CRISP-DM (Cross-Industry Standard Process for Data Mining).

This methodology is an open-standard model, developed in the EU, and widely used. It provides a structured approach to planning and executing data analysis tasks. This iterative process helps organizations systematically tackle complex data mining challenges, ensuring that projects remain aligned with business objectives while maintaining a focus on producing reliable and actionable results

The experiment will consist of 6 steps in line with the CRISP-DM methods:

1. Business Understanding
2. Data Understanding,
3. Data Preparation,
4. Modeling,
5. Evaluation, and
6. Deployment.

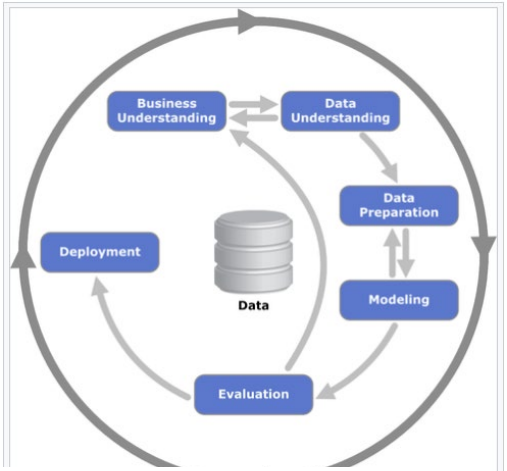


Figure 17 Process diagram showing the relationship between the different phases of CRISP-DM. Source: Wikipedia

The experience aims to enrich the data by using metadata related to the file itself (e.g. title, author, security level, etc.) but also related to paragraphs (e.g. name of the heading) to enrich the text. This study utilizes publicly accessible datasets and open-source tools to ensure reproducibility and accessibility. The tools used – Knime and Python scripts – are also available for public administrations through the Big Data Test Infrastructure (BDTI).

4.1.1 Business Understanding

The experiment aims to address the critical need for trustworthy AI services in the European Commission, specifically, the challenge of integrating large language models into policy-making processes while ensuring reliability and transparency. The objectives include developing a system that can efficiently process and manage (possibly sensitive and frequently updated) documents, maintain data security, and provide up-to-date information for RAG pipelines.

The documents in the experience are in the domain of substances of human origin (SOHO).[10] SoHO plays a vital role in modern medicine but its availability depends on voluntary donations; there is a European legislation and a strong collaboration of sharing expertise and information. The author's familiarity with the documents and the domain allowed a quick assessment of the outcomes of the experiment.

4.1.2 Data understanding

The following documents, all published on ec.europa.eu, serve as the primary data sources for subsequent processing. These documents were selected for their status as typical, key reference materials for EU policies, characterized by their high quality, precise formatting, and well-structured content. These are all public documents available to all. The corpus includes:

- Proposal for a **Regulation** on standards of quality and safety for substances of human origin intended for human application and repealing Directives 2002/98/EC and 2004/23/EC
- SOHO **Evaluation** preparing the above regulation
- SOHO **Impact Assessment** supporting the above regulation (parts 1-3)
- Clinical Trials FAQ
- Test document from Sharepoint

The revision of a **regulation** is a significant event in any policy domain, typically occurring only once every 10 to 20 years. This process is preceded by an in-depth **evaluation** of the sector and the existing regulation, followed by a comprehensive **impact assessment**. The impact assessment estimates the costs and benefits of the proposed regulation, providing a crucial evidence base for decision-making. This selection of documents offers a comprehensive overview of the regulatory process, from initial evaluation to final proposal.

4.1.3 Data preparation

The Data Preparation phase in the CRISP-DM framework is crucial for ensuring that the data is in the optimal condition for analysis.

In this phase, the open source softwares were used.

[10] Substances of Human Origin (SoHO) refer to materials derived from the human body that are used for medical applications, such as blood, plasma, bone marrow, cornea, etc.

The tools/softwares selected are as follows: Knime, Python, Google colab, and DCAT ML.

- Knime allows users to create data workflows with a visual interface, making it accessible for users with varying levels of technical expertise. Its robust capabilities in handling large datasets, integration with various data sources make it an ideal tool for the European Commission to streamline data-driven decision-making processes and improve operational efficiency.
- Python scripts were also used for pre-processing. Python was selected partly because of its open-source nature. Equally important are Python's extensive libraries and frameworks for machine learning, making it the most widely used programming language in AI applications.
- Google Colab was used to run the scripts due to its ease of access and open-source nature, which facilitates quick sharing and collaboration. Although Google Colab is not recommended for internal applications for security reasons, the Python scripts developed in Google Colab can be effortlessly extracted and deployed within internal systems.
- DCAT-AP (Data Catalog Vocabulary Application Profile) was chosen for metadata management due to its status as an EU standard used for describing public sector datasets; including the European Data Portal (now part of data.europa.eu) and the European Commission's internal Data Catalogue. This standardized approach facilitates interoperability, enhances data discoverability, and provides essential information regarding access rights and security levels, making it particularly suitable for managing datasets within the complex regulatory landscape of the EU. There is a specialised set of vocabulary for machine learning projects, MLDCAT AP. MLDCAT-AP (Machine Learning Data Catalog Application Profile) is an extension of DCAT-AP (Data Catalog Application Profile) specifically designed for describing machine learning datasets in the European public sector. While DCAT-AP provides a standard specification for describing public sector datasets in general, MLDCAT-AP includes additional categories and properties tailored to machine learning datasets. While the current experience works with a limited number of fields, many properties could be relevant for the deployment of a similar pipeline.

In this experience, due to the very limited number and high quality of documents there was no need for **data cleaning**. It should be noted though that in many applications this can be a work-intensive phase.

In the **data construction and formatting phase**, the selected documents were processed to create JSON files with content and metadata from the documents.

Steps:

1. Selected documents were parsed (read) from a local folder.
2. The metadata was aligned with the MLDCAT AP description, using the following elements (cf table1).

3. A JSON[11] file was created with the content and the metadata of the document

The workflow can be found under Annex 1.

Table 1 Selected metadata for corpus from MLDCAT AP

Property	Definition	Comment
title	A name given to the resource	The file name was not picked up automatically. It was extracted from the file path.
creator	This property refers to the entity responsible for producing the dataset.	The creator was picked automatically in 3 of the 7 test documents.
collectionDate	The date the data was originally collected, given by the uploader.	
Language		Language was picked automatically 6 times from the 7 test documents, but only 3 were correct.
access rights	This property refers to information that indicates whether the Dataset is open data, has access restrictions or is not public. The recommended controlled vocabulary is the Access Rights Named Authority List.	All documents are in the category 'public'
description	A free-text account of the resource.	Added 'legal text' and 'question and answer' to the relevant documents
dataset distribution		We used the file path. Could be replaced by the webpage (URL) Included correctly for all 7 documents.

[11] JSON (JavaScript Object Notation) is a lightweight, human-readable data format that can represent structured information, tables, or just text. JSON is widely used to exchange data on the internet and is particularly beneficial for machine learning because it can easily handle complex data structures.

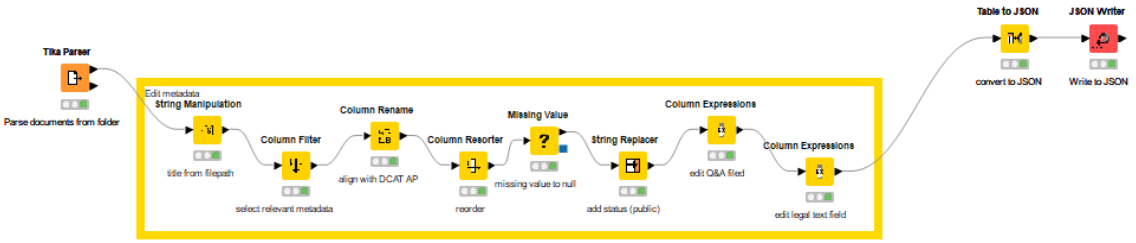


Figure 18 Knime workflow

The code in Annex 2 takes a document (SOHO Impact Assessment) and creates multiple files – each one containing one chapter. These files can be run through the above KNIME workflow to generate a JSON file in which the text of the chapters is stored as separate objects along with the title of the heading. This process ensures that the machine has accessible information on what text belongs under which heading (e.g., problems, objectives, etc.). This approach is called semantic chunking; it should be noted that most chapters are longer than the maximum length and will be split further.

4.1.4 Modelling: RAG pipeline

In the experiment, we have constructed a RAG pipeline (cf Annex 3) to demonstrate a process called Retrieval-Augmented Generation (RAG). It takes the JSON files containing the content and metadata of selected files, processes this data, and creates a system that can answer questions about these topics.

Selection of tools and libraries:

The selection of tools is based on a tutorial authored (Roucher, n.d.) modified for the AI4GOV Workshop (2024 May).

1. Pandas: Pandas is a powerful data manipulation library for Python. It is commonly used for handling structured data and is useful for managing the JSON files and dataframes created during the data preparation phase.

2. FAISS (Facebook AI Similarity Search): FAISS is used to create a FAISS-based knowledge vector database for efficient similarity search. This tool is particularly well-suited for the task as it is fast and widely used.

3. HuggingFace Embeddings: The script uses HuggingFaceEmbeddings for converting text into numerical format (embeddings). HuggingFace provides state-of-the-art pre-trained models for various NLP tasks, including text embedding, making it a reliable choice for this step. The embeddings are normalized for cosine similarity, which is important for the subsequent similarity search.

4. LangChain is an open-source framework designed to simplify the development of applications powered by large language models (LLMs), LangChain is a key component in the RAG pipeline implementation; a number of features that could be used for further developments in the experience.

The pipeline begins by loading and preparing the data. We turn the dataset into a Langchain object, keeping its metadata.

```
metadata={
    "AccessRights": doc.get("AccessRights"),
    "Creator": doc.get("Creator"),
    "Language": doc.get("Language"),
    "Title": doc.get("Title"),
    "collectionDate": doc.get("collectionDate"),
    "Filepath": doc.get("Filepath"),
```

The embedding process uses the model from HuggingFace to generate chunks from the text. We used Langchain's tool RecursiveCharacterTextSplitter with a chunk size of 3000 and an overlap of 100 characters. We have also used MarkdownTextSplitter to split the text along paragraphs. Given that the documents are well-structured, we anticipate that each paragraph encapsulates a distinct idea. By using paragraphs as the basis for chunking, we leverage the intentional organization of the original text, resulting in more semantically meaningful chunks. As a next step, we adjust the length of the chunk to 512 tokens, the maximum length for the embedding model. The embeddings are normalized to optimize for cosine similarity in the vector search.

The FAISS vector database is configured to use cosine similarity as its distance metric (to compare the angle between vectors – direction of 'arrows' in the multidimensional space). Cosine similarity was chosen because it effectively measures semantic similarity between text embeddings (as it uses vector directions, it can handle well embeddings that have different length), works well with normalized vectors, and is computationally efficient.

LangChain's configuration includes setting up a ChatOpenAI model (GPT-3.5-turbo) for generating responses. LangChain offers additional features that could be used to further develop the experiment, agents for more complex, multi-step reasoning tasks, or tools for integrating external data sources or APIs.

The implementation also includes a reranking step using the RAGatouille library, which can improve the relevance of retrieved documents before they are passed to the language model.

This question-answering system, when given a question, searches through the processed data (embeddings) to find relevant information and uses the custom prompt and an AI language model (GPT 3.5) to generate a human-like response based on that information.

The code modifies an existing pipeline[12].There are two versions: Annex 2a is the plain version loading entire documents in a JSON format[13]; Annex 2b is adapted to using an input file, where the JSON structure is built around the chapters, and information (metadata) on the individual chapters is recorded and used. Further details related to the technical approach taken in the script is clearly explained in the annotations.

4.1.5 Evaluation

For the evaluation of the pipeline, the following approach is suggested:

1. Define a set of four test questions in the domain that varies in complexity.
2. Set up different configurations:
    a. No RAG (base LLM response)
    b. RAG without semantic chunking (vanilla)
    c. RAG with semantic chunking
3. Run each question through each configuration and record the responses.
4. Evaluate the responses based on the following criteria:

- Faithfulness: the alignment of the generated responses with the retrieved data. (Method: value provided by RAGAS in the script)
- Answer Relevance: how well the generated answer addresses the initial question. (Method: Human assessment on a scale of 1-5[14])
- Context Relevance: how relevant the context provided to the language model is for answering the question. (Method: Human assessment on a scale of 1-5)

This framework addresses the challenge of assessing the performance and quality of RAG systems, which rely on the combination of two distinct components, retrieval mechanisms and language generation modules.

[12] written by Aymeric Roucher](https://huggingface.co/m-ric); modified by Steven Schockaert & Luis Espinosa-Anke for the AI4GOV Workshop in Milan 9th May 2024
[13] Such plain versions are often called vanilla.
[14] Given the expected high standards, it is recommended include evaluation based on human feedback for such RAG systems. While this approach is resource-intensive, requiring significant time and effort from human evaluators, it offers better insights into the performance and usability of the system. Human feedback helps to identify subtle inaccuracies that automated metrics might overlook, thereby enhancing the overall trustworthiness of the service. To mitigate the high human resource demands, innovative strategies like crowdsourcing can be employed.

Table 2 Testing the RAG pipeline

	No RAG	RAG vanilla	RAGwith semantic chunking +cust prompt
Method	*Simple query to the LLM (GPT 3.5) without further context.*	*Query to the LLM (GPT 3.5) using all 8 documents in the corpus in the 'vanilla' pipeline.*	*Query to the LLM (GPT 3.5) using the pipeline where the processed Impact Assessment was introduced (chopped based on chapters; chapter title included in metadata), with a custom prompt (see above)*
Primary findings	The model is not responding based on domain-specific information. In one instance, it responded assuming that SOHO is Small Office/Home Office.	Domain specific information was used for generating the answer, which dramatically improved its relevance. The reranking substantially increased the ranking of two key parts of the text.	Domain specific information was used for generating the answer, which dramatically improved its relevance. The reranking substantially increased the ranking of two key parts of the text. Compared to the vanilla pipeline, the responses are more precise and closer to the original source documents E.g. for a question on the objectives of the regulation, it did not only accurately respond with the objectives, but it distinguished between general and specific objectives as in the source document.
Faithfulness (0-1)	NA	0.75	0.98
Answer relevance (0-5)	0.67	4.25	4.67
Context relevance (0-5)	NA	4.67	4.76

The tests during the experimentation confirmed that the pipeline works as expected. While very limited, tests indicated that RAG strongly outperforms the no RAG version. Moreover, the RAG using semantic chunking leads to better quality (more comprehensive) responses. Higher scores were observed for faithfulness (the alignment of the generated responses with the retrieved data) and answer relevance (how well the generated answer addresses the initial question) (see Annex 4).

4.1.6 Further work

In the future, further work can be done to **map the structure of documents**. While such mapping is not feasible for all different documents in the same pipeline, it could be beneficial for high-quality, comprehensive reference documents (e.g., legal texts, impact assessments, evaluations, legal interpretations) or multiple documents in the same format (e.g., briefings).

Semantic chunking can also be applied to documents containing question-and-answer pairs related to interpreting legislation. By using this approach, each Q&A pair can be transformed into a JSON object that includes the text and relevant metadata, such as an ID number, date of publication (to account for Q&A pairs published on different dates), or reference to the related provisions in the legal text. This method enables targeted searches for Q&A pairs related to specific provisions or those published after a certain date.

The **metadata** presented in table 1 could be extended as necessary (e.g. ID number of Q&A pairs, dates of updates, etc.)

Adjusting the **prompt** is also expected to improve the accuracy of the answers. Prompting frameworks, such as COAST (Context, Objective, Actions, Scenario, Task) or RTF (Role, Task, Format) specifies the framework for executing the task.

Example of a prompt:

Context:

You are a highly skilled policy officer at the European Commission. You always write using the spelling conventions of British English. You value correct, formal writing, but you take into account that your readers are usually non-native English speakers. Your writing is concise but provides a pleasurable reading.

Task:

Generate a detailed analysis of [...] Your analysis should cover the following aspects: [...]

Instructions:

Use relevant data and examples to support your analysis. If available, use the heading names in the metadata to identify the most relevant text.

Ensure your analysis is comprehensive and covers all the specified aspects.

Use bullets and formatting to structure your response. Source: EC Prompt Library.

Another exciting possibility with high potentials is connecting a the LLM to a **knowledge graph**. Knowledge graphs represent data as nodes (entities) and edges (relationships), which allows for efficient navigation and retrieval of interconnected information. A great advantage of knowledge graphs is that they combine seamlessly structured and unstructured data into one resource.

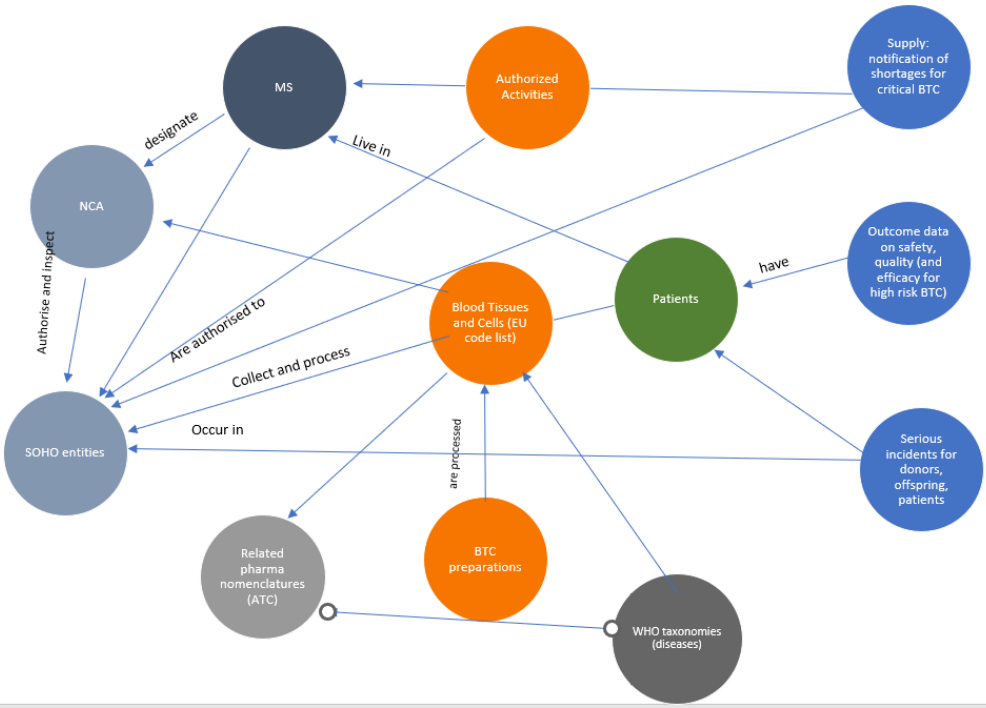


Figure 19 Knowledge Graph of the SOHO domain

By integrating an LLM with a knowledge graph, the LLM can extract relevant sub-graphs that provide contextually rich data to answer complex queries. This process could use LangChain's `LLMGraphTransformer`. The knowledge graph can then be queried to retrieve relevant triples, which are used to generate responses to user queries. To improve the query, an LLM agent could translate the original prompt into a query for the knowledge graph using a specialised language, such as Cypher. (Neo4J, 2024). Existing knowledge graphs in the SOHO domain or the boarder domain of health could be leveraged.

Agents could also be employed to do specific tasks in the pipeline (cf. Case study 5). One could even imagine a concept where users can create their own RAG agent(s) that query a specific knowledge base. This approach allows for multiple agents with specific access rights.

Evaluation of the approach should be completed. The preliminary test results in Annex 4 prove that the system works. A well designed comprehensive evaluation would be needed to assess how well it works; however, this is out of scope of this thesis.

5 People

Despite the exciting possibilities and new horizons in technology, it is essential to put the most important element at the centre of all efforts: people. In discussions about the implementation of Artificial Intelligence (AI) in the public sector, human-centric AI and the upskilling and training of staff are central components. This chapter presents an approach to foster open and constructive discussions on the topic and provides training materials specifically designed for the European Commission on generative AI.

5.1 Participatory engagement and co-creation

As we work on (generative) AI-based services, it is crucial to ensure that they are co-designed with the users and deployed in an ethical, trustworthy, and user-centric manner. The participatory approach, which actively involves diverse stakeholders throughout the AI development lifecycle, offers a framework and provides practical tools for addressing these concerns.

AI development teams often **lack representation** from diverse backgrounds, including gender, race, ethnicity, and culture.[15] This homogeneity creates a high risk for overlooking the needs, perspectives, and experiences of underrepresented groups, leading to biased AI systems that may discriminate against or marginalize certain communities. Adopting a participatory approach can help mitigate these issues by bringing together diverse perspectives and fostering a shared understanding as creating safe spaces and fostering non-violent communication allows for the expression of concerns and for all voices to be heard – even the non-mainstream perspectives. The concept of **collective intelligence**, often referred to as the "wisdom of crowds," has been shown to outperform expert opinions, especially if participants are diverse (Surowiecki, 2004). By integrating this diversity, teams can identify the risks of adverse effects and bias early in the AI development process. Diverse, multidisciplinary teams can also bring different considerations from the policy and technology domains, including legal experts, people with experience in the policy implementation, business analysts, data experts, GDPR coordinators, security experts, etc.

We live in *volatile, uncertain, complex, and ambiguous times* (VUCA) – but even by the standards of today's world, the developments of AI are unpredictable and disturbingly quick. (Bennis & Nanus, 1985). The VUCA model calls for a more agile and collaborative skill set. The **Diamond of Participation** model gives a practical map and reassurance for navigating through divergent viewpoints and uncertainties to achieve collective

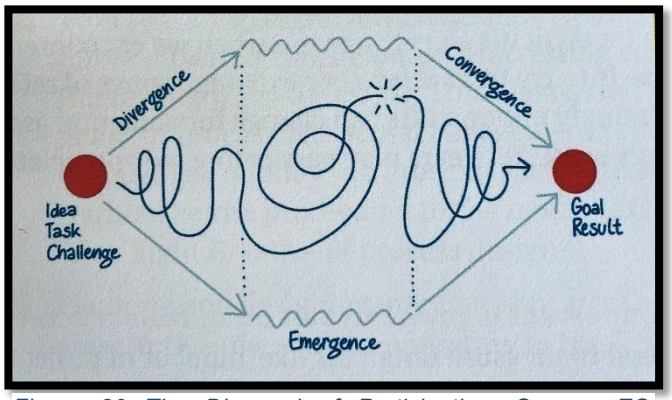


Figure 20 The Diamond of Participation. Source: EC, Participatory Leadership Practitioners' Guide.

[15] According to the AI Now Institute report, over 80% of AI professors are men. At major tech companies like Facebook, only 15% of AI researchers are women, and at Google, it's only 10%. In 2018, less than 25% of PhDs in computer science were awarded to women and minorities (Klave, 2024)

understanding and action (Kaner, 1998).

In a participatory AI development, these phases could involve (1) stakeholders in exploring diverse perspectives (divergence), (2) identifying emergent patterns and concerns (emergence), and (3) collectively shaping the AI system's design and implementation (convergence).

Storytelling, as highlighted by the **Storycatcher** concept, can be a powerful tool for fostering empathy, sharing knowledge, and communicating ethical principles in AI development. Techniques of the participatory approach, such as sitting in circles and passing around a 'talking stick' evoke age-old traditions of human storytelling. Other methods use tools, such as cards, to bring in a variety of perspectives.

Figure 21 Tarot Cards of Tech. source: Artefact Group

There are further frameworks to understand and operate in uncertainty, such as the **Cynefin** (Snowden & Boone, 2007) or the **Chaordic Age** (Hock, 1999) concepts.

There is an active community of practice around 'Participatory Leadership' and the 'Art of Hosting Conversations that Matter' within the European Commission. Specifically in the digital sector, the Digital Innovation Lab (iLab) promotes co-innovation through the rapid development of prototypes. These initiatives have already achieved significant success by bringing together diverse, multidisciplinary groups and creating the space for conversations that matter. Their engagement and the use of their established practices are essential for the effective design, development, and implementation of AI services.

5.2 Training materials on Generative AI in the EC

This chapter presents a structured approach for a 1.5-hour training session on generative AI, tailored for European Commission staff.

5.2.1 Theoretical Framework: Merrill's First Principles of Instruction

The program is built upon Merrill's (2002) First Principles of Instruction, a comprehensive instructional design model focused on creating effective, problem-centered learning experiences.

- Problem-Centered: Learning is centered around real-world problems or tasks, ensuring relevance and practical application.; the training presents a use case in a domain that participants regularly work with to ensure relevance and practicality.
- Activation: Emphasizes activating learners' existing knowledge and providing a structure for new information. The training provides a clear explanation of technical concepts
- Demonstration: Shows learners what they need to know through practical examples, tasks that they need to do every day, in their own domain of expertise. .
- Application: Provides opportunities for learners to practice and apply new knowledge.
- Integration: Encourages learners to integrate new knowledge into their everyday lives and reflect on their learning.

5.2.2 Format of the Training Programme

The training programme is composed of a formula:

Technical explanations + demonstration + discussion

1. Technical Explanations
    - Provide a foundational overview of generative AI
    - Cover capabilities, limitations, and potential applications in the public sector
    - Explain key concepts (detailed materials related to AI, LLMs, Computers, RAG, etc. available in Chapter 3 and annexes 5)

2. Demonstration
    - Collaborate with the unit or directorate to develop a relevant use case
    - Showcase how available AI tools and platforms can be used on a real-life example
    - Demonstrate best practices and potential pitfalls

- Present the European Commission guidelines for safe and ethical use of AI technologies

3. Discussion
    - Facilitate an open dialogue on concerns and shortcomings, building on the philosophy and approaches of participatory meetings
    - Address ethical, legal, and practical issues
    - Encourage participants to voice apprehensions and questions

This training program has been delivered three times in both hybrid and in-person formats, incorporating interactive elements such as Slido for real-time feedback and questions. The training materials are in Annex 5.

It is crucial that this training is delivered to teams, units, or directorates rather than to individuals. Although individual training caters to personal needs and interests, group training creates an opportunity where participants can engage in meaningful discussions about the content and collectively drive the implementation of changes. This group dynamic enhances the learning experience and creates a shared momentum for improvement.

6 Design for User Engagement

This chapter proposes a user workshop for designing two RAG services for EC staff. In the frame of this paper, describing such workshop serves as a practical demonstration of how participatory methods and design tools can be effectively applied in service design and co-creation. It also allows us to show a glimpse into how RAG services could look like.

1. Service 1, Prepare My Documents for RAG, aims to process documents for a RAG pipeline.
2. Service 2, Call My RAG Agents, presents the use of RAG agents in the generation process.

Service design plays a pivotal role in the development of AI applications, particularly in the context of European policy-making where trustworthy and user-centric AI is paramount. This chapter builds upon the technical solutions explored in Chapter 4 and the people-centric approach discussed in Chapter 5, demonstrating how design for user engagement is crucial in developing AI services that are both effective and aligned with EU values. Such a human-centric approach aligns with the principles advocated by Don Norman (2013).

This chapter provides a framework and practical tools to address some of the key challenges and limitations of Large Language Models (LLMs) mentioned in earlier chapters, such as bias and user trust. By emphasizing collaborative processes and utilizing tools like service blueprints and user journey maps, service design enables the integration of AI into existing workflows and user experiences in a way that enhances trust and mitigates potential biases.

Co-creation, a core tenet of service design, emphasizes the active involvement of customers and end-users in the value creation process. In the context of AI, co-creation enables designers to engage with non-technical users, leveraging their knowledge and experiences to shape AI-enabled services that better align with customer needs (Sanders & Stappers, 2008). This collaborative approach is particularly valuable given the unpredictability and adaptability of AI systems, allowing for continuous value co-creation and co-production during use.

As highlighted in the chapter above, participatory methods are excellent tools for engaging users and experts in the co-creation of services. AI can also be used to support the various phases of the design process (Paetzold, & Lindemann 2023). In the initial phases, AI can analyse user experiences. It can be used for creating user personas or for rapid prototyping to produce mock-ups, wireframes[16], visual design to help users imagine and discuss solutions. In the later phases it can analyse feedback and support in the drafting of specifications. It can effectively translate between domain specific language (registers with specific format and vocabulary, such as in design, in business analysis or project management) and everyday language.

[16] Wireframes are like rough sketches or blueprints for websites or apps. They show the basic layout and structure of a page without any fancy colors or graphics.

6.1.1 Preliminary description of the RAG Services

Service 1: Prepare My Documents for RAG

This service aims to process documents for use in a Retrieval Augmented Generation (RAG) pipeline. It involves preparing and structuring documents to be effectively used in AI-powered information retrieval and generation systems. The service likely includes tasks such as:

- Document upload: I could upload various types of documents (PDFs, Word files, etc.) that I want to use as a knowledge base.
- Automatic metadata extraction: The service would automatically extract important metadata like title, author, date, and security classification from my documents.
- Content chunking: The service would intelligently break down my documents into smaller, meaningful chunks that are optimized for RAG systems.
- Semantic tagging: It would add semantic tags or labels to different sections of my documents, making it easier to retrieve relevant information later.
- Security and access control: I would be able to set access permissions for my documents, ensuring sensitive information is properly protected.
- Integration with RAG tools: The prepared documents would be easily exportable or directly usable in various RAG-powered applications within the European Commission.
- Version control: The service might offer version control, allowing me to update documents while maintaining a history of changes.

Service 2: Call My RAG Agents

This service focuses on the use of RAG agents in the text generation process. It likely involves deploying multiple specialized AI agents that work together to retrieve relevant information and generate responses. These agents may perform tasks such as query analysis, information retrieval, answer generation, and quality assurance within the RAG framework.

Key features of this service could include:

- Multiple specialized agents: Users can access various AI agents, each with a specific role (e.g., summarization, fact-checking, style editing).
- Customizable workflows: Users can create and save custom workflows by chaining together different agents for specific tasks.
- Interactive querying: Users can engage in a dialogue with the agents, asking follow-up questions or requesting clarifications.
- Source attribution: The service provides clear references to the source documents used in generating responses.
- Explanation of reasoning: Agents can provide step-by-step explanations of their reasoning process.
- Collaborative features: Multiple users can work together on the same project, sharing agents and workflows.

6.1.2 Workshop format and agenda

To elicit discussion and feedback on this service, we suggest to engage users through a workshop. This workshop should ideally take place in person over a minimum of 2 hours to allow time for good conversation. The format of the workshop depends on the objective. For the agenda and practical organisation, refer to Annex 6.

Method	Open Space Technology	World Café	Ritual Dissent
Objective	Open Space Technology is ideal for early phases of service design, as it allows for self-organized, flexible, and inclusive discussions where participants can propose topics of interest, form discussion groups, and move freely between them. This method is effective when there are complex issues with no predefined solutions and a need for diverse input.	The World Café method is ideal for bringing together multiple viewpoints. It encourages the sharing of diverse perspectives and the co-creation of knowledge through small group discussions. This method is particularly effective when exploring complex issues that benefit from a variety of viewpoints and collaborative thinking as well as action planning.	Ritual Dissent is a structured method designed to rigorously test ideas and assumptions through critical feedback. This method is effective when developing robust solutions to complex problems, as it allows participants to challenge and refine each other's proposals constructively. It ensures that the final framework is well thought out and resilient to various critiques.

Table 3 Participatory methods and objectives

6.1.3 Workshop content and materials

The following materials are created to elicit discussion of user needs and features of the two RAG services (Service_1, *Prepare My Documents for RAG* and Service_2, *Call my RAG Agents).* It builds on the concept of walkthrough, which elicitates feedback on a service idea at very early stages, by walking users through the process. The service card gives a simple visual overview that helps users imagine the concept in practice and engage with it. A draft list of user needs is provided using a classical format (*I want to... so that*) to launch the discussion and help users articulate their needs. Finally, low-fidelity mock-ups were developed to support the explanation and the discussions. These materials can be found in Annex 7.

Participants should be introduced to these materials and engage with them during the session to work towards a collaborative definition of user needs and features. The agenda and the materials should be adjusted according to the specific project. ***It should be noted that the materials are not a description of user needs for a RAG service, but a first draft to use for discussing and defining these needs.***

6.1.4 Service 1: Prepare My Documents for RAG

Service_1, *Prepare My Documents for RAG*, aims to process and manage documents for a RAG pipeline efficiently and securely.

The users are EC staff that would like to use their own documents or another knowledge base in a retrieval augmented generation service.

6.1.5 Service 2: Call my Agents

Service_2, *Call my RAG Agents* integrates RAG for advanced querying based on documents or datasets. It provides access to multiple agents based on access rights and ensures secure information management.

The users are EC staff that would like to use their own documents or another knowledge base in a retrieval augmented generation service.

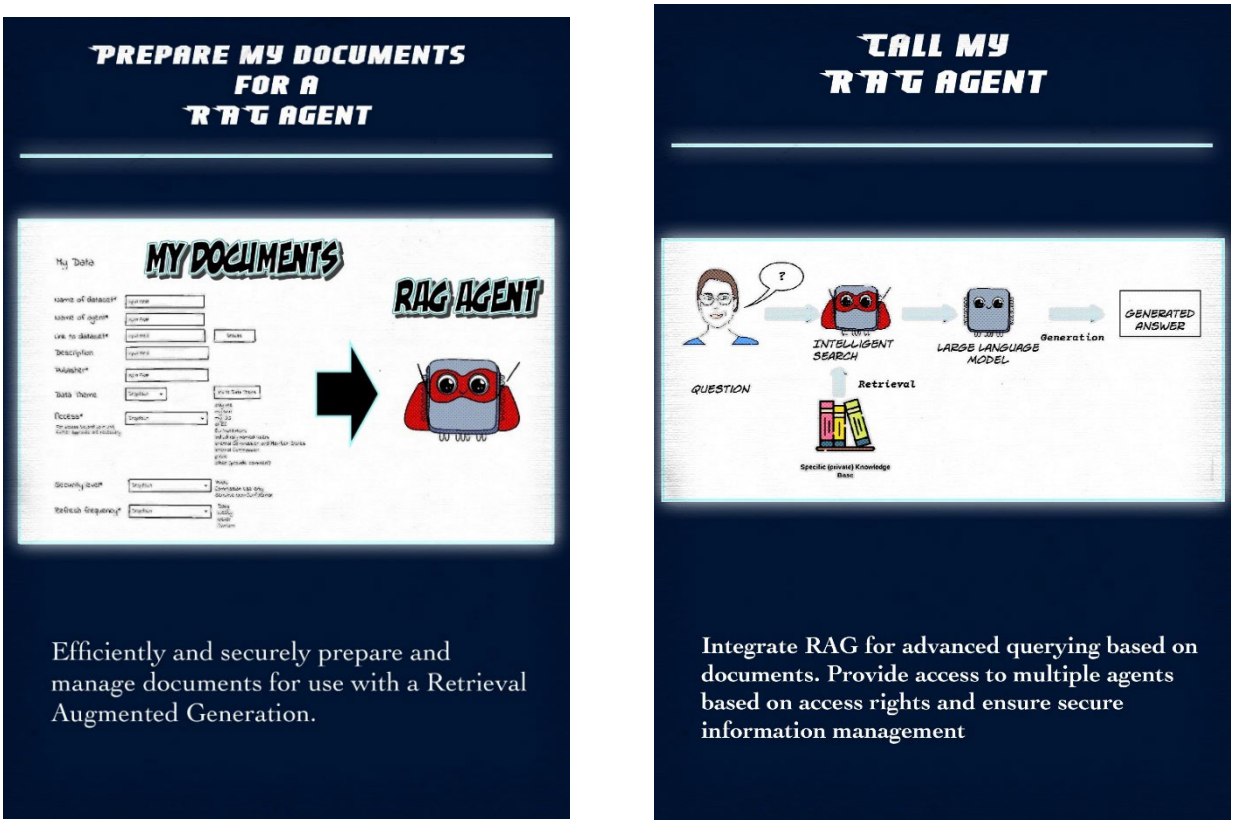


Figure 22 Service Cards. inspired by oblo (for discussion)

6.1.6 After the Workshop

The harvested insights into user needs should shape the products or services so that they align with the needs, preferences, and contexts of the intended users. This approach was defined as *semanticization* by Luciano Floridi (2018), which emphasizes the importance of capturing and leveraging the meaning and significance. Through user input, designers can gain insights into the semantic layers that users ascribe to their experiences, behaviours, and interactions with existing systems or products (Kolko, 2010). This semantic information can then be analysed and interpreted through the lens of systems thinking, which recognizes the interconnectedness and interdependencies within complex systems.

By adopting a system thinking approach, designers can explore how user input and semantic information fit into the broader ecosystem, considering the interplay between various components, stakeholders, and environmental factors (Sanders & Stappers, 2008). In the case of LLMs supported by RAG in the EC, one should consider EU policies and guidelines on data and AI, the existing IT infrastructure and architecture, as well as current administrative policies, processes, and practices.

7 Conclusion: Actionable insights for the Integration of Large Language Models in European Policy-Making

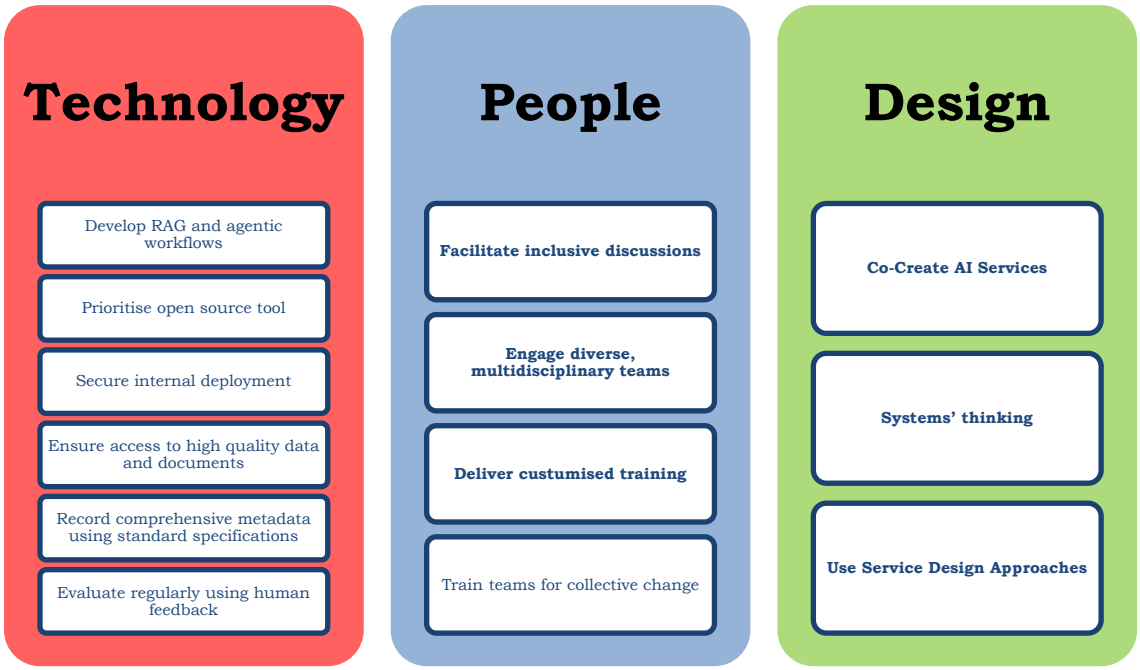


Figure 23 Actionable insights for the Integration of Large Language Models in European Policy-Making

This thesis aimed to explore how to build trustworthy AI services in the European Commission. The following are the proposed practical approaches in three domains:

Technical Domain:

- **Implement Retrieval-Augmented Generation (RAG) and Agentic Workflows:** Enhance the accuracy and reliability of LLM outputs by rooting generated text in reliable, up-to-date documents and datasets. RAG improves transparency and accountability by creating direct links to the data used for generating responses. Being able to verify the precise reuse of source documents in the generated content is one of the key elements of trustworthy LLMs. Agentic workflows bring in reasoning, allow for breaking down complex tasks into smaller, manageable ones and improve the quality of the text.
- **Prioritize Open-Source Tools:** Utilize tools like Knime, which is accessible for all European Commission users and available to the public sector through BDTI, for pre-processing documents. Python scripts are flexible tools to develop pipelines and allow access to open resources, such as Langchain for orchestrating processes and resources. Using open-source tools contributes to

a thriving tech ecosystem in the EU and reduces dependency on external service providers, controlling costs associated with extensive token processing.

- **Secure Internal Models:** Adapt existing models to operate securely on internal servers without external communication to handle Commission Use and Sensitive Non-Confidential (SNC) information. This approach, already used for eBriefing or GPT@JRC mitigates data breach risks, keeps sensitive data within the EU's digital ecosystem and strengthens the strategic sovereignty of the AI ecosystem used in the EC.
- **Data Focus:** Ensure LLMs access high-quality, relevant data specific to European policy domains, either through RAG or fine-tuning. Respect privacy, transparency, and fairness in training data and AI development.
- **Metadata** Use metadata standards like DCAT AP to describe key features of the data, including access rights and security levels, and ensure generated text inherits these metadata properties.
- **Establish Clear Benchmarks for Evaluation:** Develop clear evaluation benchmarks and incorporate human assessments for continuous improvement. Experiment with using LLMs for evaluations.

People:

- **Facilitate Inclusive Discussions:** Implement techniques such as Open Space Technology, World Café, and Ritual Dissent to promote collaborative discussions.
- **Engage diverse, multidisciplinary teams:** Use participatory methods and co-creation workshops to bring together a diverse, multidisciplinary team. Integrating experts from both policy and technology domains is essential to ensure that the systems are user-centric, feasible, secure, and compliant with legal requirements.
- **Customized Training:** Provide training materials and upskilling opportunities on generative AI tailored to the needs of European Commission staff. Demonstrate on domain specific materials and use cases.
- **Train teams for collective change**: Organise joint training sessions for teams and users to facilitate the adoption of new practices.

Service Design:

- **Co-Create AI Services:** Develop AI services in collaboration with end-users to ensure solutions are user-friendly and meet policy-makers' needs. This aligns with human-centric AI principles, emphasizing the importance of aligning AI tools with administrative processes and user needs.
- **Systems' thinking:** AI systems should serve as a **tool to support policy work** – and not replace humans doing it. They should complement and support existing policy and administrative process. Agentic workflows can bring in

reasoning to AI service. Keeping humans in the loop must be essential element in the design of any AI tool supporting public administration.

- **Use Service Design Approaches:** Integrate user needs and preferences into AI services, making them practical and effective. Employ design tools to visualize services in practice, experiment with small prototypes, and iterate based on feedback. Don't be afraid to fail—learn, improve, and try again.

Integrating Large Language Models into European policy-making offers significant potential for enhancing decision-making processes. However, ensuring the effectiveness, trustworthiness, and inclusivity of these systems requires a multifaceted approach. By adhering to robust data management practices, fostering participatory development processes, and aligning AI systems with European values and regulations, the EU can harness the power of AI while maintaining trust and reliability. A holistic approach towards people, technology, and design ensures that the AI solutions are contextually relevant, meaningful, and sustainable, ultimately enhancing the user experience and fostering long-term adoption and satisfaction.

8 Bibliography

3blue1brown. (2024, April). But what is a GPT? Visual intro to transformers [Video]. *YouTube*. https://www.youtube.com/watch?v=wjZofJX0v4M&t=285s

Amazon Web Services. (n.d.). What is a GPU? - Graphics Processing Unit explained. Retrieved from https://aws.amazon.com/what-is/gpu/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3442188.3445922

Bennis, W. G., & Nanus, B. (1985). *Leaders: The strategies for taking charge*. Harper & Row.

Bratanič, T. (2023, June 6). Knowledge graphs & LLMs: Fine-tuning vs. retrieval-augmented generation. *Neo4j Developer Blog*. https://neo4j.com/developer-blog/fine-tuning-retrieval-augmented-generation/

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies.* W. W. Norton & Company.

Capital.com. (2023). NVIDIA stock 5-year forecast. *Capital.com.* https://img.capital.com/imgs/articles/1140xx/NVIDIA-Stock-5-Year-Forecast-MCT-2333-EN-2.png

Chomsky, N. (1957). *Syntactic structures.* Mouton and Company.

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review.*

D'Ignazio, C., & Klein, L. F. (2020). *Data feminism.* MIT Press.

Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press.

European Commission Decision (2015). (EU, Euratom) 2015/443 of 13 March 2015 on Security in the Commission.

European Commission. (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data.

European Commission. (2022) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down measures for a high level of public sector interoperability across the Union (Interoperable Europe Act) COM/2022/720 final.

European Commission. (2023) Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the European Health Data Space COM/2022/197 final.

European Commission. (2023) Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828.

European Commission. (2024). Artificial Intelligence in the European Commission (AI@EC): A strategic vision to foster the development and use of lawful, safe and trustworthy Artificial Intelligence systems in the European Commission (C(2024) 380 final).

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Jiang, L. (2024). LLMCarbon: Modeling the end-to-end carbon footprint of large language models. *arXiv*. https://arxiv.org/abs/2309.14393

Floridi, L. (2018a). Soft ethics and the governance of the digital. *Philosophy & Technology*, 31(1), 1-8.

Greenpeace. (2019). *Clicking clean Virginia: The dirty energy powering data center alley.* Retrieved from https://www.greenpeace.org/usa/reports/click-clean-virginia/

Greenpeace. (2021). *Clean cloud 2021: Tracking renewable energy use in China's tech industry.* Retrieved from https://www.greenpeace.org/static/planet4-eastasia-stateless/2021/04/03a3ce1a-clean-cloud-english-briefing.pdf

Greyling, C. (2023, March 16). What are realistic GPT-4 size expectations? *Medium*. https://cobusgreyling.medium.com/what-are-realistic-gpt-4-size-expectations-73f00c39b832

Hildebrandt, M. (2019). *Law for computer scientists and other folk.* Oxford University Press.

HIX.AI. (2024). GPT-4 parameters explained. https://hix.ai/hub/chatgpt/gpt-4-parameters

Investing.com. (n.d.). NVIDIA facts and statistics. Retrieved from https://www.investing.com/academy/statistics/nvidia-facts-and-statistics/

Klawe, M. (2020, July 16). Why diversity in AI is so important. *Forbes*. https://www.forbes.com/sites/mariaklawe/2020/07/16/why-diversity-in-ai-is-so-important/

Kaner, S. (1998). *Facilitator's guide to participatory decision-making*. New Society Publishers.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv.

Knight, W. (2023, April 3). The cost of training AI could soon become a major issue. *Yahoo Finance*. https://finance.yahoo.com/news/cost-training-ai-could-soon-101348308.html?guccounter=1

Lappin, S. (2023). Assessing the strengths and weaknesses of large language models. *Computational Linguistics*, 1-11.

Lambert, N., Castricato, L., von Werra, L., & Havrilla, A. (2022). Illustrating Reinforcement Learning from Human Feedback (RLHF). Hugging Face Blog.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Language Teaching Publications.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. arXiv preprint *arXiv*:2005.11401.

Li, D. (2024). More agents is all you need. *arXiv*, abs/2402.05120. https://arxiv.org/abs/2402.05120

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building artificial intelligence we can trust.* Pantheon Books.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think.* Houghton Mifflin Harcourt.

Merrill, M. D. (2002). First principles of instruction. Educational Technology Research and Development, 50(3), 43-59. https://doi.org/10.1007/BF02505024

Norman, D. A. (2013). *The design of everyday things*: Revised and expanded edition. Basic books.

Neo4j. (2024a). Knowledge Graphs & LLMs: Multi-Hop Question Answering. Neo4j Developer Blog. https://neo4j.com/developer-blog/knowledge-graphs-llms-multi-hop-question-answering/

Neo4j. (2024b). LLM Knowledge Graph Builder: From Zero to GraphRAG in Five Minutes. Neo4j Developer Blog. https://neo4j.com/developer-blog/graphrag-llm-knowledge-graph-builder

OpenAI. (2022, November 30). Introducing ChatGPT. *OpenAI.* https://openai.com/index/chatgpt/

OpenAI. (2023). ChatGPT prompt engineering for developers [Online course]. *Coursera*. Retrieved from https://www.coursera.org/learn/chatgpt-prompt-engineering-for-developers

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., & Amodei, D. (2022). Training language models to follow instructions with human feedback. *arXiv* preprint arXiv:2203.02155. https://arxiv.org/abs/2203.02155

Oblo. (n.d.). AI Opportunity Landscape Canvas. Retrieved from [https://miro.com/app/board/uXjVKMEsonk=/

Publications Office of the European Union. (2023). Access right [Name authority list]. EU Vocabularies. http://publications.europa.eu/resource/dataset/access-right

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., & Irving, G. (2021). Scaling language models: Methods, analysis & insights from training gopher. *arXiv* preprint arXiv:2112.11446.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning* (pp. 8748-8763). PMLR.

Roucher, A. (n.d.). Advanced RAG Techniques. Hugging Face Learn. Retrieved from https://huggingface.co/learn/cookbook/advanced_rag

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *Co-design*, 4(1), 5-18.

Snowden, D. J., & Boone, M. E. (2007). A leader's framework for decision making. *Harvard Business Review,* 85(11), 68-76.

Surowiecki, J. (2004). *The Wisdom of Crowds.* Doubleday.

Stanford University. (2024). AI Index 2024 Annual Report. *Stanford Institute for Human-Centered Artificial Intelligence (HAI).*

Tannen, D., & Schriffin, D. (eds) (1987). *Discourse Makers*. Cambridge University Press.

Van Dijk, T. A. (2008). *Discourse and power.* Palgrave Macmillan.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Wang, P.-H., et al. (2020). Contextual temperature for language modeling. *arXiv* preprint arXiv:2012.13575. https://arxiv.org/abs/2012.13575

Yang, X., Yang, K., Cui, T., Chen, M., & He, L. (2022). A study of text vectorization method combining topic model and transfer learning. *Processes*, 10(2), 350. https://doi.org/10.3390/pr10020350

Annex 1. KNIME Workflow for Document Pre-processing (separate file)

Annex 2. RAG Pipeline Implementation (Python Code in Google Colab) (separate file)

Annex 3. Separating PDFs by Chapters (Python Script in Google Colab) (separate file)

Annex 4. RAG Pipeline Test: Evaluation Table and Selected Examples

Evaluation table

Q	Configuration	Faithfulness (0-1)	Answer relevance (0-5)	Context relevance (0-5)	Overall assessment
		how well the generated response aligns with the retrieved data. Calculated by RAGAS	how well the generated answer addresses the initial question. Human assessment	relevance of the context provided to the language model for answering the question. Human assessment	
Q1	No RAG	NA	1	NA	Not relevant. generic answer on regulation - did not pick up the topic
	RAG vanilla	1 reranked (0.89 before reranking)	3	5	Good, basic response
	RAG with semantic chunking +custom prompt	1	5	5	Excellent response; picked up on the responses between general and specific objectives, provided precise anwer
Q2	No RAG	NA	0	NA	Not relevant. The answer did not even relate to the topic (SOHO- Small Office/Home Office)
	RAG vanilla	1	4	5	Good, basic response
	RAG with semantic chunking +custom prompt	1	5	5	Excellent. Longer, more structured answer

Q3	No RAG	NA	1	NA	The response was related to the topic, and included a correct definition. However, the number provided was incorrect
	RAG vanilla	0	5	5	Short but good response. Correct number provided, but counterintuitive value for faithfulness
	RAG with semantic chunking +custom prompt	1	5+		Excellent. Correct number provided plus additional breakdown
Q4	No RAG	NA	1	NA	Not correct
	RAG vanilla	1	5	5	Short but correct answer
	RAG with semantic chunking +custom prompt	0.92 reranked (0.0.77 before reranking)	4	5	Too broad. The answer included a number of topics related to diseases that can be transmitted by SOHO - some of the elements were too broad. Note that the relevance of the context provided was indeed much weaker before reranking, referring to the adverse effects related to SOHO instead of diseases
Average	**No RAG**		**0.67**		
	RAG vanilla	**0.67**	**4.25**	**4.67**	
	RAG with semantic chunking +custom prompt	**0.98**	**4.67**	**4.67**	

Question	No RAG	RAG vanilla	RAG with semantic chunking +custom prompt
Method	*Simple query to the LLM (GPT*	*Query to the LLM (GPT 3.5) using all*	*Query to the LLM (GPT 3.5) using the pipeline where the*

	3.5) without further context.	*8 documents in the corpus in the 'vanilla' pipeline.*	*processed Impact Assessment was introduced (chopped based on chapters; chapter title included in metadata), with a custom prompt (see above)*
Primary findings	The model is not responding based on domain-specific information. In one instance, it responded assuming that SOHO is Small Office/Home Office.	Domain specific information was used for generating the answer, which dramatically improved its relevance. The reranking substantially increased the ranking of two key parts of the text.	Domain specific information was used for generating the answer, which dramatically improved its relevance. The reranking substantially increased the ranking of two key parts of the text. Compared to the vanilla pipeline, the responses are more precise and closer to the original source documents (E.g. for objectives, it did not only accurately respond with the objectives of the regulation, but it distinguished between general and specific objectives.) In everyday use, it is preferred to answer with text that is close to the original source. The system was able to identify the most relevant parts of the documents.

What are the objectives of the regulation?	Regulations are put in place to achieve various objectives, including: 1. Protecting public health and safety 2. Ensuring fair competition in the marketplace 3. Safeguarding the environment 4. Promoting consumer rights and interests 5. Preventing fraud and abuse 6. Fostering economic growth and stability 7. Maintaining social order and public welfare These objectives help ensure that regulations serve the best interests of society as a whole.	The objectives of the regulation include ensuring safety and quality for patients treated with BTC therapies, ensuring safety and quality for BTC donors and children born from donated eggs, strengthening and harmonizing oversight practices among Member States, facilitating the development of safe and effective innovative BTC therapies, and improving the resilience of the sector to mitigate the risk of shortages. These objectives aim to ensure a high level of health protection for EU citizens and provide access to safe and effective therapies (Source: Document 1)	General objectives: - Ensure a high level of health protection for EU citizens. - Ensure access to safe and effective therapies. - Enhance the future framework to be more effective, crisis-resistant, and agile. - Identify areas for simplification and efficiency improvements in legislation. Specific objectives: 1. Ensure safety and quality for patients treated with BTC therapies. 2. Ensure safety and quality for BTC donors and children born from donated eggs, sperm, or embryos. 3. Strengthen and harmonise oversight practices among Member States. 4. Facilitate the development of safe and effective innovative BTC therapies. 5. Improve the resilience of the sector to mitigate the risk of shortages.

How is the SOHO regulation impacting people?	The Small Office/Home Office (SOHO) regulation can impact people in various ways, depending on the specific rules and requirements set by local authorities. Some potential impacts may include:\n\n1. Zoning restrictions: SOHO regulations may dictate where home offices are allowed to operate within residential areas, potentially limiting the options for individuals looking to work from home.\n\n2. Business licensing requirements: Individuals operating a home office may be required to obtain specific business licenses or permits to comply with SOHO regulations, adding administrative and financial burdens.\n\n3. Tax implications: SOHO regulations may affect tax obligations for individuals running a business from home, such as the ability to deduct home office expenses or qualify for certain tax benefits.\n\n4. Compliance with building codes: Home-based businesses may need to adhere to certain building codes or safety	People are impacted by the SOHO regulation through various measures proposed in the policy options. These measures aim to provide legal clarity, ensure patient protection, improve crisis prevention and resilience, facilitate innovation, and enhance oversight efficiency in the BTC sector. The regulation addresses challenges related to unclear classifications of substances, aims to streamline regulatory practices across the EU, and mobilizes scientific knowledge for updating guidance in the sector. The establishment of an EU-level committee for advisory purposes and a common data infrastructure are expected to positively impact stakeholders affected by the SOHO regulation.	**Analysis of the Impact of SOHO Regulation on People:** - **Lack of Legal Clarity:** - Stakeholders, including professionals and authorities, have expressed concerns about the lack of legal clarity in the regulation of Substances of Human Origin (SOHO). - The uncertainty in regulatory classification has significant consequences on safety, cost, availability, and access to therapies using SOHO. - **Borderline Challenges:** - Stakeholders have highlighted challenges related to unclear regulatory classifications and the need for more precise guidelines on which legal framework applies to different substances/products. - The complexity of determining applicable frameworks has led to inconsistencies in regulatory practices across Member States, creating barriers to cross-border exchanges. - **Patient Protection & Safety:** - The inadequacy of safety and quality requirements, outdated technical rules, and insufficient risk assessment pose risks to patients treated with BTC (Blood, Tissues, Cells). - Ensuring timely information for risk management, consistency in regulatory practices, and mobilizing relevant scientific knowledge are crucial for patient protection. - **Access and Affordability:** - Regulatory ambiguities and overlaps with existing frameworks can hinder patient access to innovative

	regulations outlined in SOHO guidelines, which could require modification		therapies and increase costs. - Unclear legal frameworks may restrict the availability of therapies and limit the affordability of treatments for patients. - **Innovation and Compliance:** - Unclear regulatory classifications and lack of legal clarity can impede innovation in the BTC sector, affecting the development and compliance of novel therapies and technologies. - Establishing a graded approach to oversight and providing legal clarity through EU-wide mechanisms can facilitate innovation and ensure compliance with regulatory requirements. - **Crisis Preparedness & Response:** - Strengthening supply monitoring, implementing measures for crisis management, and enhancing emergency plans are essential for ensuring the resilience of the sector and mitigating the risk of shortages during crises like the COVID-19 pandemic. - **Collaboration and Coordination:** - Establishing EU-level structures or committees to advise Member States on regulatory classifications and providing guidance on safety and quality requirements are essential for addressing the challenges posed by the current regulatory framework. - **Data Management & Digitalization:** - Improving data management systems, enhancing interoperability, and ensuring data quality and resilience of

			digital platforms like SOHO-X are critical for supporting evidence-based decision-making and efficient regulatory reporting. In conclusion, the impact of SOHO regulation on people is multifaceted, ranging from patient safety and access to therapies to innovation, compliance, and crisis management. Addressing the challenges related to legal clarity, regulatory classifications, and oversight is crucial for ensuring the safety, quality, and accessibility of treatments using Substances of Human Origin

How many blood a tissue establishments in the EU?	As of the latest available data, there are approximately 1,300 blood and tissue establishments in the European Union. These establishments play a crucial role in ensuring the safety and availability of blood and tissue products for medical use across the EU member states.	There are 4,563 blood and tissue establishments in the EU.	To determine the number of blood and tissue establishments in the EU, we can refer to the provided text for relevant information. Here are the key points related to the number of establishments: - Blood and Tissue Establishments (BTEs) are responsible for various aspects of the collection, testing, processing, storage, and distribution of human blood or blood components intended for transfusion. - The text mentions that there are around 4.658 blood and tissue establishments in the EU. - Additionally, it is mentioned that there are 4.563 Blood and Tissue Establishments without Medium S&Q Impact. - New establishments are numbered at 304, while new entities (former establishments) are listed as 750. - There are specifically 1.400 establishments responsible for the collection and/or preparation for transfusion of blood components. - For tissue establishments, there are about 3.258 establishments. Therefore, based on the information provided in the text, the total number of blood and tissue establishments in the EU is approximately 4,658 establishments.

What are diseases you can get from substances of human origin?		You can get communicable diseases from substances of human origin, such as blood, tissues, and cells. These diseases can be transmitted through these substances and may lead to serious adverse reactions or events in donors or patients, including life-threatening conditions or prolonged hospitalization (Source: Document 7)	### Diseases from Substances of Human Origin - **Risk of Infectious Diseases**: - Substances of human origin (SoHO) can carry communicable diseases, leading to various health risks. - Diseases such as hepatitis B and C, Ebola, syphilis, and tick-borne encephalitis can be transmitted through SoHO. - **New Diseases and Developments**: - Continuous advancements in technology and emerging diseases may require new testing on blood, tissues, and cells to prevent disease transmission. - **COVID-19 Pandemic Impact**: - The COVID-19 pandemic highlighted the importance of SoHO preparedness plans, such as for West Nile Virus and Zika. - Measures for preventing the spread of COVID-19 through SoHO were crucial during the pandemic. - **Role of ECDC**: - The European Centre for Disease Prevention and Control (ECDC) plays a vital role in providing guidance and risk assessments related to infectious disease transmission risks of SoHO. - **Regulatory Ambiguity**: - Lack of clarity in regulatory classification of various substances like cord blood, placental tissue, human milk, and isolated cells can lead to challenges in ensuring safety and quality.

			- **Potential Threats**: - The lack of clear regulatory frameworks for substances such as faecal microbiota transplantation and serum eye drops can pose threats to patient access and safety. - **Need for Harmonisation**: - Greater harmonisation of technical standards and regulatory frameworks is essential to ensure the safety and quality of blood, tissues, and cells across the EU. By addressing these challenges and ensuring robust regulatory frameworks, the EU aims to mitigate the risks associated with diseases transmitted through substances of human origin, ultimately safeguarding public health and patient safety.

Annex 5. European Commission Training Materials on Generative AI (ppt and visual materials in separate files)

Annex 6. Participatory Workshop Planning: Agenda and Logistics

Method	Open Space Technology	The World Café	Ritual Dissent
Objective	Open Space Technology is ideal for early phases of service design, as it allows for self-organized, flexible, and inclusive discussions where participants can propose topics of interest, form discussion groups, and move freely between them. This method is effective when there are complex issues with no predefined solutions and a need for diverse input.	The World Café method is ideal for bringing together multiple viwepoints. It encourages the sharing of diverse perspectives and the co-creation of knowledge through small group discussions. This method is particularly effective when exploring complex issues that benefit from a variety of viewpoints and collaborative thinking as well as action planning.	Ritual Dissent is a structured method designed to rigorously test ideas and assumptions through critical feedback. This method is effective when developing robust solutions to complex problems, as it allows participants to challenge and refine each other's proposals constructively. It ensures that the final framework is well thought out and resilient to various critiques.
Agenda	**Introduction:** to the specific topic, the policy/administrative context and processes, the technology and the objectives and choreography of the workshop. **Agenda Setting:** At the start of the workshop, participants are asked to propose topics related to main focus of the session. **Formation of Discussion Groups**:	**Introduction:** to the specific topic, the policy/administrative context and processes, the technology and the objectives and choreography of the workshop. **First Round of Discussion**: Participants are divided into small groups, and each table starts with a specific question .	**Introduction:** to the specific topic, the policy/administrative context and processes, the technology and the objectives and choreography of the workshop. **Developing Proposals:** Each group is tasked with developing a proposal (alternative: proposals defined before workshop) **Presentation and Dissent Round 1:** One

	Participants form discussion groups around the topics they are most passionate about. Each group appoints a facilitator to guide the conversation and a recorder to capture key points. **Discussion and Reporting**: Groups engage in deep discussions, exploring various perspectives, potential solutions, and challenges. Participants can move between groups to cross-pollinate ideas. At the end of the session, each group presents their findings to the entire assembly. **Harvesting and Action Planning**: The final part of the workshop focuses on synthesizing the insights gathered and developing concrete action plans and recommendations for the ethical AI framework.	**Sharing and Rotation**: After 20-30 minutes, participants are invited to switch tables and join a new group. One person remains as the table host to summarize the previous conversation for the new group. This process repeats for several rounds, with each round addressing a different but related question. **Harvesting Collective Wisdom:** Following the small group discussions, participants reconvene as a whole group. Table hosts share the key insights and themes that emerged from their discussions. A facilitator captures these insights on a large board visible to all. **Synthesis and Action Planning**: The final session focuses on synthesizing the collected ideas and developing concrete recommendations and action plans	group presents their proposal to the rest of the participants, who listen in silence. After the presentation, the presenters turn their backs to the audience, and the audience provides critical feedback, identifying weaknesses, potential risks, and areas for improvement. The feedback is candid and direct, without the presenters responding or defending their ideas during this phase. **Reflection and Revision:** The presenting group reflects on the feedback and revises their proposal based on the critiques received. This step encourages deep consideration and iterative improvement. **Repeat for All Groups**: **Final Synthesis**: The workshop reconvenes to synthesize the refined ideas. **Action Planning**: The final session involves developing a concrete action plan to implement the ethical AI framework. This includes identifying responsible parties, setting timelines, and establishing monitoring mechanisms.

Annex 7. Co-creation Tools for Designing RAG Services

features and low-fi wireframes. These could be adjusted and used together with service cards (cf Chapters 6.1.3 and 6.1.4) to facilitate the co-creation process. These drafts are based on the discussions with users during the trainings, not a structured process. ***Therefore, these materials are not a description of user needs for a RAG service, but a first draft to use for discussing and defining these needs.***

Table 1 Service_1 User needs (for discussion)

I want to be able to ask questions or generate text based on my own collection of documents **so that** I can efficiently utilize my data for various tasks.
I want to keep my documents in a folder on my local (U:) drive, restricted to commission-use only, **so that** I maintain data privacy and control access to colleagues in my team/DG/Unit.
I want to get to the original document with two clicks **so that** I can quickly verify the source.
I want to always have access to the latest versions of the documents **so that** I ensure accuracy and up-to-date information.
I do not want the content of these documents to pop up in any AI systems in the future **so that** I protect the confidentiality and integrity of my data.
I want to publish my documents/data and allow people to query it **so that** I can share public data effectively.
I want to publish my documents/data and allow a select group of people to query it **so that** I can securely share sensitive, non-confidential data with approved users.

Figure 1 Service_1 User needs (for discussion)

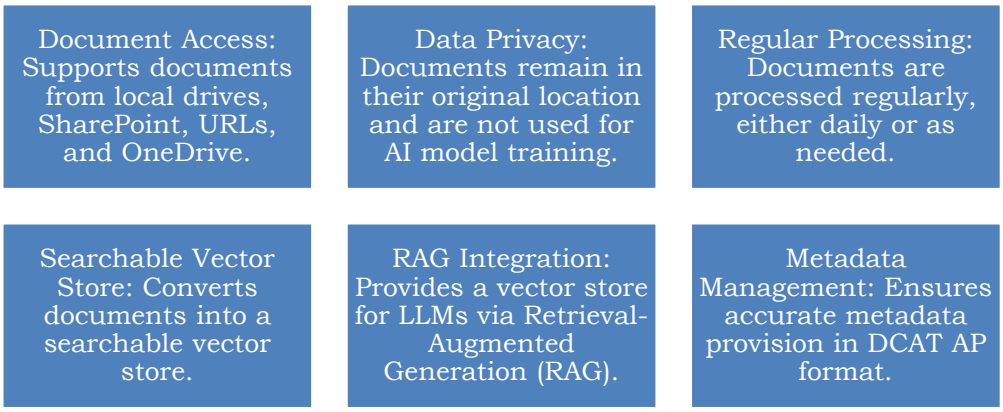
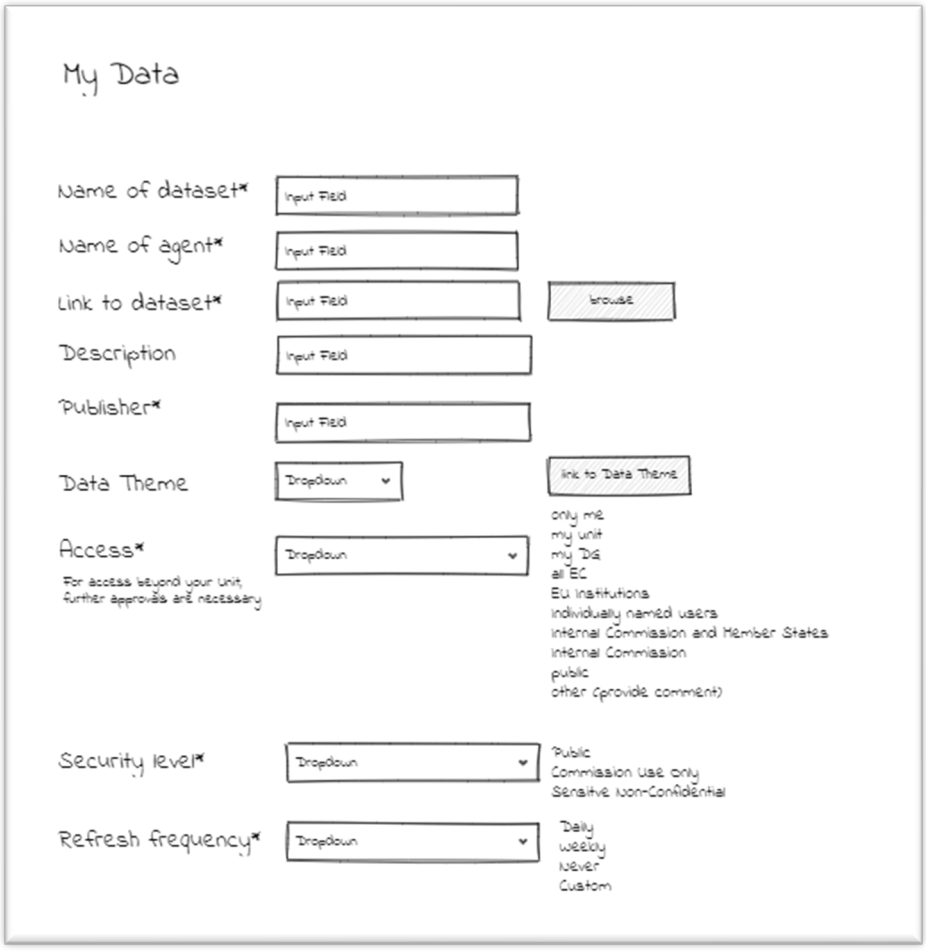


Figure 2 Service_1 Low-fi wireframe for documents for RAG agents



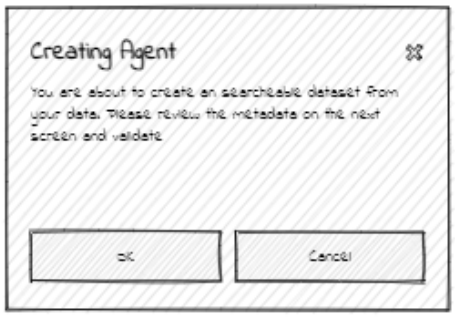


Table 2 Service_2 User needs (for discussion)

I want to use trusted data sources for generating drafts for my work **so that** I ensure the accuracy and reliability of the generated content.
I want to access a variety of data, including public, commission-use-only, or SNC documents, and different formats like text, tables, and figures **so that** I have comprehensive resources for my drafts.
I want to get to the original data with two clicks **so that** I can quickly verify the source.
I want to see the references in the generated text s**o that** I can ensure transparency and credibility.
I want to see which documents have been used even before reading the generated text **so that** I can understand the basis of the content.
I want clear indications on the security and access rights of the generated text **so that** I can manage data privacy and comply with regulations.

Figure 4 Service_2 User needs (for discussion)

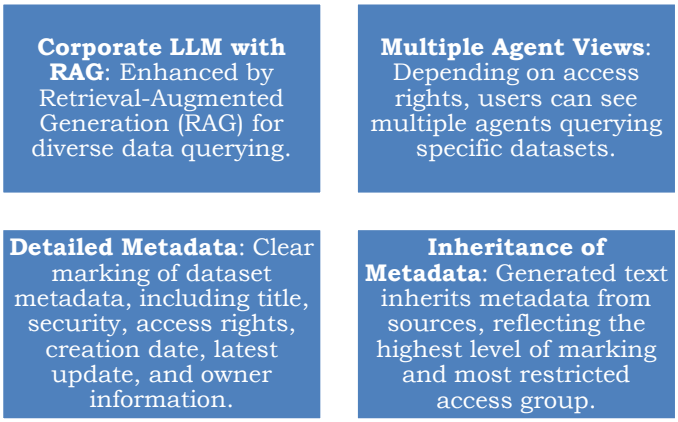


Figure 5 Service_2 Low-fi wireframes for calling RAG agents