





Article

The Proof Is in the Eating: Lessons Learnt from One Year of Generative AI Adoption in a Science-for-Policy Organisation

Bertrand De Longueville ^{1,*} , Ignacio Sanchez ², Snezha Kazakova ¹, Stefano Luoni ² , Fabrizio Zaro ² ,
Kalliopi Daskalaki ¹  and Marco Inchingolo ¹

¹ European Commission—Joint Research Centre, CDMA Building, 21 Rue du Champ de Mars/Marsveldstraat 21, B-1050 Brusselss, Belgium; snezha.kazakova@ec.europa.eu (S.K.); kalliopi.daskalaki@ec.europa.eu (K.D.); marco.inchingolo@ec.europa.eu (M.I.)

² European Commission—Joint Research Centre, Via Enrico Fermi 2749, I-21027 Ispra, VA, Italy; ignacio.sanchez@ec.europa.eu (I.S.); stefano.luoni@ec.europa.eu (S.L.); fabrizio.zaro@ec.europa.eu (F.Z.)

* Correspondence: bertrand.de-longueville@ec.europa.eu

Abstract: This paper presents the key results of a large-scale empirical study on the adoption of Generative AI (GenAI) by the Joint Research Centre (JRC), the European Commission’s science-for-policy department. Since spring 2023, the JRC has developed and deployed GPT@JRC, a platform providing safe and compliant access to state-of-the-art Large Language Models for over 10,000 knowledge workers. While the literature highlighting the potential of GenAI to enhance productivity for knowledge-intensive tasks is abundant, there is a scarcity of empirical evidence on impactful use case types and success factors. This study addresses this gap and proposes the JRC GenAI Compass conceptual framework based on the lessons learnt from the JRC’s GenAI adoption journey. It includes the concept of AI-IQ, which reflects the complexity of a given GenAI system. This paper thus draws on a case study of enterprise-scale AI implementation in European public institutions to provide approaches to harness GenAI’s potential while mitigating the risks.

Keywords: Artificial Intelligence; Generative AI; Large Language Models; LLMs; AI governance; organisational transformation; technology adoption; knowledge workers; AI technology; public sector innovation



Academic Editor: Demos T. Tsahalīs

Received: 15 April 2025

Revised: 27 May 2025

Accepted: 29 May 2025

Published: 17 June 2025

Citation: De Longueville, B.; Sanchez, I.; Kazakova, S.; Luoni, S.; Zaro, F.; Daskalaki, K.; Inchingolo, M. The Proof Is in the Eating: Lessons Learnt from One Year of Generative AI Adoption in a Science-for-Policy Organisation. *AI* **2025**, *6*, 128.
<https://doi.org/10.3390/ai6060128>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

“The proof of the pudding is in the eating”—Miguel de Cervantes, *Don Quixote*.

A tension defines our attitudes towards technology. On the one hand, we view it as a promise for continuous improvement of the human condition; on the other, we recoil at the prospect that it might diminish human flourishing by depriving us of central aspects of our agency, jobs, creativity, etc. [1] and also raise moral concerns [2].

AI is no exception. Recent developments in Large Language Model (LLM) technologies, rooted in Nobel-Prize-winning neural network techniques (<https://www.nobelprize.org/prizes/physics/2024/popular-information/> (accessed on 12 November 2024)), have brought forth the promise of completing a variety of tasks at a seemingly human-like level of performance. However, unlike other types of general-purpose technologies (i.e., technological innovations that find applications in a multiplicity of contexts), which have had the biggest impact on mechanical and routine tasks (e.g., the invention of the steam engine), these systems bear a direct impact on distinctively knowledge-intensive aspects of work [3].

In recent years, LLMs have become capable of matching and even surpassing human performance in an impressive variety of domains, showcasing the breadth of applications of these technologies across a variety of cognitive tasks, from competitive examinations for undergraduate and graduate admission to top universities, such as the SAT and GRE, to challenging professional tests such as the US Medical Licensing Exam and the Uniform Bar Examination [4]. Even for complex tasks such as synthesising knowledge from decades of medical research, AI systems are said to be competing with—if not surpassing—human abilities [5].

The most recent scholarly research in the field has only begun to scratch the surface of what this new technology may do for us [6]. Unsurprisingly, the same tension seems to govern the insights produced.

Some of the emerging evidence highlights the positive impact of LLMs on productivity. For instance, Brynjolfsson, Li and Raymond (2023) [7] found that the introduction of AI-based tools in the customer service industry led to a 14% average increase in productivity. Notably, novice and low-skilled workers experienced a 34% productivity boost. Similarly, research by Noy and Zhang (2023) [8] revealed that ChatGPT (at the time powered by GPT 3.5) increased the quality and speed of writing in professional contexts, primarily by substituting human effort rather than complementing worker skills.

In contrast, other scholars are tempering the pervading enthusiasm with a more cautionary approach. Messeri and Crockett (2024) [9] claim that although AI promises to enhance scientific productivity and objectivity, it also risks promoting scientific monocultures and instilling in its users the illusion of understanding. Similarly, Wirtz et al. [10] highlights ethical concerns related to AI adoption in business organisations at *micro* (e.g., loss of privacy, discrimination), *meso* (e.g., unfair competition, loss of independence of economic actors), and *macro* (e.g., rise in unemployment and social inequalities) levels. Exploring future venues of AI in science-for-policy advice, Tyler et al. (2023) [11] highlight the vast potential of these tools but also caution against several important hazards associated with their deployment. These risks include the well-documented issue of AI-generated “hallucinations”, challenges in ensuring the validity and unbiased nature of AI outputs, the need for transparency and governance in AI tools development, and the threat of disinformation.

Kreitmeir and Raschky (2024) [12] show that the effectiveness of ChatGPT in enhancing productivity may be dependent on the complexity of the intended task and on the size and quality of the underlying training data.

Similarly, Dell’Acqua et al. (2023) [13] provide evidence that productivity gains facilitated by AI tools such as ChatGPT in knowledge work are largely limited to tasks at the so-called “jagged technological frontier” of these tools. Conversely, for tasks that fall beyond this frontier, AI tools can “backfire”, even leading to a decrease in the quality of workers’ outputs.

What this early work shows is that the main question surrounding the uptake of Generative AI (GenAI) technologies cannot be reduced to a simple yes-or-no answer. The challenge of GenAI adoption encompasses a range of questions related to where, how, and for what purpose we deploy it. Rather than a binary decision, it is a matter of refining our understanding of GenAI’s impact across its various contexts of application and use cases.

More importantly, accurately predicting which tasks AI will handle proficiently and which will prove to be more difficult is extremely challenging. Given the uncertainty surrounding the precise boundaries of AI’s capabilities, or its “technological frontier”, a more effective approach is to delineate it through concrete case studies.

This challenge is the focus of the present paper, which draws on the deployment of Generative AI within the Joint Research Centre (JRC) of the European Commission. We gathered empirical data from multiple sources: over 10,000 user registration forms, user

surveys, and community feedback gathered through events and support channels. This approach enabled us to capture a rich, contextualised understanding of how GenAI tools can be integrated into a complex organisational context, identify emerging use cases, and propose transferable insights for practitioners. We believe that our experience in implementing GenAI at a large science-for-policy organisation offers a valuable and instructive case study to analyse and learn from. As an integral part of the European Commission, a prominent European public institution, the JRC is tasked with the critical function of aggregating and synthesising vast amounts of scientific knowledge and translating it into actionable insights that inform EU policymaking. The JRC's role is thus pivotal in bridging the gap between complex scientific research and the practical needs of policy development, ensuring that policy decisions are underpinned by robust scientific evidence.

Furthermore, as a public institution, the JRC is subject to the dual demands of accountability to the public and adherence to strict transparency requirements, as well as compliance with internal corporate rules and applicable regulatory frameworks (e.g., data protection).

Given the JRC's unique position at the intersection of science and policymaking, our experience with implementing GenAI at the JRC can offer valuable insights for other public administrations and research bodies worldwide. By documenting our journey and sharing our reflections and learnings, we also aim to contribute to the broader topic of how digital transformation impacts the public sector. We present a case study that aims to spark further reflection, dialogue, and practical experimentation with GenAI. We focus on the unique challenges and opportunities that arise at the intersection of science, technology, and public policy.

Practitioners in the public sector interested in experimenting with similar technological solutions can benefit from the lessons we have learnt and the JRC GenAI Compass, a conceptual tool that we introduce in this paper. So, while the case study presented here is specific to a given organisation, its conclusions are designed such that they are transposable to public institutions in EU Member States and beyond.

The purpose of this work is to facilitate a more nuanced and collaborative dialogue among scholars, practitioners, and public officials. Such a dialogue can contribute to a shared long-term vision for safely and successfully harnessing the potential of GenAI to help modernise public administration and improve productivity among knowledge workers.

2. Context and Case

2.1. JRC's Approach to Leveraging GenAI

As a subtype of Artificial Intelligence, GenAI can be seen as both a disruptive innovation [14] and a general-purpose technology [15], given its potential to disrupt traditional business models and create value across multiple fields. Although transformer models, on which GenAI is largely based, represent a relatively recent approach to natural language processing [16], their impact on traditional change management approaches in organisations remains to be explored [17]. Building on its pre-existing expertise in the field, the JRC is uniquely positioned to capitalise on the "GenAI revolution".

Immediately following the launch of Open AI's ChatGPT 3.5, AI scientists at the JRC began working on a prototype designed to evaluate the technology and its potential. Unlike third-party cloud solutions, which may pose information security risks, this prototype aimed to create a safe research environment, allowing for a higher degree of control and deeper understanding of the technology. In mid-2023 the JRC launched the GPT@JRC prototype, a safe environment for JRC staff to experiment with GenAI technology. Initially, the project provided a limited number of staff members access to state-of-art, proprietary, open-source, and open-weights LLM through a custom interface. As the project evolved

into a full pilot, its primary objective became to deepen our understanding of GenAI's opportunities and risks by facilitating, documenting, and evaluating its use within the JRC to bring value to our unique context as a science-for-policy organisation and later more broadly to EU institutions. We adopted a practical experimental approach, building on LLMs' basic text generation capabilities and improving output quality through other techniques. Our incremental, exploratory, and multidisciplinary approach can serve as a valuable guide or compass, helping other organisations to navigate the technology's complexities and unlock its long-term potential to transform key work processes, drive innovation, and enhance productivity.

2.2. *Anatomy of a State-of-the-Art GenAI System: The Case of GPT@JRC*

2.2.1. Design Fundamentals

The GPT@JRC system was designed and developed within a short timeframe, leveraging the JRC's pre-existing expertise in working with Large Language Models and operating state-of-the-art on-premises research infrastructures. To address the Commission's requirements for information systems, the project team adopted a set of design principles deeply ingrained within the project. These principles are outlined below to provide readers with a better understanding of the strategic elements at play in the development and deployment of a general-purpose GenAI scientific system in the corporate environment of our organisation.

Privacy and data protection by design.

GPT@JRC was designed and developed following a privacy and data protection-by-design approach [18]. This approach includes several key technical and organisational measures, including on-premises hosting and incorporating contractual measures for third-party AI models' providers to prevent unnecessary data collection and processing by third parties. The system was designed following the data minimisation principle and providing multilayered measures to ensure data security. This includes using secure encryption for data in transit and at rest, as well as minimising the data stored at the server side (an example of the latter is our implementation of the user's chat history—disabled by default—that was carried out on the client side using the browser's local storage). This privacy-first and data-protection-first spirit that guided the design and operation of the system was highly appreciated by our user community, playing a significant role in creating a trust-based atmosphere that stimulated the adoption of the platform.

Information security.

GPT@JRC was also designed following a security-by-design approach, ensuring compliance with the Commission's cybersecurity requirements to allow for the processing of non-classified internal data up to and including the Sensitive Non-Classified (SNC) level.

The entire GPT@JRC infrastructure is hosted on the premises, including open-source and open-weight AI models, which run on a dedicated JRC Graphics Processing Unit (GPU) cluster. The only exception made was for the commercial models supported, which were run in the cloud within Europe through a dedicated contract with the providers, ensuring that our data cannot be further processed or stored.

Notably, GPT@JRC is designed to ensure that SNC data is never processed in the cloud. The authorisation to process SNC data was a key requirement of the project, as it allowed for the exploration of specific use cases that would not have been possible otherwise.

Modularity and hybrid nature.

Given the fast-evolving nature of GenAI technology, we opted for a future-proof architecture that allows each component of the system to be upgraded and modified independently. To achieve this we adopted a modular approach, relying on functional "boxes" with clearly defined inputs and outputs that run on independent containers.

As a result GPT@JRC is designed to seamlessly integrate with any Large Language Model, whether hosted locally or in the cloud (with considerations for sensitive data handling as explained above). This modular approach is also applied to other components beyond the LLMs themselves (see Section 2.2.2).

GPT@JRC features a web interface for users. It also provides an Application Programming Interface (API) to enable the usage of the GPT@JRC services by other authorised IT systems or research teams within the organisation. Use cases that benefit from API access will be discussed in Section 3.1.2.

Independence.

Considering that the primary driver for building GPT@JRC was experimentation and the evaluation of GenAI technologies, independence from specific vendors was a key concern. In practice, this means being able to conduct tests on given technologies while mitigating the “black-box” effect (e.g., by being able to parametrise specific features) and avoiding dependence on third-party services whenever possible.

2.2.2. High-Level Architecture

The purpose of this section is to provide a high-level overview of the anatomy of GPT@JRC in non-technical terms. We consider the architecture of the system representative of state-of-the-art general-purpose GenAI systems. The goal of the current section is not to provide a detailed technical description but rather to focus on functional components, providing a basis for further discussions in this paper.

As illustrated in Figure 1, an interaction with GPT@JRC can be schematically represented as a flow that begins with a text input by the human user (referred to as a “prompt”) and concludes with a text output by the **Large Language Model (LLM)**. For the purposes of this paper, we can describe LLMs as artificial neural networks trained on vast corpora of textual data that model the probability distribution of these data by learning patterns and creating internal models. As a result, LLMs possess remarkable abilities to comprehend human language, handle unstructured data, and allow for interaction in natural language.

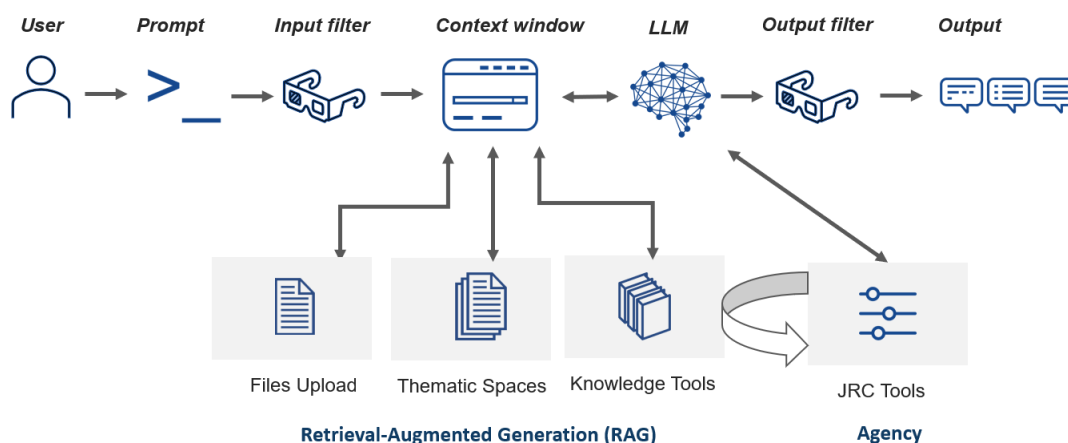


Figure 1. GPT@JRC high-level architecture diagram.

LLMs can be fine-tuned to follow instructions (i.e., instruction-following models) and support interaction in the form of multi-turn conversations. A Large Language Model has a **context window** of a pre-set size, which is designed to contain all the information related to the ongoing conversation. This context window can include the previous turns in the conversation, but it can also be populated with external information such as uploaded text files or the results of a search from a knowledge base. This allows the model to access additional context, similar to how a human might consult a book or conduct a Web search to gather more information before answering a complex question.

The Retrieval-Augmented Generation (RAG) component is what enables the system to populate the context window of the LLM with relevant information from a knowledge base, which can include, for example, all the scientific papers ever published by an organisation. The RAG component functions by retrieving knowledge to enhance (“augment”) the quality of the text generation (i.e., the LLM) by filling its context window with relevant information. As opposed to other techniques aiming to adapt a model permanently to a new set of data (such as fine-tuning), RAG is not computing-intensive and can be performed in near-real-time and in an ad hoc fashion. It is thus a suitable solution for use cases requiring a dynamic adaptation to context knowledge.

Input and output filters are also a common feature in GenAI systems as one of the essential AI safety components in a multilayered strategy. As their name suggests, these components carefully examine user input to ensure the absence of any potentially harmful content. A typical example often cited is the potential for GenAI systems to provide instructions on how to perform illicit or dangerous activities, such as building a bomb using domestic ingredients [19]. This risk can be mitigated by a “guardrail” component that can take the form of a dedicated specialised LLM designed for this type of content. The system can then be configured to intercept the flow in and out of the LLM, preventing the system from providing an unsafe response. GPT@JRC features a “guardrail” input filter designed and evaluated as part of its ongoing research activities.

While the previous description provides a schematic overview of how a state-of-the-art system such as GPT@JRC works, the immediate future holds more complex scenarios, such as collaboration and deliberation among multiple specialised LLMs to formulate the most accurate response. Additionally, LLMs may be empowered with access to specialised tools (e.g., statistical analysis software like R or MATLAB, a Python code interpreter, or a weather forecasting service) that it knows how to communicate with and consult based on the user’s request. The ability of a GenAI system to leverage specialised LLM tools and other types of software components iteratively, in an intelligent way, to reach the goal as described in the prompt is usually referred to as “agency”. The **agency component** in GPT@JRC is currently under development and will only be considered theoretically in the analysis described in the next section. Agency is an important notion to consider when evaluating the AI-IQ of a given GenAI system.

2.2.3. Introducing the AI-IQ: A Simplified Measure of GenAI System Capabilities

In this section, we introduce the notion of AI-IQ, one of the key conceptual contributions of this paper to the state-of-the-art of GenAI adoption within large organisations.

We define AI-IQ as a qualitative measure of the capabilities of a GenAI system based on its functional architecture. As such, rather than a genuine attempt to evaluate the cognitive abilities of LLMs, as explored in Wang et al. 2024 [20], the AI-IQ model is designed to serve as a tool to facilitate communication between experts and non-experts. The evaluation of cognitive abilities represents a significant field of research, but it falls outside the scope of this paper.

AI-IQ is meant to serve as a relative scale for comparing the capabilities of different GenAI systems, as well as evaluating the complexity of the GenAI solution required to address a specific real-life use case.

It is essential to note that the AI-IQ refers to the GenAI system as a whole, encompassing not only the LLM(s) that power(s) it but also the various internal components that contribute to its overall performance. GenAI systems typically comprise a diverse range of elements, many of which are not machine learning models per se but rather serve critical supporting functions that impact the system’s AI-IQ, such as retrieval-augmented generation (RAG) or agentic loops. By “agentic” in this context, we refer to AI systems

designed to operate autonomously, making decisions and taking actions independently to achieve specific goals.

As GenAI systems continue to evolve, it is anticipated that the AI-IQ scale will be expanded to include additional levels. Table 1 provides a definition of the current levels considered at this stage of Generative AI technology development, characterising each in terms of both GenAI system configuration and capabilities.

Table 1. Definition of the AI-IQ levels in terms of GenAI system configuration and capabilities.

| AI-IQ Level | GenAI System Capabilities | Typical GenAI System Configuration |
|-------------|---|--|
| 0 | Non-conversational—Limited to answering the predefined questions anticipated by the system developer. | Unlike open-ended interactive conversational LLMs, this level represents a more basic semantic search system that retrieves answers to similar questions predetermined in advance. Examples of this level include traditional chatbots found on airline websites, which provide pre-scripted responses to common user inquiries. |
| 1 | Conversational—Capable of engaging in open-ended conversations and generating text on demand. The accuracy and reliability of the responses may be limited by outdated knowledge and a lack of validation mechanisms, and there is a risk of hallucinations. | The system runs an LLM with some relevant information provided in the context window, such as a custom system prompt (i.e., an instruction that is always provided to the LLM at the beginning of each conversation, made by its programmer and not always shown to the user). |
| 2 | Basic RAG—Capable of providing answers based on specific, proprietary knowledge from the organisation that owns the system, complementing the GenAI's (e.g., LLM's) general knowledge. | The GenAI system employs a basic form of retrieval-augmented generation (RAG), enabling the retrieval of knowledge from organisation-specific documents or knowledge bases and making this information available to the LLM to inform its responses. |
| 3 | Advanced RAG—The GenAI system is capable of accessing, interpreting, and synthesising knowledge from the organisation's proprietary sources in a targeted and optimal way, specifically tailored to meet the requirements of identified use cases. | Multiple knowledge bases are made available to the GenAI system as specialised knowledge tools, each optimised for specific use cases through tailored RAG modalities. For example, one knowledge tool might be designed to generate a digest of the latest relevant research papers. The user chooses which knowledge tool(s) to activate in the context of the interaction with the GenAI system, allowing it to focus on the most relevant information for the task at hand. Advanced RAG is also characterised by the level of sophistication and optimisation of the RAG system, e.g., for dividing documents into information chunks, converting such chunks into machine-understandable embeddings, or ranking the relevance of selected information. |
| 4 | Basic Agentic—Enables the GenAI system to proactively leverage specialised data sources and tools, automatically selecting the most relevant ones from a predefined set of corporate resources and tools that have been made available with all necessary information for the system to utilise them. | A predefined set of tools and corporate resources is made available to the GenAI system, complete with all necessary information for their utilisation. An agentic system component leverages one or multiple LLMs to interpret user requests and understand how to effectively use the available tools. A RAG-based approach is used by the system using the relevant tools. This may involve tasks such as launching API requests, generating Python code that is executed in a sandbox, or taking other actions that enable the system to harness the capabilities of the provided tools. |
| 5 | Full Agentic—The GenAI system possesses advanced agentic features with some level of autonomy, enabling it to iteratively utilise available tools and resources to not only complete the provided task but also to adapt and improve its approach in response to changing circumstances or new information. This level of agentic capability allows the system to operate with increased independence and flexibility, pursuing multiple lines of inquiry and incorporating new data or insights as it works to achieve its objectives. | The GenAI system incorporates agentic components that enable it to iteratively complete complex tasks by breaking them down into smaller, more manageable tasks, leveraging available resources and tools to execute each step. The GenAI system benefits from a degree of autonomy in exploring various avenues, identifying an optimal solution, and planning various execution steps. Ultimately, the GenAI system can execute routine business tasks with the appropriate level of human oversight, freeing-up resources for more strategic and high-value activities. |
| 6 | Multi-Agentic Systems—The GenAI system functions as a swarm of agents, each with advanced agentic capabilities, working collaboratively to meet complex objectives. This level represents a significant leap forward in GenAI capabilities, enabling the system to tackle intricate tasks that require coordination, negotiation, and collective problem-solving. The swarm intelligence allows the system to adapt and respond to changing circumstances, leveraging collective knowledge and the ecosystem of available tools and data. | This level of GenAI system is characterised by a decentralised architecture, in which multiple agents operate autonomously, interacting with each other and leveraging the available tools and data to achieve the goal requested by the user. Individual agents contribute to the collective objectives of the swarm. Advanced knowledge management and sharing mechanisms can facilitate the exchange of information and expertise among agents, potentially enabling them to learn from each other and improve overall performance through continuous collaboration. |

In the subsequent sections, the notion of AI-IQ will be applied in practice to the various GenAI use cases identified within the JRC in recent months. Before presenting the results and conclusion, it is important to provide the relevant contextual information regarding the community of users who interact with GPT@JRC.

2.3. The Human Factor in GenAI: Meet the GPT@JRC Community

2.3.1. Community Approach

GPT@JRC started as a large-scale GenAI adoption experiment rather than the simple rollout of an innovative IT tool. We put in place a series of mechanisms to ensure the project team was in close contact with users, and the users with each other, to foster a mutual learning mindset.

All registered users were part of a community, featuring different themes based on users' interests. Regular events and sharing sessions were organised, focused on users' engagement and feedback (see Section 2.3.3).

The project team included communication and change management experts as well as community managers. Weekly coordination meetings were held with the technical leads, ensuring that the project's development and communication efforts were driven both by the AI and by human factors.

2.3.2. User Adoption and Onboarding Trajectory

Initially the project was limited to JRC staff, a population of about 3000 knowledge workers with varied profiles: scientists, policy analysts, IT experts, and administrative support staff. During the initial prototype phase, the platform accommodated over 500 users consisting solely of JRC staff.

Access to GPT@JRC was granted individually based on a registration form and agreement to the terms and conditions. Users were onboarded in weekly waves in which new users were also added to the GPT@JRC community and received basic guidance on how to use the interface.

After the transition of the project to a pilot phase, access was extended to members of other European Commission DGs and EU institutions. This precipitated a rapid expansion of the user community, which exceeded 12,000 users by November 2024.

The growth in the number of registered users is depicted in the following Figure 2, which illustrates the cumulative number of users granted access over successive onboarding waves. The onboarding took place on a first-come, first-served basis. Due to the rapid growth of our user base, we did not introduce any selection requirements or randomisation into the onboarding process. At the time of analysis, our population was large enough (for the JRC, it represented more than half of the entire organisation) to mitigate any initial self-selection bias in the data.

In addition to individual user access, the project granted API access to project teams, based on specific requests for a more advanced use of GPT@JRC or for integration into other information systems. In total, 134 projects were granted access to the GPT@JRC API service at the time of writing this paper.

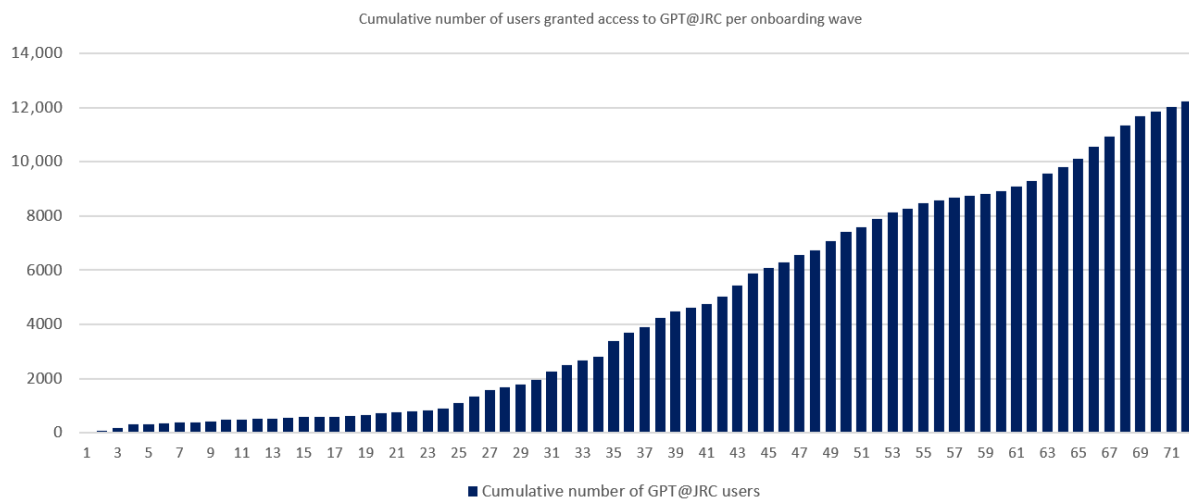


Figure 2. Cumulative number of users granted access, per onboarding wave (week 1 on the horizontal axis is 30 April 2023).

2.3.3. How We Learnt from Our Users: Our Research Material

Collecting feedback and understanding GPT@JRC usage was essential to inform our findings on how GenAI technologies are perceived, approached, and utilised by knowledge workers. This section presents an overview of the sources used to obtain the insights shared in this paper.

GPT@JRC Access request form. This short survey was designed to collect information on demographic variables, user expertise and intentions in using GPT@JRC. Questions ranged from respondents' unit designation and role within the unit to their perceived level of expertise with GenAI and their motivations for engaging with the GPT@JRC platform. The objective was to understand the composition and motivations of the user community. As of November 2024, 12,314 users had applied using the registration form, allowing extensive quantitative data analysis. Data from this form represents the full population of users of GPT@JRC and was the main source for the analysis and mapping of use cases presented in Section 3.1.1.

API service registration forms. This survey was completed by staff intending to use the GPT@JRC API service for a specific project that required this functionality. It was designed to capture the perceived technical expertise of the applicants, the specific projects or information systems for which it would be employed, and the nature of the data to be processed. Additionally, it inquired about the expected number of API calls per day, serving as a proxy for anticipated usage intensity. As of November 2024, a total of 284 projects had applied for GPT@JRC API access, with 134 actively using it.

In April 2024, a dedicated **user survey** was distributed to GPT@JRC users to collect feedback on their experiences with the platform (performance, usability, impact on daily productivity, and suggestions for improvement). A total of 560 valid responses were received (response rate = 10.2%). This feedback was useful to understand user satisfaction and emerging needs and to identify areas for improvement to guide the strategic development of the GPT@JRC platform. Due to the low response rate and relatively small sample size, which can introduce significant bias, we do not report analyses from this survey in the current paper. An **API user survey** was also launched in June 2024, with a response rate of 18% of all registered API users at the time.

GPT@JRC **user community events** were held in July and October 2023 and in April 2024. Before each event, a call for testimonials was made, and selected users were invited to present their user stories focusing, on specific challenges, how GPT@JRC assisted in them solving them, and what they learned. Informal conversations held with users before and

after the event allowed the project team to obtain a qualitative picture of the most common and impactful uses of GenAI among the GPT@JRC community.

The GPT@JRC team also took part in several key **events related to AI adoption** across the EU institutions and bodies. In total, about 75 presentations of GPT@JRC were delivered, reaching an estimated 6000 colleagues, who had the opportunity to engage directly with the project teams via Q&A sessions. The participation of the project team in events organised by the AI@EC Network—a network of professionals aiming to promote the efficient and compliant use of AI within the EU institutions—was particularly fruitful.

Lastly, due to the wide exposure of the project, about 100 **requests for specific support** were received by the project team. A first screening was performed to see if the request could be addressed by sending project documentation, if this was irrelevant. With the remaining requests (45) their characteristics were recorded in a ticketing system, and most were followed-up on by one or more formal exchanges.

Our findings are based on distilling the rich and varied insights obtained through these user interaction mechanisms. For the sake of conciseness and clarity we will not detail the qualitative and quantitative analyses that were performed, but will focus on their outcomes: a GenAI use case mapping three distinct levels of technological complexity, and a compass for navigating GenAI adoption in public organisations.

2.3.4. Analysis of User Information for Identifying Use Cases

The findings presented in the following section are the result of various efforts to synthesise “what GenAI is useful for” in knowledge-intensive tasks.

In our efforts to formally describe the use cases that emerged via engaged experimentation by thousands of users, we followed a six-step method to identify families of out-of-the-box LLM use cases (see Section 3.1.1):

1. Collection of intended use case data and first classification: data was collected through the GPT@JRC Access request form, focusing on intended use and specific research focus or policy areas. We identified seven broad use case categories through an initial analysis of user-reported applications, followed by a qualitative comparison with GenAI literature review findings to establish a refined set of generic use cases.
2. Use of GPT@JRC API for classification: an API with prompts was used to categorise the data under the identified generic use cases. The users’ responses were processed through this API with the help of two different AI models, LLaMa3 and Mistral-7b-OpenOrca, to classify each entry. The models were used as-is, without fine-tuning, with a prompt requiring the LLM to behave like a classifier and “label” each use case with the categories defined in step 1.
3. Data insertion and quality check: the classified data was then inserted back into the database. A rigorous quality check was conducted manually by our team, using filters and pivot tables to ensure the accuracy of the classification using human oversight.
4. Manual corrections: after the quality assessment, a total of 362 entries (6.6% of the classification) were manually corrected to adjust any misclassifications by the AI models.
5. Selection of a classification model: based on the manual checks, a comparison could be drawn between the accuracy of the two models used, LLaMa3 and Mistral-7b-OpenOrca. The classification performed by the LLaMa3 model was ultimately selected for use in further analysis, as it was estimated to offer the best accuracy and consistency.
6. Cross-tabulation and statistical analysis: finally pivot tables were created, and several cross-tabulations were performed to analyse user distribution across variables such as role, expertise level, and motivational factors.

3. Results and Findings

As a platform designed for knowledge workers to safely experiment with GenAI to assist them in the performance of daily tasks, the GPT@JRC project saw a consistent user growth and high levels of community engagement. This was in line with our expectations due to the GenAI hype observed worldwide since the launch of commercial services based on the latest generation of Large Language Models in Autumn 2022.

The results presented in this section draw on the analysis of the user feedback material described in the previous section, as well as on the personal experience of the project team members in their daily contacts with users.

We chose to present our findings in a form that can benefit readers engaged in similar projects and/or interested in the uptake of GenAI for knowledge-intensive business processes. We organise our findings into two conceptual models: a GenAI resource value pyramid, supported by a use case mapping at each level, and a JRC GenAI Compass leveraging the notion of AI-IQ introduced in Section 2.2.3.

3.1. A Resource Value Pyramid for Mapping GenAI Use Cases

As the project started with many ad hoc support requests, an efficient triage of these requests was important to develop a common sense of the effort necessary to address them. Three inter-related factors were identified that allowed us to distinguish between different types of GenAI use cases: technical complexity, the expertise/AI maturity of the requesting team, and the resource investment required. Many use cases exhibited relatively low values for these factors, while a small number exhibited high values. Therefore, the concept of a pyramid of use cases emerged, with clearly identifiable levels that could facilitate GenAI use case management (see Figure 3).

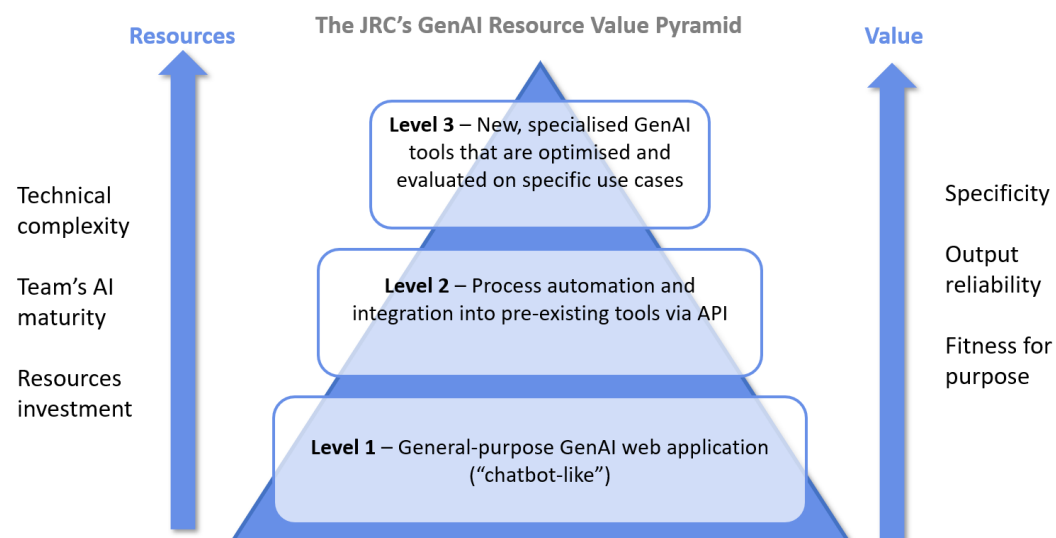


Figure 3. The JRC's GenAI use case pyramid.

At the lowest level of complexity, we map **Level 1 use cases** linked to “general-purpose GenAI” accessed via a custom user interface that allows for direct access to a selection of proprietary and open-source LLMs (the GPT@JRC platform in our case). This first level allows any user, regardless of their level of expertise, to start safely exploring LLMs’ basic capabilities. Over time, several custom features were added to the interface, such as the option to upload one or more files to the chat or to save recurrent prompts into a “prompt template” and re-use them later. These features allowed users to personalise the application to some extent. Still, Level 1 use cases rely exclusively on the capability of LLMs to process the input provided in the context window and generate a response, following a form of

reasoning and recalling knowledge from the model's training. Typical Level 1 use cases are discussed in Section 3.1.1.

Level 2 use cases can take the capabilities and impact of LLMs a step further by automating processes and/or by allowing for the integration of existing tools with the LLMs. This is accomplished by providing an API technology that allows for programmatic access to specific LLMs and embedding them at the input (prompting, providing context) and output (processing, evaluation, structure, etc) stages. Engaging with this level is only possible if the project team has enough technical expertise coupled with a sound understanding of the strengths and limitations of LLMs. The added value of this approach resides in the fact that LLMs' capabilities are combined with project-specific business processes. Examples of Level 2 use cases are discussed in Section 3.1.2.

Level 3 use cases leverage existing bodies of knowledge with more advanced AI techniques to create new, specialised AI tools that serve highly specific use cases. These tools can benefit from a high level of customisation. The accuracy and consistency of these specialised systems in performing specific tasks need to be formally evaluated and benchmarked. Naturally, they require a higher level of technical expertise and more resources to build. They also require domain-specific expertise to evaluate and ensure the desired level of output accuracy. This approach has the potential to bring significantly higher value, as new tools can be designed to tackle specific complex use cases addressing challenges considered essential and impactful for the organisation.

3.1.1. Level 1: Out-of-the-Box LLMs

Qualitative analyses of the data collected from our community concerning the use that staff intended to make of the GPT@JRC platform revealed seven broad use case categories (see Section 2.3.3) that can be considered largely mutually exclusive. They range from more generic tasks (text enhancement, creative assistance, learning) to more specific ones linked to the needs of a science-for-policy organisation (coding assistance, data analysis and interpretation, scientific literature review). Table 2 provides an overview of the Level 1 use case types identified, while Figure 4 presents their overall importance to our user community, based on their intended use declared during the registration process.

Table 2. Mapping of Level 1 GPT@JRC use cases.

| Use Case Type | Short Description | Detailed Description and Examples |
|------------------------|--|---|
| Text enhancement | Proofreading, summarising, translating, and drafting assistance. | Text enhancement involves the augmentation of written text to improve clarity, coherence, and overall quality. This includes activities such as proofreading, summarising, translating, and providing writing assistance. Examples of applications include the following: <ul style="list-style-type: none"> - Drafting and summarising project reports, policy briefings, speeches, and meeting minutes. - Streamlining repetitive administrative tasks, such as email drafting and document management. - Assisting in the creation of communication materials such as social media posts, newsletters, and web content. |
| Programming assistance | Helping with specific programming tasks and automatic documentation. | Programming assistance includes support for various programming-related activities aimed at increasing programming speed and efficiency. Examples of using AI for programming assistance are as follows: <ul style="list-style-type: none"> - Writing and debugging code or optimising existing code in various programming languages such as Python, R, VBA, SQL, Stata, and JavaScript (latest versions). - Generating unit tests. - Translating code between different programming languages. |

Table 2. *Cont.*

| Use Case Type | Short Description | Detailed Description and Examples |
|---------------------------------------|--|--|
| Text/data analysis and interpretation | Analysing documents, extracting specific information and help drawing conclusions. | <p>Text/data analysis and interpretation focuses on extracting and analysing information from diverse textual sources to inform decision-making processes. Examples of these applications include the following:</p> <ul style="list-style-type: none"> - Extracting information from technical, legal, and project documents. - Analysing tender documents and reports for assessment and compliance checks. - Researching EU policies and regulations in view of comparing policy documents and identifying discrepancies; investigating policy options and drafting policy papers; or analysing legislation for obligations and compliance. |
| Literature review | Discovering and summarising scientific literature. | <p>Literature review aims to streamline the process of identifying and summarising relevant scientific literature across various fields. It assists in the drafting of scientific papers, synthesising information to identify research gaps, and exploring new scientific questions. Examples of using AI to assist literature review tasks are as follows:</p> <ul style="list-style-type: none"> - Accelerating the screening and summarising of scientific literature across various fields. - Finding and summarising existing research, reports, and policy documents, including EU legislation and programmes. - Synthesising information from scientific articles and technical reports to identify key messages and research gaps. |
| Learning | Answering questions about anything. | <p>Learning encompasses a broad range of activities designed to expand knowledge on topics which can range from science to policy analysis, as well as the AI itself, aiming to understand its underlying technology and explore ethical considerations, such as bias in AI, data privacy, and the implications of AI-generated content. Examples of applications include the following:</p> <ul style="list-style-type: none"> - Utilising the AI for learning new topics quickly, especially those outside of the user's area of expertise. - Using the AI as a research assistant to gather information, compile data, or understand complex material. - Asking the AI to explain difficult concepts in simpler terms. |
| Project/process management | Automating tasks, help in planning, and supporting decision-making. | <p>Project/process management involves the use of AI to automate administrative tasks such as document management, planning, and supporting decision-making in project-related activities. Examples of using AI to assist project management tasks include the following:</p> <ul style="list-style-type: none"> - Organising and structuring input from various stakeholders; assisting in drafting terms of reference; evaluating project proposals; managing schedules; setting-up workflows for contract approvals. - Assisting with administrative processes related to public procurement and budget planning. |
| Creative assistance/critical review | Generating ideas, role-playing, and crafting scenarios. | <p>For creative assistance, users leverage AI to generate ideas, facilitate role-playing, and craft scenarios for various creative and strategic purposes. Examples of using AI to help in performing creative tasks are as follows:</p> <ul style="list-style-type: none"> - Assisting in brainstorming sessions and generating questions, test scenarios, and content in various fields, including producing concepts for policy options and exploring future scenarios. - Finding solutions to business problems and exploring process improvement opportunities. - Generating alternatives for decision-making and trouble-shooting. - Providing critical review of a user's ideas and analyses. |

| | Text Enhancement | Data Analysis and Interpretation | Learning | Coding Assistance | Project Management | Literature review | Creative Assistance |
|----------------|------------------|----------------------------------|----------|-------------------|--------------------|-------------------|---------------------|
| Administration | 726 | 195 | 236 | 115 | 163 | 33 | 62 |
| Managerial | 574 | 214 | 195 | 50 | 156 | 51 | 58 |
| Policymaker | 572 | 203 | 106 | 46 | 56 | 55 | 33 |
| Scientist | 273 | 168 | 93 | 170 | 38 | 72 | 20 |
| Support | 312 | 99 | 96 | 98 | 56 | 18 | 30 |

Figure 4. Heatmap of self-declared GenAI usage by profile of user (source: GPT@JRC access request form—5482 responses were received when the analysis took place).

At this early stage of GenAI adoption, which remains widely exploratory, the expected value from Level 1 use cases is challenging to quantify robustly. Efficiency gains and quality improvements can vary significantly depending on both the nature of the task and on user characteristics, such as the level of understanding and trust in the technology, domain expertise, and skills in interacting with GenAI models (prompt engineering). Even at this level, we see an impressive breadth in the potential that GenAI holds as a general-purpose technology to transform the way we work. The main limitations were observed for tasks that require in-depth understanding of specific concepts and issues (e.g., in the medical or legal domains), which is consistent with recent research (e.g., Kandpal et al. 2023 [21]).

Given the JRC’s focus on science and technology, it is not surprising to see coding assistance scoring relatively highly among the types of successful use cases, reflecting LLMs’ proficiency, which extends to programming languages (such as Python or R). We zoom into these capabilities, as they appear crucial when measuring early productivity gains through GenAI and will be instrumental when considering advanced use cases involving agentic AI systems (see Section 2.2.3 for a description of agentic AI systems). Further research through specific surveys would be necessary to better capture the impact of coding assistance within our organisation, but schematically, we can distinguish two use case types. For software engineers, GenAI provides code snippets from natural language description (“write a function that . . .”), assists in debugging (“what is wrong with this function aiming at . . .?”), or generates documentation/explanations in natural language for code (“can you describe that function?”). Thanks to the Application Programming Interface, software engineers can integrate LLM capabilities in their development environment, thus further increasing convenience and productivity. The other use case type is for scientists, who often have to adapt to new frameworks and toolkits based on their data analytics requirements. For them, the productivity gain is to smooth the learning curve (“can you tell me what is the function name in MATLAB R2024b to do with. . .?”), and speed up ad hoc tasks such as data cleaning (“please write a VBA script to do the following operation on all lines of an Excel sheet . . .”).

3.1.2. Level 2: Enhancing Existing Tools and Processes

The user group engaged in Level 2 use cases is narrower than that for Level 1, consisting of technically skilled users (e.g., data scientists, IT engineers). As the project was conceived to address the needs of the JRC, the scientific profile is over-represented at

this level, compared to that for IT professionals working in other parts of EU institutions and bodies.

The analysis of Level 2 use cases is based on the self-declared intended use from the API request form ($n = 256$), enriched by insights from pre-onboarding meetings and many direct exchanges with expert and technically skilled users. We discovered that users use the API for three main purposes: (a) sending batches of requests to the LLMs to process a large amount of data or implement advanced workflows; (b) integrating GenAI functionality within an existing IT system; and (c) carrying out experimental research with GenAI. The GPT@JRC API is employed across a spectrum of scientific and technical projects, from bulk data-processing to system enhancement and exploratory research, each contributing to the advancement of AI applications within the European Commission.

Table 3 presents the results of this analysis, while Figure 5 shows their frequency by type of use case, based on API users' intentions declared during the registration process.

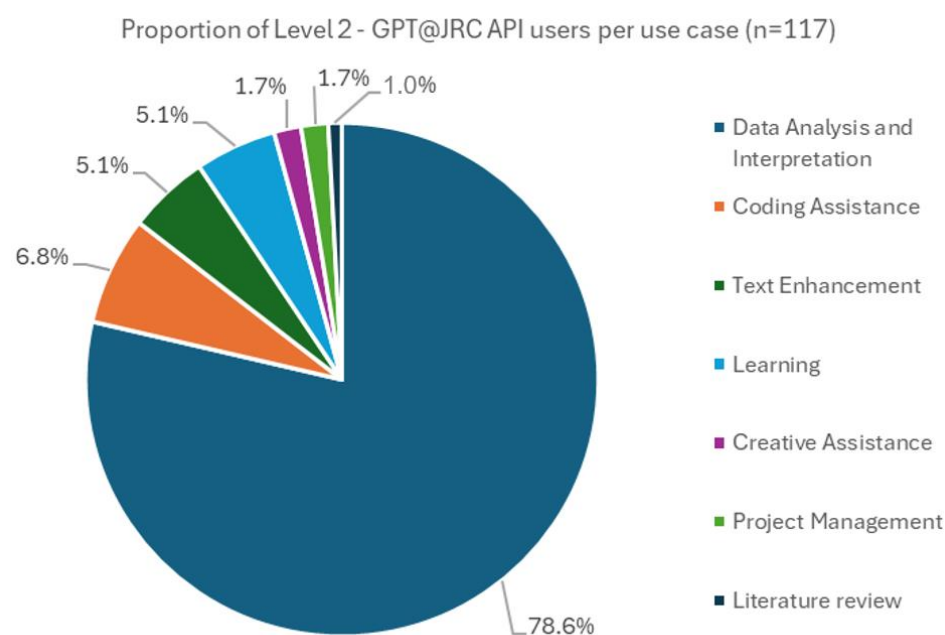


Figure 5. Proportion of GPT@JRC use cases per use case type (self-declared, based on 117 API access forms received at the time of the analysis).

Considering the resource value pyramid, the group of Level 2 use cases can be seen as narrower in scope, as fewer users are involved given that specific skills are required to benefit from API access. However, these use cases tend to score higher in terms of impact, quality, and specificity. Indeed, in a typical Level 2 use case, the leader of the beneficiary project ensures that the LLM has in its context window exactly the information needed to perform the task. For example, this could be a selection of news articles relevant to a given security event or guidance documents that all financial officers must be aware of to successfully perform their tasks. This is not always the case for Level 1 use cases due to a lack of user expertise or technical constraints. For example, a user may assume the LLM “knows” everything it needs to know (including non-public or recent information that was not available to train it) or that the LLM’s context window, the amount of input it can take from the user, is functionally limited by the user interface (e.g., uploading documents with a conservative size limit).

For these reasons, Level 2 use cases have successfully brought LLMs’ capabilities to a higher level and showed that important processes within the organisation can become faster, cheaper and more accurate by combining GenAI abilities with human expertise. This

is, for example, the case for public consultation analysis, systematic scientific literature reviews, and threat analysis in the (cyber) security domain.

Table 3. Mapping of Level 2 GPT@JRC use cases.

| API Use Case Type | Description | Examples |
|---|---|---|
| Sending batches of requests to GPT models | Users leverage this approach to facilitate large-scale text processing and analysis within scientific projects. This method involves programmatic access, typically via Python scripts, to dispatch multiple inquiries to LLM and/or embedding models. The applications range from content summarising, information extraction and classification, to more complex tasks such as generating synthetic data, conducting adversarial robustness tests, and exploring model explainability. This batch processing enables the handling of extensive datasets, allowing for the extraction of structured information from unstructured text, the comparison of model performance, and the augmentation of existing data analysis pipelines. | <p>Examples of engaging with the API to send batches of requests to GPT models include the following:</p> <ul style="list-style-type: none"> - Summarising content of news articles and other textual sources to extract essential information, for example for the monitoring of terrorism-related events worldwide. - Analysing public consultation results and performing trending topic extraction, similar-answer clustering, and sentiment analysis of large sets of free-text responses. - Translation of chemical formulations from product registries to standardise and compute toxicities. - Generation of embeddings for patent abstracts to facilitate comparison with other models and enhance data analysis capabilities. - Automating parts of systematic scientific literature reviews by evaluating the relevance of a large number of published scientific papers to a given research question. |
| Integrating GPT functionality within an IT system | This integration aims to enhance existing information systems by embedding GenAI capabilities to enable new functionalities. Use cases include the automation of tasks such as text classification, metadata assignment, and summarising within various domains such as finance, customs, and scientific research. The integration extends to the development of conversational chatbots, the augmentation of user interfaces with natural language processing features, and the provision of immediate user support by bypassing first-level query responses. The overarching goal is to streamline workflows, automate tasks, improve user experience, and leverage AI to provide assistance in complex processes. | <p>Examples of engaging with the API to integrate GPT GenAI functionality within an IT system include the following:</p> <ul style="list-style-type: none"> - Development of a conversational chatbot to address queries in the finance and contract domain based on legislation, guidelines, and “how-to” wikis. - Automation of text classification and metadata assignment for scientific documents to improve information retrieval and management. - Enhancement of user interfaces with plain text command capabilities for interacting with complex datasets such as satellite images. - Provision of automated filing suggestions by combining user documents with existing metadata and data sources to improve record management compliance. - Automatically drafting answers to enquiries based on past similar enquiries and reference documentation. |
| Experimenting | The experimental use of the API serves as a preliminary phase for testing and development purposes and to carry out experimental research. It encompasses the investigation of research on GenAI capabilities and limitations and broad experimentation with potential GenAI applications in scientific projects. Experimentation also includes the development of pilot projects to assess the feasibility and performance of GPT@JRC in various scenarios. This phase is crucial for understanding the potential and limitations of the API before committing to full-scale implementation or integration within IT systems. | <p>Examples of experimenting with the API include the following:</p> <ul style="list-style-type: none"> - Testing the integration of GPT@EC Pilot with GPT@JRC to evaluate compatibility and performance. - Investigating the potential of the API in the field of cybersecurity. - Checking the API’s ability to speed-up work on ongoing and future scientific projects related to energy systems and markets. - Exploring the use of the API for compliance checks on text descriptions to assess adherence to regulatory standards. |

Interestingly, the vast majority (about 80%) of Level 2 use cases were declared as text/data analysis and interpretation, suggesting that automation (sending batches of text

to LLMs) seems to be a preferred approach compared to integration within existing IT systems. This finding could be partially explained by the fact that our API user population is biased towards scientific profiles and we are still in an experimentation phase. We expect that Level 2 use cases will become more numerous and impactful as skilled users' understanding of GenAI capabilities increases, supported by AI literacy initiatives within our institution.

To move beyond Level 2 use cases towards the design of specialised GenAI systems, the validation of their fitness-for-purpose and rigorous evaluation are key to success. We therefore see a continuum between the informed and impactful use of GenAI through API (Level 2) and the successful development of specialised GenAI systems (Level 3).

3.1.3. Level 3: Creating New, Specialised GenAI Systems

The JRC's experience with GPT@JRC's large-scale deployment within EU institutions and bodies, as well as high-level exchanges with other public institutions currently exploring GenAI (e.g., CERN, World Economic Forum), suggest that the most promising long-term value strategy for GenAI adoption lies in what we define as Level 3 use cases. Level 3 implies the development of specialised GenAI systems, carefully engineered and evaluated to reliably tackle complex tasks within the organisation. Recent studies [22,23] seem to suggest that this empirical observation is also true for other knowledge-intensive organisations.

To explain in non-technical terms what we mean by “carefully engineered and evaluated” specialised GenAI systems, we propose taking a look at a series of examples distilled from numerous real-life questions and observations from our population of over 10,000 GPT@JRC users.

As a first example, let us consider a difficult question: which of the EU Member States of the Mediterranean region has seen the biggest relative growth in its GDP in the last five years? Most of the LLMs are likely capable of translating the first part of the question into a list of countries with some room for interpretation (does the Mediterranean region include Portugal?). They are also capable of interpreting the acronym “GDP” given the context: “growth” and “country” go well with the notion of gross domestic product. Then, the LLM needs to “know” which year it is, usually available in an automatically generated part of the system (e.g., “Today's date is 13 October 2024”). In this case, the LLM should have no difficulties understanding that “in the last five years” means 2020, 2021, 2022, 2023, and 2024, unless it chooses not to take into account the ongoing year. The most challenging task would be to ensure that the LLM retrieves the actual GDP value in each country for each year via the system's RAG component. Such data would be typically available in table format within various documents and reports. Interpreting tables is a challenging task for LLMs, and particular care is needed when engineering the RAG component to ensure all information is available in a form the LLM understands with minimal ambiguity. And even if the information per year and per country can be fed in context in a clear and non-ambiguous way, the LLM would have to avoid a final trap, in which many bachelor students would fall: we are asking to compare growth in relative terms, i.e., not the absolute growth in billion EUR but the positive difference over time in percentage points.

A second example is a system that facilitates systematic scientific literature reviews: work-intensive and time-consuming tasks usually performed by scientists. For this GenAI use case, it is crucial to define objective criteria for the selection and prioritisation of certain scientific publications over others to inform the design of the RAG component (e.g., should we prioritise the most influential research outputs over more “trivial”, less relevant ones?), specifying what objective criteria to use for the prioritisation of some papers over others. Also, as terminology in science is often complex, evolving, and domain-specific, it must be ensured that the LLM does not “confound” key scientific concepts and terminology.

As a final example, we can consider a GenAI system that has all the required knowledge in the right format and is equipped with the necessary lexical and reasoning resources to tackle complex issues, and yet suggests unethical/ flawed solutions to them. For example, when asked to advise on the best way to demonstrate the effectiveness of parachutes, a GenAI system may suggest performing a double-blind study with the justification that it would allow us to formally exclude any placebo effect of such gravity mitigation devices. In a different scenario, since a war is an absolute tragedy for those who take part in it, wars should be always avoided, even if such a decision contradicts signed treaties or international laws. Such extreme cases illustrate the need for alignment of GenAI systems.

Alignment is a process that aims to ensure that an AI system consistently performs according to the intended goals, preferences, and ethical principles of its creator. Alignment is widely studied (see e.g., Hagendorff 2024 [24] for a comprehensive overview), often through the angle of legal liability (i.e., preventing a GenAI system from facilitating any illegal or criminal activity) and of non-discrimination (preventing GenAI from perpetuating the many biases already present in society). Some use cases may require finer alignment to ensure the right choice of wording (e.g., law, diplomacy) or lines to take on specific issues (e.g., communication).

As these examples illustrate, developing fit-for-purpose GenAI systems that can reliably tackle complex questions involves addressing numerous business and technical issues related to information retrieval, terminology, specialised reasoning and alignment, among others.

For this reason, building Level 3 GenAI systems is resource-intensive and requires a longer-term approach to build the necessary capabilities within a given organisation. In other words, the high AI-IQ of Level 3 solutions comes at a cost—the cost of complexity—and dodging that cost can lead to failure to fulfil decision-makers' expectations of GenAI, and to reputational damage.

It is crucial that the Level 3 use cases selected by any organisation as having the highest potential impact are well defined in terms of target user and value and have sufficient resources to progress quickly. Moreover, it is essential to have the culture and skills of GenAI evaluation within the organisation.

At the JRC, we have identified several Level 3 use cases with high potential. We envision a JRC virtual scientific assistant, specialised in scientific tasks such as performing systematic scientific literature reviews, generating “digests” of the most influential recent research in a given field, or engaging in advanced brainstorming on research topics. We also intend to invest in GenAI evaluation capabilities (see Section 4).

3.2. A Compass for Innovating with GenAI

3.2.1. From the Diamond Model to a Checklist for Your GenAI Use Case Journey

Whether GenAI will live up to its promise of bringing value to organisations depends not only on the evolution of the technology itself but also on how effectively each organisation manages the process of organisational change and the implementation and deployment of the technology.

We considered the case of GenAI through the lens of Leavitt's widely applied Diamond model, also known as the systems model [25], focusing on its four intricately connected dimensions: task, people, structure, and technology. In our context, *task* refers to the specific use case and its characteristics. *People* are represented here by the users of GenAI technology within the organisation, who perform the specific task or use case. *Structure* refers to the type of organisation in which the technology is being implemented and the context in which it operates. Lastly, the *technology* dimension is captured by the AI-IQ concept introduced in Section 2.2.3, as a measure of the technical complexity of GenAI

systems designed to perform a given *task*. The AI-IQ concept is therefore a component of the JRC GenAI Compass, allowing for the balancing of system complexity with user expertise.

There is an inherent tension, a trade-off between the four dimensions. For example, the expected value of a GenAI use case can be offset by the reputational and other risks if the outputs are of insufficient quality. Similarly, a lack of user expertise can reduce the value of use cases at all three levels, acting as a barrier to technical implementation (e.g., data science skills at Level 2), adoption (e.g., prompt engineering skills at Level 1), and evaluation (e.g., domain expertise at Level 3). Reflecting on these trade-offs can be a useful exercise in planning GenAI implementation.

Below, we introduce each dimension in more detail and propose an initial set of questions that can serve as a starting point for considering the implementation of specific use cases based on the Diamond model. While this list is not exhaustive, it can be a useful tool to facilitate initial discussions for any organisation reflecting on how to approach the implementation of GenAI.

I. Use case characteristics (*Task*). This dimension refers to the use case itself, its alignment with the core tasks of the organisation, and the auditability of the output produced by GenAI tools. Some questions to ask when reflecting on this are as follows:

1. Can the output be verified or evaluated for quality and/or accuracy (auditability)? Responses include: Yes, by anyone using the tool; Yes, but it requires domain expertise; Yes, but additional technical implementation is required; No, extensive manual verification is required.
2. To what extent does this use case fit with the core tasks of the organisation?
3. Are there ethical and/or regulatory concerns about relying, even partially, on AI in this use case (e.g., the requirements from the AI Act in the EU applicable to AI systems that fall under the category of high risk in the regulation)?

II. User characteristics (*people*). This dimension refers to the level of user expertise required for a specific use case, encompassing both Generative AI and domain-specific expertise. Other user characteristics such as resistance to change, digital literacy, and trust are also important to consider. Relevant questions to ask in this respect include the following:

1. What is the required level of user expertise with Generative AI to successfully leverage this use case within the organisation?
2. Is the level of digital literacy across the organisation sufficient to ensure that user expertise in GenAI can be acquired quickly?
3. Is the level of trust in the technology sufficient for user acceptance, or is it likely to bring about resistance to change?

III. Organisation (*structure*). This dimension refers to factors that are inherent to the nature of the organisation, such as whether it is private or public, the economic or societal domain in which it operates, its main product or value proposition. Some questions for reflection in this area centre around expected value and risks.

1. What is the potential of this use case to bring value to the organisation?
2. What are the potential associated risks (reputational, financial, etc) in the case of poor quality or insufficient accuracy?

IV. Technical approach (*technology*). This dimension refers to the level of technical complexity of a GenAI solution suitable for the use case, exemplified by the AI-IQ axis of the GenAI Compass. We address this dimension last, as it requires an organisation to make an informed decision regarding the technical approach to take, drawing on analysis and reflections from the previous three dimensions. We can categorise use cases across the organisation at the three levels of the pyramid by asking the following questions:

1. Is the use case described in accurate enough terms to allow for the design of a GenAI system that would most likely address it?
2. If so, which AI-IQ score would the GenAI solution require?
3. Does the organisation have the necessary financial resources, skills, and time to implement and evaluate a solution with such an AI-IQ, either independently or with the assistance of external partners (e.g., outsourcing certain aspects of the work)?

3.2.2. Making People Work Together with the Technology: A Driver for GenAI Adoption

In their widely cited and commented-upon book *Human + AI: Reimagining work in the age of AI*, Daugherty and Wilson (2018) [26] emphasised human–machine collaboration as the key success factor for organisations to successfully harness AI in their business processes.

Following this line of thought, and capitalising on our experience analysing a large number of GenAI use cases across EU Institutions and bodies, we introduce the JRC GenAI Compass as a tool to give a sense of direction for organisations exploring the largely uncharted territory of GenAI adoption for their business processes.

As with a real-life compass, the JRC GenAI Compass is bi-dimensional, focusing on the human and technological dimension of the Diamond model. Like a real-life compass, it will not tell you where to go but will allow you to gauge your direction in any context and circumstances. The technological dimension is represented by the AI-IQ concept we introduced earlier, while the human dimension aims to capture the users' required AI literacy. In the light of the experiences we shared in this paper, we believe that thinking of any potential GenAI use case in terms of these two axes can greatly facilitate strategic decision-making for GenAI adoption within an organisation. One of the reasons is that representing the complex organisational reality in these two dimensions will, in most cases, provide a good enough approximation. This is because they tend to correlate well with numerous organisational factors, such as available skills, maturity of data/knowledge management processes, technical capabilities, budget availability, pre-existing optimisation of business processes, etc. The choice of people as a dimension to capture AI awareness and readiness of the user seems coherent with the findings of Flavián et al. [27], which demonstrated it as a key factor in AI adoption among service industry customers.

Figure 6 represents an instantiation of the JRC GenAI Compass, in which selected GPT@JRC use cases are mapped for illustration purposes.

An alternative formulation of the JRC GenAI Compass can be based on pairing other dimensions of the Diamond model. While factors linked to the user dimension, such as AI literacy, are flexible and can be changed in a favourable direction, use case characteristics and organisational factors are more fixed and difficult to change. Therefore, they should be introduced as early as possible and carry more weight during the reflection process. Evaluating GenAI use cases along the *task* dimension of Leavitt's Diamond model can be useful to determine whether specific use cases are a good fit for implementation at higher levels of AI-IQ, where more resources need to be committed to development needs.

To demonstrate this, we present a mapping of use cases along two axes: AI-IQ and auditability (see Figure 7). Along the x-axis, to reflect task characteristics, we propose the term “auditability”, defined as the degree to which the output produced by a GenAI system can be verified or evaluated by the user to ascertain its level of quality and/or accuracy. This is not a yes/no dimension but a continuous one, reflecting the level of effort (and time) needed to perform an evaluation before an output can be used in practice. This can also be reflected by the need for a domain expert to evaluate the output. Use cases for which outputs are easy to audit present potentially better value for a given level of investment. For example, a scientist with expertise in a specific field would be able to

quickly evaluate the quality of output from a scientific assistant responding to a specific scientific question. Conversely, output from a grant management assistant that is designed to check compliance with a set of predefined criteria would require a substantial amount of effort to be evaluated. Tools at high AI-IQ levels should therefore be designed with a clear strategy for ensuring high accuracy without the need for user verification.

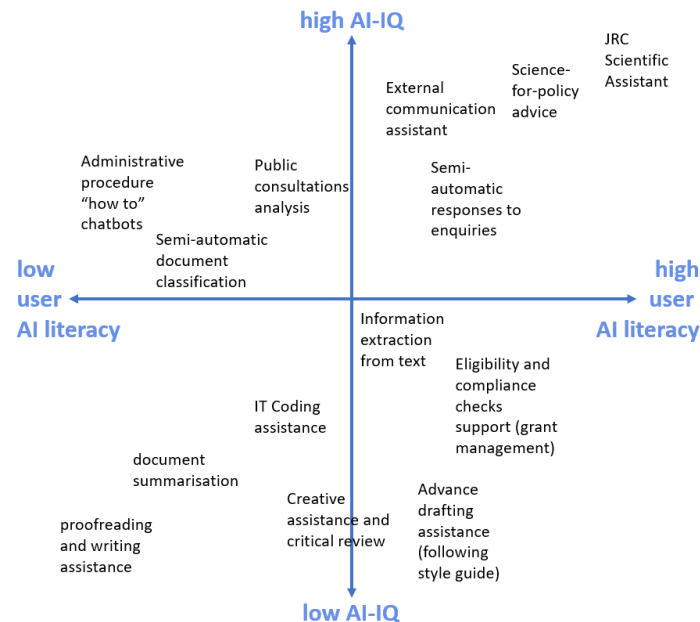


Figure 6. Mapping selected GPT@JRC use cases according to the JRC GenAI Compass with people and technology as key dimensions.

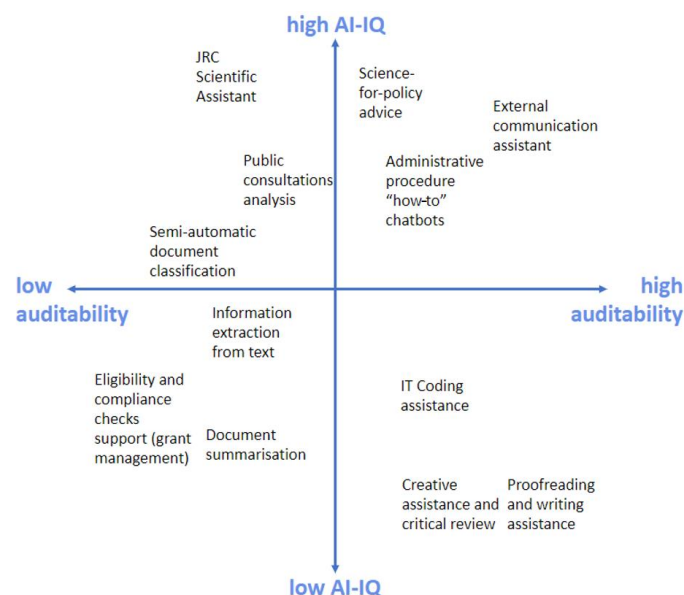


Figure 7. Mapping selected GPT@JRC use cases according to the JRC GenAI Compass with task and technology as key dimensions.

4. Discussion: Perspectives for GenAI Adoption

The organisational factor.

The introduction of GenAI is often compared to the invention of electricity or the emergence of the Internet in terms of its expected transformative impact on society. The hype surrounding it is palpable, even if we are already seeing some signs of it subsiding. As in the private sector, public organisations worldwide are facing increased pressure to

leverage Generative AI to increase their efficiency and help mitigate the risks posed by increased workloads, shrinking resources, and increasingly complex societal challenges. This context can easily tip the decision scale in favour of off-the-shelf solutions, whether they are integrated with existing collaborative tools or not. GenAI technology is unique in its breadth of potential applications and flexibility to assist in almost any knowledge-intensive task. However, our experience confirms that LLMs alone are not suitable to be applied out-of-the-box to complex tasks. There is no magic in AI, and disappointment comes quickly when unrealistic expectations are not met.

We identified and mapped high-value use cases specific to public organisations at three levels of technological complexity. At all levels, LLMs' basic functionalities can be enriched with internal knowledge that reflects each organisation's unique context to improve output quality. The highest level of customisation requires the implementation of complex GenAI system designs, as well as a high level of maturity in the organisation's business and knowledge management processes. Our experience confirms that the impact of the organisations' AI experts can be maximised by surrounding them with like-minded business process and knowledge management experts.

Measuring impact is another key dimension that is often overlooked or underdeveloped in organisational change initiatives. The first step is to define what measurable impact means through a mix of targeted quantitative and qualitative indicators. Some dimensions to consider are increases in efficiency through time savings or productivity gains. For a science-of-policy organisation, this can mean reducing the time from an initial policy question to actionable policy recommendation based on scientific insights, reducing the time researchers spend on literature reviews, and performing scenario generation or more repetitive tasks. Other factors are linked to improved output quality, such as increased accuracy, comprehensiveness, and impact of science-for-policy briefs or higher stakeholder satisfaction.

Ultimately, our experience suggests that a successful strategy for leveraging GenAI in a large organisation with a wide variety of use cases would be a semi-decentralised one, with a core team having high levels of GenAI expertise (understanding both its capabilities and limitations, and how to operate it technically) providing access and advice to/about the LLM, while the local project leaders with business expertise take care of the actual implementation in specific use cases.

The human factor.

Our GenAI Compass also underscores the critical importance of the people dimension, which is often overlooked in organisational change initiatives driven by technological innovation. This oversight can bring significant risks, as the human aspect plays a critical role in the ultimate success or failure of any new technology. Over the last year and a half, our enthusiastic, often self-trained users have already managed to acquire an impressive empirical understanding of how LLMs function and how to leverage them to increase their productivity. More importantly, they have been sharing their insights regularly with the rest of the community. A phrase we hear often within the GPT@JRC community is "we're all in this together, we need to learn from each other". Early adopters, when given the right mandate and opportunity, can act as essential multipliers of adoption, helping to increase both the trust and literacy of other users. This can be achieved by establishing, growing, and continuously nurturing a dedicated community of practice, in which the JRC's *Communities of Practice Playbook* [28] can provide evidence-informed guidance on the approach to take.

The technology factor.

Insights into the human understanding of GenAI technology are key to its successful adoption in any knowledge-intensive organisation. In our numerous interactions with

GenAI users from different backgrounds and with different levels of expertise, a consistent pattern emerged: when faced with GenAI, users intuitively tend to approach it by drawing analogies with human intelligence. Such analogies are particularly useful in facilitating communication between AI experts and non-experts, especially in elucidating key concepts that are not easy to understand, such as the probabilistic nature of LLMs and their limited context windows.

While parallels to human intelligence are useful, our findings support early evidence that LLMs fall short in replicating the cognitive abilities of humans, especially in complex domains [29]. A main source of these limitations lies in their inherent design. While LLMs have demonstrated impressive capabilities in understanding and generating natural language, they still struggle to disentangle language from knowledge and knowledge from reasoning. When trying to explain the gap between human intelligence and GenAI, we often say to users “it talks, it knows, and it reasons, but not exactly in the same way we do”. This statement is intentionally meant to raise more questions than it answers. It is in line with recent research on GenAI’s emerging reasoning capabilities [29] and captures an important distinction between language production, knowledge retrieval, and more complex abilities envisioned for future general-purpose AI systems [30].

“It talks” refers to LLMs’ impressive ability to understand and produce natural language, an ability that has captivated human imagination for centuries—from the mechanical head (also known as the brazen head) in the Middle Ages, allegedly destroyed by Saint Thomas Aquinas (who saw it as an offense to God), to the character of HAL 9000 in the movie *2001: A space Odyssey*. Our human intuition “sees” the ability to engage in brilliant conversations on any topic as a hallmark of intelligence. However, while this may be true for humans, it is not necessarily the case for LLMs, sometimes described as powerful stochastic parrots [31]. It is imperative to distinguish their purely linguistic abilities from other commonly recognised manifestations of intelligence.

“It knows” refers to LLMs being trained on vast corpora of knowledge in the form of billions of internet pages, fora, news articles, scientific literature, books, essays, novels, and more [32]. Despite this impressive breadth of input, hallucinations arise unexpectedly and in any context, so it would be naïve to assume that the LLM “knows everything”. Similarly to the way that human memory can be unreliable [33], not all knowledge used to train LLMs is readily available at all times and in all contexts. Our findings confirm that it is the coexistence of knowledge acquired during training and contextual knowledge (e.g., when a RAG component is included) that enhances the knowledge capabilities of LLMs. As illustrated in this paper, the successful adoption of GenAI requires a revision of an organisation’s knowledge management processes to make them AI-ready, combined with an in-depth understanding by all stakeholders (GenAI users, managers, IT experts) of how LLMs come to “know” things.

“It reasons” refers to a prominent topic in current AI research assessing the reasoning abilities of LLMs [34]. By design, LLMs are word-predicting machines without formally programmed reasoning abilities. The extent to which these abilities are emerging is highly debated. For example, while commonly recognised benchmarks exist to measure the textual and computational reasoning of LLMs, some argue that these benchmarks are too shallow and simple [35]. Our experience with Level 3 use cases confirms that much like knowledge extraction, the ability of LLMs to “reason” and make informed decisions comes without any guarantee of accuracy, especially when an in-depth understanding of complex domains is required.

LLMs may not (yet) be able to accurately capture knowledge and reason correctly. Still, human innovation has already produced formally verified reasoning machines that excel in specific tasks with high success rates: computer programmes. By enabling LLMs to interact

with or even create specialised computer programmes, we can potentially create GenAI systems that leverage the strengths of both approaches, leading to enhanced reasoning abilities and higher accuracy.

This approach can help us produce agentic systems, in which LLMs have access to the right knowledge and are able to choose which computer programmes to utilise for which purpose [36,37]. In the long-term, we believe this can be highly transformative in terms of the power and potential uses of GenAI in knowledge-intensive organisations, with early signs of its potential already emerging [38].

Evaluation, evaluation, and more evaluation.

Our findings highlight the need for robust evaluation methodologies to assess the effectiveness of GenAI systems. Since the advent of highly capable LLMs, the trend has been to complement their language generation abilities with contextual knowledge (e.g., via RAG approaches). As a result, GenAI systems are becoming increasingly complex: going from a single chat interface to the emergence of multi-step flows with specialised components involving multiple LLMs, called sequential or parallel mode. Managing such technological complexity is a key challenge for next-generation GenAI applications, as the multiplication of components will also lead to the multiplication of factors explaining the success or failure of a task. Tackling these challenges will require a reliance on robust GenAI system evaluation methodologies and tools that allow for fine-grained task-specific evaluations, which mainly remain to be developed (an interesting contribution was made recently by Bommasani, Liang, and Lee 2023 [39]). For example, while off-the-shelf solutions are usually seen as performant for text summarisation tasks, they can fail dramatically when the text includes some quantitative statements, so an ad hoc evaluation would be required, for example for a system summarising survey result interpretations (containing, e.g., statements such as “80% of respondents agreed, among which 35% claimed to have no prior experience”).

We consider use-case-specific GenAI evaluation to be an essential prerequisite for successful Gen AI implementation. Organisations should focus on developing a culture of understanding and evaluating GenAI technology in the context of their own use cases and business needs. We continuously invest in reusable tools and methodologies to perform such evaluations as an integral part of our organisation’s GenAI adoption process. The second iteration of the GenAI Compass we proposed can be particularly useful in prioritising use cases that easily lend themselves to evaluation methods by focusing on the “AI auditability” dimension, instead of “users’ AI literacy”.

Limitations.

While we believe that our findings might be useful for similar organisations in devising their GenAI adoption strategy, it is crucial to embed them within broader considerations that, in our view, provide the required intellectual guardrails for the purposeful and effective deployment of such technologies.

Firstly, it is inherently difficult to draw actionable, generalisable conclusions on human–AI collaboration in a context of fast technological evolution. This is so for two main reasons. On the one hand, little research has been conducted on what kinds of tasks are better performed by human–AI systems as opposed to humans or AI alone (see, e.g., Dell’Acqua et al. 2023 [13]). On the other hand, studies dedicated to human–AI collaboration usually deploy a variety of AI systems for a variety of tasks, making it hard to identify common trends and usage patterns. For example, recent research suggests that AI systems can act as mediators between humans [40]. However, this possibility has been tested on a specific AI system, making it impossible to draw a general lesson to apply to the larger community of users (e.g., “use AI to mediate disagreements”) unless we provide users with the exact same system that Tessler et al. experimented with.

Secondly, and closely related to the point above, every application of AI systems within organisations needs to be accompanied by robust experimentation, not only in the sense of further experimental research but also of experimentation in specific contexts where these AI systems are intended to be deployed. For example, an accurate analysis of European legislation may require a specific set-up, as the terminology used may differ from the one used in other policy contexts, while a virtual scientific assistant may require specific tuning in the choice of sources, following an organisation's policies on the matter. This approach is indispensable to ensure that these technologies have the expected outcomes in real-life scenarios. In this context, the auditability dimension introduced in the JRC GenAI Compass aims to facilitate the understanding of limitations—and subsequent risks—in deploying AI systems with insufficient ability to evaluate their task-specific performance.

5. Conclusions

After more than a year of experimenting with and building GenAI solutions internally, we begin to see glimpses of the real potential for change. Taking stock of our large-scale applied experimentation efforts, we can see more clearly beyond the hype. We now have a better understanding not only of the value that GenAI can bring but also of the potential risks, that call for intellectual vigilance and rigorous evaluation before implementing plans and change management initiatives.

One of the main lessons we draw from our experience is that quickly deciding to move to ready-made solutions is unlikely to prove a sound long-term strategy. This is especially true for large organisations whose main source of value is the knowledge they produce, manage and disseminate. Successfully implementing GenAI in complex contexts requires good alignment with the strategic objectives of the organisation and rigorous technological development (AI-IQ) accompanied by targeted organisational change efforts that focus on staff upskilling, clear communication, and multipliers through engaged communities of practice. By approaching change holistically, organisations can harness Generative AI's transformative potential to enhance policy insights, operational agility, and organisational resilience.

This is the path taken by the JRC and documented in this paper. We experimented (safely) within a concrete professional setting and closely monitored the uptake patterns of our users to inform our approach and adapt the technology as needed. In other words, we ate the pudding. It is our hope that our reflections can contribute to further research focusing more broadly on the adoption of GenAI by the public sector.

In a way, deploying AI systems in the work environment with an experimental mindset is like testing a new employee under probation. We “hired” a new type of collaborator (in the spirit of Anthony, Bechky, and Fayard 2023 [41]), based on some preliminary knowledge about them (the equivalent of what would be a CV etc for human collaborators), which allowed us to build some expectations of how they were going to perform. Then, in the same way that human collaborators remain largely a “black box” (i.e., we do not “reverse engineer” their thinking processes but rather rely on some widely recognised characteristics such as diplomas or the results of personality tests) when it comes to knowing how they are actually going to perform on the job, we assessed their added value by putting them to work and saw how they integrated with our existing teams and processes (in the spirit of Burton et al. 2024 [42]), bringing their additional performance. And the proof of the pudding was in the eating.

A final concluding remark: as this technology advances it is imperative to keep an open mind, continue experimenting and innovating, and continuously re-evaluate. An organisation that solely focuses on “productising” GenAI might miss the next big opportunity in this rapidly evolving field.

Author Contributions: Conceptualization, B.D.L., I.S. and S.K.; methodology, B.D.L., I.S., S.K. and K.D.; software, I.S., S.L. and F.Z.; validation, B.D.L., I.S. and S.K.; formal analysis, B.D.L., I.S., S.K. and K.D.; investigation, B.D.L., I.S. and S.K.; resources, B.D.L., I.S. and S.K.; data curation, K.D.; writing—original draft preparation, B.D.L., I.S., S.K., S.L., F.Z., K.D. and M.I.; writing—review and editing, B.D.L., I.S., S.K. and M.I.; visualization, B.D.L.; supervision, B.D.L., I.S. and S.K.; project administration, B.D.L., I.S. and S.K.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JRC's institutional budget and received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Declaration of Generative AI in Scientific Writing: During the preparation of this work, the author(s) used the Joint Research Centre's internal platform called GPT@JRC, leveraging several AI models to assist with selected research tasks as described in the manuscript. Additionally, the authors used GPT@JRC to support the writing process, thereby improving the readability and language of the manuscript. The author(s) have carefully reviewed and edited the text as needed across multiple iterations with GPT@JRC and take full responsibility for the published article's content.

References

1. Bostrom, N. *Deep Utopia: Life and Meaning in a Solved World*; Ideapress Publishing: 2024. Available online: <https://books.google.be/books?id=HSeX0AEACAAJ> (accessed on 10 December 2024).
2. Belanche, D.; Belk, R.W.; Casaló, L.V.; Flavián, C. The dark side of Artificial Intelligence in services. *Serv. Ind. J.* **2024**, *44*, 149–172. [CrossRef]
3. Mollick, E.; Mollick, L. Instructors as Innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv* **2024**. [CrossRef]
4. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using Large Language Models. *PLOS Digit. Health* **2023**, *2*, e0000198. [CrossRef]
5. Luo, X.; Recharadt, A.; Sun, G.; Nejad, K.K.; Yáñez, F.; Yilmaz, B.; Lee, K.; Cohen, A.O.; Borghesani, V.; Pashkov, A.; et al. Large Language Models surpass human experts in predicting neuroscience results. *Nat. Hum. Behav.* **2024**, *9*, 305–315. [CrossRef]
6. Vaccaro, M.; Almaatouq, A.; Malone, T. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.* **2024**, *8*, 2293–2303. [CrossRef] [PubMed]
7. Brynjolfsson, E.; Li, D.; Raymond, L. *Generative AI at Work*; National Bureau of Economic Research: Cambridge, MA, USA, 2023; p. w31161. [CrossRef]
8. Noy, S.; Zhang, W. Experimental evidence on the productivity effects of Generative Artificial Intelligence. *Science* **2023**, *381*, 187–192. [CrossRef]
9. Messeri, L.; Crockett, M.J. Artificial Intelligence and illusions of understanding in scientific research. *Nature* **2024**, *627*, 49–58. [CrossRef] [PubMed]
10. Wirtz, J.; Patterson, P.G.; Kunz, W.H.; Gruber, T.; Lu, V.N.; Paluch, S.; Martins, A. Brave new world: Service robots in the frontline. *JOSM* **2018**, *29*, 907–931. [CrossRef]
11. Tyler, C.; Akerlof, K.L.; Allegra, A.; Arnold, Z.; Canino, H.; Doornenbal, M.A.; Goldstein, J.A.; Budtz Pedersen, D.; Sutherland, W.J. AI tools as science policy advisers? The potential and the pitfalls. *Nature* **2023**, *622*, 27–30. [CrossRef]
12. Kreitmeir, D.; Raschky, P.A. The Heterogeneous Productivity Effects of Generative AI. *arXiv* **2024**. [CrossRef]
13. Dell'Acqua, F.; McFowland, E.; Mollick, E.R.; Lifshitz-Assaf, H.; Kellogg, K.; Rajendran, S.; Kraymer, L.; Candelon, F.; Lakhani, K.R. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Working Paper No. 24-013, The Wharton School Research Paper. *SSRN J.* **2023**. [CrossRef]
14. Christensen, C.M.; McDonald, R.; Altman, E.J.; Palmer, J.E. Disruptive Innovation: An Intellectual History and Directions for Future Research. *J. Manag. Stud.* **2018**, *55*, 1043–1078. [CrossRef]
15. Crafts, N. Artificial Intelligence as a general-purpose technology: An historical perspective. *Oxf. Rev. Econ. Policy* **2021**, *37*, 521–536. [CrossRef]
16. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**. [CrossRef]

17. Kanitz, R.; Gonzalez, K.; Briker, R.; Straatmann, T. Augmenting Organizational Change and Strategy Activities: Leveraging Generative Artificial Intelligence. *J. Appl. Behav. Sci.* **2023**, *59*, 345–363. [\[CrossRef\]](#)
18. Huang, K.; Zhang, F.; Li, Y.; Wright, S.; Kidambi, V.; Manral, V. Security and Privacy Concerns in ChatGPT. In *Beyond AI*; Huang, K., Wang, Y., Zhu, F., Chen, X., Xing, C., Eds.; Future of Business and Finance; Springer Nature: Cham, Switzerland, 2023; pp. 297–328. [\[CrossRef\]](#)
19. Robey, A.; Wong, E.; Hassani, H.; Pappas, G.J. Smoothllm: Defending Large Language Models against jailbreaking attacks. *arXiv* **2023**, arXiv:2310.03684.
20. Wang, X.; Yuan, P.; Feng, S.; Li, Y.; Pan, B.; Wang, H.; Hu, Y.; Li, K. CogLM: Tracking Cognitive Development of Large Language Models. *arXiv* **2024**. [\[CrossRef\]](#)
21. Kandpal, N.; Deng, H.; Roberts, A.; Wallace, E.; Raffel, C. Large Language Models Struggle to Learn Long-Tail Knowledge. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; PMLR, 2023. Volume 202, pp. 15696–15707. Available online: <https://proceedings.mlr.press/v202/kandpal23a.html> (accessed on 5 November 2024).
22. Jiang, M.; Karanasios, S.; Breidbach, C. Generative AI In The Wild: An Exploratory Case Study of Knowledge Workers. In Proceedings of the European Conference on Information Systems (ECIS), Paphos, Cyprus, 13–19 June 2024; Available online: https://aisel.aisnet.org/ecis2024/track04_impactai/track04_impactai/7 (accessed on 12 December 2024).
23. Al Naqbi, H.; Bahroun, Z.; Ahmed, V. Enhancing Work Productivity through Generative Artificial Intelligence: A Comprehensive Literature Review. *Sustainability* **2024**, *16*, 1166. [\[CrossRef\]](#)
24. Hagendorff, T. Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. *arXiv* **2024**. [\[CrossRef\]](#)
25. Leavitt, H.J.; March, J.G. *Applied Organizational Change in Industry: Structural, Technological and Humanistic Approaches*; Carnegie Institute of Technology, Graduate School of Industrial Administration: Pittsburgh, PA, USA, 1962; Available online: https://books.google.be/books?id=P_KZNQAACAAJ (accessed on 17 December 2024).
26. Daugherty, P.R.; Wilson, H.J. *Human + Machine: Reimagining Work in the Age of AI*; Harvard Business Press: Boston, MA, USA, 2018.
27. Flavián, C.; Pérez-Rueda, A.; Belanche, D.; Casaló, L.V. Intention to use analytical Artificial Intelligence (AI) in services—The effect of technology readiness and awareness. *JOSM* **2022**, *33*, 293–320. [\[CrossRef\]](#)
28. European Commission, Joint Research Centre. *The Communities of Practice Playbook: A Playbook to Collectively Run and Develop Communities of Practice*; Publications Office: Luxembourg, 2021; Available online: <https://data.europa.eu/doi/10.2760/443810> (accessed on 28 November 2024).
29. Lowe, S.C. System 2 Reasoning Capabilities Are Nigh. *arXiv* **2024**. [\[CrossRef\]](#)
30. Dawid, A.; LeCun, Y. Introduction to latent variable energy-based models: A path toward autonomous machine intelligence. *J. Stat. Mech.* **2024**, *2024*, 104011. [\[CrossRef\]](#)
31. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, Canada, 3–10 March 2021; pp. 610–623.
32. Kalyan, K.S. A survey of GPT-3 family Large Language Models including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2024**, *6*, 100048. [\[CrossRef\]](#)
33. Bobrow, D.G.; Norman, D.A. Some principles of memory schemata. In *Representation and Understanding*; Elsevier: Amsterdam, The Netherlands, 1975; pp. 131–149.
34. Hosseini, A.; Sordoni, A.; Toyama, D.; Courville, A.; Agarwal, R. Not All LLM Reasoners Are Created Equal. *arXiv* **2024**. [\[CrossRef\]](#)
35. Li, Z.; Cao, Y.; Xu, X.; Jiang, J.; Liu, X.; Teo, Y.S.; Lin, S.-W.; Liu, Y. LLMs for Relational Reasoning: How Far are We? In Proceedings of the 1st International Workshop on Large Language Models for Code, Lisbon, Portugal, 20 April 2024; ACM: Lisbon, Portugal, 2024; pp. 119–126. [\[CrossRef\]](#)
36. Händler, T. Balancing Autonomy and Alignment: A Multi-Dimensional Taxonomy for Autonomous LLM-powered Multi-Agent Architectures. *arXiv* **2023**. [\[CrossRef\]](#)
37. Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A Survey on Large Language Model based Autonomous Agents. *arXiv* **2023**. [\[CrossRef\]](#)
38. Chawla, C.; Chatterjee, S.; Gadadinni, S.S.; Verma, P.; Banerjee, S. Agentic AI: The building blocks of sophisticated AI business applications. *J. AI Robot. Workplace Autom.* **2024**, *3*, 196–210. [\[CrossRef\]](#)
39. Bommasani, R.; Liang, P.; Lee, T. Holistic Evaluation of Language Models. *Ann. New York Acad. Sci.* **2023**, *1525*, 140–146. [\[CrossRef\]](#)
40. Tessler, M.H.; Bakker, M.A.; Jarrett, D.; Sheahan, H.; Chadwick, M.J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D.C.; et al. AI can help humans find common ground in democratic deliberation. *Science* **2024**, *386*, eadq2852. [\[CrossRef\]](#)

41. Anthony, C.; Bechky, B.A.; Fayard, A.-L. “Collaborating” with AI: Taking a System View to Explore the Future of Work. *Organ. Sci.* **2023**, *34*, 1672–1694. [\[CrossRef\]](#)
42. Burton, J.W.; Lopez-Lopez, E.; Hechtlinger, S.; Rahwan, Z.; Aeschbach, S.; Bakker, M.A.; Becker, J.A.; Berditchevskaia, A.; Berger, J.; Brinkmann, L.; et al. How Large Language Models can reshape collective intelligence. *Nat. Hum. Behav.* **2024**, *8*, 1643–1655. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Reproduced with permission of copyright owner. Further reproduction
prohibited without permission.